Identification of biomarkers for the management of human prostate cancer

Bogdan-Alexandru Luca

This dissertation is submitted for the degree of

Doctor of Philosophy

University of East Anglia
School of Computing Sciences

May 2017

I would like to dedicate this thesis to my loving family…

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Bogdan-Alexandru Luca
May 2017

Acknowledgements

I am grateful to Vincent Moulton, Professor of Computational Biology, School of Computing Sciences, University of East Anglia, for his guidance and constant support throughout the last four years. I am also grateful to Colin Cooper, Professor of Cancer Genetics, Faculty of Medicine and Health Sciences, University of East Anglia, for his vision and dedication in supervising my research. I am indebted to Dr Dan Brewer, Senior Bioinformatician, Faculty of Medicine and Health Sciences, University of East Anglia for his advice in performing data analysis and help in preparation of this thesis.

I am thankful to Dr Jeremy Clark, Senior Scientist, Cooper Laboratory, Faculty of Medicine and Health Sciences, University of East Anglia for his biological insights into my analyses. I am also indebted to Professor Dylan Edwards, Chair of Cancer Studies, Faculty of Medicine and Health Sciences, University of East Anglia for his help in interpreting the biological results.

I am grateful to Liviu Ciortuz, Associate Professor, Computer Science Department, A.I. Cuza University, for introducing me to bioinformatics and preparing me for the pursuit of this PhD.

I would like to thank all the colleagues and friends who have made my stay in Norwich enjoyable. I am especially grateful to my housemates for all the parties, dinners and good time that we spent together.

Finally, I would like to express my gratitude to my family for their love and support in all these years. Thank you.

Abstract

A critical problem in the clinical management of prostate cancer is that it shows high intra- and inter-tumoural heterogeneity. As a result, accurate prediction of individual cancer behaviour is not achievable at the time of diagnosis, leading to substantial overtreatment. It remains an enigma that, in contrast to other cancers, no molecular biomarkers which define robust subtypes of prostate cancer with distinct clinical outcomes have been discovered.

In the first part of this study, using data from exon microarrays, we developed a novel method that can identify transcriptional alterations within genes. The alterations might be the result of chromosomal rearrangements, such as translocations, and deletions, or of other abnormalities, such as read-through transcription and alternative transcriptional initiation sites. Using data from two independent datasets we identify several candidate alterations that are constantly correlated with the biochemical failure or that are linked to the development of metastasis.

In the second part of the study we illustrate the application of an unsupervised Bayesian procedure, which identifies a subtype of the disease in five prostate cancer transcriptome datasets. Cancers assigned to this subtype (designated DESNT cancers) are characterized by low expression of a core set of 45 genes. For the four datasets with linked PSA failure data following prostatectomy, patients with DESNT cancer exhibited poor outcome relative to other patients ($p = 2.65 \cdot 10^{-5}$, $p = 4.28 \cdot 10^{-5}$, $p = 2.98 \cdot 10^{-8}$ and $p = 1.22 \cdot 10^{-3}$). The DESNT cancers are not linked with the presence of any particular class of genetic mutation, including *ETS* gene status. However, the methylation analysis reveals a possible role of epigenetic changes in the generation of the DESNT subtype. Our results demonstrate the existence of a novel poor prognosis category of human prostate cancer and will assist in the targeting of therapy, helping avoid treatment-associated morbidity in men with indolent disease.

Table of contents

List of figures

List of tables

Abbreviations/Acronyms

ADT	androgen deprivation therapy
AJCC	America Joint Committee on Cancer
APT	Affymetrix Power Tools
AUC	area under the curve
BCR	biochemical recurrence
BH	Benjamini-Hochberg
BPH	benign prostatic hypertrophy
CCP	cell cycle progression
CDC	Centre for Disease Control and Prevention
cDNA	complementary DNA
CI	confidence interval
CNA	copy number alteration
CNV	copy number variation
CRPC	castration resistant prostate cancer
CT	computed tomography
DABG	detected above the background
DHT	dihydrotestosterone
DMR	differentially methylated region
DNA	deoxyribonucleic acid

DRE	digital rectal examination
EAU	European Association of Urology
ECE	extracapsular extension
FDR	false discovery rate
FF	fresh-frozen
FFPE	formalin-fixed paraffin-embedded
FISH	fluorescence *in situ* hybridization
fRMA	frozen RMA
GC	genomic classifier
GEO	Gene Expression Omnibus
GO	Gene Ontology
GOF	goodness of fit
GPS	genomic prostate score
HGPIN	high-grade prostatic intraepithelial neoplasia
HPV	human papillomavirus
HR	hazard ratio
ICGC	International Cancer Genome Consortium
ICR	Institute of Cancer Research
IQR	interquartile range
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kaplan-Meier
LASSO	least absolute shrinkage and selection operator
LDA	latent Dirichlet allocation
lncRNA	long non-coding RNA

LPD	latent process decomposition
LVI	lymphovascular invasion
MAB	maximum androgen blockade
MAD	mean of absolute deviation
MAP	maximum posterior
MDS	multidimensional scaling
miRNA	micro RNA
MLE	maximum likelihood
MRI	magnetic resonance imaging
mRNA	messenger RNA
MSKCC	Memorial Sloan-Kettering Cancer Centre
MVB	marginalised variational Bayes
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
ncRNA	non-coding RNA
NED	no evidence of disease
NHS	National Health System
NICE	National Institute for Health and Care Excellence
ONS	Office for National Statistics
OR	odds ratio
PCA	principal component analysis
PCR	polymerase chain reaction
PH	proportional hazards
PLIER	probe logarithmic intensity error

pre-mRNA	precursor mRNA
PSA	prostate specific antigen
PSM	positive surgical margins
QA	quality assessment
qPCR	quantitative PCR
RACE	rapid amplification of cDNA ends
RCT	randomised clinical trial
RF	random forests
RLE	relative log expression
RMA	robust multiarray analysis
RNA	ribonucleic acid
ROC	receiver operating characteristic
RR	risk ratio
rRNA	ribosomal RNA
RT-PCR	real time PCR
RT-qPCR	real time qPCR
SCNA	somatic CNA
siRNA	small interfering RNA
snRNA	small nuclear RNA
STI	sexually transmitted infection
SVI	seminal vesicle invasion
SYS	systematic progression
TCGA	The Cancer Genome Atlas
TNM	tumour node metastasis

tRNA	transfer RNA
TRUS	transurethral ultrasound
UICC	International Union for Cancer Control
VB	variational Bayes

Chapter 1

Introduction

Prostate cancer is the most common cancer in males [1, 2]. In 2013, over 1.4 million men were diagnosed with prostate cancer worldwide and 293,000 died due to the disease [1]. In England, one in four diagnosed cancers is a prostate cancer [2] (Figure 1.1).

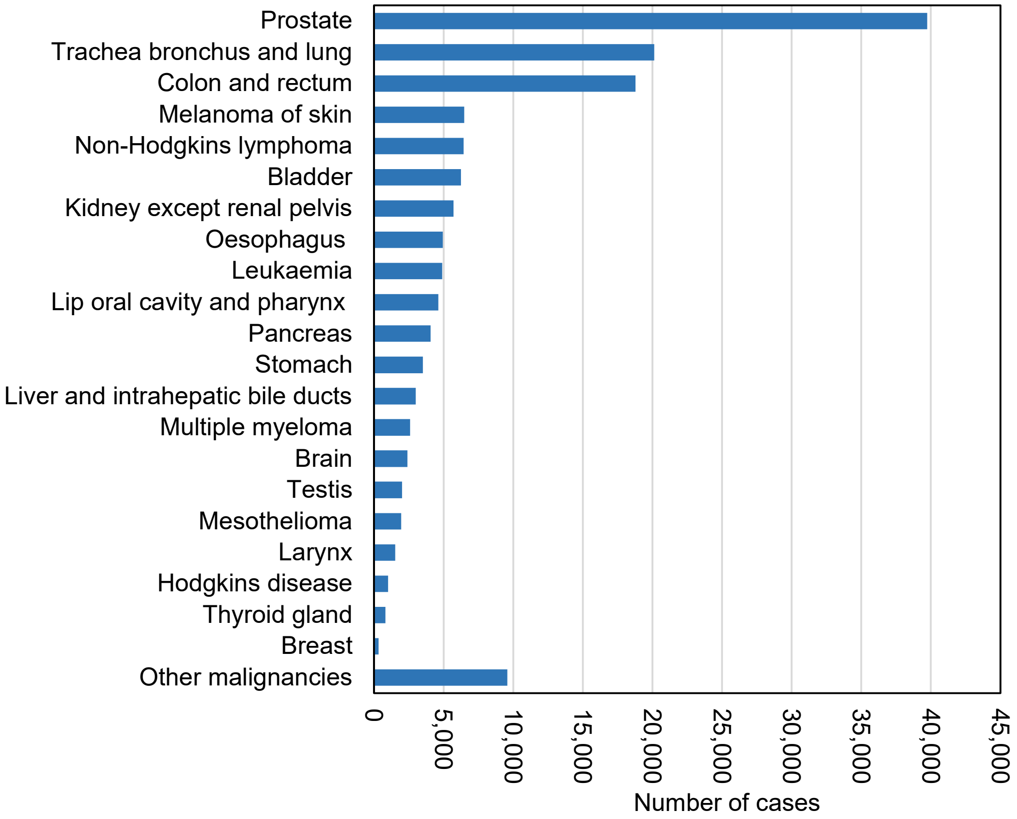

Figure 1.1 The number of male cancers diagnosed in England in 2014. Adapted from ONS [2].

Cancer is a group of conditions characterized by uncontrolled division of cells, which invade and destroy the surrounding healthy tissue. In later stages cancer may spread to other parts of the body, destroying the functions of different organs. A group of modified cells that show abnormal growth is referred to as *tumour*.

The prostate is a gland of the male reproductive system, which is the size of a chestnut, situated underneath the bladder and surrounding the urethra. Its main function is to produce a fluid that is a major constituent of semen [3].

Prostate cancer is a disease usually affecting men that are over 50 years old, with the majority of cases being diagnosed between the ages of 50-80 [4]. Studies have shown that one third of men aged 50 years or more have evidence of prostate cancer [5–10], but in 80% of cases, the disease is low grade and is clinically insignificant [11].

1.1 Biomarkers

According to the National Cancer Institute, a biomarker is *"a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease."* [12]. Cancer biomarkers can be anything that predicts the presence, progression or treatment response of the disease, from proteins and urine markers to genomic alterations, abnormal gene expression and epigenetic modifications [13].

Recent technological developments such as expression microarrays, gene cloning and sequencing technology have led to the discovery of a multitude of new biomarkers with important roles in the management of cancer. For example in leukaemia, genetic alterations are used as biomarkers in the selection of treatment. It has been found that patients with a *PML-RARA* translocation respond to a certain treatment (all-*trans* retinoic acid) while those with a *BCR-ABL1* translocation respond to a different one (imatinib mesylate)[14]. In breast cancer, the Mammaprint test [15] uses the molecular profile of 70 genes to identify patients with high risk of cancer recurrence after the tumour has been surgically removed.

1.2 Prostate cancer biomarkers

In prostate cancer, the detection rates sharply increased with the introduction of the PSA biomarker at the end of 1980s [16]. The PSA test measures the levels of the prostate specific antigen (PSA) serum-marker in peripheral blood. High levels of PSA have been associated with the presence of prostate cancer [17]. However, it has been estimated that PSA screening (testing all males over 50 in the USA) leads to the detection of up to 50% of cancers that are clinically irrelevant [18–20] - that is cancers that would never have caused symptoms in a man's lifetime in the absence of screening.

Prostate cancer shows high inter- and intra-tumour heterogeneity. Many times patients from the same risk category have quite different clinical behaviours [21]. Some patients have indolent forms of the disease, that never require treatment, while others have lethal forms [22]. Even within the same tumour, multiple linages with different properties very often coexist [23]. The intra-tumour heterogeneity can further complicate the molecular profiling of cancer and the identification of distinct molecular subtypes [24].

In contrast to other types of cancer, prostate cancer currently lacks a consistent classification in subtypes that differ in prognosis or treatment response [25]. For example, in breast cancer there are at least three subtypes (*ERBB2* overexpressing, basal and luminal subgroups), with different outcomes [26, 27]. The current risk stratification strategy of early stage prostate cancer [28], based on clinical indicators, cannot accurately predict the outcome of the patients. As accurate prediction of individual prostate cancer behaviour at the time of diagnosis is not possible, immediate radical treatment for all cases has been a common approach.

Radical treatment, which consists of surgical removal of the prostate (prostectomy) or radiotherapy, has a significant risk of complications [29] and, also, has a high impact on quality of life, as most men report urinary incontinence and impotence [30]. Radical treatment of early stage prostate cancer should ideally be targeted to men with significant cancers, so that the remainder, with biologically irrelevant disease, are spared the side-effects of treatment.

Around 35% of men experience recurrence of the disease within 10 years of radical treatment [31]. It is therefore important to also predict the recurrence of the disease as early as possible, so that the patients with risk of recurrence are referred to secondary treatments, increasing chances of survival.

With the advent of new technologies a multitude of recurrent genetic alterations have been identified, but they do not seem to constitute reliable biomarkers for predicting disease outcome. The most common genomic alteration, the *TMPRSS2-ERG* fusion, which occurs in around 50% [32] of cancers is not robustly associated with cancer aggressiveness [33–35]. Some other frequent genomic abnormalities such as *TP53* and *PTEN* alterations have been linked with poor outcome. However, their role as independent predictors has not been proven [36].

Several molecular biomarkers have been also discovered. Noteworthy are the PCA3 gene overexpression in urine, which detects prostate cancer with high sensitivity and specificity [37] and several commercially available gene panels such as, Prolaris [38], OncotypeDx [39] and Dechipher [40], which can predict aggressive prostate cancer. Even though these advancements have the potential to improve the clinical management of the disease and to benefit the patients, so far they have failed to reach a widespread clinical use.

Hence, there remains in the field a need for reliable biomarkers for prostate cancer that could better assist in distinguishing between aggressive cancer, which may require treatment, and non-aggressive cancer, which can be left untreated and spare the patient any side effects from unnecessary interventions.

1.3 Thesis aims

The purpose of the work presented in this thesis is to derive more reliable prostate cancer biomarkers, that could help better stratify patients and distinguish aggressive prostate cancers from indolent cancers. We focused our efforts in two main directions.

In the first part of the thesis we present a novel method that can identify transcriptional abnormalities within genes. We then present several candidate genes for which the transcriptional abnormalities correlate with aggressive prostate cancer.

In the second part of the thesis we describe the analysis that led to the identification of a molecular subtype of prostate cancer, designated DESNT, with poor clinical outcome. We also present the main characteristics of the DESNT cancers. Additionally, we illustrate the derivation of a gene signature with high predictive power for DESNT, which has the potential to be used for a better stratification of prostate cancer in a clinical setting.

1.4 Chapter summaries

We now briefly summarise the chapters presented in the rest of this thesis:

- In **Chapter 2** we introduce the key biological concepts used in this thesis. We also present the current recommendations for managing prostate cancer in the clinic and the exon microarray technology used to measure gene expression.

- In **Chapter 3** we present the computational approaches used in this thesis. We focus in particular on describing the Latent Process Decomposition (LPD) model, which is the basis of the analysis presented in Chapter 5, and the algorithms used to perform survival analysis.

- In **Chapter 4** we describe a novel method that we developed to identify possible transcriptional abnormalities within a gene, using exon microarrays. We describe how this approach works, how it was set up and how it compares with previous approaches. We also present the results obtained by applying the method on three prostate cancer datasets. We further focus on several candidate genes, for which the transcriptional alterations are correlated with clinical outcome of patients.
- In **Chapter 5** we apply the LPD algorithm on five prostate cancer datasets and identify a molecular subtype of cancer with poor prognosis, denoted DESNT. We then describe the analysis that led to the definition of a core set of genes that characterise the DESNT cancers. In the second part of the chapter we present the derivation of a 20 gene signature that can predict DESNT membership. Finally, we correlate the DESNT cancers with mutational and methylation data.
- In **Chapter 6** we conclude our findings and discuss several possible future directions for this research area.

Chapter 2

Biomedical background

2.1 Summary

In this chapter we present an overview of the main biological concepts and medical approaches relevant for the management and research of prostate cancer. We briefly describe the central dogma of molecular biology, to introduce some basic concepts used throughout the thesis such as exon, gene and gene expression. We then present the general characteristics of cancer and the main types of mutations associated with cancer. Next, we describe the current understanding and management of prostate cancer and also present the emerging biomarkers, that try to address some of the current challenges in the field. Besides this, we briefly describe some biological and bioinformatical approaches used in the study of prostate cancer, such as the use of microarrays.

2.2 Genes and gene expression

As presented in Kuriyan et al. [41], organisms store genetic information in DNA (deoxyribonucleic acid), a helicoidal, double-stranded macromolecule of nucleic acids. The central dogma of molecular biology [42] states that the information in DNA is *transcribed* into RNA (ribonucleic acid), which is then *translated* into proteins, that make up the structural and functional elements of the cell (Figure 2.1).

A *gene* is the unit of heredity, made of DNA and regulatory elements, that encodes a protein or a functional non-coding RNA (ncRNA) molecule [44]. In eukaryotes, the protein-coding genes are transcribed into *pre-mRNA* (precursor messenger RNA). The pre-mRNA is then processed in a mechanism named *RNA splicing*, in which some portions of the molecule, known as *introns*, are cleaved out, and the remaining portions, known as *exons* are ligated together, resulting in *mRNA* (messenger RNA). *Alternative*

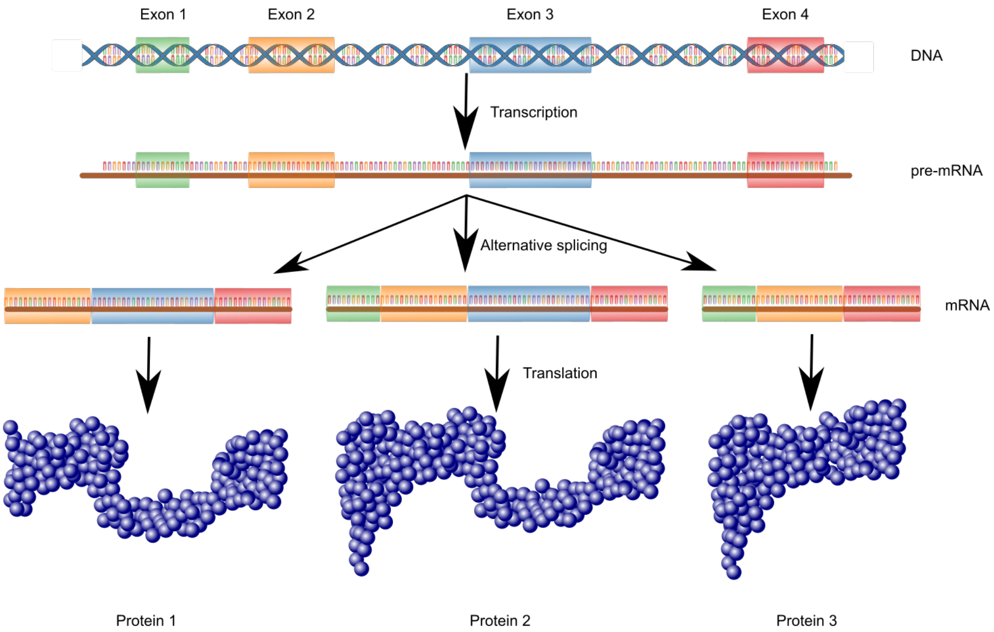

Figure 2.1 Central dogma of molecular biology (adapted from Wikipedia [43]).

splicing is a process that can result in the selective inclusion of some exons, as well as in the modification of the transcription start and/or end locus, allowing the creation of several different mRNAs from the same pre-mRNA. The different versions of mRNA created via alternative splicing are referred to as *isoforms*. The mRNA nucleotide sequence is then translated into the amino acids sequence of a protein via a mapping known as *genetic code*, which associates three consecutive nucleotides to a *codon*, that is specific to a given amino acid. Notably different splice isoforms may encode proteins with different sequences and distinct functions.

The non-coding RNA molecules, on the other hand, are not translated into proteins, but have important functional roles in the cell. They are roughly divided into two categories based on size: small ncRNAs (<200 base-pairs) and long ncRNAs (>200 base-pairs) [45]. Small ncRNAs include transfer RNA (tRNA) and ribosomal RNA (rRNA), essential for the cell machinery, micro RNA (miRNA), small nuclear RNA (snRNA) and small interfering RNA (siRNA), with roles in gene regulation [45]. The long ncRNAs (lncRNA) are divided into intergenic and intragenic lncRNA. The lncRNAs are mainly involved in transcription regulation [45], and their aberrant activity has been linked with various diseases, including prostate cancer [46].

The process of synthesising a functional gene product (proteins or ncRNAs) using the information encoded in a gene is usually referred to as *gene expression*, and the amount of RNA transcribed as *gene expression level*. The mechanisms that control gene expression in a cell are very complex and are not completely understood. The expression level change dynamically in time, as it depends on many factors such as tissue type, external stimuli or internal needs.

2.3 Cancer

Cancer is characterised by uncontrolled cell division. The progression from normal cells to cancer cells in known as tumorigenesis and is a multistage process, during which a significant number of redundant cell control mechanisms are disabled or bypassed.

Hanahan and Weinberg [47] described six essential capabilities that cancer cells need to acquire in order to multiply and spread: self-sufficiency in growth signals, insensitivity to growth-inhibitory (anti-growth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis (growth of blood vessels), and tissue invasion and metastasis. In a newer version of the study [48], they add two more capabilities: reprogramming of energy metabolism and evading immune destruction.

Firstly, cancer cells need to be able to proliferate independent of external growth signals, a set of complex extra-cellular stimuli without which normal cells do not grow; and also they need to become unresponsive to growth suppressors. Another key capability is the ability to avoid programmed cell death (also known as apoptosis), that occurs in normal cells due to either external stimuli, or if internal abnormalities, such as DNA damage, are detected. Besides these capabilities, the cancer cells gain a survival advantage when they are able to evade the senescence mechanism, an independent machinery that controls the maximum numbers of divisions a cell can undergo.

Once the tumour starts to grow in size, it needs to develop a system of blood vessels, to supply the cells with oxygen and nutrients and to evacuate the waste. Usually tumours hijack the mechanism that is responsible for angiogenesis (the sprouting of new vessels from existing ones) in normal cells to supply this need.

Most of the cancers will invade the surrounding sites and later move to more distant locations, developing metastasis, which is the main cause of cancer-related death. Metastasis is the result of a series of stages starting with local invasion of the surrounding tissue, then the transfer of cancer cells into blood and lymph, which transports them to distant sites where they form new settlements.

Cancers adjust also the energy metabolism mechanisms to fuel faster growth. Furthermore, it is currently thought that the immune system constantly monitors cells and eliminates the incipient tumours if irregularities are observed. Therefore, in order for the cancer to develop it needs to evade the immune system. The exact mechanisms to do this are not well understood.

2.4 Genetic and epigenetic alterations in cancer

There are multiple ways in which the function of key regulatory genes are altered, in the evolution of cancer. The most common abnormalities are *genetic alterations*, such as *point mutations*, *indels*, *structural* and *numerical alterations* of chromosomes, and *epigenetic alterations* such as *DNA methylation*.

The multitude of combinations in which these alterations can occur and the genes they target account for the heterogeneity observed in cancers. However, there have been identified a significant number of alterations which are recurrent in some diseases. Their discovery lead to an improved understanding of the disease, definition of genetic subtypes (often with different clinical outcomes), the development of new diagnostic tests and therapeutic targets.

The genetic alterations can be classified into two main categories, based on the effect on chromosome structure: small-scale mutations and chromosomal abnormalities [49].

2.4.1 Small-scale mutations

Small-scale mutations are genetic alterations that affect one or several nucleotides within a gene, and include point mutations (the substitution of a DNA nucleotide with another) and indels (the deletion or the insertion of DNA nucleotides) [49]. They are the result of either exogenous factors (chemicals, ultraviolet light, radiations etc.) or endogenous factors (such as mitotic errors or errors in DNA repair) [50].

If the mutation takes place within a protein coding gene, then the protein encoded might have different properties, leading to its malfunction. Alternatively the mutation might lead to a truncation of the protein encoded. Mutations can also affect the functionality of non-coding genes, leading to expression aberrations and promoting tumorigenesis [51].

It is estimated that most cancer harbour between 1,000 and 20,000 point mutations and indels [50]. Some of the most relevant mutated genes in cancer are *TP53* (36.1%), *PIK3CA* (14.3%), and *BRAF* (10%) [50]. *TP53* is a tumours suppressor gene, while *PIK3CA* and *BRAF* are involved in cell growth, therefore all play an important role in the development of cancer.

2.4.2 Chromosomal abnormalities

Chromosomal abnormalities are a complex class of alterations that result in changes in the structure or number of chromosomes. They have been originally thought to be specific to blood cancers and leukaemia, but in the recent years they have been also found in almost all major solid tumours.

Fröhling and Döhner [52] identify two main categories of chromosomal abnormalities: *balanced chromosomal rearrangements* and *chromosomal imbalances*.

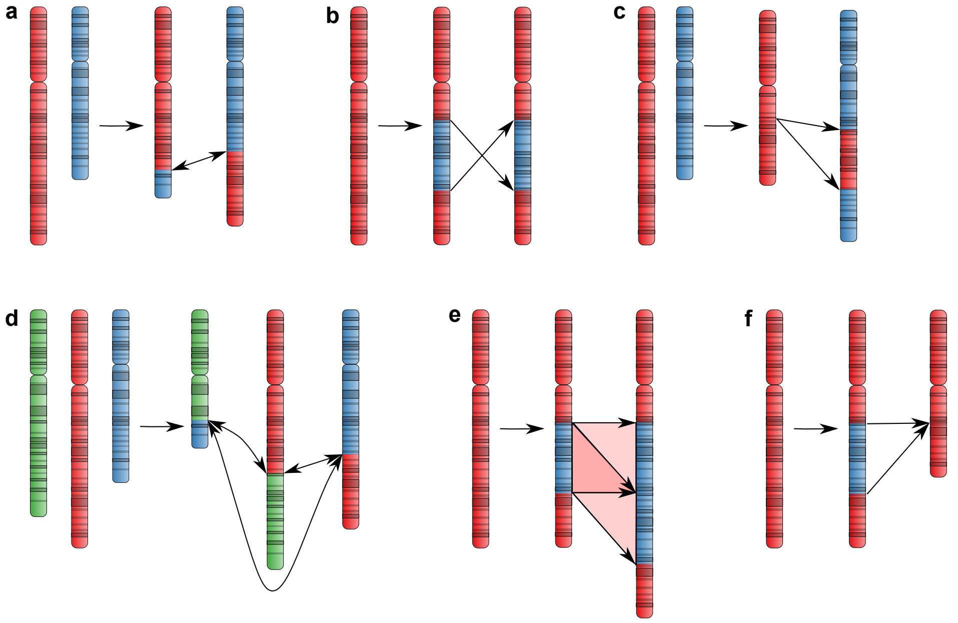

Figure 2.2 Chromsomal abnormalities: a) reciprocal translocation, b) inversion, c) insertion, d) chromoplexy, e) duplication, f) deletion.

Balanced chromosomal rearrangements are those alterations that result in the modifications in the structure of the chromosomes, but do not result in an increase or decrease in the number of copies of a gene. The typical balanced chromosomal rearrangements are *reciprocal translocations* (two chromosomes exchange a portion), *inversions* (a portion of a chromosome is inverted) and *insertions* (a portion of a chromosome is inserted into another chromosome) (Figure 2.2a-c). In recent years a new complex chromosomal rearrangement, referred to as *chromoplexy* has been identified in prostate cancer [53, 54] (Figure 2.2d). Chromoplexy refers to a chain of translocations and deletions that involve simultaneously several genomic intra and inter-chromosomal locations.

Chromosomal imbalances are those abnormalities that result in gains or losses of genetic material. The amount of genetic material can range from entire chromosomes to chromosome arms, portions of chromosome, to intragenic deletions or duplications [52] (Figure 2.2e,f).

Sometimes, the breakpoints of chromosomes involved in rearrangements are within or in the proximity of genes. This might result in the formation of *gene fusions*, in the deregulation of the expression of a structurally normal gene, due to its translocation close to the promoter of another gene or to the truncation of the gene transcript. A gene fusion is the juxtaposition of parts of two distinct genes, leading to the formation of a chimeric gene with altered or new functionality [52].

One of the first gene fusions discovered was the Philadelphia chromosome [55], the fusion between the *BCR* gene, located on chromosome 22, and the *ABL1* gene, located on the chromosome 9. The *BCR-ABL1* fusion is the result of a reciprocal translocation and is found in almost all chronic myeloid leukaemia (CML - a type of leukaemia) cases. The protein encoded by *ABL1* is involved in the cell division process and under normal conditions its activity is carefully regulated. When fused to *BCR*, the *ABL1* function is preserved, but the resulting hybrid protein is not responsive to control mechanisms anymore, leading to uncontrolled cell proliferation [56].

Genomic losses or gains can also have a significant contribution to tumorigenesis. For example, the amplification of the part of chromosome 17, which occurs in 30% of the breast cancers, leads to the up-regulation of the *ERBB2* gene, which, in turn, results in an increased proliferation of cancer cells [52]. Or, the deletions occurring in chromosome 10 leads to loss of the *PTEN* tumour suppressor gene. *PTEN* loss leads to deregulation in the PIK3/Akt pathway, which has an important role in controlling, among other things, cell growth, cell proliferation and apoptosis [57].

2.4.3 DNA methylation

There is another class of processes that alters the activity of genes without changing the DNA sequence, called epigenetic modifications, which are essential to many organism functions. However, if they occur improperly, they can contribute to the promotion of various diseases, such as cancer [58]. There are at least three known epigenetic processes: DNA methylation, histone modification, and RNA silencing [59].

Here we focus on DNA methylation, which is a process that adds a methyl (CH_3) group to DNA in regions known as CpG sites (regions where a cytosine nucleotide is followed by a guanine). Usually the CpG sites are more dense around the promoter of genes (these regions are called CpG islands), leading to interference with the transcription mechanisms [59].

The promoters of genes have in normal conditions a certain level of methylation. Both hypomethylation (a lower level of methylation than normal) and hypermethylation (a higher level of methylation than normal) can lead to significant changes in expression. In breast cancer, for example, the hypomethylation of the *SATR1* gene has been linked with the early stages development of the tumour [60]. Conversely hypermethylation has been proven to have important roles in tumour progressions; hypermethylation of the *CDH13* promoter leads to progression of non-small cell lung cancers [61].

2.5 Prostate cancer

2.5.1 The prostate

The prostate is a fibromuscular and a glandular organ which weighs about 18 grams and measures approximately 3 cm in length, 4 cm in width in 2 cm in depth [3]. It is located underneath the bladder and surrounds the lower part of the posterior urethra (Figure 2.3a). Prostate is composed of around 70% glandular elements and 30% fibromuscular stroma and is surrounded by a capsule composed of collagen, elastin and smooth muscle [3].

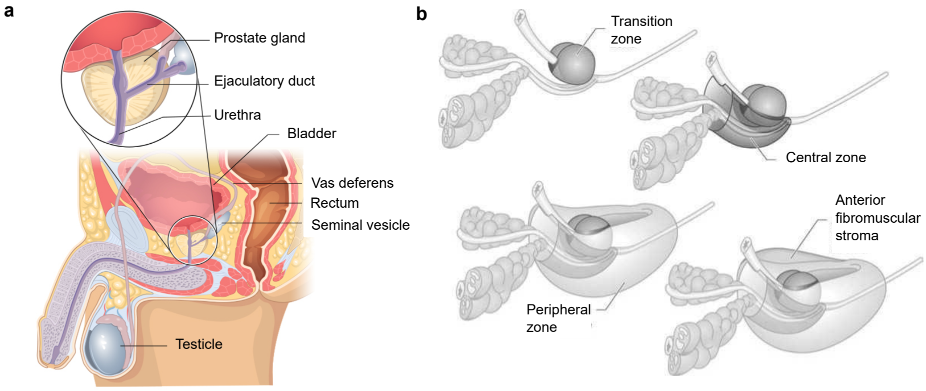

Figure 2.3 The prostate gland: a) the location of prostate (adapted from CDC [62]); b) the zonal architecture of prostate (adapted from Aibolita [63])

The prostate is part of the male reproductive system. It along with the seminal vesicles, ampullae, Cowper's gland and glands of Littre form the sex accessory tissues. The sex accessory tissues are responsible for producing seminal plasma, which is the major component of the ejaculate. The prostate is contributing to 0.5 mL to the total volume of ejaculate (of about 3 mL). The prostate secretions are rich in citric acid, polyamines and zinc, along with some secretory proteins such as PSA [64].

The prostate is divided into four zones: *transitional zone*, *central zone*, *peripheral zone* and *anterior fibromuscular stroma* (Figure 2.3b). The transitional zone accounts for 5-10% of the glandular tissue. It is believed that 20% of prostate cancers originate in this area. The central zone, which constitutes 25% of the glandular tissue of prostate, is structurally different from the rest of the prostate tissue. Only 1-5% of prostate cancers originate in this area. The peripheral zone, which makes up around 70% of the prostate glands is where 70% of cancers develop. The anterior fibromuscular stroma makes up to a third of the prostate. This zone is rarely involved in prostate cancer [3].

Also, in clinical practice, prostate is regarded as having two lateral lobes, separated by a central sulcus and a middle lobe. This division does not correspond to anatomical components of the prostate [3].

2.5.2 Risk factors

There are several established risk factors associated with the development of prostate cancer including age, race/ethnicity and positive family history of prostate cancer [65, 66]. Other factors that have been reported as having some association with the risk of prostate cancer are consumption of lycopene, consumption of milk protein, sexually transmitted diseases, infections, smoking, obesity and environmental factors [65].

The risk of prostate cancer increases sharply in men after 50 years, with the majority of cases being diagnosed between ages 50-79 [4]. In the US, Leitzmann and Rohrmann [65] reports an incidence of about 9 case per 100,000 per year in men aged 40-44 years, while for men aged 70-74 years the incidence increases sharply to 985 cases per 100,000 men per year. Also, autopsy studies performed in various locations on different races and ethnicities reported a prevalence of prostate cancer of 25%-40% in men over 70 years and 33%-87% in men over 80 years old [5–10].

Afro-American and Afro-Caribbean men have a higher risk than Caucasians of developing prostate cancer, and tend to have worse prognosis and higher chance of recurrence following prostectomy [67–69]. Asian men, on the other hand, tend to have lower incidence and mortality rates than Asian-Americans, which in turn have lower rates than Caucasians [70]. The difference between Asian-Americans and Asians might be explained by environmnetal factors, particularly Western diet [70].

Family history of prostate cancer increases the risk. Men with a first degree relative diagnosed with prostate cancer have a 2.5 times higher chance of developing the diseases themselves [71]. The risk increases with the number of first degree relatives with prostate cancer and can reach an odds ratio of 17.7 for men with three brothers diagnosed with the disease [72].

Sexually transmitted infections (STIs) have not been consistently associated with the risk of prostate cancer. Some studies reported elevated risk in patients with history of STIs, especially syphilis, gonorrhoea and human papillomavirus (HPV) [73, 74]. However, they are mainly case-control retrospective studies that are susceptible to recall biases [75]. Subsequent large-cohort, retrospective and prospective studies failed to identify associations between the most common STIs and the risk of prostate cancer [75–78].

There is no commonly accepted consensus about the effects of dietary factors on the development prostate cancers. A number of food categories such as vegetables, meat, fish and diary products have been reported as either increasing or decreasing the risk of prostate cancer [65]. However the results are heterogeneous and the results have not always been supported by independent studies.

One of the most notable dietary factors that might play a role in prostate cancers is the intake of lycopene. The main source of lycopene are tomatoes, watermelons and pink grapefruits [79]. Several studies reported an inverse correlation between the intake of lycopene and prostate cancer risk, progression and mortality, but others have not found any association [79]. Noteworthy is also the inverse correlation between the intake of cruciferous vegetables, such as broccoli, and the risk of developing prostate cancer [80].

2.5.3 Screening and early detection

Screening means the application of the test diagnostics to patients which are at risk of developing a disease, but which do not present symptoms, in order to identify the disease at an early stage, and therefore to reduce the disease-specific mortality rate, while trying to have as little impact as possible on the patient's quality of life [81].

In the context of prostate cancer, the detection rates sharply increased with the introduction of the PSA screenings at the end of 1980s. In the United States, for example, the incidence of prostate cancer increased 12% per year from 1986, the year before the PSA was reported, until 1992, when it peaked [16]. The PSA test measures the levels of the prostate specific antigen (PSA) serum-marker in peripheral blood. High levels of PSA have been associated with the presence of prostate cancer [17].

PSA screening generates a great deal of debate on its benefits compared to its disadvantages. Currently there is no consistent evidence that PSA screening reduces the cancer-specific mortality rates. The randomised study of Schröder et al. [82], performed on 162,387 men, identified a reduction with 20% in the rate of death for patients screened. On the other hand, the meta-analysis of Ilic et al. [83] of five randomised clinical trials (RCT), comprising of a total of 341,342 participants found no survival difference between the patients who received screening and those who did not (risk ratio (RR) 1.00, 95% confidence interval (CI) 0.86 - 1.17). Moreover the subgroup analysis found that the mortality was not influenced by the age at which screening was given. Ilic et al. [83] also reported a higher proportion of localised prostate cancers in the patients that received screening (RR 1.79, 95% CI 1.19 - 2.70) and a lower proportion of advanced cancers (RR 0.80, 95% CI 0.73 - 0.87).

PSA screenings do lead to significant overdiagnosis and overtreatment [81, 83]. Overdiagnosis refers to the detection of indolent disease, that would have not shown any clinical symptoms during the lifetime of the patient. It has been estimated that the PSA screening led to an overdiagnosis rate of around 50% [18–20]. The immediate adverse effects of the overdiagnosis are the anxiety associated with the diagnosis and the possible complications due to invasive tests, such as biopsy.

Also, many times the overdiagnosis leads to unnecessary treatment. Besides the quality of life impact the overtreatment has, there is a a significant risk of treatment-related complications. Around 10-15% of the patients who undergo radical prostectomy report urinary incontinence and about 70% have erectile problems [30]. There have been also reported surgery-related complications such as infections, respiratory and cardiac problems in about 20-25% of patients who underwent surgery, with about 0.5% of patients dying due to surgery [84, 85].

In the light of the current evidence the European Association of Urology (EAU) guidelines on prostate cancer [81] recommended against population-wide screening for prostate cancer. Instead a risk-adapted strategy for screening men at risk, that have a life expectancy of more than 10-15 years has been devised. Broadly the main categories of risk identified are men over 50 years old, men over 45 years old with family history of prostate cancer and Afro-American and Afro-Caribbean men [81]. However, these strategies are also prone to overdiagnosis and overtreatment.

2.5.4 Diagnosis

Patients suspected of having prostate cancer are first referred for a prostate specific antigen (PSA) test and a digital rectal examination (DRE).

PSA as an indicator for prostate cancer that lacks both sensitivity and specificity. Many patients with prostate cancer have low values of PSA [86]. On the other hand, high levels of PSA might also be associated with benign prostatic hypertrophy (BPH), prostatis or other non-malignant conditions.

Most cancers are located in the peripheral zone, and can be detected by DRE if the tumour is larger than 0.2 mL [81]. DRE can complement the PSA screening, as it has been reported that in about 18% cases prostate cancer was indicated solely by DRE, irrespective of PSA levels [87].

The EAU and NICE (National Institute for Health and Care Excellence) guidelines on prostate cancer recommend that the definitive diagnosis should be made based on a needle biopsy [81, 88]. NICE recommends that the decision of referring a patient to biopsy should be made based on the results of the PSA screeing, DRE results, risk factors (age, race, family history), therapeutic risks, general health. etc [88].

The standard biopsy is a transrectal ultrasound (TRUS)-guided biopsy which collects 10-12 cores [81]. The TRUS-guided biopsies might miss 20-30% of the clinically relevant cancers [89]. If the first biopsy is negative, but the PSA levels remain high or other clinical factors still indicate the possibility of having prostate cancer, a repeat biopsy might be performed [81]. NICE recommends that the second biopsy is performed if the multiparametric magnetic resonance imaging (MRI) comes back positive. It is recommended that the second biopsy is a template biopsy, aiming to collect more than 20 cores [81], for a more accurate sampling.

2.5.5 Classification criteria

Once the prostate cancer is diagnosed, several clinical factors are evaluated in order to decide the most suitable way of managing the disease. Besides the PSA and DRE results, CT scan and multiparametric MRI can be used to determine the extent of the disease [81]. A management strategy of the disease is developed by assessing the PSA levels, DRE results, Gleason score, TNM stage (clinical stage), and several other pathological features.

2.5.5.1 Gleason score

Gleason score [90, 91] is a grading system based on the architectural patterns of the tumour and is one of the most important prognostic indicators available for prostate cancer. Gleason score is calculated as the sum of two grades, each one taking values between 1 and 5. The primary grade is assigned to the most common tumour pattern. The secondary grade is assigned to the second most common pattern, with the condition that it is present in at least 5% of the total patterns. If the condition is not met, then the primary grade is doubled. Grade 1 is assigned to well-differentiated tissue, that resembles the normal tissue, while grade 5 is assigned to poorly-differentiated tissue, that is very dissimilar to normal tissue.

Gleason sum can range between 2 and 10. A higher Gleason sum is associated with a poorer outcome. Within the Gleason sum 7 category, score 4 + 3 (primary grade 4 and secondary grade 3) has a significantly worse outcome than score 3 + 4 [92–95].

2.5.5.2 TNM stage

TNM (Tumour Node Metastasis) classification [96] is the standard system for staging malignant tumours, maintained by the AJCC (American Joint Committee on Cancer) and UICC (International Union for Cancer Control). As described in Leslie et al. [96], it comprises of three components:

- **T**: describes the spread of the primary tumour;
- **N**: describes the presence/absence and the extent of regional lymph node metastasis;
- **M**: describes the presence/absence of distant metastasis.

For prostate cancer, the TNM classification is made as follows [96]:

T - primary tumour:

TX - primary tumour could not be assessed;

T0 - no evidence of primary tumour;

T1 - clinically inapparent tumour, neither palpable nor visible by imaging:

T1a - tumour incidental histological finding in 5% or less of tissue resected;

T1b - tumour incidental histological finding in more than 5% of tissue resected;

T1c - tumour identified by needle biopsy;

T2 - tumour confined within prostate:

T2a - tumour involves one-half of one lobe or less;

T2b - tumour involves more than one-half of one lobe, but not both lobes;

T2c - tumour involves both lobes;

T3 - tumour extends through the prostatic capsule:

T3a - extracapsular extension;

T3b - tumour invades seminal vesicle(s);

T4 - tumour is fixed or invades adjacent structures other than seminal vesicles: external sphincter, rectum, levator muscles, and/or pelvic wall;

N - regional lymph nodes:

NX - regional lymph nodes cannot be assessed;

N0 - no regional lymph node metastasis;

N1 - regional lymph node metastasis;

M - distant metastasis:

M0 - no distant metastasis;

M1 - distant metastasis:

M1a - non-regional lymph nodes;

M1b - bones;

M1c - other sites.

2.5.5.3 Clinical vs. pathological classification

The extent of cancer is evaluated at diagnosis, using clinical indicators, such as Gleason grade and TNM stage. However, due to sampling errors, biopsies might miss some cancer foci, leading to a suboptimal calculation of the Gleason grade. Also, CT scans and other technologies used to measure the spread of tumour might underestimate the extension of cancer.

Sometimes a more accurate classification is obtained following a radical prostectomy, when the whole prostate can be analysed by a pathologist. This might yield different classification than the one obtained previously. In order to distinguish between these two situations, TNM stage calculated at diagnosis, or before the prostectomy is usually referred to as *clinical stage*, while the TNM stage evaluated after prostectomy is referred to as *pathological stage*. For Gleason grade, the score evaluated after prostectomy is referred to as *pathological Gleason grade/score*.

2.5.5.4 Additional pathological features

Following radical prostectomy, the resected prostate and the surrounding tissue is analysed in order to determine the extent of cancer. Besides the classical pathological

Gleason score and pathological stage, several additional criteria associated with clinical outcome are evaluated. These criteria are: *seminal vesicle invasion*, *lymphovascular involvement*, *extracapsular extension*, and *positive surgical margins*.

2.5.5.4.1 Seminal vesicle invasion (SVI) *Seminal vesicle invasion* is the spread of cancer to the muscular wall surrounding the seminal vesicles [97]. This pathological feature is assessed at prostectomy and is associated with poor prognosis [97]. It has been found that SVI positive patients have a 7-year survival rate of 32.2%, compared with 66.6% in SVI negative patients.

2.5.5.4.2 Lymphovascular invasion (LVI) Prostate cancer spreads through lymphatic channels. Therefore the microscopic analysis performed to detect lymphovascular invasion is a widely used procedure in analysing radical prostectomy specimens [98]. In a study on 1709 men who underwent radical prostectomy [99], LVI has been reported in 7% of cases. LVI is a significant clinical predictor for recurrence, tumour grade, volume and several other pathological features [99]. Biochemical progression for men with LVI was around 34%, compared to 10% of men without.

2.5.5.4.3 Extracapsular extension (ECE) Extracapsular extension is the spread of prostate cancer in the tissue surrounding the prostate. There have been reported several ways of subgrouping patients with extracapsular extension [100–103]. One of the most commonly used is the classification of Epstein et al. [100], which divide ECE into two categories: focal (the cancer spread to a lesser extent outside the prostate) and established (a more extensive penetration). The patients with established extension seem to have a higher risk of progression. In patients without seminal vesicles invasion or lymph node involvement, 5 years after radical prostectomy, the progression-free rate is 87% for the men without capsular extension, 73% for the patients with focal extension and 42% for the patients with established extension [102].

2.5.5.4.4 Positive surgical margins (PSM) At prostectomy, some adjacent tissue surrounding the prostate is removed as well. Positive surgical margins refers to the detection of cancer cells on the surface of the removed tissue [104]. About a third of prostectomies have positive surgical margins [105]. Progression free rate of patients with positive surgical margins is 58-64%, while for the patients with negative margins the rate is 81-83% [105].

2.5.5.5 ICGC risk stratification after prostectomy

In this thesis we will use a stratification of prostate cancer patients who undergo prostectomy into three risk groups of progression, based on the UK International Cancer Genome Consortium (ICGC) consensus (Professor Chris Foster, personal communication). The criteria are presented in Table 2.1.

Table 2.1 ICGC risk stratification of prostate cancer after prostectomy.

Low risk	PSA <= 10ng/ml AND (Gleason = 3+3 OR (Gleason = 3+4 AND no extra capsular extension))
Medium risk	10ng/ml < PSA <= 20ng/ml OR (Gleason = 4+3 AND no extra capsular extension) OR (Gleason = 3+4 AND extra capsular extension)
High risk	PSA > 20ng/ml OR Gleason sum > 7 OR (Gleason = 4+3 AND extra capsular extension) OR Seminal vesicle invasion

2.5.6 Localised prostate cancer

Patients with localised prostate cancer (clinical stage T1/T2) are stratified at diagnosis into risk categories using D'Amico stratification [28] presented in Table 2.2.

The low risk men usually do not receive immediate treatment, but rather they are enrolled to active surveillance or watchful waiting programmes. If the disease seems to progress, patients can receive radical treatments with the intent of cure. The most common radical treatments are prostectomy, radiotherapy or brachytherapy, which are discussed below. The intermediate or high risk patients are usually referred for radical treatments.

Table 2.2 D'Amico risk stratification for men with localised prostate cancer [28].

Level of risk	PSA		Gleason score		Clinical stage
Low risk	$<$10 ng/ml	and	$\leq$6	and	T1–T2a
Intermediate risk	10–20 ng/ml	or	7	or	T2b
High risk	$>$20 ng/ml	or	8–10	or	$\geq$T2c

2.5.6.1 Active surveillance and watchful waiting

Active surveillance aims to avoid or at least delay the treatment for the patients with low-risk prostate cancer, with the purpose of reducing overtreatment, without influencing the cancer-specific survival. Instead of receiving treatment, the low-risk patients are closely monitored. PSA and DRE screenings are performed at regular intervals (every 3 months in the first 2 years and every 6 months afterwards), and repeat biopsies are done every 6-18 months [106, 107]. If there is indication of the cancer progression, such as PSA doubling time less than 2-3 years, histological progression (Gleason $\geq$ 4+3 at repeat biopsies), clinical progression [106, 107], or at patient's request, the active surveillance is interrupted and treatment is given. Most patients who leave active surveillance receive curative intended treatment such as radical prostectomy, radiotherapy and brachytherapy.

The outcome of men initially managed with active surveillance is good in general. It is estimated that the 10-year prostate cancer specific survival is greater than 96% [108–110].

Watchful waiting is usually a strategy for patients with a life expectancy of less than 10 years that, due to general health conditions, might not be fit for receiving radical treatment. Their disease is regularly monitored with PSA screenings and digital rectal examinations. If there is sign of progression they are offered hormone therapy.

2.5.6.2 Prostectomy, radiotherapy and brachytherapy

For the men with localised prostate cancer, there are several treatment options available, offered with curative intent. The most commonly used in clinical practice are radical prostectomy, radiotherapy and brachytherapy. Besides them, the EAU guidelines also suggest new alternative treatments such as cryosurgery (the use of freezing techniques to induce cell death) and high intensity focused ultrasound [81].

Radical prostectomy refers to the surgical removal of the prostate gland, seminal vesicles and a portion of the surrounding tissue, in order to obtain a negative margin [81]. It is mostly suitable for patients with localised prostate cancer and a higher life expectancy.

Patients who have undergone radical prostectomy have a relatively good outcome. The proportion of patients free from cancer progression at 5, 10 and 15 years after prostectomy has been estimated at 82-84%, 74-77% and 66-75% [111, 112]. Also, the cancer-specific survival at 5, 10 and 15 years was 99%, 95-96% and 89-90% [111, 112]. Unfortunately a significant proportion of the patients that undergo prostectomy experience permanent erectile dysfunction and urinary incontinence.

External beam radiotherapy, followed by androgen deprivation therapy is an alternative way of treating patients with localised prostate cancer, especially the patients that are less suitable for surgery. Radiotherapy can still result in urinary incontinence and impotence too. Besides that, the patients experience radiotherapy side-effects such as fatigue, irritation, nausea, etc.

Currently there are no large randomized clinical trials comparing the efficiency of radical prostectomy and radiotherapy, but the data from 16 retrospective studies and one small randomized study suggests that patients with high-risk prostate cancer treated with radical prostectomy have a higher prostate cancer specific survival than those treated with radiotherapy [113]. Sun et al. [114] also reported higher overall survival in patients with a life expectancy of $\geq$10 years, treated with radical prostectomy, relative to radiotherapy. For the patients with <10 years life expectancy, the survival rates seem comparable [114].

Another treatment option for patients with early stage localised prostate cancer is brachytherapy. Brachytherapy is a treatment in which radioactive seeds are implanted into the prostate. This therapy is recommended for low risk patients. Brachytherapy can also be used in combination with external beam radiotherapy to maximise the treatment efficacy.

Because brachytherapy is a minimal invasive technique, the side effects associated with surgery are minimised. Besides that, it has been reported that men who receive brachytherapy have significantly less urinary problems compared to men who undergo prostectomy [115]. Also about 80% of patients who receive report satisfactory erectile function, compared to 50% of the men who receive prostectomy [115]. There is little evidence, however, about the efficacy of brachytherapy relative to the other treatment options [116].

2.5.6.3 Biochemical recurrence (BCR)

After radical treatment, the PSA level of patients is constantly monitored. An increase in PSA levels following radical treatment indicates biochemical recurrence of the disease. In general, a patient is considered to have BCR if two consecutive PSA measures, yield values above 0.2 ng/mL [81]. Around 35% of men who undergo radical treatment experience BCR within 10 years time [31].

The biochemical recurrence precedes metastasis by around seven years and prostate-cancers specific death by 15 years [117] and is usually a trigger for secondary therapy. However, not all patients patients with BCR progress to metastasis. It is estimated that BCR leads to metastasis only in around 35% of cases [117].

2.5.7 Androgen deprivation therapy (ADT)

Androgens are steroid hormones that control the development and the function of the male sexual organs and male characteristics. The major circulating androgen is testosterone, produced by testicles [118]. Another important androgen is dihydrotestosterone (DHT) produced by the peripheral tissue, which is even more potent than testosterone [118].

Androgens also play an important role in the development of prostate cancer. Reducing the levels of androgens or preventing them from binding to the androgen receptors often results in tumour shrinkage, or slows down its development. This strategy of handling prostate cancer is called androgen deprivation therapy, or hormone therapy. Androgen deprivation is achieved by surgical or chemical castration, the use of androgen blockers (compounds that block the androgen receptors, so that the androgen cannot bind to them), or a combination of both methods, referred to as maximum androgen blockade (MAB).

ADT is usually used along with radiotherapy for the patients with localised prostate cancer. Also it can be used if the cancer spread outside the prostate and the radical treatment will not be effective anymore, or if the patients are unfit for receiving curative treatment. Androgen deprivation therapy might also be an option if the disease relapses after radical treatment [119].

Hormone therapy comes with significant adverse effects such as muscular and bone mass loss, hot flushes, depression, erectile dysfunction, anaemia, cardiovascular and endocrinological problems [120].

ADT improves the survival of patients with locally advanced prostate cancer when used as adjuvant to radiotherapy. A randomized study on more than 400 patients [121] reported that patients who received hormone therapy following radiotherapy had a disease free survival rate of 85% (CI 78-92%), compared to the group that received only radiotherapy - 48% (CI 38-58%). A follow-up study [122] reported similar results, with a 5-year disease free survival of 74% (CI 67-81%) for the radiotherapy plus hormone therapy group, compared to 40% (CI 32-48%) in the radiotherapy-alone group.

2.5.7.1 Castration resistant prostate cancer (CRPC)

Hormone therapy leads to remission for a short period of time, but after that virtually all patients become unresponsive to treatment. This stage is referred to as castration resistant prostate cancer (CRPC) or androgen-independent prostate cancer. CRPC is characterised by continuous rise in the levels of PSA, progression of the pre-existing disease and/or appearance of new metastases [123].

Sharifi et al. [124] reported a median time of 13.1 months to developing CRPC in patients with metastatic disease that are given ADT. The patients that had no evidence of metastatic disease at the beginning hormone therapy, developed CRPC in 19.3 months and metastasis in 37.8 months.

2.5.8 Metastatic disease

Initially, the patients with metastatic prostate cancer are given hormone therapy, that delays the progression for about a year. After that the disease progresses to androgen-independent stage and the disease progression resumes. At that point several palliative treatments are available such as abiraterone, chemotherapy and immunotherapy. Also because most CRPC prostate cancers develop metastasis to bones, bone-targeted therapies can be tried, in order to alleviate the pain [123]. One example of such treatment is radium-223, which reduces the bone pain and increases the overall survival from 11.2 months to 14 months [125].

2.5.9 Genetic alterations

2.5.9.1 *TMPRSS2-ETS* fusions

Fusions between *TMPRSS2* and genes from the *ETS* transcription factors family have been reported in about half of prostate cancers [32, 126]. The most common member of *ETS* family fused to *TMPRSS2* is *ERG*, occurring in 40-55% of prostate cancers [32, 127–131]. Besides *ERG* fusions, fusions between *TMPRSS2-ETV1* have been reported in 5-10% of prostate cancers [32] and *TMPRSS2-ETV4* with much lower frequency (about 2% [132]).

TMPRSS2 is an androgen-regulated gene located on chromosome 21, highly expressed in prostate [133]. Androgen-regulated genes are important for the normal function of the prostate, and also have significant contribution to the development of the prostate cancer [133]. *ETS* is a family of genes, consisting of at least 27 genes, with role in transcription regulation [134]. *ETS* genes regulate around 400 other genes, some of them with role in cell proliferation and apoptosis [134].

Both TMPRSS2 and ERG are located on chromosome 21, only 3 million base pairs apart, on the same strand. The most well-characterised mechanism that leads to *TMPRSS2-ERG* fusion is the deletion of the chromosome between TMPRSS2 and ERG loci, occurring in about 60% of fusions [131]. In some cases copy-number analysis does not show a loss of the portion between the two genes, suggesting that the fusion might be a result of balanced translocations. However the balanced rearrangements leading to

TMPRSS2-ERG fusions are largely unexplored. Recently Berger et al. [53] described some complex balanced translocations where sets of 3 and 4 genes are involved in closed-chain translocations, called chromoplexy, that include *TMPRSS2-ERG* fusion.

The most common *TMPRSS2-ERG* fusion is between exon 1 of *TMPRSS2* and exon 4 of *ERG*, occuring in about 44% of fusion positive samples, and TMPRSS2 exon 1 with *ERG* exon 5, in 4% of cases [135].

The *ETS* gene fusions seem to be an early event in the development of prostate cancer. It is found in a high percentage of HGPIN (high-grade prostatic intraepithelial neoplasia - a precursor of prostate cancer) [136], but seem to be insufficient to induce the formation of prostate cancer [137].

The usefulness of *TMPRSS2-ETS* fusions as a biomarker in predicting the clinical outcome is controversial. A number of studies have reported an association between the *TMPRSS-ERG* fusions and poor outcome [138–140]. However some other studies found no association [33–35]. Also, it is reported that *TMPRSS2-ERG* is not a good predictor for the response to radiotherapy [141].

2.5.9.2 *PTEN loss*

PTEN is a tumour suppressor gene located on arm q of chromosome 10. It negatively regulates the PIK3/Akt pathway, which has an important role in controlling, among other things, cell growth, cell proliferation and apoptosis [57]. This means that when *PTEN* is inactivated, PIK3/Akt pathway becomes over-active, leading to uncontrolled cell division and decreased apoptosis [142].

The functional loss of *PTEN* might be induced by several events such as point mutations, reported in around 16% of prostate cancers [143, 144], and methylation. However, the most common cause are deletions occurring in the chromosome arm 10q, which is a very frequent chromosomal rearrangement in prostate cancer and other malignancies [142]. Deletions of *PTEN* have been reported in 30-60% of adenomacarcinomas, of which 10-30% are homozygous deletions [145–149], with a higher frequency in CPRC (deletions in 77% of samples and homozygous deletions in 43% [150]).

Inactivation of *PTEN* in general is linked to progression of prostate cancer and significantly worse survival outcome [151, 152], while homozygous deletions are associated with much faster biochemical recurrence [147].

There is a significant association between *PTEN* loss and *TMPRSS2-ERG* fusions. In one study all samples with *PTEN* deletions also harboured *TMPRSS2-ERG* fusions [149]. It has been hypothesised that the interaction between *PTEN* deletions and *TMPRSS2-ERG* fusions prostate is a significant driver for prostate cancer development and progression [153]. Bismar et al. [153] suggested that initial hemizygous loss of *PTEN* would promote genomic instability and facilitate gene fusions, leading to the formation of prostate cancer. Subsequent *PTEN* homozygous loss, would trigger further progression, to the invasive disease.

Disregulation of the PIK3/Akt pathway, mainly due to *PTEN* loss, seems to be critical to prostate cancer proliferation. Therefore, inhibiting key genes from this pathway might prove to be a viable therapeutic strategy for managing the disease. Currently several trials aiming to test various inhibitors are in progress [154].‘

2.5.9.3 *SPINK1* overexpression

SPINK1 overexpression has been reported in 8-11% of prostate cancers [155–157]. Initially, Tomlins et al. [158] found that *SPINK1* overexpression was mutually exclusive to *ETS* fusions and was associated with significantly poorer prognosis following prostectomy. Moreover, the *SPINK1* overexpression could be detected non-invasively, in urine. However, later studies failed to find association between *SPINK1* overexpression and *ETS* fusion status [156, 157] or significant biochemical or mortality difference in patients treated with radical prostectomy between the *SPINK1* positive and negative cases [157].

2.5.9.4 *SPOP* mutation

Barbieri et al. [159] discovered a subtype of prostate cancer characterised by mutations of the *SPOP* gene in 6-15% of cancers. The samples that harbour *SPOP* mutations have a distinct pattern of genomic alterations. The *SPOP* mutations are mutually exclusive to the *ETS* family rearrangements, and are highly association with *CHD1* deletions. However it is not clear yet the clinical usefulness of *SPOP* in predicting biochemical recurrence or survival. Blattner et al. [160] found no difference in biochemical recurrence in patients with *SPOP* mutations.

2.5.10 Emerging biomarkers and clinical tests

In recent years, with the advance of the new technologies such as microarrays and high throughput sequencing, a wealth of publications have reported new biomarkers for prostate cancer. Despite this, only a few have reached clinical practice. Here we described the most important biomarkers that are currently available in clinical practice or at an advanced validation stage.

2.5.10.1 *PCA3*

PCA3 is a non-coding mRNA that shows high values in over 90% of prostate tumours, while not expressing elevated values in the normal prostate [161]. It is detectable in urine collected following a DRE.

PCA3 score seems to increase with Gleason grade and with the probability of a positive biopsy, but it is nor correlated with prostate volume, age and PSA levels [162]. Also *PCA3* is correlated with the tumour volume, pathological stage and Gleason score of the resected prostate [163].

PCA3 has higher sensitivity and specificity than PSA in predicting the outcome of the first biopsy and, also, of the repeat biopsies [164, 165]. The accuracy of *PCA3* can be further improved by combining it with PSA, prostate volume and the DRE [164, 165].

In the UK, the *PCA3* test is not available yet in the National Health System (NHS), but only in a few private clinics. It is currently under assessment by NICE and in the future might be introduced into general clinical practice together with PSA to help doctors decide if a biopsy is necessary [166].

2.5.10.2 *AMACR*

AMACAR (α-Methylacyl coenzyme A racemase) is an enzyme that can be used as a diagnostic biomarker for prostate cancer and can also be a therapeutic target. Some studies suggested that the mRNA levels of *AMACR* are 9-fold higher in prostate cancer compared with normal tissue and are also strongly correlated with metastatic and androgen independent diseases [167].

AMACR seems to have high sensitivity and specificity for detecting prostate cancer. Rubin et al. [168] obtained 97% sensitivity and 100% specificity by measuring protein levels, while Luo et al. [167] obtained 95% sensitivity and 96% specificity.

Moreover, it seems that the expression of *AMACR* is functionally important for the growth of prostate cancer cells, at least in-vitro [169]. Reducing the expression of *AMACR* impaired the proliferation of prostate cancer cells in the LAPC-4 cell-line. As *AMACR* is not regulated by androgens [170], these results suggest that targeting *AMACR* as a complementary therapy might improve the efficiency of androgen deprivation. In LAPC-4 the combination of *AMACR* knock-down with androgen deprivation led to better results than either treatment alone [169].

2.5.10.3 Prolaris

Prolaris [38] is a biomarker based on the expression of 31 cell cycle progression (CCP) genes, designed to predict the outcome of prostate cancer. CCP genes are genes involved in essential cell cycle processes, whose expressed vary at different stages of cell cycle [171]. The expression of these genes seems to be correlated with the proliferation of tumours [171]. Before Prolaris, CCP genes had been successfully used in the prognosis of breast, lung and brain cancers [38, 172–174].

Prolaris was developed on data from two cohorts of patients: one cohort of 336 patients who underwent prostectomy and one group of 337 patients diagnosed from tissue extracted by TURP (transurethral resection of the prostate - a type of surgery designed to remove prostate tissue that causes urinary symptoms due to enlarged prostate). The expression levels of 31 CCP genes and 15 housekeeping genes were assessed using quantitative RT-PCR (a biological technique used to measure the expression levels of RNA). The expression of the CCP genes were measured relative to the expression of the housekeeping genes and the measures were combined to obtain a CCP score. The score is correlated with the expression of the CCP genes, an increase with one unit corresponding to a doubling in the expression.

For the radical prostectomy cohort the endpoints considered were time to biochemical recurrence (BCR) and death after progression. In the univariate analysis evaluating the association between CCP score and time to BCR, an increase of one unit of CCP score had a hazard ratio 1.89 (95% CI 1.69 - 5.28), indicating that patients with higher CCP scores progress faster to BCR. Also the CCP score seems to be associated with the risk of death due to disease progression, HR 2.99 (95% CI 1.69 - 5.28). The result was similar in the multivariate analysis, with CCP score and PSA value being the most important predictors.

In the TURP cohort only the time to death from prostate cancer was used as endpoint. The univariate hazard ratio 2.92 (95% CI 2.38 - 3.57) and multivariate ratio 2.56 (CI 1.85 - 3.53) were in line with the radical prostectomy cohort, suggesting that Prolaris is a robust biomarker.

2.5.10.4 Oncotype DX

Oncotype DX [39] is a a multi-gene RT-PCR array that measures the expression levels of 12 cancer-related genes and 5 reference genes from tissue extracted from biopsies, which predicts the aggressiveness of early-stage prostate cancer. The 12 cancer-related genes correspond to four biological pathways: androgen signaling pathway (*AZGP1*, *KLK2*, *SRD5A2* and *FAM13C*), cellular organization (*FLNC*, *GSN*, *TPM2*, and *GSTM2*), proliferation (*TPX2*) and stromal response (*BGN*, *COL1A1*, and *SFRP4*).

The expression of the 12 genes is normalised relative to reference genes and the normalised expressions are combined to obtain a GPS score (genomic prostate score), that ranges between 0 and 100, with higher score corresponding to more aggressive prostate cancer.

The score has been validated as an independent predictor of adverse pathology in patients with low/intermediate risk prostate cancer, based on biopsy samples. It has been reported as a significant predictor of pathological outcome. Moreover, it seems that GPS can be computed using very low sample volumes, which makes it suitable for biopsy extracted samples [175].

2.5.10.5 Dechipher

Dechipher [40] is a gene signature that predicts the risk of developing metastasis following radical prostectomy. The signature has been developed using *Affymetrix GeneChip Human Exon 1.0 ST Arrays*. It consists of 22 probesets corresponding to coding and non-coding RNA sequences.

The study was designed as a nested control-case study. The patients have been initially classified intro three groups: no evidence of disease (NED) - patients with no sign of BCR after 7 years, PSA-recurrence group (PSA) - patients with BCR, but no signs of metastasis within 5 years and systematic progression (SYS) - patients with metastasis within 5 years from radical prostectomy.

Initial screenings found little molecular differences between NED and PSA groups, but large differences between these two groups and SYS group, therefore NED and PSA groups have been combined into a single group (the control cases).

From the 545 samples that had RNA available for hybridisation on microarray, 359 samples were selected for training and 186 were held out for validation. Several pre-processing steps were performed in order to select the 22 most informative probesets from the total of 18,902 probesets differentially expressed between cases and controls. The 22 features have been assembled into a random forest classifier (a machine learning technique that will be discussed later in Section 3.2.2.1). The parameters of random forest have been optimised, resulting in a genomic classifier (GC) that outputs values between 0 and 1, increasing with the probability of developing metastasis.

The classifier obtained an AUC (area under the curve) of 0.9 in the training set and 0.75 in the validation set, more than a clinical-only classifier, built using only clinical variables (Gleason score, PSA, SVI, ECE, etc.). If split into GC > 0.5 and GC ≤ 0.5, in univariate logistic regression, the GC classifier outperforms 17 classifiers based on previously published signatures or individual biomarkers. Also in the univariate survival analysis it seems that GC is a significant predictor of the risk of dying from cancer (p-value 0.003).

2.6 Microarrays

Micorarrays are genomic tools that can be used to simultaneously measure the expression levels of thousands of genes or other transcripts. The first microarrays were created by Schena et al. [176] in the mid-1990s and revolutionised biological and medical research.

Microarrays are small glass or silicon slides that contain from a few tens to millions of *probes* (or spots). At each spot there are attached millions of copies of the same single-stranded DNA sequence, corresponding to a region of interest in the genome. These regions of interest might be part of a transcript or a whole transcript.

The RNA extracted from cells is amplified and converted to cDNA (complementary DNA) in a reaction called reverse transcription. The resulting material is then labelled with a fluorescent dye and is injected onto the microarray. Depending on the expression level of the genomic area (exon, gene) interrogated by the probe, a larger or a smaller number of complementary sequences hybridize to each probe. The microarray is then scanned with a laser, which measures the luminosity of each spot. The image resulting from scanning is processed using software that converts the luminosity of each spot to a number. The resulting numbers are mapped to corresponding sequence and the data is normalised, to mitigate possible sources of non-biological bias. Normalisation algorithms take into account several sources of variation such as probe affinity, background noise, and position on the slide. The resulting normalised data can then be analysed using various bioinformatical approaches, depending on the purpose of the study.

Trevino et al. [177] describes the most common applications of microarray technology, including the identification of differentially expressed genes between two groups of samples, biomarker detection, study of the relationship between the molecular profile and biological manifestation and identification of genes associated with risk and survival.

The identification of differentially expressed genes refers to finding those genes that are up-regulated or down-regulated between two conditions, such as different treatment conditions, cancer vs. non-cancer tissue, patients with different outcomes, and samples before and after a certain treatment. There are several statistical methods available such as linear models, *t*-tests and gene-set enrichment analysis that can identify the differentially expressed genes. Usually this type of analysis leads to the selection of a subset of genes for further analyses.

One of the most useful applications of microarrays is the derivation of *gene signatures* that can be used as biomarkers for various purposes, such as diagnosis, risk stratification, response to treatment prediction, etc. A gene signature is a set of genes whose expression levels can be used to discriminate between two or more conditions, and which also have good predictive power for reliably classifying new samples [177]. Usually gene signatures are derived using supervised methods. Labelled samples (samples for which the outcome of interest is known) are used by a machine learning model to identify and optimise the genes with the biggest discriminative power. The signature is then validated on independent sets of data, that have not been used for training, in order make sure that it is robust and has good predicative power.

Unsupervised methods are also often used for analysing microarray data. These methods are used to identify groups of similar samples without using any labelled data. In cancer, unsupervised analysis can reveal underlying mechanisms that can explain clinical outcomes and can offer insights for understanding the heterogeneity of the disease.

2.6.1 Exon microarrays

The analysis in this thesis is mainly based on data obtained using the *Affymetrix GeneChip Human Exon 1.0 ST Arrays* platform, which are one of the highest resolution microarrays currently available. Throughout this thesis we will refer to this type of microarrays as *exon microarrays*.

Exon microarrays contain over 5.5 million probes, grouped in 1.4 million probesets, interrogating over 1 million known or predicted exons, therefore offering a very comprehensive coverage of the genome. They contain on average 4 probes per exon and 40 probes per gene. This allows two main type of analyses, exon level and gene level.

As presented on the manufacturer's website [178], the original purpose of exon microarrays was to allow the study of alternative splicing events such as intron retention, exon skipping and alternative promoter usage. However in this thesis we illustrate the use of exon microarrays to detect gene fusions and transcriptional deregulations within genes. Besides the exon level analyses, exon microarrays can also be used as to measure the expression levels of genes, just as standard microarrays do.

Usually the standard microarrays contain a mismatch probe for every perfect match probe (probes that interrogate the genome). The mismatch probes have the same sequence as the corresponding perfect match probe, except for one nucleotide in the middle. The purpose of these probes is to estimate the non-specific hybridization levels, and therefore to help correct for background noise. Exon microarrays lack mismatch probes, and thus several normalisation algorithms that rely on mismatch probes can not be used. However there are some algorithms such as RMA, fRMA and PLIER that can be still used and which will be presented in the following sections. These algorithms estimate the background signal using two types of probes that exon microarrays provide: genomic background probes - probes from regions of genome which are very unlikely to be transcribed and anti-genomic background probes - probes not found in genome.

Having only one probe to interrogate a genomic region, might be unreliable due to various sources of technical bias, such as sequence hybridization affinity, position on the microarray, and background noise. Therefore, microarrays have in general several probes measuring the expression of the same biological sequence of interest (exons in this case). These probes usually map to slightly different locations within the same genomic region, to protect from sequence specific effects and also are placed on different positions on the chip, to protect from local variations within the array. During the normalisation phase, the estimates provided by each probes are adjusted and summarised together, obtaining a single estimate. The group of probes that interrogate the same region of interest is referred to as *probeset*.

Probesets from exon microarrays are classified into five confidence categories, depending on the quality of evidence supporting the transcription of the genomic sequence (see Table 2.3). Throughout the analyses, depending on the purpose of the analysis, we will use various categories of probesets. We will describe which probesets we used for each individual analysis in the following chapters.

2.6.2 Exon microarray normalisation

The raw microarray signals need to be pre-processed in order to correct the effects and biases that occur during the experimental procedures. There are several algorithms for normalising exon microarrays such as *RMA (Robust multiarray analysis)* [180, 181], *Frozen robust multiarray analysis (fRMA)* [182] and *PLIER (Probe Logarithmic Intensity Error*, proposed by Affymetrix*)*. However the most commonly used algorithms in practice are RMA and, its slightly modified versions, such as fRMA. PLIER was reported as being technically biased and numerically unstable [183], and is not very much used.

Table 2.3 Exon microarrays probeset classification by confidence level [179].

Evidence Level	Description
Core	Refers to probesets that are supported by the most reliable evidence from RefSeq and full-length mRNA GenBank records containing complete CDS information.
Extended	Refers to probesets that are supported by other cDNA evidence beyond what is used to support core probesets. Extended evidence comes from other Genbank mRNAs not annotated as full-length, EST sequences, ENSEMBL gene collections, synthetically mapped mRNA from Mouse, Rat, or Human, mitoMap mitochondrial genes, microRNA registry genes, vegaGene, and vegaPseudoGene records.
Full	Refers to probesets that are supported by computational gene prediction evidence only. They are supported by gene and exon prediction algorithms including GeneID, GenScan, GenScanSubOptimal, exoniphy, RNAGene, sgpGene and Twinscan.
Free	Refers to probesets that are supported by annotations which were merged such that no single annotation (or evidence) contains the probeset.
Ambiguous	Refers to probesets that cannot be unambiguously assigned to a particular transcript cluster.

2.6.2.1 Robust multiarray analysis (RMA)

The RMA algorithm proposed by Irizarry et al. [180] is one of the most commonly used normalisation methods for exon microarrays. The main advantage of RMA is that it uses only perfect match probes. RMA normalisation consists of three steps: background correction, quantile normalisation and summarisation.

The first step of RMA is background correction. The purpose of background normalisation is to correct for non-specific binding, i.e. the hybridisation of sequences that are not complementary to microarray probes. The model assumes that the observed probes intensities are a combination of the true signal and background noise. More specifically, as presented in Bolstad [181]:

$$S = X + Y, \tag{2.1}$$

where S is the observed signal intensities of the probes, X is the true signal (assumed to follow an exponential distribution) and Y is the background noise, normally distributed and truncated at 0 to avoid negative values. Under this model the background corrected values are given by the expectation $\mathbb{E}(X|S)$.

Next, the probe intensities are quantile normalised. Quantile normalisation [184] is a method designed to make the distribution of probe intensities the same. This is achieved by transforming the intensities so that the corresponding quantiles across all microarrays are equal.

The third step of RMA normalisation is the summarisation of the intensities of probes within a probeset in order to obtain a single value, the probeset estimate expression level. Li and Wong [185] observed that the variation of the intensities of probes from the same probeset can be very large, due to probe-specific effects (or affinities). Sometimes the variation due to probe-specific effects was larger than the variance across microarrays [185]. Fortunately these probe-specific effects are reproducible, predictable and can be reliably accounted for. RMA uses the following linear additive model to account for probes affinities, when estimated the probeset expression:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + e_{ijn}, \text{ with } i = 1,...,I, j = 1,...,J, n = 1,...,N, \tag{2.2}$$

where i is the index of the microarray, j represents the probe index in the probeset, and n is the probeset index in the microarray. Y_{ijn} represents the $\log_2$ background-adjusted and quantile normalised expression level of a probe j from probeset n from the array i, μ_{in} is the $\log_2$ expression level of the probeset n in array i, α_{jn} is the probe affinity of the probe j from probeset n and e_{ijn} an independent identically distributed error term with mean 0 [180].

The parameters of the above model are estimated using the median polish algorithm [186], which is robust to outliers. In the end we are interested in the value of μ_{in}, which represents the probeset expression level, after we corrected for probe affinities.

2.6.2.2 Frozen robust multiarray analysis (fRMA)

Frozen robust multiarray analysis (fRMA) [182] is an extension of the RMA algorithm. The main difference is that the reference distribution used in quantile normalisation, the probe-effects and the error variances necessary for RMA are not computed locally from a set of microarrays anymore, but have been precomputed using a large number of microarrays available in the public databases and frozen. This allows fRMA to process single arrays or small batches separately, and to obtain in the end comparable arrays.

More specifically, the background correction for fRMA is the same as for RMA, as background correction is a single-array procedure anyway. For quantile normalisation, probe intensities of the single-array/batch are forced to the frozen reference distribution.

Intuitively, one would expect the probe effects to be constant across studies. However McCall et al. [182] discovered that the probe-specific effects were variable in samples coming from different batches. Also McCall et al. [182] noted the variance of the error e_{ijn} (see Equation 2.2) within batches is different. Therefore, the summarisation probe-level model has been extended to account for batch-effects and to allow the error variability to depend on batch as well. The updated summarisation model is:

$$Y_{ijkn} = \mu_{in} + \alpha_{jn} + \gamma_{jkn} + e_{ijkn}, \text{ with } i = 1,...,I, j = 1,...,J, n = 1,...,N, k = 1,...,K, \tag{2.3}$$

where a k has been added to notation to represent batch. We note that compared to Model 2.2 this model has a new term, γ, that accounts for batch-specific variability and that the error term depends now on the batch as well.

The performance of fRMA is comparable to the performance of RMA. When data was processed together RMA slightly outperformed fRMA, while when processing the data in separate batches, fRMA slightly outperformed RMA [182].

2.6.3 Quality assessment of exon microarrays

Affymetrix produced a series of metrics for identifying the outlier microarrays, due to technical effects. These metrics are described in the Quality Assessment of Exon and Gene Arrays whitepaper [187].

There are presented three probeset summarisation based metrics, useful in determining the general quality of the data, i.e. *positive controls vs. negative controls area under the curve (AUC)*, the *mean of the absolute deviation (MAD) of the residuals from the median* and the *mean absolute relative log expression (RLE)*. Out of these three metrics, the mean absolute RLE metrics is useful only in experiments where the same sample has been run over many chips. This is not the case in our analyses, as we have samples from a big variety of RNA sources and hence we did not use it.

Besides the above three metrics, in the whitepaper there are presented a series of other metrics, useful in troubleshooting various aspects of the microarray analysis protocol, in order to identify the source of problems in the outlier microarrays. Also these metrics have not been used in our analyses.

2.6.3.1 The positive controls vs. negative controls AUC metric

The positive controls vs. negative controls AUC metric is based on the intensity of two groups of special probesets: positive controls and negative controls. The positive controls are a set of probesets mapping to the exons of about 100 putative housekeeping genes, and therefore are expected to have high intensities. The negative controls on the other hand correspond to intronic regions of the housekeeping genes. The negative controls are expected to have low intensities in general.

The positive controls are considered to estimate the true positive rate, while the negative controls are considered to estimate the false positive rate. Using these two estimates a receiving operating characteristic (ROC) curve is generated. In this case, the ROC curve evaluates how well the probeset signals separate the positive controls from negative controls.

The positive controls vs. negative controls AUC metric corresponds to the AUC value corresponding to the ROC curve. The AUC gives an indication of the separation, with values close to 1 indicating a good separation, and values close to 0.5 indicating a poor separation.

2.6.3.2 The MAD of the residuals from the median

The MAD of the residuals from the median is a summary statistic based on the deviation from the estimated probe-specific affinities. More specifically, as described in Section 2.6.2.1, each probe has a different hybridisation affinity, that can result in big variations in signal intensities across different probes. However these probe-specific affinities are usually predictable and can be estimated by the RMA algorithm.

The MAD of the residuals from the median aims to identify problematic chips by assessing if a large number of probes are behaving differently than predicted by RMA. This is achieved by calculating for each probe the difference (residual) between the predicted intensity and observed intensity. The mean of all the residuals is then calculated, resulting in a metric that has large values in poorly performing microarrays.

2.7 Primary tissues and cell cultures

In cancer research there are two commonly used approaches to obtain tissue samples for analysis. The first one is to extract tissue from clinical samples, to store it and analyse it using various techniques such as microarrays, sequencing, or methylation arrays. The second one is to grow cancer cells *in vitro* or *in vivo*, eventually using different experimental conditions, and to study their behaviour. Cell cultures can also be subjected to -omic analysis.

Both approaches have advantages and disadvantages. The clinical samples can give an accurate image of the state of disease. Mutations and recurrent alterations can be detected, the expression level of genes can be determined and the histology of the tumour can be evaluated. However they provide only indirect information about the processes happening inside the cell and how the cells evolve. Cell cultures, on the other hand, can give a dynamical insight into how the cancer cells are proliferating and how they are responding to treatment, but, they might not reflect accurately what is happening *in vivo*.

Clinical samples can be obtained from primary tumour, normal or metastatic tissue. The extracted tissues are *fresh frozen (FF)* or *formalin-fixed paraffin-embedded (FFPE)*, in order to protect them from degradation. FFPE tissues are obtained in two steps: formalin fixation - the freshly extracted tissues are treated with formalin, a solution that preserves the tissue permanently, with a minimum impact on its structure, and paraffin embedding - the encapsulation of the fixed tissue into a wax called paraffin, that supports the tissue and enables researchers to cut microscopic sections when needed. FF tissues are obtained by submerging the fresh sample into liquid nitrogen.

FFPE tissues can be stored at room temperature for many years, therefore are cost-effective and convenient, while FF samples need dedicated freezer storage [188]. Moreover, there are available large archives of FFPE tissues, many of them with long follow-up time, which makes them very useful for retrospective studies. Also FFPE tissues are preserved in a form more suitable for morphological analysis [188].

The main advantage of FF samples is that the quality of RNA is much better than the RNA obtained from FFPE samples, making them more suitable for molecular analysis [188]. Despite this, a large number of studies reported promising results when performing sequencing and microarray profiling on genetic material obtained from FFPE specimens [189–194]. One the other hand, the processing of FF tissues is much faster than FFPE. Also the FFPE protocol is not standardized, and this can lead to biases [188]

Another way of understanding cancer is to grow cancer cells outside the source organism. The initial culture is referred to as primary culture, which undergoes multiple sub-cultures in order to produce *cell-lines* [195]. The cell-lines can be grown *in vitro*, given proper medium, or can be injected into immunodeficient mice, to obtain *xenografts*, which are efficient *in vivo* models.

The primary culture is the closest model to the original tissue [196]. The main disadvantage of using primary culture is that the cells do not always grow in culture or die after few replications. Also the primary cultures are less well characterised.

An easier approach is to work with immortalised cell-lines. Immortalised cell-lines are cultures derived from tumours which acquired capability to reproduce indefinitely [195]. The first human cancer cell-line, named HeLa, was established in 1951 [197]. Ever since, cell-lines have been established in all types of cancer.

Cell-lines revolutionised medical research and are used for drug development, study of gene function, generation of artificial tissues and synthesis of biological compounds [198]. They are cost effective, easy to use, provide an unlimited supply of material, bypass ethical concerns and are well characterised [198]. The main drawback of cell-lines is that they might not reflect accurately what is happening in the original tissue, due to different growth conditions and mutations acquired by cells during the multiple passages. Several studies evaluated the differences between the cell-lines and primary tumours [199–205]. Most of these studies identify that cell-lines have similar recurrent mutations and alterations compared with tumours, but, overall, cell-lines have more mutations [206]. It is not clear yet the effect of the cell-line specific alterations in producing differences between *in vitro* and *in vivo* models [206].

2.8 Discussion

In this chapter we have introduced the basic biological and medical concepts necessary for understanding the analyses and the results presented in this thesis. We presented the central dogma of molecular biology, the general characteristics of cancer and the overview of the management of prostate cancer. We also described how microarrays work and rationale for using various types of tissues in the research of cancer. We shall now describe the bioinformatics methods used to analyse this type of data.

Chapter 3

Computational background

3.1 Summary

In this chapter we introduce the computational approaches that were used in this thesis. We present a review of several machine learning approaches that were applied to define groups of patients with different mutational and gene expression profiles. We also introduce survival analysis models, that helped us compare the clinical outcomes of patients from different groups. We further present methods used for pathway analysis, to study the biological functionality of various sets of genes identified by our analyses.

3.2 Machine learning

Machine learning is a form of artificial intelligence that uses example data or past experiences to learn the parameters of a mathematical or statistical model, that is then used to partition (new) data into classes of objects with similar characteristics. Machine learning techniques are used nowadays in a wide variety of applications, from speech and face recognition to classification of cancers into subtypes. In machine learning there are two main approaches for partitioning data objects into classes: *supervised methods* and *unsupervised methods*.

Supervised methods classify objects based on models that are trained on a set of objects (also referred to as instances, samples, data points, or, simply, points) for which the class is known *a priori*. The objects for which the class is known are referred to as *labelled objects*, while the objects for which the class is unknown are referred to as *unlabelled objects*. We will present some supervised methods in Section 3.2.2.

The other main category of machine learning approaches is unsupervised methods. Unsupervised methods partition the data into classes, or more commonly referred to as *clusters* of objects, with similar characteristics. The difference to supervised methods is that the labels should not be known *a priori*, as they are derived from the data. We will discuss three unsupervised methods in Section 3.2.1.

There are also semi-supervised methods that work with models trained using both labelled and unlabelled data. However these methods are beyond the scope of this work and will not be discussed here.

3.2.1 Clustering methods

In this section we discuss two commonly used clustering methods, hierarchical clustering and *k*-means, based on Chapter 8 of Tan et al. [207] and Part III of Maimon and Rokach [208] and also a relatively new method, called latent process decomposition (LPD) based on the work of Rogers et al. [209].

Unsupervised classification, also referred to as cluster analysis, groups the objects into clusters with common characteristics. The aim is for the objects in a cluster to be similar to one another and different to objects in other clusters. The separation of the objects from a dataset into clusters is referred to as a *clustering*.

One can distinguish between different types of clusterings such as hierarchical vs. partitional and exclusive vs. overlapping vs. fuzzy.

Partitional clustering refers to the division of objects into non-overlapping groups, with each object assigned to exactly one group. *Hierarchical clustering*, on the other hand, allows the clusters to be further divided into subclusters. The clusters in a hierarchical clustering can be organised into a tree with nodes representing clusters and their children representing subclusters (see Figure 3.1).

Exclusive clustering assigns each sample exclusively to one cluster. However, there are situations when objects need to be assigned to more than one group. This behaviour can be modelled using *overlapping clustering*, which allows samples to be assigned to several clusters simultaneously. Sometimes objects can be assigned to clusters with a certain membership weight. This is called *fuzzy clustering*. One of the common instances of fuzzy clustering is *probabilistic clustering*, where a sample is assigned to a cluster with a certain probability. The probability of the sample to belong to a cluster is a number between 0 and 1, with a further constraint that the sum of probabilities for all clusters add up to 1.

3.2.1.1 Hierarchical clustering

As discussed earlier, hierarchical clustering is a technique of clustering where clusters are organised as a tree, with nodes representing clusters and children subclusters. The root of the tree is a cluster comprising all samples, while leaves are clusters that contain a single object.

There are two approaches for generating hierarchical clustering: *agglomerative* (or bottom-up) and *divisive* (or top-down). Agglomerative clustering starts with each objects assigned to a separate cluster and at each steps merges the closest pair of clusters, until only one cluster is obtained. The agglomerative hierarchical clustering algorithm is schematically presented in Algorithm 1. Divisive clustering, on the other hand, starts with all objects assigned to a single clusters and at each step splits a cluster into two subclusters. For the moment we will focus only on the agglomerative clustering approach, as it is the technique used in this thesis. We will further refer to agglomerative hierarchical clustering simply as hierarchical clustering.

Algorithm 1 The agglomerative hierarchical clustering algorithm.

```
Assign each data point to a separate cluster.
repeat
    Merge the two most similar clusters into a single cluster.
until A single cluster is obtained.
```

In order to determine the closest clusters, a proximity measure is used. The proximity measure gives an indication of how similar two clusters are. There are several ways of defining the proximity measure. The complete link (or maximum linkage) defines the proximity of two clusters as the maximum distance between any point from the first cluster to any point from the other cluster. Single link (or minimum linkage) is the minimum distance between two points from the separate clusters, while average link defines the proximity as the mean distance between all possible pairs of points from the two clusters.

Another very important aspect of clustering is how the distance between a pair of points is computed. Hierarchical clustering can work with either distance measures (also referred to as metrics), or similarity/dissimilarity measures, depending on the nature of the data. For example, in the case of points from a metric space, Manhattan, Euclidean or Minkowski distances can be used, while for objects with binary or nominal attributes the Jaccard coefficient might be more suitable. For genetic expression profiles on the other hand it might be more suitable to work with some similarity measures, such as Pearson's correlation, in order to assess how similar the expression patterns between two tissues are.

Hierarchical clustering can be visually represented in a tree-like structure, called a *dendrogram* (Figure 3.1b). A dendrogram depicts both the relationship between clusters and its subclusters and the order in which they were merged. Each leaf represents a data point. A subtree represents the cluster that contains all the data points which are leaves in the subtree. The height from the bottom of the tree to the horizontal line that connects two subtrees indicates the degree of dissimilarity between the two clusters represented by the subtrees. A dendrogram corresponding to the hierarchical clustering of the points from a synthetic dataset containing points from the two-dimensional space is presented in Figure 3.1.

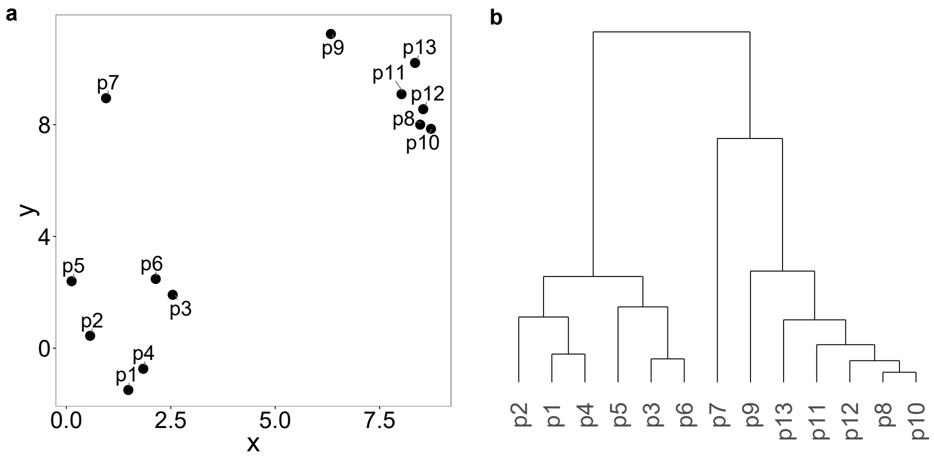

Figure 3.1 Hierarchical clustering on a synthetic dataset: a) a scatter plot depicting points in a two-dimensional space; b) a dendrogram corresponding to the hierarchical clustering of the points in Figure 3.1a, using the average linkage method and Euclidean distance.

Hierarchical clustering algorithms cannot intrinsically determine the number of optimal clusters in a dataset. This number needs to be evaluated separately using some external numerical criteria, by visual inspection of the dendrogram, or using some *a priori* knowledge about the data. Once the number of clusters is chosen, the clusters are determined by cutting the dendrogram into the desired number of subtrees.

3.2.1.2 *k*-means

k-means is a member of the partitional clustering algorithms family. It splits the data points into a predefined number *k* of non-overlapping clusters. Each data point is assigned exclusively to a single cluster.

The *k*-means algorithm is also a prototype-based clustering technique. This means that each cluster is represented by a prototype, i.e. a representative data point, usually defined as the mean or the median of the points in the cluster. In the case of *k*-means, the prototype is referred to as *centroid*. Each point is assigned to the centroid closest to the point, therefore a cluster is represented by the set of points assigned to a centroid.

k-means works by initializing *k* centroids, most commonly at random positions in the data space. Each point is then assigned to closest centroid. Next, the centroids are moved to the middle of all points assigned to that centroid (the cluster), i.e. to the mean of all points in the cluster. Because the position of centroids changes it is possible that some points become closer to other centroids than previously assigned. Therefore, the previous two steps are repeated until the process convergences. Convergence means that between two iterations no points changed the assignment and thus the position of the centroids are stable. The *k*-means algorithm is schematically presented in Algorithm 2 and is visually illustrated in Figure 3.2.

Algorithm 2 The *k*-means algorithm.

```
Initialise the $k$ centroids at random positions.
repeat
    Assign each point to the closest centroid.
    Update the position of centroid.
until The position of centroids does not change anymore (convergence).
```

k-means is a particular case of the EM algorithm, which is mathematically proven to converge to a local maxima [210]. Thus, it is certain that in a finite number of iterations, the algorithm will stop.

As in the case of hierarchical clustering, the distance between a pair of points can be computed using different measures, depending on the *k*-means application. Most commonly used distances are the Euclidean distance, Manhattan distance, cosine similarity and Bregman divergence.

When applying the *k*-means algorithm to multidimensional data, a common practice is to first apply a dimensionality reduction technique on the data, in order to make it easier to visualize. We will discuss in Section 3.4 such a technique, called principal component analysis.

3.2.1.3 Latent process decomposition (LPD)

In this section we present a hierarchical Bayesian technique called *latent process decomposition (LPD)*, which is the basis of the analysis presented in Chapter 5. LPD is an extension of the latent Dirichlet allocation (LDA) approach [211] and is fully described in Rogers et al. [209].

LPD is a probabilistic clustering of microarray data. This means that LPD allows objects to have partial membership to more than one cluster, reflecting the fact that a given object can share some characteristics with a group of objects, but in the same time it can share other characteristics with a different group of objects.

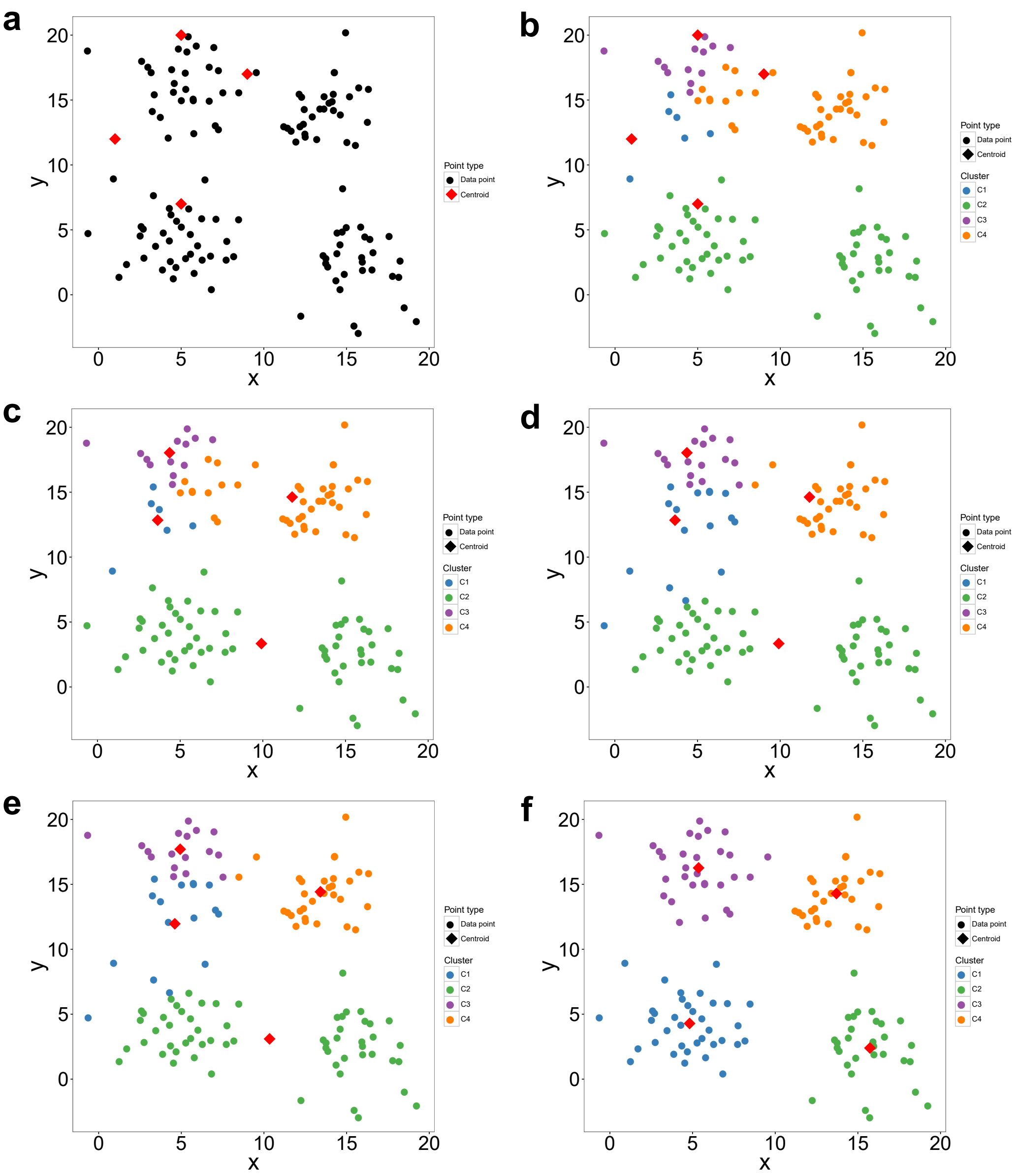

Figure 3.2 Illustration of the *k*-means algorithm on a synthetic dataset generated by sampling four bivariate normal distributions with means (5, 5), (5, 16), (16, 3) and (14, 14) and standard deviation (2, 2): a) the sampled points, prior to being assigned to clusters (the round black points) and the randomly initiated centroids (the four red diamonds); b) the initial assignment of the points to clusters, based on the initial position of the centroids; c) the first recalculation of the centroid positions. We can see that the centroids have been moved to the weight centre of each cluster; d) the second iteration of the *k*-means algorithm - the points have been reassignment to clusters represented by the closest centroid. We note that the top-right orange cluster is already well defined; e) the second iteration of the *k*-means algorithm - the centroids have been moved to the middle of the newly defined clusters; f) the final *k*-means result. In this case, the algorithm converged after 5 iterations.

In the context of prostate cancer we assume that a cluster represents a biological process that leads to a certain expression pattern. As prostate cancer is a highly heterogeneous disease, and often several foci are present in the same sample [23], it is possible that several distinct processes are simultaneously present and are jointly contributing to the expression profile of a given sample.

LPD determines for each process an expression profile, that describes the expected expression level of each gene due to the process. Then, for a given sample it estimates how well the expression profile of each process is reflected in the expression levels of the genes in the sample. Alternatively, we say that LPD determines the contribution of each process to the expression profile of a sample. More specifically, in a given dataset D, for which the number, K, of processes is known in advance, LPD considers that each gene g in a given set of genes G has a specific distribution in each process. The distribution of each gene g in process k, is assumed to follow a normal distribution (a Gaussian distribution) with mean μ_{gk} and variance σ_{gk}.

LPD assumes that for a given sample there is a specific distribution of processes, θ, that contribute to its observed expression profile. The distribution θ is a K-dimensional vector whose elements θ_k are mixture components which take values between 0 and 1, and which sum to 1. These values reflect the probability of each process being involved in the generation of the expression profile of a sample, i.e. the contribution of each process to the sample. The distribution θ, in its turn, is assumed to come from a dataset-specific Dirichlet distribution, $Dir(\alpha)$, which reflects how the mixture of components θ vary across the samples in the dataset.

From a generative perspective, the model works as follows: For each sample a, a multinomial distribution, θ, is sampled from the Dirichlet distribution, $Dir(\alpha)$. Then, for each gene, g, a process k is drawn from the distribution θ with probability θ_k. The expression level of the gene g in sample a, e_{ga}, is then sampled from the Gaussian distribution corresponding to process k, which has the mean μ_k and variance σ_k.

The graphical representation of the structure of the model is presented in Figure 3.3.

3.2.1.3.1 Parameter estimation In general, the Bayesian models, such as LPD, work with observed data D and a set of parameters, H which are unknown (or hidden) and need to be estimated. In our case, the set of parameters is $H = \{\alpha, \mu, \sigma, \theta\}$, where μ denotes the set of parameters μ_{gk} and σ, the set σ_{gk}. When fitting a model to a given observed dataset D, we are interested in estimating the values for the parameters H for which the *posterior probability* $p(H|D)$, that is the probability of parameters given the data, is maximised. The maximum $p(H|D)$ is usually referred to as the *maximum posterior (MAP).*

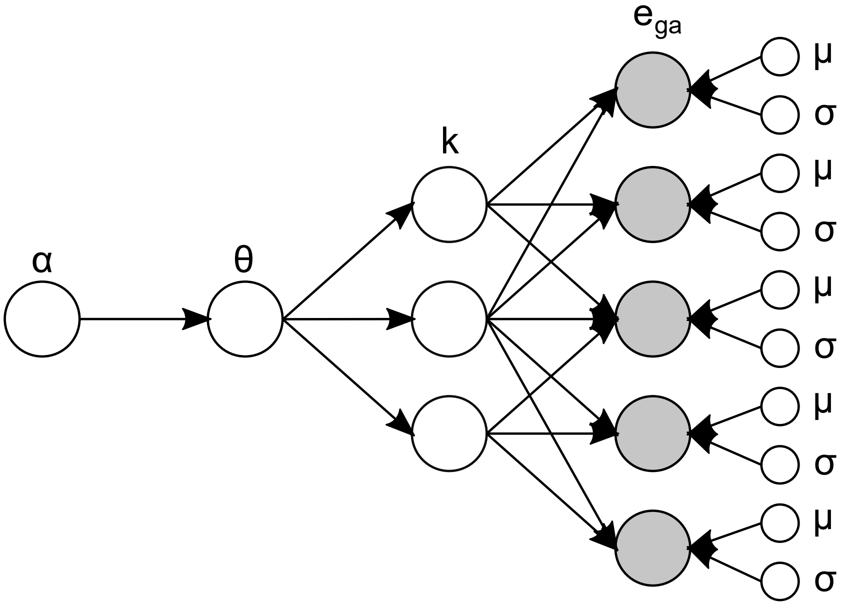

Figure 3.3 A schematic illustration of the LPD model, adapted from Rogers et al. [209]. Each circle corresponds to a variable. The empty circles correspond to hidden variables (variables for which we do not observe the values) and filled circles correspond to observed variables. The arrows represent conditional dependencies between variables.

In order to estimate the MAP, Bayes' rule can be employed. Bayes' rule specifies that:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}. \tag{3.1}$$

The factor $p(D|H)$ is known as the *likelihood*, and represents the likelihood of the data given the parameters, while $p(H)$ is the *prior*, which, intuitively, encodes any prior knowledge (or belief) about the data, before seeing it.

We are interested in finding H for which the posterior probability is maximised, and therefore we can ignore the denominator from the above equation, as it does not depend on H. Following from this, we say that:

$$p(H|D) \propto p(D|H)p(H), \tag{3.2}$$

meaning that the MAP is proportional with the product between the likelihood and the prior.

If no prior belief is held about the data, we say that we have an *uninformative (or uniform) prior*, and we consider the probability $p(H)$ constant across H. In this case, finding the MAP solution is equivalent with finding the *maximum likelihood (MLE)* solution (the values of H for which the likelihood $p(D|H)$ is maximised).

Depending on the nature of the data, and the type of model used, there are many approaches for finding the global or local maximum likelihood solutions. However, a big issue with the MLE is that often it leads to *over-fitting*. Over-fitting refers to training a model which fits too tightly to the training data, and which does not work well on any new data. One way of dealing with this issue is to perform *cross-validation*, a technique in which a proportion of samples are held out in turn and then used to check if the model trained on the rest of samples has a good generalisation power.

In a Bayesian setting, however, the over-fitting problem can be solved by defining suitable *informative (non-uniform) priors*. This means that instead of using uniform distributions for the priors, we specify prior distributions that reflect our belief about the expected form of the parameters. Using the informative priors together with likelihood, leads to maximum posterior solutions (MAP).

LPD provides implementations for both MLE and MAP solutions. In our analysis, the MLE approach is very useful in determining the optimal number of processes, as we will illustrate later. The MAP solution, also helps in determining the number of processes, but, more importantly, it is the approach used for setting the final model parameters and classifying the samples.

3.2.1.3.2 The MLE solution For the MLE solution, the likelihood can be expressed as:

$$p(D|\mu,\sigma,\alpha) = \prod_{a=1}^{A} \int_{\theta} p(a|\mu,\sigma,\theta)p(\theta|\alpha)d\theta, \tag{3.3}$$

where A is the number of samples. As the log function is a monotonous increasing function, finding the maximum likelihood is equivalent with finding the maximum log-likelihod, defined as $logp(D|H)$. In practice, it is usually easier to estimate the maximum log-likelihood. Expanding from the above likelihood definition, the log-likelihood for each sample, a, can be expressed as:

$$logp(a|\mu,\sigma,\alpha) = log \int_{\theta} \left\{ \prod_{g=1}^{G} \sum_{k=1}^{K} \mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk}) \right\} p(\theta|\alpha)d\theta, \tag{3.4}$$

where $\mathcal{N}$ denotes the normal distribution.

The presence of the summation over k inside the logarithm, makes the the log-likelihood intractable for exact parameter calculation. There are, however, several parameter approximation techniques, that can be employed. In the implementation we used, provided by Rogers et al. [209], the parameters have been estimated using the Bayesian variational inference framework.

Two sets of variational parameters, Q_{kga} anf γ_{ak}, are introduced in order to estimate a lower bound [212], for the log-likelihood. Informally, a lower bound is a function that approximates the log-likelihood function, and which is mathematically guaranteed to be lower or equal with the log-likelihood at any point. The lower bound is introduced as it can be more easily maximised, and its maximums usually provide good approximations for the model parameters. The variational parameters above are defined as:

$$Q_{gka} = \frac{\mathscr{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk}) exp\{\psi(\gamma_{ak})\}}{\sum_{k=1}^{K} \mathscr{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk}) exp\{\psi(\gamma_{ak})\}}, \tag{3.5}$$

where $\psi(x)$ is the digamma function, and:

$$\gamma_{ak} = \alpha_k + \sum_{g=1}^{G} Q_{kga}. \tag{3.6}$$

Using the variational parameters, the model parameters can be iteratively computed as:

$$\mu_{gk} = \frac{\sum_{a=1}^{A} Q_{gka} e_{ga}}{\sum_{a'=1}^{A} Q_{gka'}} \tag{3.7}$$

$$\sigma_{gk}^2 = \frac{\sum_{a=1}^{A} Q_{gka}(e_{ga} - \mu_{gk})^2}{\sum_{a'=1}^{A} Q_{gka'}}. \tag{3.8}$$

and:

$$\alpha_{new} = \alpha_{old} - \mathbf{H}(\alpha_{old})^{-1} \mathbf{g}(\alpha_{old}), \tag{3.9}$$

where $\mathbf{H}(x)$ is a Hessian matrix and $\mathbf{g}(x)$ is a gradient, both described in Rogers et al. [209], α_{new} is the updated value of the Dirichlet α parameter and α_{old} is value of α obtained at the previous iteration.

3.2.1.3.3 The MAP solution As we described previously, informative priors for the parameters can be introduced, in order to avoid the over-fitting problems, that can occur in the MLE estimations. The informative priors reflect our beliefs about the form of the parameters. For example, if for a dataset for which the expression level of each gene has been normalised across samples to a normal distribution with mean 0 and variance 1, we can assume that the parameter μ_{gk} (which represents the mean expression level of a gene, g, in process k) comes from a normal distribution $\mathscr{N}(0, \sigma_\mu)$. This encodes our prior belief the bulk of genes are not expected to be differentially expressed in a given process, and only some, i.e. the outliers from this distribution, will be have a process-specific distribution.

Similarly, we can assume that the variance parameters, σ^2_{gk} will tend to be close to 1. We can also design the prior over the variance parameter σ^2_{gk} such way that we make sure that the variances will never be 0, as it would make the Gaussians collapse into a single point, leading to numerical instability and model over-fitting.

Therefore the following priors over the parameters are set to:

$$p(\mu_{gk}) \propto \mathcal{N}(0, \sigma_\mu) \tag{3.10}$$

and:

$$p(\sigma^2_{gk}) \propto exp\left\{-\frac{s}{\sigma^2_{gk}}\right\}, \tag{3.11}$$

which specifies that the μ_{gk} come from a normal distribution and that σ^2_{gk} come from an (improper) exponential distribution, which have the desired properties. The quantities σ_μ and s are called *hyper-parameters*. In full Bayesian models the hyper-parameters are treated the same as the other parameters of the model and estimated together with them. LPD, however, is what we call an Empirical Bayes model, which means a model for which the hyper-parameters are estimated independently, using the data at hand. The estimated values for the parameters are then provided for the model parameter training.

Introducing informative priors, changes the equations for the MAP estimation of μ_{gk} and σ^2_{gk} parameters, as follows:

$$\mu_{gk} = \frac{\sigma^2_\mu \sum_{a=1}^{A} Q_{gka} e_{ga}}{\sigma^2_{gk} + \sigma^2_\mu \sum_{a'=1}^{A} Q_{gka'}}, \tag{3.12}$$

$$\sigma^2_{gk} = \frac{\sum_{a=1}^{A} Q_{gka}(e_{ga} - \mu_{gk})^2 + 2s}{\sum_{a'=1}^{A} Q_{gka'}}. \tag{3.13}$$

The other equations remain the same a for the MLE solution.

3.2.1.3.4 The LPD algorithm

As we can see above, the update equations for variational parameters Q_{gka} and γ_{ak} are inter-dependent with the model parameters. The variational parameters depend on the model parameters, and also the model parameters depend on the variational parameters. In order to estimate both sets of parameters, an iterative update procedure, similar to the EM algorithm, can be employed.

More specifically, each of these parameters are initialised to suitable starting values. The μ_{gk} parameters are set to the average expression of the gene g across all samples in the datasets, the variances σ^2_{gk} are set to the variance of gene g across all samples, the values for α are all set to 1, while the values for the γ_{ak} parameters are initialised to a positive random number.

The values of the hyper-parameters σ_μ are always set to 0.1 as it has been empirically determined that they have a negligible effect on the results [209]. The other hyper-parameter, s, however, seems to have a strong effect on the performance and needs to be carefully chosen [209]. One of the ways of doing it is to estimate the 10-fold cross-validation likelihood of the model for various values of s and to choose the value for which the maximum likelihood is obtained. We will illustrate this aspect in more detail in Chapter 5.

Having set the values for the hyper-parameters and the initial values for the parameters, an iterative two-step update procedure is performed, similar to the E and M steps in the EM-algorithm. In the first step the values of the variational parameters Q_{gka} and γ_{ak} are updated, as described by Equations 3.5 and 3.6, based on the current values of the other parameters. In the second step the values for μ_{gk}, σ^2_{gk} and α are updated, as indicated by Equations 3.7, 3.8 and 3.9 for the MLE solutions and by Equations 3.10, 3.11 and 3.9 for the MAP solution. The second steps uses the values of the variational parameters estimated in the first step. This iterative algorithm is guaranteed to converge after a finite number of iterations [210].

In the end the parameters μ_{gk} and σ_{gk} will describe the distribution of the expression level of the gene g in process k. In our analysis, based on these estimates, we can describe the genetical characteristics of each subtype of prostate cancer. We can, for example, identify which genes are up-regulated or down-regulated in a given process, or we can see if different subtypes of prostate cancer result in similar or different expression profiles.

Another very important set of results are the values for γ_{ak}, which are approximations of the mixture distribution, θ_k. For each sample, a, the value of γ_{ak} indicates an estimated contribution of the process k to the expression profile of the sample (Figure 3.4).

3.2.1.3.5 Choosing LPD parameters The MLE approach is a simpler version of the LPD model, which, given a dataset and the number of processes believed to underlay the data, is able to estimate for each sample the contribution of each process to its observed expression. However if the number of processes given as input to the model is larger than the number of processes inherent in the data, the model can fail to find the best representation for each process. It is said that the model overfits the data. Conversely, if the number of processes provided to the model is lower than the real number of processes, the model also can fail to find a good representations of the data. In this case it is said that the model underfits the data.

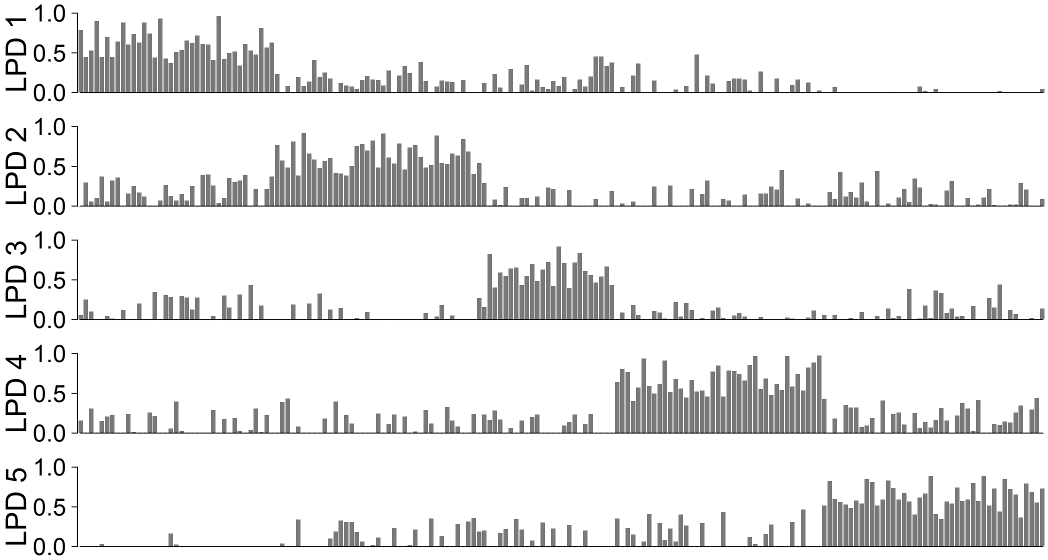

Figure 3.4 An illustration of the LPD classification on a prostate cancer dataset. The γ_{ak} values obtained after fitting the MAP version of an LPD model with 5 processes to a prostate cancer dataset, that will be described in detail later. Each horizontal panel, denoted LPD1, to LPD5, correspond to one of the 5 LPD processes. For each panel, k, corresponding to the LPD process k, the x-axis represent samples $1 \leq a \leq A$, while the y-axis represents the estimated value for γ_{ak}. The values of γ_{ak} have been normalised such that, for a given sample a, $0 \leq \gamma_{ak} \leq 1$ and $\sum_{k=1}^{K} \gamma_{ak} = 1$.

The MAP model, on the other hand, is slightly more elaborate than the MLE version. Besides the number of processes, it incorporates several additional parameters, which, if suitably chosen, protect the model from overfitting. We note, however, that the additional parameters do not prevent the model from underfitting. Numerical experiments [209] have indicated that, amongst the additional parameters, one in particular needs to be carefully chosen, while the others are set to some predefined values, as they have little impact on the results. This is the parameter s, described in Section 3.2.1.3, which is the prior for the variances σ_{gk}. Throughout the remainder of this chapter we will refer to this parameter as *sigma*.

The MAP version is more suitable for performing the final classification of the data, as in general it gives better solutions than the MLE model, as we will illustrate later. However, it needs to be provided with two parameters, the number of processes and a parameter which we denote *sigma* - the s parameter in Section 3.2.1.3.4. In order to estimate them, a MAP model needs to be trained for each of possible combination of the two parameters. More specifically, for each of the two parameters we set a range of values which are probable to be satisfactory for the model and, for each combination of values, we fit a model. The model that fits the best, i.e. the model that yields the maximum hold-out log-likelihood estimate, is then chosen for the final classification.

However, LPD is a quite computationally intensive method. In our evaluations, for example, an average performance computer needs around 24 hours to fit a single LPD model an a dataset of 300 samples. Therefore, varying both parameters in the same time is computationally difficult, due to large number of models that need to be fit.

A more efficient approach to deal with to this issue is to split the choice of parameters into two steps. First, we employ the MLE model to estimate the number of processes, as it does not need the sigma parameter. Once the number of processes is chosen, it is easier to determine the value of sigma alone, using the MAP model.

More specifically, the first step consists in determining the number of processes underlying each dataset by fitting a MLE model for different choices of the number of processes. In our case, we assumed that each dataset can have between 2 and 15 inherent processes. For each of these numbers we calculate the hold-out log-likelihood of the MLE model, which gives an indication about how well the model fits the data. We then select the number of processes at which the log-likelihood peaks.

Once we determined the number of processes, the second step consists in choosing a suitable value for the sigma parameter. The MAP model can be used for this undertaking. More specifically, we set a range of possible values for sigma. Then, for each value in the range we fit a MAP model. As before, we choose the value of sigma for which we obtain the maximum hold-out log-likelihood.

For a more robust choice of the parameters that are to be used for the final model, a third step can be derived. At this stage, as we have found a satisfactory value for sigma, we can fit a MAP model to all possible number of processes (2-15 in our case) to see how this compares to the MLE estimation. As sigma prevents the MAP model from over-fitting, but not from under-fitting, we would expect to see an increase of the likelihood up to a point, at which the MAP model does not under-fit the data anymore. After this point we would expect the likelihood to remain at about the same level, as the model is prevented by sigma to over-fit.

The step three can be useful to validate the number of processes chosen at step one. The process at which the MAP likelihood reaches the plateau indicates the probable number of processes inherent in the data.

3.2.2 Supervised learning methods

3.2.2.1 Random forests

We base the discussion in this subsection on the work of Breiman [213] and Breiman and Cutler [214].

The random forests algorithm is an ensemble classification method. Ensemble classification methods, also referred to as meta-classifiers, classify objects by aggregating the results of a collection (ensemble) of independent predictors. The aim of meta-classifiers is to obtain a more accurate classification than the component classifiers alone.

Random forests work by growing an ensemble of decision trees (a supervised classification technique that we will briefly describe below). Each tree provides a classification of a new sample. Informally, it is said that each tree votes for a class. The sample is assigned to the class that obtains most votes.

3.2.2.1.1 Decision trees Decision trees are a supervised classification method that work by organising a set of attributes (features) in a rooted tree structure. Each non-leaf node represents one or several attributes being tested. Each branch represents the outcome of the test, while the leaf nodes represent the class labels. An illustration is shown in Figure 3.5. The decision tree determines if a day is suitable for playing tennis, based on three weather characteristics: outlook, humidity and wind.

In the case of decision trees, a new object is classified by evaluating its attributes using the rules encoded by the tree. The evaluation starts with the root note. At each non-leaf node, one ore several attributes are tested. Depending on the value of the attributes being tested on each node, the evaluation continues on one of the branches, until it reaches a leaf node. The leaf node in which the classification stops represents the

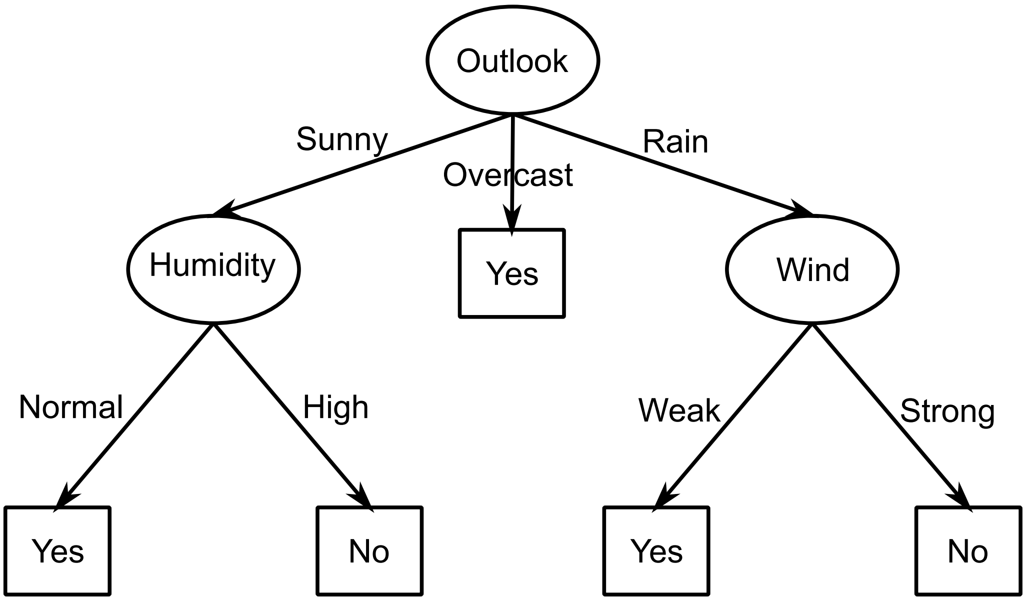

Figure 3.5 Example of a decision tree. The elliptic nodes represent the attribute being tested. The square nodes represent the class label. Adapted from Mitchell et al. [215].

class to which the object is assigned. For example, a day characterised by rainy outlook, high humidity and weak wind is classified as suitable for tennis (Figure 3.5). This decision is made by first evaluating the attribute in the root node, i.e. the *outlook*. In this particular case the outlook is *rain*, therefore the classification continues on the right branch. Next, the *wind* attribute is evaluated. It has the value *weak*, thus the evaluation continues on the left branch. That branch leads to a leaf node labelled *Yes*, indicating that the day is suitable for tennis.

3.2.2.1.2 Random forests algorithm Given a dataset set with N samples, each one with M attributes, random forests build each of their decision trees as follows:

1. N samples are selected, at random, with replacement, which will be used as training set for growing the tree;
2. from the list of M attributes, $m << M$ attributes are selected at random;
3. using the m attributes and the N samples a decision tree is constructed.

One of the key features of random forests is the selection with replacement of the training samples, used in the construction of each tree. When sampling N times with replacement from a set of N samples, some samples are selected more than once, while about a third of samples are not selected at all [216]. The collection of samples not selected is referred to as *oob (out-of-bag) data* and is used by random forests to calculate an unbiased classification error estimate. This is important as, by using the oob error rate, there is no need to perform cross-validation, as is the case with most supervised classification algorithms. The oob error is also used to calculate variable importances, i.e. a measure which tells how important a variable is for the overall classification.

The error rate of random forest depends on two aspects: the *correlation* between the trees and the *strength* of trees. The correlation estimate measures how similar are the classifications on average yielded by each pair of trees, across all samples in the dataset. Intuitively this tells if the trees output redundant classifications. Strength, on the other hand, tells how accurately each tree is classifying. Decreasing correlation and increasing strength lead to the decrease of error rate.

Both correlation and strength increase when the *m* parameter (also referred to as the *mtry* parameter - the number of attributes sampled for each tree) increases. Therefore, when using random forest it is important to choose a value for $mtry$ that gives a good trade-off between strength and correlation. The default value for this parameter is $\sqrt{M}$ for classification, and $M/3$ for regression. Another important parameter is the number of decision trees to grow (the *ntree* parameter). If too few trees are grown, the model might underfit the data. The default value for *ntree* is 500.

Random forests can be adapted to handle imbalanced datasets, i.e. datasets for which there is a significant difference in the size of classes. This is an issue for the classification algorithms as usually they try to optimize the overall error rate. Most of the times this will keep the error for the larger classes low, while letting the error of the small classes, which contributes little to the overall error, to be high.

For random forests there are two commonly used techniques for addressing this problem. One is to assign class weights inversely proportional to the class size, which are then used to weigh the contribution of the samples to the overall error. The other approach is to use stratified sampling, i.e. an equal number of samples is drawn from each class, regardless of the class size. This can be achieved by either over-sampling the smaller classes or down-sampling the larger classes.

3.2.3 Logistic regression

Regression models are some of the most popular techniques used for modelling the relationship between a continuous or discrete *outcome* (target) and a set of *predictors*. Depending on the types of the outcome and the predictor variables, different types of regression analysis are suitable. For example, *linear regression* is useful in modelling a continuous outcome variable as a linear combination of the predictors, while the *logistic regression*, which we will discuss next, based on Cox [217], is useful in modelling a binary target variable as a linear combination of a set of predictors.

Given a set of N mutually independent target random variables, $Y_1, Y_2, ..., Y_N$ taking the values $0, 1$ and a set of N vectors $X_1, X_2, ..., X_N$, where each vector X_i is a set of K predictor variables, $X_i = (X_{i1}, X_{i2}, ..., X_{iK})$, the logistic regression is concerned with modelling the relationship between $\theta_i = P(Y = 1)$ and the set of predictors X_i.

Since θ_i is a probability, and therefore takes values in the interval $[0, 1]$, its direct representation as a linear combination of predictors is unsuitable, as the linear combinations generally result in values in the interval $(-\infty, \infty)$. However, the values of θ_i can be mapped in the interval $(-\infty, \infty)$ using a link function. There are several functions that can be used for this transformation, but the one most commonly used is the *logit function*, defined as:

$$logit(p) = log\left(\frac{p}{1-p}\right) \tag{3.14}$$

Using the logit function, the logistic regression models the relationship between the outcome variable and predictors as:

$$logit(\theta_i) = log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \sum_{k=1}^{K} \beta_k X_{ik}, \tag{3.15}$$

which is the *logarithm of the odds ratio* (log odds ratio, for short). The parameters $\beta_1, \beta_2, ..., \beta_K$ are the regression coefficients, which describe how the log odds ratio modify with an unit increase in the corresponding predictor variable, and α is an intercept. By exponentiating the log odds ratio we obtain the *odds ratio* (OR), which is another measure very often used to describe the association between an outcome of interest and the predictors.

For a fitted model, given an instance of the predictors, the probability of the classes can be calculated as:

$$p(Y_i = 1) = \theta_i = \frac{exp\{\alpha + \sum_{k=1}^{K} \beta_k X_{ik}\}}{1 + exp\{\alpha + \sum_{k=1}^{K} \beta_k X_{ik}\}}, \tag{3.16}$$

or:

$$p(Y_i = 0) = 1 - \theta_i = \frac{1}{1 + exp\{\alpha + \sum_{k=1}^{K} \beta_k X_{ik}\}}. \tag{3.17}$$

3.3 LASSO

Shrinkage (regularization) methods are a form of penalising over-complex regression models by imposing restriction on the coefficients, forcing them to take lower values. One of the most popular shrinkage methods is the *LASSO* (least absolute shrinkage and selection operator) technique proposed by Tibshirani [218].

The LASSO technique imposes restrictions on the regression coefficients, such that only the coefficients corresponding to the most informative variables for the outcome are set to values different from 0, while the coefficients corresponding to less useful or redundant variables are set to 0. A reduced set of variables can improve the prediction accuracy, and also offers an easier interpretation of the model by selecting only a few variables Tibshirani [218].

LASSO is essentially a form of feature selection and can be applied to a wide range of models such as linear regression, logistic regression and the Cox model, which we will present later.

As presented in Friedman et al. [219], the LASSO technique can be incorporated in the estimation of the parameters of logistic regression. Given a logistic regression model with the predictor vectors $X \in \mathbb{R}^p$ and response variable $Y = \{0, 1\}$, defined as:

$$P(Y = 1|X = x) = \frac{1}{1 + exp\{\alpha + x^T \beta\}}, \tag{3.18}$$

the parameters $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ can be estimated by optimizing the objective function:

$$\min_{\alpha,\beta} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i(\alpha + x_i^T \beta) - log(1 + exp\{\alpha + x_i^T \beta\}) \right] + \lambda ||\beta||_1 \tag{3.19}$$

where the term in square brackets corresponds to the original log-likelihood of the logistic regression and the term $\lambda||\beta||_1$ corresponds to the LASSO regularization term. The factor $||\beta||_1$ of the LASSO regularization term corresponds to the L^1-norm of a p-dimensional vector $\beta = (\beta_1, \beta_2, ..., \beta_p)$, defined as $\sum_{i=1}^{p} |\beta_i|$.

The scalar $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage that is applied to the coefficients. When $\lambda = 0$, no coefficients are forced to 0. As λ increases, the number of coefficients forced to 0 increases.

The value for λ needs to be separately determined and supplied to the model. The most common approach in choosing it is to evaluate the k-fold cross validation prediction error at various values of λ and to select the value that provides the lowest cross-validation error.

3.4 Principal component analysis (PCA)

This section is based on Chapter I of Jolliffe [220]. Dimensionality reduction refers to that geometric technique that takes as input data into a multidimensional space and maps it to a lower dimensional space. Usually this mapping is irreversible, as during the conversion phase some features of the data are lost. Dimensionality reduction has many applications, the most common ones being the visualisation of the data in two or three dimensions and the extraction of the most informative features from data.

One of the most commonly used dimensionality reduction techniques is *principal component analysis* or PCA, for short. The main objective of PCA is to take data in a high dimensional space (i.e. a dataset containing objects with many attributes) and to map it to a lower dimensionality space, while retaining as much of the variance as possible. This is done by transforming the input set of variables to a new, smaller set of variables, which are linear combinations of the original variables. The new set of variables, referred to as *principal components*, are uncorrelated and are sorted decreasingly by the amount of variance from the original data which they explain. The first principal component contains the maximum amount of variance that can be projected into one direction (not necessarily parallel to the axes), the second principal component contains the next highest amount of variation, and so on.

Mathematically, the principal components correspond to the *eigenvectors* of the covariance matrix, Σ, of the original data. Briefly, given a dataset containing n-dimensional objects, with the directions $x_1, x_2, ..., x_n$, the covariance matrix, Σ is a $n \times n$ matrix, that describes the spread of data and the direction in which the it is spread. Each component $\Sigma_{ij} = cov(x_i, x_j) = \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}\left[x_j\right])^T\right]$, describes the covariance of variables x_i and x_j, while the components $\Sigma_{ii} = var(x_i) = \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_i - \mathbb{E}[x_i])^T\right]$ describe the variance of the components x_i, where $\mathbb{E}$ denotes expected value.

Given a general square matrix A, a vector $\vec{v}$ is called eigenvector if $A\vec{v} = \lambda\vec{v}$, where λ is a scalar called, *eigenvalue*. Eigenvectors and eigenvalues come in pairs. For a square matrix $n \times n$, there are n pairs of eigenvectors and eigenvalues. By convention eigenvectors are scaled, without loss of generality, so that they have length 1.

In the particular case of covariance matrices, the eigenvectors are orthogonal (perpendicular to each other) and point to the directions in which the data is most spread, while the eigenvalues are non-negative and are proportional to the variance of data in the direction indicated by the corresponding eigenvector. Therefore, the first principal component analysis corresponds to the pair of eigenvectors/eigenvalues with the maximum eigenvalue, the second component to the second largest eigenvalues and so on.

Intuitively, PCA projects the data on the directions indicated by the first k principal components, where k is the number of dimensions in which the data is to be mapped, which is less than or equal to the number of dimensions of the original data. We illustrate in Figure 3.6 the PCA decomposition of a synthetic dataset obtained by sampling points from two bivariate normal distributions with means (3, 3) and (6, 6) and covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$. We note that the original points are linearly separable. In Figure 3.6b the PCA is applied on the original data and the resulting data is projected on both principal components. We note that in this case all correlations between data points are preserved, due to the fact that the number of principal components equals the number of original dimensions and therefore they explain all the variance in data. The only transformation is the translation of the points into a new coordinate system with the x-axis corresponding to the first principal component and the y-axis corresponding to the second principal components. In Figure 3.6d the PCA transformed data is projected on the first principal component. Because the points have been projected on the dimension with maximum variance the structure of the original data is preserved. After this transformation the points from the two distributions are still separable. In Figure 3.6c the original points have been projected in one dimension, without being PCA transformed. Because the projection is not optimal, some useful information from the y-axis has been lost and the points are not separable anymore.

3.5 Survival analysis

The information presented in this section and its subsections is based on Chapters 1-6 of Kleinbaum and Klein [221].

Survival analysis is a set of statistical techniques concerned with studying data for which the outcome of interest is the *time* until an *event* occurs. In medical research some examples of common outcomes of interest include time to biochemical recurrence of the disease (this is the outcome of interest in our analysis), time to death, length of stay in the hospital, and time until a transplant is made. In practice, irrespective of the endpoint of the study, for simplicity, the time until the event occurs is referred to as *survival time*, while the occurrence of the event of interest is referred to as *failure*.

Usually, the participants in a study are monitored for a period of time, following some initial state of interest such as the date of diagnosis or the date of surgery (as is the case in our analysis). Some patients will experience failure during the observation period, but some will not. The patients who do not experience failure in the period

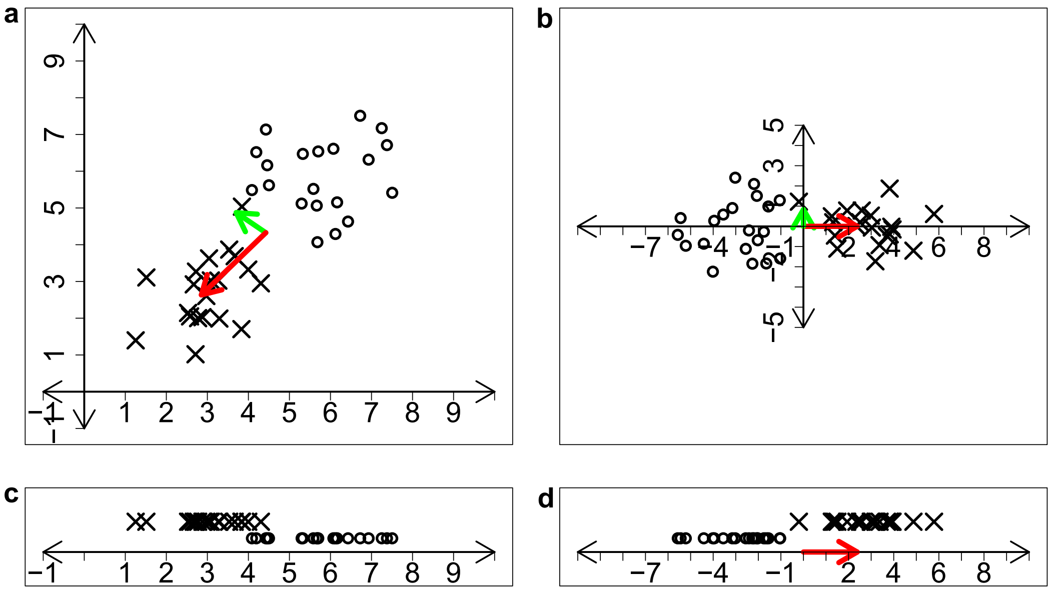

Figure 3.6 An illustration of PCA on a synthetic set of points: a) circles and crosses correspond to the sampled points from the two bivariate distributions, while the red and green arrow correspond to the direction of eigenvectors and the length of the arrow corresponds to the magnitude of the eigenvalues; b) PCA in which we project the data on both principal components; c) the projection of the original 2-dimensional points on the *x* dimension; d) PCA in which the data has been projected on the first principal component.

in which they are monitored are said to be *censored* at the last time they were under observation. The patients could still experience failure at some unknown point.

One important assumption when working with censored data, is that the censoring is a random effect that is not correlated with the outcome of interest.

There are four main reasons why censoring occurs: (i) the person does not experience failure before the end of the study, (ii) the person is lost to follow-up, (iii) the person withdraws at some point from the study, (iv) the person dies before the end of the study (if death is not the outcome of interest).

3.5.1 Kaplan-Meier (KM) survival curves

One way of modelling survival data is through the Kaplan-Meier (KM) survival curves, which are a representation of the survival probability as a function of time. Survival probability, denoted as $S(t)$, represents the probability that a participant survives past the time t (denoted $P(T > t)$). Theoretically the time t can take values between 0 and ∞, while $S(t)$ takes values between 1 when $t = 0$, and decreases towards 0 when t tends to ∞. In practice, however, we work with an estimate of this function denoted $\hat{S}(t_{(j)})$ (Figure 3.7), which is a step-function, rather than a smooth curve, due to the finite number of patients in a study. This makes the function remain constant in the intervals between two consecutive failure times in the dataset. Also, because the time t is never infinite, the function might not decrease all the way to 0.

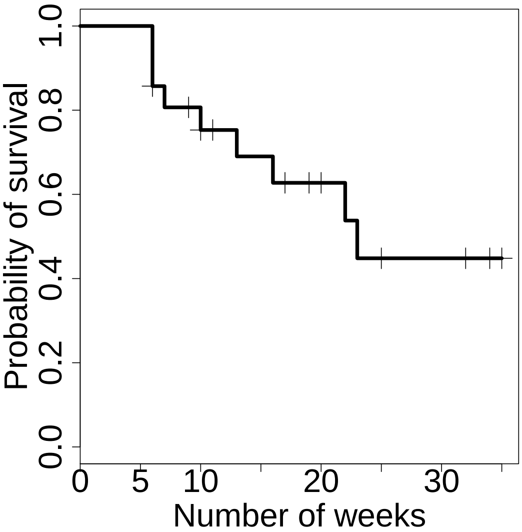

Figure 3.7 The KM plot corresponding to the $\hat{S}(t_{(j)})$ function calculated for the data in Table 3.1. The thin crosses represent time at which an observation has been censored.

We illustrate how the $\hat{S}(t_{(j)})$ function is estimated using an example dataset from Kleinbaum and Klein [221], presented in Table 3.1. The first column, $t_{(j)}$, contains the distinct time points at which failures occurred, sorted increasingly. The time can be counted using different time scales (hours, weeks, months, etc.) starting from an initial time of interest, which can be, for example, the date of diagnosis, or the date of surgery. However, it is important that the time is calculated in the same way for all the participants in the study. For the sake of illustration we consider that in this example the time is expressed in weeks since the initial event.

Table 3.1 Example of survival data, represented in a layout suitable for the computation of the KM curves: $\mathbf{t_{(j)}}$ represents the survival time, $\mathbf{n_j}$ represents the number of persons in the risk set, $\mathbf{m_j}$ the number of failures at each distinct time point and $\mathbf{q_j}$ represents the number of persons censored at each time point.

$\mathbf{t_{(j)}}$	$\mathbf{n_j}$	$\mathbf{m_j}$	$\mathbf{q_j}$	$\mathbf{\hat{S}(t_{(j)})}$
0	21	0	0	1
6	21	3	1	$1 \times 18/21 = .8571$
7	17	1	1	$.8571 \times 16/17 = .8067$
10	15	1	2	$.8067 \times 14/15 = .7529$
13	12	1	0	$.7529 \times 11/12 = .6902$
16	11	1	3	$.6902 \times 10/11 = .6275$
22	7	1	0	$.6275 \times 6/7 = .5378$
23	6	1	5	$.5378 \times 5/6 = .4482$
>23	0	-	-	-

The second column, n_j represents the number of patients still in the study at time $t_{(j)}$, including the patients that failed at time $t_{(j)}$. The patients still in the study at a specific time are referred to as the *risk set*. The third column, m_j, represents the number of patients that failed at time $t_{(j)}$, while the fourth column, q_j represents the number of observations censored starting from time $t_{(j)}$, up to, but not including $t_{(j+1)}$. We note that the first row, corresponding to week 0, is included even though there are no failures at that time. This row is always included because there might exist observations that were censored before the earliest failure time.

As described earlier, the function $\hat{S}(t_{(j)})$ represents the survival probability past time $t_{(j)}$. This probability can be expressed as product of two factors:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \times P(T > t_{(j)} | T \geq t_{(j)}), \tag{3.20}$$

The first factor represents the probability surviving past the previous failure, while the second factor represents the probability of surviving past the time $t_{(j)}$, given survival to at least time $t_{(j)}$.

The probability of surviving past time 0 is always 1. As we can see in the second row of the Table 3.1, there are 3 patients failing at the earliest failure time, $t_{(j)} = 6$, out of the total of 21 patients. This gives a probability of surviving past week 6 of $18/21 = .8571$. The next failure time is $t_{(j)} = 7$. We note that besides the 3 patients that failed at week 6, there is another observation censored between week 6 and week 7 (Table 3.1, row 2, column 4). This reduces the size of the risk set to 17. Out of the 17 patients remaining in the risk set, 1 fails at week 7. This means that the probability of surviving past week 7 conditioned on surviving to at least week 7 is 16/17, while the probability of survival past week 6 is .8571. The multiplication of these two factors gives the probability of surviving past week 7, which is $.8571 \times 14/15 = .8067$. The algorithm is repeated until no patients are left in the risk set.

3.5.2 Log-rank test

One of the main aims of the survival analysis is to assess if the KM survival curves corresponding to two or more groups of participants to a study are statistically significantly different, i.e. that they have significantly different rates of failure. It is for example important to know if a group of patients that are given a treatment have a significantly better outcome that the patients on placebo, or to know if the patients with a subtype of cancer are likely to develop recurrence at a faster rate than other patients.

The log-rank test is one of the most commonly used statistical tests that tests the hypothesis that several survival curves are statistically equivalent. It is essentially a version of χ^2 test that uses as test criterion a statistic based on the overall comparison of the KM curves.

The log-rank statistic is based on the difference between the observed and expected cell counts of each category, where categories represent the ordered failure times in the dataset. In Table 3.2 we present an example taken form Kleinbaum and Klein [221], containing the observations of 42 leukaemia patients, split into two groups, group 1 corresponding to 21 patients in placebo and group 2 corresponding to 21 patients on treatment. The data is ordered on the time of failure, $t_{(j)}$. The columns n_{ij} represent the risk set size of group i at time $t_{(j)}$, while columns m_{ij} represent the number of failed patients from the group i at time $t_{(j)}$. The columns m_{ij} represent the observed failures for each category. The expected cell counts from columns e_{ij} represent the expected failures for group i at time $t_{(j)}$ and are computed as:

$$e_{ij} = \left(\frac{n_{ij}}{\sum_i n_{ij}} \right) \times \sum_i m_{ij}, \tag{3.21}$$

which is the proportion of subjects in group i at time $t_{(j)}$ multiplied by the total number of failures at time $t_{(j)}$.

Table 3.2 An illustration of the steps involved in computation of the log-rank statistic.

		Risk set		**O**		**E**		**O − E**	
j	$\mathbf{t_{(j)}}$	$\mathbf{n_{1j}}$	$\mathbf{n_{2j}}$	$\mathbf{m_{1j}}$	$\mathbf{m_{2j}}$	$\mathbf{e_{1j}}$	$\mathbf{e_{2j}}$	$\mathbf{m_{1j} - e_{1j}}$	$\mathbf{m_{2j} - e_{2j}}$
1	1	21	21	0	2	$(21/42) \times 2$	$(21/42) \times 2$	-1.00	1.00
2	2	21	19	0	2	$(21/40) \times 2$	$(19/40) \times 2$	-1.05	1.05
3	3	21	17	0	1	$(21/38) \times 1$	$(17/38) \times 1$	-0.55	0.55
4	4	21	16	0	2	$(21/37) \times 2$	$(16/37) \times 2$	-1.14	1.14
5	5	21	14	0	2	$(21/35) \times 2$	$(14/35) \times 2$	-1.20	1.20
6	6	21	12	3	0	$(21/33) \times 3$	$(12/33) \times 3$	1.09	-1.09
7	7	17	12	1	0	$(17/29) \times 1$	$(12/29) \times 1$	0.41	-0.41
8	8	16	12	0	4	$(16/28) \times 4$	$(12/28) \times 4$	-2.29	2.29
9	10	15	8	1	0	$(15/23) \times 1$	$(8/23) \times 1$	0.35	-0.35
10	11	13	8	0	2	$(13/21) \times 2$	$(8/21) \times 2$	-1.24	1.24
11	12	12	6	0	2	$(12/18) \times 2$	$(6/18) \times 2$	-1.33	1.33
12	13	12	4	1	0	$(12/16) \times 1$	$(4/16) \times 1$	0.25	-0.25
13	15	11	4	0	1	$(11/15) \times 1$	$(4/15) \times 1$	-0.73	0.73
14	16	11	3	1	0	$(11/14) \times 1$	$(3/14) \times 1$	0.21	-0.21
15	17	10	3	0	1	$(10/13) \times 1$	$(3/13) \times 1$	-0.77	0.77
16	22	7	2	1	1	$(7/9) \times 2$	$(2/9) \times 2$	-0.56	0.56
17	23	6	1	1	1	$(6/7) \times 2$	$(1/7) \times 2$	-0.71	0.71
Total		0	0	9	21	19.26	10.74	-10.26	10.26

The computations of log-rank statistic uses the sum of observed failures minus expected failures, i.e. $O_i - E_i = \sum_j (m_{ij} - e_{ij})$, depicted in the last two columns of the last row of the Table 3.2. The log-rank statistic for two groups is computed as:

$$\text{Log-rank statistic} = \frac{(O_i - E_i)^2}{Var(O_i - E_i)}, \tag{3.22}$$

where i represents one of the groups. It does not matter which of the two groups is selected as they yield exactly the same results. The calculation of the log-rank statistic for more than three groups can be extended by generalizing the above equation. However we will not get into further details.

Having computed the statistic, a p-value for the test can be easily derived, as the log-rank statistic is approximately χ^2 with $G - 1$ degrees of freedom, where G is the number of groups.

3.5.3 Cox proportional hazards (PH) model

The survival time for an observation might be influenced by more than one variable. For example, in our study, besides the assignment of patients to a more aggressive subgroup, other factors could also significantly influence survival, such as the Gleason grade or pathological stage. One might be interested in accounting for these factors when studying the effect of a variable of interest.

The Cox proportional hazards (PH) model [222] is one of the most popular statistical models for performing multivariate survival analysis. It is a regression model designed to investigate simultaneously the effects of several explanatory variables on the survival time.

The use of the Cox PH model can achieve three main aims: (i) to test if a variable is a statistically significant factor that influences survival, after adjusting for the effects of other covariates; (ii) to provide a point estimate, called *hazard ratio*, that describes how the survival is impacted when the value of a variable changes; (iii) and to provide a confidence interval for the hazard ratio.

Central to the Cox PH model is a function, called the *hazard function*, defined as:

$$h(t,\mathbf{X}) = h_0(t)exp\left\{\sum_{i=1}^{p}\beta_i X_i\right\}, \tag{3.23}$$

where $\mathbf{X} = (X_1, X_2, ..., X_p)$ is a set of p explanatory variables and $(\beta_1, \beta_2, ..., \beta_p)$ are a set of p coefficients corresponding to them. This function models the *hazard rate* of an individual with a specific set of values for the explanatory variables as a function of time. The right side of the above equation has two factors. The first one, $h_0(t)$, called the *baseline hazard function*, is only a function of time (it does not involve any X). Intuitively, it explains how the hazard changes as a function of time prior to considering the explanatory variables (the hazard function equals the baseline hazard when all Xs are 0). The second factor is a function of explanatory variables, which does not involve the time. It is basically the exponential of a linear combination of the explanatory variables.

For a given dataset, the parameters $(\beta_1, \beta_2, ..., \beta_p)$ of the model can be estimated using a partial maximum-likelihood approach. The estimated parameters will be further denoted as $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$.

The hazard ratio describes the hazard rate of one individual with a specific instance of the explanatory variables, relative to another one with another instance of the explanatory variables. Using the above definition for the hazard function, the hazard ratio can be easily computed for two set of instances, X^* and X, of the explanatory variables as:

$$\widehat{HR} = \frac{\hat{h}(t,X^*)}{\hat{h}(t,X)} = exp\left\{\sum_{i=1}^{p} \hat{\beta}_i(X_i^* - X_i)\right\}. \tag{3.24}$$

Following from the above equation, the formula $exp\left\{\hat{\beta}_i\right\}$ gives the hazard ratio corresponding to the explanatory variable X_i, after adjusting for the effect of other covariates. Informally, this hazard ratio tells us what are the odds of experiencing faster failure with every one unit increase in the value of X_i, after adjusting for the effects of the other covariates. In the case of categorical variables, the hazard ratio gives the odds of experiencing faster failure for individuals from one category, relative to a baseline category, after adjusting for other covariates.

In Table 3.3 we illustrate an example of multivariate analysis performed on a prostate cancer dataset, which we will describe later, using the Cox PH model. We assess the influence of a binary variable of interest, denoted as *DESNT/non-DESNT*, on the survival time, after adjusting for the effects of three dichotomised covariates (Pathological Stage, Gleason grade and PSA). The values in the second column represent the hazard ratios for each variable, while the values in the second and third column represent the 95% confidence intervals for the hazard ratios. The values in the fourth column represent the p-values that indicate if each covariate is significantly associated with the survival time. We note that for the variable of interest, the hazard ratio is 3.8041, indicating that the odds of experiencing failure is almost four times higher in the patients from the *DESNT* group, compared to the patients in the *non-DESNT* group. As indicated in the last column, this is the only variable significantly associated with the survival time (p-value ≤ 0.05), after correcting for the effects of the other covariates.

Table 3.3 An illustration of the Cox regression model on a prostate cancer dataset, where we assess the influence of the binary variable *DESNT/non-DESNT*, adjusting for the effects of three dichotomised covariates (Pathological Stage, Gleason grade and PSA).

Covariate	**Hazard Ratio**	**CI lower 0.95**	**CI upper 0.95**	***p*-value**
Non-DESNT/DESNT	3.8041	1.889	7.661	0.000183
Path Stage: T1-T2/T3-T4	1.6947	0.8409	3.415	0.14014
Gleason: $\leq 7/ > 7$	2.0393	0.9881	4.209	0.05392
PSA: $\leq 10/ > 10$	1.9233	0.9753	3.793	0.059042

3.5.3.1 Cox PH model assumption and testing that assumption

The validity of the Cox PH model depends on the assumption that the hazard ratio is constant over time, or alternatively, that the hazard for one individual is proportional to the hazard for any other individual over time, referred to as the *PH assumption*. This implies that the effect of the explanatory variables is constant in time, i.e. that the variables are *time-independent*.

There are three approaches to check if the PH assumption is met: (i) graphical, (ii) goodness of fit and (iii) the use of time-dependent variables. We will briefly discuss each one next.

3.5.3.1.1 Graphical approaches For graphical approaches, there are two informative plots that can help determining if the PH assumption is respected. The first type of plot is the *log-log survival plot*. Log-log plots are built based on the *Cox model survival function*, which is closely related to the hazard function, and is defined as:

$$S(t,\mathbf{X}) = [S_0(t)]^{exp\{\sum_{i=1}^{p} \beta_i X_i\}}, \tag{3.25}$$

where $S_0(t)$ is a baseline survival function. Taking $-ln\,[-ln\ S]$, we obtain the *log-log survival curves*:

$$-\,ln\,[-ln\ S(t,\mathbf{X})] = -\sum_{i=1}^{p} \beta_i X_i - ln\,[-ln\ S_0(t,\mathbf{X})]\,. \tag{3.26}$$

When considering two individuals $\mathbf{X}_1 = (X_{11}, X_{12}, ..., X_{1p})$ and $\mathbf{X}_2 = (X_{21}, X_{22}, ..., X_{2p})$, taking the difference between their log-log survival curves, we obtain:

$$-\,ln\,[-ln\ S(t,\mathbf{X}_1)] = -ln\,[-ln\ S(t,\mathbf{X}_2)] + \sum_{i=1}^{p} \beta_i X_i. \tag{3.27}$$

Following from the above equation, if we plot the log-log survival curves as a function time, we should obtain two parallel curves, as the term $\sum_{i=1}^{p} \beta_i X_i$ is independent of time. If the curves converge, or intersect, it suggests that the PH assumption is not respected. We present two example of two log-log survival curves taken from our data in Figure 3.8a,c.

The second graphical approach is to build observed vs. expected survival curves for each category of a variable. If the PH assumption is met, the observed vs. expected survival curves should be "close" to each other. The observed survival curves are plotted using the KM approach, while for the expected curves we fit a Cox model with only

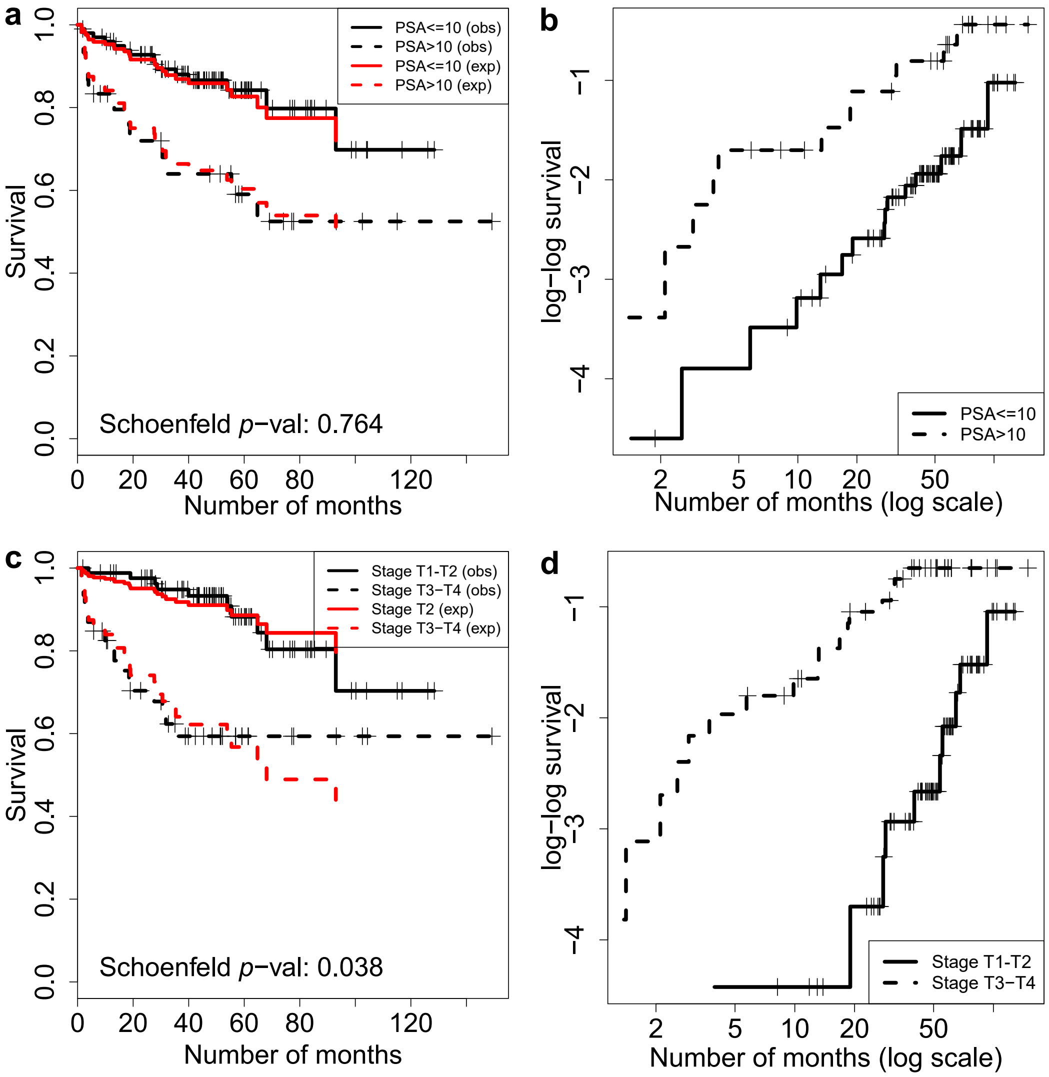

Figure 3.8 The use of graphical approaches in verifying the PH assumption for a variable (PSA) for which the assumption is valid (a,b), and for a variable (stage) for which the assumption is not valid (c,d): a) the log-log survival curves corresponding to the two levels of the PSA variable; b) the observed vs. the expected survival curves corresponding to the two levels of the PSA variable; c) the log-log survival curves corresponding to the two levels of the stage variable; d) the observed vs. the expected survival curves corresponding to the two levels of the stage variable. Note that in the first case the log-log survival curves are approximately parallel, while the observed vs. expected curves are close to each other, suggesting that the PH assumption is met. In the second case the log-log survival curves converge and also the lower observed curve departs from the expected curve.

one predictor, i.e. the variable being assessed. We present two examples of observed vs. expected survival plots in Figure 3.8b,d.

A big drawback of the graphical approaches is the subjectivity involved in determining how "parallel", or "close" are the survival curves.

3.5.3.1.2 The goodness of fit approach (GOF)

The GOF approach has the advantage that it provides a statistical way of assessing the PH assumption for a specific variable.

One of the most commonly used GOF approaches is based on Schoenfeld residuals [223]. Briefly, for each individual that experiences failure at time t, a Schoenfeld residual at time t is calculated for each covariate. When the PH assumption is met, the Schoenfeld residuals corresponding to the examined variable should be uncorrelated with the survival time. Therefore, to test the PH assumption of a variable, a statistical test that tests the null hypothesis that the correlation between the Schoenfeld residuals and survival time is 0 is performed.

The GOF approach provides a more objective way of testing the PH assumption. On the other hand, graphical approaches are useful in determining the underlying causes of departure from the PH assumption. Thus, it is generally recommended to use both procedures when evaluating the PH assumption

3.5.3.1.3 The use of a time-dependent variable

Another approach to test the PH assumption of a variable is to incorporate in the Cox model an interaction term between the variable being assessed and time. Then it is tested if the interaction term is significantly associated with the outcome. If it is significant, this suggests that the PH assumption is not respected, i.e. that the variable is time-dependent.

More specifically, we extend the Cox model for a predictor X to include a term that is the product of the predictor with a function of time, $g(t)$. This yields the hazard function:

$$h(t,X) = h_0(t)exp\{\beta X + \delta X \times g(t)\} \tag{3.28}$$

When $\delta = 0$, the hazard function of this extended model reduces to its form in the basic version of the Cox PH model, suggesting that the PH assumption is met.

3.5.4 The extended Cox model for time-dependent variables

As shown earlier, the Cox model can be extended to include both time-dependent variables and time-independent covariates. The new hazard function takes the form:

$$h(t,\mathbf{X}) = h_0(t)exp\left\{\sum_{i=1}^{p_1}\beta_i X_i + \sum_{i=1}^{p_2}\delta_i X_i(t)\right\}, \tag{3.29}$$

where $(X_1, X_2, ..., X_{p1})$ represent the time-independent covariates with coefficients $(\beta_1, \beta_2, ..., \beta_{p1})$, $X_1(t), X_2(t), ..., X_{p2}(t)$ represent the time-dependent covariates, with coefficients $(\delta_1, \delta_2, ..., \delta_{p2})$. This leads to the definition of the hazard ratio for to sets of covariates, $\mathbf{X}^*$ and $\mathbf{X}$, defined as:

$$\widehat{HR} = exp\left\{\sum_{i=1}^{p1}\hat{\beta}_i(X_i^* - X_i) + \sum_{i=1}^{p2}\hat{\delta}_i(X_i^*(t) - X_i(t))\right\}. \tag{3.30}$$

In the extended Cox model the PH assumption is no longer satisfied. However, an important assumption of the extended model is that the effect of a time-dependent variable $X_i(t)$ at time t depends only on the value of the variable measured at time t, not later or earlier.

3.6 Pathway analysis

One of the major challenges in the experiments that produce lists of genes that have different properties between two conditions (such as genes that are differentially expressed, or which exhibit different proportions of mutations and fusions between two conditions) is to extract information about their biological functionality. For this purpose several bioinformatical approaches have been developed. One of the most common technique, known as over-representation analysis, or *pathway analysis* for short, identifies over-represented categories of genes that share a similar function [224].

Pathway analysis relies on existing gene annotation databases, such as Gene ontology (GO) [225, 226], Kyoto encyclopaedia of genes and genomes (KEGG) [227] and Reactome [228, 229], that contain information about genes and gene products, together with the relationships between them. The GO database organises the functionalities into three domains: cellular components, molecular functions and biological process. Of particular interest for the current project is the biological process domain, which provides information about the processes involved in the functioning of cells, tissues, organs and organisms. KEGG and Reactome store manually curated biological pathways.

For each functional category (pathway, process) in the database the *background frequency* is compared with the *sample frequency* [230]. The background frequency is the number of genes annotated in the category, relative to the entire background set of genes (all the genes in the database) [230]. The sample frequency is the number of genes in the input set annotated to that category [230]. A statistical test (usually a χ^2 test) which assess the null hypothesis that the sample frequency is equal to the background frequency, with the alternative hypothesis that the pathway genes are under/over-represented in the input list of genes, is then performed. Since an independent test is performed for each pathway, the resulting p-values are usually adjusted for multiple comparisons.

3.7 Discussion

In this chapter we have presented the main computational and statistical approaches used in this thesis. We will illustrate in Chapter 5 how the LPD model has been used to identify several group of patients with distinct gene expression profile. We will also show how this classification compares to other unsupervised machine learning techniques presented here, such as k-means and hierarchical clustering. We will also illustrate the use of the random forest and LASSO techniques introduced here to generate a gene expression signature.

The survival analysis models described in this chapter will be used in Chapter 4 and Chapter 5 to illustrate the association between the groups of patients identified by the machine learning algorithms and their clinical outcome. The pathway analysis will be used to link sets of genes identified by the methods presented in Chapter 4 and Chapter 5 to the biological functionality.

Chapter 4

Identification of transcriptional alteration candidates using exon microarrays

4.1 Summary

Chromosomal rearrangements, read-through transcription and several other mechanisms can disrupt the normal expression of some genes, which can, in turn, lead to the development and progression of cancer. The identification of such events can improve the understanding of cancer and can help the development of new management strategies.

In this chapter we present a novel technique for identifying genes potentially involved in aberrant transcriptional events in prostate cancer, using the data provided by the exon microarrays. We describe in detail how our method works and how it compares with other existing methods developed for this purpose.

We also illustrate our new technique on three prostate cancer datasets. In these datasets our method identifies alterations in many genes previously involved in chromosomal rearrangements and read-through transcription, as well as several other novel candidates. As the datasets we analysed provide linked clinical data, we have been able to correlate some known and novel candidates with the clinical outcome of patients.

4.2 Background

Chromosomal rearrangements are a class of complex mutations, that result in changes in the structure and the number of chromosomes. Many times the chromosomal rearrangements ligate together components from two or more separate genes, resulting in events called *gene fusions*. Gene fusions, in turn, can generate chimeric transcripts with important roles in cancer progression. For example, the recurrent translocation known as the *Philadelphia chromosome*, which occurs in around 90% of chronic myelogenous leukaemia cases, results in the *BCR-ABL1* gene fusion, which encodes a hybrid protein that causes the uncontrollable multiplication of cancer cells [231].

Chromosomal rearrangements can also result in the disruption of genes with key roles in the cell. One such example is the recurrent deletion in prostate cancer of region p23 of chromosome 10, which causes the inactivation of *PTEN*, a tumour suppressor gene [57].

Additionally, it is well established that alterations in transcript splicing are associated with cancer development and several mechanism of alteration have been observed including trans-splicing (also known as read-through transcription) [232], and the use of alternative transcription start sites [233]. Trans-splicing refers to the formation of hybrid transcript, resulted from the juxtaposition of some exons from two consecutive genes through a transcriptional process that does not involve chromosomal rearrangements [234]. One such example is the fusion transcript *SLC45A3-ELK4* observed in prostate cancer [232].

In prostate cancer the *TMPRSS2-ERG* fusion occurs in 40-55% of prostate cancers [32, 127–131]. The fusion is an early event in the development of prostate cancer, as it is found in a high percentage of HGPIN (high-grade prostatic intraepithelial neoplasia - a precursor of prostate cancer) [136], but is insufficient to induce the formation of prostate cancer on its own [137]. Also, it seems to contribute to cell invasion and migration [136]. However, the usefulness of *TMPRSS2-ERG* fusions as a biomarker in predicting clinical outcome is controversial. A number of studies reported an association between the *TMPRSS-ERG* fusions and poor outcome [138–140], but others found no association [33–35].

Often gene fusions, read-through transcription and gene truncations alter the normal expression of transcripts, leading to different expression patterns along the exons of a given gene. This is the case with the *TMPRSS2-ERG* fusion in prostate cancer. *TMPRSS2* is an androgen regulated gene, and hence is highly expressed in prostate cancer. The fusion of the 5' regions of *TMPRSS2* to the 3' domain of *ERG* disrupts the normal expression of both genes. *ERG* is normally expressed at much lower levels than *TMPRSS2*. After fusion the exons located after the breakpoint, toward the 3' end, which are translocated at the *TMPRSS2* locus, become expressed at higher levels, relative to the exons located before the breakpoints. Conversely, the 3' exons of *TMPRSS2* which are either deleted or translocated to a less expressed locus, can become less expressed compared to the exons before the breakpoint, which remain regulated by the *TMPRSS2* promoter.

This concept can be generalised to fusions resulting from rearrangement or read-through transcription of two genes that in normal conditions have quite different expression levels. In the less expressed gene, the exons located towards the 3’ end, after the breakpoint, are expressed more highly than the exons towards the 5’ end, as they become regulated by the promoter of the more highly expressed gene. Conversely, for the more highly expressed gene, the expression level of the translocated exons can decrease.

Besides these events, there might be many other mechanisms, such as alternative splicing (the use of alternative transcription start sites), that can result in a different expression pattern of the exons towards the 5’ and 3’ ends of a gene.

We can identify these alterations with the help of exon microarrays. As the microarrays measure the expression level of most exons in the human genome, we can study if there is a shift in their expression along a gene. Several examples of possible scenarios in which the events described above can result in jumps are presented in Figure 4.1. The fusion of two genes resulted from balanced chromosomal rearrangements represented in Figure 4.1a, trans-splicing (Figure 4.1b) and truncation (Figure 4.1c) can all result in the expression patterns presented in Figure 4.1d.

For simplicity, throughout the remainder of this thesis we will refer to the shifts in the expression of the exons within a gene, such as the patterns illustrated in the right panels of Figure 4.1d, as *jumps*. The shifts that result in a higher expression of the exons after the breakpoints, such as the one in the bottom area of the panel, will be refer to as *step-up jumps*. While the opposite pattern will be referred to as *step-down jumps*. Also, we will refer to the genes which exhibit such jumps as *candidates*.

The established techniques for the identification of gene fusions, such as RT-PCR, FISH or RNA-seq, can probably identify fusions and other abnormalities more accurately than the methods based on exon microarray data, such as the one we will present here. However, the biological techniques such as RT-PCR or FISH, require an experiment for each gene in each sample. Therefore they are more suitable for the validation of a limited list of target genes, rather than genome-wide discovery of the fusions. Also, as each experiment is labour intensive, it is quite difficult to obtain a number of samples suitable for correlation with clinical data, especially if the fusions appear at a low frequency. RNA-seq on the other hand, has the advantage of a more reliable identification of fusions and also can be used for genome-wide identification of fusions. However, it is a relatively new technology, and currently there are few large-cohort RNA-seq datasets with associated long-term follow-up data.

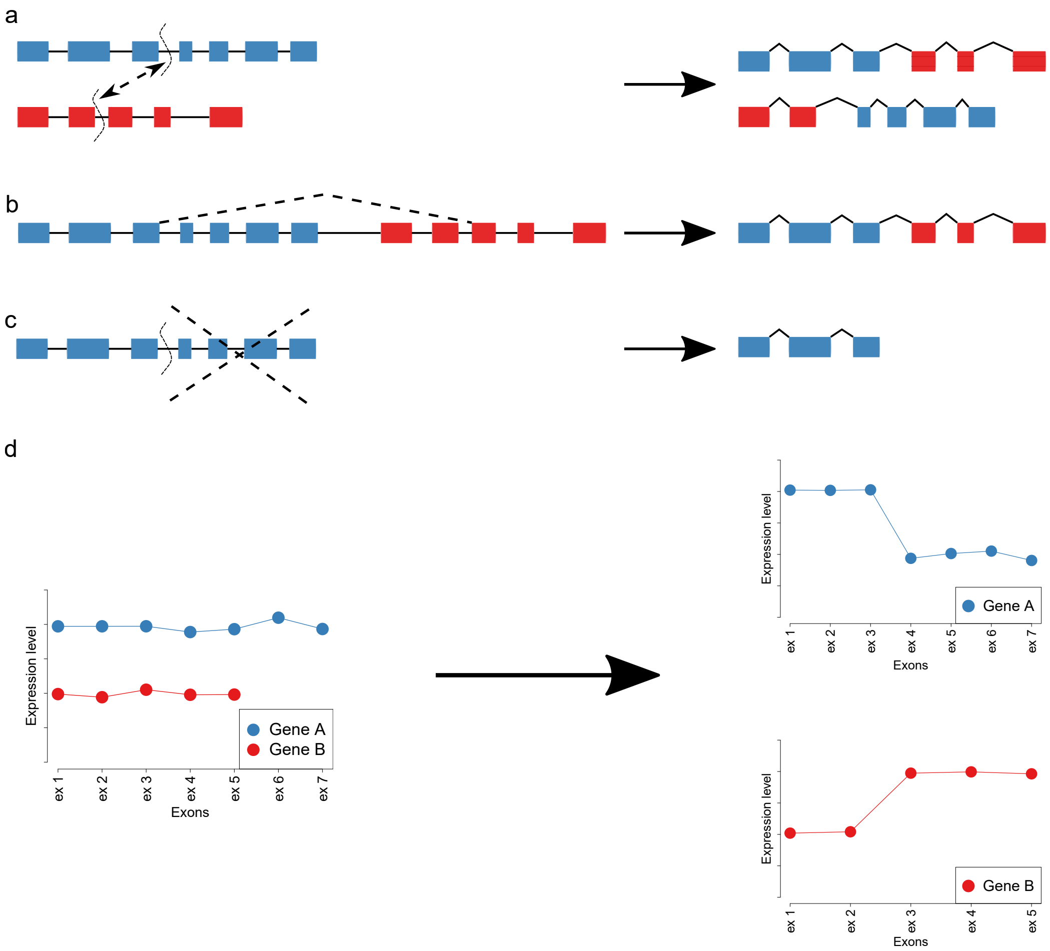

Figure 4.1 Examples of transcriptional alterations that can lead to different expression levels of the exons within a gene: a) gene fusion resulting from balanced translocation; b) trans-splicing; c) gene truncation; d) the resulting expression patterns.

The identification of fusion candidates using exon microarray data has the advantage that it can be applied genome-wide, in large existing datasets, with long term follow-up data. However, we acknowledge that it does not provide a definitive validation of the transcriptional abnormalities, and the identified candidate genes need to be further characterised using established methods. This is also true for the fusion candidates identified with RNA-seq data.

4.2.1 Previous approaches

High-density microarrays such as the *Affymetrix GeneChip*® *Human Exon 1.0 ST Arrays* (exon microarrays) were initially designed for the study of different isoforms of genes generated by alternative splicing. Later, Jhavar et al. [235] suggested that exon-level analysis using exon microarrays could also be useful to detect gene fusions candidates. The approach was focused solely on detecting *TMPRSS2-ERG* fusions in prostate cancer, and was restricted to the detection of *ERG* alterations. In this approach, the core probesets corresponding to ERG exons were normalised to a reference sample by taking the $\log_2$ ratio relative to the corresponding probesets in a reference sample. When more than one probeset mapped to an exon, the median value was taken. In order to determine which samples exhibit *ERG* jumps, two t-tests were used. One tested whether exons 2 and 3 (there are no core probesets mapping to exon 1) had a significantly lower expression than the exons 4-11, while the second t-test assessed whether the expression of exons 4-11 is greater than 0. The second test is based on the fact the $\log_2$ ratio should distribute the exons of the non-fused samples around 0.

The analysis was based on 27 malignant samples and 3 non-malignant epithelial samples. Out of the 27 cancer samples, 15 samples were discovered to have significant jumps, all confirmed by RT-PCR analysis. Out of the remaining 12 samples that did not express significant jumps, *TMPRSS2-ERG* fusions were found by the RT-PCR analysis in two samples.

Lin et al. [236] used exon microarray profiling to confirm the presence of the *EML4-ALK* fusions in non–small cell lung cancer and, additionally, to discover it in breast and colorectal cancer. The approach was more general than the method of Jhavar et al. [235]. Starting from core probesets, the intensities of each probeset were normalised to a standard normal distribution across samples. For each sample in a given gene, the probesets were ordered by their genomic position. A Student's t-statistic comparing the intensity distribution of the exons before and after the putative breakpoint was then computed. A given sample was considered as expressing jumps in the given gene and sample if the maximum t-statistic was above a pre-defined threshold. Subsequently the fusion candidates were filtered using several criteria, leading to a list of candidates amongst which there was the *ALK* gene.

Li et al. [237] profiled lung cancer samples using exon microarrays, using a technique similar to Lin et al. [236], based on t-statistics, to identify a list of about 1,000 candidate genes. From this list, the genes encoding kinases (a class enzymes with critical role in many pathways) were targeted, as the kinases are known to be the most common oncogenic drivers [237]. One of the fusion candidates from the kinase family was the *RET* gene. Further biological validation identified that the jumps in *RET* gene were generated by the *CCDC6-RET* fusion.

Giacomini et al. [238] presented a study spanning multiple cancers. They present two approaches for detecting gene fusion candidates. One of them was designed to detect fusion candidates using exon microarrays. As in previous approaches, the probesets were centred using a $\log_2$ ratio. The fusion candidates were then identified using a Student's t-test at every putative breakpoint, similar to Lin et al. [236]. The method was successful, as several known fusions, including *BCR-ABL1*, *FIP1L1-PDGFRA* and *NPM1-ALK*, were identified. Also, the analysis revealed a set of novel fusion partners including *ROS1*, *SLC1A2*, *RAF1*, *EWSR1* and *CLTC*.

Wang et al. [239] presented a score-based method that led to the discovery of the *HEY1-NCOA2* fusion in sarcoma. The probes were normalized to a consistent scale, as in the previous approaches. Then, the fusion candidate genes were identified using a model that took into account two scores.

The first one was based on a z-score, computed as:

$$z\_score = \frac{s_i - \mu}{\sigma}, \tag{4.1}$$

where s_i represents the expression of a given probeset in the sample i, μ corresponds to the mean of signal of the probeset in all samples, excepting the sample i, while σ represents the standard deviation. The second score, based on the ranking of the signal levels for a probeset, was computed as:

$$p\_score = log\left(\frac{r_i}{1 - r_i}\right) \tag{4.2}$$

where $r_i = rank(s_i)/N$.

For a given gene in a given sample, the above two probeset-level scores were aggregated to obtain two gene-level scores, computed as:

$$FS_1 = \max_k \left[\frac{mean(z\_score_{k+1},...,z\_score_K) - mean(z\_score_1,...,z\_score_k)}{\sqrt{1/k - 1/(K-k)}} \right], \quad (4.3)$$

and:

$$FS_2 = \max_k \left[\frac{mean(p\_score_{k+1},...,p\_score_K) - mean(p\_score_1,...,p\_score_k)}{\sqrt{1/k - 1/(K-k)}} \right], \quad (4.4)$$

where $k < K$ represents the index of the exon in the gene.

A non-linear logistic regression model that models the fusion outcome as a function of the scores $FS1$ and $FS2$, was defined using the formula:

$$fusion.outcome \sim FS_1 + FS_2 + FS_1 * FS_2. \quad (4.5)$$

The logistic model was fitted to a training dataset for which the fusion status of two genes was known. Based on the trained model and the FS scores, a list of candidate genes was identified. The list was further reduced using several criteria, leading to a short set of candidates, amongst which was *NCOA*2. Using RACE (Rapid amplification of cDNA ends) and other biological approaches, the fusion partner *HEY*1 was identified and confirmed.

4.2.2 Motivation for our method

All the above methods are based on the same general approach. The probesets from exon microarray are mapped to their corresponding exons and normalised across samples, in order to centre them around 0. Next, a detection method that identifies jumps in the expression intensities of the exons towards the 3' end, relative to the expression of the exons towards the 5' end is applied to a set of genes. Usually a list of hundreds or even thousands of fusion candidate genes is generated by this approach. The list is subsequently filtered using various criteria, leading to a small set of candidates which are then analysed using wet-lab techniques that can confirm the fusion and identify the fusion partners.

There are two main types of detection methods: t-test based approaches ([235–238]) and the score based method of Wang et al. [239]. The t-test approaches are mainly based on applying t-tests at the positions of putative breakpoint, comparing the distribution of expression levels of the exons before and after the breakpoint. For a given gene in a given sample the maximum t-statistic/minimum p-value gives the probable position of a breakpoint. The gene is considered fused in the given sample if the *t*-statistic or the *p*-value meet a certain threshold. The score based approach of Wang et al. [239], on the other hand, combines two scores in a non-linear logistic regression model. The model is trained on known fusions, and afterwards is used to identify novel fusion candidates.

Given the success of the above methods, we decided to apply the same general approach, but using a novel method, to identify fusion candidates in several prostate cancer exon microarray datasets with linked clinical data. Our main purpose was to identify in the list of candidates several genes for which the jumps are correlated with the clinical outcome of the patients, hoping to identify biomarkers with potential to predict aggressive prostate cancer.

4.3 Materials

4.3.1 Datasets

The analysis was based on three prostate cancer microarray datasets that will be further referred to as ICR, Cambridge and MSKCC.

The ICR and Cambridge datasets are part of the same project, referred to as CancerMap, and have been created using fresh prostate cancer specimens obtained from a systematic series of patients who had undergone prostatectomy at the Royal Marsden NHS Foundation Trust and Addenbrooke's Hospital, Cambridge, UK. The relevant local Research Ethics Committee approval was obtained for this study. Frozen prostate slices were collected [240] and RNAs were prepared as described previously [235, 241] in two centres: Institute of Cancer Research (ICR) from London, UK and CRUK Institute, Cambridge, UK. Expression profiles were determined using 1.0 Human Exon ST arrays (Affymetrix, Santa Clara, CA, USA). The microarrays were processed at The Paterson Institute for Cancer Research, Manchester, UK, according to the manufacturer's instructions.

The ICR dataset contains 124 microarrays, from 81 patients and the Cambridge dataset contains 111 microarrays from 73 patients. For each sample there is only one corresponding microarray experiment, but for each patient there may be up to 4 samples, containing variable amounts of normal, stromal and tumour tissue.

The ICR and Cambridge datasets were generated using the same protocol for choosing patients and collecting clinical data. This minimises the risk of systematic biases in the clinical data. However, since the RNA samples were extracted in different centres, there are dataset-specific effects in the microarrays.

For the first part of the analysis we decided to process each dataset separately. The micorarrays have been preprocessed and candidates have been identified independently for each dataset. However, in the second part of the analysis, when we correlated the candidates with the clinical data, we merged the results obtained for ICR and Cambridge and clinical data in a larger dataset, referred to as CancerMap, as we will describe later.

The MSKCC dataset, published in Taylor et al. [242], is publicly available on the GEO repository under accession *GSE21032*. The submission consists of exon microarrays, CGH arrays and miRNA arrays. We restricted our analysis to the subseries *GSE21034* containing 370 *Affymetrix Human Exon 1.0 ST Array* experiments. The exon microarrays were generated in duplicates from 185 RNA sources. Some of the samples correspond to benign and tumour primary prostate tissue obtained from prostate cancer patients that underwent prostectomy at Memorial Sloan-Kettering Cancer Center (MSKCC), USA. Other samples come from prostate cancer metastatic tissues, and some other come from four prostate cancer cell-lines (VCaP, PC3, LNCaP and DU145) or two LNCaP derived xenografts.

A summary of the datasets is presented in Table 4.1.

Table 4.1 Dataset description.

	Nr. Microarrays		**Tumour**		**Benign**		**Stroma**	
	Total	**Unique**	**Total**	**Unique**	**Total**	**Unique**	**Total**	**Unique**
ICR	124	81	107	72	15	7	2	2
Cambridge	111	73	102	65	9	8	0	0
MSKCC	370	185	262	131	58	29	0	0

	Metastatic		**Cell-line**		**Xenografts**	
	Total	**Unique**	**Total**	**Unique**	**Total**	**Unique**
ICR	0	0	0	0	0	0
Cambridge	0	0	0	0	0	0
MSKCC	38	19	8	4	4	2

4.3.2 Clinical data for survival analysis

In the ICR and Cambridge datasets, there might be up to four samples per patient, extracted from tissue that contained variable amounts of tumour, tissue and stroma. In order to perform a meaningful survival analysis, for each dataset we restricted the survival analysis to one microarray per patient. We considered that the observed expression profile of samples is proportionally influenced by the percentage of each type of tissue. Therefore, for each patient we selected the sample with the highest percentage of tumour tissue, as we considered them to be the most informative about the cancer-specific expression profile. The amounts of tumour, benign and stromal tissue for each sample are presented in Supplementary Table A.1.

For the MSKCC dataset, there are exactly two technical replicates from the same RNA samples for each patient. Thus, in this case one of the replicates was selected at random.

As the ICR and Cambridge and datasets are relatively small datasets, we merged the clinical data for the ICR and Cambridge samples into a larger dataset, denoted CancerMap, to increase the statistical power of the correlations between the jumps and clinical outcome. CancerMap contains 254 unique samples. The summary of the clinical data corresponding to the two datasets, MSKCC and CancerMap, is presented in Table 4.2.

For a large proportion of samples from the ICR and Cambridge datasets we have the confirmation of the *ERG* fusion status, through a FISH "break-apart" analysis (a wet laboratory technique that can identify if the 5' and 3' ends of a gene are separated). We can hence use this information to evaluate the performance of the methods which aim to identify fusion candidates using exon microarrays.

4.3.3 Probeset selection

The exon microarray probesets have been designed to measure the expression of every known or putative exon in the human genome. They have been divided into five confidence categories, i.e. *core*, *extended*, *full*, *free*, and *ambiguous* (Table 2.3), based on the quality of transcriptional evidence available at the date of microarray design (mid 2000s).

A common practice when working with exon microarrays is to restrict the analysis only to the core probesets, as they are believed to be the highest confidence probesets. However, since 2003 (when the exon microarrays were designed), several genome assemblies have been released, which have gradually improved the gene annotations. Many probesets that were included in the extended or full categories because they corresponded to genes or isoforms for which at the time there was not enough reliable biological evidence or which were predicted only by bioinformatical approaches, have been since confirmed.

This means that the probesets in the extended and full category might be useful to gather as much information as possible about the updated gene annotations. Therefore we decided to keep in the analysis all the probesets from core, extended and full categories.

Table 4.2 Clinical summaries for the MSKCC and CanerMap datasets.

	MSKCC		**CancerMap**	
	count	**%**	**count**	**%**
Pathological Gleason score				
6	41	31.30%	35	25.60%
3 + 4	54	41.20%	75	54.80%
4 + 3	22	16.80%	17	12.40%
8	8	6.10%	3	2.20%
9	7	5.30%	7	5.10%
NA	1	0.70%	0	0%
Pathological Stage				
T1C	0	0%	1	0.70%
T2A	9	6.90%	5	3.60%
T2B	47	35.90%	1	0.70%
T2C	29	22.10%	45	32.90%
T2X	0	0%	22	16.00%
T3A	28	21.40%	47	34.30%
T3B	10	7.60%	14	12.20%
T3C	2	1.50%	0	0%
T4	6	4.60%	2	1.50%
PSA				
<4	22	16.80%	8	5.90%
4<PSA<10	78	59.60%	89	64.90%
10<PSA<20	20	15.20%	33	24.00%
>20	10	7.70%	4	2.90%
NA	1	0.70%	3	2.20%
Age at diagnosis				
Median	57.99		61	
Mean	58.03		60.11	
IQR	53.53-62.11		56-65	
Range	37.3-83		21-74	
Follow-up (months)				
Median	46.49		56	
Mean	48.19		52.27	
IQR	27.73-61.44		39-64	
Range	1.38-149.2		1-129	
BCR				
Failures	27	20.60%	35	25.60%

4.3.4 Pre-processing

For each dataset the analysis began from the raw signal intensity data, represented in *CEL* file format [243]. Briefly, after the labelled RNA tissue is hybridised on the microarray, a laser excites the fluorescent dye and a scanner measures the luminosity of each spot on the microarray, resulting in an image file [244]. For a microarray probe there are several corresponding pixels on the resulting image. The probe intensity is then estimated by aggregating the luminosities of the corresponding pixels. A CEL files stores for each probe the estimated probe intensity, its standard deviation, the number of pixels used to estimate the intensity and few other useful features.

The raw signal intensities need to be normalised in order to mitigate the technical effects introduced during the microarray processing. For normalisation we used the RMA algorithm (Section 2.6.2.1), implemented in the Affymetrix Power Tools (APT) software package [245]. APT is a cross-platform open-source command line program developed by Affymetrix for the analysis of Affymetrix GeneChip® arrays, including exon microarrays.

Besides the normalised probeset estimates, the RMA analysis implemented in APT also produces for each probeset in each sample a DABG (detected above the background) estimate, which indicates if the probeset is measuring any expression in the given samples. The DABG estimate is essentially the p-value of a t-test which assesses the hypothesis that the probeset intensity comes from the same distribution as the intensities of a set of anti-genomic probesets. Anti-genomic probesets are probesets which do not align to any human genome sequence and therefore any observed intensity is expected to be generated by background noise.

A common practice when working with microarrays is to exclude the probesets that are not detected above the background in a predefined proportion of samples. We did not follow this practice as sometimes the few samples in which the probe is detected above the background might carry relevant information. However, we paid attention to this aspect when we evaluated the fusion candidates, as we will discuss in the second part of this chapter.

4.3.5 Quality assessment

APT also reports several quality assessment metrics useful in flagging outlier microarrays. In this analysis we used the *positive controls vs. negative controls area under the curve (AUC)* and the *mean of the absolute deviation (MAD) of the residuals from* *the median* metrics, described in Section 2.6.3. For brevity we will refer to these two metrics as *positive vs. negative AUC* and *MAD of the residuals* respectively.

Outlier values for these statistics indicate that the quality of the data on chips might by influenced by non-specific effects, introduced during RNA processing or microarray handling. An important issue with this approach is that there is no objective way of deciding whether an array is an outlier or not. In our case, for each of the three datasets we plotted the values of the two metrics (y-axis) across microarrays (x-axis) (Figure 4.2). We visually inspected these plots and flagged as outlier the microarrays that exhibit spikes in the values of at least one of the above two metrics.

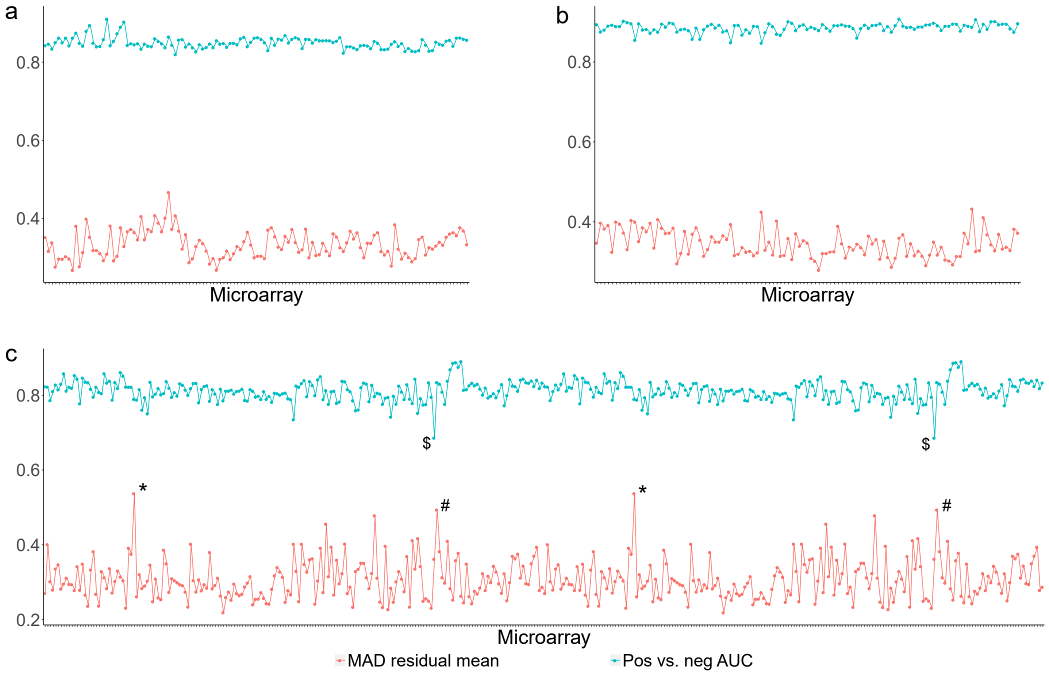

Figure 4.2 QA plots for the a) ICR, b) Cambridge and c) MSKCC datasets. The blue points correspond to the positive vs. negative AUC values for a given microarray, while the red points correspond to the MAD of the residuals values. The points annotated with the *, $^\#$ and $^\$$ symbols correspond to outlier values of the metrics.

For the ICR and Cambridge datasets (Figure 4.2a and Figure 4.2b respectively) it seems that there is no major spike, suggesting an acceptable quality for all microarrays.

For the MSKCC dataset we detected at least 6 microarrays that seem to spike, which are annotated on the plot with the *, $^\#$ and $^\$$ symbols. We note that the 6 outlier microarrays correspond in fact to three pairs of replicates, taken from three different patients. The points annotated with the * symbol correspond to the pair of replicates coming from a primary tumour sample, while the points annotated with $^\#$ and $ symbols correspond to two pairs of replicates taken from the metastatic samples of two patients.

We note that in all three pairs of microarrays, the metric values are consistent between the replicates, suggesting that any observed effects are unlikely to be generated by chip-related issues, but rather because of possible RNA quality issues or biological differences.

The MAD of the residuals values annotated with * and $^\#$ are larger than other microarrays, suggesting that a large number of probes are behaving differently than predicted by the RMA model. However, the positive vs. negative AUC seem to be normal. For the pair of microarrays annotated with $^\$$, on the other hand, the positive vs. negative AUC values seem slightly lower than the rest, while the MAD of the residuals are normal, suggesting that the probes behave as expected. Both the MAD of the residuals and positive vs. negatives AUC metrics are sensitive to the RNA quality [187], but only one or the other metric shows abnormal values. This observation suggests that the RNA quality is also unlikely to be the cause of differences.

We also note that two out of three pairs of microarrays correspond to metastatic tissue which can have a significantly different expression profile compared to the primary tissue [246, 247]. Thus, if many genes are differently expressed, the MAD of the residuals value can be significantly influenced, leading to the observed effects. Also the positive vs. negative AUC metric is sensitive to tissue type [187]. This suggests that some large scale expression differences might influence the above two metrics.

We, therefore, decided to keep all 6 samples in the analysis, as we do not have enough evidence supporting the hypothesis that the microarrays are technically biased. It seems rather that the differences are caused by some sort of biological difference, which might be useful for better understanding the behaviour of cancer.

4.3.6 Annotation

The human genome project [248] was finalised in mid 2000s and provided the complete DNA sequence of the human. However, having the complete DNA sequence is not enough to determine which areas of the genome are transcribed and to predict with certainty the structure of genes. Projects such as Ensembl [249], RefSeq [250], and UCSC [251] each created their own version of human genome annotation databases, which store information about genes, transcripts and their relationships. These databases are constantly updated, as the amount of biological evidence increases and as the prediction methods improve.

The Ensembl, RefSeq and UCSC databases have a large number of genes in common (approx. 22,000 [252]), but also contain a large number of unique genes. For example, Ensembl has around 33,000 genes not found in UCSC, nor in RefSeq. RefSeq has around 950 genes not found in the other databases and UCSC has around 5500 [252].

In order to maximise the number of genes analysed, we selected the Ensembl annotations, as it is the database with the largest number of genes. More specifically, we used the Ensembl release 76, based on the human genome assembly GRCh38 (GENECODE 20), released in August 2014. In this Ensemble there are available 58,640 gene ids.

For a each gene we aimed to select only the probesets that correspond to regions likely to be transcribed, i.e. exons. The probesets corresponding to the intronic regions carry no information about expression and, therefore, would have just added to the noise. For a given gene we aligned all the gene transcripts, as illustrated in Figure 4.3 and determined all the maximal continuous regions covered by the exons of at least one transcript. We also mapped the probesets to their corresponding genomic positions using the R package *annmap* [253]. We next excluded all probesets not lying entirely within the continuous regions determined by exons, resulting in a final set of probesets that was used for the downstream analysis.

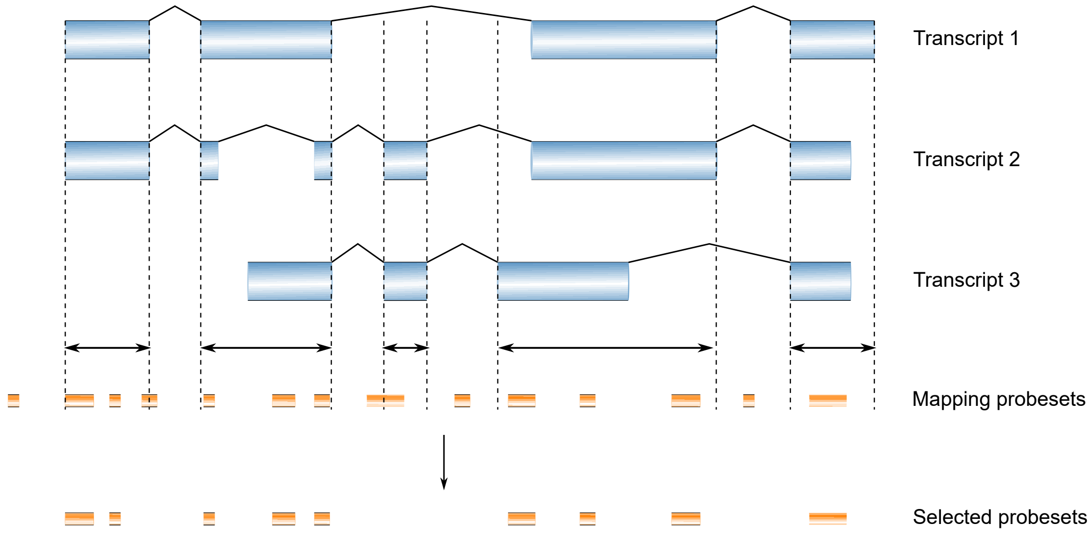

Figure 4.3 The selection of probesets for a given gene.

4.4 Methods

4.4.1 The jump detection method

4.4.1.1 Normalisation

The signal intensities of probesets within a gene can be quite variable even in normal genes, due to alternative splicing, different probeset affinities or other effects, non-related to fusions or other RNA abnormalities. The resulting variances can make the identification of jumps in expression described in Section 4.2 quite difficult, as illustrated in Figure 4.4a,c.

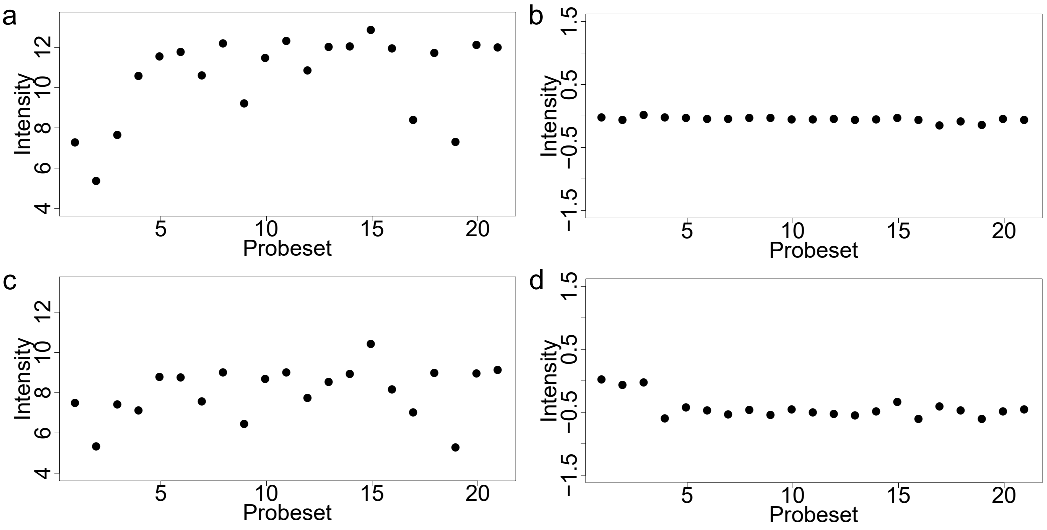

Figure 4.4 The effects of normalisation relative to a reference sample on the *TMPRSS2* gene: a) a normal *TMPRSS2* gene before normalisation; b) the same normal *TMPRSS2* gene after normalisation; c) a fused *TMPRSS2* gene before normalisation; d) the same fused *TMPRSS2* gene after normalisation;

In order to mitigate the influence of these effects, all methods for identification fusion candidates using exon microarrays, presented in Section 4.2.1, normalise the signal intensities of probesets to a consistent scale.

We identified in the literature two main approaches commonly used to normalise the probesets. The first method transforms the signal intensities of a given probeset to a standard normal distribution across samples [236], while the second method normalises the signal intensity of a given probeset relative to a reference intensity [235].

Although the first method, based on scaling the intensities of probesets to mean 0 and standard deviation 1 across samples, although worked well in initial tests, it was impacted by how many samples had an altered within-gene expression. If, for example, there are many samples for which a recurrent fusion leads to the under expression of the exons towards the 3' of a given gene, the scaling of probesets across all samples makes some samples falsely exhibit a step-up jump, when in fact they have a normal expression. This led us to use the other method, based on normalisation to a reference, which we will describe next, as it is independent on the number of samples with altered expression.

All our three datasets contain several normal tissue samples, which are less likely to contain chromosomal rearrangements or other abnormalities. Therefore, we supposed that for these samples the differences in probeset intensities for a given gene are generated by alternative splicing, different probesets affinities, noise or other possible effects, that occur in normal conditions. Also, we also made the assumption that these effects are approximately constant between samples. This means that we assume that different isoforms of a gene are expressed in the same proportions across samples, and also that the probeset affinities are constant. However we acknowledge that these assumptions might not always hold.

For a given probeset in a given dataset, we calculated a reference intensity, using only the normal samples. We then used the reference intensity to normalise the intensities of the probeset in all samples in the dataset. More specifically, given a dataset with N normal samples and P probesets, and a probeset p with the signal intensities $x_{pn}, 1 \leq n \leq N$, we estimated the reference intensity of the probeset p as:

$$r_p = \underset{1 \leq n \leq N}{median}(x_{pn}). \tag{4.6}$$

We then normalised the intensity of each probeset in each sample in a given dataset by computing the log ratio, relative to the corresponding reference intensity. Given the probeset p with the intensity x_p, we calculated its normalised intensity, y_p, as:

$$y_p = log_2\left(\frac{x_p}{r_p}\right). \tag{4.7}$$

We note that, for a probeset, the resulting normalised intensities take values close to 0 when the probeset intensity is similar to the reference intensity, it takes positive values if the probeset intensity is larger than the reference intensity and negative if it is smaller.

In Figure 4.4 we illustrate the effect of the normalisation on the *TMPRSS2* gene in a prostate cancer which does not harbour the *TMPRSS2-ERG* fusion and respectively on a prostate cancer with the *TMPRSS2-ERG* fusion. In Figure 4.4a we depict the probeset intensities of the non-fused *TMPRSS2*, and in Figure 4.4c we depict the probeset intensities in the fused *TMPRSS2* genes. We note that in both cases the probeset intensities are highly variable and that it is quite difficult to distinguish any jumps in the fused gene. The jump becomes more clear after normalisation. The normalised probeset intensities of the non-fused gene are all very closely distributed around 0 (Figure 4.4b), while for the fused gene we notice a step-down jump starting with the fourth probeset (Figure 4.4d).

4.4.1.2 Candidate identification

As shown earlier there are two main approaches when it comes to identifying jumps using exon microarrays, namely the method of Jhavar et al. [235], based on a walking Student's *t*-test which determines the position where the *t*-statistic comparing the distribution of probesets before and after takes the maximum value, and the method of Wang et al. [239] based on fitting a non-linear regression model to two scores.

In our preliminary attempts we implemented both methods and tried to estimate their performance on our datasets. As we will describe later, the *t*-test based method performs reasonable well, while the Wang et al. [239] method does not seem to generalize very well on validation data. We developed a novel method based on step functions, which will be presented next and which performs at least as well as these two methods.

As discussed in Section 4.2, gene fusions can result in jumps in the expression of the exons located after the breakpoint. We tried to verify this assumption on genes most commonly fused in prostate cancer, such as *TMPRSS2*, *ERG*, *ETV1* and even the less common *ETV4*. The *ERG* gene, with validated fusion status in the ICR and Cambridge datasets was particularly useful for this purpose.

In all these genes we identified jumps in the expression, similar to those illustrated in Figure 4.5b,d,f. *TMPRSS2* exhibits step-down jumps, which is what we would expect to observe as the exons of *TMPRSS2* located after the breakpoint are either deleted or translocated to a less transcribed locus. On the other hand, *ERG* shows step-up jumps, probably generated by the translocation of the exons after the breakpoint to a much more transcribed locus. We also observed the expected step-up patterns in the *ETV4* gene.

To programatically identify such jumps, we developed an approach based on fitting a step function to the points determined by the probesets intensities from each gene. A *step function* is a function $f : [a,b) \to \mathbb{R}$ for which there exists a sequence $a = a_0 < a_1 < a_2 < ... < a_n = b$ such that the function f is constant for each interval $[a_i, a_{i+1})$. Given a sent of points $P = \{p \mid p \in \mathbb{R}^2\}$, a step function can be fitted to P by minimizing some error measure, e.g. the maximum vertical distance, $d(P,f) = max\{d(p,f) | p \in P\}$ [254],

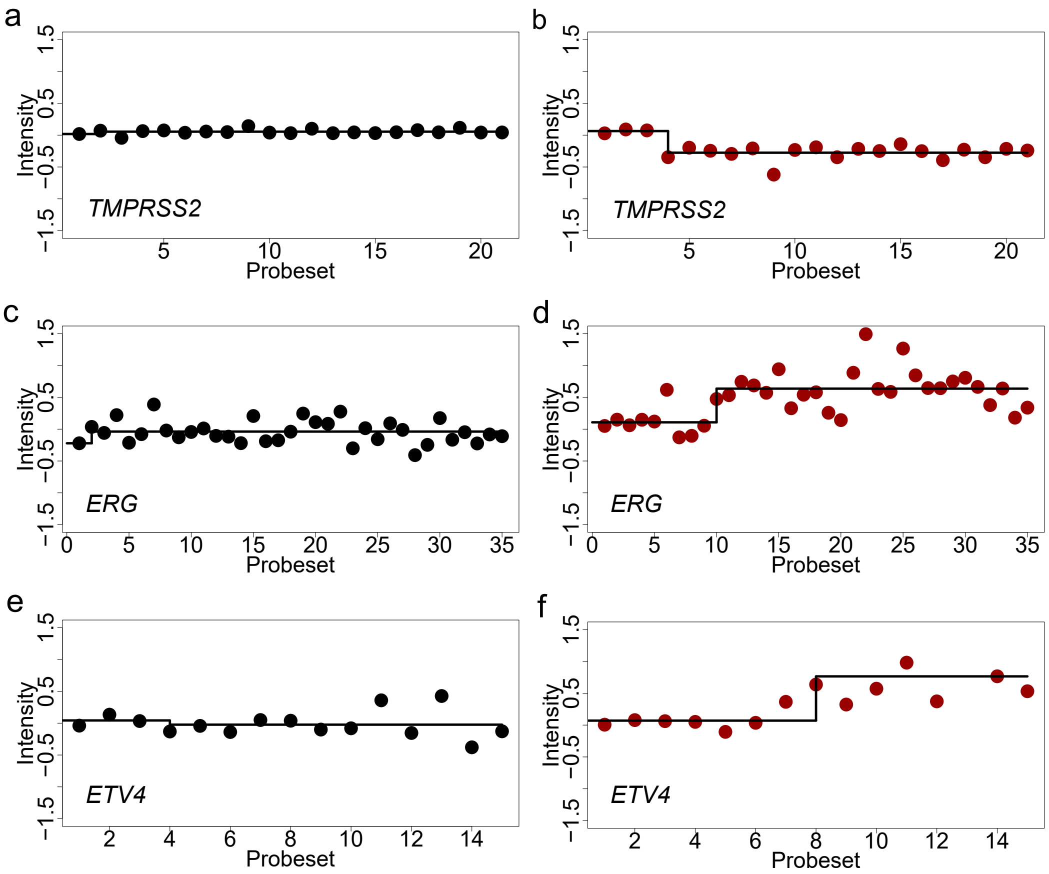

Figure 4.5 Example of samples without jumps (a, c, e) and with jumps (b, d, f) in genes commonly involved in fusions in prostate cancer, namely *TMPRSS2*, *ERG* and *ETV4*.

or the sum of squared distances $\sum_{p \in P} d(p,f)^2$ [255], where $d((x,y),f) = |f(x) - y|$, or some other error function.

In our case we need a step function with two intervals, the first one corresponding to the probesets before the putative breakpoint and second one corresponding to the probesets after the putative breakpoint. We consider the value of the step function of each interval as corresponding to the average intensity of the probesets in the interval. More specifically, given a gene for with P probesets, ordered by their 5' to 3' genomic position, their corresponding normalised intensities $y_p, 1 \leq p \leq P$ and the putative breakpoint occurring after probeset $k, 1 \leq k < P$, we define the step function $step : [1,P] \to \mathbb{R}$, as:

$$step(x) = \begin{cases} \mu_l, \; x \in [1,k] \\ \mu_r, \; x \in (k,N], \end{cases}$$

where $\mu_l = \frac{\sum_{p=1}^{k} y_p}{k}$ and $\mu_r = \frac{\sum_{p=k+1}^{N} y_p}{N-k+1}$.

We tried to determine if the two levels of the step function which best fits a given gene in a given sample reflects a jump in the probeset intensities. We designed a measure (score) aimed to reflect how "well-defined" the step function is and then to consider the gene as a fusion candidate if the score was above a predefined threshold.

A "well-defined" step function would be a function for which the step levels are as far apart as possible and for which the probesets intensities have a minimum variance around the corresponding levels. However, the probesets of different genes seems to exhibit different amounts of noise (Figure 4.5). For some genes, such as *TMPRSS2*, the intensities of probesets seem to exhibit little variation around a certain level, while for others, such as *ERG*, the variations are larger. Some other genes, such as *ETV4* seem to have an intermediary amount.

We therefore derived a score that takes values close to 0 when there are no jumps in the intensities and which increases as the distance between step levels increases or the variance of the probeset intensities around the step levels decreases. As before, we consider a gene with P probesets, ordered by their 5' to 3' genomic position, their corresponding normalised intensities $y_p, 1 \leq p \leq P$ and the putative breakpoint that occurs after probeset $k, 1 \leq k < P$. A score calculated as:

$$l\_score(k) = log\left(\frac{\frac{\sum_{p=1}^{k}(y_p-\mu_r)^2}{k}}{\frac{\sum_{p=1}^{k}(y_p-\mu_l)^2}{k}}\right), \tag{4.8}$$

corresponds to the log ratio of two quantities. The numerator is the average distance of the probesets to the left of the breakpoint relative to the level, μ_r, of probesets to the right of the breakpoint. It increases as the magnitude of jump increases. The denominator corresponds to the average distance of the probesets to the left of the breakpoint relative to their mean, i.e. the variance. The log ratio increases as the variance decreases. The resulted log-ratio takes values between 0 and ∞, as the numerator is always greater or equal than the denominator, and increases as the distance between the levels of the step functions increases or as the variance decreases, which is exactly what we need.

Analogously, we can compute a score for the probesets to the right of breakpoint, defined as:

$$r\_score(k) = log\left(\frac{\frac{\sum_{p=k+1}^{P}(y_p-\mu_l)^2}{P-k+1}}{\frac{\sum_{p=k+1}^{P}(y_p-\mu_r)^2}{P-k+1}}\right), \tag{4.9}$$

The score at a given position k is the defined as:

$$score(k) = min\begin{cases} l\_score(k) \\ r\_score(k) \end{cases}. \tag{4.10}$$

We then determine the position of the putative breakpoint as the probeset for which we obtain the maximum score, that is:

$$breakpoint = \underset{1\leq k<P}{argmax}(score(k)) \tag{4.11}$$

and the score of the gene in a given sample is defined as:

$$step\_score = score(breakpoint). \tag{4.12}$$

We will further refer to this score as the *step score*. As we illustrate in Section 4.5.1, we determined a threshold for the step score above which we consider the step function as reflecting a jump in the expression of the exon of a given gene in a given sample. Also, for simplicity, we will refer the method that detects jumps based on the step sore as the *step method*.

For a big proportion of the 58,640 genes annotated initially there are less than five probesets mapping to their exonic regions. For these genes it is difficult to distinguish the jumps, as there is not enough data to reliably determine if the intensities of probesets located on either side of the putative breakpoint have a step-like shape. Hence, for our analysis we considered only the 19,202 genes for which we could map at least five probesets.

4.4.1.3 Additional criteria for reducing false positives

The exon microarray data is quite noisy and there are several recurrent cases when the step method identifies jumps that are either generated by noise or by chance. Also sometimes the jumps identified are not consistent with a transcriptional alteration. We illustrate several such examples in Figure 4.6.

In Figure 4.6a, there is an example of a step-up jump identified in *C1GALT1*. As the gene has relatively few probesets, the small jump seems to be generated by chance. In Figure 4.6b, the data coming from the *BRWD3* gene is very noisy and it seems that the jump is the result of noise in the data.

In Figure 4.6c the intensities of all probesets (corresponding to *BPIFB1*), except the first probeset, are distributed around 0, suggesting normal expression. In the case of a transcriptional alteration we would expect the probes after the breakpoint to be over or underexpressed. Therefore, this kind of jump is inconsistent with a transcriptional alteration.

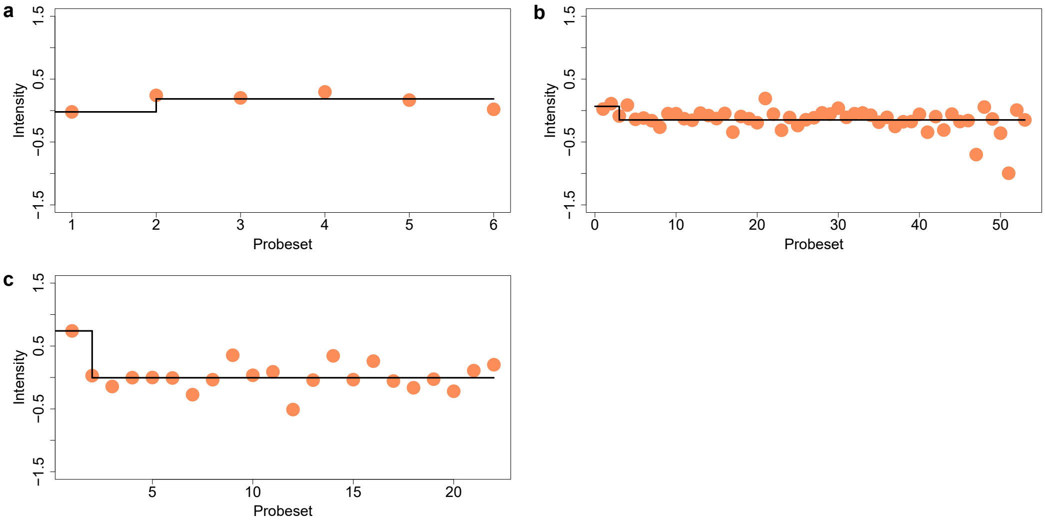

Figure 4.6 Examples when the step method identifies jumps that are either generated by noise or inconsistent with a transcriptional alteration in: a) *C1GALT1*, b) *BRWD3* and c) *BPIFB1*.

As the step score is not enough for detecting such cases, we considered three additional criteria:

1. a non-parametric statistical test that assesses if the probesets before and after the breakpoint come from the same distribution;

2. a non-parametric statistical test that verifies if the mean of probesets after the breakpoint is significantly different than 0;
3. the magnitude of the jump, i.e. the distance between the two levels of the step function.

The first two criteria are based on the detection method of Jhavar et al. [235]. The Jhavar et al. [235] method used a *t*-test that assesses if the probesets before and after the breakpoint have different distributions and another *t*-test that verifies if the mean of probesets after the breakpoint is significantly different than 0. However, the *t*-test makes the assumptions that the points come from a normal distribution. Our initial normality tests on a set of randomly selected genes indicated that this is usually not the case. Therefore we chose to perform two non-parametric tests instead, which test the same hypotheses, but do not make the normality assumption.

For assessing the first criterion we performed a two sample Mann-Whitney U independence test [256], which tests the hypothesis that two populations come from the same distribution, without making the normality assumption. For the second criterion we used the sign median test, which assesses the hypothesis that the median of a general distribution equals a specified value.

Furthermore, for a given gene we adjusted the *p*-values of the both statistical tests for multiple comparisons across samples using the false discovery rate (FDR) correction [257] at a 5% level. The criteria were considered met if the adjusted *p*-values were below 0.05.

The third criteria concerns the distance between the two levels of the step function described in Section 4.4.1.2. We imposed this criterion met if the distance was above a specific threshold.

The first criterion was designed with the purpose of filtering out the candidate genes for which the jumps are most likely generated by chance due to small number of probesets and noise in the data, such as the cases illustrated in Figure 4.6a,b. The second criterion was created for cases such as the one presented in Figure 4.6c. Because the first criterion is quite stringent, and sometimes genes with few exons have large jumps we introduced the third criterion, that can help detecting such cases. A jump detected by the step method is considered a candidate for transcriptional alterations if it meets at least two of the three criteria.

4.4.1.4 The final method

We integrated the three criteria presented in Section 4.4.1.3 with the step method described in Section 4.4.1.2 to obtain a final step method, which was used for screening the data.

We classified a gene in a given sample as a candidate for transcriptional abnormalities if the step score was above a determined threshold and, additionally, at least two of the three criteria were met. We will refer to this method as the *final step method*.

4.4.2 Genomic plots

To help visualise the mapping between jumps and their position along the gene transcripts, we built plots such as the one presented in Figure 4.7. In the top panel we illustrate a representative jump for the candidate gene (*AZGP1* in this case), where the probesets are numbered increasingly starting from 5' end of the genes. In the bottom panel we represent a gene model, created by aligning all the gene transcripts annotated in Ensembl, using a version of the GenomeGraph R package [258].

In the middle panel, the vertical lines correspond to the positions where the probesets align to the gene model. We note that for a given exon there might be none, one or several probesets aligning to it. Each read line links the intensities of two consecutive probesets in a sample with jumps in a given dataset (MSKCC in this case). The black horizontal line links the average intensities of each probeset across samples. For simplicity, we will refer to plots such as this one as *genomic plots*.

4.4.3 Survival analyses

On the candidate genes identified by the method described in the previous section we performed a survival analysis, aiming to identify candidates for which the samples with jumps have a faster (or slower) time to biochemical recurrence compared with samples without jumps.

For the survival analysis we considered only primary tumour samples coming from unique patients, chosen as described in Section 4.3.2. For each candidate gene in each dataset we performed a log-rank test (Section 3.5.2), testing if the patients which exhibit step-up have a significantly different time to BCR relative to the patients who do not show jumps. In each dataset we adjusted the log-rank p-values for multiple comparisons using the FDR method, at a 5% level.

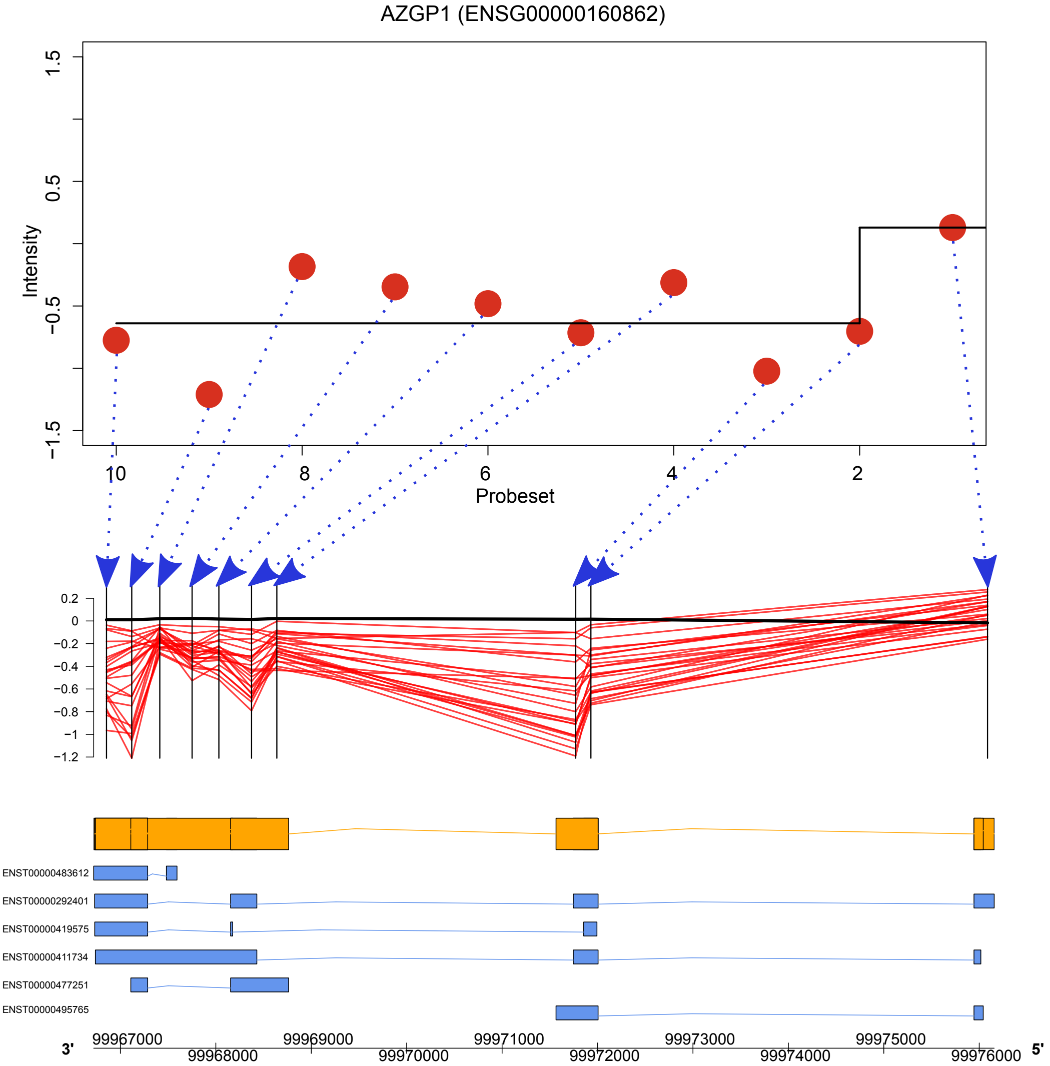

Figure 4.7 Genomic plot depicting the mapping of the jumps to the *FKBP5* gene model. In the top panel we depict a representative step-down jump. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model. Each read line links the intensities of two consecutive probesets in a sample with step-down jumps. In the bottom panel it is represented the gene model.

4.4.4 Correlation of jumps with metastasis

We also tested whether the genes which exhibit jumps are significantly over-represented (or under-represented) in the metastatic samples relative to primary prostate samples.

Since the CancerMap lacks metastatic samples, we restricted this analysis only to the MSKCC dataset. For each candidate gene we performed a Pearson's χ^2 independence test, testing the hypothesis that the jumps are proportionally distributed in the metastatic samples and the primary tissues (the normals and primary tumours), with the alternative hypothesis that in the metastatic samples the jumps of a given gene are under or over-represented.

We then adjusted the χ^2 p-values for multiple comparisons using the Benjamini-Hochberg (FDR) method at a 5% level. We considered that a candidate gene is associated with the metastatic samples if its corresponding FDR adjusted χ^2 p-value was less than 0.05.

4.4.5 Pathway analysis

For the lists of candidates with jumps significantly associated with the clinical outcome we performed pathway analysis (Section 3.6), with the purpose of identifying biological pathways for which the component genes are over/under-represented in the list.

For each set of candidates we performed an independent analysis using all pathways annotated in Gene Ontology (GO) [225] (from which we used the biological processes ontology), Kyoto Encyclopedia of Genes and Genomes (KEGG) [227] and Reactome [259]. The analyses have been performed using the *clusterProfiler* R package [260]. We adjusted the resulting p-values for multiple comparisons using the FDR method at a 5% level. We considered that a pathway is over/under-represented in a set of genes if its corresponding FDR adjusted p-value was less than 0.05.

4.4.6 Known fusion candidates

We further focused our analysis on the candidate genes significantly associated with time to BCR (Section 4.4.3) and metastasis (Section 4.4.4) which have been previously associated with fusions in prostate cancer, but which have not necessarily been associated with clinical outcome.

We obtained a set of prostate cancer-specific fusions, that have been experimentally validated, from several studies ([32, 54, 136, 261–263]). The fusions are the result of wide range of mechanisms. Some of them derive from well-established chromosomal aberrations such as translocations or deletions, but others, like for example several fusions obtained from Pflueger et al. [263], are the result of read-through transcription events. As described previously, the read-through transcription is a process which results in chimeric transcripts containing sequences from two adjacent genes.

Also, some of the fusions we retrieved from Baca et al. [54] are the effect of the relatively newly discovered complex translocation mechanism, called chromoplexy, which involves more than two genes which simultaneously cleave and rejoin the wrong ends, resulting in a chain of gene fusions. Several chromoplexy-caused fusions reported by Baca et al. [54], such as *ERG-PADI6*, *YIPF1-TMPRSS2* and *ARHGEF3-TMPRSS2*, are quite unusual, as the *ERG* gene which is usually a 3' partner, appears to be the driver genes, or *TMPRSS2*, the usual 5' participant, is fused as the 3' partner.

We separated the 5' fusion partners, from the 3' ones. For the 5' fusion participants we correlated the step-down jumps with the clinical data, while for the 3' participants we correlated the step-up jumps. Genes that were reported as 5', but also 3' partners, have been included in both lists of partners, and therefore have been screened for step-up but, also for step-down jumps. The two lists are available in the Supplementary Tables A.2 and A.3 and contain 55 and respectively 45 genes.

4.5 Step method tuning

4.5.1 Step score threshold

For the step score (Section 4.4.1.2) we need to set a threshold above which to consider that the step method identifies a jump. For this purpose, we evaluated the step scores of the *ERG* gene in samples with fusion status confirmed by FISH, from the ICR dataset. Based on this we selected a threshold which minimises the classification error. We then validated the threshold on the step scores of the *ERG* genes in the Cambridge dataset.

More specifically, we trained a logistic regression model (Section 3.2.3) on the ICR dataset, for which we consider the FISH fusion status as the target variable and the step score as the only predictor variable. The logistic model resulting after estimating the parameters is:

$$log\left(\frac{p(FISH = \text{``}fused'')}{p(FISH = \text{``}non-fused'')}\right) = 1.288 \cdot step\_score - 2.992. \tag{4.13}$$

The coefficient corresponding to step score, the only predictor variable, is positive (1.288), which indicates a positive association between the step score and the odds of the sample being fused. A positive log odds ratio for a given step score indicates that the score is more likely to correspond to a sample with confirmed FISH fusion, while a negative score indicates the opposite. Therefore, we can find a threshold above which the step score would yield positive log odds ratios. This can be easily determined by solving the inequality:

$$1.288 \cdot step\_score - 2.992 \geq 0 \implies step\_score \geq 2.32, \tag{4.14}$$

which yields the threshold 2.32.

4.5.2 Performance of the step score

In this section we evaluate the step method and compare it to the other two methods, i.e. *t*-test based method and the method of Wang et al. [239]. For this evaluation we considered the step method presented in Section 4.4.1.2, before applying the additional criteria described in Section 4.4.1.3.

The *t*-test method (Section 4.2.1) classifies a gene as a candidate if the *t*-statistic is above a predefined threshold. In order to estimate the *t*-statistic threshold, we took the same approach we took in deriving the threshold for the step score in our method, described in Section 4.5.1. As previously, we fitted to the the ICR dataset a logistic model with the *t*-statistic as the only predictor. The logistic model indicated the threshold 6.13 for the *t*-statistic.

The Wang et al. [239] method, classifies genes based on a non-linear logistic regression of two scores (Section 4.2.1). The regression coefficients were also estimated using the FISH fusion status of *ERG*, as the other two methods.

For each of these three models, we estimated the performance on the training dataset (ICR) and the validation dataset (Cambridge) (Figure 4.8).

The step method obtained a classification accuracy of 83.33% on the training dataset and an AUC for the ROC curve of 0.83, which outperforms the *t*-test method (accuracy 76.85% and AUC 0.76), and also the method of Wang et al. [239] (accuracy 70.37% and AUC 0.7).

On the test dataset the step method obtained the exactly the same classification accuracy and AUC as the *t*-test method (82.35% and respectively 0.86), and clearly outperformed the Wang et al. [239] method, which seems to perform quite poorly on the test dataset (accuracy 57.84% and AUC 0.64).

These results suggest that the step method and the *t*-test perform reasonable well on the training data and also seem to generalise well on new datasets. One limitation in this comparison is that it has only been performed on the *ERG* gene. Due to lack of additional validation data, we can not objectively assess how well the step method

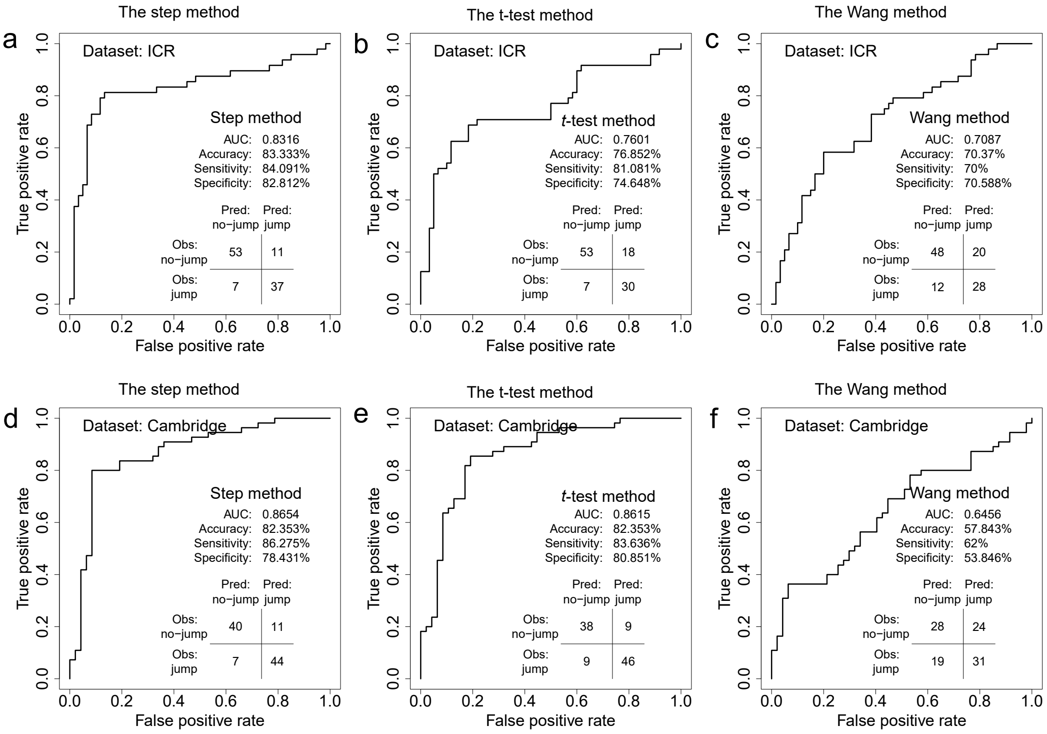

Figure 4.8 Classification performance of the three methods on the ICR dataset (a-c) and Cambridge dataset (d-f): a) the performance of the step method on the ICR dataset; b) the performance of the *t*-test method on the ICR dataset; c) the performance of the Wang et al. [239] method on the ICR dataset; d) the performance of the step method on the Cambridge dataset; e) the performance of the *t*-test method on the Cambridge dataset; f) the performance of the Wang et al. [239] method on the Cambridge dataset;

performs on other genes. However, jumps are identified in all the common fusion candidates in prostate cancer, such as *TMPRSS*, *ETV1*, and even *ETV4* in all datasets, giving indication that the method works on other genes as well.

We further investigated the reasons why our method misclassified some samples. We visually inspected each misclassified sample trying to determine if any noticeable jumps are exhibited. Of course, this procedure is subjective and prone to biases. However, we were just aiming to obtain some rough estimation of what causes the misclassification, to better understand the behaviour of our method and to assess how the jumps are correlated with the fusion status.

In the ICR dataset, out of 11 false negatives, we estimate that in around 9 samples the probeset intensities do not show any jump, as it can be seen in Supplementary Figure A.1. Furthermore, in the Cambridge dataset none of the 11 false negatives seems to exhibit any jump either (Supplementary Figure A.2). Regarding the false positives, out of 7 false positives in the ICR dataset, in at least 4 we discovered some noticeable jumps (Supplementary Figure A.3), while the rest have some visible jumps, although not as well-defined. Similarly in the Cambridge dataset, where out 7 false positives, at least 5 have jumps in the expression (Supplementary Figure A.4).

The main cause of misclassification seems to be the imperfect correlation between the fusion status identified with the FISH break-apart assays and the presence of the jumps. We estimate, based on the above data, that for the *ERG* gene around 20% of times the fusions do not result in jumps. We expect for driver genes, such as *TMPRSS2*, the percentage to be even larger, because of the expression of the other copy of the gene might reduce the magnitude of the jump.

4.5.3 Additional criteria threshold

The third criterion presented in Section 4.4.1.3, concerning the distance between the step levels, also needs a threshold above which we consider the criterion met. We determined this threshold using the same approach based on logistic regression, as the one described in Section 4.5.1. In more detail, we fitted a logistic regression model on the ICR dataset, usind a single predictor variable - the distance between the two levels of the step function, and using as outcome variable the FISH fusions status.

As before, based on the model coefficients (5.978 for the coefficient corresponding to jump magnitude and -1.415 for the intercept) we calculated the threshold value, which is 0.24.

4.5.4 Performance of the final step method

We evaluated the performance of the final step method (Section 4.4.1.4) on the *ERG* gene in samples with FISH confirmed fusion status. For this method we cannot calculate ROC curves, as the final method produces discrete outputs. However, we have been able to estimate the accuracy, which we compared with the initial step method, to make sure that after introducing the criteria the detection power of the method is not affected.

On the ICR dataset the accuracy increased from 83.33% to 84.25%, as one of the false positives was removed. On the Cambridge dataset, for which we estimated that all false positive occur exclusively because the fusion does not result in jumps, the accuracy slightly decreased from 82.35% to 81.37%, as another false negative was produced.

This method was further used for genome-wide screening of genes, in order to identify candidates for transcriptional abnormalities.

4.6 Candidate genes

We used the step method described in Section 4.4.1.4, to screen for jumps in all genes with more than five probesets, from each of the three datasets (ICR, Cambridge and MSKCC). As the ICR and Cambridge datasets are relatively small (aprox. 120 samples each) and the transcriptional abnormalities can be quite rare, we combined the screening results from the ICR and Cambridge into a larger dataset, referred to as CancerMap.

We restricted the analysis only to candidates with jumps in at least 1% of samples. In the CancerMap dataset we identified 4,839 genes that exhibit step-up jumps in at least 1% of samples. In the MSKCC we identified 5,690 step-up candidates. For the candidates with step-down jumps, the numbers are similar. For CancerMap we identified 4,332 candidates with step-down jumps and for MSKCC we found 4,889. Histograms depicting the frequency of jumps in these fusion candidates is presented in Supplementary Figure A.5.

4.6.1 Top candidates

In Supplementary Tables A.4 and A.5 we present the top 200 candidates in CancerMap and respectively MSKCC, sorted descending by the number of samples in which they exhibit step-up jumps. *ERG* gene is in the top 20 candidates in both datasets. In CancerMap it is 16th (Supplementary Table A.4) with jumps in 92 (39.3%) samples. In MSKCC it is the 10th (Supplementary Table A.5) candidate with jumps in 96 (25.9%) samples.

Other *ETS* family candidates are identified at lower frequencies. *ETV1* shows jumps in 16 (6.8%) samples in CancerMap and 28 (7.5%) samples in MSKCC. *ETV4* shows jumps in 4 (1.7%) samples in CancerMap and in 4 (1.08%) samples in MSKCC, while for *ETV5* we identify jumps in 6 (2.5%) samples in CancerMap and 12 (3.24%) in MSKCC. Moreover, we identify jumps in *FLI1*, which is a *ETS* family gene recently reported as being involved in fusions in prostate cancer [264]. We identify jumps in 18 (7.6%) samples in CancerMap and 12 (3.2%) samples in MSKCC.

In Supplementary Tables A.6 and A.7 we present the top 200 candidates in CancerMap and respectively MSKCC, sorted descending by the number of samples in which they exhibit step-down jumps. *TMPRSS2* is in the top 20 candidates in MSKCC, with 82 (22.1%) jumps. In CancerMap is is only the 160th, as it shows jumps in 37 (15.8%) samples.

4.6.2 Candidates in common

Most candidate genes are detected in both datasets. As it can be seen in Figure 4.9a, more than three quarters of the step-up candidates (3,802) are in common between the two datasets. Similarly, 3,216 step-down candidates are in common (Table 4.3, Figure 4.9b).

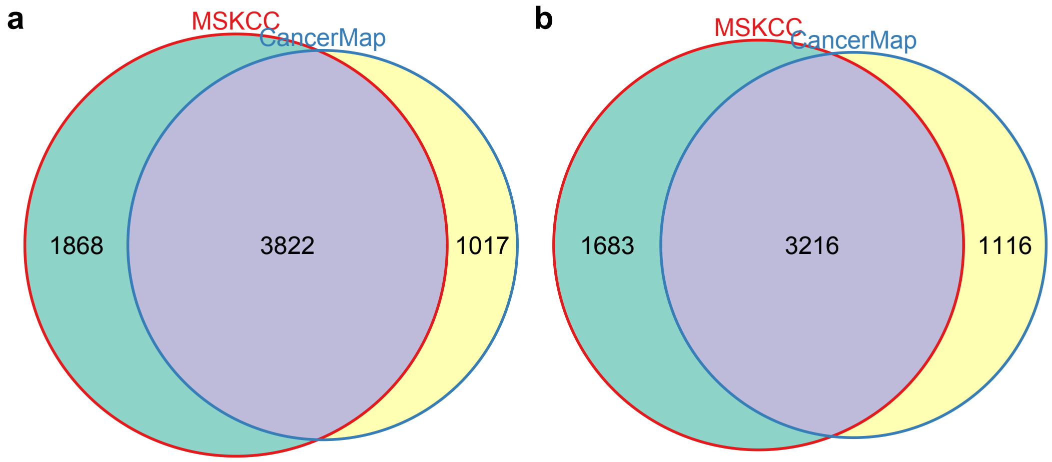

Figure 4.9 Number of genes in common between the CancerMap and MSKCC datasets which exhibit: a) step-up jump; b) step-down jumps.

Table 4.3 Top 10 candidates with the largest number of step-up jumps and respectively step-down jumps in common between CancerMap and MSKCC.

Step-up candidates				
	CancerMap		**MSKCC**	
Gene	**count**	**%**	**count**	**%**
TDRD1	135	57.69	120	32.43
C1QTNF3-AMACR	145	61.97	66	17.84
CRISP3	115	49.15	92	24.86
TMEM178A	98	41.88	96	25.95
ERG	92	39.32	96	25.95
F5	108	46.15	80	21.62
LUZP2	120	51.28	68	18.38
GCNT1	113	48.29	66	17.84
PLA2G7	64	27.35	114	30.81
SLC38A11	75	32.05	102	27.57
Step-down candidates				
	CancerMap		**MSKCC**	
Gene	**count**	**%**	**count**	**%**
OLFM4	112	47.86	140	37.84
KRT23	81	34.62	94	25.41
SYNM	74	31.62	100	27.03
CHRDL1	72	30.77	92	24.86
TP63	90	38.46	70	18.92
ANPEP	94	40.17	62	16.76
SELE	86	36.75	60	16.22
PTGS2	69	29.49	76	20.54
MME	65	27.78	76	20.54
MYBPC1	40	17.09	100	27.03

4.6.3 Survival analyses

We performed survival analyses, as described in Section 4.4.3, on the candidate genes identified by the step method, aiming to ascertain if the presence of the jumps can predict a faster (or slower) biochemical recurrence rate.

4.6.3.1 Step-up candidates

The starting points in this survival analysis were the 3,822 genes that exhibit step-up jumps in CancerMap and also in MSKCC (Section 4.6.2). For each gene we performed a log-rank test, assessing if the samples with step-up jumps have different BCR outcome, compared to those without. Then, we adjusted the *p*-values for multiple corrections using the FDR method, at 5% level.

For CancerMap we identified 213 step-up candidates (Supplementary Table A.8) with log-rank *p*-value below 0.05, before FDR adjustment. After adjustment, only 52 of these candidates remained significant. In MSKCC, we identified 138 step-up candidates (Supplementary Table A.9) with significant *p*-values, of which 17 were significant after FDR adjustment. None of the 52 significant candidates in CancerMap is in common with the 17 significant candidates in MSKCC.

4.6.3.2 Step-down candidates

We performed the same analysis as in the previous section, but for the 3,216 candidates with step-down jumps in common between CancerMap and MSKCC. For each gene in each dataset we performed a log-rank test, assessing if the samples with step-down jumps have different BCR outcome, compared to those without. Then, we adjusted the *p*-values for multiple corrections using the FDR method, at 5% level.

In this case, the log-rank test identified 356 candidates with log-rank *p*-values less than 0.05, before correction, in CancerMap (Supplementary Table A.10) and 594 in MSKCC (Supplementary Table A.11). After adjusting for multiple comparisons, we obtained 76 significant candidates in CancerMap and respectively 308 in MSKCC. Of these 9 genes were in common (Table 4.4). The KM plots corresponding to these genes are presented in Supplementary Figures A.6-A.14.

4.6.3.3 Known fusion partners

The step method identified step-up jumps in 17 out of the 45 known 3’ participants (Supplementary Table A.3; Section 4.4.6), in both CancerMap and MSKCC. The

Table 4.4 Fusion candidates with step-down jumps correlated with time to BCR in CancerMap and MSKCC.

Gene Symbol	Gene ID	CancerMap adj. *p*-value	MSKCC adj. *p*-value
AKAP7	*ENSG00000118507*	$3.65 \cdot 10^{-5}$	$4.27 \cdot 10^{-13}$
ALDH3A2	*ENSG00000072210*	$2.48 \cdot 10^{-3}$	$9.26 \cdot 10^{-4}$
ARMCX1	*ENSG00000126947*	$1.87 \cdot 10^{-8}$	$4.76 \cdot 10^{-5}$
ASPA	*ENSG00000108381*	$3.61 \cdot 10^{-2}$	$8.68 \cdot 10^{-4}$
DIXDC1	*ENSG00000150764*	$4.06 \cdot 10^{-3}$	$2.80 \cdot 10^{-3}$
HSDL2	*ENSG00000119471*	$3.05 \cdot 10^{-3}$	$2.34 \cdot 10^{-2}$
LRCH2	*ENSG00000130224*	$3.99 \cdot 10^{-2}$	$5.07 \cdot 10^{-4}$
PI15	*ENSG00000137558*	$3.05 \cdot 10^{-3}$	$2.01 \cdot 10^{-2}$
VAT1	*ENSG00000108828*	$3.48 \cdot 10^{-2}$	$3.717 \cdot 10^{-2}$

survival analysis however failed to identify any genes significantly associated with the time to BCR (Supplementary Table A.12).

The step method also identified 17 jumps in the 55 known 5' partners (Supplementary Table A.2; Section 4.4.6). None of the 17 candidates is correlated with the time to BCR in both datasets (Supplementary Table A.13). The survival analysis identified significant association in the MSKCC dataset for *YIPF1* (adjusted log-rank *p*-value $2.37 \cdot 10^{-4}$), but the results have not been reproduced in the CancerMap dataset. Conversely, the log-rank *p*-value corresponding to the *AZGP1* was significant in CancerMap (adjusted log-rank *p*-value $2.28 \cdot 10^{-2}$), but not in MSKCC.

4.6.4 Correlation with the metastatic samples

We also correlated the step-up and respectively step-down jumps with the metastasis, as described in Section 4.4.4. As presented there, this analysis is based solely on the MSKCC dataset, as CancerMap does not contain any metastatic samples.

4.6.4.1 The step-up candidates

For this analysis the starting point was the 5,690 genes with step-up jumps in MSKCC dataset (Section 4.6). 79 candidates were significantly associated with metastasis (FDR adjusted χ^2 *p*-value < 0.05, Supplementary Table A.14).

4.6.4.1.1 ***AR*** The top step-up candidate is the *AR* (androgen receptor) gene. We identified step-up jumps in this gene in 7/19 unique metastatic samples, and 1/160 unique primary tissues (FDR adjusted χ^2 *p*-value $1.05 \cdot 10^{-7}$) in *AR*. Most putative breakpoints for *AR* occur between the probesets 1 and 2 (Supplementary Figure A.15).

The first probeset maps to a region included in an alternative first exon for the transcripts ENST00000612452 and ENST00000396044, but which is spliced out in the other transcripts. The observed step-up jumps could therefore be explained by alternative splicing. If some of the transcripts that splice out the region where the first probeset maps are over-expressed in some samples, the resulting expression pattern would resemble the jumps we observe.

4.6.4.1.2 Pathway analysis On the list of 79 candidates we performed a pathway analysis (Section 4.4.5), to determine if genes involved in biological pathways with possible role in cancer are over-represented in the list of candidates. In the GO database, the pathway analysis identified 141 over-represented pathways. The top 10, with the lowest p-value are presented in Table 4.5. Also, 20 pathways annotated in the Reactome database are over-represented (of which top ten are presented also in the Table 4.5). In KEGG, only one pathway, containing genes involved in the cell cycle is over-represented (Table 4.5).

We note that most pathways identified in all three databases are related to cell cycle processes, which are known to be involved in prostate cancer progression. For example, the Prolaris [38] test predicts aggressive prostate cancer based on the expression of 31 cell cycle progression genes.

4.6.4.2 The step-down candidates

We also performed metastatic correlation on the 4,889 step-down candidates in MSKCC dataset. We found 548 genes that show step-down jumps associated with the metastatic sample in the MSKCC dataset. The top 200 genes, with the lowest χ^2 p-values are presented in Supplementary Table A.15. Further investigations are necessary to determine the source of these jumps.

4.6.4.2.1 Pathway analysis We performed a pathway analysis (Section 4.4.5) on the top 200 candidates. The analysis on the GO database identified over 300 significant pathways, of which the top 10 pathways, with the lowest p-values, are presented in Table 4.6. The analysis on the Reactome database identified three significant pathways, also presented in Table 4.6. The analysis did not identify any significant pathway in the KEGG database.

We note that both analyses identified the muscle contraction and extracellular matrix organization pathways. These pathways have also been identified as associated with

Table 4.5 The top pathways from the GO, Reactome and KEGG databases, over-represented in the set of 79 step-up candidates.

GO				
ID	**Description**	**Count**	χ^2 ***p*-val**	**Adj. *p*-val**
GO:0000278	mitotic cell cycle	22	$7.48 \cdot 10^{-12}$	$1.31 \cdot 10^{-8}$
GO:1903047	mitotic cell cycle process	20	$1.76 \cdot 10^{-11}$	$1.54 \cdot 10^{-8}$
GO:0000280	nuclear division	16	$7.39 \cdot 10^{-11}$	$4.31 \cdot 10^{-8}$
GO:0007059	chromosome segregation	12	$1.36 \cdot 10^{-10}$	$5.34 \cdot 10^{-8}$
GO:0051301	cell division	17	$1.52 \cdot 10^{-10}$	$5.34 \cdot 10^{-8}$
GO:0048285	organelle fission	16	$1.91 \cdot 10^{-10}$	$5.58 \cdot 10^{-8}$
GO:0007067	mitotic nuclear division	14	$2.69 \cdot 10^{-10}$	$6.72 \cdot 10^{-8}$
GO:0007049	cell cycle	25	$8.83 \cdot 10^{-10}$	$1.93 \cdot 10^{-7}$
GO:0044772	mitotic cell cycle phase transition	13	$1.56 \cdot 10^{-8}$	$2.88 \cdot 10^{-6}$
GO:0098813	nuclear chromosome segregation	9	$1.71 \cdot 10^{-8}$	$2.88 \cdot 10^{-6}$
Reactome				
ID	**Description**	**Count**	χ^2 ***p*-val**	**Adj. *p*-val**
68877	Mitotic Prometaphase	7	$1.58 \cdot 10^{-6}$	$3.15 \cdot 10^{-4}$
1640170	Cell Cycle	13	$3.1 \cdot 10^{-6}$	$3.15 \cdot 10^{-4}$
69278	Cell Cycle, Mitotic	11	$1.65 \cdot 10^{-5}$	$1.12 \cdot 10^{-3}$
2514853	Condensation of Prometaphase Chromosomes	3	$3.08 \cdot 10^{-5}$	$1.56 \cdot 10^{-3}$
2500257	Resolution of Sister Chromatid Cohesion	5	$1.9 \cdot 10^{-4}$	$6.91 \cdot 10^{-3}$
69481	G2/M Checkpoints	4	$2.16 \cdot 10^{-4}$	$6.91 \cdot 10^{-3}$
1538133	G0 and Early G1	3	$2.38 \cdot 10^{-4}$	$6.91 \cdot 10^{-3}$
195258	RHO GTPase Effectors	7	$3.86 \cdot 10^{-4}$	$8.92 \cdot 10^{-3}$
68886	M Phase	7	$3.96 \cdot 10^{-4}$	$8.92 \cdot 10^{-3}$
69620	Cell Cycle Checkpoints	5	$7.31 \cdot 10^{-4}$	$1.33 \cdot 10^{-2}$
KEGG				
ID	**Description**	**Count**	χ^2 ***p*-val**	**Adj. *p*-val**
hsa04110	Cell cycle	7	$6.77 \cdot 10^{-6}$	$7.72 \cdot 10^{-4}$

aggressive prostate cancer in an independent analysis, that we will describe in the next chapter.

4.6.4.3 Correlation of the known fusion partners with metastasis

None of the 45 known 3' fusion partners in Supplementary Table A.3 is in the list of 79 step-up candidates significantly correlated with the metastasis (Section 4.6.4.1).

On the other hand, 7 out of the 55 known 5' fusion partners (Supplementary Table A.2) are significantly associated with metastasis (Table 4.7), namely *AZGP1*, which also is significant in the survival analysis on CancerMap, *FOXP1*, *PTEN*, *FKBP5*, *ALG5*, *TMPRSS2* and *KLK2*.

We focused our analysis on several of these candidates. We tried to determine if the positions at which we identify jumps are consistent with the fusion breakpoints reported in literature. We also verified where possible if the step-down jumps are correlated with the jumps in their know fusion partners, as we describe next.

4.6.4.3.1 ***AZGP1*** The *AZGP1* gene is the top candidate with an FDR adjusted p-value of $2.31 \cdot 10^{-11}$ (Table 4.7), generated by the over-representation of the step-down jumps of this gene in the metastatic samples. More specifically, we identified step-down jumps in 12/19 unique metastatic samples and only 7/160 primary prostate samples. The jumps are always occurring after the first exon (marked with a red arrow in Figure 4.10).

Pflueger et al. [263] report a read-through transcription involving the exons 1-2 of *AZGP1* and the exon 1 of *GJC3*, an adjacent gene, located 50 kilobases downstream, on the same strand. The read-through transcription has been also identified by Nacu et al. [265], which found several paired RNA-seq reads spanning the exon 2 of *AZGP1* and exon 2 of *GJC3*. The position at which the gene is reported fused is represented with a blue arrow.

We note that the position where the jump starts is not matching with the position of the reported breakpoints. We also note that the DABG p-values (Section 4.3.4) corresponding to the first probeset are above 0.05 in all samples, suggesting that the probeset is not functional. As the probeset is not functional, it will have a constant intensity across samples, while the intensities of the other probesets will vary according to the expression of the exons they align to. This might explain the generation of the jumps as the ones described above.

The *GJC3* gene has only two exons, which are too few for the step method be able to identify jumps.

Table 4.6 The top pathways from the GO and Reactome databases over-represented in the set of 200 step-down candidates.

GO				
ID	**Description**	**Count**	χ^2 ***p*-val**	**Adj. *p*-val**
GO:0009888	tissue development	48	$1.4 \cdot 10^{-11}$	$2.97 \cdot 10^{-8}$
GO:0048731	system development	81	$1.89 \cdot 10^{-11}$	$2.97 \cdot 10^{-8}$
GO:0048856	anatomical structure development	89	$5.52 \cdot 10^{-11}$	$5.79 \cdot 10^{-8}$
GO:0008150	biological process	179	$3.01 \cdot 10^{-10}$	$2.36 \cdot 10^{-7}$
GO:0006936	muscle contraction	18	$4.18 \cdot 10^{-10}$	$2.62 \cdot 10^{-7}$
GO:0009653	anatomical structure morphogenesis	58	$8.5 \cdot 10^{-10}$	$4.45 \cdot 10^{-7}$
GO:0003012	muscle system process	19	$1.31 \cdot 10^{-9}$	$5.53 \cdot 10^{-7}$
GO:0030154	cell differentiation	70	$1.48 \cdot 10^{-9}$	$5.53 \cdot 10^{-7}$
GO:0007275	multicellular organismal development	83	$1.58 \cdot 10^{-9}$	$5.53 \cdot 10^{-7}$
GO:0048869	cellular developmental process	72	$4.43 \cdot 10^{-9}$	$1.3 \cdot 10^{-6}$
GO:0044707	single-multicellular organism process	102	$4.87 \cdot 10^{-9}$	$1.3 \cdot 10^{-6}$
GO:0048468	cell development	48	$4.95 \cdot 10^{-9}$	$1.3 \cdot 10^{-6}$
GO:0044767	single-organism developmental process	90	$8.6 \cdot 10^{-9}$	$2.08 \cdot 10^{-6}$
GO:0007155	cell adhesion	37	$1.87 \cdot 10^{-8}$	$3.97 \cdot 10^{-6}$
GO:0032502	developmental process	90	$1.9 \cdot 10^{-8}$	$3.97 \cdot 10^{-6}$
GO:0022610	biological adhesion	37	$2.09 \cdot 10^{-8}$	$4.1 \cdot 10^{-6}$
GO:0030198	extracellular matrix organization	18	$3.38 \cdot 10^{-8}$	$6.14 \cdot 10^{-6}$
GO:0043062	extracellular structure organization	18	$3.52 \cdot 10^{-8}$	$6.14 \cdot 10^{-6}$
GO:0032501	multicellular organismal process	102	$4.29 \cdot 10^{-8}$	$7.1 \cdot 10^{-6}$
GO:0031589	cell-substrate adhesion	15	$2.01 \cdot 10^{-7}$	$3.17 \cdot 10^{-5}$
Reactome				
ID	**Description**	**Count**	χ^2 ***p*-val**	**Adj. *p*-val**
397014	Muscle contraction	11	$3.2 \cdot 10^{-10}$	$9.71 \cdot 10^{-8}$
445355	Smooth Muscle Contraction	7	$5.29 \cdot 10^{-8}$	$8.02 \cdot 10^{-6}$
1474244	Extracellular matrix organization	11	$2.91 \cdot 10^{-4}$	$2.94 \cdot 10^{-2}$

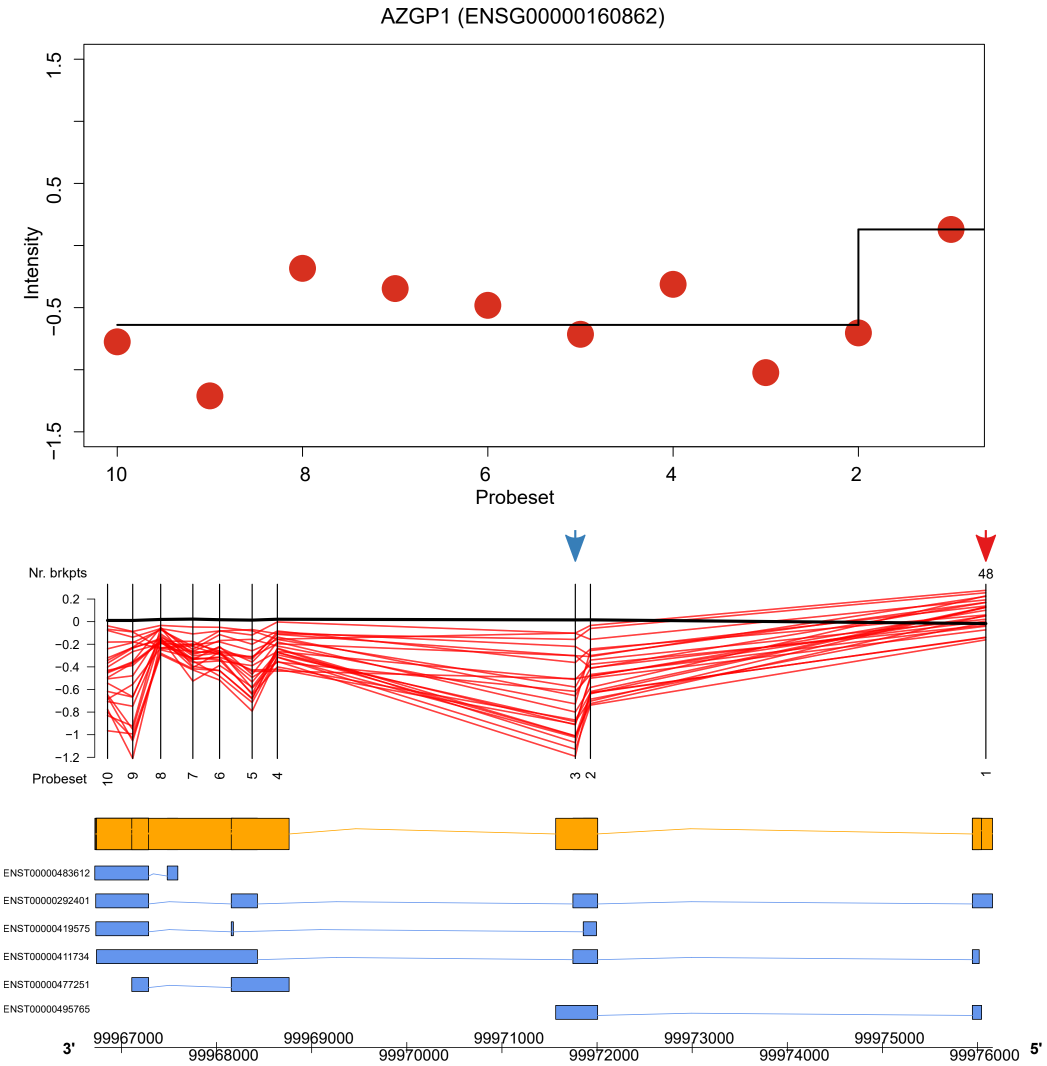

Figure 4.10 Genomic plot depicting the mapping of the jumps to the *AZGP1* gene model. In the top panel we depict a representative step-down jump. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model, while each read line links the intensities of two consecutive probesets in a sample with step-down jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The number underneath the red arrow represents the number of putative breakpoints identified at that position. The blue arrow indicates the position where the gene breakpoint have been reported in the literature.

Table 4.7 The 5' known fusion participants whose down-step jumps are significantly correlated with the metastasis in the MSKCC dataset. The **Mechanism** column specifies the mechanism reported to have produced the fusion in the source paper (chromosomal rearrangement or read-through transcription). The columns χ^2 ***p*-value** and **Adj. *p*-val** indicate the χ^2 p-value and respectively the FDR-adjusted p-value. The **Mets** column indicates the number of step-down in the metastatic samples. The **Primary** column indicates the number of jumps in the primary tissue samples.

Gene Symbol	Mechanism	χ^2 *p*-val.	Adj. *p*-val.	Mets	Primary
AZGP1	read-through	$\mathbf{7.98 \cdot 10^{-14}}$	$\mathbf{2.3 \cdot 10^{-11}}$	12/19	7/160
FOXP1	rearrangement	$\mathbf{2.56 \cdot 10^{-6}}$	$\mathbf{8.91 \cdot 10^{-5}}$	5/19	2/160
PTEN	rearrangement	$\mathbf{1.56 \cdot 10^{-5}}$	$\mathbf{4.35 \cdot 10^{-4}}$	8/19	11/160
FKBP5	rearrangement	$\mathbf{4.2 \cdot 10^{-4}}$	$\mathbf{6.87 \cdot 10^{-3}}$	8/19	16/160
ALG5	rearrangement	$\mathbf{1.85 \cdot 10^{-3}}$	$\mathbf{2.27 \cdot 10^{-2}}$	4/19	4/160
TMPRSS2	rearrangement	$\mathbf{4.62 \cdot 10^{-3}}$	$\mathbf{4.23 \cdot 10^{-2}}$	9/19	27/160
KLK2	rearrangement	$\mathbf{4.72 \cdot 10^{-3}}$	$\mathbf{4.23 \cdot 10^{-2}}$	4/19	5/160

4.6.4.3.2 ***FOXP1*** For *FOXP1* we identified step-down jumps in 5/19 unique metastatic samples and 2/160 primary tissue samples (FDR adjusted χ^2 p-value $8.91 \cdot 10^{-5}$). The breakpoints are occurring at various positions, marked with red arrows in the Figure 4.11.

The fusions involving the exons 1-11 of *FOXP1* and exons 5-12 of *ETV1* have been previously reported by Hermans et al. [266]. Also fusions between *FOXP1* and *MIPEP* [267] and *DMPK* [54] have been found. We marked the position of the confirmed breakpoints in these fusions with blue arrows. We note that some of the breakpoints identified by our method seem to be consistent with the confirmed fusions.

We also tried to determine if the step-down jumps in *FOXP1* are correlated with step-up jumps in its known fusion partners, namely *ETV1*, *MIPEP* and *DMPK*. For *DMPK* we do not detect any jump, while for *ETV1* none of the 14 step-up jumps we identify is overlapping with the step-down jumps in *FOXP1*. For *MIPEP*, we identify 2/20 step-up jumps matching with the step-down jumps of *FOXP1*, but which are not significantly correlated (χ^2 p-value 0.2072).

4.6.4.3.3 ***PTEN*** We identified step-down jumps in *PTEN* in 8/19 metastatic samples and 11/160 (FDR adjusted χ^2 p-value $4.35 \cdot 10^{-4}$). The jumps start at different probesets, such as 3, 7, 10, 14, as illustrated in Figure 4.12.

PTEN-PLCE1 fusions, involving exon 1 of *PTEN* and exon 30 of *PLCE1* have been reported by Baca et al. [54]. The probesets 1-8 align to exon 1 an therefore most of the breakpoints are consistent with the reported *PTEN-PLCE1* fusions. However, none

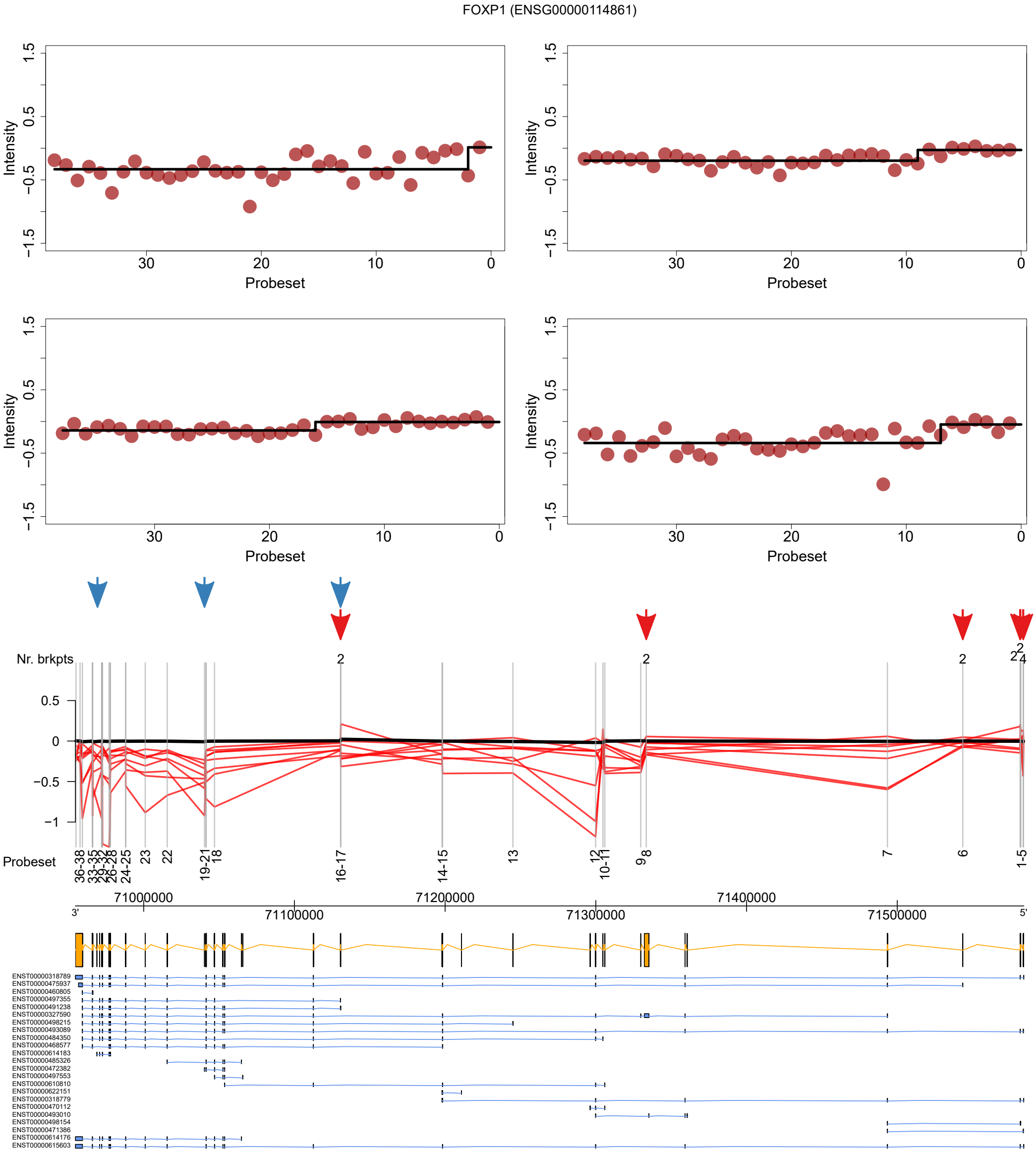

Figure 4.11 Genomic plot depicting the mapping of the jumps to the *FOXP1* gene model. In the top panel we depict several representative step-down jumps. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model. Each read line links the intensities of two consecutive probesets in a sample with step-down jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. The blue arrows represent the positions of breakpoints reported in the literature. In the bottom panel it is represented the gene model.

of the 12 step-up jumps identified in *PLCE1* occur in samples with *PTEN* step-down jumps.

4.6.4.3.4 ***FKBP5*** *FKBP5* exhibits step-down jumps in 8/19 metastatic samples and 16/160 primary samples (FDR adjusted χ^2 p-value $6.87 \cdot 10^{-3}$). The step jumps occur at various positions along the gene, as illustrated with red arrows in Supplementary Figure A.16.

For *FKBP5*, Pflueger et al. [263] discovered four variants of the triple *TMPRSS2-FKBP5-ERG* fusion, that is a fusion transcript that contains the exon 1 or exons 1-4 of *TMPRSS2*, exon 8, or exons 8-9 of *FKBP5* and exons 5-13 of *ERG*. The positions of exons 8 and 9 are depicted with blue arrows Supplementary Figure A.16. As the fusion is quite complex, it is difficult to predict the expected expression pattern across the exons. Most of the jumps occur before exons 8-9, which would be consistent with an early interruption of the gene expression.

We correlated the step-down jumps in *FKBP5* with the step-up jumps in *ERG* and step-down jumps in *TMPRSS2*. The step-up jumps in *ERG* are almost mutually exclusive with the step-downs in *FKBP5*. Out of 49 unique samples that exhibit step-up jumps in *ERG* and 24 unique samples that show step-down samples in *FKBP5*, only 2 samples are in common. We performed a χ^2 with the null hypothesis that the step-up jumps in *ERG* and step-down jumps in *FKBP5* occur independently, with the alternative hypothesis that the jumps are dependent, which yielded a marginally insignificant p-value (0.055).

Another χ^2 test, testing the correlation between the step-down jumps in *FKBP5* with the step-down jumps in *TMPRSS2* produced a very low p-value ($1.63 \cdot 10^{-5}$), suggesting a strong association between the step-down jumps in the two genes. 14/24 *FKBP5* step-down jumps occur in samples with step-down jumps in *TMPRSS2* as well.

4.6.4.3.5 ***TMPRSS2*** *TMPRSS2* step-down jumps are over-represented in the metastatic samples (χ^2 p-value $4.62 \cdot 10^{-3}$). 9/19 of metastatic samples exhibit step-down jumps and 27/133 primary samples exhibit it.

The most common *TMPRSS2-ERG* fusion is between exon 1 of *TMPRSS2* and exon 4 of *ERG*, occurring in about 44% of fusion positive samples, and *TMPRSS2* exon 1 with *ERG* exon 5, in 4% of cases [135]. Besides there are several other variants, found with a lower frequency, such as T1-E2 (that is exon 1 of *TMPRSS2* and exon 2 of *ERG*), T4-E4, T4-E5, T5-E4, T5-E5, T2-E5, T1-E3, T2-E2, T2-E4, T3-E4, T1-E6, T1-E3,5 (that is *TMPRSS2* exon 1 with *ERG* exon 3 and 4 - exon 4 being spliced out),

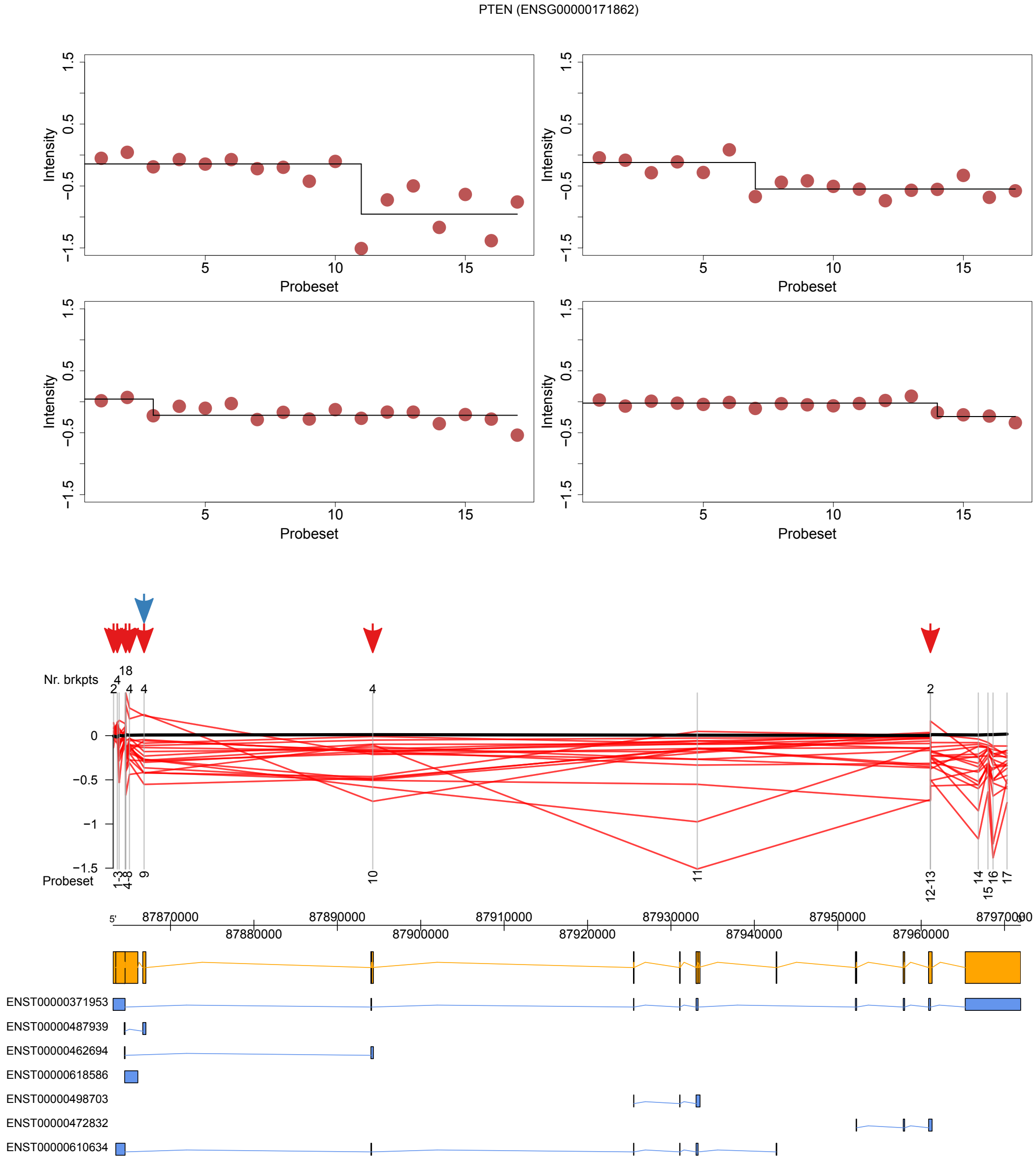

Figure 4.12 Genomic plot depicting the mapping of the jumps to the *PTEN* gene model. In the top panel we depict several representative step-down jumps. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model. Each read line links the intensities of two consecutive probesets in a sample with step-down jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. The blue arrows represent the positions of breakpoints reported in the literature. In the bottom panel it is represented the gene model.

T1-E2,3,4,6, T1-E3a4 (an alternative exon 3 of *ERG* is used), T1-E3b, and T1-E3c [126].

We mapped these breakpoints to our data and, as can be seen in Figure 4.13 and Supplementary Figure A.19, most of them are in agreement with our predictions. There are, however, some discrepancies in *TMPRSS2*, as we identify jumps after the first two probesets of *TMPRSS2*, which, to our knowledge, are not documented breakpoints. Moreover, the exons corresponding to these probes do not even appear in the UCSC annotation of *TMPRSS2*, which starts at the exon where our third probeset maps.

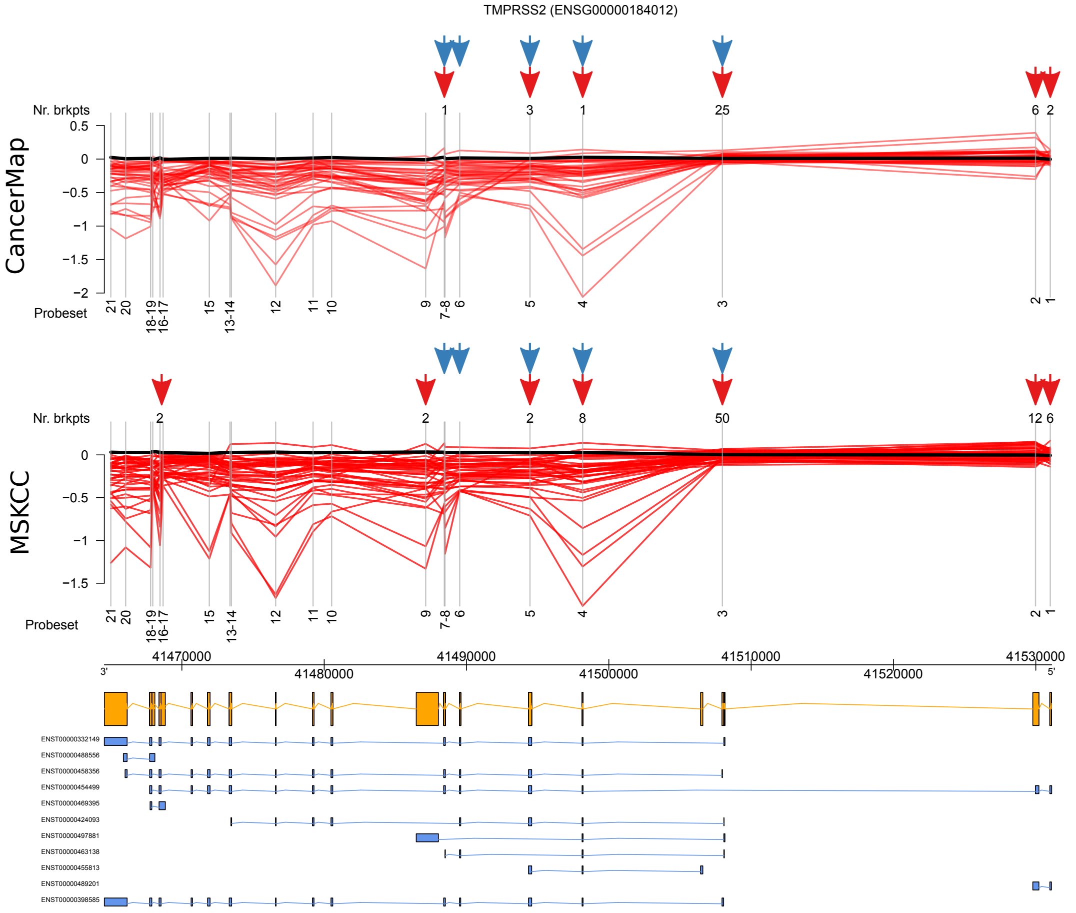

Figure 4.13 Genomic plot depicting the mapping of the jumps in CancerMap (top panel) and MSKCC (middle panel) to the *TMPRSS2* gene model. In the two top panels, the vertical lines correspond to the position where the probesets align to the gene model, while each read line links the intensities of two consecutive probesets in a sample with step-down jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. The blue arrows represent the positions of breakpoints reported in the literature. In the bottom panel it is represented the gene model.

4.7 Discussion

Here we presented a novel approach for the identification of fusion candidates based on exon microarrays. In the first part of this chapter we described the background and the motivation that lead us to develop the method. We then described in detail the approach and how it compares with other existing approaches. Based on the experimentally validated data on *ERG* gene, available in two datasets we found indications that our method performs at least as good as the existing methods and that it can be extended on new datasets. However, due to limited amount of validation data, we could not assess how well the method generalises to genes other than *ERG*, which has been used for training the model parameters. We could just get some indirect indications that the method works on other datasets, by identifying candidate genes known to be involved in fusions in prostate cancer.

We applied the method on several prostate cancer datasets in which we have been able to discover jumps in known fusion partners, such as *TMPRSSS2*, *ERG* and *ETV1*, *ETV4* and *ETV5*. The analysis, however, produced thousands of candidates. We tried to prioritise the candidates that show correlation between the jumps and the time to BCR or metastatic samples.

The survival analysis did not identify any candidates with step-up jumps robustly associated with the BCR time. The survival analysis on the step-down candidates, on the other hand, found nine candidates correlated with the time to BCR in MSKCC and also in CancerMap. Further analyses are required for establishing the cause of jumps in these candidates.

The correlation between metastatic samples and jumps, identified 79 step-up candidates and respectively 548 step-down candidates with statistically significant correlations in MSKCC. A weakness of this analysis is that it is based only on the MSKCC dataset, as CancerMap does not contain metastatic samples.

Nonetheless, the pathway analysis revealed that the candidates associated with metastasis are involved in key pathways in prostate cancer progression. Up to 25 candidates in the list of 79 candidates for which the step-up jumps are associated with the metastasis are involved in cell cycle processes. The expression of these genes seems to be correlated with the proliferation of tumours [171]. Cell cycle genes are at the base of the Prolaris test. Therefore a more in depth study of the jumps in these genes, might reveal useful information regarding aggressive prostate cancer.

Also many step-down candidates are involved in the muscle contraction and extracellular matrix organization pathways. As we will show in the next chapter, we have been able to identify an aggressive subtype of prostate cancer which is characterised by the down-regulation of the genes involved in these processes.

It is still not clear to us why for the known fusion partners the jumps do not always match with the previously reported fusion locations. One possibility is that the breakpoints occur at previously unreported positions.

4.7.1 Candidates discussion

4.7.1.1 *AR*

AR plays a key role in the male sexual development and it is also crucial for the development and the progression of prostate cancer [268]. It is also a therapeutical target in prostate cancer. The hormone therapy aims to block the androgens and the *AR*, which initially stops the progression of cancer. However, after a while the expression of *AR* increases again or *AR* acquires mutations, which make it unresponsive to anti-androgens, leading to the progression of the disease to the castrate resistant prostate cancer(CPRC) stage [268], and from there to metastasis.

As *AR* is a key gene in prostate cancer, its structure and function have been intensively studied. Despite frequent mutations being reported, to our knowledge it has not been reported as being fused, especially at such a high frequency (jumps in approx. 5% of samples).

In the light of these observations, it is highly unlikely that the step-up jumps we are identifying are produced by fusions. It seems rather that the observed jumps are generated by alternative splicing, as the first probeset maps to a region that is sometimes spliced out.

4.7.1.2 *AZGP1*

AZGP1 is a gene involved in the processing of lipids, located on the arm p of the chromomsome 7, on the "-" strand. Henshall et al. [269] indicated that low levels of *AZGP1* are an independent predictor for clinical recurrence of prostate cancer (HR 4.8, 95% CI 2.2 - 10.7) and of metastasis (HR 8.0, 95% CI 2.6 - 24.3). The results have been confirmed in subsequent studies [270–273]. Low levels of *AZGP1* seem to also predict the relapse in positive surgical margins localised prostate cancer [273], early biochemical recurrence [272], and are strongly associated with *TMPRSS2-ERG* fusions, *PTEN* deletions, Gleason score, pathological stage and positive node status [272]. Furthermore, decreased expression of *AZGP1* was also associated with poor prognosis in other types of cancer, such as gastric cancer [274].

In our analysis the jumps seem to be the product of a non-functional probeset that yields constant intensities across samples, while the functioning probesets are down-regulated in the metastatic samples, as previously reported [269].

4.7.1.3 *FOXP1*

FOXP1 is a tumour suppressor gene [275]. The 3p13 region, where *FOXP1* and several other genes, including *GPR27*, *PROK2*, *GXYLT2*, *EIF4E3*, *EIF3E4*, *RYBP* and *SHQ1*, are located is affected by recurrent deletions in prostate cancer [242]. The deletions are highly correlated with the *TMPRSS2-ERG* positive prostate cancer [242] and also linked to advanced stage (p-value < 0.0001), high Gleason (p-value = 0.0125) and early PSA recurrence (p-value < 0.0001) [276].

A large cohort study of [276] identified *FOXP1* deletions in 17% of prostate cancer. Of these, around 3% are partial deletions of either 3' or 5' end of *FOXP1*. We note that the 3' deletions have the potential to result in jumps. Also, and in around 2% of cases *FOXP1* is involved in translocations [276], as a 5' partner for fusions involving *ETV1* [266], *MIPEP* [267] and *DMPK* [54], which can also result in the step-down jumps.

Since the deleted region 3p13, spans several other genes, besides *FOXP1*, we tried to assess if there is also a significant under-expression of these genes in the samples where *FOXP1* exhibits jumps, which would be consistent with a deletion of the region. In brief, for the genes reported to be involved in the deletions of 3p13, namely *FOXP1*, *GPR27*, *GXYLT2*, *EIF4E3*, *EIF4E4*, *RYBP* and *SHQ1* genes we estimated the expression level in each sample, by averaging the intensities of all the probesets. For the *PROK2* and *EIF3E4* we could not asses the expression levels, as they were not annotated in our data.

For each of these genes we performed a Mann-Whitney U independence test, assessing the hypothesis that the expressions of these genes in samples with *FOXP1* jumps vs. no-jumps come from the same distribution, with the alternative hypothesis, that the distributions are different. As depicted in Supplementary Figure A.17, the p-values are highly significant for *FOXP1* ($2.94 \cdot 10^{-9}$), *GXYLT2* ($6.72 \cdot 10^{-7}$), *EIF4E3* ($2.35 \cdot 10^{-6}$), *RYBP* ($2.09 \cdot 10^{-4}$) and *SHQ1* ($2.94 \cdot 10^{-2}$), while it is not significant only in *GPR27* (0.37). This suggests that a significant proportions of the observed jumps might be generated by some form of chromosomal deletion.

4.7.1.4 *PTEN*

We identify step-down jumps in *PTEN* in around 10% of MSKCC samples. *PTEN* is a tumour suppressor gene located on the arm q23.3 of chromosome 10. It negatively regulates the PIK3/Akt pathway, which has an important role in controlling, among other things, cell growth, cell proliferation and apoptosis [57].

Deletions of 10q23 are one of the most common events in prostate cancer, along with *TMPRSS2-ERG* fusions, being reported in 30-60% of adenomacarcinomas, of which 10-30% are homozygous deletions [145–149], with a higher frequency in CPRC (deletions in 77% of samples and homozygous deletions in 43% [150]). The functional loss of *PTEN* might also be induced by several events such as point mutations, reported in around 16% of prostate cancers [143, 144], and methylation. However fusions of *PTEN* have been very rarely reported.

Inactivation of *PTEN* in general is linked to progression of prostate cancer and significantly worse survival outcome [151, 152], while homozygous deletions are associated with much faster biochemical recurrence [147]. There is also a significant association between *PTEN* loss and *TMPRSS2-ERG* fusions. In one study all samples with *PTEN* deletions also harboured TMPRSS2-ERG fusions [149]. It has been hypothesised that the interaction between *PTEN* deletions and *TMPRSS2-ERG* fusions prostate is a significant driver for prostate cancer development and progression [153]. Bismar et al. [153] suggested that initial hemizygous loss of *PTEN* would promote genomic instability and facilitate gene fusions, leading to the formation of prostate cancer. Subsequent *PTEN* homozygous loss, would trigger further progression, to the invasive disease.

We tried to see if the genes flanking the 10q23.3 region often deleted, namely *BMPR1A* and *FAS* [277], exhibit low-expression in samples where *PTEN* exhibits down-step jumps, consistent with a deletion of the region. As presented in Supplementary Figure A.18, there seems to be a strong association between the step-down jumps and the under-expression of all three genes. Mann-Whitney tests yielded very low p-values, namely $1.83 \cdot 10^{-18}$ for *PTEN*, $2.3 \cdot 10^{-9}$ for *BMPR1A* and $9.63 \cdot 10^{-5}$ for *FAS*, suggesting deletions of 10q23.3 in many samples with step-down jumps in *PTEN*.

4.7.1.5 *FKBP5*

The *FKBP5* gene seems to have important roles in cancer progression and drug resistance. *FKBP5* is an androgen-regulated gene which is a therapeutic target in the hormone therapy [278, 279]. Also, low levels of *FKBP5* have been associated with increased activity of the AKT pathway, with important roles in cancer proliferation, and decreased chemosensitivity [280].

In our analysis the step-down in *FKBP5* are highly correlated with step-down in *TMPRSS2* and marginally significantly correlated with the step-up jumps in *ERG*. The jumps seem to be consistent with previous reported triple fusions involving simultaneously *TMPRSS2*, *FKBP5* and *ERG*. However further validations are required to elucidate the source of the jumps.

4.7.1.6 *TMPRSS2*

TMPRSS2 is an androgen-regulated gene located on chromosome 21, highly expressed in prostate [133]. The fusions involving *TMPRSS2* and a member of *ETS* family is the most common genomic alteration in prostate cancer.occur very frequently in prostate cancers. The most common fusion is the *TMPRSS2-ERG* fusion, which occurs in 40-55% of prostate cancers [32, 127–131].

The *TMPRSS2-ERG* gene fusions seem to be an early event in the development of prostate cancer. It is found in a high percentage of HGPIN [136], but seem to be insufficient to induce the formation of prostate cancer by their own [137]. Also, the overexpression of *ETS* genes seems to contribute to cell invasion and migration [136].

The usefulness of *TMPRSS2-ETS* fusions as a biomarker in predicting the clinical outcome is controversial. A number of studies reported an association between the *TMPRSS-ERG* fusions and poor outcome [138–140]. However some other studies found no association [33–35].

The step-down jumps we identified in *TMPRSS2* do not seem to be significantly associated with the time to BCR (log-rank p-values 0.59 in MSKCC and 0.22 in CancerMap). However they are significantly correlated with metastasis (χ^2 p-value $4.62 \cdot 10^{-3}$) in MSKCC.

4.7.2 Conclusions

The analysis of these candidates suggests that in some cases, such as for *TMPRSS2* and *FKBP5*, the jumps are consistent with previously reported fusions. For other candidates, such as *PTEN* and *FOXP1*, the step-down jumps are associated with known deletions, while for others, such as *AR*, the jumps might be explained by alternative splicing. Further analysis is necessary for confirming these hypotheses.

The concordance between these results and the previous findings in the field, suggest that amongst the many candidates produced by the step method, some might reflect alterations generated by real biological processes. This is also supported by the statistically significant associations with the clinical outcomes in some candidates.

Furthermore, the pathway analysis suggests possible important roles in cancer for some of the novel candidates identified by our method. Therefore additional analyses on the top candidates might unravel important alterations in genes with key roles in aggressive prostate cancer.

Chapter 5

Latent process decomposition (LPD)

5.1 Summary

A critical problem in the clinical management of prostate cancer is that it is highly heterogeneous. Accurate prediction of individual cancer behaviour is therefore not achievable at the time of diagnosis leading to substantial overtreatment. It remains an enigma that, in contrast to many other cancer types, stratification of prostate cancer based on unsupervised analysis of global expression patterns has not been possible.

In this chapter we apply a Bayesian unsupervised technique called Latent Process Decomposition (LPD) to identify a common prostate cancer process (subtype), designated DESNT, in five different prostectomy datasets. DESNT cancers are characterized by down-regulation of a core set of 45 genes, many encoding proteins involved in the cytoskeleton machinery, ion transport and cell adhesion. For four datasets with linked PSA failure data following prostatectomy, assignment to DESNT predicted very poor outcome relative to non-DESNT patients.

Additionally, the analysis of a set of prostate cancers annotated in The Cancer Genome Atlas failed to reveal links between DESNT cancers and the presence of any particular class of genetic mutation, including *ETS* gene status. However, the correlation of the expression of the core set of 45 down-regulated genes in the DESNT cancers with methylation data, suggest possible roles of epigenetic changes in DESNT.

We also describe the derivation of a 20 gene signature which can predict with high accuracy DESNT membership. This approach simplifies the technical analysis necessary to determine if a new cancer is part of the DESNT subtype, which makes it suitable for clinical use.

5.2 Background

A common method for the diagnosis of prostate cancer is the measure of prostate specific antigen (PSA) in blood. However, as many as 50-80% of PSA-detected prostate cancers are biologically irrelevant, that is, even without treatment, they would never have caused any symptoms [18–20].

Prostate cancer is highly heterogeneous and accurate prediction of individual prostate cancer behaviour at the time of diagnosis is not currently possible. Immediate radical treatment for all cases has been a common approach. The overtreatment has a considerable impact on the quality of life and also leads to serious risks of treatment-related complications. Around 10-15% of the patients who undergo radical prostectomy, for example, report urinary incontinence and up to 70% have erectile problems [30]. It has been also reported that there were surgery-related complications such as infections, respiratory and cardiac problems in about 20-25% of patients who underwent surgery [84, 85].

A large number of prognostic biomarkers have been proposed for the stratification of prostate cancer, including many expression signatures. However, the signature based on expression profiling have been derived in a supervised fashion by comparing aggressive and non-aggressive cancers [40, 281, 282], or by selecting genes with specific biological functions [38, 283, 284]. Despite the important roles that these biomarkers could bring into a better management of the disease, they fail to define clear molecular subtypes of the diseases.

In contrast to other cancer types, such as breast cancer, stratification of prostate cancer based on unsupervised analysis of global expression patterns has not been possible so far. Our hypothesis was that the identification of robust subtypes has been unsuccessful because the commonly used, general purpose unsupervised methods, such as hierarchical clustering and k-means, are too simplistic to account for the high intra-tumoural heterogeneity of prostate cancer.

To address this issue, we employed a more realistic mathematical modelling of bulk-cell transcriptome data. More specifically, we used a Bayesian technique called Latent Process Decomposition (LPD) to deconvolute the heterogeneity of prostate cancer, and to identify intrinsic molecular subtypes, in a completely unsupervised fashion.

5.3 Materials

5.3.1 Datasets

We initially worked on five prostate cancer microarray datasets denoted MSKCC, CancerMap, CamCap, Stephenson and Klein. Later, we obtained an additional RNA-seq dataset, that we will refer to as TCGA.

The MSKCC and CancerMap datasets have been obtained as described in Section 4.3.1. In brief MSKCC has been downloaded from the GEO repository, while the CancerMap has been obtained by merging two smaller datasets, denoted ICR and Cambridge.

MSKCC is the only dataset that contains samples generated from metastatic tissue, cell-lines and xenografts. For consistency with the other datasets, for the current analysis we dropped all 50 samples which were not from primary tumour or normal prostate and performed the analyses on the remaining 320 samples.

The CamCap dataset is the result of combining two *Illumina HumanHT-12 V4.0 expression beadchip* (bead microarray) datasets, publicly available on GEO, under accession GSE70768 and GSE70769, published by Ross-Adams et al. [241], comprising of 199 and 94 microarrays respectively. The first dataset, GSE70768, consists of 186 radical prostectomy samples and 13 TURP (transurethral resection of the prostate) samples. As GSE70768 is the only dataset containing TURP samples, for consistency with the other datasets, we have removed them in our analyses. Out of the 186 prostectomy samples, 113 come from primary tumour samples and the rest of 73 are matched benign samples from a subset of patients. GSE70769 contains only 94 primary tumour samples. CamCap and CancerMap datasets have 40 patients in common and therefore are not independent datasets.

The Stephenson dataset [285] contains 89 *Affymetrix U133A human gene arrays*, from patients with clinically localised prostate cancer treated with radical prostectomy. Out of 89 samples, 78 are from primary tumour and 11 from non-malignant prostate tissue.

The fifth dataset which we refer to as Klein, published by Klein et al. [286], is available in GEO under accession GSE62667. It consists of 182 formalin-fixed and paraffin-embedded (FFPE) primary tumour samples analysed with *Affymetrix Human Exon 1.0 ST Arrays*. Unlike the other datasets presented so far, the Klein dataset does not provide clinical data.

The TCGA dataset is produced by The Cancer Genome Atlas (TCGA) Research Network and is freely available on TCGA Data Portal. From this portal we downloaded a set of RNA-seq (Illumina HiSeq 2000) data generated from 376 fresh-frozen primary prostate tissue samples, from 333 patients who underwent prostectomy. 333 samples have been extracted from tumour tissue and 43 have been obtained from the non-malignant adjacent prostate tissue of some of the 333 patients.

An appealing feature of the TCGA data is that for all 333 tumour samples and 30/46 normal samples there is also available DNA methylation data, obtained using the Illumina Infinium HumanMethylation 450K platform, which provides 485,777 probes, most of them mapping to CpG sites. Also, for the TCGA samples we were able to obtain from The Cancer Genome Atlas Research Network [287] the *ETS* fusion status in all the malignant samples, and the mutations and deletions status of the most commonly mutated genes in prostate cancers.

A summary of the datasets are presented in Table 5.1.

Table 5.1 Summary of the prostate cancer datasets used in the LPD analysis. For each dataset we present the total number of microarrays used in the analysis, the number of unique patients, the number of primary tumour samples, the number of primary samples from unique patients, the number of benign samples, the number of benign samples from unique patients, an indication if the dataset provides linked clinical data, the platform used to generate the data and the source from where the data has been retrieved.

	Nr. Samples		**Primary tumour**		**Benign**	
	Total	**Unique**	**Total**	**Unique**	**Total**	**Unique**
MSKCC	320	160	262	131	58	29
CancerMap	235	154	209	137	24	17
CamCap	280	207	207	207	73	73
Stephenson	89	89	78	78	11	11
Klein	182	182	182	182	0	0
TCGA	376	333	333	333	43	43
TCGA (methyl)	363	333	333	333	30	30

	Follow-up	**Tissue**	**Platform**	**Citation**
MSKCC	Y	FF	Affymetrix Exon 1.0 ST	[242]
CancerMap	Y	FF	Affymetrix Exon 1.0 ST	NA
CamCap	Y	FF	Illumina HT12 v4 BeadChip	[241]
Stephenson	Y	FF	Affymetrix U133A	[285]
Klein	N	FFPE	Affymetrix Exon 1.0 ST	[286]
TCGA	Y	FF	Illumina HiSeq 2000	[287]
TCGA (methyl)	NA	FF	Illumina Infinium 450K	[287]

5.3.2 Clinical data

Five datasets have associated follow-up data, namely MSKCC, CancerMap, CamCap, Stephenson and TCGA. MSKCC, CancerMap and CamCap may contain several malignant samples from the same patient. In analyses that do not depend on clinical correlations, such as the LPD clustering, we used all the samples available for each dataset. However, for the clinical correlations, or other analyses that are sensitive to over-representation of samples, we used only one sample per patient.

In the CamCap and CancerMap datasets, there are up to four samples per patient, extracted from tissue that contained variable amounts of tumour, benign tissue and stroma. For each patient we selected the sample with the highest percentage of tumour tissue. In the MSKCC dataset there are exactly two technical replicates from the same RNA samples for each patient. For the MSKCC dataset we selected at random one of the two replicates.

For all five datasets with linked follow-up data, the endpoint for clinical outcome is time to biochemical recurrence (BCR), calculated from the time of radical prostectomy. The clinical summaries of the five datasets are presented in Table 5.2.

5.3.3 Microarray pre-processing

5.3.3.1 MSKCC, CancerMap and Klein

We normalised the three exon microarray datasets, MSKCC, CancerMap and Klein using the RMA algorithm [180], described in Section 2.6.2.1. The RMA algorithm background corrected, quantile normalised and summarised the data to produce gene-level estimates. For performing this task we used the Affymetrix Expression Console software package [288], which normalised the core probesets and annotated the summarised genes to the UCSC Human Genome 19 (hg19). The effect of the RMA normalisation is depicted in Supplementary Figure B.1. We note that the normalisation seems to bring the intensities of all microarrays to relatively similar levels.

The quality assessment of the MSKCC dataset and the two datasets that make-up the CancerMap datasets has been discussed at length in Section 4.3.5. In brief, the overall quality of the microarrays is good and, even though few microarrays exhibit outlier for one of the two quality metrics we calculated, there is no strong evidence for removing any microarray.

The overall quality of the Klein dataset seems to be lower compared with the other two datasets, MSKCC and CancerMap. In the case of the MSKCC and CancerMap data the AUC values are around and above 0.8 (Figure 4.2). For Klein, the same values are

Table 5.2 Clinical summaries for the MSKCC, CancerMap, CamCap, Stephenson and TCGA datasets.

	MSKCC	CancerMap	CamCap	Stephenson	TCGA
Gleason					
4	0	0	0	1	0
5	0	0	2	0	0
6	41	35	35	15	27
3+4	54	75	102	43	116
4+3	22	17	40		61
8	8	3	13	10	42
9	7	7	11	9	84
10	0	0	1	0	3
Unknown	1	0	3	0	0
Stage					
T0	0	0	1	0	0
T1c	0	1	0	34	0
T2a	9	5	7	16	9
T2b	47	1	36	19	7
T2c	29	45	27	7	112
T2x	0	22	13	0	0
T3a	28	47	94	2	111
T3b	10	14	24	0	83
T3c	2	0	0	0	0
T4	6	2	2	0	6
Tx	0	0	3	0	5
PSA (ng/mL)					
<4	22	8	11	8	NA
4<PSA<10	78	89	136	42	NA
10<PSA<20	20	33	47	17	NA
>20	10	4	8	11	NA
Unknown	1	3	5	0%	NA
Age (years)					
Median	57.99	61	62	61.1	61.59
Mean	58.03	60.11	60.55	60.6	61.2
IQR	53.53-62.11	56-65	56-73	56.75-65.50	56.40-66.53
Range	37.3-83	21-74	41-73	44.9-72.7	43.52-76.88
Follow-up (months)					
Median	46.49	56	36.59	60.35	33.29
Mean	48.19	52.27	39.98	50.56	52.31
IQR	27.73-61.44	39-64	17.36-59.9	16.68-72.02	10.85-74.55
Range	1.38-149.2	1-129	0.36-103.4	1.40-105.7	0.08-378.4
BCR					
Failures	27	35	64	38	40

around 0.7-0.75 (blue points in the Supplementary Figure B.2). Also, the estimates for the MAD of the residuals, depicted with red dots in Supplementary Figure B.2, vary greatly across the microarray.

The poorer quality of the Klein data was expected given that the samples come from formalin-fixed paraffin-embedded tissues, which usually yield lower quality DNA, compared to fresh-frozen tissues, which have been used for the MSKCC and CancerMap datasets [188]. However, for the analysis we have used all the samples available in the Klein dataset, as none of them seems to produce hugely different estimates for the two metrics, compared with the other samples in the dataset.

5.3.3.2 CamCap

The NCBI GEO portal provides readily normalised data for the two datasets, GSE70768 and GSE70769, that make up the CamCap dataset. We therefore retrieved the normalised datasets, pre-processed as described in Ross-Adams et al. [241]. In brief, the data we downloaded has been previously log-transformed and quantile normalised using the *beadarray* R package [289]. The batch effects in each dataset have been mitigated using the ComBat algorithm, implemented in the R package *sva* [290]. We annotated the probes to UCSC hg19 using *illuminaHumanv4.db* R annotation package [291].

The bead arrays probes can be divided into three categories *perfect*, *bad* and *no match* [292]. In line with Ross-Adams et al. [241], we restricted the down-stream analysis only to *perfect* probes, as the probes from the other two categories have poor quality [292].

5.3.3.3 Stephenson

For the Stephenson dataset we only had access to the normalised data, as described in Stephenson et al. [285]. We annotated the Stephenson dataset to hg19, using the *hgu133a.db* R annotation package [293].

5.3.3.4 TCGA transcriptome

For the TCGA dataset we downloaded the level 3 data [294], generated using the Illumina HiSeq 2000 sequencing platform, processed to obtain raw counts using the RSEM algorithm [295], and annotated to UCSC hg19. We further normalised the raw count data using the Variance stabilizing transformation algorithm implemented in the DESeq2 R package [296].

5.3.3.5 TCGA methylation

For the methylation TCGA dataset we downloaded the level 3 data, which contains for each probe in each sample a beta value. Beta values take values between 0 and 1, proportional with the level of methylation measured by the probe [297]. Informally, the beta value corresponds to the percentage of methylation of the area assessed by the probe. For this dataset no further pre-processing was required, as the data was annotated to UCSC hg19.

5.3.3.6 Batch effect removal

Two of our datasets, CancerMap and CamCap, have been created by merging in each case two independent smaller datasets. The CancerMap data has been built by merging the ICR and Cambridge datasets, while CamCap is the result of the combination of GSE70768 and GSE70769 series. The microarray normalisation algorithms, such as RMA, normalise the samples in a dataset relative to each other. However, even though the differences between the microarrays within a given dataset are mitigated, the microarrays from independent datasets are usually not comparable. This can be seen in Figure 5.1a,c. We can see that the boxplots corresponding to samples from different datasets are at different levels.

To mitigate the dataset-specific effects we employed the *ComBat* algorithm, implemented in the *sva* R package [290]. ComBat transformed the intensities of probes from different samples to the same distribution across datasets. As it can be seen in Figure 5.1b and d, the resulted boxplots are at consistent levels across datasets, suggesting comparable expression levels of the samples generated in different centres.

5.4 Methods

5.4.1 Latent process decomposition (LPD)

We presented a technical description of LPD in Section 3.2.1.3. To summarize the aspects presented there, we note that LPD is a hierarchical Bayesian model that can perform probabilistic clustering of microarray data. This means that LPD allows objects to have partial membership to more than one cluster, reflecting the fact that a given object can share some characteristics with a group of objects, but in the same time it can share other characteristics with a different group of objects.

In the context of prostate cancer we assume that a cluster represents a biological processes that leads to a certain expression pattern. As prostate cancer is a highly

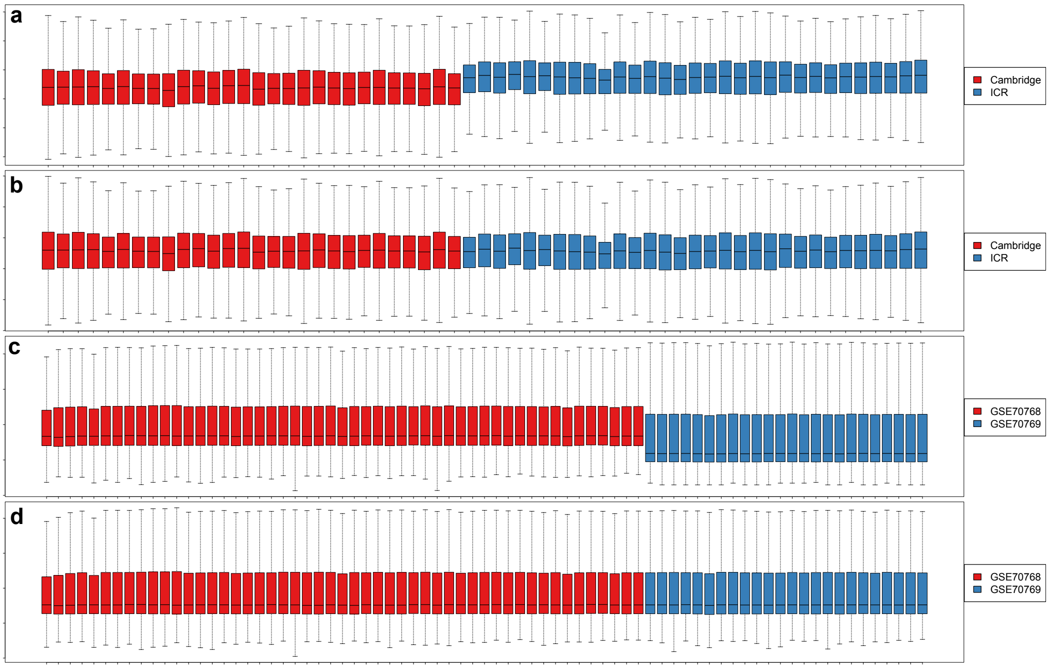

Figure 5.1 The effects of ComBat normalisation on the CancerMap and CamCap datasets: a) the CancerMap dataset before ComBat, b) the CancerMap dataset after ComBat, c) the CamCap dataset before ComBat, d) the CamCap dataset after Combat. Each boxplot corresponds to the distribution of intensities of genes in one sample. Due to limited space on page, only every fourth sample in each dataset has been plotted.

heterogeneous disease, and often several foci are present in the same sample [23], it is possible that several distinct processes are simultaneously present and are jointly contributing to the expression profile of a given sample.

LPD determines for each process an expression profile, that describes the expected expression level of each gene due to the process. Then, for a given sample it estimates how well the expression profile of each process is reflected in the expression levels of the genes in the sample. Alternatively, we say that LPD determines the contribution of each process to the expression profile of a sample.

The contributions, denoted γ_{ak} (gamma) in Section 3.2.1.3, where a is a samples and k is a process, are quantified as a number between 0 and 1, with 1 representing exclusive contribution of the process to the expression of a sample and 0 no contribution. For a given sample, the sum of gammas corresponding to all processes should sum to 1.

LPD can also be used to objectively estimate the number of processes present in a given set of samples, as we will illustrate below. Furthermore, the expression profiles of the processes can be analysed, in order to determine if different processes produce similar expression patterns, or to identify the characteristics of each process, such as the differentially expressed genes.

5.4.2 LPD parameters

LPD comes in two version, the maximum likelihood (MLE) model and the maximum posterior (MAP) model. As we described in Section 3.2.1.3.5, both models are necessary for choosing the LPD parameters. However, as we will show later, once the LPD parameters are chosen, the MAP version is more suitable for the final classification.

In short, the MAP version of LPD, used for the final classification, needs two parameters: the number of processes underlying the data and a parameter denoted *sigma* (Section 3.2.1.3.5). The selection of parameters is a three step process:

- Step 1: the number of processes underlying the data is estimated using the MLE model. The log-likelihood of the MLE model is calculated for various choices for the number of processes. In our case we tried every possible choice in the range 2-15 processes. The choices for the number of processes that produce the highest log-likelihoods are considered suitable;

- Step 2: suitable values for sigma are chosen using the MAP model. Sigma needs to be set to small negative values. Similar to Rogers et al. [209], we tried the following values: -0.01, -0.05, -0.1, -0.2, -0.3, -0.5, -0.75, -1, and -2. Using the number of processes estimated at step 1, the log-likelihood of the MAP model is calculated for each of the above values. The sigma values which produce the highest log-likelihoods were chosen;

- Step 3: the choices for the number of processes determined at step 1 are validated using the MAP model. More specifically, for each number of processes in the range 2-15 the log-likelihood of the MAP model is calculated. The number of processes at which the likelihood reaches a plateau is considered suitable.

LPD can give slightly different solutions on different restarts. This is due to the fact that some of the variables inside the LPD model are initialised with random values, that can lead to the convergence of the algorithm to different local maxima. For a robust choice of the parameters we restarted the LPD algorithm 100 times and calculated the mean log-likelihood.

5.4.3 Survival analyses

5.4.3.1 Univariate survival analysis

For the survival analyses we considered only primary tumour samples coming from unique patients, chosen as described in Section 5.3.2. For comparing the survival outcomes of two or more groups of patients we calculated a Kaplan-Meier (KM) survival curve for each group (Section 3.5.1). Then, we performed a log-rank test (Section 3.5.2), that assess the null hypothesis that all groups have similar KM survival curves, with the alternative hypothesis that at least one group has a different survival curve. For this analysis we used time to biochemical recurrence (BCR) as endpoint (Section 2.5.6.3).

5.4.3.2 Multivariate survival analysis

For multivariate survival analyses we used the Cox PH model (Section 3.5.3). We employed the multivariate analysis to assess if two groups of patients have significantly different failure times after adjusting for the effect of clinical predictors (Gleason grade at prostectomy, pathological stage and PSA level at diagnosis).

In order to apply the Cox model we stratified the Gleason score into $Gleason \leq 7$ and $Gleason > 7$. Probably a more suitable stratification would be to split the Gleason score into $Gleason \leq 3+4$ and $Gleason \leq 4+3$, as it seems to be a significant outcome difference between the two scores [92–95]. However, the Stephenson dataset, does not provide this information. Therefore, in order to keep the multivariate analysis constant across datasets, we chose 7 as the threshold for stratification. We also split the stage in $stage = T1/T2$ and $stage = T3/T4$ and PSA levels in $PSA \leq 10$ and $PSA > 10$.

5.4.4 Expression profile correlation

To investigate if two groups of samples have similar expression patterns for a given set of genes, we performed an expression profile correlation. For each group we calculate the mean expression level of each gene in the gene set, across the samples in the group. For simplicity, we will refer to the average expression levels of a set of genes in a group of samples as the expression profile of the group.

To determine if the groups have similar expressions patterns, we calculate the Pearson's correlation between the expression profiles. In order to give each gene the same contribution to the correlation, we select only one probeset for each gene. Most of genes have a single probeset anyway, but for those which have more than one probeset in a given dataset, we select a probeset at random. Before calculating the average expression of each group, we scale the expression levels of each gene to mean 0 and standard deviation 1 across all samples. This transformation was necessary to bring each gene to a comparable level.

5.4.5 Differential expression

To determine the differentially expressed genes between two groups of samples we used linear models implemented in the *limma* package [298], which is a commonly used method for differential expression analysis. In brief, *limma* fits a linear model for each gene, similarly to computing a *t*-test which verifies if the gene is differentially expressed between two conditions. The main difference is that *limma* uses some shrinkage methods, such as empirical Bayes, to borrow information across genes, and hence makes the results more robust.

The linear models produce a *p*-value for each gene, corresponding to the null hypothesis that the gene has similar expression levels in the two groups, with the alternative hypothesis that the gene is differentially expressed. The resulted *p*-values were adjusted for multiple testing using false discovery rate (FDR) at 1% level.

5.4.6 Pathway analysis

For the sets of genes differentially expressed between two groups we performed pathway analysis (Section 3.6), with the purpose of identifying biological pathways for which the component genes are over/under-represented in the set of genes.

For each gene set we performed an independent analysis using all pathways annotated in Gene Ontology (GO) [225] (from which we used the biological processes ontology), Kyoto Encyclopedia of Genes and Genomes (KEGG) [227] and Reactome [259]. The analyses have been performed using the *clusterProfiler* R package [260]. We adjusted the resulting p-values for multiple comparisons using the FDR method at a 5% level. We considered that a pathway is over/under-represented in a set of genes if its corresponding FDR adjusted p-value was less than 0.05.

5.4.7 Hierarchical clustering

We also performed a hierarchical clustering on our data to see if we could reproduce the LPD results. As described in Section 3.2.1.1, the hierarchical clustering needs a similarity measure that describes how the distance between two samples is calculated, and also a proximity measure, that specifies how the distance between two clusters is estimated.

We considered the distance between two samples as being $1 - corr$, where $corr$ is the Pearson's correlation. This similarity measure takes value 0 when the expression of two samples is perfectly correlated, and 2 when they are perfectly inversely correlated. As a proximity measure the complete link was used, which is the default proximity measure provided by R. The complete link, computes the distance between two clusters as the maximum distance between any point from one cluster to any point from the other cluster. To remove the variation across genes, in each dataset we scaled the expressions of every gene to mean 0 and variance 1 across samples.

5.4.8 Random forests

We describe the random forests algorithm in Section 3.2.2.1. In this analysis we used the random forests implementation from the *randomForest* R package [299].

The default random forest algorithm implemented in this packages tries to minimise the overall error rate. This approach leads to balanced class errors when the size of the two classes used for training is about the same. When one class is over-represented, however, the model becomes biased towards that class, resulting in higher classification errors in the smaller class. In our analysis we had to work with imbalanced classes. To correct for imbalances, we down-sampled the larger class.

The random forest models were trained using 10001 decision trees and the default mtry parameter (the square root of the total number of features).

5.4.9 LASSO

To perform feature selection, we used the LASSO logistic regression model described in Section 3.3. In brief, LASSO is a form of feature selection that, given a regression model, imposes restrictions on its coefficients, such that only the coefficients corresponding to the most informative variables for the classification are set to values different from 0. The coefficients corresponding to less useful or redundant variables are set to 0 and therefore the variables are removed from the analysis. In this analysis we used the *glmnet* R package [219].

As LASSO can give slightly different responses at different restarts, in order to obtain robust results we restarted the algorithm 100 times. At each iteration we performed a 10-fold cross validation to select the λ parameter (Section 3.3) which gives the lowest cross-validation error. The selected λ was then used to fit the LASSO and select the non-zero coefficients. For the classification we used genes that have been selected by the LASSO algorithm in at least 25 restarts.

5.4.10 Methylation analysis

We also performed a methylation analysis with the purpose of identifying differentially methylated regions. Differentially methylated regions (DMR) are essentially adjacent CpG sites that are differentially methylated between two conditions [300]. For this analysis we used an implementation available in the *methyAnalysis* R package [301].

The methylation analysis implemented in this package works with M-values, instead of beta-values, which are available in our data. An M-value is computed as the log-ratio between the methylated and unmethylated probe [297], while the beta-values are obtained by performing a logistic transformation of the M-values. Therefore, we could convert the beta-values back to M-values via a logit transformation, as described in Du et al. [297].

The *methyAnalysis* package considers only probes for which the difference in the average M-values in two groups is larger than 1. For all these probes, a t-test is performed and the resultant p-values are adjusted using the FDR correction.

A region is considered differentially methylated if the majority of probesets in the region are differentially methylated. In the default implementation of the *methyAnalysis* packages, two probes are considered part of the same DMR if the distance between them is less than 2,000 base-pairs.

5.5 LPD analysis

We performed an independent LPD analysis (Section 5.4.1), on each of the five microarray datasets, namely MSKCC, CancerMap, CamCap, Stephenson and Klein. We included in the analysis all the samples available in the dataset, including duplicates from the same patient. Therefore, we analysed all 320 samples in the MSKCC dataset, the 235 samples in CancerMap, 280 in CamCap, 89 in Stephenson and 182 in Klein.

5.5.1 Data preparation

Following the normalisations and annotations described in Section 5.3.3, we obtained for each dataset an expression matrix. The columns correspond to samples, and the rows correspond to probesets that estimate the expression level of a gene. The values in the matrix represent the normalised expressions for a probeset in a sample. For most genes there is only one probeset which measure its expression. However, some genes might have more than one probesets estimating its expression. Also the number of probesets (and the number of genes) varies depending on the microarray platform used for generating each datasets. The exon microarrays datasets (MSKCC, CancerMap and Klein) measure the expression of 17,868 probesets, mapping to 17,320 unique gene symbols. For the CamCap dataset there are 34,476 probesets mapping to 19,412 symbols and Stephenson contains 22,283 probesets mapping to 12,500 gene symbols.

Genes that have little variance across samples are of little interest for classification as they usually do not have discriminative power. Moreover, although LPD scales linearly with the number of genes and number of samples [209], for large datasets, such as MSKCC, which contains 320 samples, using a large number of genes is computationally prohibitive.

Previous applications of LPD on cancer datasets used the 500 genes which exhibit the highest variance across samples [209, 302]. The LPD classification led to good results in each case, suggesting that, in general, around 500 genes should contain enough information for the model to work well.

In line with previous LPD analyses, we selected the top 500 probesets that exhibit the highest variance across the samples in the MSKCC dataset. To keep the analysis consistent and comparable between datasets, we tried to use a similar set of probesets for the other datasets. However, as our datasets have been created using several microarray and RNA-seq platforms, and different platforms provide different probes, we could not find a direct mapping between the probesets across platform. We, therefore, determined the corresponding probesets by checking if they map to the same gene symbol.

As for some platforms some genes might have more than one corresponding probesets and for some genes there might not be any probe, the number of probes selected for LPD is slightly different across the datasets. For the exon microarrays (MSKCC, CancerMap and Klein), the top 500 probesets are mapping to 489 unique gene symbols (Supplementary Table B.1). When mapping back the 489 gene symbols to the probesets provided by exon microarray platform we obtain 507 probesets, as there are 7 extra probesets that are not in the top 500 probeset by variance, but map to the same gene symbol as some of the probesets in the top 500. For the CamCap, we identified 483 probes mapping to the reference set of genes and for Stephenson there are 609 probesets.

5.5.2 Choosing LPD parameters

We derived the LPD parameters in three steps, as described in Section 5.4.2. In the first step, for each of the five datasets (MSKCC, CancerMap, Stephenson, CamCap and Klein), for each possible number of processes we calculated the log-likelihoods corresponding to 100 restarts of the MLE model. We represent the resulting average MLE log-likelihoods as a function of the number of processes in Figure 5.2 (the red curves). We note that for the MSKCC the processes in the range 6-10 have similar likelihoods, suggesting that any choice would be satisfactory. After 10 processes, the likelihood begins to decrease, giving an indication that the model starts to over-fit the data. We set 8 as the number of processes, as this seems to be the peak value. Similarly, for CancerMap we select 8 processes, for Stephenson we choose 3 processes, for Klein 5 processes and for CamCap 6.

In the second step, we tried to determine suitable values for sigma. For each choice of sigma we run the LPD algorithm 100 times with different seeds and calculated the average log-likelihood. The average log-likelihoods of the MAP model as a function of sigma are presented in Supplementary Figure B.3. For MSKCC we chose the peak, -0.5. Similarly, for CancerMap we selected -0.5, for Stephenson -0.75, for Klein -0.3 and for CamCap -0.05.

Then, we performed the third step, which consists in evaluating the log-likelihood of the MAP model for various number of processes, using as input the values for sigma determined at step two (the blue curves in Figure 5.2). We observe that when choosing small number of processes MAP curves and MLE curves are at lower, similar, levels, suggesting that both models underfit the data. As we increase the number of processes, curves raise, with MAP curve increasing at a faster rate. As the approximative number of processes inherent in the data is reached, the MLE likelihoods start to decrease, while the MAP likelihoods tend to slow down the increase and most of the times reach

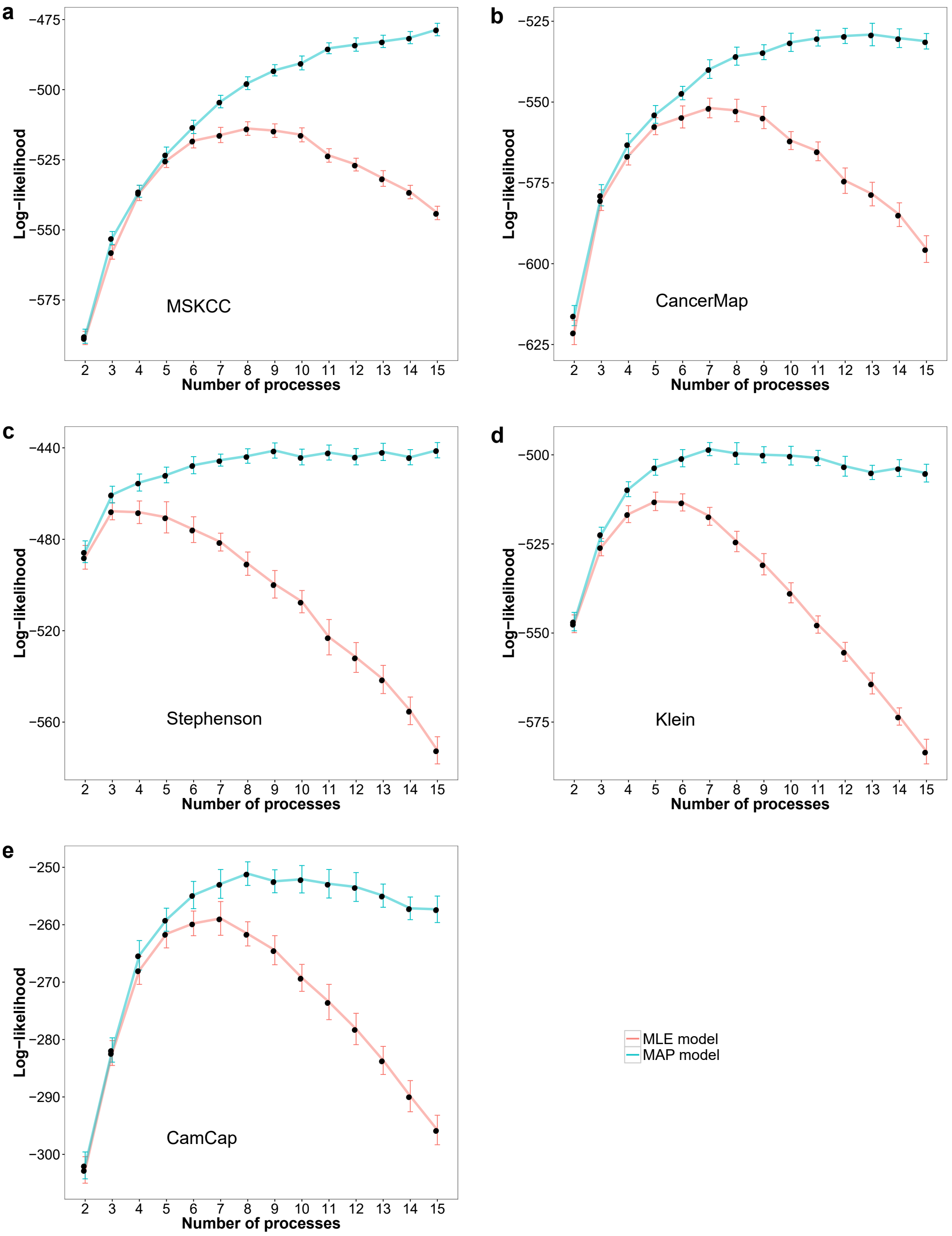

Figure 5.2 The log-likelihood (vertical axis) versus number of processes (horizontal axis) using the MLE solution (lower curve) and the MAP solution (upper curve) for each dataset. The points represent the average likelihood of 100 LPD restarts. For each point we also plot the error bars corresponding to the distribution of the likelihoods obtained in the LPD restarts. For the MLE model the peak in likelihood indicates the number of processes to use. For the MAP model the likelihood rises to a plateau after which no further gain is to be made, indicating the maximum number of processes that should be used.

a plateau. In every case the MAP curves are above the MLE curves, suggesting that the MAP model is fitting the data better and, therefore, is more suitable for the final classification.

It is still not clear to us why for the MSKCC and CancerMap dataset the MAP likelihood do not reach a plateau even after 15 processes, given that the corresponding MLE estimates suggest up to 10 processes, but more likely 7-8. Given this behaviour, we decided to continue the analysis with 8 processes, as predicted by the MLE curve. The MAP curve in the Stephenson elbows at 3 processes, which is also the number of processes we determined using the MLE curves, suggesting a good choice of the parameter. For Klein and CamCap it is less obvious the exact number of processes at which the MAP curve reaches a plateau. We considered that the previous choices, of 5 and respectively 6 processes are consistent with the MAP curves, and therefore there is no reason for updating them.

To summarise the above discussion, for each dataset we selected two sets of parameters to be used for the classification of samples described in the next section. For MSKCC and CancerMap we set the number of processes to 8 and the value of sigma to -0.5, for Stephenson we selected 3 processes and sigma -0.75, for Klein 5 processes and sigma -0.3, while for CamCap we decided to use 6 processes and -0.05 for sigma.

5.5.3 LPD classification

We employed the LPD algorithm to produce an unsupervised classification of the samples in each of the five datasets (MSKCC, CancerMap, CamCap, Stephenson and Klein). As LPD is not a deterministic algorithm, i.e. for the same settings for the parameters and input data it can give distinct results at different restarts of the algorithm, we repeated the LPD analysis 100 times for each dataset.

In Figure 5.3 we illustrate the results obtained for one of the 100 LPD runs on the MSKCC dataset. We note in particular the LPD1 process, enclosed in the red box. This process has high contributions to the expressions profile of a large group of high risk samples (the bulk of mainly green samples on the left). This indicates an association between the LPD1 process and poor outcome, which we will explore in more detail later. For the moment we just note that if we assign each sample to the process with the highest contribution to its expression profile, the high risk samples are over-represented in the LPD1 process (χ^2 p-value $4.12 \cdot 10^{-6}$).

We also note that the LPD7 process has high contributions in a group of benign samples, suggesting a non-malignant nature for this process. The χ^2 test, assessing if the benign samples are over-represented in the LPD7 process, is highly significant (p-value $3.71 \cdot 10^{-13}$).

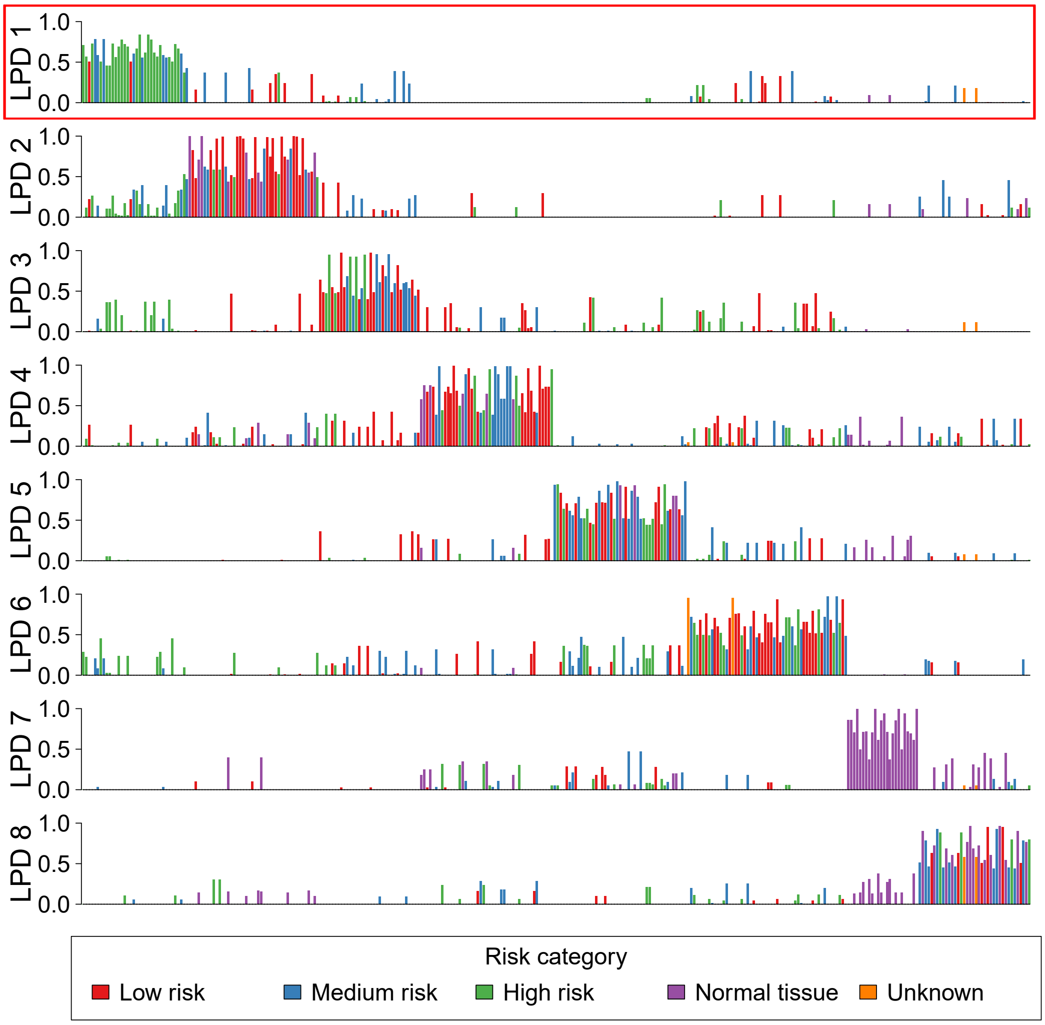

Figure 5.3 An illustration of the LPD classification on the MSKCC dataset. Each horizontal panel, denoted LPD1 to LPD8, correspond to one of the 8 LPD processes. For each panel, the x-axis represent samples, while the y-axis represents the contribution of the process to the expression of each sample. The colours correspond to the ICGC risk categories defined in Section 2.5.5.5, which indicate the risk of recurrence following prostectomy by taking into account several clinical indicators.

In CancerMap we have been able to identify a process with high contributions to the expressions of a group of mainly high and medium risk samples (the LPD5 process in Supplementary Figure B.4). The high and medium risk samples are over-represented in the group of samples to which the LPD5 has highest contributions (χ^2 p-value $5.81 \cdot 10^{-3}$).

In CamCap we have also found a process with high contributions to the high risk samples (LPD6 in Supplementary Figure B.5, χ^2 p-value $1.8 \cdot 10^{-2}$). Moreover for CamCap there are two processes, LPD3 and LPD4, who contribute to the expression of mainly benign samples. The p-values of the χ^2 test assessing if the normal samples are over-represented in these two process are $7.96 \cdot 10^{-12}$ and $4.06 \cdot 10^{-10}$ respectively.

For the Stephenson dataset the clinical data does not contain some of the clinical indicators necessary for the estimation of the ICGC risk category and therefore the pathological stage is used as proxy. We note that the LPD2 group does not have high contributions in any of the benign tissue samples, suggesting a possible malignant characteristic of the process (Supplementary Figure B.6).

For the Klein dataset (Supplementary Figure B.7), the clinical associations are not available. However, we will illustrate later how some of these processes relate to the processes in the other datasets.

5.6 Survival analyses

The LPD results suggest that in each dataset with associated clinical data some processes might contribute to poorer outcomes. To further study these assumptions we performed survival analyses. However, the survival analysis requires each sample to be exclusively assigned to a group. LPD, on the other hand, produces a probabilistic association of a sample to several processes. To convert the probabilistic to exclusive association, we assigned each sample to the process with the highest contribution to its expression profile.

For each of the four datasets with linked clinical data (MSKCC, CancerMap, CamCap and Stephenson) we performed both univariate and multivariate survival analyses, as described in Section 5.4.3, using BCR as endpoint. Our main aim was to determine if the membership of samples to certain LPD processes is a significant predictor for the time to BCR and if the membership to an LPD group is an independent predictor for the time to BCR, after correcting for the effects of other covariates, such as the Gleason grade, PSA and pathological stage.

5.6.1 Univariate survival analysis

For each of the 100 LPD restarts in each dataset, we calculated the Kaplan-Meier (KM) survival curves. Then, to assess if the KM survival curves are statistically different, we performed a log-rank test (Section 5.4.3.1).

5.6.1.1 Choosing representative runs

Some samples can change LPD membership in different runs, which can lead to slightly different survival curves and log-rank estimates. This is illustrated in Figure 5.4. A way of dealing with this variability is to select an average LPD run, which does not produce extreme log-rank p-values and to consider it as representative for a given dataset. We decided to plot the distribution of the log-rank p-values across runs (Supplementary Figure B.8), and selected the run for which the p-value is closest to the mode of the distributions, depicted with dashed vertical lines. For the Klein dataset, as there is no clinical data, we selected at random one of the 100 runs. The LPD plots corresponding to the representative runs in each dataset in Figure 5.5.

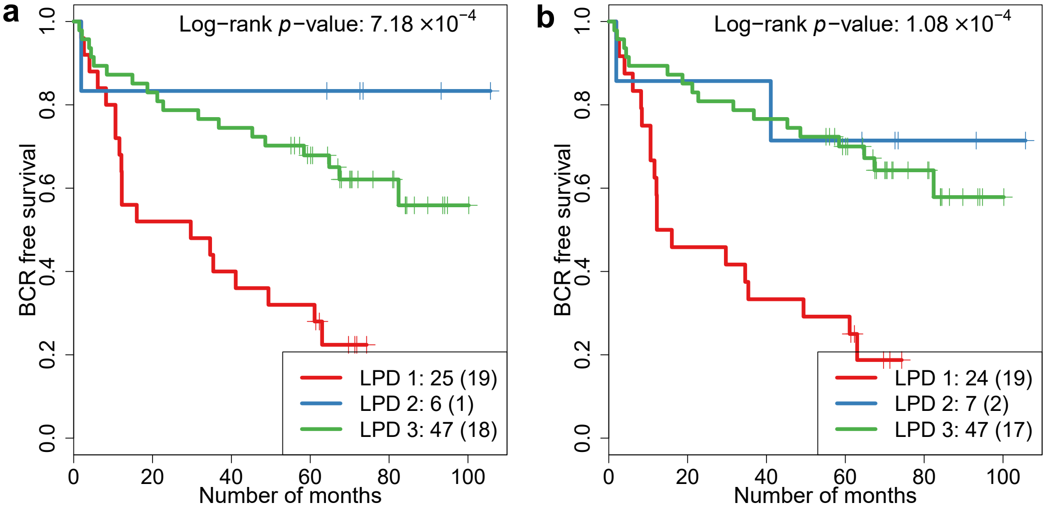

Figure 5.4 KM plots obtained at different LPD runs on the Stephenson dataset. The number of cancers in each group is indicated in the bottom right corner of each Kaplan-Meier plot. The number of patients with PSA failure is indicated in parentheses. Between two distinct LPD runs, one sample changed membership from LPD1 group, to LPD2 and, also, LPD1 and LPD3 exchanged one sample (the number of failures in the LPD3 changed, even if the number of samples in the group remained the same). This lead to a decrease of the log-rank p-value from $7.18 \cdot 10^{-4}$ to $1.08 \cdot 10^{-4}$.

5.6.1.2 Univariate survival analysis results

For each dataset, the log-rank p-values corresponding to the representative run are quite low ($4.88 \cdot 10^{-3}$ for MSKCC, $1.57 \cdot 10^{-5}$ for CancerMap, $1.75 \cdot 10^{-4}$ for Stephenson and $6.27 \cdot 10^{-3}$ for CamCap), suggesting that, for all datasets, the LPD groups have statistically different BCR failure outcomes (Figure 5.6).

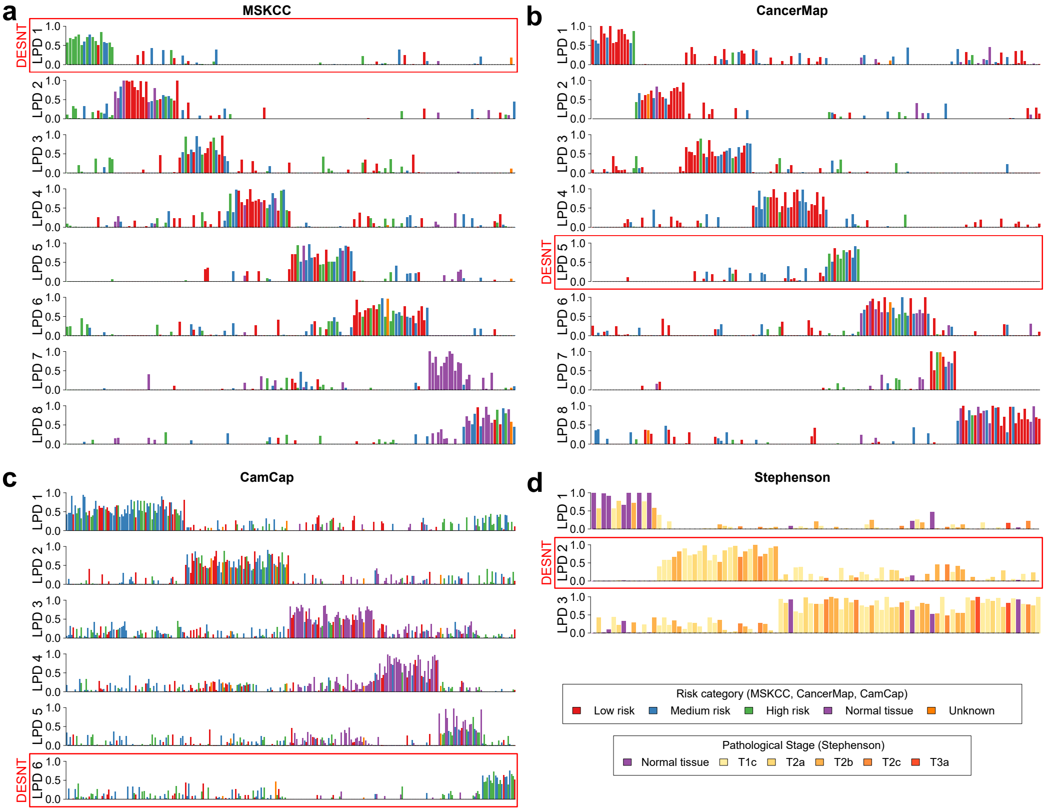

Figure 5.5 The LPD classification on the the representative runs. Each horizontal panel corresponds to a LPD process. For each panel, the *x*-axis represent samples, while the *y*-axis represents the contribution of the process to the expression of the sample. The colours correspond to the ICGC risk categories defined in Section 2.5.5.5. Duplicated samples from from the same patient were removed, as described in Section 5.3.2.

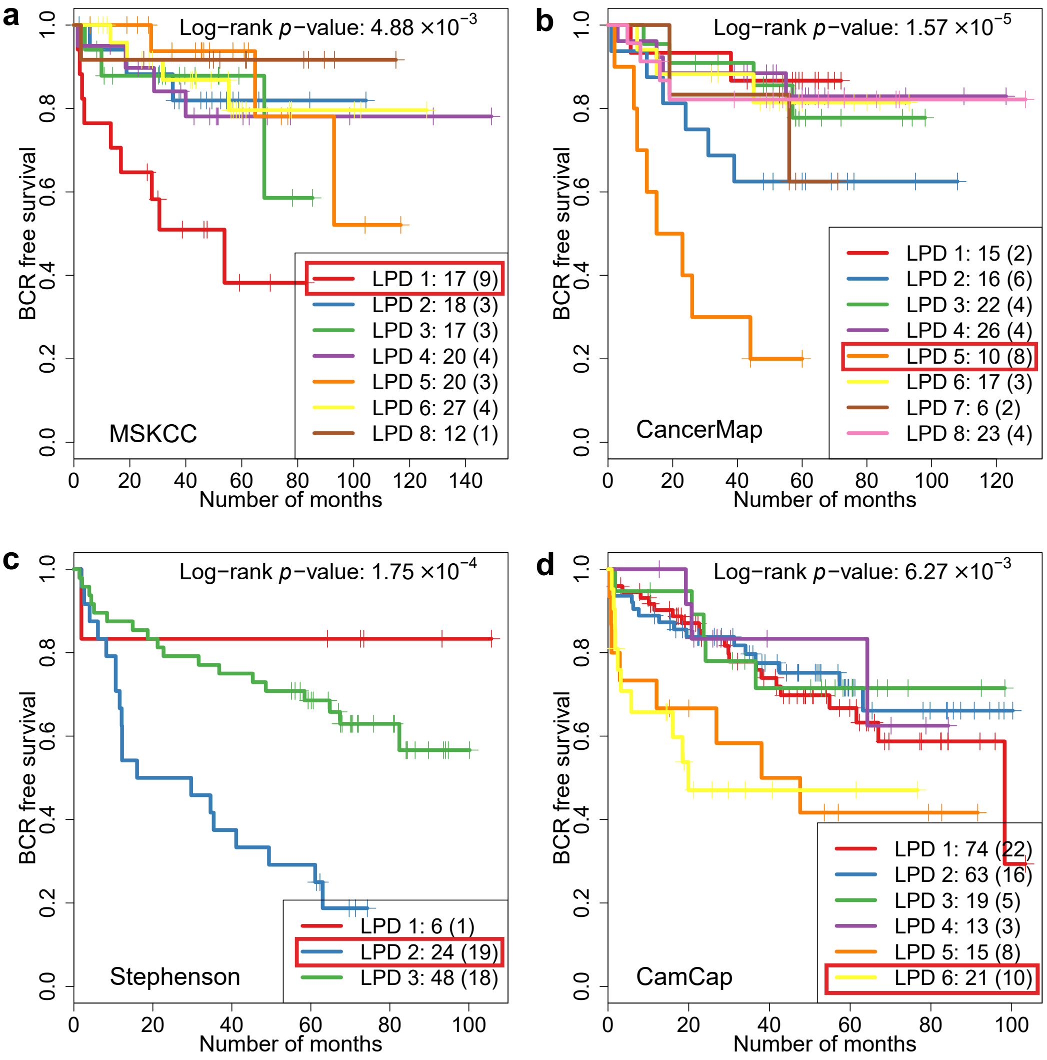

Figure 5.6 KM plots for the representative runs in MSKCC, CancerMap, Stephenson and CamCap datasets. The number of cancers in each group is indicated in the bottom right corner of each Kaplan-Meier plot. The number of patients with PSA failure is indicated in parentheses. The processes surrounded by red boxes are the DESNT processes.

For the LPD1 group in the MSKCC dataset (Figure 5.5a), for which most of the samples are in the ICGC high risk category, the corresponding survival curve (the red curve in Figure 5.6a) is lower than all other survival curves. Similarly, the LPD5 process in Figure 5.5b, corresponding to a group of samples with intermediate and high ICGC risk categories in the CancerMap dataset seems to have also an worse outcome (the orange curve in Figure 5.6b). Also, for the Stephenson dataset, the LPD2 processes, which is not contributing in high proportions to any normal samples (Figure 5.5d), has a low survival curve, as depicted in Figure 5.6c - the blue curve. And finally, one of the two poor outcome groups in the CamCap dataset (the yellow curve in Figure 5.6d), is the LPD6 process is Figure 5.5d, which contribute mainly to the expression profile of high risk samples. For convenience, in each dataset, we will refer to the group that is associated with higher risk category and poorer BCR free survival as the *DESNT* group.

So far we have determined that the DESNT groups seem to have a worse clinical outcome. However, the log-rank test allows us to test the null hypothesis that the survival curves of all groups are the same, with the alternative hypothesis that at least one group has a different survival curve compared to the other groups.

To determine if indeed the DENT cancers have a significantly worse BCR prognosis, for each dataset we merged all the non-DESNT groups into one group and recalculated the survival curves and the log-rank p-values for only two groups, DESNT and non-DESNT (Figure 5.7). For the MSKCC dataset the log-rank p-value testing the null hypothesis that the DESNT and non-DESNT groups have similar survival curves is $2.65 \cdot 10^{-5}$, indicating a statistically significant worse prognosis for the DESNT patients. In line with this result, the log-rank test for all the other datasets, suggest statistically significant BCR failure prognosis for the DESNT cancers (log-rank p-value $2.98 \cdot 10^{-8}$ for CancerMap, $4.28 \cdot 10^{-5}$ for Stephenson, and $1.22 \cdot 10^{-3}$ for CamCap).

5.6.2 Multivariate survival analysis

Having determined that the DESNT group is a significant predictor of BCR, we tried also to assess the hypothesis that it is also an independent predictor of recurrence. More specifically, we tried to verify if after adjusting for the effect of other clinical factors, the DESNT membership continues to be a statistically significant predictor of recurrence. For each dataset we performed a multivariate survival analysis using the Cox PH model (Section 5.4.3.2).

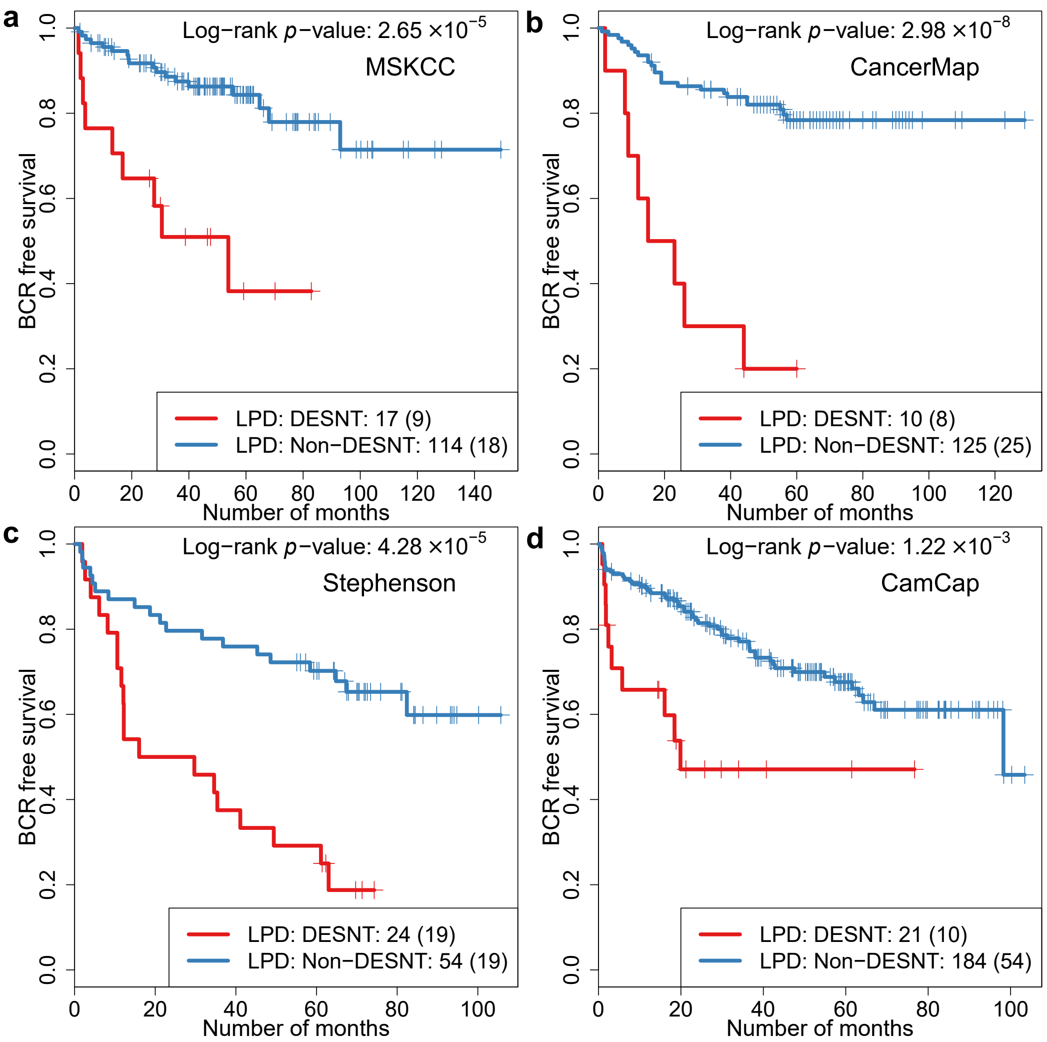

Figure 5.7 KM plots for the DESNT and non-DESNT groups in the representative runs in MSKCC, CancerMap, Stephenson and CamCap. The number of cancers in each group is indicated in the bottom right corner of each Kaplan-Meier plot. The number of patients with PSA failure is indicated in parentheses.

5.6.2.1 Cox PH assumptions

As described in Section 3.5.3.1, the Cox model is a non-parametric test, as it does not make any assumption about the shape of the survival curves. However, it depends on a very important assumption, namely that the hazard ratio is constant over time (the PH assumption).

For each dataset, we evaluated the PH assumption for every covariate used in the Cox PH model. The assumption seems to be respected for all cases, excepting for the pathological stage in the MSKCC dataset (Figure 5.8). Starting with month 35, the observed survival curve of the patients with Pathological Stage T3/T4 in MSKCC reaches a plateau. After this time none of the patients with stage T3/T4 experiences failure. The expected curve, on the other hand predicts a constant decrease of in the survival odds after this time, in line with the trend before 35 weeks. This suggests that in this case the stage is not a time-independent predictor, but rather has different behaviours for different time intervals. This observation is supported by the Figure 5.8b, as the log-log curves come closer, instead of remaining parallel. Moreover, the Schoenfeld p-value is below 0.03, bringing statistical evidence that the PH assumption is not respected in this case.

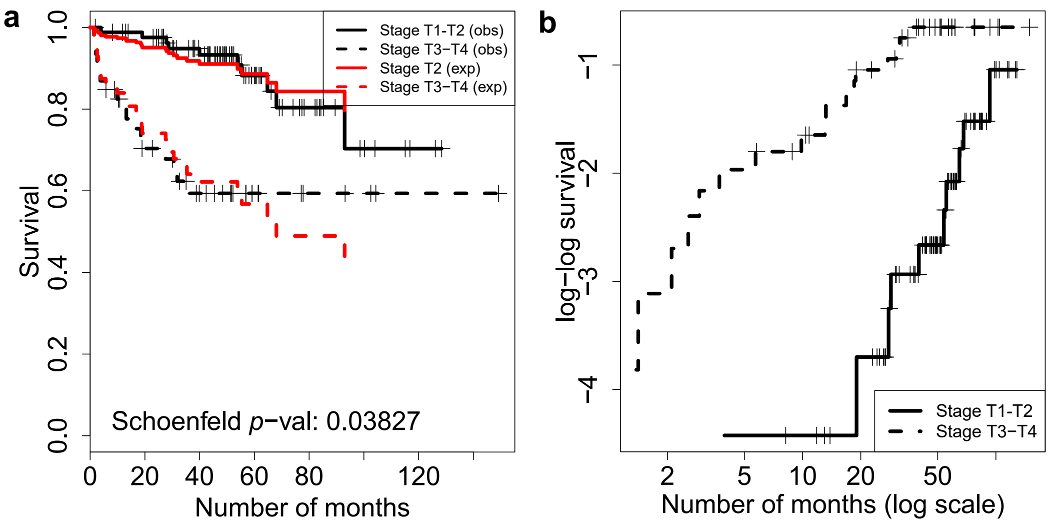

Figure 5.8 Evaluation of the PH assumption for the pathological stage variable in the MSKCC dataset using: a) log-log survival curves corresponding to stage T1/T2 and stage T3/T4.; b) the observed vs. expected survival curves corresponding to stage T1/T2 and stage T3/T4. Note that the log-log survival curves converge and also the lower observed curve departs from the expected curve.

We solved this issue, by modelling the stage in the MSKCC as a time-dependent covariate, which was then incorporated together with the other covariates into an extended Cox model. The pathological stage was split into two time intervals, $0-35$ months and > 35 months, for which the Cox model calculated different hazard ratios.

5.6.2.2 Multivariate survival analysis results

For each dataset, the extended Cox PH models was constructed using DESNT membership, the discretised Gleason score ($\leq / > 7$), the PSA ($\leq / > 10$) and the stage (T1-T2/T3-T4) (Figure 5.9, Supplementary Table B.2). For each variable the hazard ratio gives the odds of experiencing faster failure for individuals from one category, relative to the baseline category, after adjusting for other covariates. Values significantly greater than 1 suggest positive association of the category with the time to failure and values significantly lower than 1 indicate a negative association.

For example in MSKCC, the Gleason score >7 has a hazard ratio of 5.1 (95% CI 1.9-13.9). The HR is significantly greater than 1 (p-value $1.12 \cdot 10^{-3}$), suggesting a strong association of high Gleason grade with the odds of experiencing faster failure, after adjusting for the effect of other covariates. Also the HR for the T3-T4 stage is a significant independent predictor of recurrence in the first 35 months (HR 5.5, 95% CI 1.7-17.6, p-value $3.77 \cdot 10^{-3}$), but not after 35 months (HR 0.5, 95% CI 0.06-4.531, p-value $5.59 \cdot 10^{-1}$). For the MSKCC dataset the DESNT membership is not a significant predictor, after adjusting for the effect of other covariates. Even though the HR is 1.67, due to relatively low number of samples in the DESNT category (17/131 samples), the 95% CI is wide (0.59-4.65) and therefore the p-value is not significant (0.327).

For the other three datasets the DESNT membership is a significant independent predictor of recurrence (CancerMap: HR 4.29, 95% CI 1.6-11.4, p-value $3.66 \cdot 10^{-3}$; Stephenson: HR 3.80, 95% CI 1.88-7.66, p-value $1.83 \cdot 10^{-4}$; CamCap: HR 2.25, 95% CI 1.08-4.66).

As the DESNT group is a relatively small group and, as in the case of MSKCC, the statistical analysis produces quite large confidence intervals, we performed a meta-analysis across multiple datasets. The purpose was to assess if, overall, the DESNT membership is an independent predictor of BCR.

We therefore merged the covariates used for the Cox model (DESNT membership, Gleason, PSA and stage) from three datasets (MSKCC, CancerMap and Stephenson) to obtain a larger set of samples. The CamCap dataset was not included because it shares many samples with CancerMap and therefore it is not an independent dataset. This lead us to obtaining a set of 344 unique samples, out of which 51 are in the DESNT group (17 from MSKCC, 10 from CancerMap and 24 from Stephenson), on which we performed again the multivariate analysis. In this multivariate analysis, we included

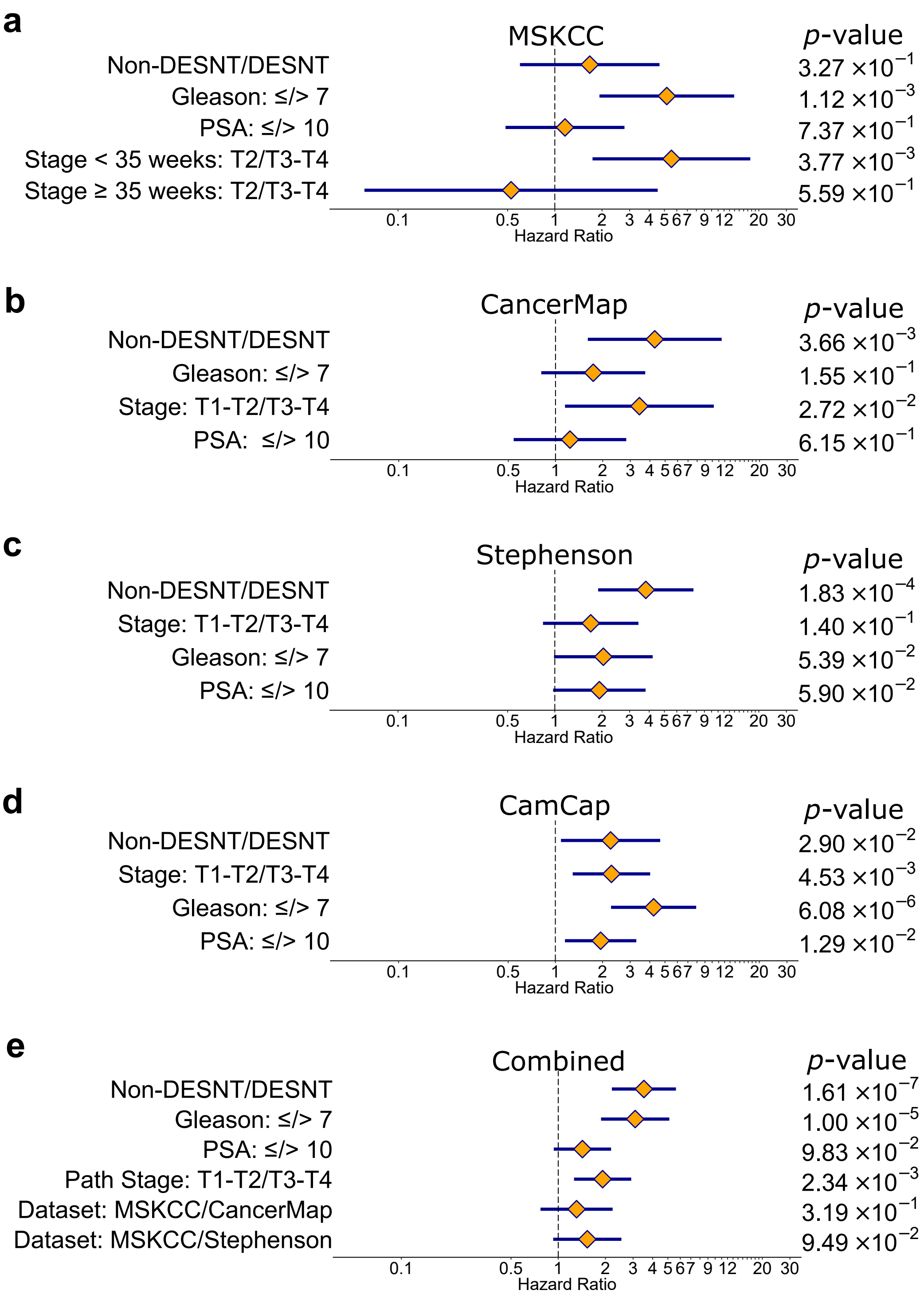

Figure 5.9 The Cox hazard ratios. The positions of the orange points on the x-axis corresponds to the Cox hazard ratios for each covariate. The blue lines represent the confidence intervals.

an additional covariate, the dataset from which each sample comes. This variable was introduced with the purpose of adjusting for any dataset-specific imbalances in the covariates. For all the covariates included in this analysis the Cox PH assumption was valid.

The DESNT membership is the highest significant predictor of BCR recurrence, after adjusting for the effect of other covariates (HR 3.51, 95% CI 2.19-5.62, p-value $1.61 \cdot 10^{-7}$) (Figure 5.9e). It outperforms the other two significant predictors, the Gleason score (HR 3.09, 95% CI 1.87-5.10, p-value $1 \cdot 10^{-5}$) and the stage (HR 1.91, 95% CI 1.26-2.91, p-value $2.34 \cdot 10^{-3}$).

These result suggest that the DESNT membership is a recurrence predictor, independent of the Gleason grade, PSA and the pathological stage.

5.7 Correlations between process expression profiles

To verify that the DESNT processes from different datasets are related to each other, we investigated if their expression profiles are correlated, as described in Section 5.4.4. A correlation of the expression profiles of two DESNT groups from two different datasets would indicate a common DESNT process in the two datasets.

As presented in Figure 5.10, the correlations are high between every possible pair of DESNT groups. Moreover we found that the LPD5 group in the Klein dataset (Supplementary Figure B.7) is highly correlated with all the other DESNT groups (Pearson's correlation 0.595 - 0.753). We considered the LPD5 Klein group a DESNT group as well.

We illustrate in Supplementary Figure B.9a-c the correlation of the expression profile of the DESNT group in MSKCC with: a) the expression profile of the LPD7 group in MSKCC (Figure 5.5a), which contains only benign samples, b) the LPD4 process in CamCap (Figure 5.5c), containing mainly normal samples and c) the LPD1 process in Stephenson (Figure 5.5d), which contains most of the normal samples. As all three processes contain benign samples, we assume a non-aggressive nature for them. We note that DESNT process is inversely correlated with all these processes (Pearson's correlation -0.64, -0.62 and -0.62). Also, for control, we present in Supplementary Figure B.9d-f, several examples of correlations of the expression profile of the MSKCC DESNT group with the expression profiles of various other LPD groups, which contain heterogeneous risk samples (Pearson's correlation < 0.25).

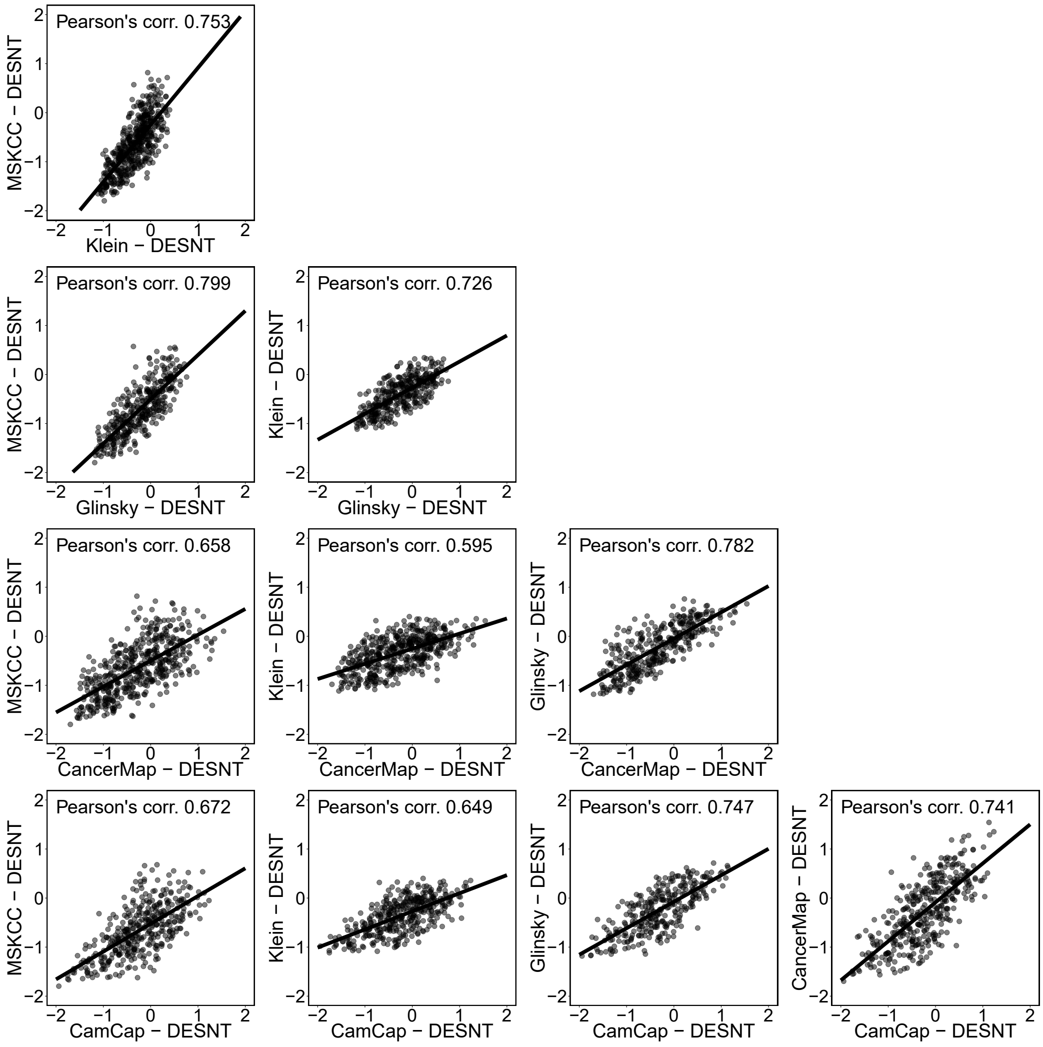

Figure 5.10 Correlations of expression profiles between cancers assigned to the DESNT process in each of the datasets MSKCC, CancerMap, Stephenson and Klein. Data from the 500 probes used in LPD are represented and ten possible comparisons are shown. The expression levels of each gene have been normalised across all samples to mean 0 and standard deviation 1.

5.8 Differentially expressed genes

Here we identify a set of genes that are constantly differentially expressed in DESNT relative to non-DESNT cancers, as described in Section 5.4.5. For this analysis we considered all probesets available, as it is possible that genes which have not been included in the list of 500 genes used for LPD training to be discriminative for the DESNT group as well.

For a more robust analysis, we used all 100 LPD restarts for each dataset. For each of the 100 LPD runs, we identified a list of genes differentially expressed in DESNT. We then selected only the genes that have been identified as differentially expressed in at least 80 of the 100 LPD runs.

6,395 differentially expressed genes were identified in MSKCC, 1,062 in CancerMap, 195 in Stephenson, 1,270 in Klein and 644 in CamCap. We intersected the resulting lists of genes corresponding to MSKCC, CancerMap, Stephenson and Klein (Figure 5.11). Again, we did not also include CamCap in the analysis as it is not independent to CancerMap. The intersection resulted in 45 genes in common between MSKCC, CancerMap, Stephenson and Klein datasets. We will further refer to this set of genes as the LPD DESNT signature.

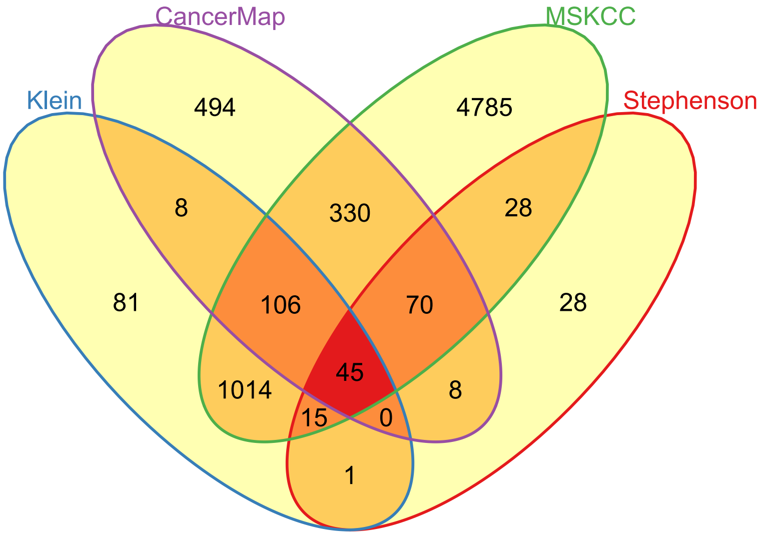

Figure 5.11 Venn diagram illustrating the intersection of differentially expressed genes in the MSKCC, CancerMap, Klein and Glinsky datasets.

All 45 genes are under-expressed in the DESNT group and at least 16 genes have been previously reported as methylated or down-regulated in prostate cancer or other cancers (Supplementary Table B.3). Many of these genes have also been linked to development and progression of various types of cancer. For example, *FBLN1* is a gene hyper-methylated in many type of cancer, including bladder, colorectal, cutaneous and tongue carcinoma [303, 304]. We also note that 38/45 of the identified genes are also differentially expressed in the CamCap, which has not been used for deriving the signature (Supplementary Table B.3).

5.9 Pathway analysis

We wished to identify a list of biological pathways for which the component genes are significantly under/over-represented in the set of 45 genes from the LPD DESNT signature. The identification of pathways significantly associated with the 45 genes signature, might unravel the biological mechanisms behind the DESNT process. We performed the pathway analysis as described in Section 5.4.6, using the Gene Ontology (GO) [225] (from which we used the biological processes ontology), Kyoto Encyclopedia of Genes and Genomes (KEGG) [227] and Reactome [259].

We identified over 200 GO biological processes over-represented in the LPD DESNT signature. The top 20 are presented in Supplementary Table B.4. For the KEGG database we found nine over-represented pathways (Supplementary Table B.5) and for Reactome we identified nine pathways as well (Supplementary Table B.6).

For the KEGG database the top five pathways over-represented in the set of 45 genes are the muscle contraction pathway, focal adhesion, adherens junction, regulation of actin cytoskeleton and leukocyte transendothelial migration pathways (Figure 5.12a). In Reactome muscle contraction pathway is again one of the top five pathways, together with RHO GTPases activate PAKs, cell-extracellular matrix interactions and cell junction organization (Figure 5.12b). In the GO database we also identified muscle contraction pathway, together with wound healing and anatomical structure morphogenesis (Figure 5.12c).

We are very grateful to Prof. Dylan Edwards, from the School of Biological Sciences, UEA, who performed an independent assessment of the possible molecular functions of the 45 genes in the LPD DESNT signature. He identified that many of the proteins encoded by these 45 genes are components of the cytoskeleton or regulate its dynamics, while others are involved in focal adhesion and ion transport (Supplementary Table B.7). Also, he provided a brief description of the possible role of these genes in the progression of prostate cancer, which we reproduce here:

"*Several signature genes encode proteins that are components of the actin cytoskeleton or which regulate its dynamics, including ACTA2, ACTG2, ACTN1, CNN, FLNA, ILK, ITGA5, LMOD1, MYLK, PALLD, VCL, CALD1, CDC42EP3, PDLIM1, SVIL,*

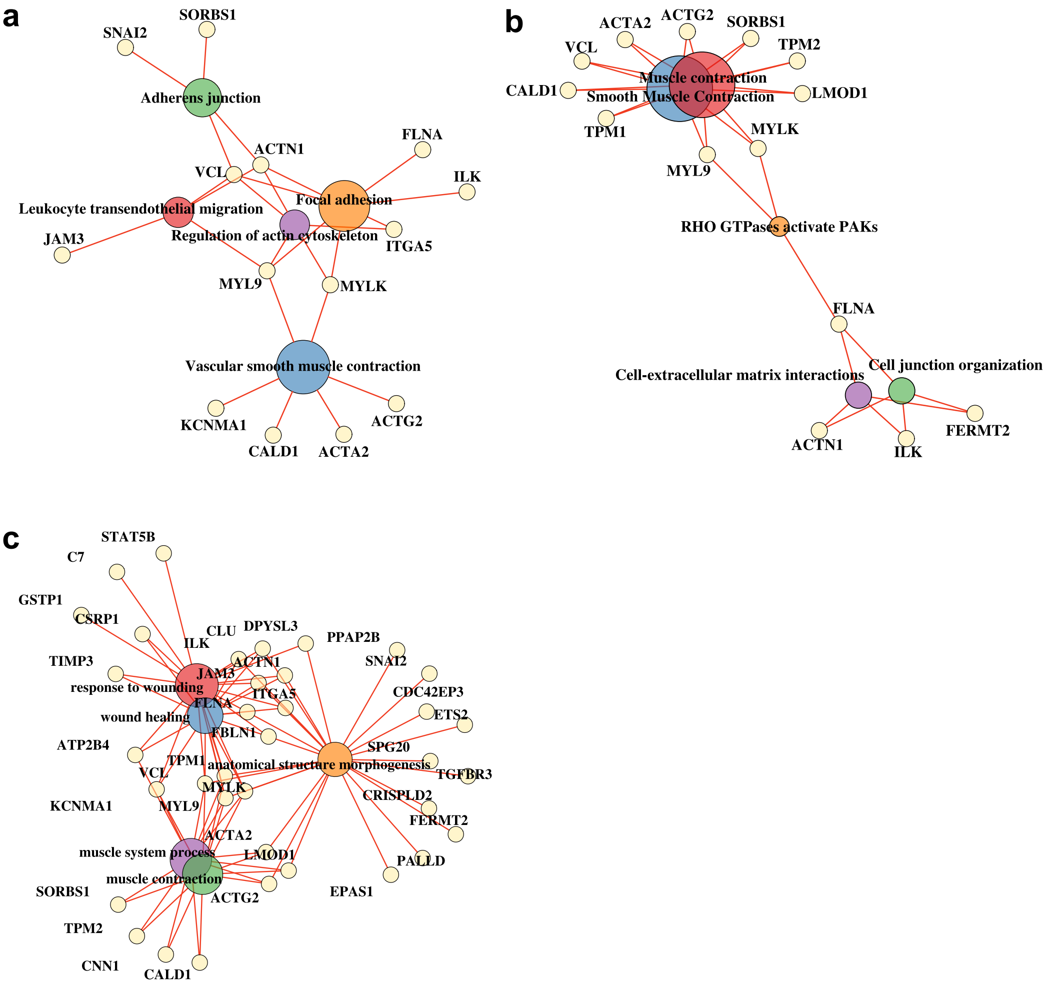

Figure 5.12 Cnet plots depicting the results of the pathway analysis on a) KEGG b) Reactome and c) GO databases. The multicolour circles correspond to the top five pathways, with the smallest *p*-values. The size of the circles is inversely proportional to the *p*-value. The small yellow circles represent genes. The edges between genes and pathways denote the involvement of gene in pathway.

TNS1, TPM1, TPM2. In particular, actomyosin contractility is highlighted by the presence of myosin light chain kinase (MLCK) and myosin light chain-9 (MYL9) and other molecules such as α-actinin (ACTN1), tensin (TNS1) and calponin (CNN1). Increased malignancy may correlate with increased cell migratory behaviour, which in turn may reflect the deployment of particular types of cell adhesion and cytoskeletal machinery. A high dependency on actomyosin contractility is recognised as a hallmark of amoeboid movement [305], and since this aspect is down-regulated in the poor prognosis signature, it would seem less likely to be the mode of migration employed.

However, also noteworthy are important focal adhesion components such as integrin α5 (ITGA5), vinculin (VCL) and integrin-linked kinase (ILK), which would be expected to be involved in mesenchymal type migration. It is thus possible that the gene signature favours a collective migration phenotype, typified by maintenance of E-cadherin mediated cell-cell adhesion mechanisms [306].

There are several signature genes (eg. ACTA2, CNN1, LMOD1) that encode proteins primarily expressed in smooth muscle cells or myofibroblasts, which is an indication of an altered tumour-stromal environment.

In the attached table (i.e. Supplementary Table B.7) *I have also highlighted genes that are important as ion channels (important in intracellular Ca homeostasis, which in turn will affect actomyosin contractility). Also too there are a few transcription factors and an RNA binding protein that will affect translation, thus there could be diverse downstream changes in genetic programmes as a result of the down-regulation of these genes. However, it is hard to predict the consequences here.*"

5.10 Intersection of LPD DESNT genes with published signatures

In this section we examined whether any of our 45 genes have been previously included in other prognostic signatures for prostate cancer.

We collected the signatures published in previous work on prostate cancer, namely Long et al. [307], Glinsky et al. [308], Planche et al. [309], Bismar et al. [310], Cuzick et al. [38], Ramaswamy et al. [311], Agell et al. [312], Bibikova et al. [313], Ross-Adams et al. [241], Wu et al. [314], Singh et al. [315], Rajan et al. [316], Erho et al. [40], Irshad et al. [317], Ramos-Montoya et al. [283], Sharma et al. [318], Knezevic et al. [175], Lalonde et al. [319], Yu et al. [320], Varambally et al. [282] and You et al. [284]. We then determined the genes in common between every pair of signatures, including the LPD DESNT signature (Figure 5.13).

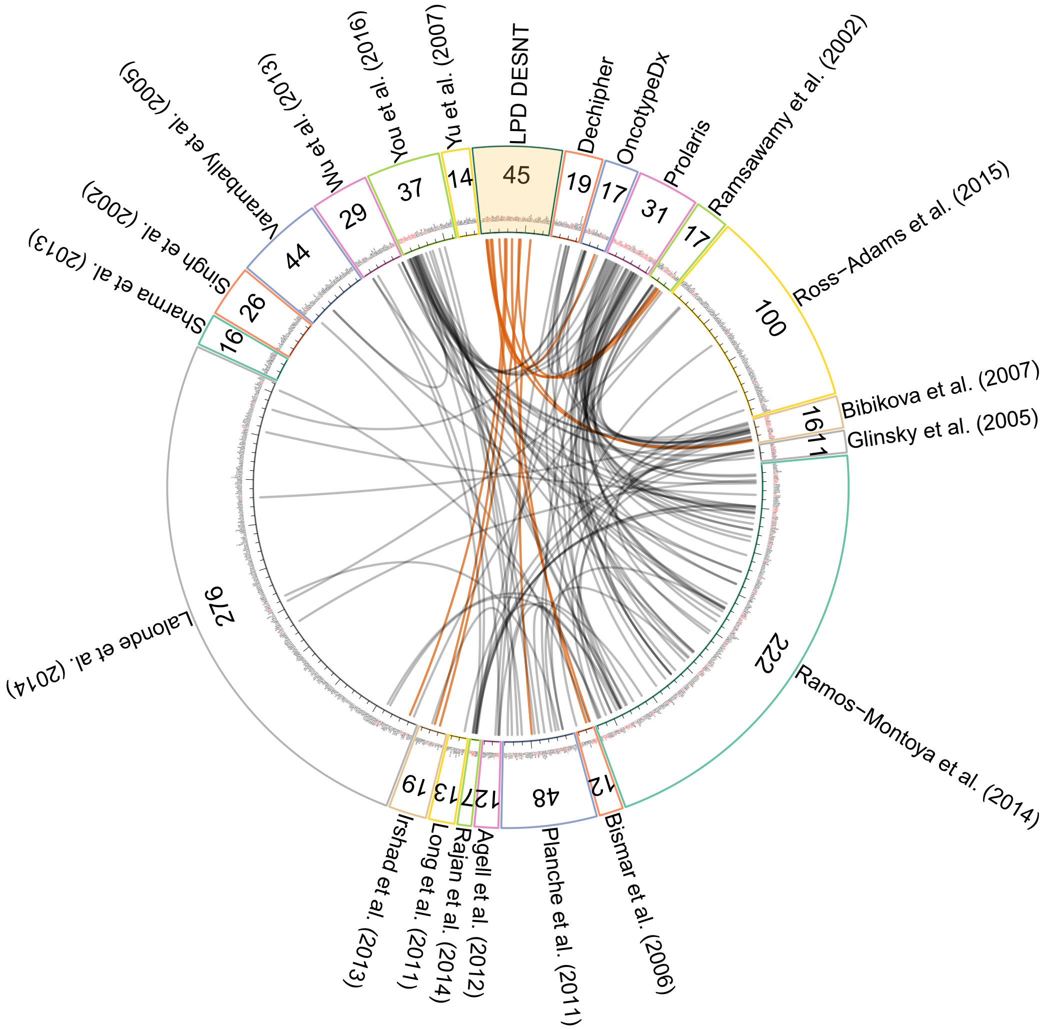

Figure 5.13 Relationship between the genes in different poor prognosis signatures for human prostate cance and the DESNT classification, represented as a circos plot. Sectors correspond to signatures, the numbers on each sector denote the number of genes in each signature. Links to the 45 commonly down-regulated genes are shown in brown.

The LPD DESNT signature shares 11/45 genes with other signatures, namely:

- *TPM2* in common with the commercial test Oncotype Dx;
- *ACTG2*, *CNN1*, *MYLK* shared with Ramaswamy et al. [311];
- *FLNA* and *ITGA5* in common with Bismar et al. [310];
- *MYLK* and *PPAP2B* also in Bibikova et al. [313];
- *CLU* and *GPX3* shared with Irshad et al. [317];

- *ACTA2* also in Lalonde et al. [319];
- *ETS2* in common with Planche et al. [309].

5.11 Comparison with traditional clustering methods

We tried to assess if other commonly used unsupervised classification methods, such as hierarchical clustering and the k-means algorithm on the PCA reduced data, would be able to robustly identify the DESNT group. In each case we tried to reproduce the LPD analysis as closely as possible. For each dataset we used exactly the same set of probes we used for LPD and, also, the same set of samples as in the LPD analysis, i.e. all the samples available. Then, for survival analysis we removed the duplicates from the same patient, as before.

5.11.1 Hierarchical clustering

The hierarchical clustering of MSKCC and CancerMap, performed as described in Section 5.4.7, revealed that the DESNT samples (Figure 5.14) do not cluster together.

We further tried to determine if the hierarchical clustering provides an alternative classification, that identifies groups of patients with different clinical outcomes. From the structure of the dendrograms it is not straight forward to objectively infer the number of clusters. Therefore, we decided to use the likely number of clusters identified with the help of LPD, namely eight for both datasets. For each dataset, we cut the dendrograms starting from the top until we obtained eight groups. The resulting groups are presented in Supplementary Figure B.10a,b, and include all the samples in the dataset, including the duplicates.

The survival analysis (Supplementary Figure B.10c, d) fails to identify any significant association between these groups and the time to BCR (log-rank p-values 0.98 for MSKCC and 0.22 for CancerMap).

These results, suggests that hierarchical clustering does not manage to robustly identify the DESNT group, or any other group of samples with poor outcome.

5.11.2 PCA and k-means

We also evaluated how the k-means algorithm (Section 3.2.1.2) performs in robustly detecting the DESNT group, or other poor outcome groups. Before applying k-means, we transformed the data, using principal component analysis (PCA), as described in Section 3.4, to reduce the dimensionality and make the data easier to visualise.

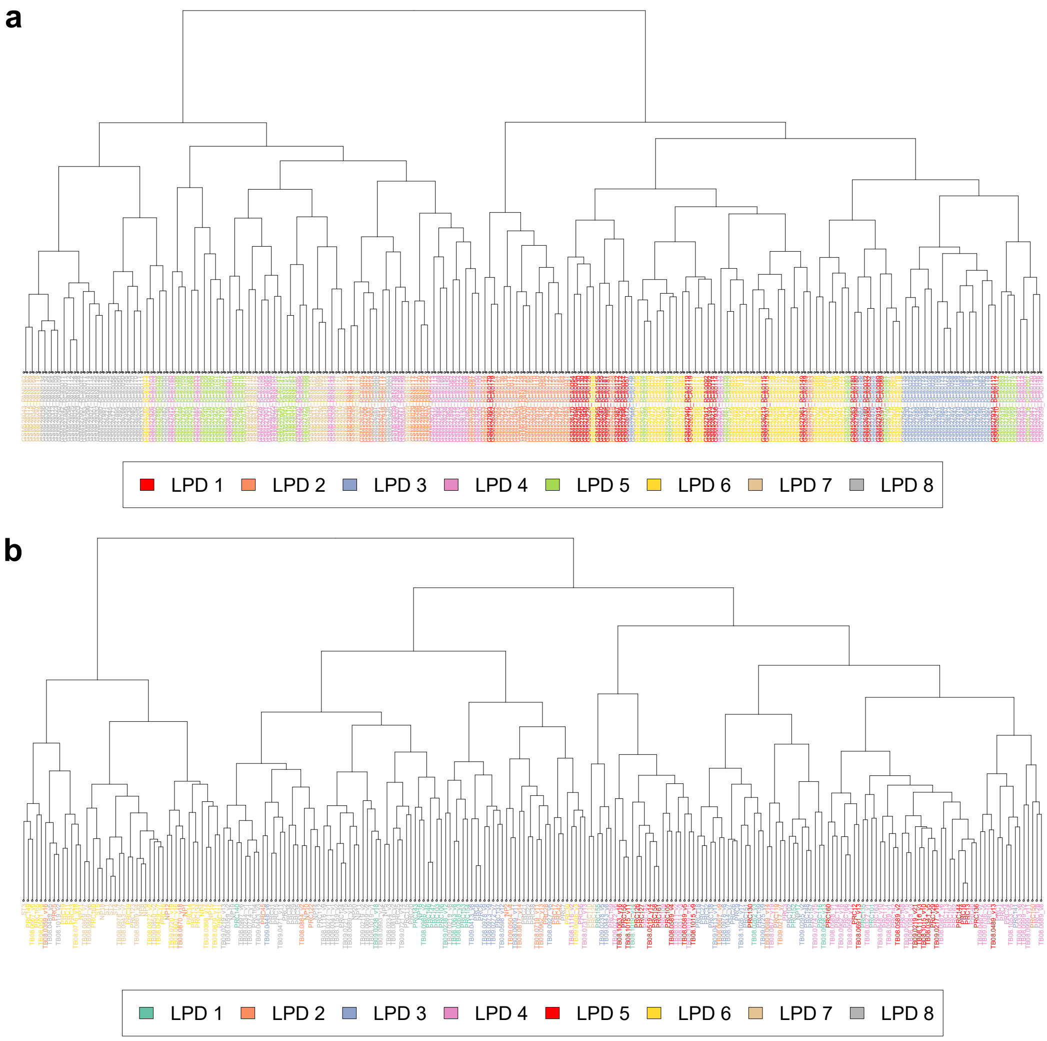

Figure 5.14 Hierarchical clustering on the MSKCC and CancerMap datasets. The colours represent LPD groups. The samples labelled with red (LPD1 in the top panel and LPD5 in the bottom panel) correspond to the DESNT groups.

We applied PCA on the same set of probes used for LPD classification, from all samples in each dataset. For each dataset, we used the first two principal components, which account for the following percentages of variance in the data: MSKCC - 39.6%, CancerMap - 32.9%, CamCap - 18.9%, and Stephenson 27.7%.

On the PCA transformed data, we then applied the k-means clustering, using for each dataset the same number of clusters as for LPD, i.e. 8 for MSKCC and CancerMap, 6 for CamCap and 3 for S6

The output for the k-means clustering depends on the initialization of the centromers and, therefore, different runs can yield different result. As previously, we restarted the algorithm 100 times and chose one representative run. However, the variance in classification of the samples is much lower than LPD. For the MSKCC dataset all runs but one yielded the same output, for the CancerMap 82 runs produced the same classification and for CamCap and Stephenson we obtained the same results every time. Therefore, for MSKCC we chose one of the 99 classifications that give the same result, for CancerMap one of the 82 equivalent results, while for CamCap and Stephenson one of the 100 runs.

For MSKCC (Figure 5.15a,b) the k-means clustering assigns the DESNT samples to two clusters (C4 and C5) together with other samples. The survival analysis did not identify any statistical significant association (log-rank p-value 0.32).

For CancerMap (Figure 5.15c,d), the DESNT samples are also included in two clusters (C5 and C6). The survival analysis in this case suggests that at least one group has significantly different outcomes (log-rank p-value $7.46 \cdot 10^{-4}$). Therefore we performed a log-rank test for each of the 8 clusters, testing if the group has significantly worse outcome than the rest of the samples taken together. The yellow group (C6), which contains four of the ten DESNT samples, is the only cluster with statistical significant association with time to BCR (log-rank p-value $1.92 \cdot 10^{-5}$).

The k-means clustering on Stephenson (Figure 5.15e,f) yielded good correlation with DESNT. All DESNT cancers have been clustered together in a group that contained several other non-DESNT samples. The survival analysis also indicates that there are groups with worse outcome (log-rank p-value $9.32 \cdot 10^{-4}$). The cluster which includes the DESNT samples has a log-rank p-value of $2.31 \cdot 10^{-4}$.

Finally, in CamCap the DESNT samples are split between three clusters. No group was associated with poor outcome (log-rank p-value 0.32).

The k-means clustering does not seem to consistently identify the DESNT group, as in only one in four datasets (Stephenson), it has been able to group the DESNT cancers together. Moreover, k-means does not consistently identify poor prognosis groups (only in two of four datasets).

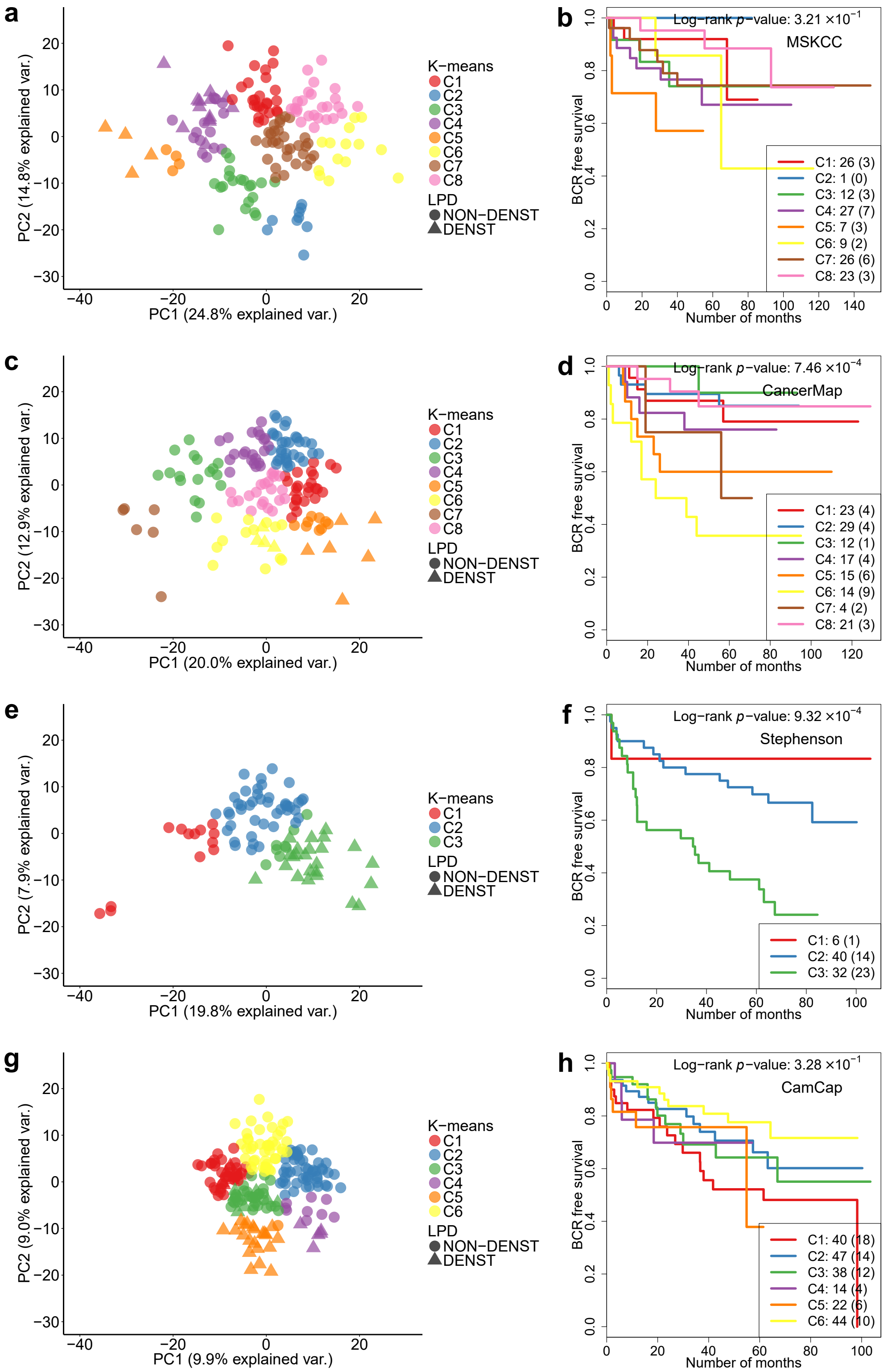

Figure 5.15 PCA analysis, followed by *k*-means analysis on a) MSKCC, c) CancerMap, e) Stephenson and g) CamCap, along with the corresponding survival analysis (b, d, f, h). The round points correspond to non-DESNT samples, while the triangular points correspond to DESNT samples. The colours represent *k*-means clusters.

5.12 Predictive signature for DESNT

With the help of LPD we have been able to define a common process, designated DESNT, in patients from five prostate cancer datasets, which leads to a significantly poorer clinical outcome. In this section we develop a way of predicting DESNT membership, which could indicate a poorer outcome, on new, individual samples, suitable for a clinical setting.

One way of doing this would be to examine the expression profile of the 500 genes in the new sample. The expression profile would need to be normalised relative to a reference dataset. Then, one can verify if the LPD model, trained on the reference dataset, is the main contributor to the expression of the genes in the sample. If the DESNT is the main contributor, the sample is considered a member of the DESNT group.

However, the set of 500 genes is quite large for a prognostic signature and not necessarily specific for the DESNT classification. The existing predictive signatures for different types of cancer use relatively small panels of genes, specific for the classification at hand. For examples in breast cancer the Mammaprint test [15] uses a panel of 70 genes to identify poor prognosis groups, while in prostate cancer the commercially available tests use panels of 31 cell-cycle progression genes (Prolaris [38]), 12 genes (Oncotype DX [39]) and 22 genes (Dechipher [40]). Therefore, in line with the other tests, we tried to find a set of genes, as small as possible, specific for DESNT, that can robustly predict if a new sample is in the DESNT group.

We assessed if a set of genes would be robust in predicting the DESNT membership by training a supervised machine learning model (a random forest classifier) on one dataset and evaluating his performance on all the other datasets.

5.12.1 Data preparation

Five of our datasets (MSKCC, CancerMap, CamCap, Stephenson and Klein) come from three different microarray platforms and TCGA has been generated from RNA-seq data. Each microarray platform uses different probes for measuring the expression of genes, while the RNA-seq estimates the expression of gene from the number of reads mapping to the transcripts.

Depending on the platform, the expression of a gene may or may not be measured on a given platform. Also, for a gene there might be a variable number of probesets available on each platform. Additionally, the gene expression levels are not globally comparable in samples from different platforms, and even across different datasets generated with the same platform, without mitigating the platform-specific and even dataset-specific differences, as we will illustrate shortly.

As we were planning to use the same random forest model on data from different datasets, which might provide different sets of features, on different scales, we worked on bringing the data to a compatible level. We kept only the probes corresponding to genes measured on all microarray platforms, namely, *Affymetrix Human Exon 1.0 ST*, *Affymetrix U133A*, *Illumina HumanHT-12 V4.0* and the genes for which the expression has been estimated in all samples from the TCGA dataset. Also, when more than one probeset was available for a gene, we kept only one of them, chosen at random. This resulted in 10,444 probesets corresponding to 10,444 genes in common between all platforms.

As illustrated in Figure 5.16a, the distribution of intensities of probesets in different datasets is quite variable. To make them comparable we applied the *ComBat* algorithm [290] implemented in the *sva* R package. The algorithm corrected the dataset-specific effects across the data (Figure 5.16b).

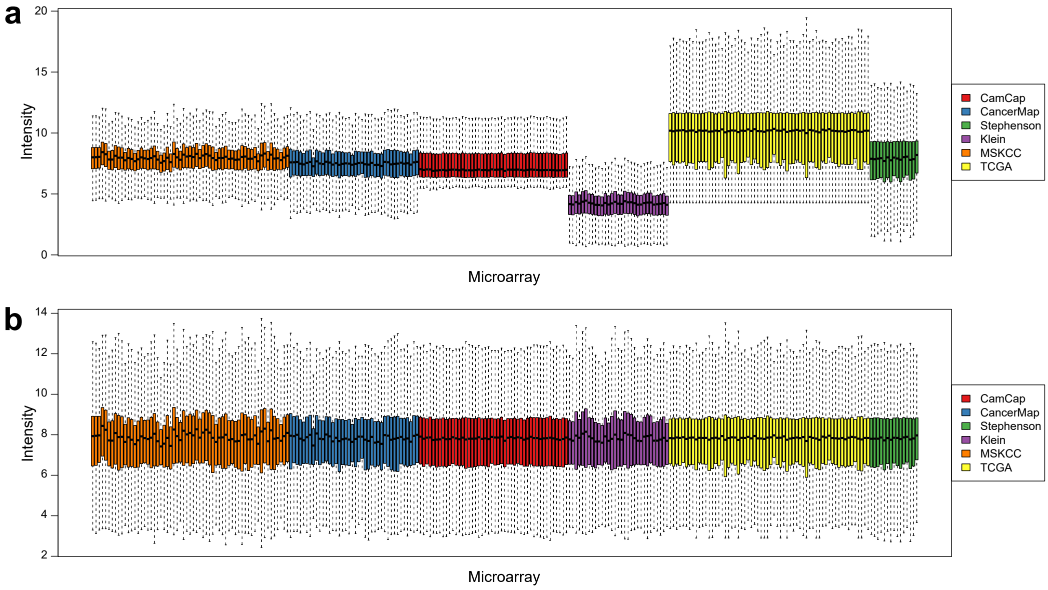

Figure 5.16 Distribution of intensities of probesets across samples from all six dataset datasets a) before and b) after ComBat. Boxes represent samples and colours represent datasets. Note that due to limited space only every other sixth sample in the datasets has been plotted.

For obtaining unbiased evaluations of the performance, we also removed all the duplicated samples from the same patient, as in the previous analyses.

5.12.2 LPD DENST predictive signature

In our first attempt of deriving a robust predictive signature for the DESNT group, we assessed how well the LPD DESNT signature, containing 45 genes (Section 5.8), performs in classifying the DESNT group.

Random forests [213], described in Section 3.2.2.1, have been successfully used before for clinical classifiers, such as, for example, in the commercial prostate cancer test *Dechipher* [40]. We, therefore, trained a random forest model which learned how to discriminate DESNT cancers from non-DESNT cancers using the expression level of the 45 genes in the LPD DESNT signature.

The proportion of samples in the DESNT group is relatively low. In MSKCC there are 17/160 (10.6%) unique samples in the DESNT group, in CancerMap 11/154 (7.1%), in CamCap 21/207 (10.1%), in Stephenson 24/89 (26.9%) and in Klein 42/182 (23.06%). Not taking into account this imbalance would lead to a very low sensitivity when training the model. Therefore, we corrected for class imbalance, by down-sampling the larger class (Section 5.4.8).

We trained a random forest model on the MSKCC dataset and tested its performance on the other datasets, namely, CancerMap, CamCap, Stephenson and Klein. As it can be seen in Figure 5.17, the validation AUC of the ROC curve is always high (0.9112-0.9821) indicating good separation of the two classes. However, even though the sensitivity is relatively good (85.71%-100%), the specificity can get as low as 65% (CamCap).

5.12.3 RF DESNT predictive signature

The LPD DESNT signature performs well in identifying DESNT cancers, but it is not specific enough for this purpose. To improve the specificity of classification, we constructed an alternative signature.

We set as a starting point all the genes identified as differentially expressed in DESNT, in at least two of the five datasets. There are in total 1,496 differentially expressed in at least two of the five dataset (MSKCC, CancerMap, CamCap, Stephenson and Klein).

To reduce this list, we used the LASSO logistic regression model, described in Section 5.4.9. LASSO shrank the regression coefficients of most of the genes to 0, selecting on average only 20 genes with non-zero coefficients. The variation between different restarts of LASSO was very small. After 100 runs, LASSO selected a total of 30 distinct genes, of which some have been selected in just 1-2 runs. To obtain a robust

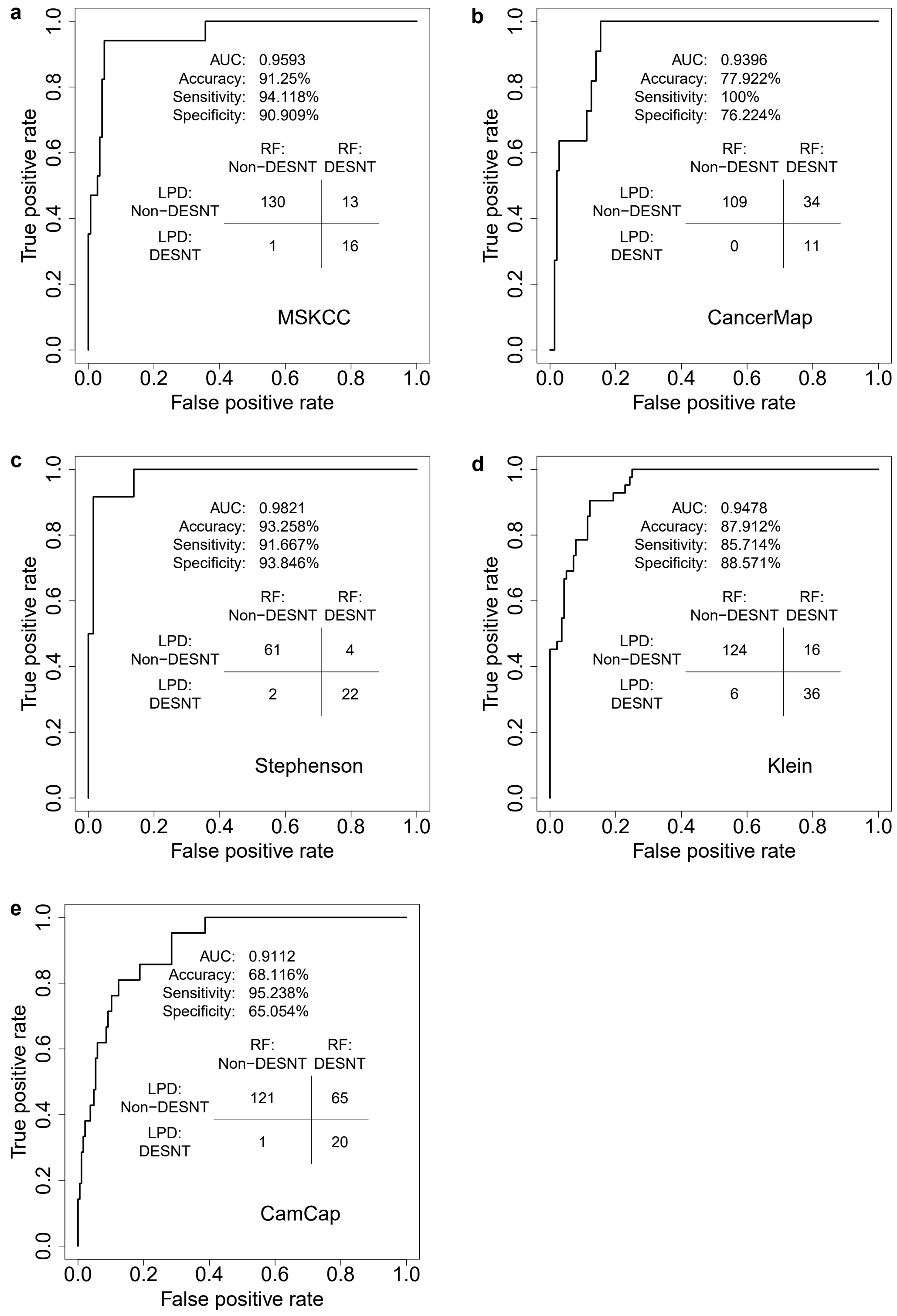

Figure 5.17 The performance of the random forest model using the LPD DESNT signature on a) the training dataset, MSKCC and the validation datasets b) CancerMap, c) Stephenson, d) Klein, e) CamCap. The samples have been assigned to the class with the highest number of votes from the decision trees.

set of predictor genes, we imposed a threshold of minimum 25 runs in which a gene needs to be selected by the LASSO model. Using this strategy, we obtained a list of 20 genes (Table 5.3). For the rest of the analysis we will refer to this 20 genes signature as the RF DESNT signature.

Table 5.3 The 20 genes that have been selected by LASSO logistic regression. The second column contains the variable importances estimated by the random forest classifier, trained on MSKCC.

Gene	**Variable importance**
DST	2.208882
CHRDL1	1.820796
THSD4	1.585727
GSTM4	1.568462
CYP27A1	1.450209
ACTG2	1.371804
RND3	1.280383
PLEKHA6	0.688459
SP100	0.669480
PARM1	0.643371
ZNF532	0.573341
ALDH2	0.528605
DLG5	0.467959
WDR59	0.461952
LDHB	0.418893
CDK6	0.330462
MME	0.268322
S100A13	0.236298
MSRA	0.228337
EPHX2	0.198256

Then, we trained a new random forest model to classify the DESNT cancers based on the expression of the 20 genes panel that make up the RF DESNT signature. As before, we trained the model the MSKCC dataset and tested its performance on the other datasets. We also adjusted for imbalances by down-sampling the larger class. The performance of classification using the RF DESNT signature improved compared to the LPD DESNT signature (Figure 5.18). The AUC of the ROC curves is in the range 0.937-0.9942, suggesting very good separation of the classes. The sensitivity remained high 78.5%-100% and the specificity increased to 82.7%-95.3%.

In the end, as the classifier seems to produce good results on training and validation data, we further used it to classify the samples from the TCGA dataset, that have not

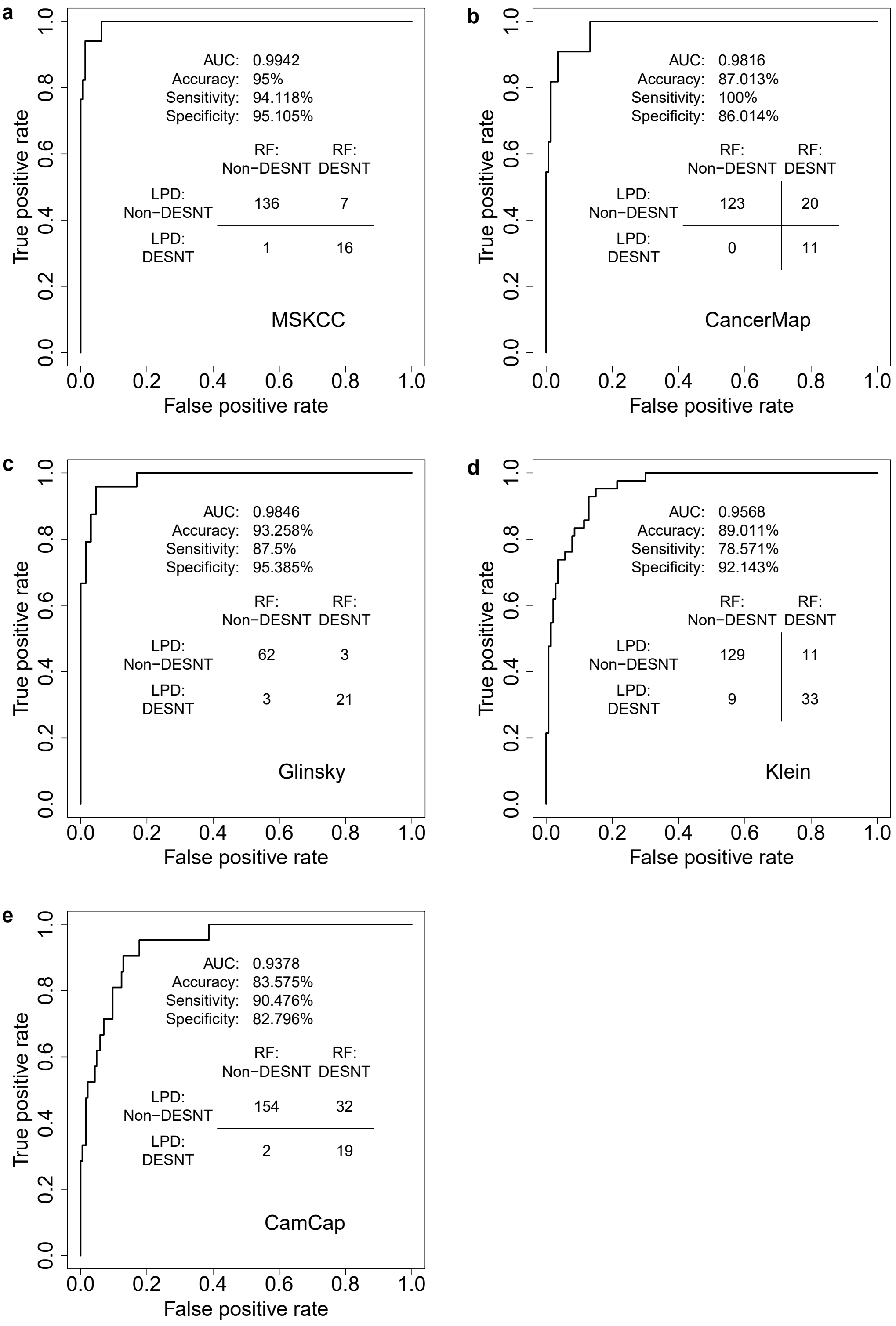

Figure 5.18 The performance of the random forest model using the RF DESNT signature on a) the training dataset, MSKCC and the validation datasets b) CancerMap, c) Stephenson, d) Klein, e) CamCap. The samples have been assigned to the class with the highest number of votes from the decision trees.

been used in the previous analyses at all. The random forest model classified 81/333 (24.3%) samples as DESNT.

The differential expression analysis identified that all 45 genes in the LPD DESNT signature are significantly down-regulated in the group of samples classified as DESNT cancers in the TCGA dataset, suggesting that random forest identified a group of DESNT cancers.

5.12.4 Survival analysis

We repeated the univariate and multivariate survival analysis, in order to determine if the membership to the DESNT group identified by the random forests classifier is a significant predictor of recurrence. For clarity we will refer to this group as the *RF DESNT* group, while for the rest of the samples we will refer to as the *RF non-DESNT* group.

Additional to the previous survival analysis, described in Section 5.6, we performed multivariate analysis on the TCGA dataset. All the covariates used in the multivariate analysis, namely the RF DESNT membership, Gleason grade and the pathological stage, satisfy the PH assumption for the TCGA dataset (Schoenfeld residuals p-values: 0.51, 1 and respectively 0.151). For this dataset we could not obtain the PSA levels at diagnosis. And, since we cannot exactly reproduce the multivariate analysis performed in the other datasets, we opted for splitting the Gleason grade into $\leq$/> 3+4, for a better stratification of the samples that are available.

The results of the both analyses are presented in Figure 5.19, while in the Supplementary Table B.8 we include the full results for the multivariate analysis.

The univariate survival analysis produced statistically significant associations with the time to BCR in all five datasets for which we have clinical data (MSKCC: log-rank p-value $1.86 \cdot 10^{-3}$; CancerMap: log-rank p-value $4.8 \cdot 10^{-4}$; Stephenson: log-rank p-value $1.73 \cdot 10^{-4}$; CamCap: log-rank p-value $1.61 \cdot 10^{-5}$; TCGA: log-rank p-value $1.86 \cdot 10^{-4}$).

The multivariate analysis identified the RF DESNT membership as independent predictor of recurrence in all datasets, except MSKCC (MSKCC: HR 1.29, 95% CI 0.49-3.4, p-value $6.05 \cdot 10^{-1}$; CancerMap: HR 2.51, 95% CI 1.19-5.25, p-value $1.45 \cdot 10^{-2}$; Stephenson: HR 3.37, 95% CI 1.71-6.672, p-value $4.56 \cdot 10^{-4}$; CamCap: HR 2.87, 95% CI 1.67-4.93, p-value $1.31 \cdot 10^{-4}$; TCGA: HR 2.1145, 95% CI 1.09-4.08, p-value $2.59 \cdot 10^{-2}$).

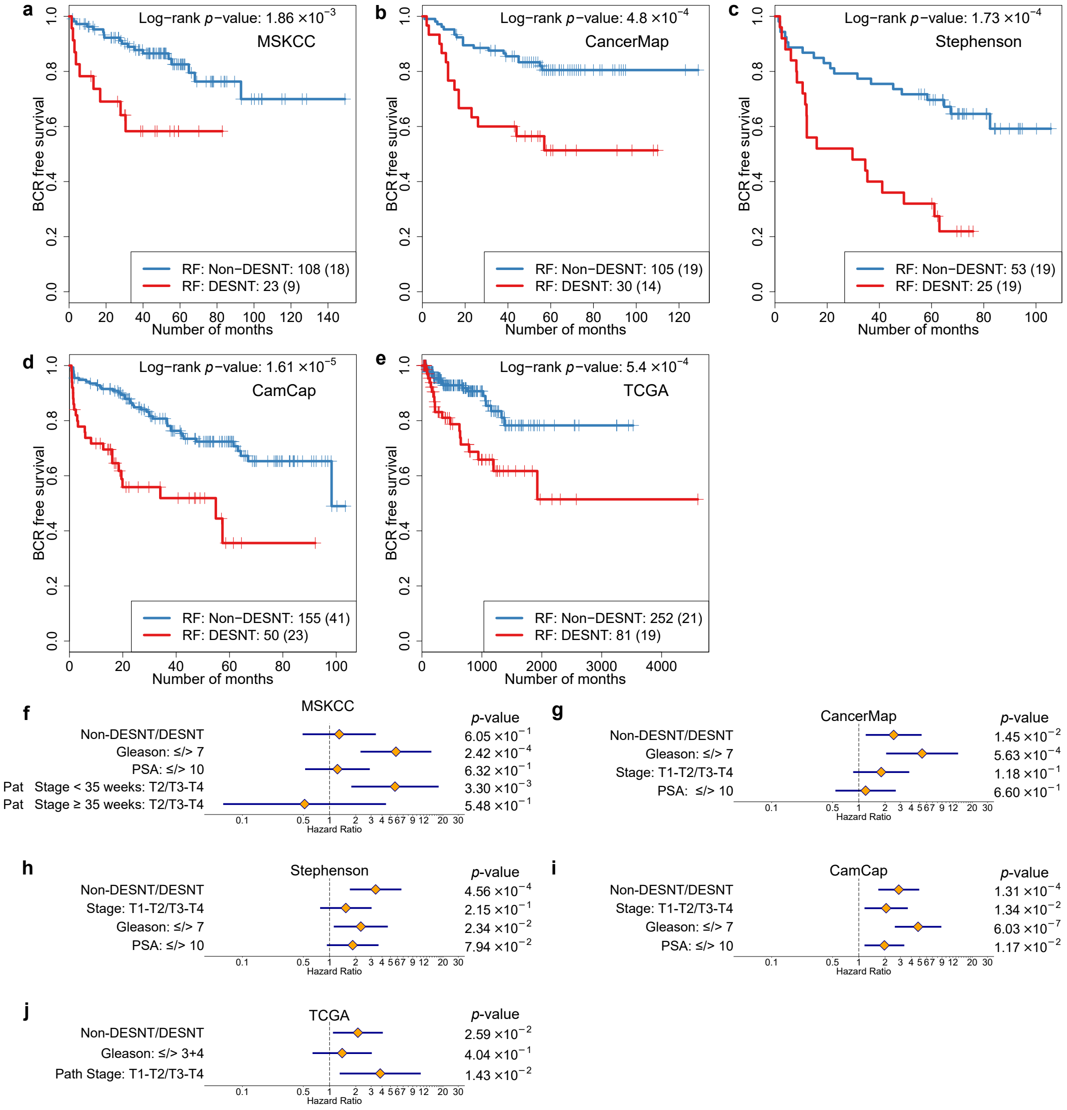

Figure 5.19 Analysis of outcome for DESNT cancers identified by RF classification. KM plots for the a) MSKCC, b) CancerMap, c) Stephenson, d) CamCap and e) TGCA datasets. For each dataset the cancers assigned to DESNT using the 20 gene RF classifier are comparing to the remaining cancers. The number of cancers in each group is indicated in the bottom right corner of each plot. The number of cancers with PSA failure is indicated in parentheses. Multivariate analyses results are depicted for the f) MSKCC, g) CancerMap, h) Stephenson, i) CamCap and j) TCGA datasets.

5.12.5 Correlation between the RF DESNT groups

We also investigated if the expression profiles of the samples in the RF DESNT groups are correlated across datasets, as in the case of LPD DESNT groups. To obtain results comparable with the previous correlation analysis (Section 5.7), we considered all the probesets (from the total of 10,444 with which we worked in this part of the analysis) that are mapping to the top 500 genes that have been used for the LPD analysis.

We calculated Pearson's correlation between every pair of RF DESNT expression profiles from MSKCC, CancerMap, CamCap, Stephenson, Klein and TCGA as presented in Section 5.4.4. Between every pair of RF DESNT groups there is a quite high correlation (Figure 5.20). The lowest correlation, between CamCap and MSKCC is 0.665, and the highest, between TCGA and CamCap is 0.846. We note in particular that the RF DESNT group in TCGA is very highly correlated with all RF DESNT groups in the other datasets (Pearson's correlation 0.731-0.846), confirming that the random forest classifier identifies for TCGA a DESNT group, which is similar to the other DESNT groups in the other datasets.

5.13 Correlation of DENST group with gene mutations

We next compared the frequency of *ETS* fusion status, *ETS* genes overexpression, mutations, and homozygous deletions in commonly altered genes in prostate cancers [287], between RF DESNT and RF non-DESNT cancers in the TCGA dataset. Also we compared the *ERG* gene rearrangement status in CancerMap, determined using the FISH break-apart assays, between RF DESNT and RF non-DESNT samples.

None of the *ETS* genes has statistically different alteration (fusion or overexpression) frequency in RF DESNT compared to RF non-DESNT in TCGA (*ERG* χ^2 *p*-value 0.29, *ETV1* χ^2 *p*-value 0.32, *ETV4* χ^2 *p*-value 0.83, *FLI1* χ^2 *p*-value 0.51). When the four *ETS* genes are taken together, the alteration frequencies are again not correlated with RF DESNT membership (χ^2 *p*-value 0.13). The frequency of *ERG* fusions in CancerMap is not significantly different between RF DESNT and RF non-DESNT cancers either (χ^2 *p*-value 0.26).

We compared the mutations and the homozygous deletions of the genes presented in Figure 5.21 between RF DESNT and RF non-DESNT cancers. No gene shows significant correlation between mutations (Supplementary Table B.10) or homozygous deletions (Supplementary Table B.11) and RF DESNT membership. When we combined together the mutation and homozygous deletion status for each gene, *TP53*, *BRCA2* and *CDK12* yielded significant unadjusted χ^2 *p*-values ($3.84 \cdot 10^{-3}$, $2.07 \cdot 10^{-2}$,

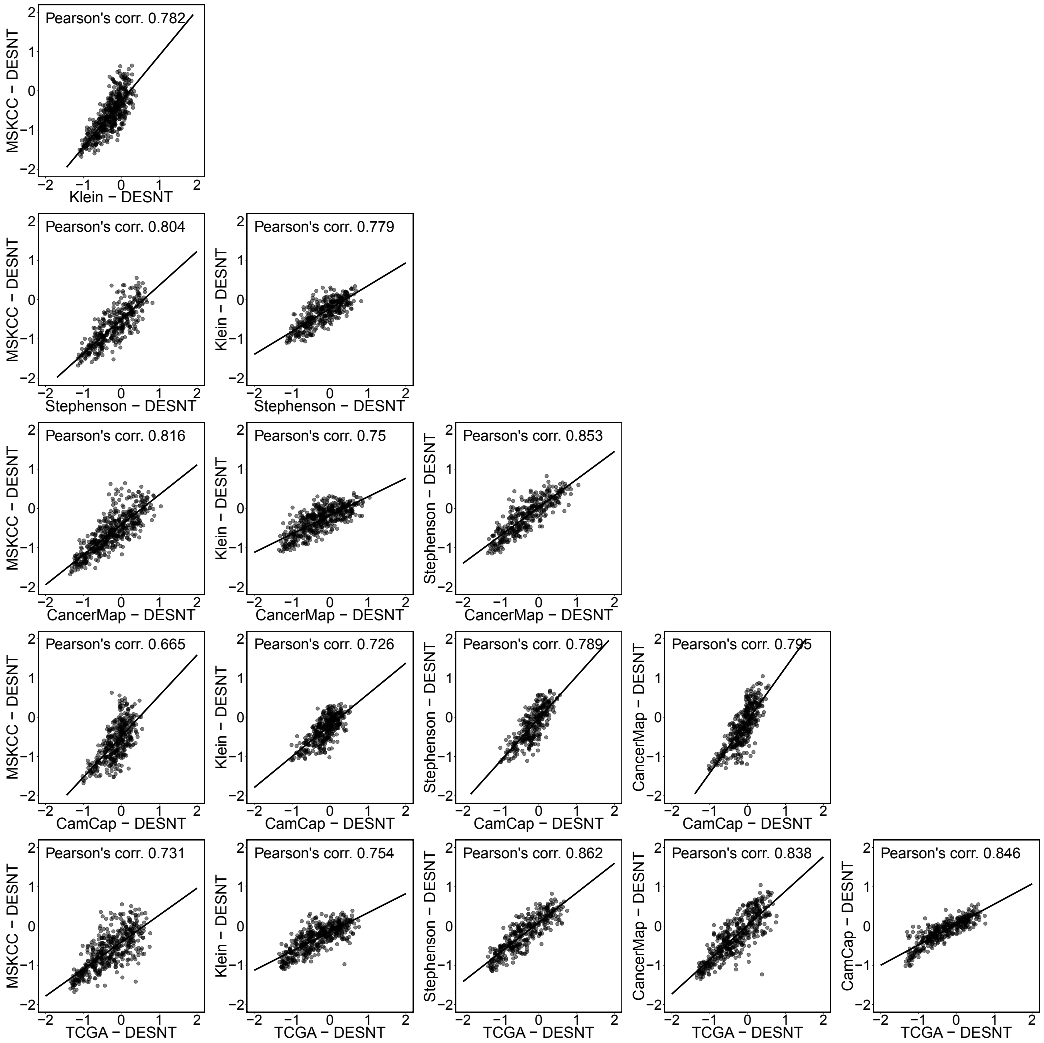

Figure 5.20 Correlations between the expression profiles of every possible pair of RF DESNT groups from the MSKCC, CancerMap, Stephenson, CamCap, Klein and TCGA datasets. The expression levels of each probesets have been normalised across all samples to mean 0 and standard deviation 1.

and respectively $4.19 \cdot 10^{-2}$; Supplementary Table B.12). However after adjusting for multiple comparisons using FDR correction at 5% level, none of the genes remained significant (FDR adjusted χ^2 p-values $8.45 \cdot 10^{-2}$, $2.28 \cdot 10^{-1}$, and $3.07 \cdot 10^{-1}$ respectively).

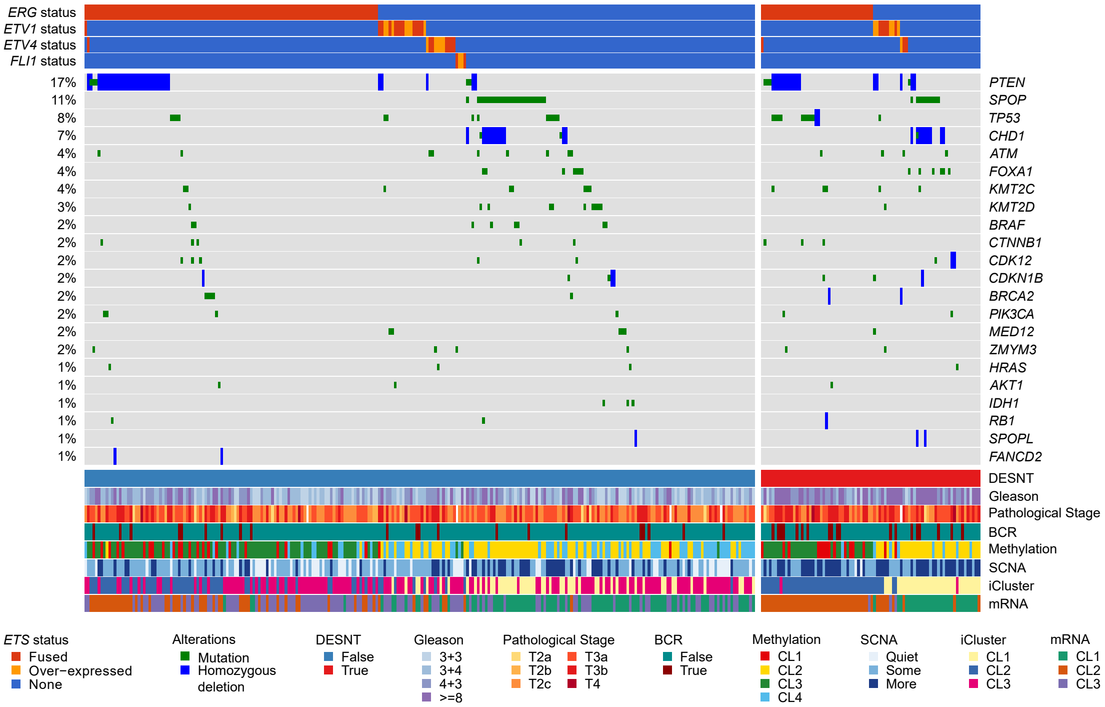

Figure 5.21 Comparison of genetic alterations in RF DESNT and RF non-DESNT cancers in the TCGA dataset. The types of genetic alteration are shown for each gene (mutations, fusions, deletions, and overexpression). Clinical parameters including biochemical recurrence (BCR) are represented at the bottom together with groups for iCluster, methylation, somatic copy number alteration (SCNA) and mRNA clusters [287].

5.14 Methylation analysis

As described in Section 2.4.3, the CpG methylation is an epigenetic alteration that can play important roles in cancer development. Therefore, we further investigated if methylation might play a role in the underexpression of the core 45 genes in the LPD DESNT signature.

We performed a differential methylation analysis (Section 5.4.10) on the set of 1,122 probes mapping to the 45 genes in the LPD DESNT signature with the purpose at identifying differentially methylated regions between RF DESNT and RF non-DESNT groups. The analysis identified 77 differentially methylated probes, corresponding to 24 out of the 45 core genes, shown in Supplementary Table B.9. These probes correspond to 43 differentially methylated regions (Supplementary Table B.13).

In Figure 5.22, we present a heatmap illustrating the beta-values of the 77 differentially methylated probes. As it can be seen in the right panel, most of the probes are hyper-methylated in the RF DESNT group relative to the non-DESNT tumours (middle-panel) and also the normal samples (left panel).

To investigate if the hyper-methylation (and also hypo-methylation) plays a role in the expression levels of the genes, for each of the 77 differentially methylated probe we calculated the Pearson's correlation between the beta-value of each sample and the expression level of the corresponding gene in the sample (Supplementary Table B.14).

For many genes the hyper-methylation of the corresponding probes (not necessarily probes from the promoter region) is strongly inversely correlated with the expression levels, and, conversely, the hypo-methylation is directly correlated with the expression levels. In particular, all probes corresponding to *GSTP1* are highly inversely correlated with the expression (Pearson's correlation between -0.72 and -0.87). Similarly for *SPG20* (Pearson's correlation between -0.7 and -0.74) and *PDLMIM1* (Pearson's correlation -0.75). In fact with few exceptions, most probes yield correlations above 0.5 or below -0.5, suggesting strong associations between methylation and expression.

The results are consistent with the possible involvement of the methylation in the underexpression of the core 45 genes, and suggest a possible role of epigenetic changes in the progression of prostate cancer. However, further work needs to be performed to elucidate this hypothesis.

5.15 Discussion

In this chapter we presented the application of LPD to prostate cancer transcriptome datasets, which has revealed the existence of a novel poor prognosis category of prostate cancer common across all prostatectomy datasets examined. The robust nature of DESNT cancers is supported by their detection in data generated using several different platforms (Illumina HT12 v4 BeadChip arrays, RNA-seq, Affymetrix arrays) and from both frozen and formalin fixed material.

The DESNT cancers are characterised by a core set of 45 down-regulated genes, many of them with role in cytoskeleton machinery, ion transport and cell adhesion. Our observations also provide clues about possible mechanisms of development of aggressive disease. For example, the down-regulation of genes determining cytoskeleton structure and involved in focal adhesion in these cancers would argue against the contri-

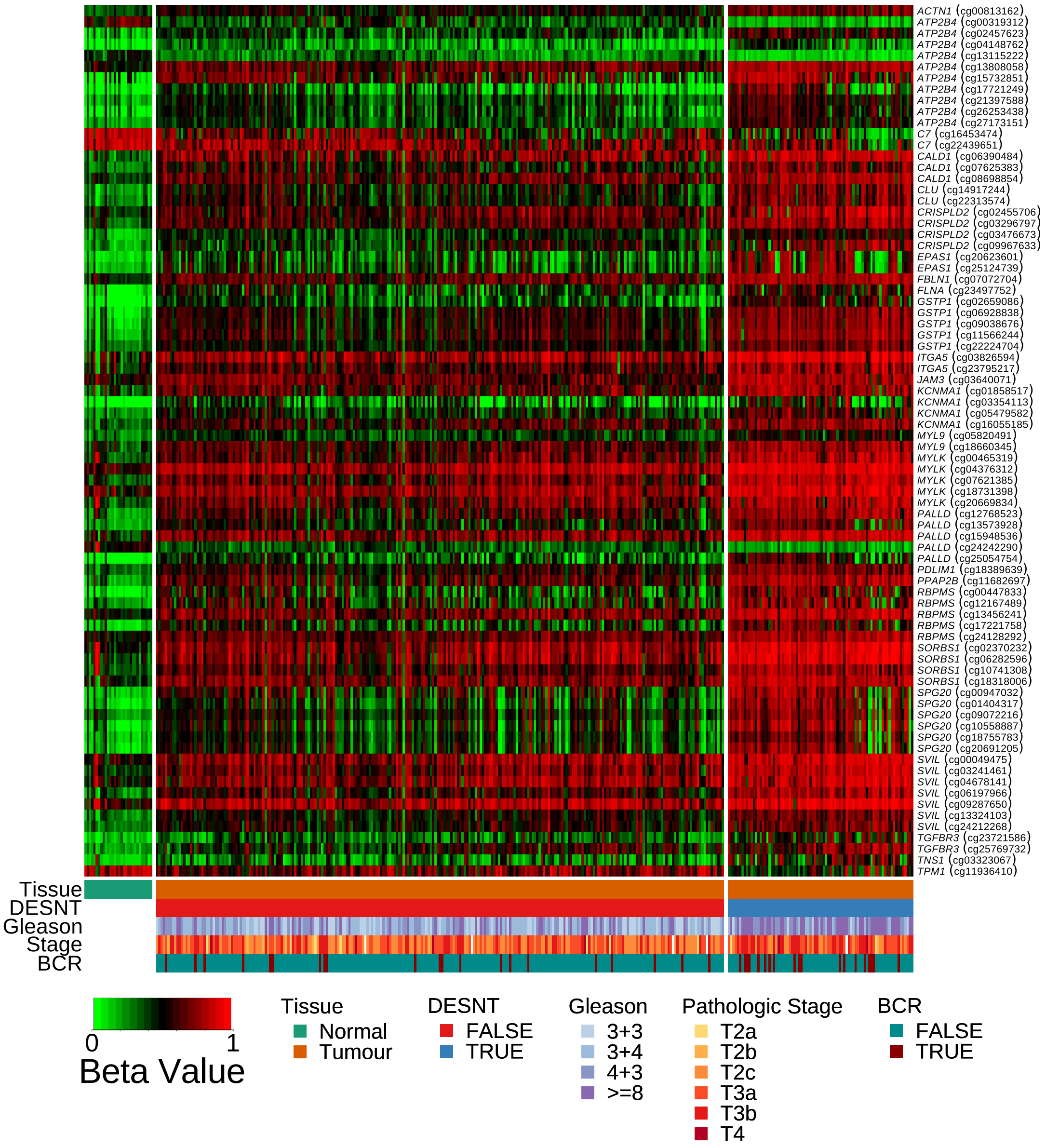

Figure 5.22 Heatmap corresponding to the differentially methylated probes. The rows represent probes, the columns represent samples and the colours correspond to the beta values. The left panel corresponds to benign samples, the middle panel to RF non-DESNT samples and the right panel to the RF DESNT samples.

butions of amoeboid-type movement and mesenchymal migration, but is consistent with involvement of the collective migration phenotype in determining cancer aggression.

The involvement of 11 of the 45 core genes in other prostate cancer signatures, including one gene in common with the commercial test Oncotype Dx, supports the association of the DESNT cancers with the aggressive behaviour. This is even more astonishing as the 45 genes have been selected in a completely unsupervised fashion, without using any previous knowledge or clinical data.

As, the core set of 45 genes does not have enough specificity for predicting the DESNT cancers, we derived an alternative set of 20 predictive genes. Using random forest classification, these 20 genes provided high specificity and sensitivity for predicting that individual cancers were DESNT in both the MSKCC training dataset and in four validation datasets. For the three validation datasets (Stephenson, CancerMap and CamCap) with linked PSA failure data the predicted cancer subgroup exhibited poorer clinical outcome in both univariate and multivariate analyses, in agreement with the results observed using LPD.

When random classification was applied to RNA-seq data from 333 prostate cancers from TCGA, which have not been used previously in the analysis, a cancer patient subgroup was identified that was confirmed as DESNT based on: (i) demonstration of overlaps of differentially expressed genes between DESNT and non-DESNT cancers with the core down-regulated gene set (45/45 genes), (ii) its poorer clinical outcome compared to non-DESNT patients and (iii) correlations of gene expression levels with DESNT cancer groups in other datasets.

Using information from TCGA we failed to find correlations between assignment as a RF DESNT cancer and the presence of specific genetic mutations. Of particular note there was no correlation to *ETS* gene status. A lack of correlation between DESNT cancers and *ERG* gene rearrangement, determined using the FISH break-apart assay, was confirmed using CancerMap samples. These observations are consistent with the lack of correlation between ERG status and clinical outcome [267]. Since *ETS* alterations, found in around half of prostate cancer [32, 127–131], are considered to be an early step in prostate cancer development [136] it is likely that changes involved in the generation of DESNT cancer represent a later event that is common to both *ETS*-positive and *ETS*-negative cancers.

For DESNT cancers some of the core down-regulated genes exhibited altered levels of CpG gene methylation compared to non-DESNT cancers suggesting a possible role in controlling gene expression. This hypothesis is supported by the inverse correlations between the methylation strength and expression levels of the targeted genes. Further supporting this idea, for 16 of the 45 core genes, epigenetic down regulation in human cancer has been previously reported including six genes in prostate cancer (*CLU*, *DPYSL3*, *GSTP1*, *KCNMA1*, *SNAI2*, and *SVIL*). CpG methylation of five of the genes (*FBLN1*, *GPX3*, *GSTP1*, *KCNMA1*, *TIMP3*) has previously been linked to cancer aggression.

Classification of a cancer as DESNT, when used together with standard clinical indicators (stage, Gleason score, PSA) should significantly enhance the ability to identify patients whose cancers will progress. In turn this will allow the targeting of radical therapies such as radiotherapy and surgery to aggressive disease avoiding the side effects of treatment, including impotence, in men with non-aggressive disease.

Chapter 6

Conclusions and future work

In this thesis we have presented two projects developed for the purpose of identifying biomarkers that could help distinguish aggressive prostate cancer from indolent cancer. We now summarise our findings and indicate some potential directions to carry on this work.

6.1 The identification of transcriptional alterations using exon microarrays

In our first approach we analysed data from exon microarrays with the purpose of identifying transcriptional alterations. We then identified a list of candidate genes, that exhibit transcriptional alterations that are correlated with the time to BCR and metastasis.

The results yielded by this approach are promising. We identified several genes with possible transcriptional abnormalities associated with aggressive disease. Of these, some are known to be involved in fusions and some are possible novel fusion candidates or genes involved in other abnormal transcription events, such as trans-splicing.

However, this approach has some shortcomings. One of the most important shortcomings is that exon microarrays do not provide enough information to determine the nature of the alterations. It is impossible to tell if the alterations are generated by fusions, other types of alterations or non-biological artefacts. Also, for fusions and trans-splicing events the fusion partner cannot be identified.

Another issue is that jumps in the expression of the exons are not always correlated with fusions or other alterations. Many times fusions do not result in jumps and jumps are not always the result of fusions. Even if jumps are generated as result of transcriptional abnormalities, sometimes they are very small. This makes it difficult to distinguish them from noise in the data.

We set the parameters of our model so that we obtain an optimal classification on genes with known fusion status. However, because the jumps generated by real fusions are sometimes quite small, we needed to set very relaxed detection thresholds. This led to the identification of a large number of candidates. Despite our efforts of reducing this number by introducing additional filtering criteria, we still obtained thousands of candidate genes.

The correlation of the breakpoints predicted by our method and the breakpoints reported in the literature gives us confidence that at least some candidates are generated by transcriptional abnormalities. However, our method, in line with the other methods developed for this purpose, only shortlists a set of candidates. These candidates need to be further validated using established methods, such as FISH and qt-PCR. Alternatively, RNA-seq analysis of the candidates could help confirming that they are involved in abnormal events.

In terms of the jump detection method per se, we are content with the performance obtained. In our evaluation, it performs at least as well as previous methods. Our investigation on the *ERG* gene concluded that in the overwhelming majority of cases the method could distinguish between samples with jumps and samples without. Most of the misclassification was caused by discordance between jumps and fusions.

As for the practical relevance of the results, many of the novel step-up candidates correlated with metastasis are involved in cell cycle progression, known to be associated with aggressive prostate cancer. Also many of the step-down candidates are involved in the muscle contraction pathway and actin cytoskeleton, which are associated withe DESNT cancers. Moreover, many of these candidates, such as *SORBS1*, *VCL*, *ACTG2*, *CALD1* and *TPM2*, are also found in the list of core down-regulated genes in DESNT. If they are proven to be involved in fusions or other events, it might enhance our knowledge about the formation of DESNT cancers and the progression of prostate cancer.

6.2 The DESNT group

In the second project we analysed six transcriptome datasets, generated from radical prostectomy samples. We applied a Bayesian model, called LPD, to robustly identify a group of aggressive cancers designated DESNT. We then developed a random forest model that, using a set of 20 genes, can predict the DESNT membership.

This is the first time that a robust genetic subtype associated with a poor outcome has been reported for prostate cancer. This is an important stepping stone in defining homogeneous subtypes of prostate cancer and a significant progress towards personalised management of the disease.

One of the results we can not totally explain is the variable number of processes underlying each dataset. This number varies between three and eight. The most probable cause is the size of the dataset. For the smallest dataset, Stephenson, LPD suggests three processes. For an intermediary size datasets, such as Klein, which contains 182 samples LPD suggests five processes, while for large ones, such as MSKCC, CancerMap and CamCap, with over 200 samples, LPD indicates 6-8 processes. For the smaller datasets the LPD does not have enough information to characterise all the processes and therefore merges together the most similar ones. We will further explore this aspect in the next section.

The current direction of a lot of prostate cancer research, and cancer research in general, is to study globally the genetic alterations found in tumour, and to try to understand cancer based on the effect of mutations on key pathways. Despite the incontestable gains that this strategy has brought into our understanding of cancer, it might not always explain the behaviour of the disease. A less explored area, but one which has gained momentum lately, is the study of epigenetic alterations. These seem to have strong effects on cancer development. In our analysis the DESNT cancers do not seem to show any correlation with *ETS* gene fusions, mutations or copy number alterations. They do, however, show strong correlation with CpG methylation patterns. A possible involvement of epigenetic alterations in the heterogeneity of prostate cancer might explain why, after many years of intense research, no genetic or set of genetic alterations is able to define robust subtypes of prostate cancer.

Besides the potential practical application, this analysis also underlines the importance of using more specialised techniques for modelling biological data. Taylor et al. [242] when produced the MSKCC dataset, tried unsuccessfully to identify poor prognosis groups using hierarchical clustering. In our attempts we have also failed to produce a clustering that robustly identifies the DESNT group or other aggressive subgroups using either hierarchical clustering or *k*-means. The progress was made when we employed LPD, which, due to a more realistic modelling of the biological data, which can take into account the cancer heterogeneity, is able to find meaningful classifications where other methods fail.

6.3 Future work on the transcriptional alterations project

We now present possible directions of research that could be pursued in continuation of the work presented in Chapter 3.

6.3.1 RNA-seq validation for the candidates

Exon microarrays give indications of expression alterations within a gene. However, the candidate genes need to be validated using supplementary data.

All studies that have discovered fusion candidates using exon microarrays validated them using biological techniques, such as break-apart FISH or RT-qPCR. As some of the tissue used for generating the CancerMap dataset is still available, biologists could go back and verify some of the candidates using these techniques. The alterations associated with recurrence or metastasis would be a good starting point in this analysis.

Alternatively, some of the tissues could be sequenced using paired-end sequencing. Software packages such as Tophat Fusion [321] could be used to identify paired reads that align to different genes, indicating fusions, trans-splicing or alternative splicing. The abnormal transcription events identified by RNA-seq, could then be used to validate the candidates. This approach would be more efficient than the biological validation, as it provides genome-wide results, allowing for more candidates to be simultaneously evaluated.

6.4 Future work on the LPD approach

We now present some possible directions of research that could be pursued in continuation of the LPD work presented in this thesis.

6.4.1 The development of a clinical test

As DESNT cancers have a poor outcome, it would be useful to develop a test that could be used in clinical practice to predict the DESNT membership. An assignment of a new cancer to DESNT would be an indication of poor outcome. However, there are many challenges that need to be considered when transferring these findings to clinical practice.

First of all, it has to be decided at what phase along the clinical management strategy this test is suitable. Is it suitable for predicting progression at diagnosis, based on biopsy samples? Is it applicable for predicting recurrence and progression following prostatectomy, based on samples obtained from the resected prostate? Is it suitable for both phases?

All six datasets that we have worked on have been generated from radical prostectomy samples and the end point available was biochemical recurrence. Since we have been able to train a model that used the 20 gene RF DESNT signature to predict the DESNT membership in five other datasets, we are confident that the test can be reliable when transferred to predict recurrence in clinical practice.

However, only around 35% of men with BCR progress to metastasis [117]. It would be therefore interesting to analyse what is the role of DESNT membership in predicting metastasis and cancer-specific mortality. The Dechipher test, for example, was proven to distinguish patients with BCR that progressed to metastasis from patients with BCR that did not progress to metastasis [21, 322]. If it is shown that DESNT can distinguish patients that progress to metastasis from those who do not, it might lead to an improved selection of the management disease strategy and the patients with good outcome can be spared the adverse effects of androgen-deprivation therapy.

At least as important as predicting recurrence, would be for the DESNT test to be able to indicate aggressive prostate cancer at the time of diagnosis. This could lead to a better stratification of the patients and reduce overtreatment. It is therefore necessary to validate the DESNT test on data generated from biopsy samples.

Besides the practical aspects, it is also important to consider the technical challenges involved in building such a test. Probably the most important two questions are: (i) what approach should be taken to determine DESNT membership and (ii) what technology should be used for measuring the expression level of genes.

To determine DESNT membership there are at least two possible approaches. One of them would be to consider a reference dataset, for which the DESNT membership of all samples is known. Then, for new samples, the expression levels of the top 500 genes used for LPD would be measured and normalised to the same level as the rest of the samples in the reference dataset. As LPD produces an expression profile for each process, these expression profiles could be compared with the expression profile of the new sample. If the DESNT process is a significant contributor to the expression of the sample, it could be considered that the sample is a DESNT cancer.

The other approach would be to use a random forest classifier that can predict the DESNT memberships using a set of specific genes, such as the RF DESNT signature. This model would be trained on the reference dataset. Then the new sample, suitably normalised, would be classified by the model.

As for the technology used, there are currently two main approaches used in the existing gene expression biomarkers. Some biomarkers, such as Mammaprint [15] and

Dechipher [40] use microarrays for detecting expression of the genes, while others such as Prolaris [38] and Oncotype DX [39] use RT-qPCR. With the development of the next-gen sequencing, RNA-seq might also be a suitable technology for the quantification of gene expression.

RT-qPCR and microarrays produce relatively concordant results in assessing the gene expression [323]. Microarray and RNA-seq technologies seem to produce similar results as well [324, 325]. Moreover, in our analysis we have been able to apply the models trained on microarray data to a RNA-seq dataset. Taken together these results suggest that either technology is suitable for developing a test.

However, RT-qPCR and custom made arrays are suitable for a small number of genes, while genome-wide microarrays and RNA-seq tests provide information about not only the genes useful for classification, but other potentially useful genes as well. This additional data can be collected into large sets of samples, that can further help improve our understanding of cancer. For example, the GenomeDx Biosciences company, which produces the Dechipher test (based on exon microarrays), collects whole-genome transcriptome data from patients that are using the Dechipher test and creates large databases of samples with clinically annotations. As a proof of usefulness of this approach, You et al. [284] published a study based on over 4,000 samples of which many were obtained from the Dechipher database.

In terms of cost, a PCR based test would be less expensive and easy to run in a hospital. Whole transcriptome on the other hand might be more expensive, and, also, would require a more complex set-up.

6.4.2 The characterisation of other LPD groups

In this thesis we prioritised the study of the DESNT group, as it has invariably shown poor prognosis. However, LPD provides description of the other processes involved in prostate cancer. Besides the DESNT cancers, there are other processes that are highly correlated across datasets.

In Figure 6.1 we present an overall image of how the LPD processes identified in the representative LPD runs relate to each other. To obtain this image we considered all LPD groups identified in all six datasets, including the TCGA dataset (for which we calculated a LPD decomposition into four processes, of which one proved to be a DESNT group). We therefore worked with 34 processes (eight from MSKCC and CancerMap, six from CamCap, five from Klein, four from TCGA and three from Stephenson). Between each possible pair of processes we calculated a distance based on the correlation between the expression profiles of the two processes, as described in Section 5.7. In this way we obtained a distance matrix. Using the multidimensional scaling (MDS) technique [326], we converted the distance matrix to a two dimensional representation of the relation between all LPD processes.

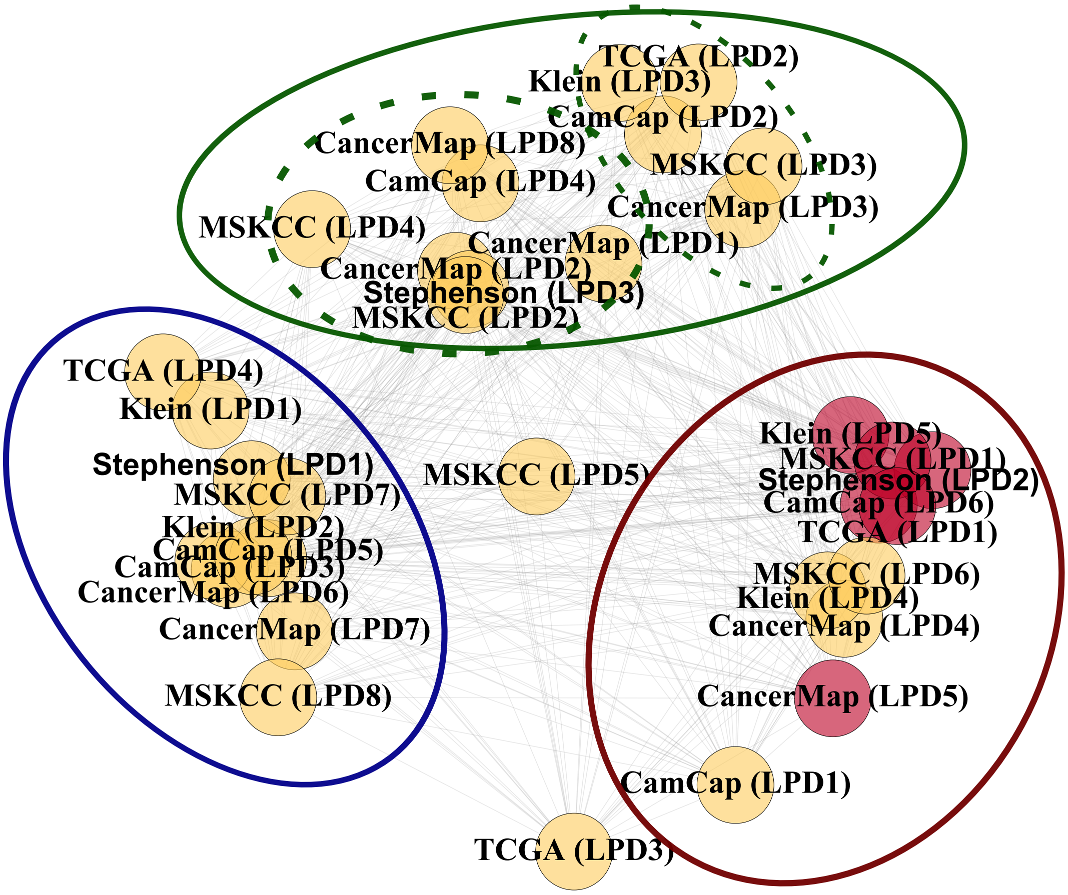

Figure 6.1 The relationship between LPD processes. Each vertex corresponds to an LPD process. The red vertices correspond to the DESNT processes, while the yellow ones correspond to the non-DESNT processes. The distance between vertices is inversely proportional to the correlation between the expression profile of the groups which they represent.

The MDS decomposition suggests 3 main clusters of related processes (the groups of processes surrounded by red, blue and green ellipses). One of them, the green one, might contain two subclusters, depicted with dashed green lines. For simplicity, we will refer to these clusters as the red, blue and green clusters.

6.4.2.1 The red cluster

The red cluster, contains all the DESNT groups. Additionally it contains four non-DESNT groups, each one from a different dataset (CamCap LPD1, CancerMap LPD4, MSKCC LPD6 and Klein LPD4). None of these groups exhibits poor prognosis (log-rank p-value: CancerMap LPD4 p=0.24, CamCap LPD1 p=0.61, MSKCC LPD6 p=0.4).

Despite the overall similarity between these groups and DESNT, the fact that LPD separated them from DESNT suggests some critical underlying differences. One hypothesis would be that the samples in these groups are in process of progressing to DESNT, but, have not acquired some key alterations which make them behave aggressively yet, as the DESNT cancers do.

Another hypothesis is that there are three big types of prostate cancer, represented by the red, blue and green clusters (or four if we consider the two green subclusters as two separate types). Each of these three types contains distinct subtypes, with different behaviour. The red cluster might contain at least two subtypes of cancer. One of them is DESNT and the other one a non-aggressive subtype.

The differences between the DESNT groups and the non-DESNT groups in the red cluster, which contain indolent cancers, might hold the key to what makes the DESNT cancers aggressive. As LPD has been able to find that these subtypes are similar, but at the same time different, the top 500 genes used for LPD would be a good starting point in this line of enquiry.

6.4.2.2 The blue and green clusters

The blue cluster contains, amongst other processes, the MSKCC LPD7 process (which contains only benign samples), the CamCap LPD3 process, also containing many benign samples and the Stephenson LPD1 process, which contains benign samples too. The presence of these groups in the blue cluster suggests an expression profile closer to the normal prostate tissue.

However, in this cluster there is another set of processes, such as CamCap LPD5, CancerMap LPD6 and MSKCC LPD8, which contain a mixture of benign samples and high risk samples. Of these, CamCap LPD5 has significantly poor outcome (log-rank p-value 0.036). In fact, the CamCap LPD5 is the only non-DESNT group that exhibits significantly poor prognosis in all datasets. The other two groups CancerMap LPD6 and MSKCC LPD 8 do not exhibit poor prognosis (log rank p-values 0.57 and respectively 0.22).

The green cluster shows a clear separation between two sets of processes. However, further analysis is required for determining the differences between these sets of processes.

6.4.3 The use of improved versions of LPD

The main motivation for choosing the LPD model proposed by Rogers et al. [209] in our analysis was the fact that our team had previously used it successfully for classifying breast cancer data [302]. We therefore hypothesised that it could yield good results for prostate cancer as well, which we consider it did.

However, several other versions of the LPD model have been developed. One of them, proposed by Ying et al. [327], uses the same specifications for the model, but an improved framework for parameter estimation (model fitting). Instead of the standard variational Bayes (VB) method, used in Rogers et al. [209], this new version uses the marginalised variational Bayes (MVB) framework [327, 328]. In essence, the idea behind these two approaches is the same, namely to estimate a lower bound for the likelihood function described in Section 3.2.1.3. The main difference between them is the strategy used for deriving the lower bound. It can be shown mathematically that the strategy behind MVB offers better solutions [327]. Also numerical experiments confirmed this aspect [327].

A further improvement of the LPD model has been proposed by Masada et al. [329]. This method improves the model solutions by using a new parameter estimation framework designated MVB+, based on the MVB method. In addition to previous approaches, this solution allows for the model hyperparameteres (such as the sigma parameter described in Sections 5.5.2 and 3.2.1.3) to be re-estimated, during model training.

Besides the fact that it produces even better model fittings than the MVB model [329], this approach also simplifies the choice of initial model parameters. As described in Sections 5.5.2, we estimate the initial LPD parameters in three steps. In this new approach there is no need for steps one and two anymore. The sigma is now intrinsically determined and we only need to estimate the number of processes underlying the data.

Another very appealing feature of the MVB and MVB+ approaches is the run time. The authors of the MVB+ approach claim that fitting this version of LPD with 10 processes to a dataset of 286 samples and 17,816 genes took only 174 minutes on an average performance computer [329]. This time is about 10% larger than for the MVB approach [329]. However, compared to the version of LPD we used in our analysis this is a significant improvement. For a dataset of 320 samples and only 500 genes we needed around 24 hours to fit an LPD model with 8 processes.

6.4.4 The application of LPD on other types of cancer

LPD has been previously used to describe four subgroups of breast cancer, two with good outcome and two with poor outcome [302]. In our analysis we have also been able to use LPD to describe a poor prognosis molecular subtype of prostate cancer. As LPD proved its utility in the study of two different types of cancer, it might be useful in studying other types of cancer, especially now that large-scale microarray and RNA-seq data is available.

One type of cancer that could probably benefit from a profiling using LPD is colorectal cancer. Colorectal cancer, as prostate cancer, is a highly heterogenous disease [330]. In the past few years several groups reported unsupervised classification that lead to the definition of molecular subtypes of colorectal cancer [331–338]. However, the results are dissimilar [330]. One of the possible reasons could be that six out of the eight studies [331–336] employed hierarchical clustering for identifying the subtypes. As our data suggests, hierarchical clustering, despite its widespread use, might not be suitable for genetic profiles. Only two other studies [337, 338] used a more advanced method, namely the non-negative matrix factorization (NMF) [339].

Recently, Guinney et al. [330] aggregated the data from several sources obtaining a set of 4,151 samples which were normalised to the same scale. Six of the above models [332–335, 337, 338] were independently applied on the set of samples. They produced six different classifications, predicting between three to six subtypes of colorectal cancer. The results were then aggregated to produce a consensus classification of colorectal into four subtypes.

As there it is a wealth of data available (over 4,000 samples), and the results produced so far have been quite heterogeneous, it would be interesting to see if LPD could produce a robust identification of different subtypes.

6.5 Conclusions

In this thesis we identified several possible transcriptional alterations that, if validated, can lead to a better understanding of aggressive prostate cancer. Also we identified a robust subtype of cancer, denoted DESNT. Classification of a cancer as DESNT, when used together with standard clinical indicators (stage, Gleason score, PSA) should significantly enhance the ability to identify patients whose cancers will progress. In turn this could allow the targeting of radical therapies such as radiotherapy and surgery to aggressive disease avoiding the side effects of treatment, including impotence, in men with non-aggressive disease.

References

[1] Christina Fitzmaurice, Daniel Dicker, Amanda Pain, Hannah Hamavid, Maziar Moradi-Lakeh, Michael F MacIntyre, Christine Allen, Gillian Hansen, Rachel Woodbrook, Charles Wolfe, et al. The global burden of cancer 2013. *JAMA Oncology*, 1(4):505–527, 2015.

[2] ONS. Cancer registration statistics, England: 2014. http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2014, 2014. Accessed August 2016.

[3] JK Anderson, JN Kabalin, and JA Cadeddu. Surgical anatomy of the retroperitoneum, adrenals, kidneys, and ureters. *Campbell-Walsh Urology*, 1:34–37, 2007.

[4] Peter D Baade, Danny R Youlden, Susanna M Cramb, Jeff Dunn, and Robert A Gardiner. Epidemiology of prostate cancer in the Asia-Pacific region. *Prostate International*, 1(2):47–58, 2013.

[5] Grant N Stemmermann, AM Nomura, Po-Huang Chyou, and Ryuichi Yatani. A prospective comparison of prostate cancer at autopsy and as a clinical event: the Hawaii Japanese experience. *Cancer Epidemiology Biomarkers & Prevention*, 1 (3):189–193, 1992.

[6] Manuel Sánchez-Chapado, Gabriel Olmedilla, Manuel Cabeza, Emilio Donat, and Antonio Ruiz. Prevalence of prostate cancer and prostatic intraepithelial neoplasia in Caucasian Mediterranean males: an autopsy study. *The Prostate*, 54 (3):238–247, 2003.

[7] Henrik Grönberg. Prostate cancer epidemiology. *The Lancet*, 361(9360):859–864, 2003.

[8] Gyorgyike Soos, Ioannis Tsakiris, Janos Szanto, Csaba Turzo, P Gabriel Haas, and Balazs Dezso. The prevalence of prostate carcinoma and its precursor in Hungary: an autopsy study. *European Urology*, 48(5):739–744, 2005.

[9] Alexandre R Zlotta, Shin Egawa, Dmitry Pushkar, Alexander Govorov, Takahiro Kimura, Masahito Kido, Hiroyuki Takahashi, Cynthia Kuk, Marta Kovylina, Najla Aldaoud, et al. Prevalence of prostate cancer on autopsy: cross-sectional study on unscreened Caucasian and Asian men. *Journal of the National Cancer Institute*, page 151, 2013.

[10] Masahito Kido, Masahito Hitosugi, Kanto Ishii, Shuichi Kamimura, and Kensuke Joh. Latent prostate cancer in Japanese men who die unnatural deaths: A forensic autopsy study. *The Prostate*, 75(9):917–922, 2015.

[11] Marc A Dall'Era, Matthew R Cooperberg, June M Chan, Benjamin J Davies, Peter C Albertsen, Laurence H Klotz, Christopher A Warlick, Lars Holmberg, Donald E Bailey, Meredith E Wallace, et al. Active surveillance for early-stage prostate cancer. *Cancer*, 112(8):1650–1659, 2008.

[12] NCI. NCI dictionary of cancer terms. http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45618, 2014. Accessed August 2016.

[13] Vathany Kulasingam and Eleftherios P Diamandis. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology*, 5(10):588–599, 2008.

[14] Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187):548–552, 2008.

[15] Marc J Van De Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[16] H Gilbert Welch and Peter C Albertsen. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *Journal of the National Cancer Institute*, 101(19):1325–1329, 2009.

[17] Thomas A Stamey, Norman Yang, Alan R Hay, John E McNeal, Fuad S Freiha, and Elise Redwine. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *New England Journal of Medicine*, 317(15):909–916, 1987.

[18] Eveline AM Heijnsdijk, Elisabeth M Wever, Anssi Auvinen, Jonas Hugosson, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Arnauld Villers, Alvaro Páez, Sue M Moss, et al. Quality-of-life effects of prostate-specific antigen screening. *New England Journal of Medicine*, 367(7):595–605, 2012.

[19] Gerrit Draisma, Rob Boer, Suzie J Otto, Ingrid W van der Cruijsen, Ronald AM Damhuis, Fritz H Schröder, and Harry J de Koning. Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. *Journal of the National Cancer Institute*, 95(12):868–878, 2003.

[20] M Zappa, S Ciatto, R Bonardi, and A Mazzotta. Overdiagnosis of prostate carcinoma by screening: an estimate based on the results of the Florence Screening Pilot Study. *Annals of Oncology*, 9(12):1297–1300, 1998.

[21] Marco Moschini, Martin Spahn, Agostino Mattei, John Cheville, and R Jeffrey Karnes. Incorporation of tissue-based genomic biomarkers into localized prostate cancer clinics. *BMC Medicine*, 14(1):1, 2016.

[22] Jonathan Shoag and Christopher E Barbieri. Clinical variability and molecular heterogeneity in prostate cancer. *Asian Journal of Andrology*, 2016.

[23] Colin S Cooper, Rosalind Eeles, David C Wedge, Peter Van Loo, Gunes Gundem, Ludmil B Alexandrov, Barbara Kremeyer, Adam Butler, Andrew G Lynch, Niedzica Camacho, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature Genetics*, 47(4):367–372, 2015.

[24] Marco Gerlinger, James W Catto, Torben F Orntoft, Francisco X Real, Ellen C Zwarthoff, and Charles Swanton. Intratumour heterogeneity in urologic cancers: from molecular evidence to clinical implications. *European urology*, 67(4): 729–737, 2015.

[25] Michael M Shen and Cory Abate-Shen. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes & Development*, 24(18):1967–2000, 2010.

[26] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[27] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100 (14):8418–8423, 2003.

[28] Anthony V D'Amico, Richard Whittington, S Bruce Malkowicz, Delray Schultz, Kenneth Blank, Gregory A Broderick, John E Tomaszewski, Andrew A Renshaw, Irving Kaplan, Clair J Beard, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA*, 280(11):969–974, 1998.

[29] Ozdal Dillioglugil, Bryan D Leibman, Neville S Leibman, Michael W Kattan, Alejandro L Rosas, and Peter T Scardino. Risk factors for complications and morbidity after radical retropubic prostatectomy. *The Journal of Urology*, 157 (5):1760–1767, 1997.

[30] David F Penson, Dale McLerran, Ziding Feng, Lin Li, Peter C Albertsen, Frank D Gilliland, Ann Hamilton, Richard M Hoffman, Robert A Stephenson, Arnold L Potosky, et al. 5-year urinary and sexual outcomes after radical prostatectomy: results from the prostate cancer outcomes study. *The Journal of Urology*, 173(5): 1701–1705, 2005.

[31] Kimberly A Roehl, Misop Han, Christian G Ramos, Jo Ann V Antenor, and William J Catalona. Cancer progression and survival rates following anatomical radical retropubic prostatectomy in 3,478 consecutive patients: long-term results. *The Journal of Urology*, 172(3):910–914, 2004.

[32] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle

Tchinda, Rainer Kuefer, et al. Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005.

[33] Anuradha Gopalan, Margaret A Leversha, Jaya M Satagopan, Qin Zhou, Hikmat A Al-Ahmadie, Samson W Fine, James A Eastham, Peter T Scardino, Howard I Scher, Satish K Tickoo, et al. *TMPRSS2-ERG* gene fusion is not associated with outcome in patients treated by prostatectomy. *Cancer Research*, 69(4):1400–1406, 2009.

[34] Jacques Lapointe, Young H Kim, Melinda A Miller, Chunde Li, Gulsah Kaygusuz, Matt van de Rijn, David G Huntsman, James D Brooks, and Jonathan R Pollack. A variant *TMPRSS2* isoform and *ERG* fusion product in prostate cancer with implications for molecular diagnosis. *Modern Pathology*, 20(4):467–473, 2007.

[35] Andreas Pettersson, Rebecca E Graff, Scott R Bauer, Michael J Pitt, Rosina T Lis, Edward C Stack, Neil E Martin, Lauren Kunz, Kathryn L Penney, Azra H Ligon, et al. The *TMPRSS2-ERG* rearrangement, *ERG* expression, and prostate cancer outcomes: a cohort study and meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*, 21(9):1497–1509, 2012.

[36] Martin Spahn, Silvan Boxler, Steven Joniau, Marco Moschini, Bertrand Tombal, and R Jeffrey Karnes. What is the need for prostatic biomarkers in prostate cancer management? *Current Urology Reports*, 16(10):1–7, 2015.

[37] Bin Hu, Hongmei Yang, and Hongwei Yang. Diagnostic value of urine prostate cancer antigen 3 test using a cutoff value of 35 μg/l in patients with prostate cancer. *Tumor Biology*, pages 1–8, 2014.

[38] Jack Cuzick, Gregory P Swanson, Gabrielle Fisher, Arthur R Brothman, Daniel M Berney, Julia E Reid, David Mesher, VO Speights, Elzbieta Stankiewicz, Christopher S Foster, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *The Lancet Oncology*, 12(3):245–255, 2011.

[39] Matthew Cooperberg, Jeffry Simko, Sara Falzarano, Tara Maddala, June Chan, Janet Cowan, Cristina Magi-Galluzzi, Athanasios Tsiatis, Imelda Tenggara-Hunter, Dejan Knezevic, et al. Development and validation of the biopsy-based genomic prostate score (GPS) as a predictor of high grade or extracapsular prostate cancer to improve patient selection for active surveillance. *J Urol*, 189 (4):e873, 2013.

[40] Nicholas Erho, Anamaria Crisan, Ismael A Vergara, Anirban P Mitra, Mercedeh Ghadessi, Christine Buerki, Eric J Bergstralh, Thomas Kollmeyer, Stephanie Fink, Zaid Haddad, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PloS One*, 8(6):e66855, 2013.

[41] John Kuriyan, Boyana Konforti, and David Wemmer. *The molecules of life: Physical and chemical principles*. Garland Science, Taylor & Francis Group, 2012. ISBN 9780815341888.

[42] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258): 561–563, 1970.

[43] Wikipedia. Alternative splicing. https://en.wikipedia.org/wiki/Alternative_splicing, 2016. Accessed June 2016.

[44] Karen Hopkin. The evolving definition of a genewith the discovery that nearly all of the genome is transcribed, the definition of a “gene” needs another revision. *BioScience*, 59(11):928, 2009. doi: 10.1525/bio.2009.59.11.3. URL +http://dx.doi.org/10.1525/bio.2009.59.11.3.

[45] Ailin Zhang, Jiawei Zhang, Arja Kaipainen, Jared M Lucas, and Hong Yang. Long non-coding rna: A newly deciphered “code” in prostate cancer. *Cancer letters*, 375(2):323–330, 2016.

[46] Anna L Walsh, Alexandra V Tuzova, Eva M Bolton, Thomas H Lynch, and Antoinette S Perry. Long noncoding rnas and prostate carcinogenesis: the missing ‘linc’? *Trends in molecular medicine*, 20(8):428–436, 2014.

[47] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100 (1):57–70, 2000.

[48] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[49] Eric Lichtfouse. *Genetics, biofuels and local farming systems*, volume 7. Springer Science & Business Media, 2011.

[50] Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[51] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108, 2016.

[52] Stefan Fröhling and Hartmut Döhner. Chromosomal abnormalities in cancer. *New England Journal of Medicine*, 359(7):722–734, 2008.

[53] Michael F Berger, Michael S Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y Sivachenko, Andrea Sboner, Raquel Esgueva, Dorothee Pflueger, Carrie Sougnez, et al. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–220, 2011.

[54] Sylvan C Baca, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Mahmoud Ghandi, et al. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, 2013.

[55] J Rowly. A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243: 290–293, 1973.

[56] John M Goldman and Junia V Melo. Chronic myeloid leukemia—advances in biology and new approaches to treatment. *New England Journal of Medicine*, 349(15):1451–1464, 2003.

[57] Vuk Stambolic, Akira Suzuki, José Luis De La Pompa, Greg M Brothers, Christine Mirtsos, Takehiko Sasaki, Jurgen Ruland, Josef M Penninger, David P Siderovski, and Tak W Mak. Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor *PTEN*. *Cell*, 95(1):29–39, 1998.

[58] Bob Weinhold. Epigenetics: the science of change. *Environmental Health Perspectives*, 114(3):A160, 2006.

[59] D Simmons. Epigenetic influence and disease. *Nature Education*, 1(1):6, 2008.

[60] Fabrício F Costa, Valeria A Paixão, Felicia P Cavalher, Karina B Ribeiro, Isabela W Cunha, José Augusto Rinck, Michael O'Hare, Alan Mackay, Fernando A Soares, Ricardo R Brentani, et al. *SATR-1* hypomethylation is a common and early event in breast cancer. *Cancer Genetics and Cytogenetics*, 165(2):135–143, 2006.

[61] Jin Seuk Kim, Joungho Han, Young Mog Shim, Joobae Park, and Duk-Hwan Kim. Aberrant methylation of H-cadherin (*CDH13*) promoter is associated with tumor progression in primary nonsmall cell lung carcinoma. *Cancer*, 104(9): 1825–1833, 2005.

[62] CDC. Prostate cancer. http://www.cdc.gov/cancer/prostate/basic_info/what-is-prostate-cancer.htm, 2016. Accessed August 2016.

[63] Aibolita. The structure of the prostate. http://aibolita.com/mens-diseases/46533-the-structure-of-the-prostate.htm, 2016. Accessed June 2016.

[64] R Veltri, R Rodriguez, et al. Molecular biology, endocrinology, and physiology of the prostate and seminal vesicles. *Campbell-Walsh Urology: Saunders*, 85, 2007.

[65] Michael F Leitzmann and Sabine Rohrmann. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. *Clin Epidemiol*, 4: 1–11, 2012.

[66] Otis W Brawley. Prostate cancer epidemiology in the United States. *World Journal of Urology*, 30(2):195–200, 2012.

[67] Chad R Ritch, Belinda F Morrison, Greg Hruby, Kathleen C Coard, Richard Mayhew, William Aiken, Mitchell C Benson, and James M McKiernan. Pathological outcome and biochemical recurrence-free survival after radical prostatectomy in African-American, Afro-Caribbean (Jamaican) and Caucasian-American men: an international comparison. *BJU International*, 111(4b):E186–E190, 2013.

[68] David Schreiber, Eric B Levy, David Schwartz, Justin Rineer, Andrew Wong, Marvin Rotman, and Jeffrey P Weiss. Impact of race in a predominantly African-American population of patients with low/intermediate risk prostate cancer undergoing radical prostatectomy within an equal access care institution. *International Urology and Nephrology*, 46(10):1941–1946, 2014.

[69] Debasish Sundi, Ashley E Ross, Elizabeth B Humphreys, Misop Han, Alan W Partin, H Ballentine Carter, and Edward M Schaeffer. African American men with very low–risk prostate cancer exhibit adverse oncologic outcomes after

radical prostatectomy: should active surveillance still be an option for them? *Journal of Clinical Oncology*, 31(24):2991–2997, 2013.

[70] Kazuto Ito. Prostate cancer in Asian men. *Nature Reviews Urology*, 2014.

[71] Alice S Whittemore, Anna H Wu, Laurence N Kolonel, Esther M John, Richard P Gallagher, Geoffrey R Howe, Dee W West, Chong-Ze Teh, and Thomas Stamey. Family history and prostate cancer risk in black, white, and Asian men in the United States and Canada. *American Journal of Epidemiology*, 141(8):732–740, 1995.

[72] Kari Hemminki. Familial risk and familial survival in prostate cancer. *World Journal of Urology*, 30(2):143–148, 2012.

[73] Leslie K Dennis and Deborah V Dawson. Meta-analysis of measures of sexual activity and prostate cancer. *Epidemiology*, 13(1):72–79, 2002.

[74] Marcia L Taylor, AG Mainous, Brian J Wells, et al. Prostate cancer and sexually transmitted diseases: a meta-analysis. *Family Medicine–Kansas City*, 37(7):506, 2005.

[75] Wen-Yi Huang, Richard Hayes, Ruth Pfeiffer, Raphael P Viscidi, Francis K Lee, Yun F Wang, Douglas Reding, Denise Whitby, John R Papp, and Charles S Rabkin. Sexually transmissible infections and prostate cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 17(9):2374–2381, 2008.

[76] Tarja Anttila, Leena Tenkanen, Sonja Lumme, Maija Leinonen, Randi Elin Gislefoss, Göran Hallmans, Steinar Thoresen, Timo Hakulinen, Tapio Luostarinen, Pär Stattin, et al. Chlamydial antibodies and risk of prostate cancer. *Cancer Epidemiology Biomarkers & Prevention*, 14(2):385–389, 2005.

[77] Zoltan Korodi, Joakim Dillner, Egil Jellum, Sonja Lumme, Göran Hallmans, Steinar Thoresen, Timo Hakulinen, Pär Stattin, Tapio Luostarinen, Matti Lehtinen, et al. Human papillomavirus 16, 18, and 33 infections and risk of prostate cancer: a Nordic nested case-control study. *Cancer Epidemiology Biomarkers & Prevention*, 14(12):2952–2955, 2005.

[78] Siobhan Sutcliffe, Edward Giovannucci, Charlotte A Gaydos, Raphael P Viscidi, Frank J Jenkins, Jonathan M Zenilman, Lisa P Jacobson, Angelo M De Marzo, Walter C Willett, and Elizabeth A Platz. Plasma antibodies against Chlamydia trachomatis, human papillomavirus, and human herpesvirus type 8 in relation to prostate cancer: a prospective study. *Cancer Epidemiology Biomarkers & Prevention*, 16(8):1573–1580, 2007.

[79] Melissa Y Wei and Edward L Giovannucci. Lycopene, tomato products, and prostate cancer incidence: a review and reassessment in the psa screening era. *Journal of Oncology*, 2012, 2012.

[80] Astrid Steinbrecher, Katharina Nimptsch, Anika Hüsing, Sabine Rohrmann, and Jakob Linseisen. Dietary glucosinolate intake and risk of prostate cancer in the EPIC-Heidelberg cohort study. *International Journal of Cancer*, 125(9): 2179–2186, 2009.

[81] Axel Heidenreich, Patrick J Bastian, Joaquim Bellmunt, Michel Bolla, Steven Joniau, Theodor van der Kwast, Malcolm Mason, Vsevolod Matveev, Thomas Wiegel, F Zattoni, et al. EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent—update 2013. *European Urology*, 65(1):124–137, 2014.

[82] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo LJ Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Hans Lilja, Marco Zappa, et al. Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360(13):1320–1328, 2009.

[83] Dragan Ilic, Denise O'Connor, Sally Green, and Timothy J Wilt. Screening for prostate cancer: an updated cochrane systematic review. *BJU International*, 107 (6):882–891, 2011.

[84] Shabbir MH Alibhai, Marc Leach, George Tomlinson, Murray D Krahn, Neil Fleshner, Eric Holowaty, and Gary Naglie. 30-day mortality and major complications after radical prostatectomy: influence of age and comorbidity. *Journal of the National Cancer Institute*, 97(20):1525–1532, 2005.

[85] William T Lowrance, Elena B Elkin, Lindsay M Jacks, David S Yee, Thomas L Jang, Vincent P Laudone, Bertrand D Guillonneau, Peter T Scardino, and James A Eastham. Comparative effectiveness of prostate cancer surgical treatments: a population based analysis of postoperative outcomes. *The Journal of Urology*, 183(4):1366–1372, 2010.

[86] Ian M Thompson, Donna K Pauler, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Howard L Parnes, Lori M Minasian, Leslie G Ford, Scott M Lippman, E David Crawford, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level $\leq$ 4.0 ng per millilitre. *New England Journal of Medicine*, 350(22):2239–2246, 2004.

[87] Jerome P Richie, William J Catalona, Frederick R Ahmann, A Hudson M'Liss, Peter T Scardino, Robert C Flanigan, Jean B Dekernion, Timothy L Ratliff, Louis R Kavoussi, Bruce L Dalkin, et al. Effect of patient age on early detection of prostate cancer with serum prostate-specific antigen and digital rectal examination. *Urology*, 42(4):365–374, 1993.

[88] NICE. Prostate cancer: diagnosis and management (clinical guideline). https://www.nice.org.uk/guidance/cg175/resources/prostate-cancer-diagnosis-and-management-35109753913285, 2014. Accessed May 2016.

[89] AV Taira, GS Merrick, RW Galbreath, H Andreini, W Taubenslag, R Curtis, WM Butler, E Adamovich, and KE Wallner. Performance of transperineal template-guided mapping biopsy in detecting prostate cancer in the initial and repeat biopsy setting. *Prostate Cancer and Prostatic Diseases*, 13(1):71–77, 2010.

[90] Donald F Gleason. Histologic grading of prostate cancer: a perspective. *Human Pathology*, 23(3):273–279, 1992.

[91] Jonathan I Epstein, William C Allsbrook Jr, Mahul B Amin, Lars L Egevad, ISUP Grading Committee, et al. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 29(9):1228–1242, 2005.

[92] Theresa Y Chan, Alan W Partin, Patrick C Walsh, and Jonathan I Epstein. Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy. *Urology*, 56(5):823–827, 2000.

[93] Kris K Rasiah, Phillip D Stricker, Anne-Maree Haynes, Warick Delprado, Jennifer J Turner, David Golovsky, Phillip C Brenner, Raji Kooner, Gordon F O'Neill, John J Grygiel, et al. Prognostic significance of Gleason pattern in patients with Gleason score 7 prostate carcinoma. *Cancer*, 98(12):2560–2565, 2003.

[94] David E Kang, Nicholas J Fitzsimons, Joseph C Presti, Christopher J Kane, Martha K Terris, William J Aronson, Christopher L Amling, Stephen J Freedland, SEARCH Database Study Group, et al. Risk stratification of men with Gleason score 7 to 10 tumors by primary and secondary Gleason score: results from the SEARCH database. *Urology*, 70(2):277–282, 2007.

[95] Jennifer R Stark, Sven Perner, Meir J Stampfer, Jennifer A Sinnott, Stephen Finn, Anna S Eisenstein, Jing Ma, Michelangelo Fiorentino, Tobias Kurth, Massimo Loda, et al. Gleason score and lethal prostate cancer: does 3+4=4+3? *Journal of Clinical Oncology*, 27(21):3459–3464, 2009.

[96] H Sobin Leslie, KG Mary, and W Christian. TNM classification of malignant tumours. *Aufl. UICC International Union Against Cancer*, 2009.

[97] Steven R Potter, Jonathan I Epstein, and Alan W Partin. Seminal vesicle invasion by prostate cancer: prognostic significance and therapeutic implications. *Rev Urol*, 2(3):190–195, 2000.

[98] Jonathan Ng, Aamer Mahmud, Brenda Bass, and Michael Brundage. Prognostic significance of lymphovascular invasion in radical prostatectomy specimens. *BJU International*, 110(10):1507–1514, 2012.

[99] Stacy Loeb, Kimberly A Roehl, Xiaoying Yu, Jo Ann V Antenor, Misop Han, Sara N Gashti, Ximing J Yang, and William J Catalona. Lymphovascular invasion in radical prostatectomy specimens: prediction of adverse pathologic features and biochemical progression. *Urology*, 68(1):99–103, 2006.

[100] JI Epstein, MJ Carmichael, G Pizov, and PC Walsh. Influence of capsular penetration on progression following radical prostatectomy: a study of 196 cases with long-term followup. *The Journal of Urology*, 150(1):135–141, 1993.

[101] JohnE Mcneal, RobertA Kindrachuk, FuadS Freiha, DavidG Bostwick, EliseA Redwine, and ThomasA Stamey. Patterns of progression in prostate cancer. *The Lancet*, 327(8472):60–63, 1986.

[102] TM Wheeler. Anatomic considerations in carcinoma of the prostate. *The Urologic Clinics of North America*, 16(4):623–634, 1989.

[103] Thomas M Wheeler, Özdal Dillioglugil, Michael W Kattan, Atsushi Arakawa, Shigehiro Soh, Kazuho Suyama, Makoto Ohori, and Peter T Scardino. Clinical and pathological significance of the level and extent of capsular invasion in clinical stage t1–2 prostate cancer. *Human Pathology*, 29(8):856–862, 1998.

[104] Liang Cheng, Michael F Darson, Erik J Bergstralh, Jeff Slezak, Robert P Myers, and David G Bostwick. Correlation of margin status and extraprostatic extension with progression of prostate carcinoma. *Cancer*, 86(9):1775–1782, 1999.

[105] Puay Hoon Tan, Liang Cheng, John R Srigley, David Griffiths, Peter A Humphrey, Theodore H Van Der Kwast, Rodolfo Montironi, Thomas M Wheeler, Brett Delahunt, Lars Egevad, et al. International Society of Urological Pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 5: surgical margins. *Modern Pathology*, 24(1):48–57, 2011.

[106] C Parker. Active surveillance: an individualized approach to early prostate cancer. *BJU International*, 92(1):2–3, 2003.

[107] Laurence Klotz, Liying Zhang, Adam Lam, Robert Nam, Alexandre Mamedov, and Andrew Loblaw. Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, 28(1):126–131, 2010.

[108] Frederik B Thomsen, Klaus Brasso, Laurence H Klotz, M Andreas Røder, Kasper D Berg, and Peter Iversen. Active surveillance for clinically localized prostate cancer—-a systematic review. *Journal of Surgical Oncology*, 109(8): 830–835, 2014.

[109] Mark S Soloway, Cynthia T Soloway, Steve Williams, Rajinikanth Ayyathurai, Bruce Kava, and Murugesan Manoharan. Active surveillance; a reasonable management alternative for patients with prostate cancer: the Miami experience. *BJU International*, 101(2):165–169, 2008.

[110] Roderick CN van den Bergh, Stijn Roemeling, Monique J Roobol, Gunnar Aus, Jonas Hugosson, Antti S Rannikko, Teuvo L Tammela, Chris H Bangma, and Fritz H Schröder. Outcomes of men with screen-detected prostate cancer eligible for active surveillance who were managed expectantly. *European Urology*, 55 (1):1–8, 2009.

[111] Fernando J Bianco, Peter T Scardino, and James A Eastham. Radical prostatectomy: long-term cancer control and recovery of sexual and urinary function ("trifecta"). *Urology*, 66(5):83–94, 2005.

[112] Misop Han, Alan W Partin, Charles R Pound, Jonathan I Epstein, and Patrick C Walsh. Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy: the 15-year Johns Hopkins experience. *Urologic Clinics of North America*, 28(3):555–565, 2001.

[113] Fausto Petrelli, Ivano Vavassori, Andrea Coinu, Karen Borgonovo, Enrico Sarti, and Sandro Barni. Radical prostatectomy or radiotherapy in high-risk prostate cancer: a systematic review and metaanalysis. *Clinical Genitourinary Cancer*, 12(4):215–224, 2014.

[114] Maxine Sun, Jesse D Sammon, Andreas Becker, Florian Roghmann, Zhe Tian, Simon P Kim, Alexandre Larouche, Firas Abdollah, Jim C Hu, Pierre I Karakiewicz, et al. Radical prostatectomy vs radiotherapy vs observation among older patients with clinically localized prostate cancer: a comparative effectiveness evaluation. *BJU International*, 113(2):200–208, 2014.

[115] Juanita Mary Crook, Alfonso Gomez-Iturriaga, Kris Wallace, Clement Ma, Sharon Fung, Shabbir Alibhai, Michael Jewett, and Neil Fleshner. Comparison of health-related quality of life 5 years after SPIRIT: Surgical Prostatectomy Versus Interstitial Radiation Intervention Trial. *Journal of Clinical Oncology*, 29 (4):362–368, 2011.

[116] Frank Peinemann, Ulrich Grouven, Carmen Bartel, Stefan Sauerland, Holger Borchers, Michael Pinkawa, Axel Heidenreich, and Stefan Lange. Permanent interstitial low-dose-rate brachytherapy for patients with localised prostate cancer: a systematic review of randomised and nonrandomised controlled clinical trials. *European Urology*, 60(5):881–893, 2011.

[117] Charles R Pound, Alan W Partin, Mario A Eisenberger, Daniel W Chan, Jay D Pearson, and Patrick C Walsh. Natural history of progression after PSA elevation following radical prostatectomy. *JAMA*, 281(17):1591–1597, 1999.

[118] Chawnshang Chang. *Androgens and Androgen Receptor: Mechanisms, Functions, and Clinical Applications*. Springer Science & Business Media, 2012.

[119] American Cancer Society. Hormone therapy for prostate cancer. http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-treating-hormone-therapy, 2016. Accessed May 2016.

[120] Ravi J Kumar, Al Barqawi, and E David Crawford. Adverse events associated with hormonal therapy for prostate cancer. *Reviews in Urology*, 7(Suppl 5):S37, 2005.

[121] Michel Bolla, Dionisio Gonzalez, Padraig Warde, Jean Bernard Dubois, René-Olivier Mirimanoff, Guy Storme, Jacques Bernier, Abraham Kuten, Cora Sternberg, Thierry Gil, et al. Improved survival in patients with locally advanced prostate cancer treated with radiotherapy and goserelin. *New England Journal of Medicine*, 337(5):295–300, 1997.

[122] Michel Bolla, Laurence Collette, Léo Blank, Padraig Warde, Jean Bernard Dubois, René-Olivier Mirimanoff, Guy Storme, Jacques Bernier, Abraham Kuten, Cora Sternberg, et al. Long-term results with immediate androgen suppression and external irradiation in patients with locally advanced prostate cancer (an EORTC study): a phase III randomised trial. *The Lancet*, 360(9327):103–108, 2002.

[123] SJM Hotte and F Saad. Current management of castrate-resistant prostate cancer. *Current Oncology*, 17:S72–S79, 2010.

[124] Nima Sharifi, William L Dahut, Seth M Steinberg, William D Figg, Christopher Tarassoff, Philip Arlen, and James L Gulley. A retrospective study of the time to clinical endpoints for advanced prostate cancer. *BJU International*, 96(7): 985–989, 2005.

[125] C Parker, S Nilsson, Daniel Heinrich, Svein I Helle, JM O'sullivan, Sophie D Fosså, Aleš Chodacki, Paweł Wiechno, John Logue, M Seke, et al. Alpha emitter radium-223 and survival in metastatic prostate cancer. *New England Journal of Medicine*, 369(3):213–223, 2013.

[126] Jason St John, Katelyn Powell, M Katie Conley-LaComb, and Sreenivasa R Chinni. *TMPRSS2-ERG* fusion gene expression in prostate tumor cells and its clinical and biological significance in prostate cancer progression. *Journal of Cancer Science & Therapy*, 4(4):94, 2012.

[127] Maisa Yoshimoto, Anthony M Joshua, Isabela W Cunha, Renata A Coudry, Francisco P Fonseca, Olga Ludkovski, Maria Zielenska, Fernando A Soares, and Jeremy A Squire. Absence of *TMPRSS2*: *ERG* fusions and *PTEN* losses in prostate cancer is associated with a favorable outcome. *Modern Pathology*, 21 (12):1451–1460, 2008.

[128] Maisa Yoshimoto, Anthony M Joshua, Susan Chilton-MacNeill, Jane Bayani, Shamini Selvarajah, Andrew J Evans, Maria Zielenska, and Jeremy A Squire. Three-color FISH analysis of TMPRSS2/ERG fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. *Neoplasia*, 8(6):465–469, 2006.

[129] Cristina Magi-Galluzzi, Toyonori Tsusuki, Paul Elson, Kelly Simmerman, Chris LaFargue, Raquel Esgueva, Eric Klein, Mark A Rubin, and Ming Zhou. *TMPRSS2-ERG* gene fusion prevalence and class are significantly different in prostate cancer of caucasian, african-american and japanese patients. *The Prostate*, 71(5):489–497, 2011.

[130] Juan-Miguel Mosquera, Rohit Mehra, Meredith M Regan, Sven Perner, Elizabeth M Genega, Gerri Bueti, Rajal B Shah, Sandra Gaston, Scott A Tomlins, John T Wei, et al. Prevalence of *TMPRSS2-ERG* fusion prostate cancer among men undergoing prostate biopsy in the united states. *Clinical Cancer Research*, 15(14):4706–4711, 2009.

[131] Sven Perner, Francesca Demichelis, Rameen Beroukhim, Folke H Schmidt, Juan-Miguel Mosquera, Sunita Setlur, Joelle Tchinda, Scott A Tomlins, Matthias D Hofer, Kenneth G Pienta, et al. *TMPRSS2-ERG* fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer Research*, 66 (17):8337–8341, 2006.

[132] Scott A Tomlins, Rohit Mehra, Daniel R Rhodes, Lisa R Smith, Diane Roulston, Beth E Helgeson, Xuhong Cao, John T Wei, Mark A Rubin, Rajal B Shah, et al. *TMPRSS2-ETV4* gene fusions define a third molecular subtype of prostate cancer. *Cancer Research*, 66(7):3396–3400, 2006.

[133] Biaoyang Lin, Camari Ferguson, James T White, Shunyou Wang, Robert Vessella, Lawrence D True, Leroy Hood, and Peter S Nelson. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease *TMPRSS2*. *Cancer Research*, 59(17):4180–4184, 1999.

[134] Arun Seth and Dennis K Watson. *ETS* transcription factors and their emerging roles in human cancer. *European Journal of Cancer*, 41(16):2462–2478, 2005.

[135] Antonio Fernández-Serra, Luis Rubio, Ana Calatrava, José Rubio-Briones, Rocio Salgado, Rosario Gil-Benso, Blanca Espinet, Zaida García-Casado, and José Antonio López-Guerrero. Molecular characterization and clinical impact of *TMPRSS2-ERG* rearrangement on prostate cancer: comparison between FISH and RT-PCR. *BioMed Research International*, 2013, 2013.

[136] Delila Gasi Tandefelt, Joost Boormans, Karin Hermans, and Jan Trapman. *ETS* fusion genes in prostate cancer. *Endocrine-related Cancer*, 21(3):R143–R152, 2014.

[137] Jeremy P Clark and Colin S Cooper. *ETS* gene fusions in prostate cancer. *Nature Reviews Urology*, 6(8):429–439, 2009.

[138] Jianghua Wang, Yi Cai, Chengxi Ren, and Michael Ittmann. Expression of variant *TMPRSS2/ERG* fusion messenger RNAs is associated with aggressive prostate cancer. *Cancer Research*, 66(17):8347–8351, 2006.

[139] Francesca Demichelis, K Fall, S Perner, Ove Andrén, F Schmidt, SR Setlur, Y Hoshida, JM Mosquera, Y Pawitan, C Lee, et al. *TMPRSS2-ERG* gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene*, 26(31):4596–4599, 2007.

[140] RK Nam, L Sugar, W Yang, S Srivastava, LH Klotz, LY Yang, A Stanimirovic, E Encioiu, M Neill, DA Loblaw, et al. Expression of the *TMPRSS2-ERG* fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *British Journal of Cancer*, 97(12):1690–1695, 2007.

[141] Alan Dal Pra, Emilie Lalonde, Jenna Sykes, Fiona Warde, Adrian Ishkanian, Alice Meng, Chad Maloff, John Srigley, Anthony M Joshua, Gyorgy Petrovics, et al. *TMPRSS2-ERG* status is not prognostic following prostate cancer radiotherapy: implications for fusion status and dsb repair. *Clinical Cancer Research*, 19(18):5202–5209, 2013.

[142] Sopheap Phin, Mathew W Moore, and Philip D Cotter. Genomic rearrangements of *PTEN* in prostate cancer. *Front Oncol*, 3(240):1–9, 2013.

[143] Jin-Tang Dong, Chang-Ling Li, Tavis W Sipe, and Henry F Frierson. Mutations of *PTEN/MMAC1* in primary prostate cancers from Chinese patients. *Clinical Cancer Research*, 7(2):304–308, 2001.

[144] Jin-Tang Dong. Prevalent mutations in prostate cancer. *Journal of Cellular Biochemistry*, 97(3):433–447, 2006.

[145] Paul Cairns, Kenji Okami, Sarel Halachmi, Naomi Halachmi, Manel Esteller, James G Herman, Jin Jen, WB Isaacs, G Steven Bova, and David Sidransky. Frequent inactivation of *PTEN/MMAC1* in primary prostate cancer. *Cancer Research*, 57(22):4997–5000, 1997.

[146] Steven I Wang, Ramon Parsons, and Michael Ittmann. Homozygous deletion of the *PTEN* tumor suppressor gene in a subset of prostate adenocarcinomas. *Clinical Cancer Research*, 4(3):811–815, 1998.

[147] M Yoshimoto, IW Cunha, RA Coudry, FP Fonseca, CH Torres, FA Soares, and JA Squire. FISH analysis of 107 prostate cancers shows that *PTEN* genomic deletion is associated with poor clinical outcome. *British Journal of Cancer*, 97 (5):678–685, 2007.

[148] PCMS Verhagen, PW Van Duijn, KGL Hermans, LHJ Looijenga, RJHLM van Gurp, Hans Stoop, TH Van der Kwast, and Jan Trapman. The *PTEN* gene in locally progressive prostate cancer is preferentially inactivated by bi-allelic gene deletion. *The Journal of Pathology*, 208(5):699–707, 2006.

[149] Zahra Rafiei Fallahabadi, Mohammad Reza Noori Daloii, Reza Mahdian, Farhkondeh Behjati, Mohamad Ali Shokrgozar, Maryam Abolhasani, Mojgan Asgari, and Hossein Shahrokh. Frequency of *PTEN* alterations, *TMPRSS2-ERG* fusion and their association in prostate cancer. *Gene*, 575(2):755–760, 2016.

[150] Kanishka Sircar, Maisa Yoshimoto, Federico A Monzon, Ismael H Koumakpayi, Ruth L Katz, Abha Khanna, Karla Alvarez, Guanyong Chen, Andrew D Darnel, Armen G Aprikian, et al. *PTEN* genomic deletion is associated with p-Akt and *AR* signalling in poorer outcome, hormone refractory prostate cancer. *The Journal of Pathology*, 218(4):505–513, 2009.

[151] Ismail Turker Koksal, Ercument Dirice, Duygu Yasar, Ahter D Sanlioglu, Akif Ciftcioglu, Kemal H Gulkesen, Nidai O Ozes, Mehmet Baykara, Guven Luleci, and Salih Sanlioglu. The assessment of *PTEN* tumor suppressor gene in combination with Gleason scoring and serum PSA to evaluate progression of prostate carcinoma. In *Urologic Oncology: Seminars and Original Investigations*, volume 22, pages 307–312. Elsevier, 2004.

[152] AHM Reid, Gerhardt Attard, Laurence Ambroisine, Gabrielle Fisher, Gyula Kovacs, Daniel Brewer, Jeremy Clark, Penny Flohr, Sandra Edwards, Daniel M Berney, et al. Molecular characterisation of *ERG*, *ETV1* and *PTEN* gene loci identifies patients at low and high risk of death from prostate cancer. *British Journal of Cancer*, 102(4):678–684, 2010.

[153] Tarek A Bismar, Maisa Yoshimoto, Robin T Vollmer, Qiuli Duan, Matthew Firszt, Jacques Corcos, and Jeremy A Squire. *PTEN* genomic deletion is an early event associated with *ERG* gene rearrangements in prostate cancer. *BJU International*, 107(3):477–485, 2011.

[154] Debashis Sarker, Alison HM Reid, Timothy A Yap, and Johann S de Bono. Targeting the PI3K/AKT pathway for the treatment of prostate cancer. *Clinical Cancer Research*, 15(15):4799–4805, 2009.

[155] Scott A Tomlins, Daniel R Rhodes, Jianjun Yu, Sooryanarayana Varambally, Rohit Mehra, Sven Perner, Francesca Demichelis, Beth E Helgeson, Bharathi Laxman, David S Morris, et al. The role of *SPINK1* in *ETS* rearrangement-negative prostate cancers. *Cancer Cell*, 13(6):519–528, 2008.

[156] Katri A Leinonen, Teemu T Tolonen, Hazel Bracken, Ulf-Håkan Stenman, Teuvo LJ Tammela, Outi R Saramäki, and Tapio Visakorpi. Association of *SPINK1* expression and *TMPRSS2*: *ERG* fusion with prognosis in endocrine-treated prostate cancer. *Clinical Cancer Research*, 16(10):2845–2851, 2010.

[157] Richard Flavin, Andreas Pettersson, Whitney K Hendrickson, Michelangelo Fiorentino, Stephen Finn, Lauren Kunz, Gregory L Judson, Rosina Lis, Dyane Bailey, Christopher Fiore, et al. *SPINK1* protein expression and prostate cancer progression. *Clinical Cancer Research*, 20(18):4904–4911, 2014.

[158] Scott A Tomlins, Bharathi Laxman, Sooryanarayana Varambally, Xuhong Cao, Jindan Yu, Beth E Helgeson, Qi Cao, John R Prensner, Mark A Rubin, Rajal B Shah, et al. Role of the *TMPRSS2-ERG* gene fusion in prostate cancer. *Neoplasia*, 10(2):177–IN9, 2008.

[159] Christopher E Barbieri, Sylvan C Baca, Michael S Lawrence, Francesca Demichelis, Mirjam Blattner, Jean-Philippe Theurillat, Thomas A White, Petar Stojanov, Eliezer Van Allen, Nicolas Stransky, et al. Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nature Genetics*, 44(6): 685–689, 2012.

[160] Mirjam Blattner, Daniel J Lee, Catherine O'Reilly, Kyung Park, Theresa Y MacDonald, Francesca Khani, Kevin R Turner, Ya-Lin Chiu, Peter J Wild, Igor Dolgalev, et al. *SPOP* mutations in prostate cancer across demographically diverse patient cohorts. *Neoplasia*, 16(1):14–W10, 2014.

[161] Marion JG Bussemakers, Adrie van Bokhoven, Gerald W Verhaegh, Frank P Smit, Herbert FM Karthaus, Jack A Schalken, Frans MJ Debruyne, Ning Ru, and William B Isaacs. *DD3*: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Research*, 59(23):5975–5979, 1999.

[162] Alexandre de la Taille, Jacques Irani, Markus Graefen, Felix Chun, Theo de Reijke, Paul Kil, Paolo Gontero, Alain Mottaz, and Alexander Haese. Clinical evaluation of the *PCA3* assay in guiding initial biopsy decisions. *The Journal of Urology*, 185(6):2119–2125, 2011.

[163] Daphne Hessels, Martijn PMQ van Gils, Onno van Hooij, Sander A Jannink, J Alfred Witjes, Gerald W Verhaegh, and Jack A Schalken. Predictive value of *PCA3* in urinary sediments in determining clinico-pathological characteristics of prostate cancer. *The Prostate*, 70(1):10–16, 2010.

[164] Leonard S Marks, Yves Fradet, Ina Lim Deras, Amy Blase, Jeannette Mathis, Sheila MJ Aubin, Anthony T Cancio, Marie Desaulniers, William J Ellis, Harry Rittenhouse, et al. *PCA3* molecular urine assay for prostate cancer in men undergoing repeat biopsy. *Urology*, 69(3):532–535, 2007.

[165] Ina L Deras, Sheila MJ Aubin, Amy Blase, John R Day, Seongjoon Koo, Alan W Partin, William J Ellis, Leonard S Marks, Yves Fradet, Harry Rittenhouse, et al. *PCA3*: a molecular urine assay for predicting prostate biopsy outcome. *The Journal of Urology*, 179(4):1587–1592, 2008.

[166] CRUK. The PCA3 test for prostate cancer. http://www.cancerresearchuk.org/about-cancer/cancers-in-general/cancer-questions/the-pca3-test-for-prostate-cancer, 2016. Accessed May 2016.

[167] Jun Luo, Shan Zha, Wesley R Gage, Thomas A Dunn, Jessica L Hicks, Christina J Bennett, Charles M Ewing, Elizabeth A Platz, Sacha Ferdinandusse, Ronald J Wanders, et al. α-Methylacyl-CoA Racemase a new molecular marker for prostate cancer. *Cancer Research*, 62(8):2220–2226, 2002.

[168] Mark A Rubin, Ming Zhou, Saravana M Dhanasekaran, Sooryanarayana Varambally, Terrence R Barrette, Martin G Sanda, Kenneth J Pienta, Debashis Ghosh, and Arul M Chinnaiyan. α-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA*, 287(13):1662–1670, 2002.

[169] Shan Zha, Sacha Ferdinandusse, Simone Denis, Ronald J Wanders, Charles M Ewing, Jun Luo, Angelo M De Marzo, and William B Isaacs. α-Methylacyl-CoA racemase as an androgen-independent growth modifier in prostate cancer. *Cancer Research*, 63(21):7365–7376, 2003.

[170] Zhong Jiang, Bruce A Woda, and Ximing J Yang. α-Methylacyl coenzyme A racemase as a marker for prostate cancer. *JAMA*, 287(23):3080–3081, 2002.

[171] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13(6):1977–2000, 2002.

[172] Jonathan D Mosley and Ruth A Keri. Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. *BMC Medical Genomics*, 1(1):1, 2008.

[173] Kerby Shedden, Jeremy MG Taylor, Steven A Enkemann, Ming-Sound Tsao, Timothy J Yeatman, William L Gerald, Steven Eschrich, Igor Jurisica, Thomas J Giordano, David E Misek, et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14(8):822–827, 2008.

[174] Jing Zhang, Bing Liu, Xingpeng Jiang, Huizhi Zhao, Ming Fan, Zhenjie Fan, J Jack Lee, Tao Jiang, Tianzi Jiang, and Sonya Wei Song. A systems biology-based gene expression classifier of glioblastoma predicts survival with solid tumors. *PLoS One*, 4(7):e6274, 2009.

[175] Dejan Knezevic, Audrey D Goddard, Nisha Natraj, Diana B Cherbavaz, Kim M Clark-Langone, Jay Snable, Drew Watson, Sara M Falzarano, Cristina Magi-Galluzzi, Eric A Klein, et al. Analytical validation of the Oncotype DX prostate cancer assay–a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genomics*, 14(1):690, 2013.

[176] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[177] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Saldaña. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine*, 13(9/10):527, 2007.

[178] Affymetrix. GeneChip Human Exon ST Array. http://www.affymetrix.com/estore/browse/products.jsp?productId=131452&navMode=34000&navAction=jump&aId=productsNav#1_1, 2016. Accessed July 2016.

[179] Affymetrix. Exon and gene array glossary. http://www.affymetrix.com/support/help/exon_glossary/index.affx, 2016. Accessed May 2016.

[180] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, Terence P Speed, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[181] B.M. Bolstad. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. PhD thesis, University of California, 2004.

[182] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010.

[183] Yi Qu, Fei He, and Yuchen Chen. Different effects of the probe summarization algorithms PLIER and RMA on high-level analysis of Affymetrix exon arrays. *BMC Bioinformatics*, 11(1):1, 2010.

[184] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[185] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

[186] F. Mosteller and J.W. Tukey. *Data analysis and regression: a second course in statistics*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co., 1977. ISBN 9780201048544.

[187] Affymetrix. Quality assessment of exon and gene arrays. http://media.affymetrix.com/support/technical/whitepapers/exon_gene_arrays_qa_whitepaper.pdf, 2016. Accessed July 2016.

[188] Caitlin Smith. FFPE or frozen? working with human clinical samples. http://www.biocompare.com/Editorial-Articles/168948-FFPE-or-Frozen-Working-with-Human-Clinical-Samples/, 2016. Accessed June 2016.

[189] Michal R Schweiger, Martin Kerick, Bernd Timmermann, Marcus W Albrecht, Tatjana Borodina, Dmitri Parkhomchuk, Kurt Zatloukal, and Hans Lehrach. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number-and mutation-analysis. *PloS One*, 4(5):e5548, 2009.

[190] Jakob Hedegaard, Kasper Thorsen, Mette Katrine Lund, Anne-Mette K Hein, Stephen Jacques Hamilton-Dutoit, Søren Vang, Iver Nordentoft, Karin Birkenkamp-Demtröder, Mogens Kruhøffer, Henrik Hager, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PloS One*, 9(5):e98187, 2014.

[191] David Manson-Bahr, Richard Ball, Gunes Gundem, Krishna Sethia, Robert Mills, Mark Rochester, Victoria Goody, Elizabeth Anderson, Sarah O'Meara, Marcus Flather, et al. Mutation detection in formalin-fixed prostate cancer biopsies taken

at the time of diagnosis using next-generation DNA sequencing. *Journal of Clinical Pathology*, pages jclinpath–2014, 2015.

[192] Florenza Lüder Ripoli, Annika Mohr, Susanne Conradine Hammer, Saskia Willenbrock, Marion Hewicker-Trautwein, Silvia Hennecke, Hugo Murua Escobar, and Ingo Nolte. A comparison of fresh frozen vs. formalin-fixed, paraffin-embedded specimens of canine mammary tumors via branched-DNA assay. *International Journal of Molecular Sciences*, 17(5):724, 2016.

[193] Yuker Wang, Victoria EH Carlton, George Karlin-Neumann, Ronald Sapolsky, Li Zhang, Martin Moorhead, Zhigang C Wang, Andrea L Richardson, Robert Warren, Axel Walther, et al. High quality copy number and genotype data from FFPE samples using Molecular Inversion Probe (MIP) microarrays. *BMC Medical Genomics*, 2(1):1, 2009.

[194] Lorenza Mittempergher, Jorma J De Ronde, Marja Nieuwland, Ron M Kerkhoven, Iris Simon, J Th Emiel, Lodewyk FA Wessels, Laura J Van't Veer, et al. Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue. *PloS One*, 6 (2):e17163, 2011.

[195] Ian A Cree. Principles of cancer cell culture. *Cancer Cell Culture: Methods and Protocols*, pages 13–26, 2011.

[196] Simon P Langdon. Basic principles of cancer cell culture. *Cancer Cell Culture: Methods and Protocols*, pages 3–15, 2004.

[197] GOea Gey, Ward D Coffman, and Mart T Kubicek. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. In *Cancer Research*, volume 12, pages 264–265. AACR, 1952.

[198] Gurvinder Kaur and Jannette M Dufour. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*, 2(1):1–5, 2012.

[199] Douglas T Ross and Charles M Perou. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Disease Markers*, 17(2):99–109, 2001.

[200] Douglas T Ross, Uwe Scherf, Michael B Eisen, Charles M Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S Jeffrey, Matt Van de Rijn, Mark Waltham, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235, 2000.

[201] Silvia Domcke, Rileen Sinha, Douglas A Levine, Chris Sander, and Nikolaus Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*, 4, 2013.

[202] Hua Li, John S Wawrose, William E Gooding, Levi A Garraway, Vivian Wai Yan Lui, Noah D Peyser, and Jennifer R Grandis. Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: a rational approach to preclinical model selection. *Molecular Cancer Research*, 12(4):571–582, 2014.

[203] Dmitri Mouradov, Clare Sloggett, Robert N Jorissen, Christopher G Love, Shan Li, Antony W Burgess, Diego Arango, Robert L Strausberg, Daniel Buchanan, Samuel Wormald, et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Research*, 74(12): 3238–3247, 2014.

[204] William M Lin, Alissa C Baker, Rameen Beroukhim, Wendy Winckler, Whei Feng, Jennifer M Marmion, Elisabeth Laine, Heidi Greulich, Hsiuyi Tseng, Casey Gates, et al. Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Research*, 68(3):664–673, 2008.

[205] Martin L Sos, Kathrin Michel, Thomas Zander, Jonathan Weiss, Peter Frommolt, Martin Peifer, Danan Li, Roland Ullrich, Mirjam Koker, Florian Fischer, et al. Predicting drug susceptibility of non–small cell lung cancers based on genetic lesions. *The Journal of Clinical Investigation*, 119(6):1727–1740, 2009.

[206] Andrew Goodspeed, Laura M Heiser, Joe W Gray, and James C Costello. Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research*, 14(1):3–13, 2016.

[207] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2006.

[208] Oded Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer, 2005.

[209] Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(2):143–156, 2005.

[210] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[211] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[212] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[213] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[214] Leo Breiman and Adele Cutler. Random forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, 2016. Accessed July 2016.

[215] Tom M Mitchell et al. Machine learning. WCB, 1997.

[216] Leo Breiman. Out-of-bag estimation. Technical report, Citeseer, 1996.

[217] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[218] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[219] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

[220] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[221] David G Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006.

[222] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

[223] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

[224] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11):e79217, 2013.

[225] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[226] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[227] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[228] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2014.

[229] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, 2016.

[230] Gene Ontology Consortium. GO enrichment analysis. http://geneontology.org/page/go-enrichment-analysis, 2016. Accessed September 2016.

[231] John Groffen, John R Stephenson, Nora Heisterkamp, Annelies de Klein, Claus R Bartram, and Gerard Grosveld. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell*, 36(1):93–99, 1984.

[232] David S Rickman, Dorothee Pflueger, Benjamin Moss, Vanessa E VanDoren, Chen X Chen, Alexandre de la Taille, Rainer Kuefer, Ashutosh K Tewari, Sunita R Setlur, Francesca Demichelis, et al. *SLC45A3-ELK4* is a novel and frequent erythroblast transformation–specific fusion transcript in prostate cancer. *Cancer Research*, 69(7):2734–2738, 2009.

[233] Kasper Thorsen, Troels Schepeler, Bodil Øster, Mads H Rasmussen, Søren Vang, Kai Wang, Kristian Q Hansen, Philippe Lamy, Jakob Skou Pedersen, Asger Eller, et al. Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis. *BMC Genomics*, 12(1):1, 2011.

[234] Quan Lei, Cong Li, Zhixiang Zuo, Chunhua Huang, Hanhua Cheng, and Rongjia Zhou. Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biology and Evolution*, 8(3):562–577, 2016.

[235] Sameer Jhavar, Alison Reid, Jeremy Clark, Zsofia Kote-Jarai, Timothy Christmas, Alan Thompson, Christopher Woodhouse, Christopher Ogden, Cyril Fisher, Cathy Corbishley, et al. Detection of *TMPRSS2-ERG* translocations in human prostate cancer by expression profiling using GeneChip Human Exon 1.0 ST arrays. *The Journal of Molecular Diagnostics*, 10(1):50–57, 2008.

[236] Eva Lin, Li Li, Yinghui Guan, Robert Soriano, Celina Sanchez Rivers, Sankar Mohan, Ajay Pandita, Jerry Tang, and Zora Modrusan. Exon array profiling detects *EML4-ALK* fusion in breast, colorectal, and non–small cell lung cancers. *Molecular Cancer Research*, 7(9):1466–1476, 2009.

[237] Fei Li, Yan Feng, Rong Fang, Zhaoyuan Fang, Jufeng Xia, Xiangkun Han, Xin-Yuan Liu, Haiquan Chen, Hongyan Liu, and Hongbin Ji. Identification of *RET* gene fusion by exon array analyses in “pan-negative” lung cancer from never smokers. *Cell Research*, 22(5):928, 2012.

[238] Craig P Giacomini, Steven Sun, Sushama Varma, A Hunter Shain, Marilyn M Giacomini, Jay Balagtas, Robert T Sweeney, Everett Lai, Catherine A Del Vecchio, Andrew D Forster, et al. Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. *PLoS Genetics*, 9(4):e1003464, 2013.

[239] Lu Wang, Toru Motoi, Raya Khanin, Adam Olshen, Fredrik Mertens, Julia Bridge, Paola Dal Cin, Cristina R Antonescu, Samuel Singer, Meera Hameed, et al. Identification of a novel, recurrent *HEY1-NCOA2* fusion in mesenchymal chondrosarcoma based on a genome-wide screen of exon-level expression data. *Genes, Chromosomes and Cancer*, 51(2):127–139, 2012.

[240] Anne Y Warren, Hayley C Whitaker, Beverley Haynes, Trogon Sangan, Leigh-Anne McDuffus, Jonathan D Kay, and David E Neal. Method for sampling tissue for research which preserves pathological data in radical prostatectomy. *The Prostate*, 73(2):194–202, 2013.

[241] H Ross-Adams, AD Lamb, MJ Dunning, S Halim, J Lindberg, CM Massie, LA Egevad, R Russell, A Ramos-Montoya, SL Vowler, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. *EBioMedicine*, 2(9):1133–1144, 2015.

[242] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell*, 18(1):11–22, 2010.

[243] Affymetrix. Affymetrix cel data file format. http://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html, 2016. Accessed July 2016.

[244] M Madan Babu. Introduction to microarray data analysis. *Computational Genomics: Theory and Application*, pages 225–249, 2004.

[245] Affymetrix. Affymetrix power tools. http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx, 2016. Accessed July 2016.

[246] Adam I Riker, Steven A Enkemann, Oystein Fodstad, Suhu Liu, Suping Ren, Christopher Morris, Yaguang Xi, Paul Howell, Brandon Metge, Rajeev S Samant, et al. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Medical Genomics*, 1 (1):1, 2008.

[247] Joel A Malek, Alejandra Martinez, Eliane Mery, Gwenael Ferron, Ruby Huang, Christophe Raynaud, Eva Jouve, Jean-Paul Thiery, Denis Querleu, and Arash Rafii. Gene expression analysis of matched ovarian primary tumors and peritoneal metastasis. *Journal of Translational Medicine*, 10(1):1, 2012.

[248] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921, 2001.

[249] Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1): D710–D716, 2016.

[250] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequence project: update and current status. *Nucleic Acids Research*, 31(1):34–37, 2003.

[251] Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2015.

[252] Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):1, 2015.

[253] Tim Yates, Maintainer Tim Yates, Suggests RUnit, Bioinformatics biocViews Annotation, and OneChannel Microarray. Package ‘annmap’. 2011.

[254] Hervé Fournier and Antoine Vigneron. A deterministic algorithm for fitting a step function to a weighted point-set. *Information Processing Letters*, 2012.

[255] Hervé Fournier and Antoine Vigneron. Fitting a step function to a point set. In *Algorithms-ESA 2008*, pages 442–453. Springer, 2008.

[256] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[257] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[258] Steffen Durinck and James Bullard. *GenomeGraphs: Plotting genomic information from Ensembl*, 2015. R package version 1.30.0.

[259] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33 (suppl 1):D428–D432, 2005.

[260] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, 16(5):284–287, 2012.

[261] Joao D Barros-Silva, Paula Paulo, Anne Cathrine Bakken, Nuno Cerveira, Marthe Løvf, Rui Henrique, Carmen Jerönimo, Ragnhild A Lothe, Rolf Inge Skotheim, and Manuel R Teixeira. Novel 5' fusion partners of *ETV1* and *ETV4* in prostate cancer. *Neoplasia*, 15(7):720–IN6, 2013.

[262] Christopher A Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, Shujun Luo, Irina Khrebtukova, Terrence R Barrette, Catherine Grasso, Jindan Yu, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences*, 106(30):12353–12358, 2009.

[263] Dorothee Pflueger, Stéphane Terry, Andrea Sboner, Lukas Habegger, Raquel Esgueva, Pei-Chun Lin, Maria A Svensson, Naoki Kitabayashi, Benjamin J Moss, Theresa Y MacDonald, et al. Discovery of non-*ETS* gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Research*, 21 (1):56–67, 2011.

[264] Paula Paulo, João D Barros-Silva, Franclim R Ribeiro, João Ramalho-Carvalho, Carmen Jerónimo, Rui Henrique, Guro E Lind, Rolf I Skotheim, Ragnhild A Lothe, and Manuel R Teixeira. *FLI1* is a novel *ETS* transcription factor involved in gene fusions in prostate cancer. *Genes, Chromosomes and Cancer*, 51(3): 240–249, 2012.

[265] Serban Nacu, Wenlin Yuan, Zhengyan Kan, Deepali Bhatt, Celina Sanchez Rivers, Jeremy Stinson, Brock A Peters, Zora Modrusan, Kenneth Jung, Somasekar Seshagiri, et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Medical Genomics*, 4(1):1, 2011.

[266] Karin G Hermans, Hetty A van der Korput, Ronald van Marion, Dennis J van de Wijngaart, Angelique Ziel-van der Made, Natasja F Dits, Joost L Boormans, Theo H van der Kwast, Herman van Dekken, Chris H Bangma, et al. Truncated *ETV1*, fused to novel tissue-specific genes, and full-length *ETV1* in prostate cancer. *Cancer Research*, 68(18):7541–7549, 2008.

[267] Joachim Weischenfeldt, Ronald Simon, Lars Feuerbach, Karin Schlangen, Dieter Weichenhan, Sarah Minner, Daniela Wuttig, Hans-Jörg Warnatz, Henning Stehr,

Tobias Rausch, et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer cell*, 23(2): 159–170, 2013.

[268] MH Eileen Tan, Jun Li, H Eric Xu, Karsten Melcher, and Eu-leong Yong. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*, 36(1):3–23, 2015.

[269] Susan M Henshall, Lisa G Horvath, David I Quinn, Sarah A Eggleton, John J Grygiel, Phillip D Stricker, Andrew V Biankin, James G Kench, and Robert L Sutherland. Zinc-alpha2-glycoprotein expression as a predictor of metastatic prostate cancer following radical prostatectomy. *Journal of the National Cancer Institute*, 98(19):1420–1424, 2006.

[270] Woon Yong Jung, Chang Ohk Sung, Sang Hak Han, Kyungeun Kim, Misung Kim, Jae Y Ro, Mun Jung Kang, Hanjong Ahn, and Yong Mee Cho. *AZGP-1* immunohistochemical marker in prostate cancer: potential predictive marker of biochemical recurrence in post radical prostatectomy specimens. *Applied Immunohistochemistry & Molecular Morphology*, 22(9):652–657, 2014.

[271] James D Brooks, Wei Wei, Jonathan R Pollack, Robert B West, Jun Ho Shin, John B Sunwoo, Sarah J Hawley, Heidi Auman, Lisa F Newcomb, Jeff Simko, et al. Loss of expression of *AZGP1* is associated with worse clinical outcomes in a multi-institutional radical prostatectomy cohort. *The Prostate*, 2016.

[272] Christoph Burdelski, Sandra Kleinhans, Martina Kluth, Claudia Hube-Magg, Sarah Minner, Christina Koop, Markus Graefen, Hans Heinzer, Maria Christina Tsourlakis, Waldemar Wilczak, et al. Reduced *AZGP1* expression is an independent predictor of early PSA recurrence and associated with *ERG*-fusion positive and *PTEN* deleted prostate cancers. *International Journal of Cancer*, 138(5): 1199–1206, 2016.

[273] Hannah M Bruce, Phillip D Stricker, Ruta Gupta, Richard R Savdie, Anne-Maree Haynes, Kate L Mahon, Hui-Ming Lin, James G Kench, and Lisa G Horvath. Loss of *AZGP1* as a superior predictor of relapse in margin-positive localized prostate cancer. *The Prostate*, 2016.

[274] Chun-yu Huang, Jing-jing Zhao, Lin Lv, Yi-bing Chen, Yuan-fang Li, Shan-shan Jiang, Wei Wang, Ke Pan, Yan Zheng, Bai-wei Zhao, et al. Decreased expression of *AZGP1* is associated with poor prognosis in primary gastric cancer. *PloS One*, 8(7):e69155, 2013.

[275] Ken-ichi Takayama, Takashi Suzuki, Shuichi Tsutsumi, Tetsuya Fujimura, Satoru Takahashi, Yukio Homma, Tomohiko Urano, Hiroyuki Aburatani, and Satoshi Inoue. Integrative analysis of *FOXP1* function reveals a tumor-suppressive effect in prostate cancer. *Molecular Endocrinology*, 28(12):2012–2024, 2014.

[276] Antje Krohn, Annemarie Seidel, Lia Burkhardt, Frederic Bachmann, Malte Mader, Katharina Grupp, Till Eichenauer, Andreas Becker, Meike Adam, Markus Graefen, et al. Recurrent deletion of 3p13 targets multiple tumour suppressor genes and defines a distinct subgroup of aggressive *ERG* fusion-positive prostate cancers. *The Journal of Pathology*, 231(1):130–141, 2013.

[277] Maisa Yoshimoto, Olga Ludkovski, Dave DeGrace, Julia L Williams, Andrew Evans, Kanishka Sircar, Tarek A Bismar, Paulo Nuin, and Jeremy A Squire. *PTEN* genomic deletions that characterize aggressive prostate cancer originate close to segmental duplications. *Genes, Chromosomes and Cancer*, 51(2):149–160, 2012.

[278] Greg L Shaw, Hayley Whitaker, Marie Corcoran, Mark J Dunning, Hayley Luxton, Jonathan Kay, Charlie E Massie, Jodi L Miller, Alastair D Lamb, Helen Ross-Adams, et al. The early effects of rapid androgen deprivation on human prostate cancer. *European Urology*, 2015.

[279] Elena V Fernandez, Kelie M Reece, Ariel M Ley, Sarah M Troutman, Tristan M Sissung, Douglas K Price, Cindy H Chau, and William D Figg. Dual targeting of the androgen receptor and hypoxia-inducible factor 1α pathways synergistically inhibits castration-resistant prostate cancer cells. *Molecular Pharmacology*, 87 (6):1006–1012, 2015.

[280] L Li, Zhenkun Lou, and L Wang. The role of *FKBP5* in cancer aetiology and chemoresistance. *British Journal of Cancer*, 104(1):19–23, 2011.

[281] Gennadi V Glinsky, Anna B Glinskii, Andrew J Stephenson, Robert M Hoffman, and William L Gerald. Gene expression profiling predicts clinical outcome of prostate cancer. *The Journal of Clinical Investigation*, 113(6):913–923, 2004.

[282] Sooryanarayana Varambally, Jianjun Yu, Bharathi Laxman, Daniel R Rhodes, Rohit Mehra, Scott A Tomlins, Rajal B Shah, Uma Chandran, Federico A Monzon, Michael J Becich, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, 8(5): 393–406, 2005.

[283] Antonio Ramos-Montoya, Alastair D Lamb, Roslin Russell, Thomas Carroll, Sarah Jurmeister, Nuria Galeano-Dalmau, Charlie E Massie, Joan Boren, Helene Bon, Vasiliki Theodorou, et al. *HES6* drives a critical *AR* transcriptional programme to induce castration-resistant prostate cancer through activation of an *E2F1*-mediated cell cycle network. *EMBO Molecular Medicine*, page e201303581, 2014.

[284] Sungyong You, Beatrice S Knudsen, Nicholas Erho, Mohammed Alshalalfa, Mandeep Takhar, Hussam Al-deen Ashab, Elai Davicioni, R Jeffrey Karnes, Eric A Klein, Robert B Den, et al. Integrated classification of prostate cancer reveals a novel luminal subtype with poor outcome. *Cancer Research*, pages canres–0902, 2016.

[285] Andrew J Stephenson, Alex Smith, Michael W Kattan, Jaya Satagopan, Victor E Reuter, Peter T Scardino, and William L Gerald. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, 104(2):290–298, 2005.

[286] Eric A Klein, Kasra Yousefi, Zaid Haddad, Voleak Choeurng, Christine Buerki, Andrew J Stephenson, Jianbo Li, Michael W Kattan, Cristina Magi-Galluzzi, and Elai Davicioni. A genomic classifier improves prediction of metastatic disease within 5 years after surgery in node-negative high-risk prostate cancer patients

managed by radical prostatectomy without adjuvant therapy. *European Urology*, 67(4):778–786, 2015.

[287] Cancer Genome Atlas Research Network et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.

[288] Affymetrix. Affymetrix expression console. http://www.affymetrix.com/estore/browse/level_seven_software_products_only.jsp?productId=131414#1_1, 2016. Accessed August 2016.

[289] Mark J Dunning, Mike L Smith, Matthew E Ritchie, and Simon Tavaré. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, 23(16): 2183–2184, 2007.

[290] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1): 118–127, 2007.

[291] M Dunning, A Lynch, and M Eldridge. illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). *R package version*, 2(0), 2013.

[292] Nuno L Barbosa-Morais, Mark J Dunning, Shamith A Samarajiwa, Jeremy FJ Darot, Matthew E Ritchie, Andy G Lynch, and Simon Tavaré. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Research*, 38(3):e17–e17, 2010.

[293] M Carlson. hgu133a.db: Affymetrix Human Genome U133 set annotation data (chip hgu133a). *R package version 3.2.2.*, 2(0), 2013.

[294] TCGA. Data level. https://wiki.nci.nih.gov/display/TCGA/Data+level, 2016. Accessed August 2016.

[295] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):1, 2011.

[296] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15 (12):550, 2014.

[297] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2010.

[298] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, page gkv007, 2015.

[299] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/.

[300] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

[301] P Du and R Bourgon. methyanalysis: Dna methylation data analysis and visualization. *R package version*, 1(0), 2014.

[302] Luke Carrivick, Simon Rogers, Jeremy Clark, Colin Campbell, Mark Girolami, and Colin Cooper. Identification of prognostic signatures in breast cancer microarray data using Bayesian techniques. *Journal of The Royal Society Interface*, 3(8):367–381, 2006.

[303] Wei Xiao, Ji Wang, Heng Li, Ding Xia, Gan Yu, Weimin Yao, Yang Yang, Haibing Xiao, Bin Lang, Xin Ma, et al. Fibulin-1 is epigenetically downregulated and related with bladder cancer recurrence. *BMC Cancer*, 14(1):1, 2014.

[304] Zhiying Xu, Hui Chen, Deliang Liu, and Jirong Huo. Fibulin-1 is downregulated through promoter hypermethylation in colorectal cancer: A consort study. *Medicine*, 94(13):e663, 2015.

[305] Victoria Sanz-Moreno, Gilles Gadea, Jessica Ahn, Hugh Paterson, Pierfrancesco Marra, Sophie Pinner, Erik Sahai, and Christopher J Marshall. *Rac* activation and inactivation control plasticity of tumor cell movement. *Cell*, 135(3):510–523, 2008.

[306] Peter Friedl, Joseph Locker, Erik Sahai, and Jeffrey E Segall. Classifying collective cancer cell invasion. *Nature Cell Biology*, 14(8):777–783, 2012.

[307] Qi Long, Brent A Johnson, Adeboye O Osunkoya, Yu-Heng Lai, Wei Zhou, Mark Abramovitz, Mingjing Xia, Mark B Bouzyk, Robert K Nam, Linda Sugar, et al. Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *The American Journal of Pathology*, 179(1): 46–54, 2011.

[308] Gennadi V Glinsky, Olga Berezovska, and Anna B Glinskii. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *Journal of Clinical Investigation*, 115(6):1503, 2005.

[309] Anne Planche, Marina Bacac, Paolo Provero, Carlo Fusco, Mauro Delorenzi, Jean-Christophe Stehle, and Ivan Stamenkovic. Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PloS One*, 6(5):e18640, 2011.

[310] Tarek A Bismar, Francesca Demichelis, Alberto Riva, Robert Kim, Sooryanarayana Varambally, Le He, Jeff Kutok, Jonathan C Aster, Jeffery Tang, Rainer Kuefer, et al. Defining aggressive prostate cancer using a 12-gene model. *Neoplasia*, 8(1):59–68, 2006.

[311] Sridhar Ramaswamy, Ken N Ross, Eric S Lander, and Todd R Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1):49–54, 2003.

[312] Laia Agell, Silvia Hernández, Lara Nonell, Marta Lorenzo, Eulàlia Puigdecanet, Silvia de Muga, Nuria Juanpere, Raquel Bermudo, Pedro L Fernández, José A Lorente, et al. A 12-gene expression signature is associated with aggressive histological in prostate cancer: *SEC14L1* and *TCEB1* genes are potential markers of progression. *The American Journal of Pathology*, 181(5):1585–1594, 2012.

[313] Marina Bibikova, Eugene Chudin, Amir Arsanjani, Lixin Zhou, Eliza Wickham Garcia, Joshua Modder, Monica Kostelec, David Barker, Tracy Downs, Jian-Bing Fan, et al. Expression signatures that correlated with Gleason score and relapse in prostate cancer. *Genomics*, 89(6):666–672, 2007.

[314] Chin-Lee Wu, Brock E Schroeder, Xiao-Jun Ma, Christopher J Cutie, Shulin Wu, Ranelle Salunga, Yi Zhang, Michael W Kattan, Catherine A Schnabel, Mark G Erlander, et al. Development and validation of a 32-gene prognostic index for prostate cancer progression. *Proceedings of the National Academy of Sciences*, 110(15):6121–6126, 2013.

[315] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[316] Prabhakar Rajan, Jacqueline Stockley, Ian M Sudbery, Janis T Fleming, Ann Hedley, Gabriela Kalna, David Sims, Chris P Ponting, Andreas Heger, Craig N Robson, et al. Identification of a candidate prognostic gene signature by transcriptome analysis of matched pre-and post-treatment prostatic biopsies from patients with advanced prostate cancer. *BMC Cancer*, 14(1):977, 2014.

[317] Shazia Irshad, Mukesh Bansal, Mireia Castillo-Martin, Tian Zheng, Alvaro Aytes, Sven Wenske, Clémentine Le Magnen, Paolo Guarnieri, Pavel Sumazin, Mitchell C Benson, et al. A molecular signature predictive of indolent prostate cancer. *Science Translational Medicine*, 5(202):202ra122–202ra122, 2013.

[318] Naomi L Sharma, Charlie E Massie, Antonio Ramos-Montoya, Vincent Zecchini, Helen E Scott, Alastair D Lamb, Stewart MacArthur, Rory Stark, Anne Y Warren, Ian G Mills, et al. The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer Cell*, 23(1):35–47, 2013.

[319] Emilie Lalonde, Adrian S Ishkanian, Jenna Sykes, Michael Fraser, Helen Ross-Adams, Nicholas Erho, Mark J Dunning, Silvia Halim, Alastair D Lamb, Nathalie C Moon, et al. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *The Lancet Oncology*, 15(13):1521–1532, 2014.

[320] Jindan Yu, Jianjun Yu, Daniel R Rhodes, Scott A Tomlins, Xuhong Cao, Guoan Chen, Rohit Mehra, Xiaoju Wang, Debashis Ghosh, Rajal B Shah, et al. A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Research*, 67(22):10657–10663, 2007.

[321] Daehwan Kim and Steven L Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):1, 2011.

[322] Mohammed Alshalalfa, Anamaria Crisan, Ismael A Vergara, Mercedeh Ghadessi, Christine Buerki, Nicholas Erho, Kasra Yousefi, Thomas Sierocinski, Zaid Haddad, Peter C Black, et al. Clinical and genomic analysis of metastatic prostate cancer progression with a background of postoperative biochemical recurrence. *BJU International*, 116(4):556–567, 2015.

[323] K Allanach, M Mengel, G Einecke, B Sis, LG Hidalgo, T Mueller, and PF Halloran. Comparing microarray versus RT-PCR assessment of renal allograft biopsies: Similar performance despite different dynamic ranges. *American Journal of Transplantation*, 8(5):1006–1015, 2008.

[324] Daniel Bottomly, Nicole AR Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J Buck, Robert P Searles, Michael Mooney, Shannon K McWeeney, and Robert Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. *PloS One*, 6(3): e17820, 2011.

[325] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

[326] Florian Wickelmaier. An introduction to MDS. *Sound Quality Research Unit, Aalborg University, Denmark*, 46, 2003.

[327] Yiming Ying, Peng Li, and Colin Campbell. A marginalized variational bayesian approach to the analysis of array data. In *BMC proceedings*, volume 2, page 1, 2008.

[328] Yee W Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006.

[329] Tomonari Masada, Tsuyoshi Hamada, Yuichiro Shibata, and Kiyoshi Oguri. Bayesian multi-topic microarray analysis with hyperparameter reestimation. In *International Conference on Advanced Data Mining and Applications*, pages 253–264. Springer, 2009.

[330] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 2015.

[331] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[332] Paul Roepman, Andreas Schlicker, Josep Tabernero, Ian Majewski, Sun Tian, Victor Moreno, Mireille H Snel, Christine M Chresta, Robert Rosenberg, Ulrich Nitsche, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer*, 134(3):552–562, 2014.

[333] Eva Budinska, Vlad Popovici, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, Pu Yan, John Graeme Hodgson, Scott Weinrich, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology*, 231(1): 63–76, 2013.

[334] E Melo Felipe De Sousa, Xin Wang, Marnix Jansen, Evelyn Fessler, Anne Trinh, Laura PMH de Rooij, Joan H de Jong, Onno J de Boer, Ronald van Leersum, Maarten F Bijlsma, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, 19(5):614–618, 2013.

[335] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*, 10(5):e1001453, 2013.

[336] Beatriz Perez Villamil, Alejandro Romera Lopez, Susana Hernandez Prieto, Guillermo Lopez Campos, Antonio Calles, Jose Antonio Lopez Asenjo, Julian Sanz Ortega, Cristina Fernandez Perez, Javier Sastre, Rosario Alfonso, et al. Colon cancer molecular subtypes identified by expression profiling and associatedto stroma, mucinous type and different clinical behavior. *BMC Cancer*, 12(1): 1, 2012.

[337] Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics*, 5(1):1, 2012.

[338] Anguraj Sadanandam, Costas A Lyssiotis, Krisztian Homicsko, Eric A Collisson, William J Gibb, Stephan Wullschleger, Liliane C Gonzalez Ostos, William A Lannon, Carsten Grotzinger, Maguy Del Rio, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, 19(5):619–625, 2013.

[339] Renaud Gaujoux and Cathal Seoighe. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 11(1):1, 2010.

[340] Mei Jiang, Yunsheng Ma, Congcong Chen, Xuping Fu, Shu Yang, Xia Li, Guohua Yu, Yumin Mao, Yi Xie, and Yao Li. Androgen-responsive gene database: integrated knowledge on androgen-responsive genes. *Molecular Endocrinology*, 23(11):1927–1933, 2009.

[341] Federica Rizzi and Saverio Bettuzzi. Clusterin (*CLU*) and prostate cancer. *Advances in Cancer Research*, 105:1–19, 2009.

[342] André Fujita, Luciana Rodrigues Gomes, João Ricardo Sato, Rui Yamaguchi, Carlos Eduardo Thomaz, Mari Cleide Sogayar, and Satoru Miyano. Multivariate gene expression analysis reveals functional connectivity changes between normal/tumoral prostates. *BMC Systems Biology*, 2(1):1, 2008.

[343] X Gao, J Pang, LY Li, WP Liu, JM Di, QP Sun, YQ Fang, XP Liu, XY Pu, D He, et al. Expression profiling identifies new function of collapsin response mediator protein 4 as a metastasis-suppressor in prostate cancer. *Oncogene*, 29 (32):4555–4566, 2010.

[344] Agnieszka Anna Rawłuszko-Wieczorek, Karolina Horbacka, Piotr Krokowicz, Matthew Misztal, and Paweł Piotr Jagodziński. Prognostic potential of DNA methylation and transcript levels of *HIF1A* and *EPAS1* in colorectal cancer. *Molecular Cancer Research*, 12(8):1112–1127, 2014.

[345] Julia Ciampa, Meredith Yeager, Laufey Amundadottir, Kevin Jacobs, Peter Kraft, Charles Chung, Sholom Wacholder, Kai Yu, William Wheeler, Michael J Thun, et al. Large-scale exploration of gene–gene interactions in prostate cancer using a multistage genome-wide association study. *Cancer Research*, 71(9):3287–3295, 2011.

[346] Jing-Dong Zhou, Dong-Ming Yao, Ying-Ying Zhang, Ji-Chun Ma, Xiang-Mei Wen, Jing Yang, Hong Guo, Qin Chen, Jiang Lin, and Jun Qian. *GPX3* hypermethylation serves as an independent prognostic biomarker in non-m3 acute myeloid leukemia. *American Journal of Cancer Research*, 5(5):1786, 2015.

[347] Hua Zhao, Jingyi Li, Xin Li, Chao Han, Yi Zhang, Lili Zheng, and Mingzhou Guo. Silencing *GPX3* expression promotes tumor metastasis in human thyroid cancer. *Current Protein and Peptide Science*, 16(4):316–321, 2015.

[348] Leonel Maldonado, Mariana Brait, Myriam Loyo, Lauren Sullenberger, Kevin Wang, Sarah B Peskoe, Eli Rosenbaum, Roslyn Howard, Antoun Toubaji, Roula Albadine, et al. *GSTP1* promoter methylation is associated with recurrence in early stage prostate cancer. *The Journal of Urology*, 192(5):1542–1548, 2014.

[349] JJH Eijsink, A Lendvai, V Deregowski, HG Klip, G Verpooten, L Dehaspe, GH de Bock, H Hollema, W van Criekinge, E Schuuring, et al. A four-gene methylation marker panel as triage test in high-risk human papillomavirus positive patients. *International Journal of Cancer*, 130(8):1861–1869, 2012.

[350] Donkena Krishna Vanaja, Mathias Ehrich, Dirk Van den Boom, John C Cheville, R Jeffrey Karnes, Donald J Tindall, Charles R Cantor, and Charles YF Young. Hypermethylation of genes for diagnosis and risk stratification of prostate cancer. *Cancer Investigation*, 27(5):549–560, 2009.

[351] Lu Chen, Liping Su, Jianfang Li, Yanan Zheng, Beiqin Yu, Yingyan Yu, Min Yan, Qinlong Gu, Zhenggang Zhu, and Bingya Liu. Hypermethylated *FAM5C* and *MYLK* in serum as diagnosis and pre-warning markers for gastric cancer. *Disease markers*, 32(3):195–202, 2012.

[352] Silvia Esposito, Marco V Russo, Irma Airoldi, Maria Grazia Tupone, Carlo Sorrentino, Giulia Barbarito, Serena Di Meo, and Emma Di Carlo. *SNAI2/Slug* gene is silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor and pluripotency gene expression. *Oncotarget*, 6(19): 17121, 2015.

[353] GE Lind, RI Skotheim, MF Fraga, VM Abeler, M Esteller, and RA Lothe. Novel epigenetically deregulated genes in testicular cancer include homeobox genes and *SCGB3A1* (*HIN-1*). *The Journal of Pathology*, 210(4):441–449, 2006.

[354] Kim Andresen, Kirsten Muri Boberg, Hege Marie Vedeld, Hilde Honne, Peter Jebsen, Merete Hektoen, Christopher A Wadsworth, Ole Petter Clausen, Knut EA Lundin, Vemund Paulsen, et al. Four DNA methylation biomarkers in biliary brush samples accurately identify the presence of cholangiocarcinoma. *Hepatology*, 61(5):1651–1659, 2015.

[355] Hai-Ning Chen, Kefei Yuan, Na Xie, Kui Wang, Zhao Huang, Yan Chen, Qianhui Dou, Min Wu, Edouard C Nice, Zong-Guang Zhou, et al. *PDLIM1* stabilizes the e-cadherin/β-catenin complex to prevent epithelial–mesenchymal transition and metastatic potential of colorectal cancer cells. *Cancer Research*, 76(5): 1122–1134, 2016.

[356] Jiang-Liu Yu, Ping Lv, Jing Han, Xin Zhu, Lian-Lian Hong, Wang-Yu Zhu, Xin-Bao Wang, Yi-Chen Wu, Pei Li, and Zhi-Qiang Ling. Methylated *TIMP-3* DNA in body fluids is an independent prognostic factor for gastric cancer. *Archives of Pathology and Laboratory Medicine*, 138(11):1466–1473, 2014.

[357] Asha M Das, Senada Koljenović, Charlotte MC Oude Ophuis, Thom van der Klok, Boris Galjart, Alex L Nigg, Wiggert A van Cappellen, Vincent Noordhoek Hegt, Winand NM Dinjens, Peggy N Atmodimedjo, et al. Association of *TIMP3* expression with vessel density, macrophage infiltration and prognosis in human malignant melanoma. *European Journal of Cancer*, 53:135–143, 2016.

[358] Suhu Liu, Suping Ren, Paul Howell, Oystein Fodstad, and Adam I Riker. Identification of novel epigenetically modified genes in human melanoma via promoter methylation gene profiling. *Pigment Cell & Melanoma Research*, 21(5):545–558, 2008.

[359] Maryam Zare, Ferdous Rastgar Jazii, Zahra-Soheila Soheili, and Mohamad-Mehdi Moghanibashi. Downregulation of tropomyosin-1 in squamous cell carcinoma of esophagus, the role of *Ras* signaling and methylation. *Molecular Carcinogenesis*, 51(10):796–806, 2012.

Appendix A

Supplementary data for Chapter 4

Sample	Patient	FISH rearr.	Benign	Stroma	Tumour
TB08.0234_v1	TB08.0234		80	20	0
TB08.0234_v3	TB08.0234		80	20	0
TB08.0262_v3	TB08.0262	N	0	38	75
TB08.0268_v3	TB08.0268	N	15	78	5
TB08.0271_v1	TB08.0271	N	33	68	10
TB08.0311_v2	TB08.0311	Y	5	70	33
TB08.0311_v3	TB08.0311	Y	15	73	10
TB08.0327_v1	TB08.0327	Y	19	46	30
TB08.0341_v1	TB08.0341		50	50	0
TB08.0341_v5	TB08.0341	N	25	40	25
TB08.0359_v16	TB08.0359		40	60	0
TB08.0359_v2	TB08.0359	N	20	70	0
TB08.0368_v14	TB08.0386	Y	35	60	0
TB08.0429_v7	TB08.0429		29	69	3
TB08.0489_v5	TB08.0489		44	56	0
TB08.0489_v13	TB08.0489	Y	40	40	30
TB08.0501_v8	TB08.0501	N	7	58	33
TB08.0519_v14	TB08.0519	Y	3	23	75
TB08.0533_v6	TB08.0533	N	10	40	50
TB08.0588_v1	TB08.0588	Y	10	50	40
TB08.0589_v1	TB08.0589	N	8	54	36
TB08.0589_v2	TB08.0589	N	0	0	10
TB08.0589_v4	TB08.0589	N	15	85	0
TB08.0589_v5	TB08.0589	N	0	83	8
TB08.0598_v12	TB08.0598	N	5	45	45
TB08.0609_v11	TB08.0609	Y	18	66	15
TB08.0667_v9	TB08.0667	N	19	40	40

TB08.0667_v6	TB08.0667		30	70	0
TB08.0689_v14	TB08.0689	Y	28	33	40
TB08.0689_v15	TB08.0689	Y	20	10	70
TB08.0689_v2	TB08.0689	Y	24	53	21
TB08.0689_v8	TB08.0689	N	20	48	33
TB08.0691_v13	TB08.0691	Y	5	43	50
TB08.0716_v18	TB08.0716	N	15	85	0
TB08.0719_v11	TB08.0719	N	5	43	50
TB08.0731_v13	TB08.0731	Y	15	83	3
TB08.0816_v2	TB08.0816	Y	18	63	18
TB08.0817_v14	TB08.0817	N	18	46	34
TB08.0848_v10	TB08.0848	Y	10	55	35
TB08.0869_v4	TB08.0869	Y	0	0	5
TB08.0869_v6	TB08.0869	Y	0	80	15
TB08.0869_v7	TB08.0869	Y	5	73	15
TB08.0870_v18	TB08.0870	N	10	70	8
TB08.0872_v2	TB08.0872	Y	10	68	20
TB08.0877_v19	TB08.0877	Y	25	35	40
TB08.0879_v11	TB08.0879	Y	25	65	5
TB08.0884_v2	TB08.0884	N	60	40	0
TB08.0927_v5	TB08.0927	N	15	65	20
TB08.0943_v7	TB08.0943	N	10	90	0
TB08.0958_v12	TB08.0958	Y	18	23	55
TB08.0958_v13	TB08.0958	Y	28	28	45
TB08.0973_v9	TB08.0973	N	15	60	23
TB08.0978_v7	TB08.0978	N	0	75	20
TB08.0978_v8	TB08.0978	N	10	40	45
TB08.0978_v9	TB08.0978	N	3	66	29
TB08.0986_v2	TB08.0986	Y	30	65	38
TB08.0987_v6	TB08.0987	N	6	43	49
TB08.0993_v12	TB08.0993	Y	34	61	4
TB08.0997_v6	TB08.0997		20	80	0
TB08.0999_v11	TB08.0999	N	30	33	30
TB08.0999_v2	TB08.0999	Y	25	48	48
TB08.1015_v10	TB08.1015	Y	5	15	78
TB08.1015_v11	TB08.1015	Y	5	18	78
TB08.1015_v9	TB08.1015	Y	1	44	50
TB08.1019_v1	TB08.1019	Y	20	70	10
TB08.1019_v14	TB08.1019	Y	19	68	10
TB08.1019_v15	TB08.1019	Y	30	48	20
TB08.1019_v2	TB08.1019	Y	0	0	30
TB08.1026_v17	TB08.1026	N	5	13	78
TB08.1044_v7	TB08.1044	N	18	65	40
TB08.1053_v5	TB08.1053	Y	5	43	48

TB08.1063_v16	TB08.1063	Y	5	40	50
TB08.1063_v8	TB08.1063	N	19	48	31
TB08.1083_v3	TB08.1083	Y	11	54	33
TB08.1116_v2	TB08.1116	Y	16	69	15
TB08.1116_v3	TB08.1116	Y	3	40	56
TB08.1116_v9	TB08.1116	Y	18	53	30
TB08.1159_v2	TB08.1159	Y	50	50	0
TB08.0601_v16	TB08.601		30	60	NA
TB09.0217_v16	TB09.0217	Y	3	30	63
TB09.0217_v7	TB09.0217	N	33	40	28
TB09.0219_v13	TB09.0219	N	15	75	10
TB09.0219_v2	TB09.0219	Y	29	60	11
TB09.0219_v21	TB09.0219	Y	10	33	57
TB09.0219_v8	TB09.0219	N	14	80	4
TB09.0238_v12	TB09.0238	N	20	80	0
TB09.0238_v18	TB09.0238	Y	10	40	50
TB09.0238_v5	TB09.0238	N	5	50	25
TB09.0272_v6	TB09.0272	Y	5	45	65
TB09.0272_v7	TB09.0272	N	20	45	35
TB09.0295_v2	TB09.0295	N	5	25	70
TB09.0413_v11	TB09.0413	N	10	35	68
TB09.0413_v8	TB09.0413	N	15	80	5
TB09.0443_v3	TB09.0443	Y	18	80	2
TB09.0443_v8	TB09.0443	N	0	35	65
TB09.0448_v8	TB09.0448	N	15	53	33
TB09.0462_v7	TB09.0462	Y	13	80	8
TB09.0471_v11	TB09.0471	Y	20	60	20
TB09.0504_v4	TB09.0504	N	10	40	50
TB09.0550_v15	TB09.0550	Y	3	38	55
TB09.0606_v3	TB09.0606	N	15	61	18
TB09.0706_v5	TB09.0706	Y	6	36	54
TB09.0720_v19	TB09.0720	Y	5	73	23
TB09.0721_v14	TB09.0721	N	13	75	10
TB09.0721_v15	TB09.0721	Y	30	68	3
TB09.0725_v9	TB09.0725	N	5	25	68
TB09.0774_v1	TB09.0774	Y	15	85	0
TB09.0774_v15	TB09.0774	N	35	55	10
TB09.0850_v2	TB09.0850	Y	5	90	5
TB09.0962_v13	TB09.0962	N	15	60	23
TB09.0962_v16	TB09.0962	Y	5	18	75
NP1	ICR_38	N	NA	NA	NA
NP10	ICR_47	N	NA	NA	NA
NP11	ICR_50	N	NA	NA	NA
NP12	ICR_58	N	NA	NA	NA

NP14	ICR_35	N	NA	NA	NA
NP15	ICR_65	N	NA	NA	NA
NP16	ICR_69	N	NA	NA	NA
NP17	ICR_51	N	35	65	0
NP18	ICR_66	N	15	85	0
NP19	ICR_73	N	25	75	0
NP2	ICR_37	N	65	35	0
NP20	ICR_57	N	40	60	0
NP21	ICR_56	N	5	95	0
NP4	ICR_47	N	45	50	0
NP5	ICR_59	N	75	25	0
NP8	ICR_34	N	NA	NA	NA
NP9	ICR_54	N	NA	NA	NA
PRC140	ICR_68	Y	35	55	10
PRC101	ICR_44	Y	40	20	40
PRC102	ICR_34	N	20	20	60
PRC103	ICR_43	N	60	15	20
PRC105	ICR_54	N	25	30	45
PRC106	ICR_54	N	35	50	15
PRC109	ICR_49	Y	0	40	60
PRC10	ICR_28	Y	NA	NA	NA
PRC110	ICR_49	Y	2	50	55
PRC111	ICR_49	N	50	25	20
PRC112	ICR_60	N	60	30	NA
PRC113	ICR_63	N	10	15	70
PRC114	ICR_41	Y	30	30	40
PRC115	ICR_41	Y	40	30	30
PRC116	ICR_17	Y	30	20	50
PRC117	ICR_17	Y	50	20	20
PRC118	ICR_50	N	3	6	90
PRC119	ICR_59	Y	45	25	30
PRC11	ICR_22	Y	5	35	60
PRC122	ICR_17	Y	70	27	3
PRC123	ICR_40	N	50	30	5
PRC124	ICR_61	N	40	40	20
PRC125	ICR_40	N	5	50	45
PRC126	ICR_48	Y	0	25	70
PRC127	ICR_48	Y	NA	30	50
PRC128	ICR_55	Y	60	25	15
PRC129	ICR_55	Y	2	25	70
PRC12	ICR_4		0	10	85
PRC130	ICR_58	N	0	25	70
PRC133	ICR_35	N	0	10	90
PRC134	ICR_35	N	50	50	0

PRC135	ICR_68	Y	5	35	60
PRC136	ICR_71	N	0	30	70
PRC137	ICR_65	N	15	55	30
PRC138	ICR_69	N	5	35	60
PRC139	ICR_69	N	0	30	70
PRC13	ICR_25	Y	35	40	25
PRC141	ICR_67	Y	5	35	60
PRC142	ICR_73	N	50	50	0
PRC143	ICR_57	Y	35	60	5
PRC144	ICR_45	Y	0	30	70
PRC145	ICR_56	N	40	55	5
PRC146	ICR_70	N	38	60	2
PRC147	ICR_70	Y	35	60	5
PRC148	ICR_39	N	25	40	35
PRC149	ICR_72	N	35	60	5
PRC14	ICR_2	N	50	50	0
PRC150	ICR_53	Y	30	40	30
PRC151	ICR_64	N	10	40	50
PRC152	ICR_33	N	25	60	15
PRC153	ICR_33	N	30	50	20
PRC154	ICR_1	N	0	35	65
PRC155	ICR_62	N	5	30	65
PRC156	ICR_74	Y	0	50	50
PRC157	ICR_8	N	0	15	85
PRC158	ICR_80	N	0	30	70
PRC159	ICR_79	N	20	40	40
PRC15	ICR_7	N	50	50	0
PRC160	ICR_76	N	5	20	75
PRC161	ICR_80	N	10	30	60
PRC162	ICR_81	Y	20	30	50
PRC163	ICR_73	N	20	30	50
PRC164	ICR_3	Y	35	25	40
PRC165	ICR_36	Y	40	30	30
PRC166	ICR_19	Y	10	25	65
PRC167	ICR_78	Y	5	25	70
PRC168	ICR_77	Y	0	30	70
PRC169	ICR_75	Y	20	70	10
PRC16	ICR_23		40	60	0
PRC17	ICR_6	Y	50	40	10
PRC18	ICR_25		NA	NA	NA
PRC19	ICR_27		0	95	5
PRC1	ICR_20	Y	25	30	45
PRC20	ICR_82	Y	35	50	15
PRC21	ICR_82	Y	45	40	15

PRC22	ICR_24		60	40	0
PRC23	ICR_26		60	40	0
PRC24	ICR_12	Y	25	45	30
PRC25	ICR_29	Y	30	45	35
PRC26	ICR_30	N	20	65	15
PRC27	ICR_13		15	35	50
PRC28	ICR_15		55	40	5
PRC29	ICR_18	N	50	35	15
PRC2	ICR_2	Y	60	30	10
PRC30	ICR_22	Y	NA	NA	NA
PRC31	ICR_14		95	0	5
PRC32	ICR_21		50	45	5
PRC34	ICR_5	N	40	60	0
PRC35	ICR_5	Y	40	60	0
PRC36	ICR_12	Y	50	45	5
PRC38	ICR_11	Y	25	60	15
PRC39	ICR_32		30	60	10
PRC3	ICR_7	Y	20	30	50
PRC40	ICR_20	Y	0	30	70
PRC42	ICR_10	Y	50	45	5
PRC45	ICR_14		60	40	0
PRC4	ICR_9	Y	35	40	25
PRC5	ICR_16	Y	47	50	3
PRC6	ICR_23		10	10	80
PRC7	ICR_10		50	0	50
PRC8	ICR_23		10	10	80
PRC9	ICR_31		35	35	30
ST1	ICR_48	Y	0	100	0
ST2	ICR_46	N	0	100	0
ST3	ICR_52	N	5	95	0
ST4	ICR_66	N	0	100	0
ST5	ICR_76	N	0	100	0

Table A.1 The break-apart FISH status and the percentages for each tissue type in the CancerMap samples used for clinical correlation.

5’ partner	AR	Citation	Rearr.
ACSL3	Y	Tandefelt et al. [136]	Y
AK311452	N	Pflueger et al. [263]	N
ALG5	N	Pflueger et al. [263]	Y
ARHGEF3	Y	Baca et al. [54]	Y
ATP1A4	N	Baca et al. [54]	Y
AX747630	N	Maher et al. [262]	Y
AZGP1	Y	Pflueger et al. [263]	N
BRAF	N	Baca et al. [54]	Y
C15ORF21	Y	Tandefelt et al. [136]	Y
CANT	Y	Tandefelt et al. [136]	Y
CDKN1A	Y	Pflueger et al. [263]	Y
CSMD2	N	Baca et al. [54]	Y
DDX5	N	Tandefelt et al. [136]	Y
EIF4E2	N	Maher et al. [262]	Y
ERG	N	Baca et al. [54]	Y
ESRP1	N	Pflueger et al. [263]	Y
EST14	Y	Tandefelt et al. [136]	Y
FKBP5	Y	Pflueger et al. [263]	Y
FOXP1	N	Tandefelt et al. [136]	Y
GSK3B	N	Baca et al. [54]	Y
HARS2	Y	Pflueger et al. [263]	N
HERPUD1	Y	Maher et al. [262]	Y
HERV-K_22q11.23	Y	Tandefelt et al. [136]	Y
HERVK17	Y	Tandefelt et al. [136]	Y
HJURP	N	Maher et al. [262]	Y
HNRPA2B1	N	Tandefelt et al. [136]	Y
KIF2A	N	Baca et al. [54]	Y
KLK2	Y	Tandefelt et al. [136]	Y
LMAN2	Y	Pflueger et al. [263]	Y
LMBR1	N	Baca et al. [54]	Y
MIER2	N	Pflueger et al. [263]	Y
MIPOL1	N	Pflueger et al. [263]	Y
NDRG1	Y	Tandefelt et al. [136]	Y
NUP35	N	Baca et al. [54]	Y
NXPH1	N	Baca et al. [54]	Y
OR15E2	N	Barros-Silva et al. [261]	Y
PDZRN3	Y	Baca et al. [54]	Y
PTEN	Y	Baca et al. [54]	Y
RC3H2	N	Pflueger et al. [263]	Y
SLC45A3	Y	Pflueger et al. [263]	Y/N
SMG5	N	Pflueger et al. [263]	N
ST6GALNAC6	N	Pflueger et al. [263]	N

STRN4	N	Pflueger et al. [263]	Y
TBC1D12	N	Baca et al. [54]	Y
TIA1	Y	Maher et al. [262]	Y
TMPRSS2	Y	Tomlins et al. [32]	Y
TNPO1	N	Pflueger et al. [263]	Y
UBTF	Y	Barros-Silva et al. [261]	Y
USP10	N	Maher et al. [262]	Y
VMAC	N	Pflueger et al. [263]	N
XKR4	N	Baca et al. [54]	Y
YIPF1	N	Baca et al. [54]	Y
ZDHHC7	N	Maher et al. [262]	Y
ZNF649	N	Pflueger et al. [263]	N
ZNF772	N	Pflueger et al. [263]	N

Table A.2 Known 5’ fusion partners in prostate cancer. The **AR** column indicates if the gene is androgen regulated, as determined based on the ARGDB database [340]. The **Rearr.** column indicates if the gene is involved in rearrangements (Y) or read-through transcriptions (N).

3’ partner	**Citation**	**Rearr.**
ABCB9	Pflueger et al. [263]	Y
AK094188	Pflueger et al. [263]	N
Ak1	Pflueger et al. [263]	N
AP3S1	Pflueger et al. [263]	Y
BRAF	Pflueger et al. [263]	Y
CAPS	Pflueger et al. [263]	N
CD9	Pflueger et al. [263]	Y
CYP2A6	Baca et al. [54]	Y
DGKB	Pflueger et al. [263]	Y
DIRC2	Maher et al. [262]	Y
ELK4	Pflueger et al. [263]	Y/N
ERG	Tomlins et al. [32]	Y
ETV1	Tomlins et al. [32]	Y
ETV4	Tandefelt et al. [136]	Y
ETV5	Tandefelt et al. [136]	Y
FAF1	Baca et al. [54]	Y
FKBP5	Pflueger et al. [263]	Y
FLI1	Tandefelt et al. [136]	Y
FOXP1	Baca et al. [54]	Y
GJC3	Pflueger et al. [263]	N
GPSN2	Pflueger et al. [263]	Y
GUCY2C	Baca et al. [54]	Y
HJURP	Maher et al. [262]	Y

IKBKB	Pflueger et al. [263]	Y
INPP4A	Maher et al. [262]	Y
LYN	Baca et al. [54]	Y
MLL3	Baca et al. [54]	Y
NRF1	Baca et al. [54]	Y
PADI6	Baca et al. [54]	Y
PAQR6	Pflueger et al. [263]	N
PDE4D	Baca et al. [54]	Y
PIGU	Pflueger et al. [263]	Y
PLCE1	Baca et al. [54]	Y
PRKG1	Baca et al. [54]	Y
RAF1	Pflueger et al. [263]	Y
RGS3	Pflueger et al. [263]	Y
RSRC2	Pflueger et al. [263]	Y
SLC14A2	Baca et al. [54]	Y
SURF4	Baca et al. [54]	Y
TBL1XR1	Baca et al. [54]	Y
TMPRSS2	Baca et al. [54]	Y
VN1R1	Pflueger et al. [263]	N
ZDHHC7	Maher et al. [262]	Y
ZMAT2	Pflueger et al. [263]	N
ZNF577	Pflueger et al. [263]	N

Table A.3 Known 3’ fusion partners in prostate cancer. The **Rearr.** column indicates if the gene is involved in rearrangements (Y) or read-through transcriptions (N).

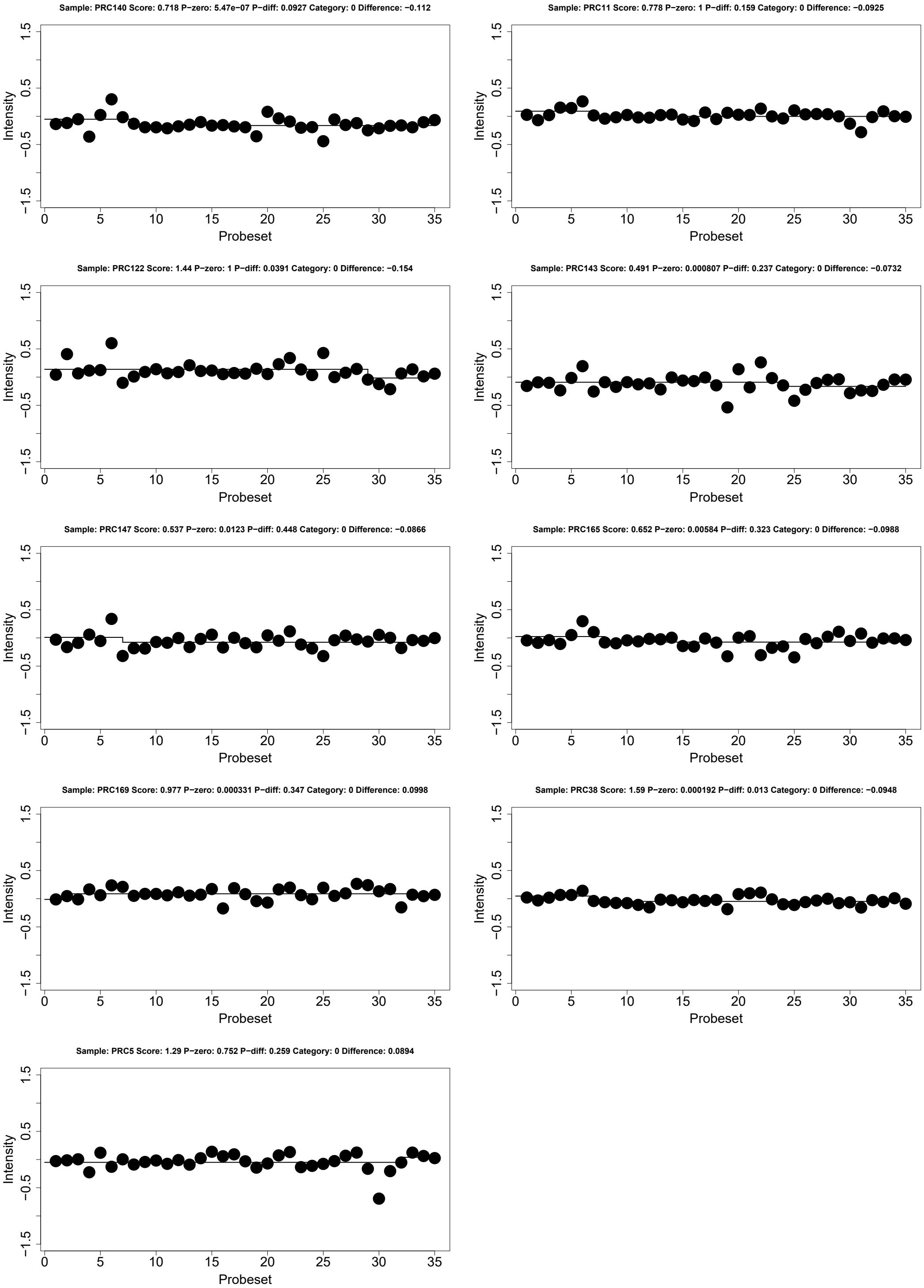

Figure A.1 The false negatives of the step method for the *ERG* gene in the ICR dataset.

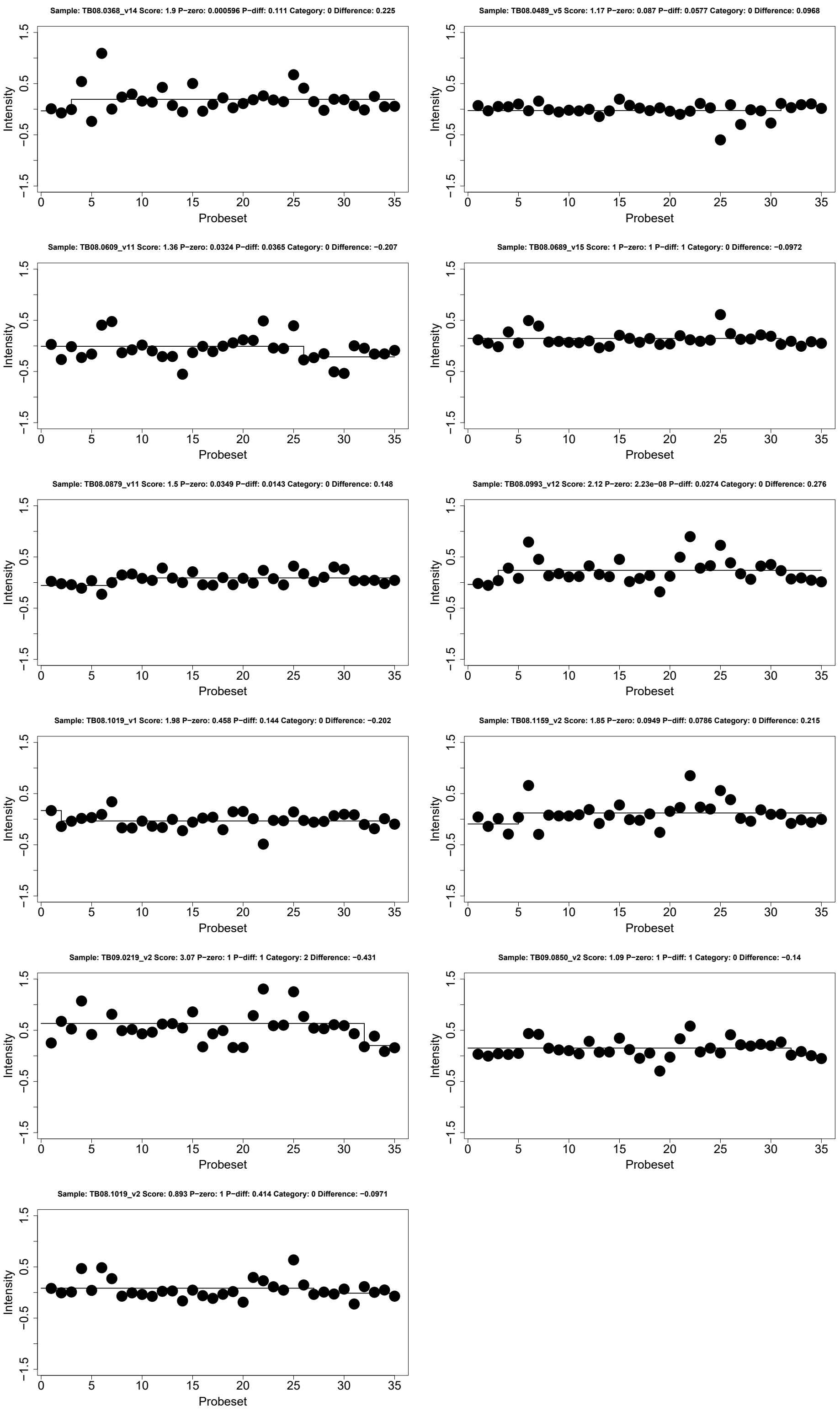

Figure A.2 The false negatives of the step method for the *ERG* gene in the Cambridge dataset.

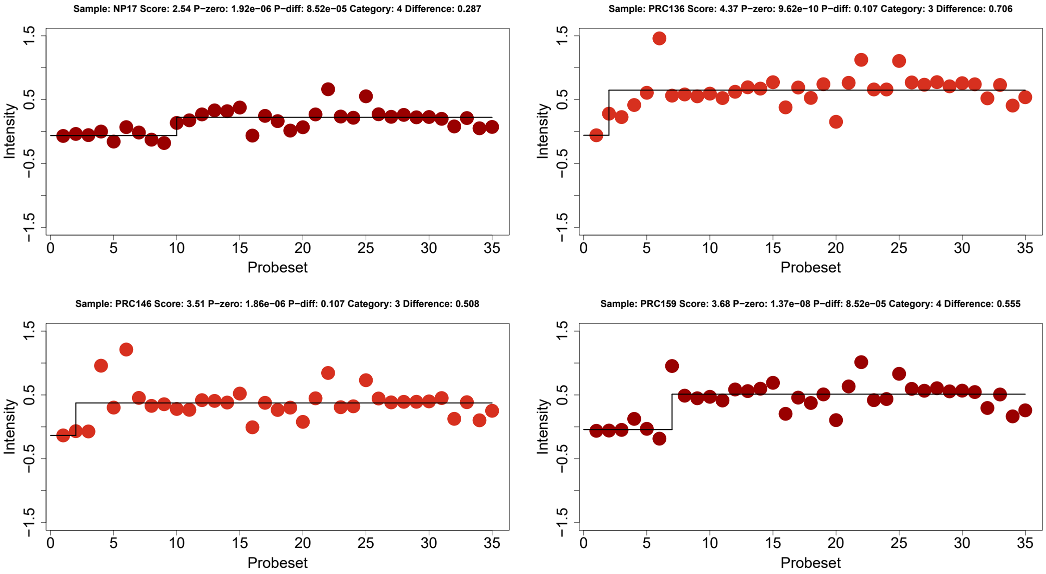

Figure A.3 The false positives of the step method for the *ERG* gene in the ICR dataset.

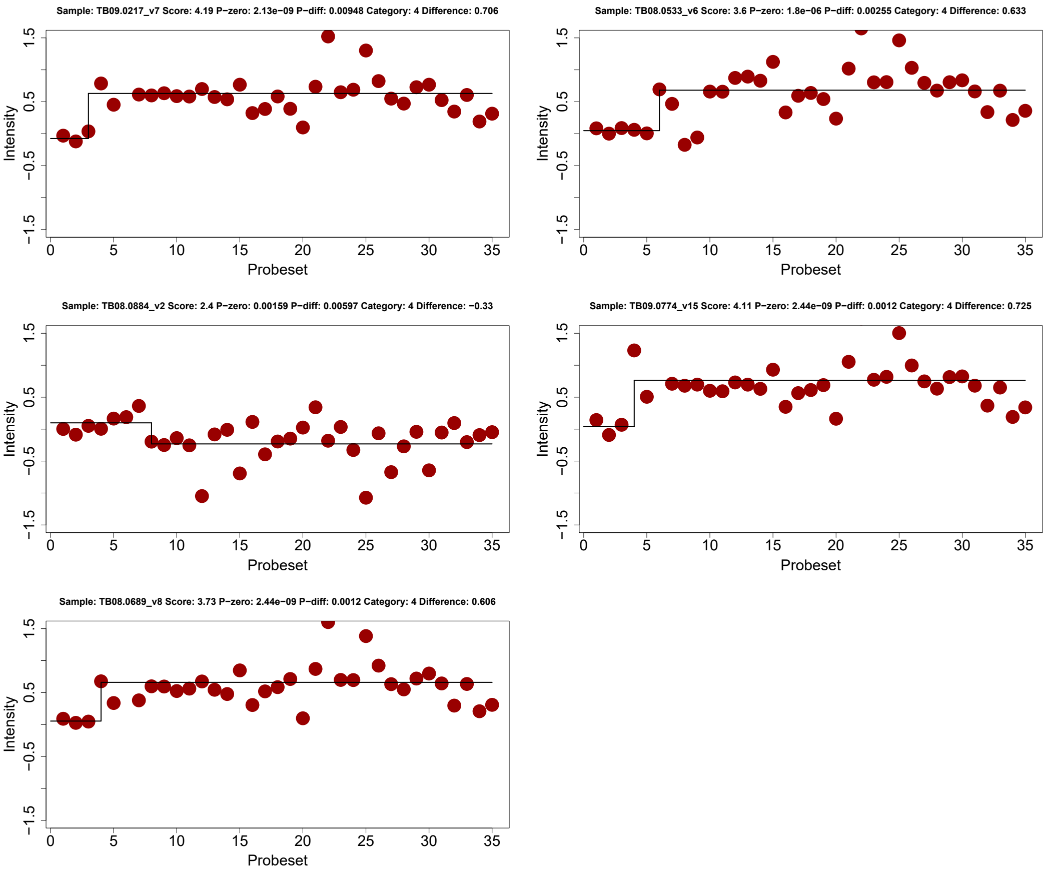

Figure A.4 The false positives of the step method for the *ERG* gene in the Cambridge dataset.

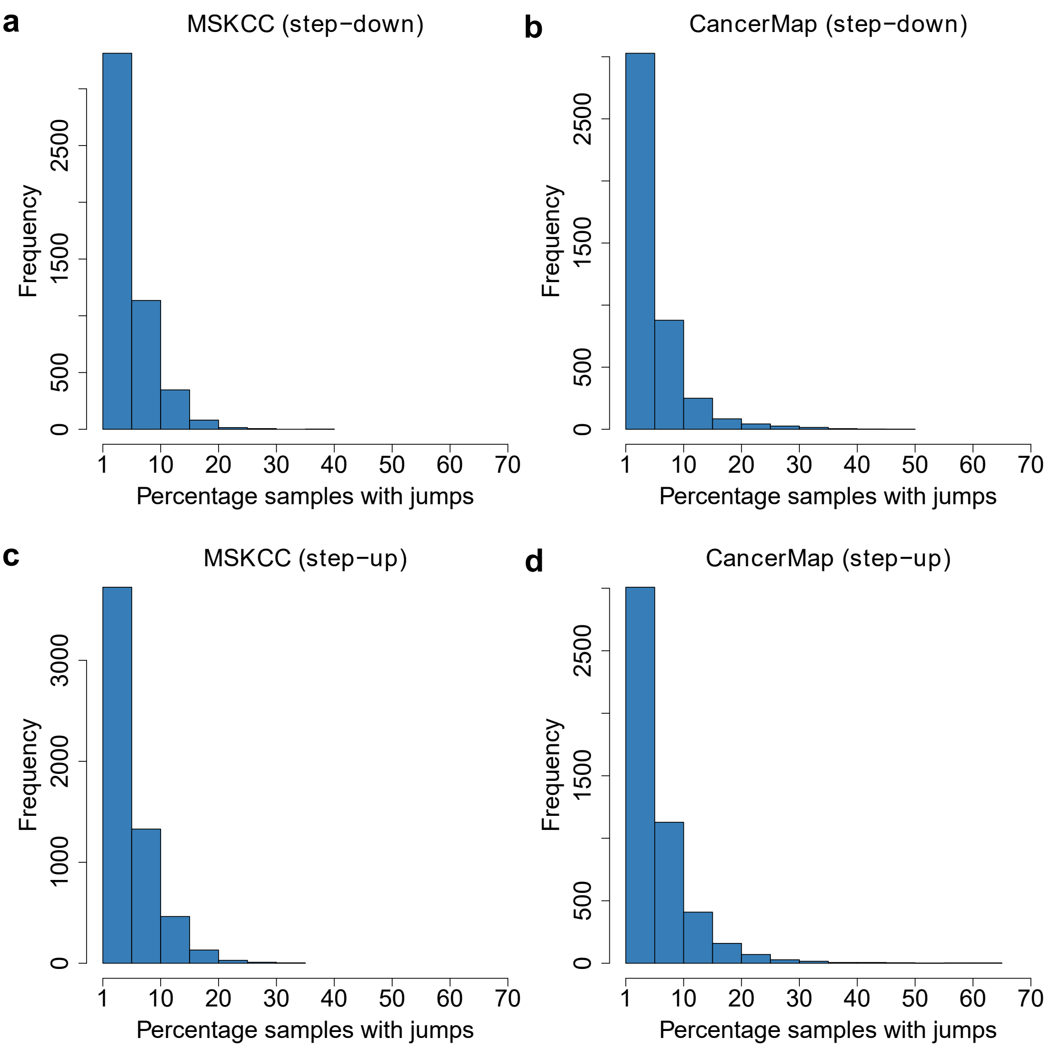

Figure A.5 Histograms depicting the distribution of percentages of samples in which the jumps are identified.

Rank	Gene	Gene ID	Count	Percent.
1	*C1QTNF3-AMACR*	ENSG00000273294	145	61.97
2	*GABRB3*	ENSG00000166206	145	61.97
3	*ARHGEF26*	ENSG00000114790	135	57.69
4	*TDRD1*	ENSG00000095627	135	57.69
5	*LUZP2*	ENSG00000187398	120	51.28
6	*CRISP3*	ENSG00000096006	115	49.15
7	*GCNT1*	ENSG00000187210	113	48.29
8	*NEK5*	ENSG00000197168	112	47.86
9	*F5*	ENSG00000198734	108	46.15
10	*KCNC2*	ENSG00000166006	104	44.44
11	*ACSM3*	ENSG00000005187	102	43.59

12	*AMACR*	ENSG00000242110	101	43.16
13	*BEND4*	ENSG00000188848	100	42.74
14	*TMEM178A*	ENSG00000152154	98	41.88
15	*NETO2*	ENSG00000171208	94	40.17
16	*ERG*	ENSG00000157554	92	39.32
17	*GPR160*	ENSG00000173890	89	38.03
18	*ZNF385B*	ENSG00000144331	87	37.18
19	*GLYATL1*	ENSG00000166840	83	35.47
20	*TMEM150C*	ENSG00000249242	83	35.47
21	*TGM3*	ENSG00000125780	82	35.04
22	*EPCAM*	ENSG00000119888	81	34.62
23	*F2R*	ENSG00000181104	80	34.19
24	*TOX3*	ENSG00000103460	78	33.33
25	*GAA*	ENSG00000273259	76	32.48
26	*ZYG11A*	ENSG00000203995	76	32.48
27	*PANK3*	ENSG00000120137	75	32.05
28	*PDE3B*	ENSG00000152270	75	32.05
29	*SLC38A11*	ENSG00000169507	75	32.05
30	*TRPM4*	ENSG00000130529	75	32.05
31	*RP11-597D13.9*	ENSG00000248429	74	31.62
32	*TAF1D*	ENSG00000166012	74	31.62
33	*AGPAT5*	ENSG00000155189	73	31.20
34	*ABCC4*	ENSG00000125257	72	30.77
35	*LRRIQ1*	ENSG00000133640	72	30.77
36	*SLC5A1*	ENSG00000100170	72	30.77
37	*DNAH8*	ENSG00000124721	71	30.34
38	*HSPA8*	ENSG00000109971	69	29.49
39	*ZSCAN12*	ENSG00000158691	67	28.63
40	*CASD1*	ENSG00000127995	66	28.21
41	*GNAI1*	ENSG00000127955	66	28.21
42	*MBOAT2*	ENSG00000143797	66	28.21
43	*NCALD*	ENSG00000104490	66	28.21
44	*BANK1*	ENSG00000153064	65	27.78
45	*DSC2*	ENSG00000134755	65	27.78
46	*FNIP2*	ENSG00000052795	65	27.78
47	*SLITRK4*	ENSG00000179542	65	27.78
48	*GNE*	ENSG00000159921	64	27.35
49	*PLA2G7*	ENSG00000146070	64	27.35
50	*TMEM45B*	ENSG00000151715	64	27.35
51	*CHRM3*	ENSG00000133019	63	26.92
52	*EFCAB13*	ENSG00000178852	63	26.92
53	*MYB*	ENSG00000118513	63	26.92
54	*PPM1E*	ENSG00000175175	63	26.92
55	*FMN1*	ENSG00000248905	62	26.50

56	*SCIN*	ENSG00000006747	62	26.50
57	*SNHG3*	ENSG00000242125	62	26.50
58	*STK19*	ENSG00000204344	61	26.07
59	*AGTR1*	ENSG00000144891	60	25.64
60	*HPN*	ENSG00000105707	60	25.64
61	*SLC26A4*	ENSG00000091137	60	25.64
62	*ATP8A2*	ENSG00000132932	59	25.21
63	*C5orf30*	ENSG00000181751	59	25.21
64	*DICER1*	ENSG00000100697	59	25.21
65	*PPP1R9A*	ENSG00000158528	59	25.21
66	*KIAA0101*	ENSG00000166803	58	24.79
67	*PIK3R2*	ENSG00000268173	58	24.79
68	*TMEM181*	ENSG00000146433	58	24.79
69	*TOP2A*	ENSG00000131747	58	24.79
70	*UAP1*	ENSG00000117143	57	24.36
71	*CNTNAP2*	ENSG00000174469	56	23.93
72	*NPR3*	ENSG00000113389	56	23.93
73	*TMC5*	ENSG00000103534	56	23.93
74	*DMRTA1*	ENSG00000176399	55	23.50
75	*NEFL*	ENSG00000277586	55	23.50
76	*PGM2L1*	ENSG00000165434	55	23.50
77	*PSD3*	ENSG00000156011	55	23.50
78	*ABHD14A-ACY1*	ENSG00000114786	54	23.08
79	*ATP8A1*	ENSG00000124406	54	23.08
80	*CADM2*	ENSG00000175161	54	23.08
81	*CSTF3*	ENSG00000176102	54	23.08
82	*ECT2*	ENSG00000114346	54	23.08
83	*SFRP4*	ENSG00000106483	54	23.08
84	*SLC9A2*	ENSG00000115616	54	23.08
85	*SPON2*	ENSG00000159674	54	23.08
86	*GAS5*	ENSG00000234741	53	22.65
87	*GUCY1B3*	ENSG00000061918	53	22.65
88	*KCNN2*	ENSG00000080709	53	22.65
89	*MARC1*	ENSG00000186205	53	22.65
90	*CADM1*	ENSG00000182985	52	22.22
91	*CHRNA5*	ENSG00000169684	52	22.22
92	*GXYLT1*	ENSG00000151233	52	22.22
93	*TAF4B*	ENSG00000141384	52	22.22
94	*TMEM26*	ENSG00000196932	52	22.22
95	*TUSC3*	ENSG00000104723	52	22.22
96	*WT1*	ENSG00000184937	52	22.22
97	*FAM111B*	ENSG00000189057	51	21.79
98	*GATA6*	ENSG00000141448	51	21.79
99	*NBN*	ENSG00000104320	51	21.79

100	*NET1*	ENSG00000173848	51	21.79
101	*PIGW*	ENSG00000277161	51	21.79
102	*PPAT*	ENSG00000128059	51	21.79
103	*REPS2*	ENSG00000169891	51	21.79
104	*SLC43A1*	ENSG00000149150	51	21.79
105	*SLCO1A2*	ENSG00000084453	51	21.79
106	*ST8SIA6*	ENSG00000148488	51	21.79
107	*TMEM106C*	ENSG00000134291	51	21.79
108	*TTK*	ENSG00000112742	51	21.79
109	*AC005062.2*	ENSG00000243004	50	21.37
110	*CCDC110*	ENSG00000168491	50	21.37
111	*ENTPD5*	ENSG00000187097	50	21.37
112	*HOOK1*	ENSG00000134709	50	21.37
113	*RP11-115D19.1*	ENSG00000251095	50	21.37
114	*SLC26A5*	ENSG00000170615	50	21.37
115	*AGR2*	ENSG00000106541	49	20.94
116	*FAM3D*	ENSG00000198643	49	20.94
117	*OPRK1*	ENSG00000082556	49	20.94
118	*PRR16*	ENSG00000184838	49	20.94
119	*DNAH5*	ENSG00000039139	48	20.51
120	*EIF4A2*	ENSG00000156976	48	20.51
121	*FAM57A*	ENSG00000167695	48	20.51
122	*MCMDC2*	ENSG00000178460	48	20.51
123	*PDLIM5*	ENSG00000163110	48	20.51
124	*POPDC3*	ENSG00000132429	48	20.51
125	*RMI1*	ENSG00000178966	48	20.51
126	*SLCO1B3*	ENSG00000257046	48	20.51
127	*DDX43*	ENSG00000080007	47	20.09
128	*DNAH14*	ENSG00000185842	47	20.09
129	*KHDRBS3*	ENSG00000131773	47	20.09
130	*MAP2K6*	ENSG00000108984	47	20.09
131	*NAALADL2*	ENSG00000177694	47	20.09
132	*NCL*	ENSG00000115053	47	20.09
133	*RAPGEF4*	ENSG00000091428	47	20.09
134	*THBS4*	ENSG00000113296	47	20.09
135	*TRNT1*	ENSG00000072756	47	20.09
136	*CDK19*	ENSG00000155111	46	19.66
137	*CLGN*	ENSG00000153132	46	19.66
138	*DDX60L*	ENSG00000181381	46	19.66
139	*DNMBP*	ENSG00000107554	46	19.66
140	*GALNT3*	ENSG00000115339	46	19.66
141	*HTATSF1*	ENSG00000102241	46	19.66
142	*LYZ*	ENSG00000090382	46	19.66
143	*PIAS2*	ENSG00000078043	46	19.66

144	*RAB3IP*	ENSG00000127328	46	19.66
145	*TMEM209*	ENSG00000146842	46	19.66
146	*ANKRD32*	ENSG00000133302	45	19.23
147	*DOPEY2*	ENSG00000142197	45	19.23
148	*MGAT4A*	ENSG00000071073	45	19.23
149	*NDC1*	ENSG00000058804	45	19.23
150	*PLA2G2A*	ENSG00000188257	45	19.23
151	*RIMKLA*	ENSG00000177181	45	19.23
152	*THSD7A*	ENSG00000005108	45	19.23
153	*URI1*	ENSG00000105176	45	19.23
154	*ZNF277*	ENSG00000198839	45	19.23
155	*AADAT*	ENSG00000109576	44	18.80
156	*CCDC138*	ENSG00000163006	44	18.80
157	*CRISPLD1*	ENSG00000121005	44	18.80
158	*DAPK1*	ENSG00000196730	44	18.80
159	*FBP1*	ENSG00000165140	44	18.80
160	*MELK*	ENSG00000165304	44	18.80
161	*MSR1*	ENSG00000038945	44	18.80
162	*SLC9A7*	ENSG00000065923	44	18.80
163	*SNX16*	ENSG00000104497	44	18.80
164	*TSPAN8*	ENSG00000127324	44	18.80
165	*DDX1*	ENSG00000079785	43	18.38
166	*HGF*	ENSG00000019991	43	18.38
167	*KIAA1324L*	ENSG00000164659	43	18.38
168	*LEMD3*	ENSG00000174106	43	18.38
169	*LPL*	ENSG00000175445	43	18.38
170	*MGAT2*	ENSG00000168282	43	18.38
171	*NRCAM*	ENSG00000091129	43	18.38
172	*PAWR*	ENSG00000177425	43	18.38
173	*PHOSPHO2*	ENSG00000144362	43	18.38
174	*PNN*	ENSG00000100941	43	18.38
175	*PTPRN2*	ENSG00000155093	43	18.38
176	*RAB4A*	ENSG00000168118	43	18.38
177	*RCN2*	ENSG00000117906	43	18.38
178	*RP11-296A16.1*	ENSG00000262560	43	18.38
179	*TMEM2*	ENSG00000135048	43	18.38
180	*TPD52*	ENSG00000076554	43	18.38
181	*ANKRD37*	ENSG00000186352	42	17.95
182	*ARHGAP28*	ENSG00000088756	42	17.95
183	*CCDC15*	ENSG00000149548	42	17.95
184	*CCNG1*	ENSG00000113328	42	17.95
185	*CMPK1*	ENSG00000162368	42	17.95
186	*FMNL2*	ENSG00000157827	42	17.95
187	*GNPNAT1*	ENSG00000100522	42	17.95

188	*IGJ*	ENSG00000132465	42	17.95
189	*INTS8*	ENSG00000164941	42	17.95
190	*KIFAP3*	ENSG00000075945	42	17.95
191	*SCGN*	ENSG00000079689	42	17.95
192	*SEC22C*	ENSG00000093183	42	17.95
193	*SFPQ*	ENSG00000116560	42	17.95
194	*TDO2*	ENSG00000151790	42	17.95
195	*UNC13B*	ENSG00000198722	42	17.95
196	*ASB5*	ENSG00000164122	41	17.52
197	*CCT3*	ENSG00000163468	41	17.52
198	*CENPN*	ENSG00000166451	41	17.52
199	*DAB2*	ENSG00000153071	41	17.52
200	*DDX58*	ENSG00000107201	41	17.52

Table A.4 Top 200 candidates with the largest number of step-up jumps in the CancerMap dataset.

Rank	**Gene**	**Gene ID**	**Count**	**Percent.**
1	*OLFM4*	ENSG00000102837	122	32.97
2	*TDRD1*	ENSG00000095627	120	32.43
3	*CHRDL1*	ENSG00000101938	116	31.35
4	*PLA2G7*	ENSG00000146070	114	30.81
5	*SYNM*	ENSG00000182253	104	28.11
6	*SLC38A11*	ENSG00000169507	102	27.57
7	*SRD5A2*	ENSG00000277893	100	27.03
8	*MAN1A1*	ENSG00000111885	98	26.49
9	*SLC22A3*	ENSG00000146477	98	26.49
10	*ERG*	ENSG00000157554	96	25.95
11	*TIMP3*	ENSG00000100234	96	25.95
12	*TMEM178A*	ENSG00000152154	96	25.95
13	*F3*	ENSG00000117525	94	25.41
14	*HSD17B6*	ENSG00000025423	94	25.41
15	*CRISP3*	ENSG00000096006	92	24.86
16	*NR4A2*	ENSG00000153234	92	24.86
17	*SLC19A2*	ENSG00000117479	90	24.32
18	*EDNRB*	ENSG00000136160	88	23.78
19	*MYBPC1*	ENSG00000196091	88	23.78
20	*NPR3*	ENSG00000113389	88	23.78
21	*PANK3*	ENSG00000120137	88	23.78
22	*DSC3*	ENSG00000134762	86	23.24
23	*LTBP1*	ENSG00000049323	86	23.24
24	*PRKCD*	ENSG00000163932	86	23.24
25	*SLC12A2*	ENSG00000064651	86	23.24

26	*TIPARP*	ENSG00000163659	86	23.24
27	*ATF3*	ENSG00000162772	84	22.70
28	*MME*	ENSG00000196549	84	22.70
29	*PDK4*	ENSG00000004799	84	22.70
30	*SESN3*	ENSG00000149212	84	22.70
31	*ATP1B1*	ENSG00000143153	80	21.62
32	*F5*	ENSG00000198734	80	21.62
33	*FHL1*	ENSG00000022267	80	21.62
34	*SPG20*	ENSG00000133104	80	21.62
35	*TMEM150C*	ENSG00000249242	80	21.62
36	*USP1*	ENSG00000162607	80	21.62
37	*ZNF655*	ENSG00000197343	80	21.62
38	*ACSL5*	ENSG00000197142	78	21.08
39	*ETS2*	ENSG00000157557	78	21.08
40	*PTGS2*	ENSG00000073756	78	21.08
41	*CDS1*	ENSG00000163624	76	20.54
42	*ITGA5*	ENSG00000161638	76	20.54
43	*SLC36A1*	ENSG00000123643	76	20.54
44	*CCBL2*	ENSG00000137944	74	20.00
45	*CFB*	ENSG00000244255	74	20.00
46	*DDX60L*	ENSG00000181381	74	20.00
47	*FBP1*	ENSG00000165140	74	20.00
48	*LIFR*	ENSG00000113594	74	20.00
49	*SERPINE1*	ENSG00000106366	74	20.00
50	*ZDHHC17*	ENSG00000186908	74	20.00
51	*AOX1*	ENSG00000138356	72	19.46
52	*KCNC2*	ENSG00000166006	72	19.46
53	*PAMR1*	ENSG00000149090	72	19.46
54	*SLC39A10*	ENSG00000196950	72	19.46
55	*TRAM1*	ENSG00000067167	72	19.46
56	*ALDH1A3*	ENSG00000184254	70	18.92
57	*BEND4*	ENSG00000188848	70	18.92
58	*ELF3*	ENSG00000163435	70	18.92
59	*FADS1*	ENSG00000149485	70	18.92
60	*KDR*	ENSG00000128052	70	18.92
61	*MAOB*	ENSG00000069535	70	18.92
62	*REST*	ENSG00000084093	70	18.92
63	*SCIN*	ENSG00000006747	70	18.92
64	*TP63*	ENSG00000073282	70	18.92
65	*CAB39*	ENSG00000135932	68	18.38
66	*ERLIN1*	ENSG00000107566	68	18.38
67	*FAM13C*	ENSG00000148541	68	18.38
68	*GNAI1*	ENSG00000127955	68	18.38
69	*GUCY1A3*	ENSG00000164116	68	18.38

70	*INSIG1*	ENSG00000186480	68	18.38
71	*LUZP2*	ENSG00000187398	68	18.38
72	*PLA1A*	ENSG00000144837	68	18.38
73	*PTPRC*	ENSG00000081237	68	18.38
74	*RAB27A*	ENSG00000069974	68	18.38
75	*RMST*	ENSG00000255794	68	18.38
76	*RRAGD*	ENSG00000025039	68	18.38
77	*SLC35A3*	ENSG00000117620	68	18.38
78	*TSPAN8*	ENSG00000127324	68	18.38
79	*BRWD1*	ENSG00000185658	66	17.84
80	*C1QTNF3-AMACR*	ENSG00000273294	66	17.84
81	*CASD1*	ENSG00000127995	66	17.84
82	*CDK8*	ENSG00000132964	66	17.84
83	*CKAP2*	ENSG00000136108	66	17.84
84	*CSRP1*	ENSG00000159176	66	17.84
85	*FAM3B*	ENSG00000183844	66	17.84
86	*GCNT1*	ENSG00000187210	66	17.84
87	*GUCY1A2*	ENSG00000152402	66	17.84
88	*KLK11*	ENSG00000167757	66	17.84
89	*NR4A1*	ENSG00000123358	66	17.84
90	*PICALM*	ENSG00000073921	66	17.84
91	*SLC26A4*	ENSG00000091137	66	17.84
92	*SNX25*	ENSG00000109762	66	17.84
93	*STK17B*	ENSG00000081320	66	17.84
94	*ACSM3*	ENSG00000005187	64	17.30
95	*AFF3*	ENSG00000144218	64	17.30
96	*EPB41L5*	ENSG00000115109	64	17.30
97	*GCNT2*	ENSG00000111846	64	17.30
98	*IREB2*	ENSG00000136381	64	17.30
99	*NPTN*	ENSG00000156642	64	17.30
100	*PAPD4*	ENSG00000164329	64	17.30
101	*RGS1*	ENSG00000090104	64	17.30
102	*SCUBE2*	ENSG00000175356	64	17.30
103	*SMARCA1*	ENSG00000102038	64	17.30
104	*SMC6*	ENSG00000163029	64	17.30
105	*SYT1*	ENSG00000067715	64	17.30
106	*TMEM30A*	ENSG00000112697	64	17.30
107	*ZEB1*	ENSG00000148516	64	17.30
108	*ZNF385B*	ENSG00000144331	64	17.30
109	*ACADSB*	ENSG00000196177	62	16.76
110	*CPM*	ENSG00000135678	62	16.76
111	*DNASE2B*	ENSG00000137976	62	16.76
112	*GALNT3*	ENSG00000115339	62	16.76
113	*HIPK3*	ENSG00000110422	62	16.76

114	*HIVEP1*	ENSG00000095951	62	16.76
115	*KIAA1324*	ENSG00000116299	62	16.76
116	*NTN4*	ENSG00000074527	62	16.76
117	*PDE11A*	ENSG00000128655	62	16.76
118	*PGR*	ENSG00000082175	62	16.76
119	*TSC22D3*	ENSG00000157514	62	16.76
120	*DSC2*	ENSG00000134755	60	16.22
121	*FAM169A*	ENSG00000198780	60	16.22
122	*FHL2*	ENSG00000115641	60	16.22
123	*GMPS*	ENSG00000163655	60	16.22
124	*GPR160*	ENSG00000173890	60	16.22
125	*IL17RD*	ENSG00000144730	60	16.22
126	*KRT23*	ENSG00000108244	60	16.22
127	*LEPREL1*	ENSG00000090530	60	16.22
128	*LTF*	ENSG00000012223	60	16.22
129	*MMP2*	ENSG00000087245	60	16.22
130	*OGFRL1*	ENSG00000119900	60	16.22
131	*PAWR*	ENSG00000177425	60	16.22
132	*PPTC7*	ENSG00000196850	60	16.22
133	*RUFY3*	ENSG00000018189	60	16.22
134	*SBSPON*	ENSG00000164764	60	16.22
135	*SERINC2*	ENSG00000168528	60	16.22
136	*ZNF652*	ENSG00000198740	60	16.22
137	*CALD1*	ENSG00000122786	58	15.68
138	*CD2AP*	ENSG00000198087	58	15.68
139	*COBLL1*	ENSG00000082438	58	15.68
140	*COCH*	ENSG00000100473	58	15.68
141	*DMD*	ENSG00000198947	58	15.68
142	*EPT1*	ENSG00000138018	58	15.68
143	*GTPBP10*	ENSG00000105793	58	15.68
144	*LDLR*	ENSG00000130164	58	15.68
145	*MBOAT1*	ENSG00000172197	58	15.68
146	*SGK1*	ENSG00000118515	58	15.68
147	*SLC5A1*	ENSG00000100170	58	15.68
148	*SMC4*	ENSG00000113810	58	15.68
149	*ST8SIA6*	ENSG00000148488	58	15.68
150	*URI1*	ENSG00000105176	58	15.68
151	*VAT1*	ENSG00000108828	58	15.68
152	*ZNF639*	ENSG00000121864	58	15.68
153	*ACBD3*	ENSG00000182827	56	15.14
154	*ANTXR2*	ENSG00000163297	56	15.14
155	*CCNG1*	ENSG00000113328	56	15.14
156	*CHRNA5*	ENSG00000169684	56	15.14
157	*CMTM4*	ENSG00000183723	56	15.14

158	*CNTN1*	ENSG00000018236	56	15.14
159	*FBXL5*	ENSG00000118564	56	15.14
160	*FERMT2*	ENSG00000073712	56	15.14
161	*FKBP5*	ENSG00000096060	56	15.14
162	*GCLC*	ENSG00000001084	56	15.14
163	*HNRNPLL*	ENSG00000143889	56	15.14
164	*KL*	ENSG00000133116	56	15.14
165	*NAMPT*	ENSG00000105835	56	15.14
166	*PI15*	ENSG00000137558	56	15.14
167	*PIGR*	ENSG00000162896	56	15.14
168	*RMI1*	ENSG00000178966	56	15.14
169	*ROCK2*	ENSG00000134318	56	15.14
170	*SNX16*	ENSG00000104497	56	15.14
171	*SPARC*	ENSG00000113140	56	15.14
172	*TMPRSS2*	ENSG00000184012	56	15.14
173	*TRPC4*	ENSG00000133107	56	15.14
174	*WDR36*	ENSG00000134987	56	15.14
175	*ADAM17*	ENSG00000151694	54	14.59
176	*ARID4A*	ENSG00000032219	54	14.59
177	*CNTNAP2*	ENSG00000174469	54	14.59
178	*ETNK1*	ENSG00000139163	54	14.59
179	*FAM117B*	ENSG00000138439	54	14.59
180	*FASTKD2*	ENSG00000118246	54	14.59
181	*FYTTD1*	ENSG00000122068	54	14.59
182	*GXYLT1*	ENSG00000151233	54	14.59
183	*IP6K2*	ENSG00000068745	54	14.59
184	*ITGB6*	ENSG00000115221	54	14.59
185	*KATNAL1*	ENSG00000102781	54	14.59
186	*NCOA7*	ENSG00000111912	54	14.59
187	*OPRK1*	ENSG00000082556	54	14.59
188	*PDIA6*	ENSG00000143870	54	14.59
189	*PIGK*	ENSG00000142892	54	14.59
190	*REV3L*	ENSG00000009413	54	14.59
191	*SELE*	ENSG00000007908	54	14.59
192	*STEAP2*	ENSG00000157214	54	14.59
193	*STX3*	ENSG00000166900	54	14.59
194	*TMEM181*	ENSG00000146433	54	14.59
195	*TMX4*	ENSG00000125827	54	14.59
196	*TRPC1*	ENSG00000144935	54	14.59
197	*VGLL3*	ENSG00000206538	54	14.59
198	*VRK2*	ENSG00000028116	54	14.59
199	*ACSL1*	ENSG00000151726	52	14.05
200	*BANK1*	ENSG00000153064	52	14.05

Table A.5 Top 200 candidates with the largest number of step-up jumps in the MSKCC dataset.

Rank	Gene	Gene ID	Count	Percent.
1	*OLFM4*	ENSG00000102837	112	47.86
2	*PROM1*	ENSG00000007062	98	41.88
3	*ANPEP*	ENSG00000166825	94	40.17
4	*TP63*	ENSG00000073282	90	38.46
5	*NCOA7*	ENSG00000111912	87	37.18
6	*SELE*	ENSG00000007908	86	36.75
7	*CFB*	ENSG00000244255	84	35.90
8	*ADD3*	ENSG00000148700	82	35.04
9	*ATP8B4*	ENSG00000104043	81	34.62
10	*KRT23*	ENSG00000108244	81	34.62
11	*TGM4*	ENSG00000163810	78	33.33
12	*LTF*	ENSG00000012223	75	32.05
13	*GABRP*	ENSG00000094755	74	31.62
14	*SYNM*	ENSG00000182253	74	31.62
15	*CFTR*	ENSG00000001626	73	31.20
16	*ETS2*	ENSG00000157557	73	31.20
17	*CHRDL1*	ENSG00000101938	72	30.77
18	*CYP3A5*	ENSG00000106258	72	30.77
19	*ITGB6*	ENSG00000115221	72	30.77
20	*LRRC9*	ENSG00000131951	72	30.77
21	*SLC18A2*	ENSG00000165646	72	30.77
22	*ACSS3*	ENSG00000111058	71	30.34
23	*ANXA2*	ENSG00000182718	71	30.34
24	*CCDC80*	ENSG00000091986	70	29.91
25	*NRK*	ENSG00000123572	69	29.49
26	*PDE11A*	ENSG00000128655	69	29.49
27	*PTGS2*	ENSG00000073756	69	29.49
28	*SBSPON*	ENSG00000164764	69	29.49
29	*SCUBE2*	ENSG00000175356	69	29.49
30	*C1S*	ENSG00000182326	68	29.06
31	*FMO2*	ENSG00000094963	67	28.63
32	*LEPREL1*	ENSG00000090530	67	28.63
33	*GCNT2*	ENSG00000111846	65	27.78
34	*KIAA1210*	ENSG00000250423	65	27.78
35	*MME*	ENSG00000196549	65	27.78
36	*NRG4*	ENSG00000169752	65	27.78
37	*TSC22D3*	ENSG00000157514	65	27.78
38	*DDR2*	ENSG00000162733	63	26.92
39	*DMD*	ENSG00000198947	63	26.92

40	*AFF3*	ENSG00000144218	61	26.07
41	*ITGA8*	ENSG00000077943	61	26.07
42	*SELP*	ENSG00000174175	61	26.07
43	*NLRP2*	ENSG00000022556	60	25.64
44	*RARRES1*	ENSG00000118849	60	25.64
45	*VGLL3*	ENSG00000206538	60	25.64
46	*CYP4B1*	ENSG00000142973	59	25.21
47	*KIT*	ENSG00000157404	59	25.21
48	*PGAP1*	ENSG00000197121	59	25.21
49	*ROCK2*	ENSG00000134318	59	25.21
50	*ERAP2*	ENSG00000164308	58	24.79
51	*NPR3*	ENSG00000113389	58	24.79
52	*ST8SIA6*	ENSG00000148488	58	24.79
53	*KDR*	ENSG00000128052	57	24.36
54	*CHL1*	ENSG00000134121	56	23.93
55	*SGK1*	ENSG00000118515	56	23.93
56	*SRD5A2*	ENSG00000277893	56	23.93
57	*FHL1*	ENSG00000022267	55	23.50
58	*LINC00668*	ENSG00000265933	55	23.50
59	*NR4A1*	ENSG00000123358	55	23.50
60	*TNRC6C*	ENSG00000078687	55	23.50
61	*C6ORF174*	ENSG00000255330	54	23.08
62	*FADS1*	ENSG00000149485	54	23.08
63	*RNF128*	ENSG00000133135	54	23.08
64	*SAMSN1*	ENSG00000155307	54	23.08
65	*SOX5*	ENSG00000134532	54	23.08
66	*CCDC68*	ENSG00000166510	53	22.65
67	*PRIM2*	ENSG00000146143	53	22.65
68	*TGFB3*	ENSG00000119699	52	22.22
69	*ADHFE1*	ENSG00000147576	51	21.79
70	*ANTXR2*	ENSG00000163297	51	21.79
71	*DSC3*	ENSG00000134762	51	21.79
72	*KCTD14*	ENSG00000151364	51	21.79
73	*RCAN3*	ENSG00000117602	51	21.79
74	*ZDHHC8P1*	ENSG00000133519	51	21.79
75	*RMST*	ENSG00000255794	50	21.37
76	*ZNF655*	ENSG00000197343	50	21.37
77	*ALOX12P2*	ENSG00000262943	49	20.94
78	*FAM83D*	ENSG00000101447	49	20.94
79	*GBP2*	ENSG00000162645	49	20.94
80	*LRCH2*	ENSG00000130224	49	20.94
81	*PRKCD*	ENSG00000163932	49	20.94
82	*STON1-GTF2A1L*	ENSG00000068781	49	20.94
83	*STXBP5L*	ENSG00000145087	49	20.94

84	*ALDH1A2*	ENSG00000128918	48	20.51
85	*COCH*	ENSG00000100473	48	20.51
86	*FERMT2*	ENSG00000073712	48	20.51
87	*HEPH*	ENSG00000089472	48	20.51
88	*TMEM178A*	ENSG00000152154	48	20.51
89	*CCDC178*	ENSG00000166960	47	20.09
90	*SPG20*	ENSG00000133104	47	20.09
91	*TLR1*	ENSG00000174125	47	20.09
92	*ZDHHC17*	ENSG00000186908	47	20.09
93	*ASPA*	ENSG00000108381	46	19.66
94	*IGSF1*	ENSG00000147255	46	19.66
95	*MSMO1*	ENSG00000052802	46	19.66
96	*CALD1*	ENSG00000122786	45	19.23
97	*COL14A1*	ENSG00000187955	45	19.23
98	*CPNE4*	ENSG00000196353	45	19.23
99	*LSAMP*	ENSG00000185565	45	19.23
100	*MAN1A1*	ENSG00000111885	45	19.23
101	*NNMT*	ENSG00000166741	45	19.23
102	*GREB1*	ENSG00000196208	44	18.80
103	*ITGA5*	ENSG00000161638	44	18.80
104	*MYZAP*	ENSG00000263155	44	18.80
105	*PAX9*	ENSG00000198807	44	18.80
106	*SLC22A3*	ENSG00000146477	44	18.80
107	*ATF3*	ENSG00000162772	43	18.38
108	*CLIP4*	ENSG00000115295	43	18.38
109	*PHACTR4*	ENSG00000204138	43	18.38
110	*PRKG1*	ENSG00000185532	43	18.38
111	*RP11-307N16.6*	ENSG00000273167	43	18.38
112	*SHISA9*	ENSG00000237515	43	18.38
113	*C2orf88*	ENSG00000187699	42	17.95
114	*CYP4F11*	ENSG00000171903	42	17.95
115	*ETS1*	ENSG00000134954	42	17.95
116	*INSIG1*	ENSG00000186480	42	17.95
117	*LPAR1*	ENSG00000198121	42	17.95
118	*NDC80*	ENSG00000080986	42	17.95
119	*POF1B*	ENSG00000124429	42	17.95
120	*PPARGC1A*	ENSG00000109819	42	17.95
121	*SV2B*	ENSG00000185518	42	17.95
122	*AF131217.1*	ENSG00000232855	41	17.52
123	*ARHGAP20*	ENSG00000137727	41	17.52
124	*PDK4*	ENSG00000004799	41	17.52
125	*PI15*	ENSG00000137558	41	17.52
126	*PTPLA*	ENSG00000165996	41	17.52
127	*SCNN1A*	ENSG00000111319	41	17.52

128	*VCAM1*	ENSG00000162692	41	17.52
129	*ACSL5*	ENSG00000197142	40	17.09
130	*ATRNL1*	ENSG00000107518	40	17.09
131	*C3*	ENSG00000125730	40	17.09
132	*CSRP1*	ENSG00000159176	40	17.09
133	*DPYS*	ENSG00000147647	40	17.09
134	*GCLC*	ENSG00000001084	40	17.09
135	*MYBPC1*	ENSG00000196091	40	17.09
136	*SCIN*	ENSG00000006747	40	17.09
137	*SFMBT2*	ENSG00000198879	40	17.09
138	*SLC14A1*	ENSG00000141469	40	17.09
139	*SLC2A14*	ENSG00000173262	40	17.09
140	*SPAG6*	ENSG00000077327	40	17.09
141	*CELF2*	ENSG00000048740	39	16.67
142	*CPLX3*	ENSG00000213578	39	16.67
143	*CRYAB*	ENSG00000109846	39	16.67
144	*INPP4B*	ENSG00000109452	39	16.67
145	*KL*	ENSG00000133116	39	16.67
146	*MR1*	ENSG00000153029	39	16.67
147	*NOSTRIN*	ENSG00000163072	39	16.67
148	*PLAGL1*	ENSG00000118495	39	16.67
149	*RAB27A*	ENSG00000069974	39	16.67
150	*RGS1*	ENSG00000090104	39	16.67
151	*TSC22D1*	ENSG00000102804	39	16.67
152	*CD38*	ENSG00000004468	38	16.24
153	*CNTNAP2*	ENSG00000174469	38	16.24
154	*FREM2*	ENSG00000150893	38	16.24
155	*MUC13*	ENSG00000173702	38	16.24
156	*NRXN3*	ENSG00000021645	38	16.24
157	*ABCB1*	ENSG00000085563	37	15.81
158	*F13A1*	ENSG00000124491	37	15.81
159	*HSPA4L*	ENSG00000164070	37	15.81
160	*NTN4*	ENSG00000074527	37	15.81
161	*PDE1A*	ENSG00000115252	37	15.81
162	*RNF150*	ENSG00000170153	37	15.81
163	*SYT1*	ENSG00000067715	37	15.81
164	*TMPRSS2*	ENSG00000184012	37	15.81
165	*VMP1*	ENSG00000062716	37	15.81
166	*AKAP7*	ENSG00000118507	36	15.38
167	*BIRC3*	ENSG00000023445	36	15.38
168	*CCDC169-SOHLH2*	ENSG00000250709	36	15.38
169	*COG6*	ENSG00000133103	36	15.38
170	*F3*	ENSG00000117525	36	15.38
171	*LMAN1L*	ENSG00000140506	36	15.38

172	*PAH*	ENSG00000171759	36	15.38
173	*PON1*	ENSG00000005421	36	15.38
174	*PRKAR2B*	ENSG00000005249	36	15.38
175	*RHOF*	ENSG00000139725	36	15.38
176	*SSBP2*	ENSG00000145687	36	15.38
177	*ANO4*	ENSG00000151572	35	14.96
178	*CFB*	ENSG00000243649	35	14.96
179	*GATM*	ENSG00000171766	35	14.96
180	*JAM3*	ENSG00000166086	35	14.96
181	*LPHN2*	ENSG00000117114	35	14.96
182	*NAALAD2*	ENSG00000077616	35	14.96
183	*PDE4B*	ENSG00000184588	35	14.96
184	*RSPO2*	ENSG00000147655	35	14.96
185	*SOD2*	ENSG00000112096	35	14.96
186	*ANXA1*	ENSG00000135046	34	14.53
187	*ATP1B1*	ENSG00000143153	34	14.53
188	*CAV2*	ENSG00000105971	34	14.53
189	*CPA6*	ENSG00000165078	34	14.53
190	*IL7R*	ENSG00000168685	34	14.53
191	*KIAA1324L*	ENSG00000164659	34	14.53
192	*PAPD4*	ENSG00000164329	34	14.53
193	*PDE4D*	ENSG00000113448	34	14.53
194	*RP11-624L4.1*	ENSG00000259345	34	14.53
195	*SGIP1*	ENSG00000118473	34	14.53
196	*SLC16A5*	ENSG00000170190	34	14.53
197	*SYNE1*	ENSG00000131018	34	14.53
198	*ARG2*	ENSG00000081181	33	14.10
199	*CLGN*	ENSG00000153132	33	14.10
200	*GPX8*	ENSG00000164294	33	14.10

Table A.6 Top 200 candidates with the largest number of step-down jumps in the CancerMap dataset.

Rank	**Gene**	**Gene ID**	**Count**	**Percent.**
1	*OLFM4*	ENSG00000102837	140	37.84
2	*TDRD1*	ENSG00000095627	134	36.22
3	*MYBPC1*	ENSG00000196091	100	27.03
4	*SYNM*	ENSG00000182253	100	27.03
5	*ATF3*	ENSG00000162772	96	25.95
6	*C1QTNF3-AMACR*	ENSG00000273294	94	25.41
7	*KRT23*	ENSG00000108244	94	25.41
8	*PLA2G7*	ENSG00000146070	94	25.41
9	*CHRDL1*	ENSG00000101938	92	24.86

10	*F3*	ENSG00000117525	92	24.86
11	*TIMP3*	ENSG00000100234	92	24.86
12	*EDNRB*	ENSG00000136160	86	23.24
13	*HSD17B6*	ENSG00000025423	86	23.24
14	*CSRP1*	ENSG00000159176	82	22.16
15	*TIPARP*	ENSG00000163659	82	22.16
16	*TMPRSS2*	ENSG00000184012	82	22.16
17	*SLC22A3*	ENSG00000146477	80	21.62
18	*SLC38A11*	ENSG00000169507	80	21.62
19	*FAM3B*	ENSG00000183844	76	20.54
20	*MAN1A1*	ENSG00000111885	76	20.54
21	*MME*	ENSG00000196549	76	20.54
22	*PTGS2*	ENSG00000073756	76	20.54
23	*RMST*	ENSG00000255794	76	20.54
24	*CPM*	ENSG00000135678	74	20.00
25	*GCNT1*	ENSG00000187210	74	20.00
26	*KIAA1324*	ENSG00000116299	74	20.00
27	*MAOB*	ENSG00000069535	74	20.00
28	*SESN3*	ENSG00000149212	74	20.00
29	*SRD5A2*	ENSG00000277893	74	20.00
30	*LUZP2*	ENSG00000187398	72	19.46
31	*ZNF655*	ENSG00000197343	72	19.46
32	*FHL1*	ENSG00000022267	70	18.92
33	*LIFR*	ENSG00000113594	70	18.92
34	*PI15*	ENSG00000137558	70	18.92
35	*TP63*	ENSG00000073282	70	18.92
36	*ACSL5*	ENSG00000197142	68	18.38
37	*DDX60L*	ENSG00000181381	68	18.38
38	*DSC3*	ENSG00000134762	68	18.38
39	*PTPRC*	ENSG00000081237	68	18.38
40	*ROCK2*	ENSG00000134318	68	18.38
41	*SLC39A10*	ENSG00000196950	68	18.38
42	*ZDHHC17*	ENSG00000186908	68	18.38
43	*AFF3*	ENSG00000144218	66	17.84
44	*PDIA6*	ENSG00000143870	66	17.84
45	*PRKCD*	ENSG00000163932	66	17.84
46	*PUM2*	ENSG00000055917	66	17.84
47	*SLC12A2*	ENSG00000064651	66	17.84
48	*SLC35A3*	ENSG00000117620	66	17.84
49	*SNX25*	ENSG00000109762	66	17.84
50	*ACSM3*	ENSG00000005187	64	17.30
51	*ATP1B1*	ENSG00000143153	64	17.30
52	*ETS2*	ENSG00000157557	64	17.30
53	*INSIG1*	ENSG00000186480	64	17.30

54	*PDK4*	ENSG00000004799	64	17.30
55	*SLC14A1*	ENSG00000141469	64	17.30
56	*TC2N*	ENSG00000165929	64	17.30
57	*TMEM150C*	ENSG00000249242	64	17.30
58	*TSPAN8*	ENSG00000127324	64	17.30
59	*ACADL*	ENSG00000115361	62	16.76
60	*ANPEP*	ENSG00000166825	62	16.76
61	*ANTXR2*	ENSG00000163297	62	16.76
62	*CRISP3*	ENSG00000096006	62	16.76
63	*FERMT2*	ENSG00000073712	62	16.76
64	*LDLR*	ENSG00000130164	62	16.76
65	*MORF4L1*	ENSG00000185787	62	16.76
66	*NR4A1*	ENSG00000123358	62	16.76
67	*PANK3*	ENSG00000120137	62	16.76
68	*PDE11A*	ENSG00000128655	62	16.76
69	*SLC19A2*	ENSG00000117479	62	16.76
70	*STEAP2*	ENSG00000157214	62	16.76
71	*THSD7A*	ENSG00000005108	62	16.76
72	*TSC22D3*	ENSG00000157514	62	16.76
73	*ZEB1*	ENSG00000148516	62	16.76
74	*ALDH1A1*	ENSG00000165092	60	16.22
75	*ALDH1A3*	ENSG00000184254	60	16.22
76	*CALD1*	ENSG00000122786	60	16.22
77	*CDS1*	ENSG00000163624	60	16.22
78	*FAM135A*	ENSG00000082269	60	16.22
79	*IREB2*	ENSG00000136381	60	16.22
80	*LPHN2*	ENSG00000117114	60	16.22
81	*MMP2*	ENSG00000087245	60	16.22
82	*NR4A2*	ENSG00000153234	60	16.22
83	*SELE*	ENSG00000007908	60	16.22
84	*SPARCL1*	ENSG00000152583	60	16.22
85	*VPS26A*	ENSG00000122958	60	16.22
86	*BEND4*	ENSG00000188848	58	15.68
87	*CNN1*	ENSG00000130176	58	15.68
88	*CNTN1*	ENSG00000018236	58	15.68
89	*EPHA3*	ENSG00000044524	58	15.68
90	*LY75-CD302*	ENSG00000248672	58	15.68
91	*RGS1*	ENSG00000090104	58	15.68
92	*SMARCA1*	ENSG00000102038	58	15.68
93	*SPG20*	ENSG00000133104	58	15.68
94	*TMEM178A*	ENSG00000152154	58	15.68
95	*USP1*	ENSG00000162607	58	15.68
96	*WWTR1*	ENSG00000018408	58	15.68
97	*ACSL1*	ENSG00000151726	56	15.14

98	*ATL3*	ENSG00000184743	56	15.14
99	*GNAI1*	ENSG00000127955	56	15.14
100	*NTN4*	ENSG00000074527	56	15.14
101	*NUP205*	ENSG00000155561	56	15.14
102	*PLA1A*	ENSG00000144837	56	15.14
103	*SLC30A9*	ENSG00000014824	56	15.14
104	*SRPRB*	ENSG00000144867	56	15.14
105	*ATP2B4*	ENSG00000058668	54	14.59
106	*CAB39*	ENSG00000135932	54	14.59
107	*CAV2*	ENSG00000105971	54	14.59
108	*CCBL2*	ENSG00000137944	54	14.59
109	*CPNE4*	ENSG00000196353	54	14.59
110	*DDX1*	ENSG00000079785	54	14.59
111	*EPCAM*	ENSG00000119888	54	14.59
112	*FASTKD2*	ENSG00000118246	54	14.59
113	*GALNT3*	ENSG00000115339	54	14.59
114	*IL1R1*	ENSG00000115594	54	14.59
115	*KLK11*	ENSG00000167757	54	14.59
116	*MFSD8*	ENSG00000164073	54	14.59
117	*NFKBIZ*	ENSG00000144802	54	14.59
118	*PDE4B*	ENSG00000184588	54	14.59
119	*RAB27A*	ENSG00000069974	54	14.59
120	*SLC36A1*	ENSG00000123643	54	14.59
121	*SPARC*	ENSG00000113140	54	14.59
122	*STX3*	ENSG00000166900	54	14.59
123	*VRK2*	ENSG00000028116	54	14.59
124	*WDR36*	ENSG00000134987	54	14.59
125	*WEE1*	ENSG00000166483	54	14.59
126	*ACADSB*	ENSG00000196177	52	14.05
127	*ACTG2*	ENSG00000163017	52	14.05
128	*ANXA1*	ENSG00000135046	52	14.05
129	*AOX1*	ENSG00000138356	52	14.05
130	*BRWD1*	ENSG00000185658	52	14.05
131	*COL14A1*	ENSG00000187955	52	14.05
132	*DDR2*	ENSG00000162733	52	14.05
133	*EPB41L5*	ENSG00000115109	52	14.05
134	*EPT1*	ENSG00000138018	52	14.05
135	*FADS1*	ENSG00000149485	52	14.05
136	*FAM169A*	ENSG00000198780	52	14.05
137	*GPR160*	ENSG00000173890	52	14.05
138	*GUCY1A2*	ENSG00000152402	52	14.05
139	*GUCY1A3*	ENSG00000164116	52	14.05
140	*LTBP1*	ENSG00000049323	52	14.05
141	*MAOA*	ENSG00000189221	52	14.05

142	*NDRG1*	ENSG00000104419	52	14.05
143	*NPTN*	ENSG00000156642	52	14.05
144	*PALLD*	ENSG00000129116	52	14.05
145	*PARVA*	ENSG00000197702	52	14.05
146	*RABGGTB*	ENSG00000137955	52	14.05
147	*SFRP4*	ENSG00000106483	52	14.05
148	*SGK1*	ENSG00000118515	52	14.05
149	*SLC17A5*	ENSG00000119899	52	14.05
150	*SMC6*	ENSG00000163029	52	14.05
151	*STON1*	ENSG00000243244	52	14.05
152	*TMX4*	ENSG00000125827	52	14.05
153	*TRAM1*	ENSG00000067167	52	14.05
154	*TRIM29*	ENSG00000137699	52	14.05
155	*ADD3*	ENSG00000148700	50	13.51
156	*CLHC1*	ENSG00000162994	50	13.51
157	*CLIP4*	ENSG00000115295	50	13.51
158	*CWH43*	ENSG00000109182	50	13.51
159	*DCTD*	ENSG00000129187	50	13.51
160	*DES*	ENSG00000175084	50	13.51
161	*ENDOD1*	ENSG00000149218	50	13.51
162	*FHL2*	ENSG00000115641	50	13.51
163	*FYTTD1*	ENSG00000122068	50	13.51
164	*GREB1*	ENSG00000196208	50	13.51
165	*LEPREL1*	ENSG00000090530	50	13.51
166	*NAMPT*	ENSG00000105835	50	13.51
167	*NFIB*	ENSG00000147862	50	13.51
168	*OMA1*	ENSG00000162600	50	13.51
169	*PHF6*	ENSG00000156531	50	13.51
170	*PICALM*	ENSG00000073921	50	13.51
171	*SMAD4*	ENSG00000141646	50	13.51
172	*TAF1B*	ENSG00000115750	50	13.51
173	*TMEM181*	ENSG00000146433	50	13.51
174	*ABCC4*	ENSG00000125257	48	12.97
175	*ACBD3*	ENSG00000182827	48	12.97
176	*AZGP1*	ENSG00000160862	48	12.97
177	*DSC2*	ENSG00000134755	48	12.97
178	*EDNRA*	ENSG00000151617	48	12.97
179	*ERLIN1*	ENSG00000107566	48	12.97
180	*F2R*	ENSG00000181104	48	12.97
181	*FBP1*	ENSG00000165140	48	12.97
182	*FBXL5*	ENSG00000118564	48	12.97
183	*FKBP5*	ENSG00000096060	48	12.97
184	*GBP2*	ENSG00000162645	48	12.97
185	*GCNT2*	ENSG00000111846	48	12.97

186	*HIPK3*	ENSG00000110422	48	12.97
187	*HNRNPLL*	ENSG00000143889	48	12.97
188	*ITGA1*	ENSG00000213949	48	12.97
189	*ITGA5*	ENSG00000161638	48	12.97
190	*KCNC2*	ENSG00000166006	48	12.97
191	*KIAA1033*	ENSG00000136051	48	12.97
192	*MAP1B*	ENSG00000131711	48	12.97
193	*MBOAT2*	ENSG00000143797	48	12.97
194	*NR3C1*	ENSG00000113580	48	12.97
195	*PDE8B*	ENSG00000113231	48	12.97
196	*PTBP3*	ENSG00000119314	48	12.97
197	*RAB3B*	ENSG00000169213	48	12.97
198	*RABEP1*	ENSG00000029725	48	12.97
199	*RBM27*	ENSG00000091009	48	12.97
200	*SLC30A4*	ENSG00000104154	48	12.97

Table A.7 Top 200 candidates with the largest number of step-down jumps in the MSKCC dataset.

Gene	**Log-rank *p*-val**	**FDR adj. *p*-val**
DEPDC1B	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{6.99 \cdot 10^{-13}}$
KIF15	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{6.99 \cdot 10^{-13}}$
TF	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{6.99 \cdot 10^{-13}}$
BUB1	$\mathbf{1.94 \cdot 10^{-12}}$	$\mathbf{1.84 \cdot 10^{-9}}$
PSMA3	$\mathbf{1.02 \cdot 10^{-9}}$	$\mathbf{7.67 \cdot 10^{-7}}$
PDK4	$\mathbf{1.86 \cdot 10^{-8}}$	$\mathbf{1 \cdot 10^{-5}}$
SLC22A23	$\mathbf{1.86 \cdot 10^{-8}}$	$\mathbf{1 \cdot 10^{-5}}$
UBAP2L	$\mathbf{1.59 \cdot 10^{-7}}$	$\mathbf{7.48 \cdot 10^{-5}}$
MEGF9	$\mathbf{1.97 \cdot 10^{-7}}$	$\mathbf{8.27 \cdot 10^{-5}}$
KIF14	$\mathbf{2.76 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
SYCP2	$\mathbf{4.55 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
BTBD8	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
CNOT2	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
DDHD2	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
DPP8	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
HNRNPDL	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
HTATIP2	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
MAPK8	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
NCOR1	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
SLBP	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
TUBE1	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
VPS4B	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$
ZC3H14	$\mathbf{6.29 \cdot 10^{-7}}$	$\mathbf{1.03 \cdot 10^{-4}}$

LRIG2	**$7.28 \cdot 10^{-7}$**	**$1.15 \cdot 10^{-4}$**
PCNT	**$5.55 \cdot 10^{-6}$**	**$8.38 \cdot 10^{-4}$**
POFUT1	**$1.3 \cdot 10^{-5}$**	**$1.82 \cdot 10^{-3}$**
ZNF217	**$1.3 \cdot 10^{-5}$**	**$1.82 \cdot 10^{-3}$**
RACGAP1	**$1.6 \cdot 10^{-5}$**	**$2.16 \cdot 10^{-3}$**
GTPBP10	**$1.95 \cdot 10^{-5}$**	**$2.54 \cdot 10^{-3}$**
EXO1	**$2.08 \cdot 10^{-5}$**	**$2.62 \cdot 10^{-3}$**
GFM2	**$2.94 \cdot 10^{-5}$**	**$3.58 \cdot 10^{-3}$**
PRPF40A	**$3.14 \cdot 10^{-5}$**	**$3.69 \cdot 10^{-3}$**
MYBL2	**$3.22 \cdot 10^{-5}$**	**$3.69 \cdot 10^{-3}$**
RPL6	**$7.44 \cdot 10^{-5}$**	**$8.26 \cdot 10^{-3}$**
FBN1	**$1.29 \cdot 10^{-4}$**	**$1.32 \cdot 10^{-2}$**
KIF5C	**$1.29 \cdot 10^{-4}$**	**$1.32 \cdot 10^{-2}$**
PDE9A	**$1.29 \cdot 10^{-4}$**	**$1.32 \cdot 10^{-2}$**
CENPP	**$1.6 \cdot 10^{-4}$**	**$1.59 \cdot 10^{-2}$**
RGS5	**$1.81 \cdot 10^{-4}$**	**$1.75 \cdot 10^{-2}$**
ZKSCAN1	**$1.87 \cdot 10^{-4}$**	**$1.76 \cdot 10^{-2}$**
CTC.360G5.8	**$2.21 \cdot 10^{-4}$**	**$2 \cdot 10^{-2}$**
ATP6V0A2	**$2.22 \cdot 10^{-4}$**	**$2 \cdot 10^{-2}$**
GSR	**$2.54 \cdot 10^{-4}$**	**$2.23 \cdot 10^{-2}$**
FAM65B	**$3 \cdot 10^{-4}$**	**$2.57 \cdot 10^{-2}$**
CPNE4	**$3.38 \cdot 10^{-4}$**	**$2.84 \cdot 10^{-2}$**
KLHDC3	**$3.56 \cdot 10^{-4}$**	**$2.86 \cdot 10^{-2}$**
NNT	**$3.56 \cdot 10^{-4}$**	**$2.86 \cdot 10^{-2}$**
EPHA3	**$4.45 \cdot 10^{-4}$**	**$3.43 \cdot 10^{-2}$**
SREK1IP1	**$4.45 \cdot 10^{-4}$**	**$3.43 \cdot 10^{-2}$**
ZNF614	**$4.54 \cdot 10^{-4}$**	**$3.43 \cdot 10^{-2}$**
BUB1B	**$6.62 \cdot 10^{-4}$**	**$4.82 \cdot 10^{-2}$**
TP53	**$6.64 \cdot 10^{-4}$**	**$4.82 \cdot 10^{-2}$**
IFT88	**$7.13 \cdot 10^{-4}$**	$5.08 \cdot 10^{-2}$
DHX57	**$8.06 \cdot 10^{-4}$**	$5.64 \cdot 10^{-2}$
PRKAR1A	**$8.4 \cdot 10^{-4}$**	$5.77 \cdot 10^{-2}$
PLCL1	**$8.67 \cdot 10^{-4}$**	$5.84 \cdot 10^{-2}$
SKA3	**$1.12 \cdot 10^{-3}$**	$7.4 \cdot 10^{-2}$
ARID5B	**$1.18 \cdot 10^{-3}$**	$7.68 \cdot 10^{-2}$
ERBB2	**$1.23 \cdot 10^{-3}$**	$7.84 \cdot 10^{-2}$
CHEK1	**$1.35 \cdot 10^{-3}$**	$8.49 \cdot 10^{-2}$
CPNE1	**$1.41 \cdot 10^{-3}$**	$8.74 \cdot 10^{-2}$
C1R	**$1.47 \cdot 10^{-3}$**	$8.92 \cdot 10^{-2}$
MBNL1	**$1.56 \cdot 10^{-3}$**	$9.32 \cdot 10^{-2}$
ASPM	**$1.61 \cdot 10^{-3}$**	$9.49 \cdot 10^{-2}$
ADAMTS18	**$1.78 \cdot 10^{-3}$**	$1.03 \cdot 10^{-1}$
TGM4	**$1.8 \cdot 10^{-3}$**	$1.03 \cdot 10^{-1}$
FBXL17	**$1.84 \cdot 10^{-3}$**	$1.04 \cdot 10^{-1}$

SMARCD2	$\mathbf{1.88 \cdot 10^{-3}}$	$1.04 \cdot 10^{-1}$
GOSR1	$\mathbf{2.01 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}$
BBS10	$\mathbf{2.06 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}$
ETAA1	$\mathbf{2.06 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}$
ZNF561	$\mathbf{2.06 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}$
SNAP25	$\mathbf{2.25 \cdot 10^{-3}}$	$1.16 \cdot 10^{-1}$
MTHFD1L	$\mathbf{2.4 \cdot 10^{-3}}$	$1.22 \cdot 10^{-1}$
MCM3	$\mathbf{2.47 \cdot 10^{-3}}$	$1.23 \cdot 10^{-1}$
SGOL1	$\mathbf{2.47 \cdot 10^{-3}}$	$1.23 \cdot 10^{-1}$
PARPBP	$\mathbf{2.61 \cdot 10^{-3}}$	$1.28 \cdot 10^{-1}$
MROH7.TTC4	$\mathbf{2.72 \cdot 10^{-3}}$	$1.3 \cdot 10^{-1}$
UBE2T	$\mathbf{2.72 \cdot 10^{-3}}$	$1.3 \cdot 10^{-1}$
KCNJ3	$\mathbf{2.76 \cdot 10^{-3}}$	$1.3 \cdot 10^{-1}$
HNRNPM	$\mathbf{2.92 \cdot 10^{-3}}$	$1.36 \cdot 10^{-1}$
CDCA7	$\mathbf{3.12 \cdot 10^{-3}}$	$1.43 \cdot 10^{-1}$
FANCB	$\mathbf{3.18 \cdot 10^{-3}}$	$1.43 \cdot 10^{-1}$
AFF1	$\mathbf{3.19 \cdot 10^{-3}}$	$1.43 \cdot 10^{-1}$
CDC6	$\mathbf{3.27 \cdot 10^{-3}}$	$1.44 \cdot 10^{-1}$
PRC1	$\mathbf{3.29 \cdot 10^{-3}}$	$1.44 \cdot 10^{-1}$
CDC25A	$\mathbf{3.63 \cdot 10^{-3}}$	$1.56 \cdot 10^{-1}$
CENPU	$\mathbf{3.64 \cdot 10^{-3}}$	$1.56 \cdot 10^{-1}$
ANLN	$\mathbf{3.68 \cdot 10^{-3}}$	$1.56 \cdot 10^{-1}$
NCAPG2	$\mathbf{3.81 \cdot 10^{-3}}$	$1.6 \cdot 10^{-1}$
SNAP91	$\mathbf{4.35 \cdot 10^{-3}}$	$1.8 \cdot 10^{-1}$
ESCO2	$\mathbf{4.67 \cdot 10^{-3}}$	$1.92 \cdot 10^{-1}$
OSBPL10	$\mathbf{4.98 \cdot 10^{-3}}$	$2.02 \cdot 10^{-1}$
SLC10A7	$\mathbf{5.25 \cdot 10^{-3}}$	$2.11 \cdot 10^{-1}$
CXorf22	$\mathbf{5.56 \cdot 10^{-3}}$	$2.21 \cdot 10^{-1}$
POLQ	$\mathbf{5.92 \cdot 10^{-3}}$	$2.31 \cdot 10^{-1}$
DTWD1	$\mathbf{5.98 \cdot 10^{-3}}$	$2.31 \cdot 10^{-1}$
NTNG1	$\mathbf{6.05 \cdot 10^{-3}}$	$2.31 \cdot 10^{-1}$
RPN1	$\mathbf{6.05 \cdot 10^{-3}}$	$2.31 \cdot 10^{-1}$
DKC1	$\mathbf{6.88 \cdot 10^{-3}}$	$2.6 \cdot 10^{-1}$
ATP12A	$\mathbf{7.14 \cdot 10^{-3}}$	$2.67 \cdot 10^{-1}$
PRAME	$\mathbf{7.5 \cdot 10^{-3}}$	$2.75 \cdot 10^{-1}$
BEND3	$\mathbf{7.5 \cdot 10^{-3}}$	$2.75 \cdot 10^{-1}$
ACADM	$\mathbf{7.73 \cdot 10^{-3}}$	$2.78 \cdot 10^{-1}$
SPATA20	$\mathbf{7.73 \cdot 10^{-3}}$	$2.78 \cdot 10^{-1}$
C11orf73	$\mathbf{8.11 \cdot 10^{-3}}$	$2.87 \cdot 10^{-1}$
BLOC1S5.TXNDC5	$\mathbf{8.12 \cdot 10^{-3}}$	$2.87 \cdot 10^{-1}$
NCAPG	$\mathbf{8.36 \cdot 10^{-3}}$	$2.91 \cdot 10^{-1}$
GIGYF1	$\mathbf{8.4 \cdot 10^{-3}}$	$2.91 \cdot 10^{-1}$
PKP2	$\mathbf{8.66 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
NUP133	$\mathbf{9.04 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$

SRP72	$\mathbf{9.05 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
RAP1A	$\mathbf{9.33 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
DACT1	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
IL13RA1	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
MYADM	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
PKN1	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
RP11.216L13.19	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
UFSP2	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
ZWINT	$\mathbf{9.35 \cdot 10^{-3}}$	$2.94 \cdot 10^{-1}$
FMR1	$\mathbf{9.88 \cdot 10^{-3}}$	$3.08 \cdot 10^{-1}$
HPRT1	$\mathbf{1.01 \cdot 10^{-2}}$	$3.12 \cdot 10^{-1}$
CSK	$\mathbf{1.03 \cdot 10^{-2}}$	$3.15 \cdot 10^{-1}$
CPHL1P	$\mathbf{1.05 \cdot 10^{-2}}$	$3.2 \cdot 10^{-1}$
CENPI	$\mathbf{1.06 \cdot 10^{-2}}$	$3.2 \cdot 10^{-1}$
ATG3	$\mathbf{1.07 \cdot 10^{-2}}$	$3.2 \cdot 10^{-1}$
RGS11	$\mathbf{1.08 \cdot 10^{-2}}$	$3.22 \cdot 10^{-1}$
EGLN3	$\mathbf{1.12 \cdot 10^{-2}}$	$3.29 \cdot 10^{-1}$
OSGEPL1	$\mathbf{1.15 \cdot 10^{-2}}$	$3.36 \cdot 10^{-1}$
TRAF5	$\mathbf{1.21 \cdot 10^{-2}}$	$3.52 \cdot 10^{-1}$
SON	$\mathbf{1.22 \cdot 10^{-2}}$	$3.52 \cdot 10^{-1}$
XYLB	$\mathbf{1.23 \cdot 10^{-2}}$	$3.52 \cdot 10^{-1}$
TRPM8	$\mathbf{1.24 \cdot 10^{-2}}$	$3.53 \cdot 10^{-1}$
SYPL1	$\mathbf{1.27 \cdot 10^{-2}}$	$3.57 \cdot 10^{-1}$
JADE3	$\mathbf{1.28 \cdot 10^{-2}}$	$3.58 \cdot 10^{-1}$
HPS3	$\mathbf{1.34 \cdot 10^{-2}}$	$3.73 \cdot 10^{-1}$
TRAC	$\mathbf{1.42 \cdot 10^{-2}}$	$3.92 \cdot 10^{-1}$
ITPR3	$\mathbf{1.5 \cdot 10^{-2}}$	$4.09 \cdot 10^{-1}$
SPAG1	$\mathbf{1.52 \cdot 10^{-2}}$	$4.14 \cdot 10^{-1}$
SLC25A30	$\mathbf{1.58 \cdot 10^{-2}}$	$4.26 \cdot 10^{-1}$
UBAP2	$\mathbf{1.7 \cdot 10^{-2}}$	$4.55 \cdot 10^{-1}$
NOX4	$\mathbf{1.71 \cdot 10^{-2}}$	$4.56 \cdot 10^{-1}$
CSDE1	$\mathbf{1.9 \cdot 10^{-2}}$	$5.01 \cdot 10^{-1}$
PDHA1	$\mathbf{1.99 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
QPCT	$\mathbf{1.99 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
TGFB3	$\mathbf{2.01 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
DLGAP5	$\mathbf{2.04 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
RFX6	$\mathbf{2.05 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
METTL17	$\mathbf{2.06 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
RIOK2	$\mathbf{2.06 \cdot 10^{-2}}$	$5.17 \cdot 10^{-1}$
ONECUT2	$\mathbf{2.21 \cdot 10^{-2}}$	$5.47 \cdot 10^{-1}$
VWDE	$\mathbf{2.21 \cdot 10^{-2}}$	$5.47 \cdot 10^{-1}$
MTUS2	$\mathbf{2.22 \cdot 10^{-2}}$	$5.47 \cdot 10^{-1}$
HDAC9	$\mathbf{2.23 \cdot 10^{-2}}$	$5.47 \cdot 10^{-1}$
HTT	$\mathbf{2.25 \cdot 10^{-2}}$	$5.48 \cdot 10^{-1}$

RAD18	**$2.27 \cdot 10^{-2}$**	$5.49 \cdot 10^{-1}$
BIRC5	**$2.3 \cdot 10^{-2}$**	$5.52 \cdot 10^{-1}$
CHORDC1	**$2.31 \cdot 10^{-2}$**	$5.53 \cdot 10^{-1}$
RORC	**$2.43 \cdot 10^{-2}$**	$5.78 \cdot 10^{-1}$
ZBTB16	**$2.48 \cdot 10^{-2}$**	$5.85 \cdot 10^{-1}$
NDUFAF6	**$2.52 \cdot 10^{-2}$**	$5.9 \cdot 10^{-1}$
GREB1	**$2.58 \cdot 10^{-2}$**	$6.01 \cdot 10^{-1}$
TNRC6C	**$2.66 \cdot 10^{-2}$**	$6.16 \cdot 10^{-1}$
CCNA2	**$2.68 \cdot 10^{-2}$**	$6.17 \cdot 10^{-1}$
KIF18A	**$2.77 \cdot 10^{-2}$**	$6.33 \cdot 10^{-1}$
F5	**$2.9 \cdot 10^{-2}$**	$6.59 \cdot 10^{-1}$
PRR16	**$2.92 \cdot 10^{-2}$**	$6.59 \cdot 10^{-1}$
BRCA1	**$3.06 \cdot 10^{-2}$**	$6.84 \cdot 10^{-1}$
NKTR	**$3.06 \cdot 10^{-2}$**	$6.84 \cdot 10^{-1}$
NR5A2	**$3.09 \cdot 10^{-2}$**	$6.85 \cdot 10^{-1}$
ATXN3	**$3.13 \cdot 10^{-2}$**	$6.88 \cdot 10^{-1}$
JAG1	**$3.13 \cdot 10^{-2}$**	$6.88 \cdot 10^{-1}$
EBNA1BP2	**$3.19 \cdot 10^{-2}$**	$6.96 \cdot 10^{-1}$
GUCY1A3	**$3.23 \cdot 10^{-2}$**	$7.01 \cdot 10^{-1}$
SHCBP1	**$3.33 \cdot 10^{-2}$**	$7.17 \cdot 10^{-1}$
FAM76B	**$3.34 \cdot 10^{-2}$**	$7.17 \cdot 10^{-1}$
YES1	**$3.37 \cdot 10^{-2}$**	$7.18 \cdot 10^{-1}$
ADSL	**$3.45 \cdot 10^{-2}$**	$7.31 \cdot 10^{-1}$
APEH	**$3.47 \cdot 10^{-2}$**	$7.31 \cdot 10^{-1}$
MUC3A	**$3.5 \cdot 10^{-2}$**	$7.34 \cdot 10^{-1}$
CNTN4	**$3.54 \cdot 10^{-2}$**	$7.39 \cdot 10^{-1}$
SMARCA1	**$3.66 \cdot 10^{-2}$**	$7.55 \cdot 10^{-1}$
ZNF76	**$3.66 \cdot 10^{-2}$**	$7.55 \cdot 10^{-1}$
SHISA9	**$3.83 \cdot 10^{-2}$**	$7.83 \cdot 10^{-1}$
DCAF13	**$3.84 \cdot 10^{-2}$**	$7.83 \cdot 10^{-1}$
CENPN	**$3.97 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
CFB	**$4 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
VHL	**$4.03 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
TMC5	**$4.03 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
ZSCAN12	**$4.09 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
CCNE2	**$4.11 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
ELMO1	**$4.11 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
ELMSAN1	**$4.11 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
MKX	**$4.11 \cdot 10^{-2}$**	$8 \cdot 10^{-1}$
DUS4L	**$4.2 \cdot 10^{-2}$**	$8.09 \cdot 10^{-1}$
ZNF566	**$4.2 \cdot 10^{-2}$**	$8.09 \cdot 10^{-1}$
OSGIN2	**$4.24 \cdot 10^{-2}$**	$8.12 \cdot 10^{-1}$
TGDS	**$4.32 \cdot 10^{-2}$**	$8.19 \cdot 10^{-1}$
GPR75.ASB3	**$4.34 \cdot 10^{-2}$**	$8.19 \cdot 10^{-1}$

ZNF229	$\mathbf{4.34 \cdot 10^{-2}}$	$8.19 \cdot 10^{-1}$

Table A.8 Top 200 step-up candidates significantly associated with the time to BCR in the CancerMap dataset, sorted by the log-rank p-value.

Gene	**Log-rank p-val**	**FDR adj. p-val**
PAPPA2	**0**	**0**
STRIP2	**0**	**0**
SLC29A1	$\mathbf{1.44 \cdot 10^{-15}}$	$\mathbf{1.76 \cdot 10^{-12}}$
MGA	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{4.46 \cdot 10^{-12}}$
DAB1	$\mathbf{8.56 \cdot 10^{-13}}$	$\mathbf{6.26 \cdot 10^{-10}}$
ARHGAP19.SLIT1	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{4.64 \cdot 10^{-8}}$
DIAPH3	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{4.64 \cdot 10^{-8}}$
LINC00535	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{4.64 \cdot 10^{-8}}$
USP6NL	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{4.64 \cdot 10^{-8}}$
DGKH	$\mathbf{1.64 \cdot 10^{-7}}$	$\mathbf{6 \cdot 10^{-5}}$
ASPN	$\mathbf{3.97 \cdot 10^{-6}}$	$\mathbf{1.32 \cdot 10^{-3}}$
SRGAP1	$\mathbf{1.51 \cdot 10^{-5}}$	$\mathbf{4.6 \cdot 10^{-3}}$
KIAA1467	$\mathbf{7.31 \cdot 10^{-5}}$	$\mathbf{1.93 \cdot 10^{-2}}$
TTN.AS1	$\mathbf{7.4 \cdot 10^{-5}}$	$\mathbf{1.93 \cdot 10^{-2}}$
LPL	$\mathbf{1.71 \cdot 10^{-4}}$	$\mathbf{3.91 \cdot 10^{-2}}$
PDLIM5	$\mathbf{1.78 \cdot 10^{-4}}$	$\mathbf{3.91 \cdot 10^{-2}}$
LOXL2	$\mathbf{1.82 \cdot 10^{-4}}$	$\mathbf{3.91 \cdot 10^{-2}}$
NRP2	$\mathbf{2.59 \cdot 10^{-4}}$	$5.27 \cdot 10^{-2}$
SEPT8	$\mathbf{3.92 \cdot 10^{-4}}$	$7.47 \cdot 10^{-2}$
LEPR	$\mathbf{4.09 \cdot 10^{-4}}$	$7.47 \cdot 10^{-2}$
PTER	$\mathbf{4.92 \cdot 10^{-4}}$	$7.82 \cdot 10^{-2}$
ST18	$\mathbf{4.92 \cdot 10^{-4}}$	$7.82 \cdot 10^{-2}$
ZNF274	$\mathbf{4.92 \cdot 10^{-4}}$	$7.82 \cdot 10^{-2}$
CR2	$\mathbf{5.96 \cdot 10^{-4}}$	$8.59 \cdot 10^{-2}$
VLDLR	$\mathbf{6.02 \cdot 10^{-4}}$	$8.59 \cdot 10^{-2}$
PARPBP	$\mathbf{6.11 \cdot 10^{-4}}$	$8.59 \cdot 10^{-2}$
PSMA1	$\mathbf{7.05 \cdot 10^{-4}}$	$9.55 \cdot 10^{-2}$
ATF3	$\mathbf{8.33 \cdot 10^{-4}}$	$1.09 \cdot 10^{-1}$
SOX5	$\mathbf{1.08 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
ENTPD3	$\mathbf{1.12 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
INHBA	$\mathbf{1.12 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
MYO10	$\mathbf{1.12 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
P4HTM	$\mathbf{1.12 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
ZFAT	$\mathbf{1.12 \cdot 10^{-3}}$	$1.21 \cdot 10^{-1}$
PLCB4	$\mathbf{1.25 \cdot 10^{-3}}$	$1.29 \cdot 10^{-1}$
THBS4	$\mathbf{1.27 \cdot 10^{-3}}$	$1.29 \cdot 10^{-1}$
NEU3	$\mathbf{1.42 \cdot 10^{-3}}$	$1.4 \cdot 10^{-1}$

ARHGEF3	**$1.76 \cdot 10^{-3}$**	$1.68 \cdot 10^{-1}$
NAALADL2	**$1.8 \cdot 10^{-3}$**	$1.68 \cdot 10^{-1}$
SLC22A15	**$1.88 \cdot 10^{-3}$**	$1.72 \cdot 10^{-1}$
C5orf30	**$2.22 \cdot 10^{-3}$**	$1.96 \cdot 10^{-1}$
RFX2	**$2.25 \cdot 10^{-3}$**	$1.96 \cdot 10^{-1}$
ONECUT2	**$2.82 \cdot 10^{-3}$**	$2.4 \cdot 10^{-1}$
ACACA	**$3.51 \cdot 10^{-3}$**	$2.92 \cdot 10^{-1}$
C12orf29	**$3.85 \cdot 10^{-3}$**	$2.93 \cdot 10^{-1}$
EIF3B	**$3.85 \cdot 10^{-3}$**	$2.93 \cdot 10^{-1}$
LSM14B	**$3.85 \cdot 10^{-3}$**	$2.93 \cdot 10^{-1}$
TRMT10A	**$3.85 \cdot 10^{-3}$**	$2.93 \cdot 10^{-1}$
HMGCLL1	**$3.99 \cdot 10^{-3}$**	$2.98 \cdot 10^{-1}$
CENPP	**$4.16 \cdot 10^{-3}$**	$3.04 \cdot 10^{-1}$
IDH3A	**$4.5 \cdot 10^{-3}$**	$3.21 \cdot 10^{-1}$
FAP	**$4.56 \cdot 10^{-3}$**	$3.21 \cdot 10^{-1}$
TFAM	**$4.91 \cdot 10^{-3}$**	$3.33 \cdot 10^{-1}$
C3orf14	**$4.92 \cdot 10^{-3}$**	$3.33 \cdot 10^{-1}$
CAPN5	**$5.26 \cdot 10^{-3}$**	$3.5 \cdot 10^{-1}$
GALNTL6	**$5.73 \cdot 10^{-3}$**	$3.67 \cdot 10^{-1}$
NXPE1	**$5.73 \cdot 10^{-3}$**	$3.67 \cdot 10^{-1}$
PCDHGA1	**$6.2 \cdot 10^{-3}$**	$3.91 \cdot 10^{-1}$
PGC	**$6.42 \cdot 10^{-3}$**	$3.98 \cdot 10^{-1}$
CDH10	**$6.83 \cdot 10^{-3}$**	$4.1 \cdot 10^{-1}$
IMMP1L	**$6.87 \cdot 10^{-3}$**	$4.1 \cdot 10^{-1}$
DNAH8	**$7.01 \cdot 10^{-3}$**	$4.1 \cdot 10^{-1}$
PTPRZ1	**$7.06 \cdot 10^{-3}$**	$4.1 \cdot 10^{-1}$
ALDH9A1	**$7.32 \cdot 10^{-3}$**	$4.18 \cdot 10^{-1}$
MAP3K5	**$7.58 \cdot 10^{-3}$**	$4.27 \cdot 10^{-1}$
ABCC11	**$7.76 \cdot 10^{-3}$**	$4.3 \cdot 10^{-1}$
RERG	**$9.88 \cdot 10^{-3}$**	$5.39 \cdot 10^{-1}$
GDF15	**$1.1 \cdot 10^{-2}$**	$5.77 \cdot 10^{-1}$
PLEKHA1	**$1.11 \cdot 10^{-2}$**	$5.77 \cdot 10^{-1}$
PNN	**$1.11 \cdot 10^{-2}$**	$5.77 \cdot 10^{-1}$
PDE9A	**$1.12 \cdot 10^{-2}$**	$5.77 \cdot 10^{-1}$
COL21A1	**$1.19 \cdot 10^{-2}$**	$6.05 \cdot 10^{-1}$
BLOC1S5.TXNDC5	**$1.22 \cdot 10^{-2}$**	$6.11 \cdot 10^{-1}$
TGDS	**$1.33 \cdot 10^{-2}$**	$6.55 \cdot 10^{-1}$
RFWD2	**$1.34 \cdot 10^{-2}$**	$6.55 \cdot 10^{-1}$
WDR7	**$1.4 \cdot 10^{-2}$**	$6.72 \cdot 10^{-1}$
TOP2A	**$1.52 \cdot 10^{-2}$**	$7.22 \cdot 10^{-1}$
MAN1A2	**$1.62 \cdot 10^{-2}$**	$7.51 \cdot 10^{-1}$
MRAS	**$1.62 \cdot 10^{-2}$**	$7.51 \cdot 10^{-1}$
LIFR	**$1.68 \cdot 10^{-2}$**	$7.7 \cdot 10^{-1}$
BTBD2	**$1.86 \cdot 10^{-2}$**	$7.71 \cdot 10^{-1}$

ELP4	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
FANCB	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
FLT3	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
RASGRP1	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
RP11.26J3.4	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
ZFP1	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
ZHX3	$\mathbf{1.86 \cdot 10^{-2}}$	$7.71 \cdot 10^{-1}$
PTPN21	$\mathbf{1.89 \cdot 10^{-2}}$	$7.77 \cdot 10^{-1}$
CAND1	$\mathbf{2.03 \cdot 10^{-2}}$	$8.19 \cdot 10^{-1}$
CAPN6	$\mathbf{2.04 \cdot 10^{-2}}$	$8.19 \cdot 10^{-1}$
ARHGEF26	$\mathbf{2.08 \cdot 10^{-2}}$	$8.27 \cdot 10^{-1}$
NR4A1	$\mathbf{2.14 \cdot 10^{-2}}$	$8.42 \cdot 10^{-1}$
NLRP9	$\mathbf{2.3 \cdot 10^{-2}}$	$8.95 \cdot 10^{-1}$
AMOTL1	$\mathbf{2.46 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ADAMTS18	$\mathbf{2.48 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
METTL25	$\mathbf{2.55 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
UNC5D	$\mathbf{2.57 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
FAM171B	$\mathbf{2.58 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
C1orf116	$\mathbf{2.63 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
COPB2	$\mathbf{2.74 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
FZD4	$\mathbf{2.88 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
SELP	$\mathbf{3.07 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TM9SF1	$\mathbf{3.14 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
NCOA2	$\mathbf{3.18 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ANKRD10	$\mathbf{3.22 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
EED	$\mathbf{3.22 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
FAT1	$\mathbf{3.22 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
CALD1	$\mathbf{3.3 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
RFC3	$\mathbf{3.35 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
CASP9	$\mathbf{3.35 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
SYT16	$\mathbf{3.38 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
GLYATL1	$\mathbf{3.41 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ASAH1	$\mathbf{3.44 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ELF3	$\mathbf{3.53 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
SLC25A45	$\mathbf{3.55 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
PPIP5K2	$\mathbf{3.72 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
CTSA	$\mathbf{3.76 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TLR3	$\mathbf{3.81 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
PHLPP2	$\mathbf{3.84 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TAF1D	$\mathbf{3.84 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
PSMF1	$\mathbf{3.98 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TMEM245	$\mathbf{4 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
JAG1	$\mathbf{4.09 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
SKA3	$\mathbf{4.09 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$

SETD7	$\mathbf{4.19 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
NSF	$\mathbf{4.23 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ACTR3	$\mathbf{4.27 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TRPM4	$\mathbf{4.3 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
STAM2	$\mathbf{4.56 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ZAK	$\mathbf{4.66 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
BTAF1	$\mathbf{4.7 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
CDH2	$\mathbf{4.72 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
GSTK1	$\mathbf{4.82 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
ARID5B	$\mathbf{4.89 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
RRAGB	$\mathbf{4.89 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
TPH1	$\mathbf{4.91 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$
RUNX1	$\mathbf{5 \cdot 10^{-2}}$	$9.09 \cdot 10^{-1}$

Table A.9 The step-up candidates significantly associated with the time to BCR in the MSKCC dataset, sorted by the log-rank p-value.

Gene	**Log-rank p-val**	**FDR adj. p-val**
AR	$\mathbf{0}$	$\mathbf{0}$
C15orf38.AP3S2	$\mathbf{0}$	$\mathbf{0}$
COPB1	$\mathbf{0}$	$\mathbf{0}$
NFS1	$\mathbf{0}$	$\mathbf{0}$
NPNT	$\mathbf{0}$	$\mathbf{0}$
FUBP3	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{1.98 \cdot 10^{-13}}$
LRIG1	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{1.98 \cdot 10^{-13}}$
THNSL2	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{1.98 \cdot 10^{-13}}$
TMEM220	$\mathbf{5.55 \cdot 10^{-16}}$	$\mathbf{1.98 \cdot 10^{-13}}$
TNFSF12.TNFSF13	$\mathbf{1.02 \cdot 10^{-13}}$	$\mathbf{3.29 \cdot 10^{-11}}$
ARMCX1	$\mathbf{6.43 \cdot 10^{-11}}$	$\mathbf{1.88 \cdot 10^{-8}}$
REPS2	$\mathbf{8.7 \cdot 10^{-11}}$	$\mathbf{2.33 \cdot 10^{-8}}$
ST6GAL1	$\mathbf{1.02 \cdot 10^{-9}}$	$\mathbf{2.51 \cdot 10^{-7}}$
ZSCAN20	$\mathbf{6.33 \cdot 10^{-9}}$	$\mathbf{1.45 \cdot 10^{-6}}$
MTMR12	$\mathbf{1.01 \cdot 10^{-7}}$	$\mathbf{2.17 \cdot 10^{-5}}$
AKAP7	$\mathbf{1.82 \cdot 10^{-7}}$	$\mathbf{3.65 \cdot 10^{-5}}$
SDC2	$\mathbf{4.07 \cdot 10^{-7}}$	$\mathbf{7.68 \cdot 10^{-5}}$
CREB3L2	$\mathbf{4.39 \cdot 10^{-7}}$	$\mathbf{7.83 \cdot 10^{-5}}$
RIN2	$\mathbf{4.63 \cdot 10^{-7}}$	$\mathbf{7.83 \cdot 10^{-5}}$
UST	$\mathbf{1.33 \cdot 10^{-6}}$	$\mathbf{2.14 \cdot 10^{-4}}$
TRIQK	$\mathbf{1.5 \cdot 10^{-6}}$	$\mathbf{2.3 \cdot 10^{-4}}$
PCM1	$\mathbf{3.24 \cdot 10^{-6}}$	$\mathbf{4.72 \cdot 10^{-4}}$
MON1B	$\mathbf{3.38 \cdot 10^{-6}}$	$\mathbf{4.72 \cdot 10^{-4}}$
SATB1	$\mathbf{3.87 \cdot 10^{-6}}$	$\mathbf{5.18 \cdot 10^{-4}}$
PDZRN4	$\mathbf{4.4 \cdot 10^{-6}}$	$\mathbf{5.66 \cdot 10^{-4}}$

GLIPR2	$6.76 \cdot 10^{-6}$	$8.31 \cdot 10^{-4}$
FBXO32	$7.17 \cdot 10^{-6}$	$8.31 \cdot 10^{-4}$
KCND3	$7.25 \cdot 10^{-6}$	$8.31 \cdot 10^{-4}$
AHNAK2	$9.13 \cdot 10^{-6}$	$1.01 \cdot 10^{-3}$
CYFIP1	$1.16 \cdot 10^{-5}$	$1.24 \cdot 10^{-3}$
EDNRA	$1.88 \cdot 10^{-5}$	$1.95 \cdot 10^{-3}$
ST8SIA6	$2.23 \cdot 10^{-5}$	$2.22 \cdot 10^{-3}$
ALDH1A3	$2.28 \cdot 10^{-5}$	$2.22 \cdot 10^{-3}$
ALDH3A2	$2.63 \cdot 10^{-5}$	$2.48 \cdot 10^{-3}$
NAPEPLD	$3.02 \cdot 10^{-5}$	$2.77 \cdot 10^{-3}$
EEF1A1	$4.19 \cdot 10^{-5}$	$3.74 \cdot 10^{-3}$
DIXDC1	$4.68 \cdot 10^{-5}$	$4.07 \cdot 10^{-3}$
NFIA	$5.28 \cdot 10^{-5}$	$4.37 \cdot 10^{-3}$
UBE2K	$5.31 \cdot 10^{-5}$	$4.37 \cdot 10^{-3}$
RPS6KA5	$8.35 \cdot 10^{-5}$	$6.7 \cdot 10^{-3}$
ACAD8	$1.05 \cdot 10^{-4}$	$8.05 \cdot 10^{-3}$
PMEPA1	$1.05 \cdot 10^{-4}$	$8.05 \cdot 10^{-3}$
CNTNAP2	$1.24 \cdot 10^{-4}$	$9.3 \cdot 10^{-3}$
IGFBP3	$1.3 \cdot 10^{-4}$	$9.46 \cdot 10^{-3}$
ERBB2IP	$1.75 \cdot 10^{-4}$	$1.25 \cdot 10^{-2}$
ZNF510	$1.8 \cdot 10^{-4}$	$1.26 \cdot 10^{-2}$
FAM122B	$2.01 \cdot 10^{-4}$	$1.37 \cdot 10^{-2}$
DNAJA2	$2.33 \cdot 10^{-4}$	$1.56 \cdot 10^{-2}$
EPHX2	$2.91 \cdot 10^{-4}$	$1.88 \cdot 10^{-2}$
C9orf91	$2.92 \cdot 10^{-4}$	$1.88 \cdot 10^{-2}$
AMOTL2	$2.98 \cdot 10^{-4}$	$1.88 \cdot 10^{-2}$
DES	$3.09 \cdot 10^{-4}$	$1.91 \cdot 10^{-2}$
FERMT1	$3.59 \cdot 10^{-4}$	$2.18 \cdot 10^{-2}$
LACC1	$3.85 \cdot 10^{-4}$	$2.26 \cdot 10^{-2}$
NRG4	$3.86 \cdot 10^{-4}$	$2.26 \cdot 10^{-2}$
AZGP1	$3.98 \cdot 10^{-4}$	$2.28 \cdot 10^{-2}$
SNAP29	$4.1 \cdot 10^{-4}$	$2.31 \cdot 10^{-2}$
FREM2	$5.08 \cdot 10^{-4}$	$2.82 \cdot 10^{-2}$
TRAPPC13	$5.53 \cdot 10^{-4}$	$3.01 \cdot 10^{-2}$
PI15	$5.78 \cdot 10^{-4}$	$3.06 \cdot 10^{-2}$
BEND4	$5.81 \cdot 10^{-4}$	$3.06 \cdot 10^{-2}$
HSDL2	$5.9 \cdot 10^{-4}$	$3.06 \cdot 10^{-2}$
AFF3	$6.54 \cdot 10^{-4}$	$3.34 \cdot 10^{-2}$
NHS	$6.99 \cdot 10^{-4}$	$3.49 \cdot 10^{-2}$
VAT1	$7.06 \cdot 10^{-4}$	$3.49 \cdot 10^{-2}$
FAM177A1	$7.37 \cdot 10^{-4}$	$3.59 \cdot 10^{-2}$
ASPA	$7.53 \cdot 10^{-4}$	$3.61 \cdot 10^{-2}$
LRCH2	$8.46 \cdot 10^{-4}$	$4 \cdot 10^{-2}$
ME1	$8.68 \cdot 10^{-4}$	$4.04 \cdot 10^{-2}$

PCDHGC5	**$9.55 \cdot 10^{-4}$**	**$4.38 \cdot 10^{-2}$**
KIAA1324L	**$9.87 \cdot 10^{-4}$**	**$4.47 \cdot 10^{-2}$**
APLF	**$1.03 \cdot 10^{-3}$**	**$4.58 \cdot 10^{-2}$**
CDK6	**$1.06 \cdot 10^{-3}$**	**$4.61 \cdot 10^{-2}$**
ATF7IP2	**$1.08 \cdot 10^{-3}$**	**$4.61 \cdot 10^{-2}$**
IGSF1	**$1.08 \cdot 10^{-3}$**	**$4.61 \cdot 10^{-2}$**
ATRN	**$1.09 \cdot 10^{-3}$**	**$4.61 \cdot 10^{-2}$**
SLC9A2	**$1.26 \cdot 10^{-3}$**	$5.25 \cdot 10^{-2}$
UBLCP1	**$1.34 \cdot 10^{-3}$**	$5.53 \cdot 10^{-2}$
CASP7	**$1.45 \cdot 10^{-3}$**	$5.91 \cdot 10^{-2}$
STIM2	**$1.49 \cdot 10^{-3}$**	$5.97 \cdot 10^{-2}$
INPP4B	**$1.52 \cdot 10^{-3}$**	$5.98 \cdot 10^{-2}$
LSS	**$1.53 \cdot 10^{-3}$**	$5.98 \cdot 10^{-2}$
JPX	**$1.56 \cdot 10^{-3}$**	$5.98 \cdot 10^{-2}$
RNF19B	**$1.56 \cdot 10^{-3}$**	$5.98 \cdot 10^{-2}$
CLVS2	**$1.6 \cdot 10^{-3}$**	$6.03 \cdot 10^{-2}$
MR1	**$1.62 \cdot 10^{-3}$**	$6.03 \cdot 10^{-2}$
TAB3	**$1.63 \cdot 10^{-3}$**	$6.03 \cdot 10^{-2}$
SRPX	**$1.74 \cdot 10^{-3}$**	$6.34 \cdot 10^{-2}$
PTPLA	**$1.75 \cdot 10^{-3}$**	$6.34 \cdot 10^{-2}$
SLC1A1	**$1.79 \cdot 10^{-3}$**	$6.41 \cdot 10^{-2}$
NUDCD1	**$1.91 \cdot 10^{-3}$**	$6.72 \cdot 10^{-2}$
ANXA6	**$1.92 \cdot 10^{-3}$**	$6.72 \cdot 10^{-2}$
ACADL	**$2.08 \cdot 10^{-3}$**	$7.16 \cdot 10^{-2}$
TUBA1A	**$2.09 \cdot 10^{-3}$**	$7.16 \cdot 10^{-2}$
RAB27A	**$2.12 \cdot 10^{-3}$**	$7.18 \cdot 10^{-2}$
FAM188A	**$2.24 \cdot 10^{-3}$**	$7.51 \cdot 10^{-2}$
BHMT2	**$2.28 \cdot 10^{-3}$**	$7.55 \cdot 10^{-2}$
CPPED1	**$2.36 \cdot 10^{-3}$**	$7.74 \cdot 10^{-2}$
NAA25	**$2.44 \cdot 10^{-3}$**	$7.93 \cdot 10^{-2}$
ELOVL7	**$2.51 \cdot 10^{-3}$**	$7.99 \cdot 10^{-2}$
SLC18A2	**$2.51 \cdot 10^{-3}$**	$7.99 \cdot 10^{-2}$
IFIT1	**$2.54 \cdot 10^{-3}$**	$7.99 \cdot 10^{-2}$
JUP	**$2.6 \cdot 10^{-3}$**	$8.09 \cdot 10^{-2}$
NLRP2	**$2.62 \cdot 10^{-3}$**	$8.09 \cdot 10^{-2}$
ECHDC1	**$2.69 \cdot 10^{-3}$**	$8.1 \cdot 10^{-2}$
BECN1	**$2.72 \cdot 10^{-3}$**	$8.1 \cdot 10^{-2}$
RXRA	**$2.72 \cdot 10^{-3}$**	$8.1 \cdot 10^{-2}$
SAFB2	**$2.72 \cdot 10^{-3}$**	$8.1 \cdot 10^{-2}$
ARHGEF9	**$2.85 \cdot 10^{-3}$**	$8.22 \cdot 10^{-2}$
CRNKL1	**$2.89 \cdot 10^{-3}$**	$8.22 \cdot 10^{-2}$
DNAJC24	**$2.89 \cdot 10^{-3}$**	$8.22 \cdot 10^{-2}$
METTL9	**$2.89 \cdot 10^{-3}$**	$8.22 \cdot 10^{-2}$
SRRD	**$2.89 \cdot 10^{-3}$**	$8.22 \cdot 10^{-2}$

PDS5B	$\mathbf{2.94 \cdot 10^{-3}}$	$8.3 \cdot 10^{-2}$
CFL2	$\mathbf{3.2 \cdot 10^{-3}}$	$8.93 \cdot 10^{-2}$
KLF5	$\mathbf{3.22 \cdot 10^{-3}}$	$8.93 \cdot 10^{-2}$
CD44	$\mathbf{3.29 \cdot 10^{-3}}$	$8.95 \cdot 10^{-2}$
SH3BGRL	$\mathbf{3.29 \cdot 10^{-3}}$	$8.95 \cdot 10^{-2}$
IDI1	$\mathbf{3.37 \cdot 10^{-3}}$	$9.09 \cdot 10^{-2}$
PARP8	$\mathbf{3.43 \cdot 10^{-3}}$	$9.2 \cdot 10^{-2}$
FUT8	$\mathbf{3.55 \cdot 10^{-3}}$	$9.44 \cdot 10^{-2}$
PDE11A	$\mathbf{3.63 \cdot 10^{-3}}$	$9.56 \cdot 10^{-2}$
KANK2	$\mathbf{3.73 \cdot 10^{-3}}$	$9.73 \cdot 10^{-2}$
ANO5	$\mathbf{3.76 \cdot 10^{-3}}$	$9.73 \cdot 10^{-2}$
NRK	$\mathbf{3.85 \cdot 10^{-3}}$	$9.88 \cdot 10^{-2}$
MLLT4	$\mathbf{3.87 \cdot 10^{-3}}$	$9.88 \cdot 10^{-2}$
CHRDL1	$\mathbf{3.99 \cdot 10^{-3}}$	$1.01 \cdot 10^{-1}$
FBLN1	$\mathbf{4.01 \cdot 10^{-3}}$	$1.01 \cdot 10^{-1}$
CYP4X1	$\mathbf{4.1 \cdot 10^{-3}}$	$1.02 \cdot 10^{-1}$
EIF4E3	$\mathbf{4.21 \cdot 10^{-3}}$	$1.04 \cdot 10^{-1}$
KLK4	$\mathbf{4.28 \cdot 10^{-3}}$	$1.04 \cdot 10^{-1}$
PCDHA3	$\mathbf{4.28 \cdot 10^{-3}}$	$1.04 \cdot 10^{-1}$
LY75.CD302	$\mathbf{4.43 \cdot 10^{-3}}$	$1.07 \cdot 10^{-1}$
METTL7A	$\mathbf{4.51 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}$
ZNF83	$\mathbf{4.65 \cdot 10^{-3}}$	$1.11 \cdot 10^{-1}$
PIFO	$\mathbf{4.8 \cdot 10^{-3}}$	$1.13 \cdot 10^{-1}$
RPL31	$\mathbf{4.92 \cdot 10^{-3}}$	$1.15 \cdot 10^{-1}$
TNC	$\mathbf{5.08 \cdot 10^{-3}}$	$1.18 \cdot 10^{-1}$
RAB4A	$\mathbf{5.22 \cdot 10^{-3}}$	$1.19 \cdot 10^{-1}$
LRRC28	$\mathbf{5.25 \cdot 10^{-3}}$	$1.19 \cdot 10^{-1}$
QSER1	$\mathbf{5.25 \cdot 10^{-3}}$	$1.19 \cdot 10^{-1}$
AUH	$\mathbf{5.27 \cdot 10^{-3}}$	$1.19 \cdot 10^{-1}$
BMP4	$\mathbf{5.47 \cdot 10^{-3}}$	$1.23 \cdot 10^{-1}$
GRAMD1C	$\mathbf{5.81 \cdot 10^{-3}}$	$1.3 \cdot 10^{-1}$
ABCC1	$\mathbf{6.12 \cdot 10^{-3}}$	$1.36 \cdot 10^{-1}$
SCUBE2	$\mathbf{6.19 \cdot 10^{-3}}$	$1.36 \cdot 10^{-1}$
EXOSC7	$\mathbf{6.5 \cdot 10^{-3}}$	$1.42 \cdot 10^{-1}$
CHN1	$\mathbf{6.64 \cdot 10^{-3}}$	$1.44 \cdot 10^{-1}$
IDH3A	$\mathbf{6.68 \cdot 10^{-3}}$	$1.44 \cdot 10^{-1}$
TGFBR1	$\mathbf{6.84 \cdot 10^{-3}}$	$1.46 \cdot 10^{-1}$
MYOCD	$\mathbf{6.99 \cdot 10^{-3}}$	$1.49 \cdot 10^{-1}$
STARD4	$\mathbf{7.34 \cdot 10^{-3}}$	$1.55 \cdot 10^{-1}$
C3AR1	$\mathbf{7.54 \cdot 10^{-3}}$	$1.58 \cdot 10^{-1}$
TUSC3	$\mathbf{7.69 \cdot 10^{-3}}$	$1.6 \cdot 10^{-1}$
FZD6	$\mathbf{7.84 \cdot 10^{-3}}$	$1.63 \cdot 10^{-1}$
CEACAM1	$\mathbf{8.38 \cdot 10^{-3}}$	$1.72 \cdot 10^{-1}$
CGNL1	$\mathbf{8.4 \cdot 10^{-3}}$	$1.72 \cdot 10^{-1}$

GPM6A	$\mathbf{8.44 \cdot 10^{-3}}$	$1.72 \cdot 10^{-1}$
VAPB	$\mathbf{8.52 \cdot 10^{-3}}$	$1.72 \cdot 10^{-1}$
KLHL2	$\mathbf{8.93 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}$
ARHGAP28	$\mathbf{8.95 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}$
TTC8	$\mathbf{8.95 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}$
SRD5A2	$\mathbf{8.99 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}$
PLAGL1	$\mathbf{9.05 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}$
PCDHA10	$\mathbf{9.35 \cdot 10^{-3}}$	$1.82 \cdot 10^{-1}$
TRPC1	$\mathbf{9.5 \cdot 10^{-3}}$	$1.84 \cdot 10^{-1}$
RBM4B	$\mathbf{9.69 \cdot 10^{-3}}$	$1.86 \cdot 10^{-1}$
ANXA1	$\mathbf{9.73 \cdot 10^{-3}}$	$1.86 \cdot 10^{-1}$
HLA.E	$\mathbf{0.999 \cdot 10^{-2}}$	$1.9 \cdot 10^{-1}$
CUL4B	$\mathbf{1.01 \cdot 10^{-2}}$	$1.9 \cdot 10^{-1}$
CRISPLD2	$\mathbf{1.01 \cdot 10^{-2}}$	$1.9 \cdot 10^{-1}$
SPATA6	$\mathbf{1.03 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
MAGED2	$\mathbf{1.03 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
RAPH1	$\mathbf{1.04 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
RNF111	$\mathbf{1.05 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
PPP3CB	$\mathbf{1.07 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
SNTB2	$\mathbf{1.07 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
STXBP6	$\mathbf{1.07 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
TMEM68	$\mathbf{1.07 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
TYMS	$\mathbf{1.07 \cdot 10^{-2}}$	$1.91 \cdot 10^{-1}$
MATN2	$\mathbf{1.08 \cdot 10^{-2}}$	$1.92 \cdot 10^{-1}$
ATXN10	$\mathbf{1.09 \cdot 10^{-2}}$	$1.92 \cdot 10^{-1}$
TNFRSF19	$\mathbf{1.09 \cdot 10^{-2}}$	$1.92 \cdot 10^{-1}$
TMEM41B	$\mathbf{1.11 \cdot 10^{-2}}$	$1.94 \cdot 10^{-1}$
RNF145	$\mathbf{1.12 \cdot 10^{-2}}$	$1.95 \cdot 10^{-1}$
FHOD3	$\mathbf{1.14 \cdot 10^{-2}}$	$1.96 \cdot 10^{-1}$
LRRC9	$\mathbf{1.14 \cdot 10^{-2}}$	$1.96 \cdot 10^{-1}$
TRAC	$\mathbf{1.16 \cdot 10^{-2}}$	$1.97 \cdot 10^{-1}$
DDX42	$\mathbf{1.18 \cdot 10^{-2}}$	$2 \cdot 10^{-1}$
NQO1	$\mathbf{1.19 \cdot 10^{-2}}$	$2.01 \cdot 10^{-1}$
SPG20	$\mathbf{1.28 \cdot 10^{-2}}$	$2.14 \cdot 10^{-1}$
KB.1507C5.2	$\mathbf{1.29 \cdot 10^{-2}}$	$2.14 \cdot 10^{-1}$
TNFAIP2	$\mathbf{1.29 \cdot 10^{-2}}$	$2.14 \cdot 10^{-1}$
WWTR1	$\mathbf{1.3 \cdot 10^{-2}}$	$2.14 \cdot 10^{-1}$
ACAT2	$\mathbf{1.3 \cdot 10^{-2}}$	$2.14 \cdot 10^{-1}$
C14orf37	$\mathbf{1.38 \cdot 10^{-2}}$	$2.26 \cdot 10^{-1}$
CAP2	$\mathbf{1.39 \cdot 10^{-2}}$	$2.26 \cdot 10^{-1}$
IRS2	$\mathbf{1.4 \cdot 10^{-2}}$	$2.27 \cdot 10^{-1}$
NME7	$\mathbf{1.4 \cdot 10^{-2}}$	$2.27 \cdot 10^{-1}$
POT1	$\mathbf{1.43 \cdot 10^{-2}}$	$2.3 \cdot 10^{-1}$

Table A.10 Top 200 step-down candidates significantly associated with the time to BCR in the CancerMap dataset, sorted by the log-rank p-value.

Gene	**Log-rank p-val**	**FDR adj. p-val**
AASS	**0**	**0**
BMP4	**0**	**0**
CD38	**0**	**0**
DACT1	**0**	**0**
DET1	**0**	**0**
IDO1	**0**	**0**
KCNAB1	**0**	**0**
KIAA1377	**0**	**0**
LRMP	**0**	**0**
PRELP	**0**	**0**
RIN2	**0**	**0**
RP11.296A16.1	**0**	**0**
SLC25A13	**0**	**0**
STK33	**0**	**0**
TPH1	**0**	**0**
DDX46	$\mathbf{2.22 \cdot 10^{-16}}$	$\mathbf{4.03 \cdot 10^{-14}}$
RAB30	$\mathbf{2.22 \cdot 10^{-16}}$	$\mathbf{4.03 \cdot 10^{-14}}$
AKAP7	$\mathbf{1.44 \cdot 10^{-15}}$	$\mathbf{2.48 \cdot 10^{-13}}$
ANAPC4	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
EPS15	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
FAM179B	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
HKR1	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
MTOR	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
SYDE2	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
TIMMDC1	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
TPP2	$\mathbf{4.88 \cdot 10^{-15}}$	$\mathbf{5.8 \cdot 10^{-13}}$
CLK1	$\mathbf{1.97 \cdot 10^{-13}}$	$\mathbf{2.17 \cdot 10^{-11}}$
PHF14	$\mathbf{1.97 \cdot 10^{-13}}$	$\mathbf{2.17 \cdot 10^{-11}}$
BBIP1	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
C6ORF174	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
CDCA7L	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
ETV1	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
FMR1	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
IL13RA1	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
KLHL2	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
MLLT4	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
MS4A7	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
MSH5	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$
PDPR	$\mathbf{1.14 \cdot 10^{-10}}$	$\mathbf{8.01 \cdot 10^{-9}}$

PER3	**$1.14 \cdot 10^{-10}$**	**$8.01 \cdot 10^{-9}$**
SLC2A14	**$1.14 \cdot 10^{-10}$**	**$8.01 \cdot 10^{-9}$**
SLC7A11	**$1.14 \cdot 10^{-10}$**	**$8.01 \cdot 10^{-9}$**
SNRNP48	**$1.14 \cdot 10^{-10}$**	**$8.01 \cdot 10^{-9}$**
SRGN	**$1.14 \cdot 10^{-10}$**	**$8.01 \cdot 10^{-9}$**
TSPAN7	**$4.71 \cdot 10^{-10}$**	**$3.23 \cdot 10^{-8}$**
TRIM21	**$2.21 \cdot 10^{-9}$**	**$1.49 \cdot 10^{-7}$**
DEPTOR	**$3.13 \cdot 10^{-9}$**	**$2.06 \cdot 10^{-7}$**
CNTRL	**$3.51 \cdot 10^{-9}$**	**$2.26 \cdot 10^{-7}$**
NEDD9	**$4.14 \cdot 10^{-9}$**	**$2.61 \cdot 10^{-7}$**
EFCAB4B	**$1.16 \cdot 10^{-8}$**	**$7.14 \cdot 10^{-7}$**
RBM47	**$1.37 \cdot 10^{-8}$**	**$8.28 \cdot 10^{-7}$**
CD44	**$1.83 \cdot 10^{-8}$**	**$1.09 \cdot 10^{-6}$**
MTERFD2	**$3.32 \cdot 10^{-8}$**	**$1.9 \cdot 10^{-6}$**
RSRP1	**$3.32 \cdot 10^{-8}$**	**$1.9 \cdot 10^{-6}$**
LAMP3	**$4.01 \cdot 10^{-8}$**	**$2.25 \cdot 10^{-6}$**
LNX1	**$7.52 \cdot 10^{-8}$**	**$4.15 \cdot 10^{-6}$**
ELP4	**$1.39 \cdot 10^{-7}$**	**$7.53 \cdot 10^{-6}$**
TUBE1	**$1.79 \cdot 10^{-7}$**	**$9.54 \cdot 10^{-6}$**
TRMT1L	**$2.3 \cdot 10^{-7}$**	**$1.2 \cdot 10^{-5}$**
STK19	**$3.44 \cdot 10^{-7}$**	**$1.77 \cdot 10^{-5}$**
BAZ1A	**$5.54 \cdot 10^{-7}$**	**$2.8 \cdot 10^{-5}$**
PRPF39	**$6.99 \cdot 10^{-7}$**	**$3.48 \cdot 10^{-5}$**
ELL2	**$9.6 \cdot 10^{-7}$**	**$4.02 \cdot 10^{-5}$**
COMMD3	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
GABRE	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
GPAM	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
KIAA0020	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
M6PR	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
MS4A6A	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
NPHP3	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
PCDHGA2	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
RPS13	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
SKIV2L2	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
SLAIN1	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
SNRPA1	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
THOC1	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
ZNF10	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
ZNF112	**$1.01 \cdot 10^{-6}$**	**$4.02 \cdot 10^{-5}$**
ARMCX1	**$1.22 \cdot 10^{-6}$**	**$4.77 \cdot 10^{-5}$**
EFEMP1	**$1.33 \cdot 10^{-6}$**	**$5.07 \cdot 10^{-5}$**
LPAR1	**$1.33 \cdot 10^{-6}$**	**$5.07 \cdot 10^{-5}$**
DR1	**$1.73 \cdot 10^{-6}$**	**$6.53 \cdot 10^{-5}$**
SCN7A	**$2.43 \cdot 10^{-6}$**	**$9.05 \cdot 10^{-5}$**

FGFR1OP2	$\mathbf{2.5\cdot 10^{-6}}$	$\mathbf{9.19\cdot 10^{-5}}$
LDHB	$\mathbf{2.85\cdot 10^{-6}}$	$\mathbf{1.04\cdot 10^{-4}}$
STYX	$\mathbf{3.1\cdot 10^{-6}}$	$\mathbf{1.11\cdot 10^{-4}}$
DHX36	$\mathbf{3.27\cdot 10^{-6}}$	$\mathbf{1.16\cdot 10^{-4}}$
CWC27	$\mathbf{4.54\cdot 10^{-6}}$	$\mathbf{1.59\cdot 10^{-4}}$
DMXL1	$\mathbf{4.61\cdot 10^{-6}}$	$\mathbf{1.6\cdot 10^{-4}}$
STARD4	$\mathbf{4.75\cdot 10^{-6}}$	$\mathbf{1.63\cdot 10^{-4}}$
RASA1	$\mathbf{5.04\cdot 10^{-6}}$	$\mathbf{1.71\cdot 10^{-4}}$
SREK1IP1	$\mathbf{5.25\cdot 10^{-6}}$	$\mathbf{1.76\cdot 10^{-4}}$
EIF4A2	$\mathbf{6.53\cdot 10^{-6}}$	$\mathbf{2.17\cdot 10^{-4}}$
RBM4B	$\mathbf{7.38\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
ESCO1	$\mathbf{7.45\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
CYP3A5	$\mathbf{7.66\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
AHSA1	$\mathbf{7.67\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
BBX	$\mathbf{7.67\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
VPS8	$\mathbf{7.67\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
YIPF1	$\mathbf{7.67\cdot 10^{-6}}$	$\mathbf{2.37\cdot 10^{-4}}$
ZBTB11	$\mathbf{9.15\cdot 10^{-6}}$	$\mathbf{2.8\cdot 10^{-4}}$
EFTUD2	$\mathbf{9.68\cdot 10^{-6}}$	$\mathbf{2.9\cdot 10^{-4}}$
DNAH7	$\mathbf{1.01\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
GRIA3	$\mathbf{1.01\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
NQO1	$\mathbf{1.01\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
RBPMS	$\mathbf{1.01\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
RGS22	$\mathbf{1.01\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
IDH1	$\mathbf{1.02\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
SOS1	$\mathbf{1.02\cdot 10^{-5}}$	$\mathbf{2.9\cdot 10^{-4}}$
SLC25A24	$\mathbf{1.3\cdot 10^{-5}}$	$\mathbf{3.64\cdot 10^{-4}}$
DAAM1	$\mathbf{1.32\cdot 10^{-5}}$	$\mathbf{3.68\cdot 10^{-4}}$
SETD4	$\mathbf{1.51\cdot 10^{-5}}$	$\mathbf{4.16\cdot 10^{-4}}$
RAB14	$\mathbf{1.77\cdot 10^{-5}}$	$\mathbf{4.8\cdot 10^{-4}}$
ZNF880	$\mathbf{1.77\cdot 10^{-5}}$	$\mathbf{4.8\cdot 10^{-4}}$
LRCH2	$\mathbf{1.89\cdot 10^{-5}}$	$\mathbf{5.08\cdot 10^{-4}}$
C11orf54	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
CCDC14	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
IMPA1	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
MOCS2	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
MUT	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
NPHP3.ACAD11	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
PSMD5.AS1	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
SMC5	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
ZNF507	$\mathbf{2.19\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
ZNF655	$\mathbf{2.21\cdot 10^{-5}}$	$\mathbf{5.45\cdot 10^{-4}}$
AK9	$\mathbf{2.37\cdot 10^{-5}}$	$\mathbf{5.8\cdot 10^{-4}}$
ERC1	$\mathbf{2.51\cdot 10^{-5}}$	$\mathbf{6.11\cdot 10^{-4}}$

CTD.2116N17.1	$2.67 \cdot 10^{-5}$	$6.44 \cdot 10^{-4}$
BBS2	$2.87 \cdot 10^{-5}$	$6.88 \cdot 10^{-4}$
LIN7C	$3.53 \cdot 10^{-5}$	$8.39 \cdot 10^{-4}$
ASPA	$3.69 \cdot 10^{-5}$	$8.69 \cdot 10^{-4}$
REV3L	$3.85 \cdot 10^{-5}$	$9 \cdot 10^{-4}$
ALDH3A2	$3.99 \cdot 10^{-5}$	$9.26 \cdot 10^{-4}$
UBE4A	$4.17 \cdot 10^{-5}$	$9.61 \cdot 10^{-4}$
C11orf73	$4.42 \cdot 10^{-5}$	$0.999 \cdot 10^{-3}$
KIAA1731	$4.42 \cdot 10^{-5}$	$0.999 \cdot 10^{-3}$
PCMTD1	$4.47 \cdot 10^{-5}$	$0.999 \cdot 10^{-3}$
SC5D	$4.47 \cdot 10^{-5}$	$0.999 \cdot 10^{-3}$
RSU1	$5.24 \cdot 10^{-5}$	$1.15 \cdot 10^{-3}$
TIAM1	$5.24 \cdot 10^{-5}$	$1.15 \cdot 10^{-3}$
VCAN	$5.24 \cdot 10^{-5}$	$1.15 \cdot 10^{-3}$
SLC7A6	$5.41 \cdot 10^{-5}$	$1.18 \cdot 10^{-3}$
HERC4	$5.56 \cdot 10^{-5}$	$1.2 \cdot 10^{-3}$
PCF11	$7.06 \cdot 10^{-5}$	$1.51 \cdot 10^{-3}$
TNFRSF19	$7.08 \cdot 10^{-5}$	$1.51 \cdot 10^{-3}$
AHR	$7.73 \cdot 10^{-5}$	$1.63 \cdot 10^{-3}$
ACSS3	$7.85 \cdot 10^{-5}$	$1.65 \cdot 10^{-3}$
ZNF614	$8.11 \cdot 10^{-5}$	$1.69 \cdot 10^{-3}$
TUBA1A	$8.88 \cdot 10^{-5}$	$1.84 \cdot 10^{-3}$
CLASP1	$9.7 \cdot 10^{-5}$	$2 \cdot 10^{-3}$
TBCCD1	$9.96 \cdot 10^{-5}$	$2.04 \cdot 10^{-3}$
ABLIM1	$1.15 \cdot 10^{-4}$	$2.34 \cdot 10^{-3}$
HPGD	$1.18 \cdot 10^{-4}$	$2.37 \cdot 10^{-3}$
KIAA1586	$1.24 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$
MYO5C	$1.24 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$
NAA16	$1.24 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$
DDX52	$1.29 \cdot 10^{-4}$	$2.53 \cdot 10^{-3}$
DMD	$1.3 \cdot 10^{-4}$	$2.53 \cdot 10^{-3}$
CSRNP1	$1.33 \cdot 10^{-4}$	$2.58 \cdot 10^{-3}$
DIXDC1	$1.45 \cdot 10^{-4}$	$2.81 \cdot 10^{-3}$
C12orf29	$1.6 \cdot 10^{-4}$	$3.07 \cdot 10^{-3}$
ARSD	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
CCAR2	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
CENPC	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
FGD6	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
IL10RB	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
ITGA2	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
PHACTR2	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
RIPK1	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
RNF180	$1.82 \cdot 10^{-4}$	$3.3 \cdot 10^{-3}$
RARS	$2.19 \cdot 10^{-4}$	$3.96 \cdot 10^{-3}$

ZDHHC13	$\mathbf{2.46 \cdot 10^{-4}}$	$\mathbf{4.41 \cdot 10^{-3}}$
TBCK	$\mathbf{2.51 \cdot 10^{-4}}$	$\mathbf{4.46 \cdot 10^{-3}}$
UBE3A	$\mathbf{2.51 \cdot 10^{-4}}$	$\mathbf{4.46 \cdot 10^{-3}}$
ESF1	$\mathbf{2.6 \cdot 10^{-4}}$	$\mathbf{4.58 \cdot 10^{-3}}$
TAF2	$\mathbf{2.68 \cdot 10^{-4}}$	$\mathbf{4.71 \cdot 10^{-3}}$
C12orf4	$\mathbf{2.73 \cdot 10^{-4}}$	$\mathbf{4.76 \cdot 10^{-3}}$
ALG6	$\mathbf{3.12 \cdot 10^{-4}}$	$\mathbf{5.42 \cdot 10^{-3}}$
ZNF271	$\mathbf{3.19 \cdot 10^{-4}}$	$\mathbf{5.5 \cdot 10^{-3}}$
CHRNA5	$\mathbf{3.26 \cdot 10^{-4}}$	$\mathbf{5.58 \cdot 10^{-3}}$
ADAM28	$\mathbf{3.31 \cdot 10^{-4}}$	$\mathbf{5.64 \cdot 10^{-3}}$
AKAP2	$\mathbf{3.39 \cdot 10^{-4}}$	$\mathbf{5.75 \cdot 10^{-3}}$
EDRF1	$\mathbf{3.7 \cdot 10^{-4}}$	$\mathbf{6.24 \cdot 10^{-3}}$
ZNF226	$\mathbf{3.72 \cdot 10^{-4}}$	$\mathbf{6.24 \cdot 10^{-3}}$
IFI16	$\mathbf{3.82 \cdot 10^{-4}}$	$\mathbf{6.37 \cdot 10^{-3}}$
TGM2	$\mathbf{4.46 \cdot 10^{-4}}$	$\mathbf{7.4 \cdot 10^{-3}}$
TCF7L1	$\mathbf{4.49 \cdot 10^{-4}}$	$\mathbf{7.41 \cdot 10^{-3}}$
CREBZF	$\mathbf{4.54 \cdot 10^{-4}}$	$\mathbf{7.46 \cdot 10^{-3}}$
ZFC3H1	$\mathbf{4.75 \cdot 10^{-4}}$	$\mathbf{7.75 \cdot 10^{-3}}$
KIAA0907	$\mathbf{4.9 \cdot 10^{-4}}$	$\mathbf{7.95 \cdot 10^{-3}}$
AMICA1	$\mathbf{4.92 \cdot 10^{-4}}$	$\mathbf{7.95 \cdot 10^{-3}}$
LPCAT2	$\mathbf{5.3 \cdot 10^{-4}}$	$\mathbf{8.52 \cdot 10^{-3}}$
PI4K2B	$\mathbf{5.43 \cdot 10^{-4}}$	$\mathbf{8.69 \cdot 10^{-3}}$
RBBP8	$\mathbf{5.98 \cdot 10^{-4}}$	$\mathbf{9.51 \cdot 10^{-3}}$
CYP20A1	$\mathbf{6.12 \cdot 10^{-4}}$	$\mathbf{9.69 \cdot 10^{-3}}$
PHLPP2	$\mathbf{6.3 \cdot 10^{-4}}$	$\mathbf{9.92 \cdot 10^{-3}}$
LYST	$\mathbf{6.42 \cdot 10^{-4}}$	$\mathbf{1.01 \cdot 10^{-2}}$
APPL1	$\mathbf{6.83 \cdot 10^{-4}}$	$\mathbf{1.04 \cdot 10^{-2}}$
AXL	$\mathbf{6.83 \cdot 10^{-4}}$	$\mathbf{1.04 \cdot 10^{-2}}$
CDK13	$\mathbf{6.83 \cdot 10^{-4}}$	$\mathbf{1.04 \cdot 10^{-2}}$

Table A.11 Top 200 step-down candidates significantly associated with the time to BCR in the MSKCC dataset, sorted by the log-rank p-value.

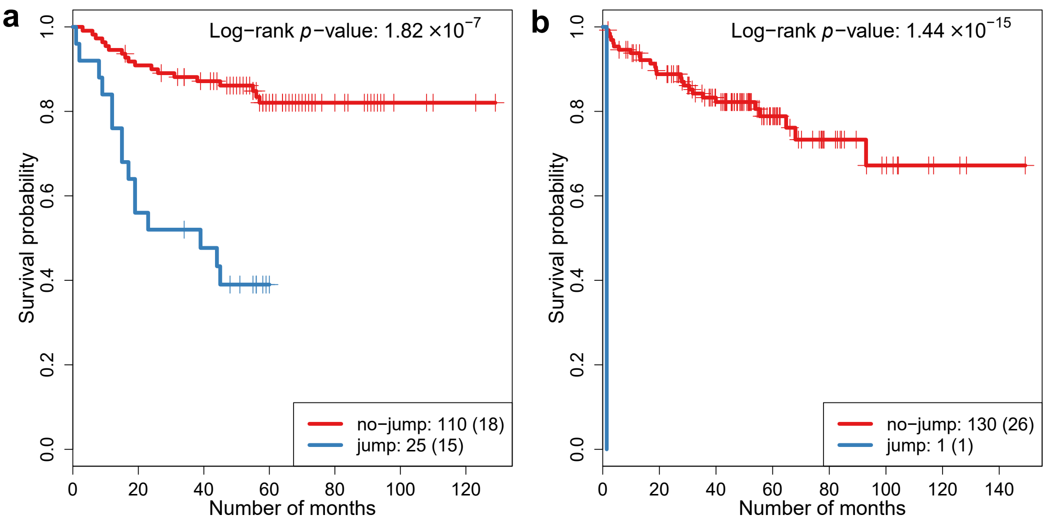

Figure A.6 KM plots for the *AKAP7* step-down jumps in a) CancerMap and b) MSKCC.

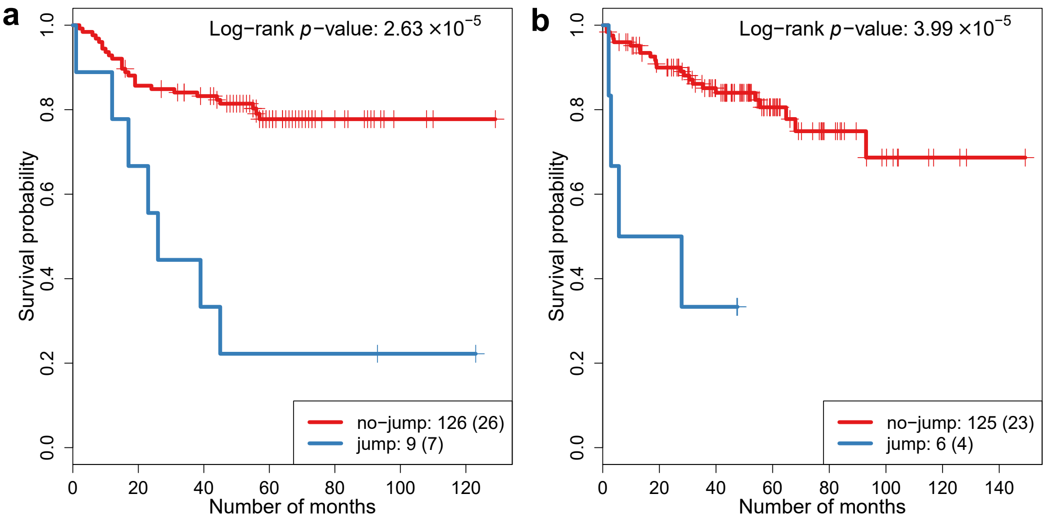

Figure A.7 KM plots for the *ALDH3A2* step-down jumps in a) CancerMap and b) MSKCC.

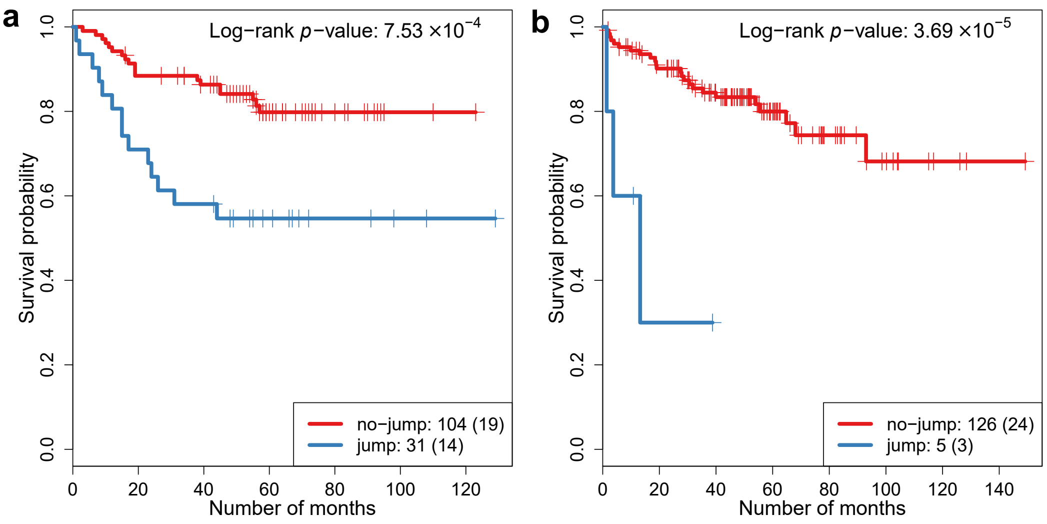

Figure A.8 KM plots for the *ASPA* step-down jumps in a) CancerMap and b) MSKCC.

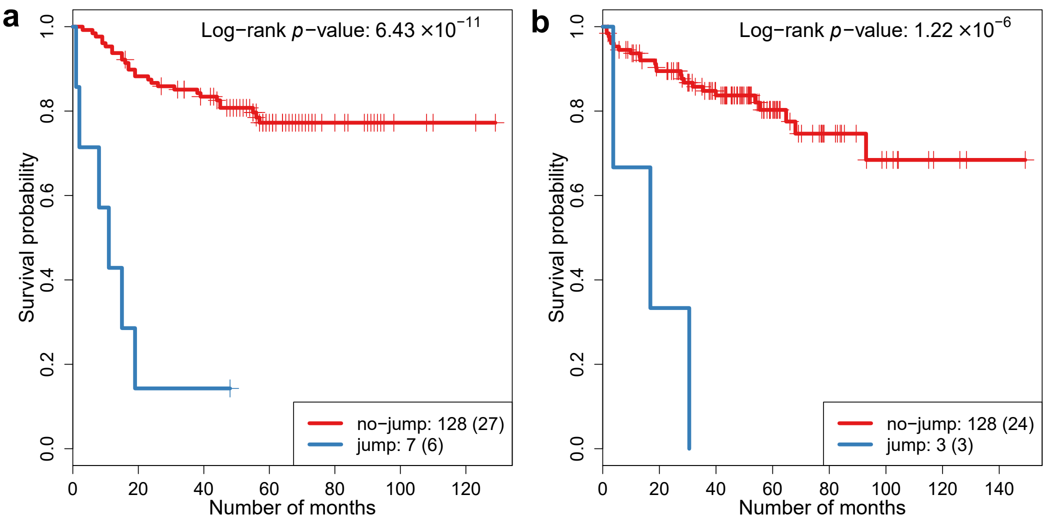

Figure A.9 KM plots for the *ARMCX1* step-down jumps in a) CancerMap and b) MSKCC.

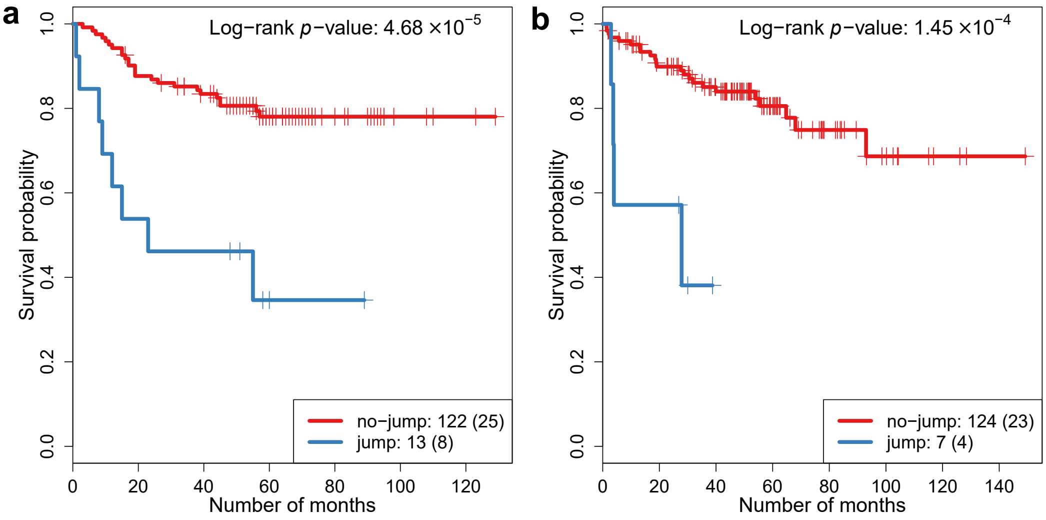

Figure A.10 KM plots for the *DIXDC1* step-down jumps in a) CancerMap and b) MSKCC.

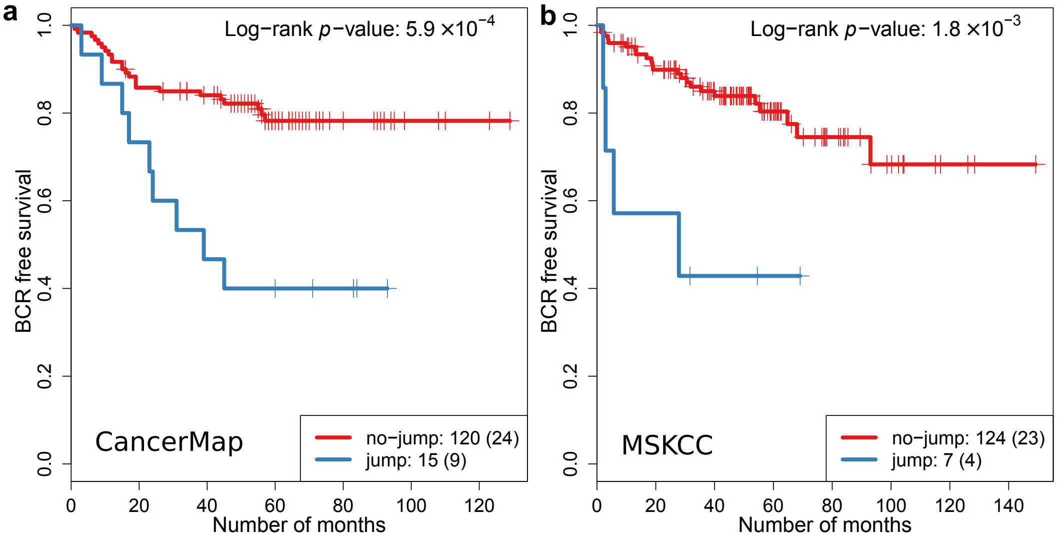

Figure A.11 KM plots for the *HSDL2* step-down jumps in a) CancerMap and b) MSKCC.

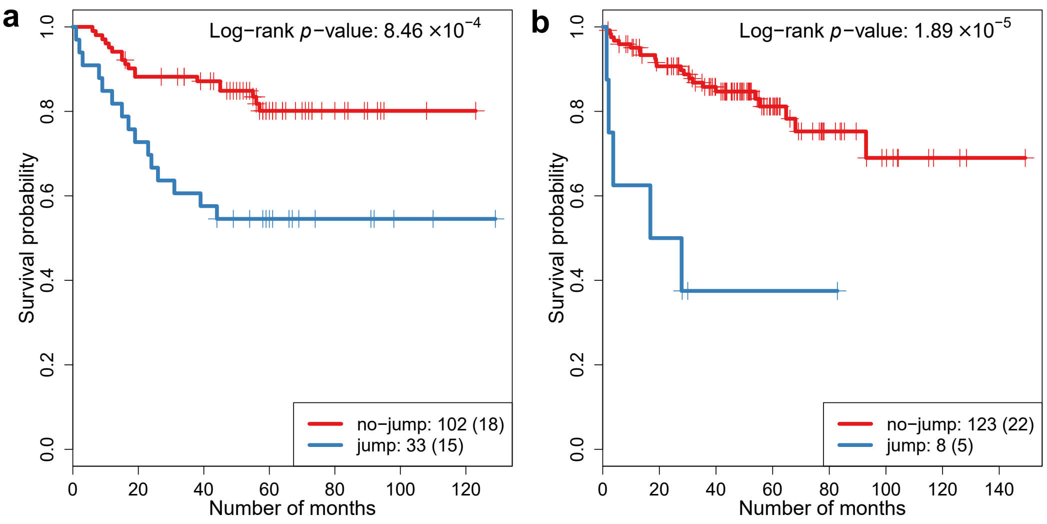

Figure A.12 KM plots for the *LRCH2* step-down jumps in a) CancerMap and b) MSKCC.

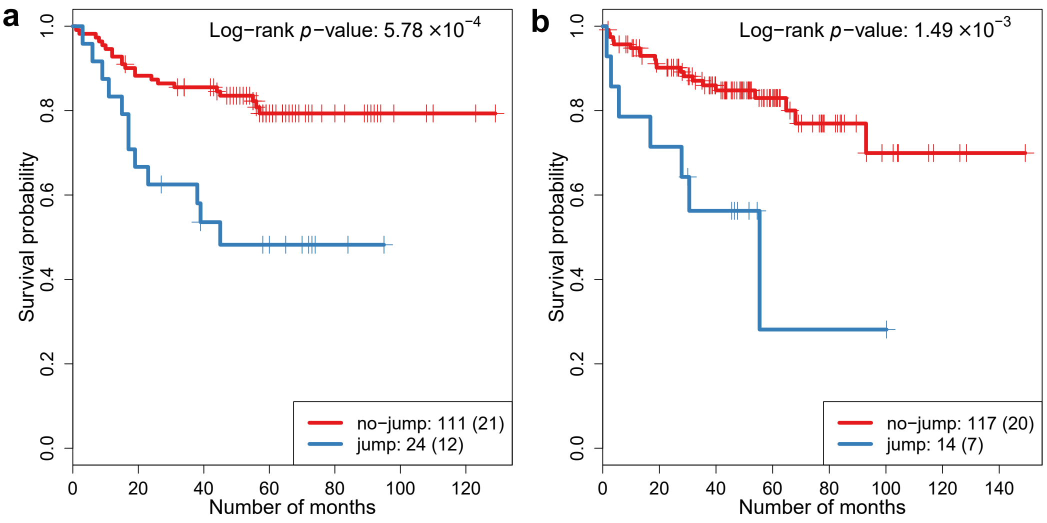

Figure A.13 KM plots for the *PI15* step-down jumps in a) CancerMap and b) MSKCC.

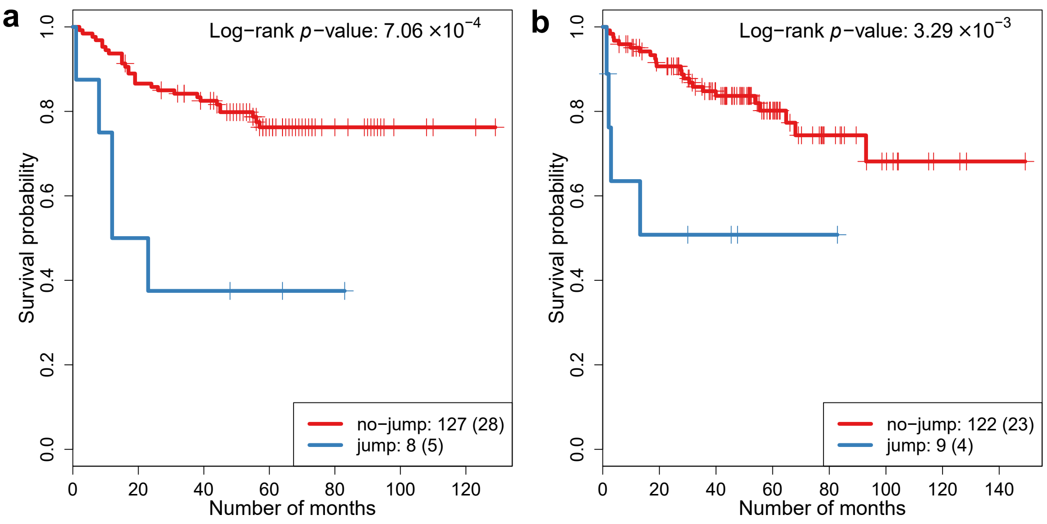

Figure A.14 KM plots for the *VAT1* step-down jumps in a) CancerMap and b) MSKCC.

3' partner	**CancerMap *p*-value**	**CancerMap *p*-adj.**	**MSKCC *p*-value**	**MKSCC *p*-adj.**
CD9	$8.81 \cdot 10^{-2}$	$9.09 \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
PIGU	$1.24 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$3.87 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
FKBP5	$2.34 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$5.91 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
FOXP1	$2.84 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$5.81 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
PDE4D	$3.59 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$8.57 \cdot 10^{-1}$	$9.77 \cdot 10^{-1}$
FAF1	$3.62 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$4.46 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
ETV1	$4.66 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$9.75 \cdot 10^{-2}$	$9.1 \cdot 10^{-1}$
ELK4	$4.84 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$4.33 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
ETV5	$4.95 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$8.72 \cdot 10^{-1}$	$9.81 \cdot 10^{-1}$
LYN	$5.81 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$8.05 \cdot 10^{-1}$	$9.65 \cdot 10^{-1}$
TMPRSS2	$5.99 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$7.51 \cdot 10^{-1}$	$9.53 \cdot 10^{-1}$
ZNF577	$6.13 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$8.73 \cdot 10^{-1}$	$9.81 \cdot 10^{-1}$
PRKG1	$6.71 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$4.33 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
BRAF	$6.9 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	$3.81 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$
ETV4	$7.22 \cdot 10^{-1}$	$9.09 \cdot 10^{-1}$	NA	NA
ERG	$7.65 \cdot 10^{-1}$	$9.32 \cdot 10^{-1}$	$7.16 \cdot 10^{-1}$	$9.45 \cdot 10^{-1}$
FLI1	$8.36 \cdot 10^{-1}$	$9.52 \cdot 10^{-1}$	$3.76 \cdot 10^{-1}$	$9.1 \cdot 10^{-1}$

Table A.12 Correlation of step-up jumps in known 3' fusion partners with the time to BCR.

5’ partner	**CancerMap** ***p*-value**	**CancerMap** ***p*-adj.**	**MSKCC** ***p*-value**	**MKSCC** **p-adj.**
YIPF1	$\mathbf{7.67 \cdot 10^{-6}}$	$\mathbf{2.37 \cdot 10^{-4}}$	$1.49 \cdot 10^{-1}$	$7.51 \cdot 10^{-1}$
AZGP1	$1.12 \cdot 10^{-1}$	$4.34 \cdot 10^{-1}$	$\mathbf{3.98 \cdot 10^{-4}}$	$\mathbf{2.28 \cdot 10^{-2}}$
FKBP5	$\mathbf{3.02 \cdot 10^{-2}}$	$1.87 \cdot 10^{-1}$	$9.67 \cdot 10^{-1}$	$9.88 \cdot 10^{-1}$
HERPUD1	$9.35 \cdot 10^{-1}$	$9.65 \cdot 10^{-1}$	$\mathbf{2.29 \cdot 10^{-2}}$	$2.96 \cdot 10^{-1}$
DDX5	$1.11 \cdot 10^{-1}$	$4.32 \cdot 10^{-1}$	$2.63 \cdot 10^{-1}$	$7.89 \cdot 10^{-1}$
KLK2	$1.59 \cdot 10^{-1}$	$5.24 \cdot 10^{-1}$	$5.79 \cdot 10^{-1}$	$8.7 \cdot 10^{-1}$
ACSL3	$3.35 \cdot 10^{-1}$	$7.14 \cdot 10^{-1}$	$4.51 \cdot 10^{-1}$	$8.07 \cdot 10^{-1}$
ESRP1	$3.67 \cdot 10^{-1}$	$7.3 \cdot 10^{-1}$	$5.82 \cdot 10^{-1}$	$8.7 \cdot 10^{-1}$
ERG	$3.73 \cdot 10^{-1}$	$7.33 \cdot 10^{-1}$	$8.15 \cdot 10^{-1}$	$9.59 \cdot 10^{-1}$
RC3H2	$4.59 \cdot 10^{-1}$	$7.61 \cdot 10^{-1}$	$2.63 \cdot 10^{-1}$	$7.89 \cdot 10^{-1}$
TMPRSS2	$5.99 \cdot 10^{-1}$	$8.11 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	$7.72 \cdot 10^{-1}$
ALG5	$6.51 \cdot 10^{-1}$	$8.17 \cdot 10^{-1}$	$6.42 \cdot 10^{-1}$	$8.87 \cdot 10^{-1}$
KIF2A	$7.33 \cdot 10^{-1}$	$8.5 \cdot 10^{-1}$	$7.54 \cdot 10^{-1}$	$9.39 \cdot 10^{-1}$
PTEN	$8.79 \cdot 10^{-1}$	$9.37 \cdot 10^{-1}$	$5.13 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$
NDRG1	$8.84 \cdot 10^{-1}$	$9.4 \cdot 10^{-1}$	$4.57 \cdot 10^{-1}$	$8.08 \cdot 10^{-1}$
TBC1D12	$9.13 \cdot 10^{-1}$	$9.54 \cdot 10^{-1}$	$4.15 \cdot 10^{-1}$	$8.07 \cdot 10^{-1}$
HARS2	$9.95 \cdot 10^{-1}$	$9.96 \cdot 10^{-1}$	$2.49 \cdot 10^{-1}$	$7.89 \cdot 10^{-1}$

Table A.13 Correlation of step-down jumps in known 5’ fusion partners with the time to BCR.

Gene Symbol	χ^2 ***p*-val.**	**Adj. *p*-val.**	**Mets**	**Primary**
AR	$\mathbf{3.22 \cdot 10^{-11}}$	$\mathbf{1.05 \cdot 10^{-7}}$	7/19	1/160
HMMR	$\mathbf{5.52 \cdot 10^{-11}}$	$\mathbf{1.05 \cdot 10^{-7}}$	6/19	0/160
INMT.FAM188B	$\mathbf{5.52 \cdot 10^{-11}}$	$\mathbf{1.05 \cdot 10^{-7}}$	6/19	0/160
TOP2A	$\mathbf{1.57 \cdot 10^{-10}}$	$\mathbf{1.79 \cdot 10^{-7}}$	8/19	3/160
TPX2	$\mathbf{1.57 \cdot 10^{-10}}$	$\mathbf{1.79 \cdot 10^{-7}}$	8/19	3/160
ANLN	$\mathbf{2.61 \cdot 10^{-9}}$	$\mathbf{2.47 \cdot 10^{-6}}$	6/19	1/160
RFX6	$\mathbf{5.06 \cdot 10^{-9}}$	$\mathbf{4.11 \cdot 10^{-6}}$	5/19	0/160
CCNB1	$\mathbf{5.48 \cdot 10^{-8}}$	$\mathbf{3.9 \cdot 10^{-5}}$	8/19	6/160
BUB1B	$\mathbf{1.91 \cdot 10^{-7}}$	$\mathbf{1.09 \cdot 10^{-4}}$	5/19	1/160
CDC6	$\mathbf{1.91 \cdot 10^{-7}}$	$\mathbf{1.09 \cdot 10^{-4}}$	5/19	1/160
EZH2	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.71 \cdot 10^{-4}}$	4/19	0/160
MCM2	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.71 \cdot 10^{-4}}$	4/19	0/160
MEX3A	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.71 \cdot 10^{-4}}$	4/19	0/160
MTERFD1	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.71 \cdot 10^{-4}}$	4/19	0/160
TTK	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.71 \cdot 10^{-4}}$	6/19	3/160
BZW2	$\mathbf{2.56 \cdot 10^{-6}}$	$\mathbf{8.58 \cdot 10^{-4}}$	5/19	2/160
MELK	$\mathbf{2.56 \cdot 10^{-6}}$	$\mathbf{8.58 \cdot 10^{-4}}$	5/19	2/160
LPL	$\mathbf{2.74 \cdot 10^{-6}}$	$\mathbf{8.65 \cdot 10^{-4}}$	6/19	4/160

CASC5	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
CHEK1	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
MCMDC2	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
NCAPG	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
NUF2	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
PAK4	**$1.23 \cdot 10^{-5}$**	**$2.91 \cdot 10^{-3}$**	4/19	1/160
SCGN	**$1.72 \cdot 10^{-5}$**	**$3.92 \cdot 10^{-3}$**	7/19	8/160
ASAP3	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
BIRC5	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
COLEC12	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
DNAH11	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
E2F5	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
EXO1	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
FOXRED2	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
GGTA1P	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
GRIN3A	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
LINC00476	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
PBLD	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
PIK3R3	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
RITA1	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
SEPT3	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
SLC17A4	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
SPDYA	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
TSHR	**$3.73 \cdot 10^{-5}$**	**$5.05 \cdot 10^{-3}$**	3/19	0/160
SERPINI1	**$4.12 \cdot 10^{-5}$**	**$5.45 \cdot 10^{-3}$**	6/19	6/160
PTGFR	**$4.43 \cdot 10^{-5}$**	**$5.73 \cdot 10^{-3}$**	7/19	9/160
ACER3	**$8.28 \cdot 10^{-5}$**	**$1.02 \cdot 10^{-2}$**	5/19	4/160
SGPP2	**$8.28 \cdot 10^{-5}$**	**$1.02 \cdot 10^{-2}$**	5/19	4/160
PPFIA2	**$1.02 \cdot 10^{-4}$**	**$1.23 \cdot 10^{-2}$**	7/19	10/160
ABCC5	**$1.13 \cdot 10^{-4}$**	**$1.27 \cdot 10^{-2}$**	4/19	2/160
CA13	**$1.13 \cdot 10^{-4}$**	**$1.27 \cdot 10^{-2}$**	4/19	2/160
NOL4	**$1.13 \cdot 10^{-4}$**	**$1.27 \cdot 10^{-2}$**	4/19	2/160
TTC21B	**$1.13 \cdot 10^{-4}$**	**$1.27 \cdot 10^{-2}$**	4/19	2/160
LAMA3	**$1.17 \cdot 10^{-4}$**	**$1.28 \cdot 10^{-2}$**	6/19	7/160
DAB1	**$2.8 \cdot 10^{-4}$**	**$2.9 \cdot 10^{-2}$**	5/19	5/160
DEGS1	**$2.8 \cdot 10^{-4}$**	**$2.9 \cdot 10^{-2}$**	5/19	5/160
SMC2	**$2.8 \cdot 10^{-4}$**	**$2.9 \cdot 10^{-2}$**	5/19	5/160
DNAH8	**$4.2 \cdot 10^{-4}$**	**$4.27 \cdot 10^{-2}$**	8/19	16/160
ABHD10	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160
COPS5	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160
CYFIP2	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160
DNAH14	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160
EEF1A2	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160
FANCI	**$5.58 \cdot 10^{-4}$**	**$4.73 \cdot 10^{-2}$**	4/19	3/160

GPR75.ASB3	$\mathbf{5.58 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	4/19	3/160
HDAC9	$\mathbf{5.58 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	4/19	3/160
ATP12A	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
CASP6	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
CES4A	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
CLPTM1	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
CXorf22	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
FANCB	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
KIF14	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
LACTB2	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
LAMC2	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
LRRC31	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
NMT2	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
NOP56	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
RACGAP1	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
RP11.26J3.4	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160
SMARCA5	$\mathbf{6.56 \cdot 10^{-4}}$	$\mathbf{4.73 \cdot 10^{-2}}$	3/19	1/160

Table A.14 Novel candidates exhibiting step-up jumps over-represented in the metastatic samples.

Gene Symbol	χ^2 *p*-val.	Adj. *p*-val.	Mets	Primary
CNN1	$\mathbf{2.44 \cdot 10^{-31}}$	$\mathbf{1.2 \cdot 10^{-27}}$	19/19	4/160
DES	$\mathbf{3.74 \cdot 10^{-30}}$	$\mathbf{9.17 \cdot 10^{-27}}$	17/19	2/160
ACTG2	$\mathbf{1.68 \cdot 10^{-28}}$	$\mathbf{2.74 \cdot 10^{-25}}$	17/19	3/160
SORBS1	$\mathbf{3.66 \cdot 10^{-22}}$	$\mathbf{4.49 \cdot 10^{-19}}$	14/19	3/160
TAGLN	$\mathbf{1.52 \cdot 10^{-21}}$	$\mathbf{1.49 \cdot 10^{-18}}$	15/19	5/160
DPP4	$\mathbf{3.76 \cdot 10^{-16}}$	$\mathbf{3.07 \cdot 10^{-13}}$	11/19	3/160
SYNPO2	$\mathbf{1.02 \cdot 10^{-15}}$	$\mathbf{7.12 \cdot 10^{-13}}$	12/19	5/160
FHL1	$\mathbf{1.23 \cdot 10^{-15}}$	$\mathbf{7.56 \cdot 10^{-13}}$	16/19	14/160
PDE5A	$\mathbf{1.44 \cdot 10^{-15}}$	$\mathbf{7.86 \cdot 10^{-13}}$	10/19	2/160
SRD5A2	$\mathbf{4.9 \cdot 10^{-15}}$	$\mathbf{2.4 \cdot 10^{-12}}$	16/19	15/160
CNTN1	$\mathbf{6.2 \cdot 10^{-15}}$	$\mathbf{2.76 \cdot 10^{-12}}$	14/19	10/160
EDNRA	$\mathbf{1.02 \cdot 10^{-14}}$	$\mathbf{4.15 \cdot 10^{-12}}$	12/19	6/160
EPHA3	$\mathbf{3.13 \cdot 10^{-14}}$	$\mathbf{1.1 \cdot 10^{-11}}$	14/19	11/160
CCND2	$\mathbf{3.14 \cdot 10^{-14}}$	$\mathbf{1.1 \cdot 10^{-11}}$	10/19	3/160
PI15	$\mathbf{5.4 \cdot 10^{-14}}$	$\mathbf{1.76 \cdot 10^{-11}}$	15/19	14/160
TRPC4	$\mathbf{7.1 \cdot 10^{-14}}$	$\mathbf{2.17 \cdot 10^{-11}}$	11/19	5/160
AZGP1	$\mathbf{7.98 \cdot 10^{-14}}$	$\mathbf{2.3 \cdot 10^{-11}}$	12/19	7/160
KL	$\mathbf{1.28 \cdot 10^{-13}}$	$\mathbf{3.47 \cdot 10^{-11}}$	9/19	2/160
TRIM29	$\mathbf{3.01 \cdot 10^{-13}}$	$\mathbf{7.77 \cdot 10^{-11}}$	13/19	10/160
MFAP4	$\mathbf{4.41 \cdot 10^{-13}}$	$\mathbf{1.08 \cdot 10^{-10}}$	10/19	4/160
SYNM	$\mathbf{4.82 \cdot 10^{-13}}$	$\mathbf{1.12 \cdot 10^{-10}}$	18/19	26/160
CSRP1	$\mathbf{5.58 \cdot 10^{-13}}$	$\mathbf{1.22 \cdot 10^{-10}}$	16/19	19/160
GALNT12	$\mathbf{5.74 \cdot 10^{-13}}$	$\mathbf{1.22 \cdot 10^{-10}}$	7/19	0/160
CHRDL1	$\mathbf{9.45 \cdot 10^{-13}}$	$\mathbf{1.93 \cdot 10^{-10}}$	17/19	23/160
SPARCL1	$\mathbf{1.36 \cdot 10^{-12}}$	$\mathbf{2.67 \cdot 10^{-10}}$	13/19	11/160
LCP1	$\mathbf{4.2 \cdot 10^{-12}}$	$\mathbf{7.91 \cdot 10^{-10}}$	11/19	7/160
ACADL	$\mathbf{5.47 \cdot 10^{-12}}$	$\mathbf{9.58 \cdot 10^{-10}}$	13/19	12/160
WWTR1	$\mathbf{5.47 \cdot 10^{-12}}$	$\mathbf{9.58 \cdot 10^{-10}}$	13/19	12/160
HSD17B6	$\mathbf{1.02 \cdot 10^{-11}}$	$\mathbf{1.68 \cdot 10^{-9}}$	16/19	22/160
TGFB3	$\mathbf{1.03 \cdot 10^{-11}}$	$\mathbf{1.68 \cdot 10^{-9}}$	8/19	2/160
SLC22A3	$\mathbf{4.15 \cdot 10^{-11}}$	$\mathbf{6.56 \cdot 10^{-9}}$	15/19	20/160
CLMP	$\mathbf{5.52 \cdot 10^{-11}}$	$\mathbf{8.44 \cdot 10^{-9}}$	6/19	0/160
FERMT2	$\mathbf{6.49 \cdot 10^{-11}}$	$\mathbf{9.63 \cdot 10^{-9}}$	13/19	14/160
PGR	$\mathbf{1.9 \cdot 10^{-10}}$	$\mathbf{2.58 \cdot 10^{-8}}$	10/19	7/160
SCN7A	$\mathbf{1.9 \cdot 10^{-10}}$	$\mathbf{2.58 \cdot 10^{-8}}$	10/19	7/160
VCL	$\mathbf{1.9 \cdot 10^{-10}}$	$\mathbf{2.58 \cdot 10^{-8}}$	10/19	7/160
EFEMP1	$\mathbf{2.32 \cdot 10^{-10}}$	$\mathbf{2.99 \cdot 10^{-8}}$	9/19	5/160
ITM2C	$\mathbf{2.32 \cdot 10^{-10}}$	$\mathbf{2.99 \cdot 10^{-8}}$	9/19	5/160
CALD1	$\mathbf{5.47 \cdot 10^{-10}}$	$\mathbf{6.88 \cdot 10^{-8}}$	13/19	16/160
AF131217.1	$\mathbf{1.45 \cdot 10^{-9}}$	$\mathbf{1.62 \cdot 10^{-7}}$	9/19	6/160
HOXA13	$\mathbf{1.45 \cdot 10^{-9}}$	$\mathbf{1.62 \cdot 10^{-7}}$	9/19	6/160
MPPED2	$\mathbf{1.45 \cdot 10^{-9}}$	$\mathbf{1.62 \cdot 10^{-7}}$	9/19	6/160

PLXDC2	$\mathbf{1.45 \cdot 10^{-9}}$	$\mathbf{1.62 \cdot 10^{-7}}$	9/19	6/160
RNF150	$\mathbf{1.45 \cdot 10^{-9}}$	$\mathbf{1.62 \cdot 10^{-7}}$	9/19	6/160
RIC3	$\mathbf{1.53 \cdot 10^{-9}}$	$\mathbf{1.67 \cdot 10^{-7}}$	8/19	4/160
ATP2B4	$\mathbf{1.75 \cdot 10^{-9}}$	$\mathbf{1.87 \cdot 10^{-7}}$	12/19	14/160
BVES	$\mathbf{2.61 \cdot 10^{-9}}$	$\mathbf{2.6 \cdot 10^{-7}}$	6/19	1/160
IGFBP6	$\mathbf{2.61 \cdot 10^{-9}}$	$\mathbf{2.6 \cdot 10^{-7}}$	6/19	1/160
MEIS1	$\mathbf{2.61 \cdot 10^{-9}}$	$\mathbf{2.6 \cdot 10^{-7}}$	6/19	1/160
MAOB	$\mathbf{3.51 \cdot 10^{-9}}$	$\mathbf{3.44 \cdot 10^{-7}}$	13/19	18/160
GPM6B	$\mathbf{3.74 \cdot 10^{-9}}$	$\mathbf{3.6 \cdot 10^{-7}}$	10/19	9/160
SLC14A1	$\mathbf{4.8 \cdot 10^{-9}}$	$\mathbf{4.52 \cdot 10^{-7}}$	12/19	15/160
EXOSC8	$\mathbf{5.06 \cdot 10^{-9}}$	$\mathbf{4.55 \cdot 10^{-7}}$	5/19	0/160
TNFSF15	$\mathbf{5.06 \cdot 10^{-9}}$	$\mathbf{4.55 \cdot 10^{-7}}$	5/19	0/160
AOX1	$\mathbf{5.11 \cdot 10^{-9}}$	$\mathbf{4.55 \cdot 10^{-7}}$	11/19	12/160
ITGA8	$\mathbf{7.25 \cdot 10^{-9}}$	$\mathbf{6.23 \cdot 10^{-7}}$	9/19	7/160
PGM5	$\mathbf{7.25 \cdot 10^{-9}}$	$\mathbf{6.23 \cdot 10^{-7}}$	9/19	7/160
TP63	$\mathbf{8.16 \cdot 10^{-9}}$	$\mathbf{6.89 \cdot 10^{-7}}$	13/19	19/160
RAB23	$\mathbf{9.13 \cdot 10^{-9}}$	$\mathbf{7.58 \cdot 10^{-7}}$	7/19	3/160
COG3	$\mathbf{1.05 \cdot 10^{-8}}$	$\mathbf{8.17 \cdot 10^{-7}}$	8/19	5/160
RCBTB2	$\mathbf{1.05 \cdot 10^{-8}}$	$\mathbf{8.17 \cdot 10^{-7}}$	8/19	5/160
STARD4	$\mathbf{1.05 \cdot 10^{-8}}$	$\mathbf{8.17 \cdot 10^{-7}}$	8/19	5/160
WLS	$\mathbf{1.05 \cdot 10^{-8}}$	$\mathbf{8.17 \cdot 10^{-7}}$	8/19	5/160
CASP7	$\mathbf{2.99 \cdot 10^{-8}}$	$\mathbf{2.22 \cdot 10^{-6}}$	9/19	8/160
FMOD	$\mathbf{2.99 \cdot 10^{-8}}$	$\mathbf{2.22 \cdot 10^{-6}}$	9/19	8/160
LPAR3	$\mathbf{2.99 \cdot 10^{-8}}$	$\mathbf{2.22 \cdot 10^{-6}}$	9/19	8/160
NR4A1	$\mathbf{3.96 \cdot 10^{-8}}$	$\mathbf{2.85 \cdot 10^{-6}}$	11/19	14/160
PALLD	$\mathbf{3.96 \cdot 10^{-8}}$	$\mathbf{2.85 \cdot 10^{-6}}$	11/19	14/160
CD38	$\mathbf{4.71 \cdot 10^{-8}}$	$\mathbf{3.21 \cdot 10^{-6}}$	6/19	2/160
DUSP5	$\mathbf{4.71 \cdot 10^{-8}}$	$\mathbf{3.21 \cdot 10^{-6}}$	6/19	2/160
NRK	$\mathbf{4.71 \cdot 10^{-8}}$	$\mathbf{3.21 \cdot 10^{-6}}$	6/19	2/160
PRELP	$\mathbf{4.71 \cdot 10^{-8}}$	$\mathbf{3.21 \cdot 10^{-6}}$	6/19	2/160
ITPR2	$\mathbf{5.48 \cdot 10^{-8}}$	$\mathbf{3.63 \cdot 10^{-6}}$	8/19	6/160
MYOF	$\mathbf{5.48 \cdot 10^{-8}}$	$\mathbf{3.63 \cdot 10^{-6}}$	8/19	6/160
OLFM4	$\mathbf{5.93 \cdot 10^{-8}}$	$\mathbf{3.88 \cdot 10^{-6}}$	18/19	46/160
DPYSL3	$\mathbf{7.13 \cdot 10^{-8}}$	$\mathbf{4.42 \cdot 10^{-6}}$	7/19	4/160
IRAK3	$\mathbf{7.13 \cdot 10^{-8}}$	$\mathbf{4.42 \cdot 10^{-6}}$	7/19	4/160
NEDD9	$\mathbf{7.13 \cdot 10^{-8}}$	$\mathbf{4.42 \cdot 10^{-6}}$	7/19	4/160
SH3BGRL	$\mathbf{7.13 \cdot 10^{-8}}$	$\mathbf{4.42 \cdot 10^{-6}}$	7/19	4/160
ANTXR2	$\mathbf{9.8 \cdot 10^{-8}}$	$\mathbf{5.93 \cdot 10^{-6}}$	11/19	15/160
LPHN2	$\mathbf{9.8 \cdot 10^{-8}}$	$\mathbf{5.93 \cdot 10^{-6}}$	11/19	15/160
FAM189A2	$\mathbf{1.06 \cdot 10^{-7}}$	$\mathbf{6.24 \cdot 10^{-6}}$	9/19	9/160
SOCS2	$\mathbf{1.06 \cdot 10^{-7}}$	$\mathbf{6.24 \cdot 10^{-6}}$	9/19	9/160
MAN1A1	$\mathbf{1.49 \cdot 10^{-7}}$	$\mathbf{8.68 \cdot 10^{-6}}$	13/19	23/160
FREM2	$\mathbf{1.91 \cdot 10^{-7}}$	$\mathbf{1.07 \cdot 10^{-5}}$	5/19	1/160
HSPA4L	$\mathbf{1.91 \cdot 10^{-7}}$	$\mathbf{1.07 \cdot 10^{-5}}$	5/19	1/160

MYH11	$\mathbf{1.91 \cdot 10^{-7}}$	$\mathbf{1.07 \cdot 10^{-5}}$	5/19	1/160
KRT23	$\mathbf{2.25 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	14/19	28/160
ACOT9	$\mathbf{2.29 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	8/19	7/160
JAM3	$\mathbf{2.29 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	8/19	7/160
MPP5	$\mathbf{2.29 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	8/19	7/160
NFE2L2	$\mathbf{2.29 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	8/19	7/160
STK19	$\mathbf{2.29 \cdot 10^{-7}}$	$\mathbf{1.21 \cdot 10^{-5}}$	8/19	7/160
FAM3B	$\mathbf{2.87 \cdot 10^{-7}}$	$\mathbf{1.5 \cdot 10^{-5}}$	12/19	20/160
ITGA1	$\mathbf{3.27 \cdot 10^{-7}}$	$\mathbf{1.68 \cdot 10^{-5}}$	9/19	10/160
PRICKLE2	$\mathbf{3.96 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	7/19	5/160
STOM	$\mathbf{3.96 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	7/19	5/160
DNAJB5	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
MEG8	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
MERTK	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
NEXN	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
PBX1	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
PXDN	$\mathbf{4.44 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	4/19	0/160
ANXA3	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
ARMCX1	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
ATRNL1	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
FADS2	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
PTER	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
SOCS2.AS1	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
SPOCK3	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
TPM1	$\mathbf{4.5 \cdot 10^{-7}}$	$\mathbf{1.98 \cdot 10^{-5}}$	6/19	3/160
ETS2	$\mathbf{4.94 \cdot 10^{-7}}$	$\mathbf{2.16 \cdot 10^{-5}}$	11/19	17/160
RMST	$\mathbf{5.61 \cdot 10^{-7}}$	$\mathbf{2.43 \cdot 10^{-5}}$	12/19	21/160
RAB27A	$\mathbf{7.37 \cdot 10^{-7}}$	$\mathbf{3.17 \cdot 10^{-5}}$	10/19	14/160
LPAR1	$\mathbf{8.04 \cdot 10^{-7}}$	$\mathbf{3.34 \cdot 10^{-5}}$	8/19	8/160
LRCH2	$\mathbf{8.04 \cdot 10^{-7}}$	$\mathbf{3.34 \cdot 10^{-5}}$	8/19	8/160
PAX9	$\mathbf{8.04 \cdot 10^{-7}}$	$\mathbf{3.34 \cdot 10^{-5}}$	8/19	8/160
SMAD9	$\mathbf{8.04 \cdot 10^{-7}}$	$\mathbf{3.34 \cdot 10^{-5}}$	8/19	8/160
PDE8B	$\mathbf{9.02 \cdot 10^{-7}}$	$\mathbf{3.71 \cdot 10^{-5}}$	9/19	11/160
TIMP3	$\mathbf{1.52 \cdot 10^{-6}}$	$\mathbf{6.22 \cdot 10^{-5}}$	13/19	27/160
SPG20	$\mathbf{1.65 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	10/19	15/160
ERAP1	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
FBXO32	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
FIP1L1	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
NEO1	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
PDS5B	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
ST8SIA6	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
TGFBR3	$\mathbf{1.69 \cdot 10^{-6}}$	$\mathbf{6.47 \cdot 10^{-5}}$	7/19	6/160
LEPREL1	$\mathbf{2.26 \cdot 10^{-6}}$	$\mathbf{8.52 \cdot 10^{-5}}$	9/19	12/160
SACS	$\mathbf{2.26 \cdot 10^{-6}}$	$\mathbf{8.52 \cdot 10^{-5}}$	9/19	12/160

DST	$\mathbf{2.43\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	8/19	9/160
MKX	$\mathbf{2.43\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	8/19	9/160
CAMK4	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
FAM63A	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
FOXP1	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
KLHDC1	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
MXI1	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
PRRG4	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
PTPN13	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
SEMA3D	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
TSPAN7	$\mathbf{2.56\cdot 10^{-6}}$	$\mathbf{8.91\cdot 10^{-5}}$	5/19	2/160
CAP2	$\mathbf{2.74\cdot 10^{-6}}$	$\mathbf{9.31\cdot 10^{-5}}$	6/19	4/160
MED4	$\mathbf{2.74\cdot 10^{-6}}$	$\mathbf{9.31\cdot 10^{-5}}$	6/19	4/160
SLC15A2	$\mathbf{2.74\cdot 10^{-6}}$	$\mathbf{9.31\cdot 10^{-5}}$	6/19	4/160
ANPEP	$\mathbf{3.45\cdot 10^{-6}}$	$\mathbf{1.15\cdot 10^{-4}}$	10/19	16/160
MMP2	$\mathbf{3.45\cdot 10^{-6}}$	$\mathbf{1.15\cdot 10^{-4}}$	10/19	16/160
PDK4	$\mathbf{3.45\cdot 10^{-6}}$	$\mathbf{1.15\cdot 10^{-4}}$	10/19	16/160
ZNF655	$\mathbf{3.78\cdot 10^{-6}}$	$\mathbf{1.25\cdot 10^{-4}}$	11/19	20/160
CTGF	$\mathbf{5.86\cdot 10^{-6}}$	$\mathbf{1.88\cdot 10^{-4}}$	7/19	7/160
EMP2	$\mathbf{5.86\cdot 10^{-6}}$	$\mathbf{1.88\cdot 10^{-4}}$	7/19	7/160
LMO7	$\mathbf{5.86\cdot 10^{-6}}$	$\mathbf{1.88\cdot 10^{-4}}$	7/19	7/160
SUGT1	$\mathbf{5.86\cdot 10^{-6}}$	$\mathbf{1.88\cdot 10^{-4}}$	7/19	7/160
ZMAT1	$\mathbf{5.86\cdot 10^{-6}}$	$\mathbf{1.88\cdot 10^{-4}}$	7/19	7/160
C12orf75	$\mathbf{6.49\cdot 10^{-6}}$	$\mathbf{2.06\cdot 10^{-4}}$	8/19	10/160
ALDH1A1	$\mathbf{6.86\cdot 10^{-6}}$	$\mathbf{2.17\cdot 10^{-4}}$	10/19	17/160
EDNRB	$\mathbf{9.4\cdot 10^{-6}}$	$\mathbf{2.95\cdot 10^{-4}}$	12/19	26/160
ANXA1	$\mathbf{1.12\cdot 10^{-5}}$	$\mathbf{3.46\cdot 10^{-4}}$	9/19	14/160
CAV2	$\mathbf{1.12\cdot 10^{-5}}$	$\mathbf{3.46\cdot 10^{-4}}$	9/19	14/160
FLNC	$\mathbf{1.2\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	6/19	5/160
GXYLT2	$\mathbf{1.2\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	6/19	5/160
PIKFYVE	$\mathbf{1.2\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	6/19	5/160
PTBP2	$\mathbf{1.2\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	6/19	5/160
SC5D	$\mathbf{1.2\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	6/19	5/160
EYA4	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
MAP3K4	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
MIR17HG	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
N4BP2L1	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
PAM	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
PARM1	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
PCDHGC5	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
PHACTR2	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
STK33	$\mathbf{1.23\cdot 10^{-5}}$	$\mathbf{3.5\cdot 10^{-4}}$	4/19	1/160
TSC22D3	$\mathbf{1.3\cdot 10^{-5}}$	$\mathbf{3.67\cdot 10^{-4}}$	10/19	18/160
MYBPC1	$\mathbf{1.56\cdot 10^{-5}}$	$\mathbf{4.35\cdot 10^{-4}}$	13/19	32/160

CRLS1	$\mathbf{1.56 \cdot 10^{-5}}$	$\mathbf{4.35 \cdot 10^{-4}}$	8/19	11/160
PTEN	$\mathbf{1.56 \cdot 10^{-5}}$	$\mathbf{4.35 \cdot 10^{-4}}$	8/19	11/160
ATL2	$\mathbf{1.72 \cdot 10^{-5}}$	$\mathbf{4.69 \cdot 10^{-4}}$	7/19	8/160
DMD	$\mathbf{1.72 \cdot 10^{-5}}$	$\mathbf{4.69 \cdot 10^{-4}}$	7/19	8/160
DNALI1	$\mathbf{1.72 \cdot 10^{-5}}$	$\mathbf{4.69 \cdot 10^{-4}}$	7/19	8/160
ZDBF2	$\mathbf{1.72 \cdot 10^{-5}}$	$\mathbf{4.69 \cdot 10^{-4}}$	7/19	8/160
APOL6	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
KCND3	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
KLHL5	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
RSRP1	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
TNC	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
YBX3	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
ZNF827	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
ZSWIM5	$\mathbf{1.81 \cdot 10^{-5}}$	$\mathbf{4.71 \cdot 10^{-4}}$	5/19	3/160
MME	$\mathbf{2.02 \cdot 10^{-5}}$	$\mathbf{5.23 \cdot 10^{-4}}$	11/19	23/160
ANK3	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
ARHGEF7	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
CRYAB	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
FOXN3	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
LATS2	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
LIMCH1	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
P2RX5.TAX1BP3	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
PCDHGA6	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
PSTPIP2	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
SGK3	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160
SMOC1	$\mathbf{3.73 \cdot 10^{-5}}$	$\mathbf{8.91 \cdot 10^{-4}}$	3/19	0/160

Table A.15 Novel candidates exhibiting step-down jumps over-represented in the metastatic samples.

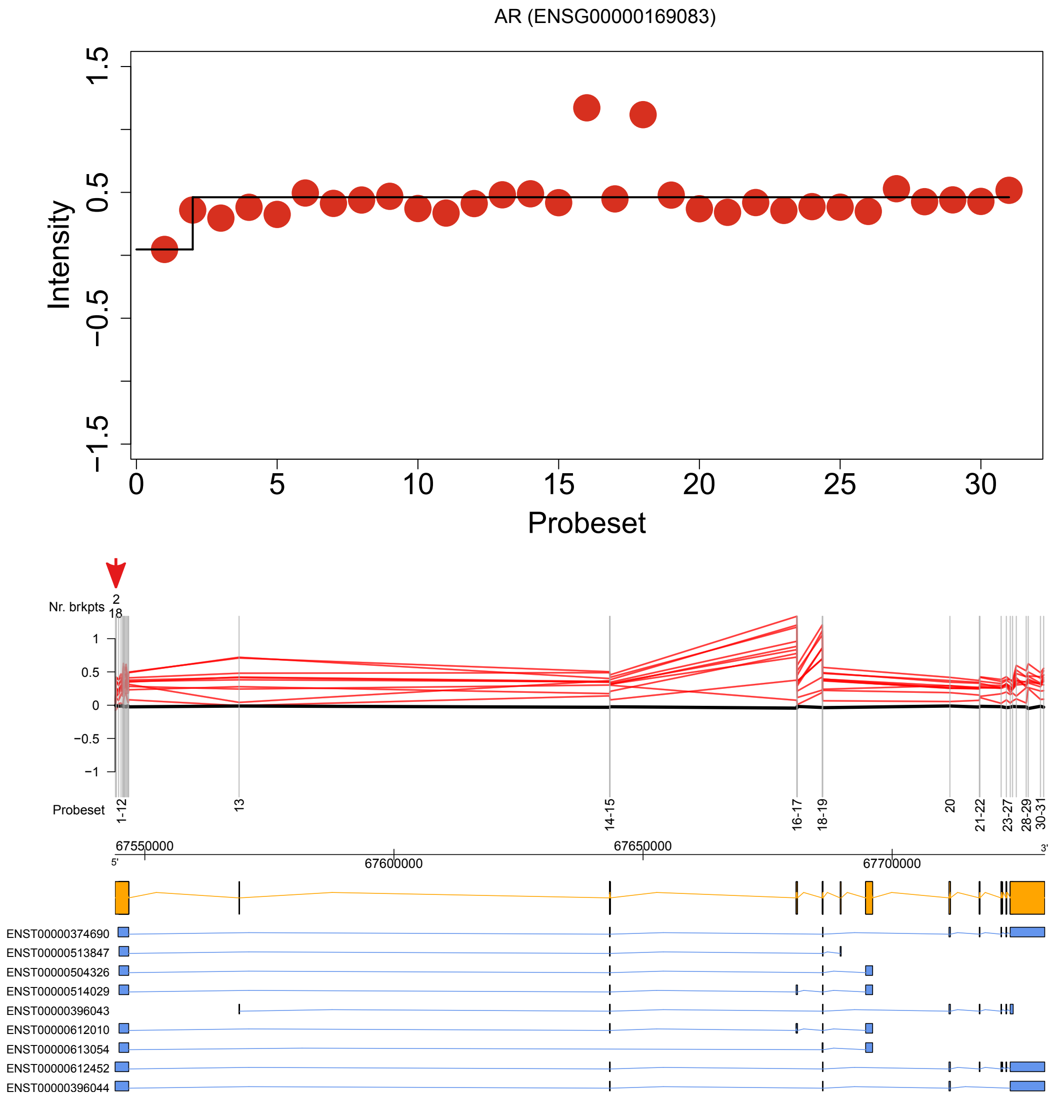

Figure A.15 Genomic plot depicting the mapping of the jumps to the *AR* gene model. In the top panel we depict several representative step-down jumps. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model, while each read line links the intensities of two consecutive probesets in a sample with step-up jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. In the bottom panel it is represented the gene model.

Gene Symbol	χ^2 ***p*-val.**	**Adj. *p*-val.**	**Mets**	**Primary**
FKBP5	$3.25 \cdot 10^{-1}$	1	1/19	27/160
ELK4	$3.73 \cdot 10^{-1}$	1	0/19	14/160
ERG	$4.05 \cdot 10^{-1}$	1	7/19	40/160
CD9	$4.11 \cdot 10^{-1}$	1	0/19	13/160

ZNF577	$5 \cdot 10^{-1}$	1	0/19	11/160
ETV4	$5.07 \cdot 10^{-1}$	1	1/19	1/160
PIGU	$5.07 \cdot 10^{-1}$	1	1/19	1/160
DIRC2	$6.13 \cdot 10^{-1}$	1	0/19	9/160
PDE4D	$6.13 \cdot 10^{-1}$	1	0/19	9/160
TBL1XR1	$6.82 \cdot 10^{-1}$	1	0/19	8/160
TMPRSS2	$7.53 \cdot 10^{-1}$	1	2/19	26/160
FAF1	$8.54 \cdot 10^{-1}$	1	0/19	6/160
FLI1	$8.54 \cdot 10^{-1}$	1	0/19	6/160
PLCE1	$8.54 \cdot 10^{-1}$	1	0/19	6/160
ETV1	$9.11 \cdot 10^{-1}$	1	2/19	11/160
FOXP1	$9.64 \cdot 10^{-1}$	1	0/19	5/160
ETV5	$1 \cdot 10^{0}$	1	1/19	5/160
LYN	$1 \cdot 10^{0}$	1	1/19	7/160
BRAF	1	1	0/19	3/160
PRKG1	1	1	0/19	3/160

Table A.16 Correlation of known 3’ fusion partners with the metastatic samples.

Gene Symbol	χ^2 *p*-val.	Adj. *p*-val.	Mets	Primary
AZGP1	**$7.98 \cdot 10^{-14}$**	**$2.3 \cdot 10^{-11}$**	12/19	7/160
FOXP1	**$2.56 \cdot 10^{-6}$**	**$8.91 \cdot 10^{-5}$**	5/19	2/160
PTEN	**$1.56 \cdot 10^{-5}$**	**$4.35 \cdot 10^{-4}$**	8/19	11/160
FKBP5	**$4.2 \cdot 10^{-4}$**	**$6.87 \cdot 10^{-3}$**	8/19	16/160
ALG5	**$1.85 \cdot 10^{-3}$**	**$2.27 \cdot 10^{-2}$**	4/19	4/160
TMPRSS2	**$4.62 \cdot 10^{-3}$**	**$4.23 \cdot 10^{-2}$**	9/19	27/160
KLK2	**$4.72 \cdot 10^{-3}$**	**$4.23 \cdot 10^{-2}$**	4/19	5/160
DDX5	**$1.84 \cdot 10^{-2}$**	$1.3 \cdot 10^{-1}$	4/19	7/160
NDRG1	**$2.66 \cdot 10^{-2}$**	$1.6 \cdot 10^{-1}$	6/19	17/160
HERPUD1	$5.25 \cdot 10^{-2}$	$2.72 \cdot 10^{-1}$	3/19	5/160
KIF2A	$1.29 \cdot 10^{-1}$	$5.32 \cdot 10^{-1}$	3/19	7/160
ARHGEF3	$2 \cdot 10^{-1}$	$6.96 \cdot 10^{-1}$	1/19	0/160
ACSL3	$2.45 \cdot 10^{-1}$	$7.8 \cdot 10^{-1}$	2/19	4/160
TNPO1	$2.95 \cdot 10^{-1}$	$9.15 \cdot 10^{-1}$	3/19	10/160
ZNF649	$4.27 \cdot 10^{-1}$	1	3/19	12/160
SMG5	$5.07 \cdot 10^{-1}$	1	1/19	1/160
YIPF1	$7.31 \cdot 10^{-1}$	1	1/19	2/160
ERG	$8.26 \cdot 10^{-1}$	1	2/19	10/160
EIF4E2	$1 \cdot 10^{0}$	1	0/19	1/160
ESRP1	$1 \cdot 10^{0}$	1	1/19	5/160
RC3H2	$1 \cdot 10^{0}$	1	1/19	11/160
BRAF	$1 \cdot 10^{0}$	1	0/19	4/160
MIPOL1	$1 \cdot 10^{0}$	1	0/19	2/160

HARS2	$1 \cdot 10^0$	1	1/19	10/160
TBC1D12	$1 \cdot 10^0$	1	1/19	7/160
PDZRN3	1	1	0/19	0/160

Table A.17 Correlation of known 5' fusion partners with the metastatic samples.

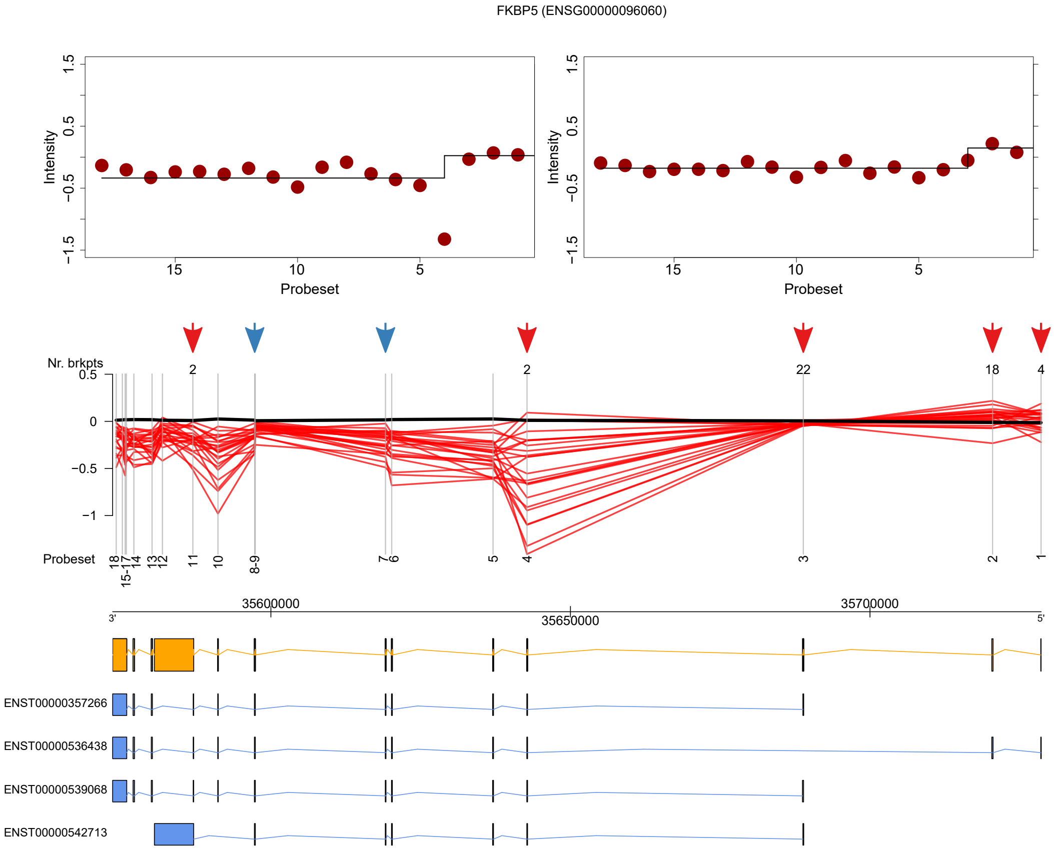

Figure A.16 Genomic plot depicting the mapping of the jumps to the *FKBP5* gene model. In the top panel we depict several representative step-down jumps. In the middle panel, the vertical lines correspond to the position where the probesets align to the gene model, while each read line links the intensities of two consecutive probesets in a sample with step-down jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. The blue arrows represent the positions of breakpoints reported in the literature. In the bottom panel it is represented the gene model.

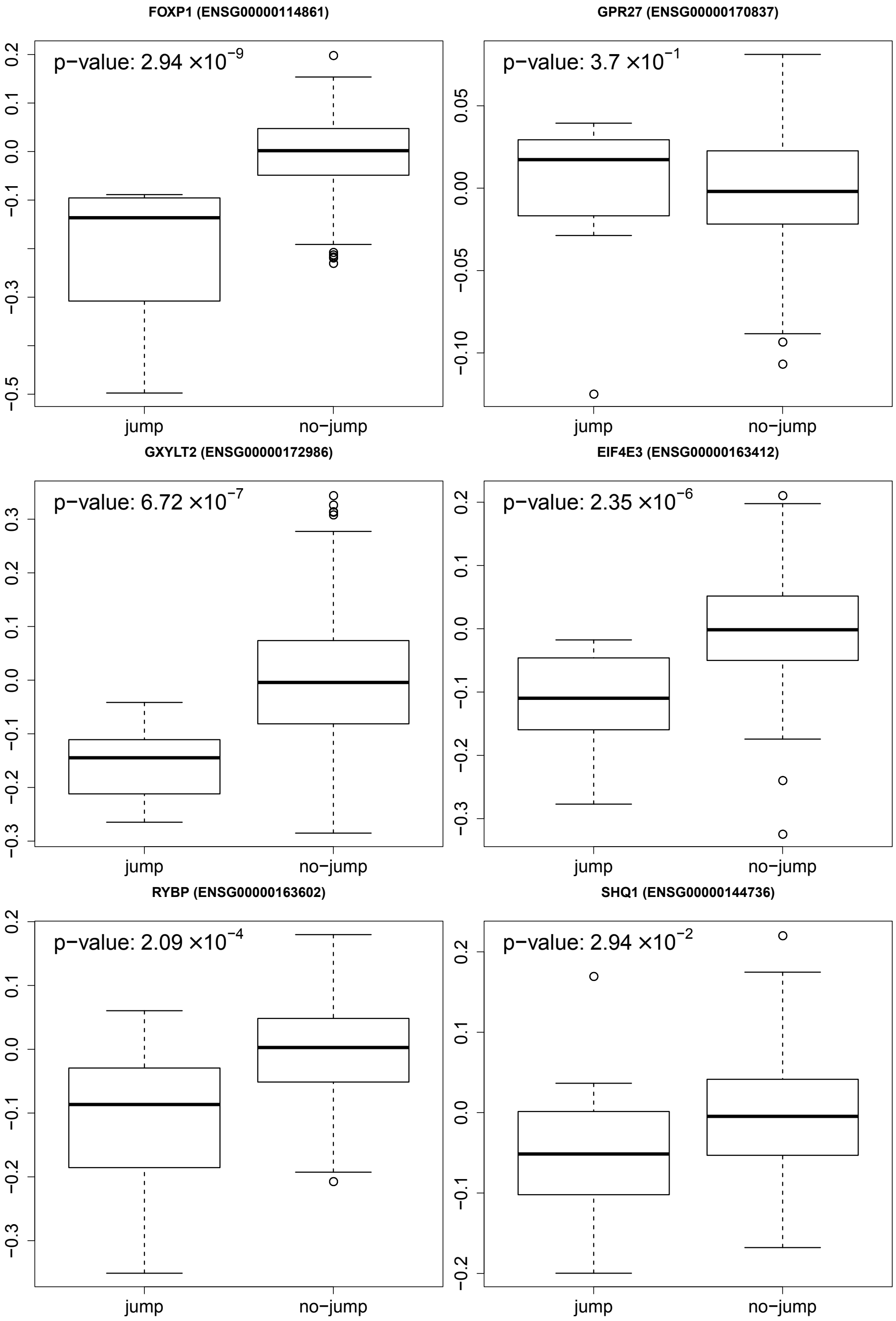

Figure A.17 Boxplots depicting the distribution of the expression levels of genes located on chromosome 3p13 (*FOXP1*, *GPR27*, *GXYLT2*, *EIF4E3*, *RYBP* and *SHQ1*) in samples with *FOXP1* jumps, and samples without.

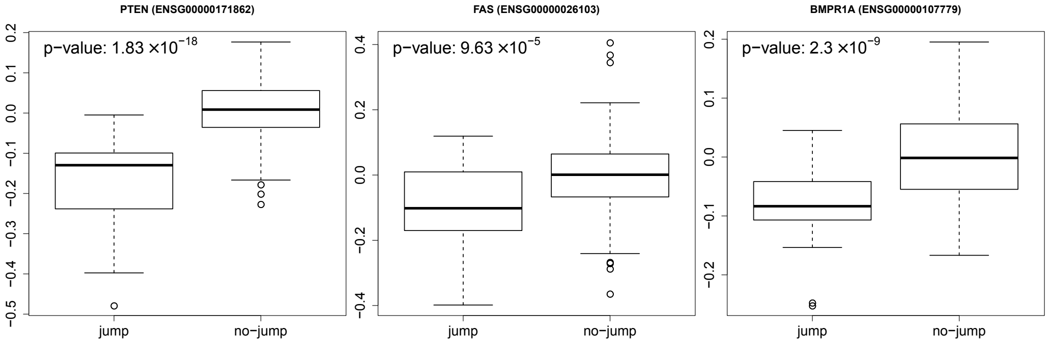

Figure A.18 Boxplots depicting the distribution of the expression levels of genes located on chromosome 10q23.3 (*PTEN*, *FAS* and *BMPR1A*) in samples with *PTEN* jumps, and samples without.

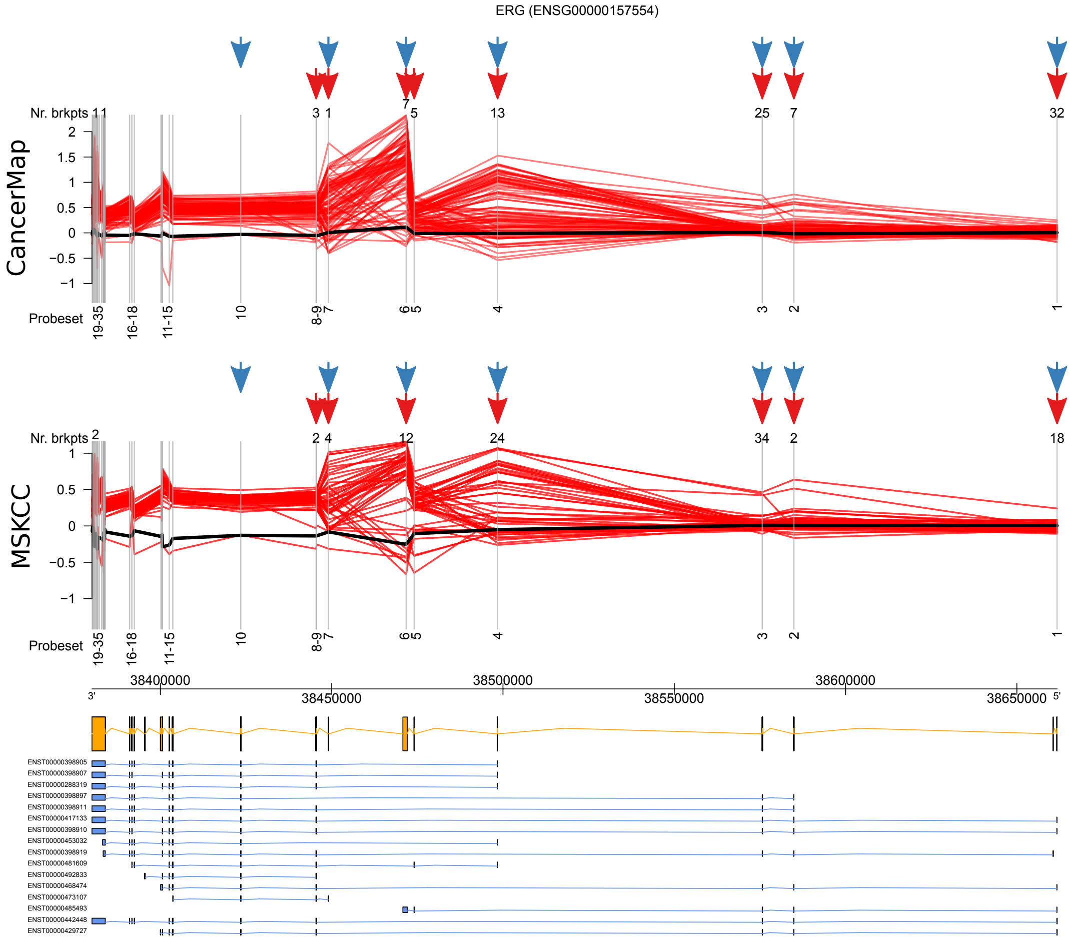

Figure A.19 Genomic plot depicting the mapping of the jumps in CancerMap (top panel) and MSKCC (middle panel) to the *ERG* gene model. In the two top panels, the vertical lines correspond to the position where the probesets align to the gene model, while each read line links the intensities of two consecutive probesets in a sample with step-up jumps. The red arrows represent the position of the putative breakpoints, i.e. the position where the step-down jumps occur. The numbers underneath the red arrows represent the number of putative breakpoints identified at that position. The blue arrows represent the positions of breakpoints reported in the literature. In the bottom panel it is represented the gene model.

Appendix B

Supplementary data for Chapter 5

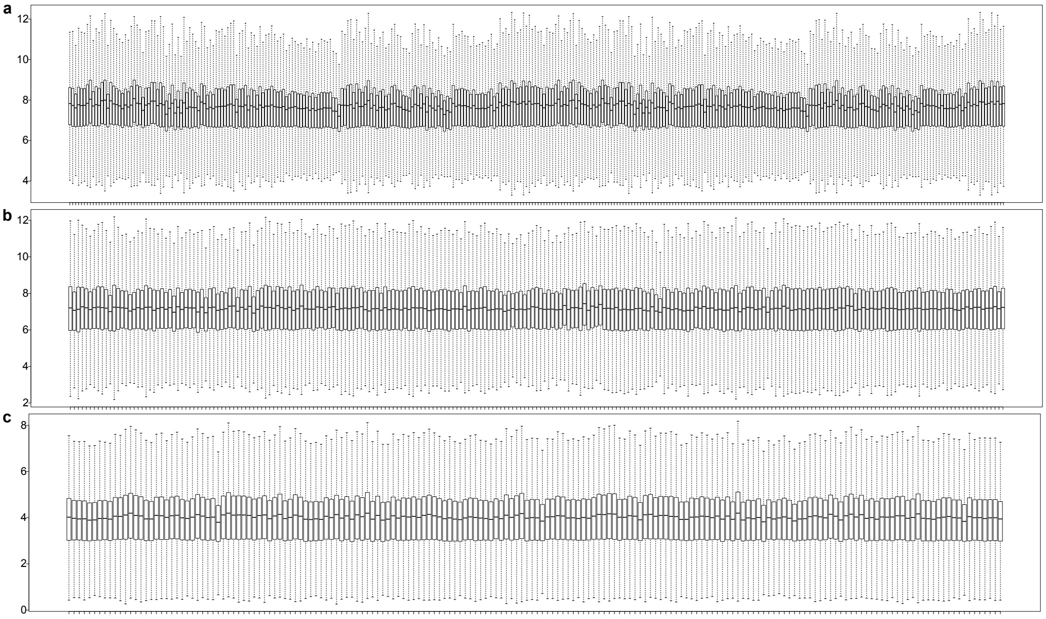

Figure B.1 Intensity boxplots for a) MSKCC, b) CancerMap and c) Klein obtained after RMA normalisation.

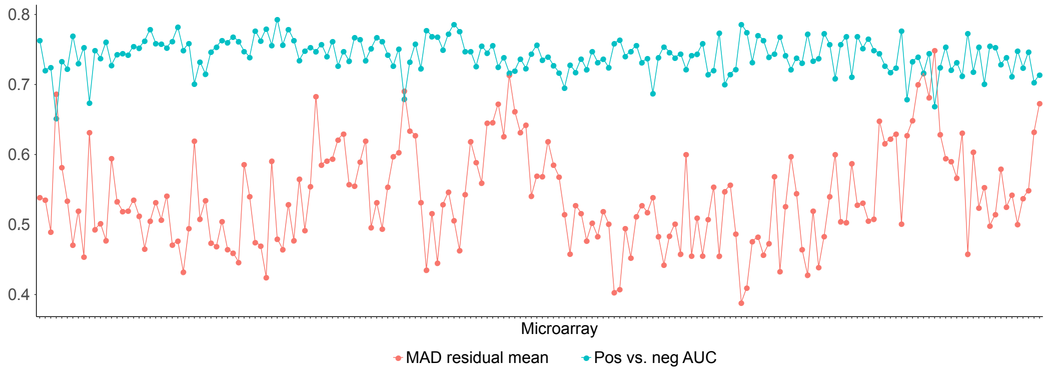

Figure B.2 QA plots for the Klein dataset. The blue points correspond to the positive vs. negative AUC values for a given microarray, while the red points correspond to the MAD of the residuals values.

TGM4	*RAET1L*	*CCL4*	*CD38*	*ARHGEF38*	*LOC100128816*
RLN1	*PCDHB3*	*RPF2*	*MMP23B*	*ACSL5*	*SYT1*
ORM1	*C1orf150*	*SLC45A3*	*OR51A7*	*SGK1*	*CPE*
OLFM4	*ALOX15B*	*SEC11C*	*CFB*	*TMEM45B*	*TRPC4*
OR51E2	*LSAMP*	*IFIT1*	*CCL2*	*AHNAK2*	*RAB27A*
SERPINB11	*SLC15A2*	*PAK1IP1*	*POTEM*	*NEDD8*	*CD69*
CRISP3	*PCP4*	*HIST1H3C*	*TPMT*	*GREB1*	*RPL17*
TDRD1	*MCCC2*	*ERRFI1*	*FAM3B*	*UBQLN4*	*PSCA*
SLC14A1	*GCNT1*	*ADAMTS1*	*FLRT3*	*SDHC*	*ATRNL1*
IGJ	*C5orf23*	*TRIM36*	*C7*	*TCEAL2*	*MYOCD*
ERG	*SCGB1D2*	*FLNA*	*NTN4*	*SLC18A2*	*MS4A8B*
GDEP	*CXCL2*	*CCND2*	*FAM36A*	*HIST1H2BE*	*TNS1*
TMEFF2	*AFF3*	*IFIT3*	*CNTNAP2*	*RARRES1*	*BAMBI*
CST1	*ATP8A2*	*FN1*	*SC4MOL*	*PLN*	*IGF1*
LTF	*PRIM2*	*PRY*	*RAP1B*	*OGN*	*RALGAPA1*
AMACR	*ADAMTSL1*	*HSPB8*	*SLC4A4*	*CLGN*	*S100A10*
SERPINA3	*NELL2*	*CD177*	*LCE2D*	*NIPAL3*	*PMS2CL*
NEFH	*RPS4Y1*	*TP63*	*EGR1*	*ACTG2*	*MMP2*
ACSM1	*CD24*	*IFI44*	*SCUBE2*	*RCAN3*	*SLC8A1*
OR51E1	*GOLGA6L9*	*COL12A1*	*FAM55D*	*KLK11*	*OAS2*
MT1G	*ZFP36*	*EDNRA*	*PDK4*	*HMGCS2*	*ARRDC3*
ANKRD36B	*TRIB1*	*PCDHB2*	*CXCL13*	*EML5*	*AMY2B*
LOC100510059	*BNIP3*	*HLA-DRA*	*CACNA1D*	*EDIL3*	*SPARCL1*
PLA2G2A	*KL*	*TUBA3E*	*GPR160*	*PIGH*	*IQGAP2*
TARP	*PDE5A*	*ASPN*	*CPM*	*GLYATL1*	*ACAD8*
REXO1L1	*DCN*	*FAM127A*	*PTGS2*	*ATP1B1*	*LPAR3*
ANPEP	*LDHB*	*DMD*	*TSPAN8*	*GJA1*	*HIGD2A*
HLA-DRB5	*PCDHB5*	*DHRS7*	*BMP5*	*PLA1A*	*NUCB2*

PLA2G7
NCAPD3
OR51F2
SPINK1
RCN1
CP
SMU1
ACTC1
AGR2
SLC26A4
IGK
MYBPC1
NPY
PI15
SLC22A3
PIGR
MME
HLA-DRB1
FOLH1
LUZP2
MSMB
GSTT1
MMP7
ODZ1
ACTB
SPON2
SLC38A11
FOS
OR51T1
HLA-DMB
KRT15
ITGA8
CXADR
LYZ
CEACAM20
C8orf4
DPP4
PGC
C15orf21
CHORDC1
LRRN1
MT1M
EPHA6
PDE11A
ACADL
ZNF99
CPNE4
CCDC144B
SLC26A2
CYP1B1
SELE
CLDN1
KRT13
SFRP2
SLC25A33
HSD17B11
HSD17B13
UGT2B4
CTGF
SCIN
C10orf81
CYR61
PRUNE2
IFI6
MYH11
PPP1R3C
KCNH8
ZNF615
ERV3
F3
TTN
LYRM5
FMOD
NEXN
IL28A
FHL1
CXCL10
SPOCK1
GSTP1
OAT
HIST2H2BF
ACSM3
GLB1L3
SLC5A1
OR4N4
MAOB
BZW1
IFI44L
ANO7
MEIS1
TSPAN1
CNTN1
TRIM22
GSTA2
SORBS1
GPR81
CSRP1
C3orf14
FGFR2
SNAI2
CALCRL
MON1B
PVRL3
VGLL3
SULF1
LIFR
C12orf75
GNPTAB
CALM2
KLF6
C7orf58
RDH11
NR4A1
RWDD4
ABCC4
GABRE
SLC16A1
DEGS1
CLDN8
HAS2
ODC1
REEP3
LYRM4
PPFIA2
PGM3
ZDHHC8P1
C6orf72
HIST1H2BD
TES
PDE8B
DNAJB4
RGS5
GOLGA8A
OR4N2
FAM135A
DYNLL1
DSC3
C4orf3
HIST1H2BK
LCN2
STEAP4
RPS27L
TRPM8
ID2
LUM
EDNRB
PGM5
SFRP4
STEAP1
FADS2
CXCL11
CWH43
SNRPN
GPR110
THBS1
APOD
HPGD
LEPREL1
LCE1D
GSTM5
SLC30A4
SEMA3D
CACNA2D1
GPR116
C7orf63
FAM198B
SCD
NR4A2
ARG2
ZNF385B
RGS1
DNAH5
NPR3
RAB3B
CHRDL1
MBOAT2
MPPED2
AMD1
EMP1
PRR16
CNN1
GHR
ALDH1A1
TRIM29
IFNA17
TAS2R4
SEPP1
GREM1
RASD1
C1S
CLSTN2
DMXL1
HIST1H2BC
NRG4
ARL17A
GRPR
PART1
CYP3A5
KCNC2
SERPINE1
SLC6A14
EIF4A1
MYOF
PHOSPHO2
GCNT2
AOX1
CCDC80
ATP2B4
UGDH
GSTM2
MEIS2
RGS2
PRKG2
FIBIN
FDXACB1
SOD2
SEPT7
PTPRC
GABRP
CBWD3
HLA-DPA1
SLITRK6
TPM2
REPS2
EAF2
CAV1
TMEM178
MFAP4
SYNM
EFEMP1
RND3
SCNN1A
B3GNT5
LMOD1
UBC
LMO3
LOX
NFIL3
C11orf92
C11orf48
BCAP29
EPCAM
PTGDS
ASB5
TUBA1B
SERHL
ITGA5
SPARC
LOC286161
NAALADL2
TMPRSS2
SERPINF1
EPHA7
SDAD1
SOX14
RPL35
HSPA1B
MSN
MTRF1L
PTN
CAMKK2
RBM7
OR52H1
C1R

TMSB15A	*KRT5*	*EPHA3*	*ATF3*	*TOR1AIP2*	*CHRNA2*
LYPLA1	*SCN7A*	*COX7A2*	*ST6GAL1*	*CXCR4*	*MRPL41*
FOSB	*GOLM1*	*MT1H*	*GDF15*	*OR51L1*	*PROM1*
F5	*HIST4H4*	*HIST2H2BE*	*ANXA1*	*SLC12A2*	*LPAR6*
C15orf48	*IL7R*	*TGFB3*	*C4B*	*AGAP11*	*SAMHD1*
MIPEP	*CSGALNACT1*	*VEGFA*	*ELOVL2*	*SLC27A2*	*SCNN1G*
HSD17B6	*A2M*	*CRISPLD2*	*GSTM1*	*AZGP1*	*DNAJC10*
SLPI	*LRRC9*	*TFF1*	*GLIPR1*	*VCAN*	*MOXD1*
MYO6	*KRT17*	*ID1*	*C3*	*ERAP2*	*HIST1H2BG*

Table B.1 The 489 genes used for LPD classification.

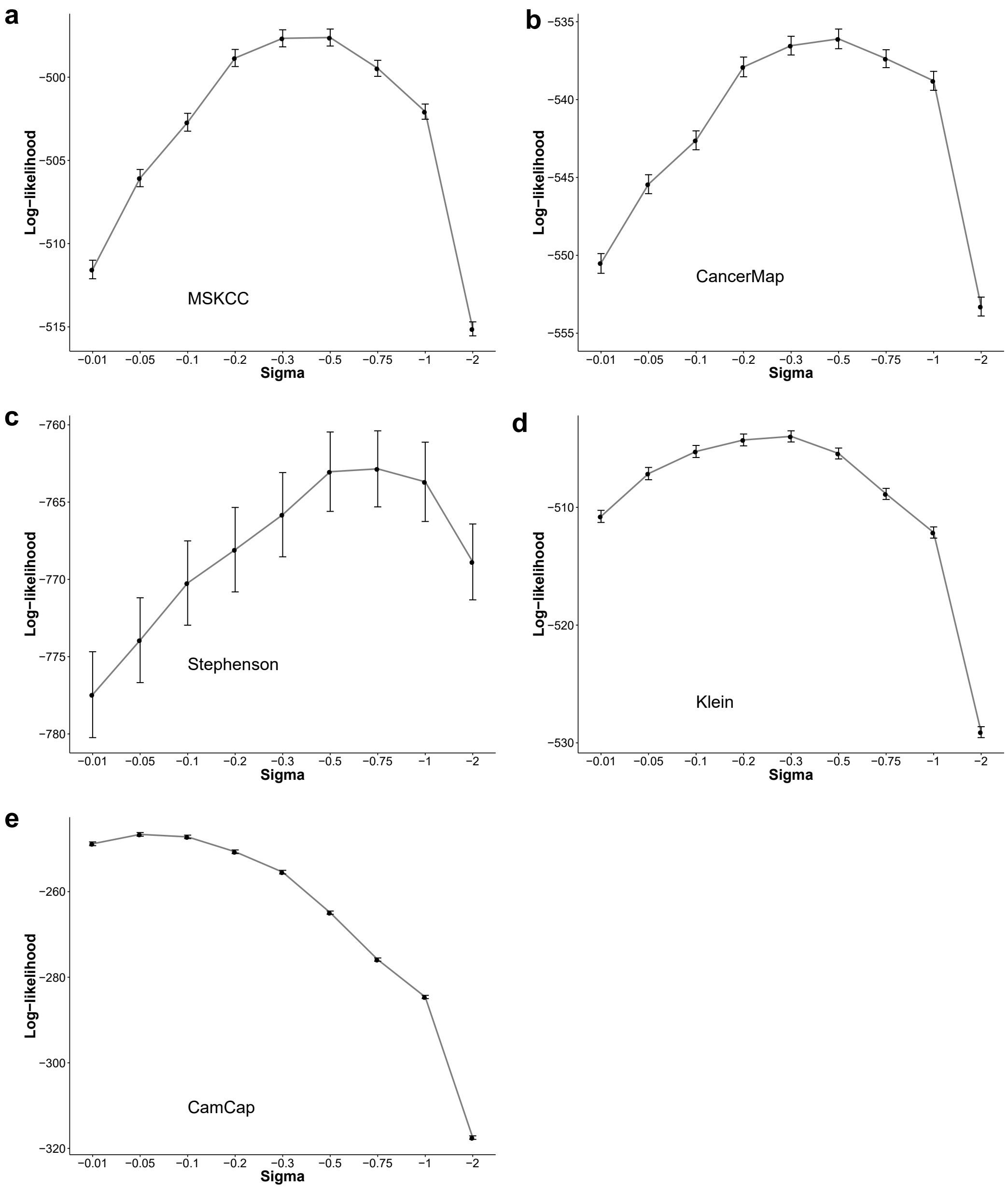

Figure B.3 The average log-likelihood of 100 LPD restarts (vertical axis) versus various choices for sigma (horizontal axis) using the MAP solution. The peak in log-likelihood indicates the optimal value for sigma.

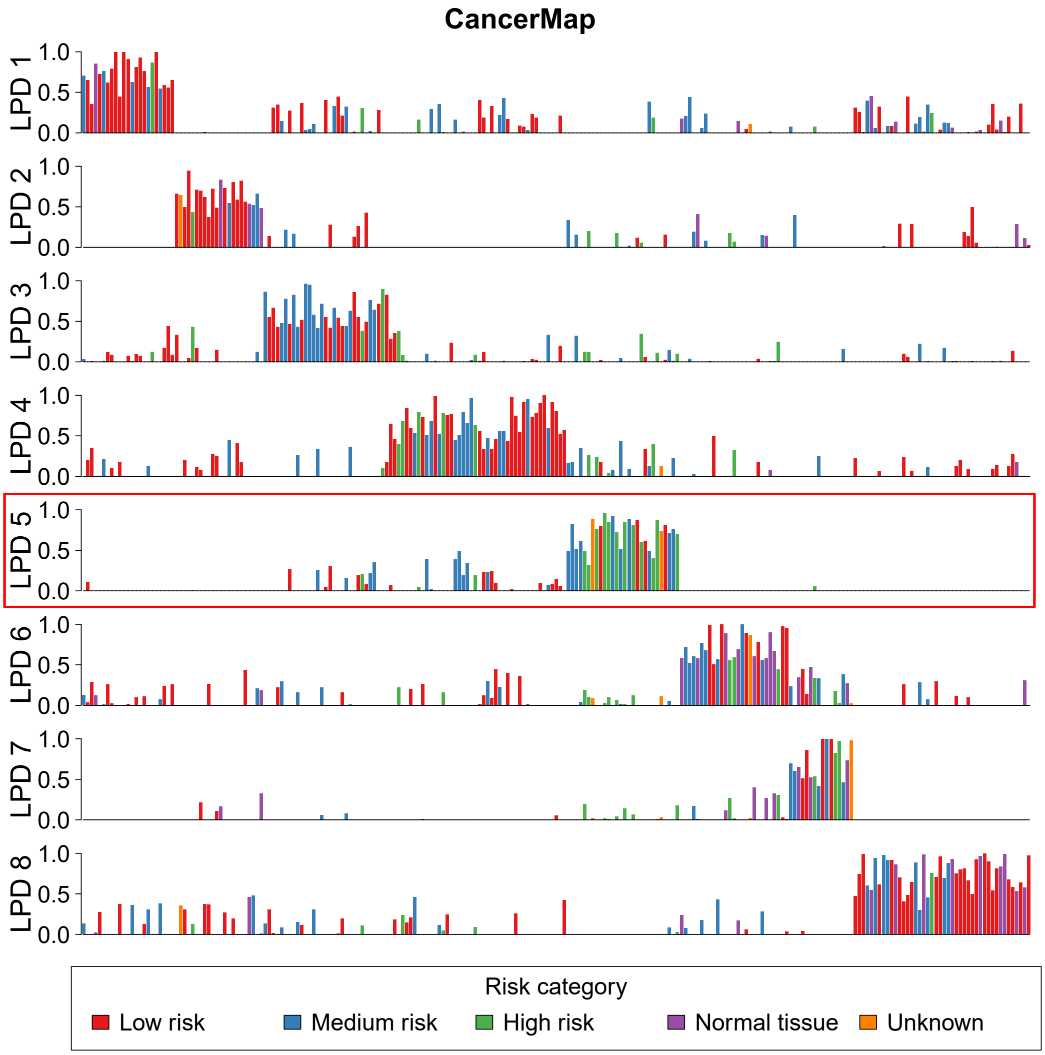

Figure B.4 An illustration of the LPD classification on the CancerMap dataset. Each horizontal panel corresponds to a LPD process. For each panel, the *x*-axis represent samples, while the *y*-axis represents the contribution of the process to the expression of the sample. The colours correspond to the ICGC risk categories defined in Section 2.5.5.5.

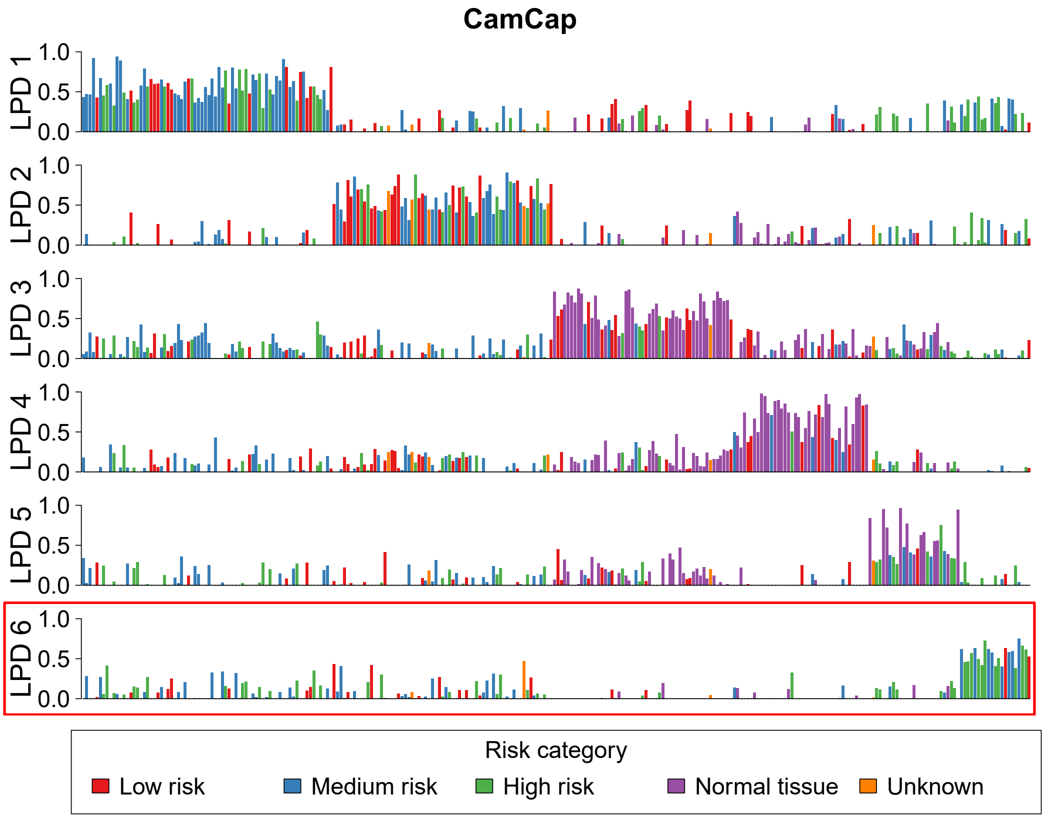

Figure B.5 An illustration of the LPD classification on the CamCap dataset. Each horizontal panel corresponds to a LPD process. For each panel, the *x*-axis represent samples, while the *y*-axis represents the contribution of the process to the expression of the sample. The colours correspond to the ICGC risk categories defined in Section 2.5.5.5.

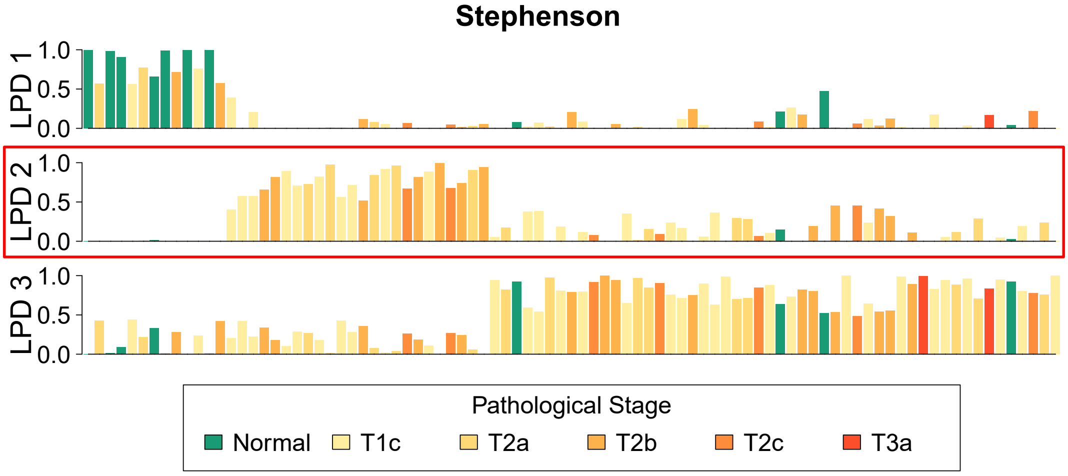

Figure B.6 An illustration of the LPD classification on the Stephenson dataset. Each horizontal panel corresponds to a LPD process. For each panel, the *x*-axis represent samples, while the *y*-axis represents the contribution of the process to the expression of the sample. The colours correspond to the pathological stage categories defined in Section 2.5.5.2.

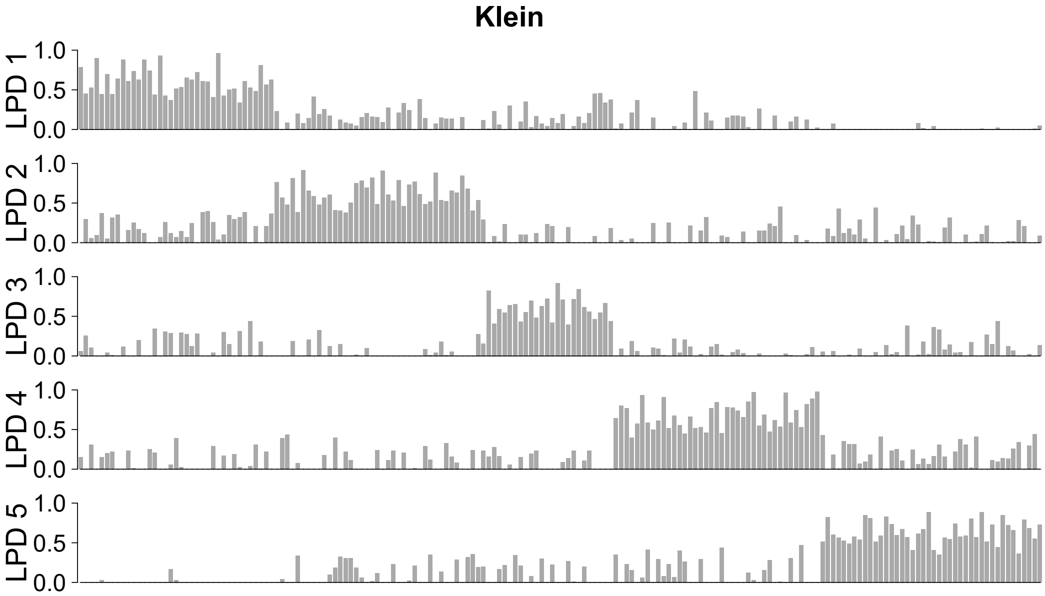

Figure B.7 An illustration of the LPD classification on the Klein dataset. Each horizontal panel corresponds to a LPD process. For each panel, the *x*-axis represent samples, while the *y*-axis represents the contribution of the process to the expression of the sample.

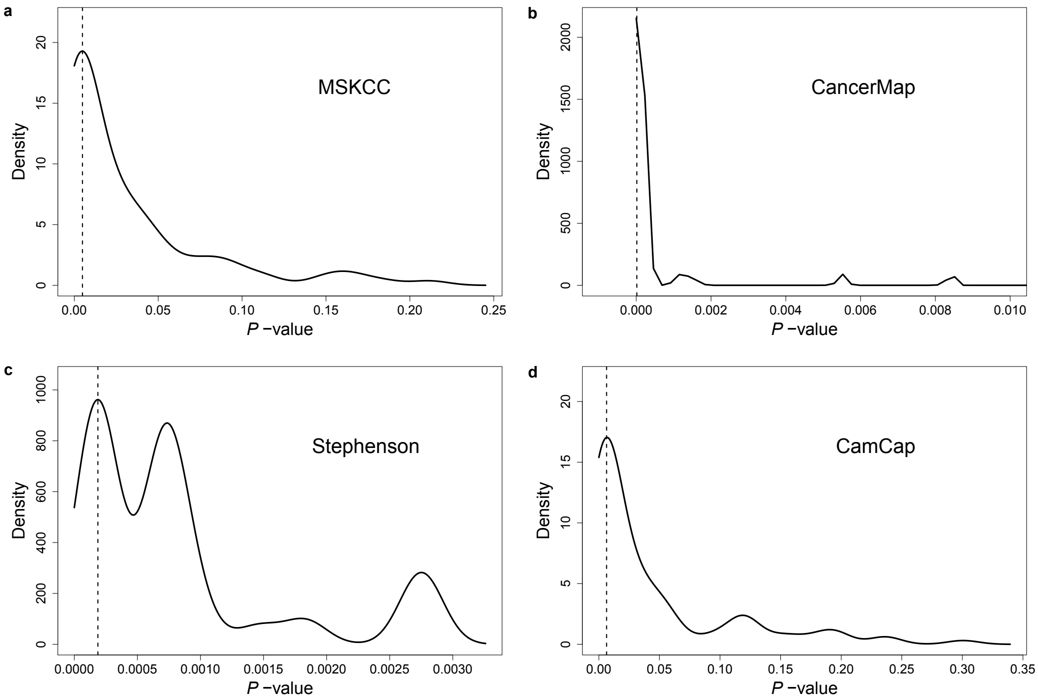

Figure B.8 The distribution of the PSA failure log-rank *p*-values of 100 LPD restarts with random seeds, for: a) MSKCC, b) CancerMap, c) CamCap and d) Stephenson. The vertical dashed lines correspond to the mode of the distribution.

MSKCC

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	1.67	0.59941	4.653	$\mathbf{3.27\cdot 10^{-1}}$
Gleason: $\leq / > 7$	5.1787	1.92652	13.921	$\mathbf{1.12\cdot 10^{-3}}$
PSA: $\leq / > 10$	1.1616	0.48496	2.782	$\mathbf{7.37\cdot 10^{-1}}$
Path Stage < 35 weeks: T2/T3-T4	5.5411	1.73982	17.648	$\mathbf{3.77\cdot 10^{-3}}$
Path Stage $\geq$ 35 weeks: T2/T3-T4	0.5261	0.06108	4.531	$\mathbf{5.59\cdot 10^{-1}}$

CancerMap

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	4.2938	1.6071	11.472	$\mathbf{3.66\cdot 10^{-3}}$
Gleason: $\leq / > 7$	3.4283	1.1492	10.227	$\mathbf{2.72\cdot 10^{-2}}$
Path Stage: T1-T2/T3-T4	1.7402	0.8111	3.734	$\mathbf{1.55\cdot 10^{-1}}$
PSA: $\leq / > 10$	1.2363	0.5409	2.826	$\mathbf{6.15\cdot 10^{-1}}$

Stephenson

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	3.8041	1.889	7.661	$\mathbf{1.83\cdot 10^{-4}}$
Path Stage: T1-T2/T3-T4	1.6947	0.8409	3.415	$\mathbf{1.4\cdot 10^{-1}}$
Gleason: $\leq / > 7$	2.0393	0.9881	4.209	$\mathbf{5.39\cdot 10^{-2}}$
PSA: $\leq / > 10$	1.9233	0.9753	3.793	$\mathbf{5.9\cdot 10^{-2}}$

CamCap

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	2.2504	1.086	4.662	$\mathbf{2.9\cdot 10^{-2}}$
Path Stage: T1-T2/T3-T4	2.2786	1.29	4.024	$\mathbf{4.53\cdot 10^{-3}}$
Gleason: $\leq / > 7$	4.2327	2.265	7.909	$\mathbf{6.08\cdot 10^{-6}}$
PSA: $\leq / > 10$	1.9441	1.151	3.283	$\mathbf{1.29\cdot 10^{-2}}$

Combined

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	*p*-value
Non-DESNT/DESNT	3.5158	2.1967	5.627	$\mathbf{1.61\cdot 10^{-7}}$
Gleason: $\leq / > 7$	3.0926	1.8739	5.104	$\mathbf{1\cdot 10^{-5}}$
PSA: $\leq / > 10$	1.4261	0.9363	2.172	$\mathbf{9.83\cdot 10^{-2}}$
Path Stage: T1-T2/T3-T4	1.9184	1.2609	2.919	$\mathbf{2.34\cdot 10^{-3}}$
Dataset: MSKCC/CancerMap	1.3087	0.7709	2.222	$\mathbf{3.19\cdot 10^{-1}}$
Dataset: MSKCC/Stephenson	1.5318	0.9287	2.527	$\mathbf{9.49\cdot 10^{-2}}$

Table B.2 Summary of the extended Cox PH model using the DESNT membership, Gleason score ($\leq / > 7$), PSA ($\leq / > 10$) and stage (T1-T2/T3-T4) as predictors of recurrence.

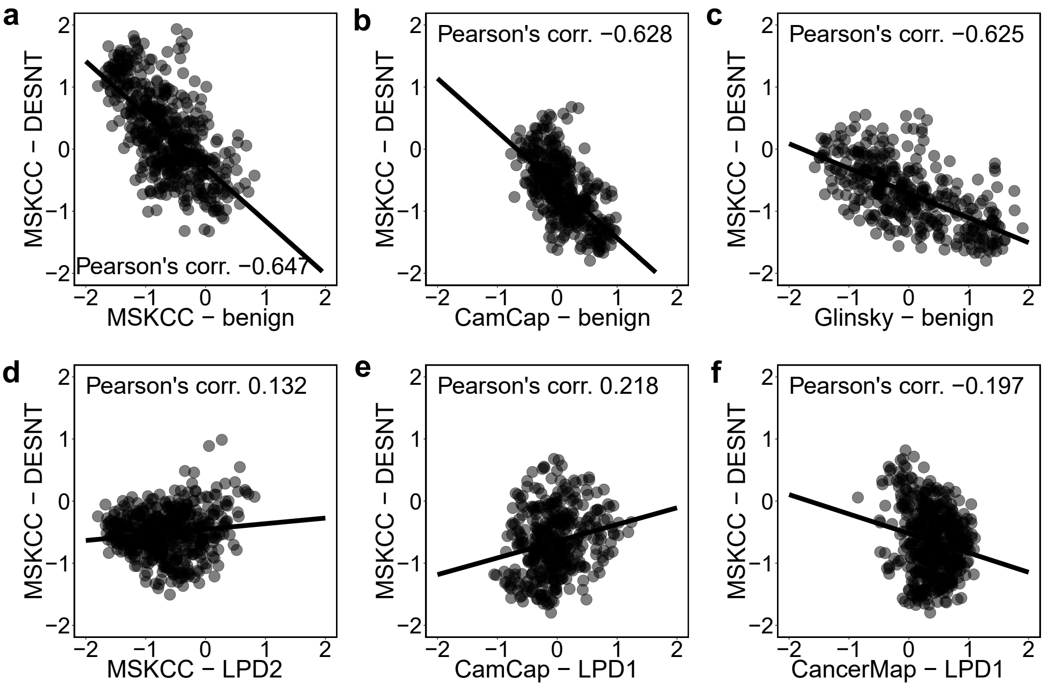

Figure B.9 Correlations of expression profiles between cancers assigned to the DESNT process in MSKCC and a) the LPD7 process in MSKCC, where all samples are benign, b) the LPD4 process in CancerMap, containing mainly normal samples, c) LPD1 process in Stephenson, where most of the benign samples are assigned, d) the LPD2 process in MSKCC, containing a mixture of risk categories, e) the CamCap LPD1 process, f) the CancerMap LPD1 process. Data from the 500 probes used in LPD are represented and ten possible comparisons are shown. The expression levels of each gene have been normalised across all samples to mean 0 and standard deviation 1

Genes	**MSKCC**	**CancerMap**	**Glinsky**	**Klein**	**CamCap**	**Reference**
ACTA2	100	92	100	98	90	
ACTG2	100	98	100	98	92	
ACTN1	100	92	100	100	67	
ATP2B4	100	92	100	100	69	
C7	100	89	100	100	74	
CALD1	100	92	92	100	40	
CDC42EP3	100	92	100	95	0	
*CLU***	100	92	100	100	0	[341]
CNN1	100	92	100	98	97	
CRISPLD2	100	92	100	98	9	
CSRP1‡*	100	93	100	100	98	[342]
*DPYSL3***	100	92	100	86	100	[343]
*EPAS1**‖	100	92	100	100	0	[344, 345]
ETS2	100	92	100	100	18	

FBLN1*†	100	92	100	100	97		[303, 304]
FERMT2	100	92	100	100	100		
FLNA	100	92	100	98	90		
GPX3*†	100	92	100	100	28		[346, 347]
GSTP1**†	100	92	100	81	91		[348]
ILK	100	92	100	100	99		
ITGA5	100	92	100	100	74		
JAM3*	92	85	100	100	96		[349]
KCNMA1**†	100	92	100	99	100		[350]
LMOD1	100	92	100	91	94		
MYL9	100	92	100	98	96		
MYLK*‡	100	92	100	98	97		[342, 351]
PALLD	100	92	100	100	98		
PCP4	100	92	100	100	80		
PDK4	100	83	100	96	0		
PDLIM1	100	91	100	81	12		
PLP2	100	92	100	100	15		
PPAP2B	100	92	100	100	5		
RBPMS	100	92	100	100	62		
SNAI2**	100	93	100	91	6		[352]
SORBS1*	100	92	100	98	96		[353]
SPG20*	100	92	100	100		NA	[354]
STAT5B	100	92	100	100	0		
STOM	100	92	100	100	80		
SVIL**	100	83	100	100	10		[355]
TGFBR3	100	92	93	87	99		
TIMP3*†	100	92	100	97	17		[356, 357]
TNS1	100	92	100	100	35		
TPM1*	100	92	100	100	1		[358, 359]
TPM2	100	92	100	80	100		
VCL	100	92	100	100	98		

Table B.3 The LPD DESNT signature. Each number corresponds to the number of LPD runs in which each gene has been found as differentially expressed. Symbols: * - down regulation by CpG methylation in cancer, ** - down regulation by CpG methylation in prostate cancer, † - CpG methylation associated with poor outcome, ‡ - prostate cancer functional connectivity hub, ‖ - gene-gene interaction focus for prostate cancer,

ID	**Description**	**GeneRatio**	***p*-value**	***p*-adjust**
GO:0009611	response to wounding	19/44	$3.2 \cdot 10^{-13}$	$6.33 \cdot 10^{-10}$
GO:0003012	muscle system process	13/44	$9.3 \cdot 10^{-13}$	$9.2 \cdot 10^{-10}$
GO:0006936	muscle contraction	12/44	$2.2 \cdot 10^{-12}$	$1.45 \cdot 10^{-9}$
GO:0042060	wound healing	15/44	$4.16 \cdot 10^{-11}$	$2.06 \cdot 10^{-8}$
GO:0030029	actin filament-based process	13/44	$5.31 \cdot 10^{-10}$	$1.92 \cdot 10^{-7}$
GO:0009653	anatomical structure morphogenesis	24/44	$5.81 \cdot 10^{-10}$	$1.92 \cdot 10^{-7}$
GO:0048856	anatomical structure development	31/44	$2.04 \cdot 10^{-9}$	$5.43 \cdot 10^{-7}$
GO:0034329	cell junction assembly	9/44	$2.33 \cdot 10^{-9}$	$5.43 \cdot 10^{-7}$
GO:0030036	actin cytoskeleton organization	12/44	$2.47 \cdot 10^{-9}$	$5.43 \cdot 10^{-7}$
GO:0044707	single-multicellular organism process	35/44	$3.36 \cdot 10^{-9}$	$6.53 \cdot 10^{-7}$
GO:0007596	blood coagulation	12/44	$3.88 \cdot 10^{-9}$	$6.53 \cdot 10^{-7}$
GO:0007599	hemostasis	12/44	$4.29 \cdot 10^{-9}$	$6.53 \cdot 10^{-7}$
GO:0050817	coagulation	12/44	$4.29 \cdot 10^{-9}$	$6.53 \cdot 10^{-7}$
GO:0050878	regulation of body fluid levels	13/44	$5.24 \cdot 10^{-9}$	$7.41 \cdot 10^{-7}$
GO:0034330	cell junction organization	9/44	$6.24 \cdot 10^{-9}$	$8.24 \cdot 10^{-7}$
GO:0032501	multicellular organismal process	35/44	$1.01 \cdot 10^{-8}$	$1.23 \cdot 10^{-6}$
GO:0048468	cell development	20/44	$1.05 \cdot 10^{-8}$	$1.23 \cdot 10^{-6}$
GO:0032989	cellular component morphogenesis	17/44	$1.87 \cdot 10^{-8}$	$2.06 \cdot 10^{-6}$
GO:0003008	system process	19/44	$2.29 \cdot 10^{-8}$	$2.39 \cdot 10^{-6}$
GO:0031589	cell-substrate adhesion	9/44	$2.65 \cdot 10^{-8}$	$2.63 \cdot 10^{-6}$

Table B.4 Top 20 GO pathways over-represented in the LPD DESNT signature.

ID	**Description**	**Gene Ratio**	***p*-value**	***p*-adjust**
hsa04270	Vascular smooth muscle contraction	6/26	$3.86 \cdot 10^{-6}$	$1.99 \cdot 10^{-4}$
hsa04510	Focal adhesion	7/26	$6.13 \cdot 10^{-6}$	$1.99 \cdot 10^{-4}$
hsa04520	Adherens junction	4/26	$1.43 \cdot 10^{-4}$	$3.11 \cdot 10^{-3}$
hsa04670	Leukocyte transendothelial migration	4/26	$8.56 \cdot 10^{-4}$	$1.28 \cdot 10^{-2}$
hsa04810	Regulation of actin cytoskeleton	5/26	$9.85 \cdot 10^{-4}$	$1.28 \cdot 10^{-2}$
hsa05100	Bacterial invasion of epithelial cells	3/26	$2.86 \cdot 10^{-3}$	$2.78 \cdot 10^{-2}$
hsa04022	cGMP-PKG signaling pathway	4/26	$3.08 \cdot 10^{-3}$	$2.78 \cdot 10^{-2}$
hsa05410	Hypertrophic cardiomyopathy (HCM)	3/26	$3.42 \cdot 10^{-3}$	$2.78 \cdot 10^{-2}$
hsa05414	Dilated cardiomyopathy	3/26	$4.3 \cdot 10^{-3}$	$3.1 \cdot 10^{-2}$

Table B.5 KEGG pathways over-represented in the LPD DESNT signature.

ID	Description	Gene Ratio	p-value	p-adjust
445355	Smooth Muscle Contraction	10/28	$4.67 \cdot 10^{-18}$	$4.2 \cdot 10^{-16}$
397014	Muscle contraction	10/28	$3.24 \cdot 10^{-14}$	$1.46 \cdot 10^{-12}$
446353	Cell-extracellular matrix interactions	4/28	$8.8 \cdot 10^{-7}$	$2.64 \cdot 10^{-5}$
5627123	RHO GTPases activate PAKs	3/28	$1.07 \cdot 10^{-4}$	$2.41 \cdot 10^{-3}$
446728	Cell junction organization	4/28	$2.6 \cdot 10^{-4}$	$4.46 \cdot 10^{-3}$
114608	Platelet degranulation	4/28	$3.16 \cdot 10^{-4}$	$4.46 \cdot 10^{-3}$
109582	Hemostasis	8/28	$3.47 \cdot 10^{-4}$	$4.46 \cdot 10^{-3}$
76005	Response to elevated platelet cytosolic Ca2+	4/28	$3.98 \cdot 10^{-4}$	$4.48 \cdot 10^{-3}$
1500931	Cell-Cell communication	4/28	$1.63 \cdot 10^{-3}$	$1.63 \cdot 10^{-2}$

Table B.6 Reactome pathways over-represented in the LPD DESNT signature.

Cytoskeleton		
ACTA2	*FLNA*	*PDLIM1*
ACTG2	*LMOD1*	*SPG20*
ACTN1	*MYL9*	*SVIL*
CALD1	*MYLK*	*TNS1*
CDC42EP3	*PALLD*	*TPM1*
CNN1	*PCP4*	*TPM2*
Adhesion, integrins and extracellular matrix		
DPYSL3	*ILK*	*TIMP3*
FBLN1	*ITGA5*	*VCL*
FERMT2	*TGFBR3*	
Transcription factors and translation regulators		
EPAS1	*RBPMS*	*STAT5B*
ETS2	*SNAI2*	
Ion channel related		
ATP2B4	*KCNMA1*	*STOM*
Mixed		
C7	*GPX3*	PLP2
CLU	*GSTP1*	*SORBS1*
CRISPLD2	*JAM3*	*PPAP2B*
CSRP1	*PDK4*	

Table B.7 The pathway involvement of the genes in the LPD DESNT signature as determined by Prof. Dylan Edwards.

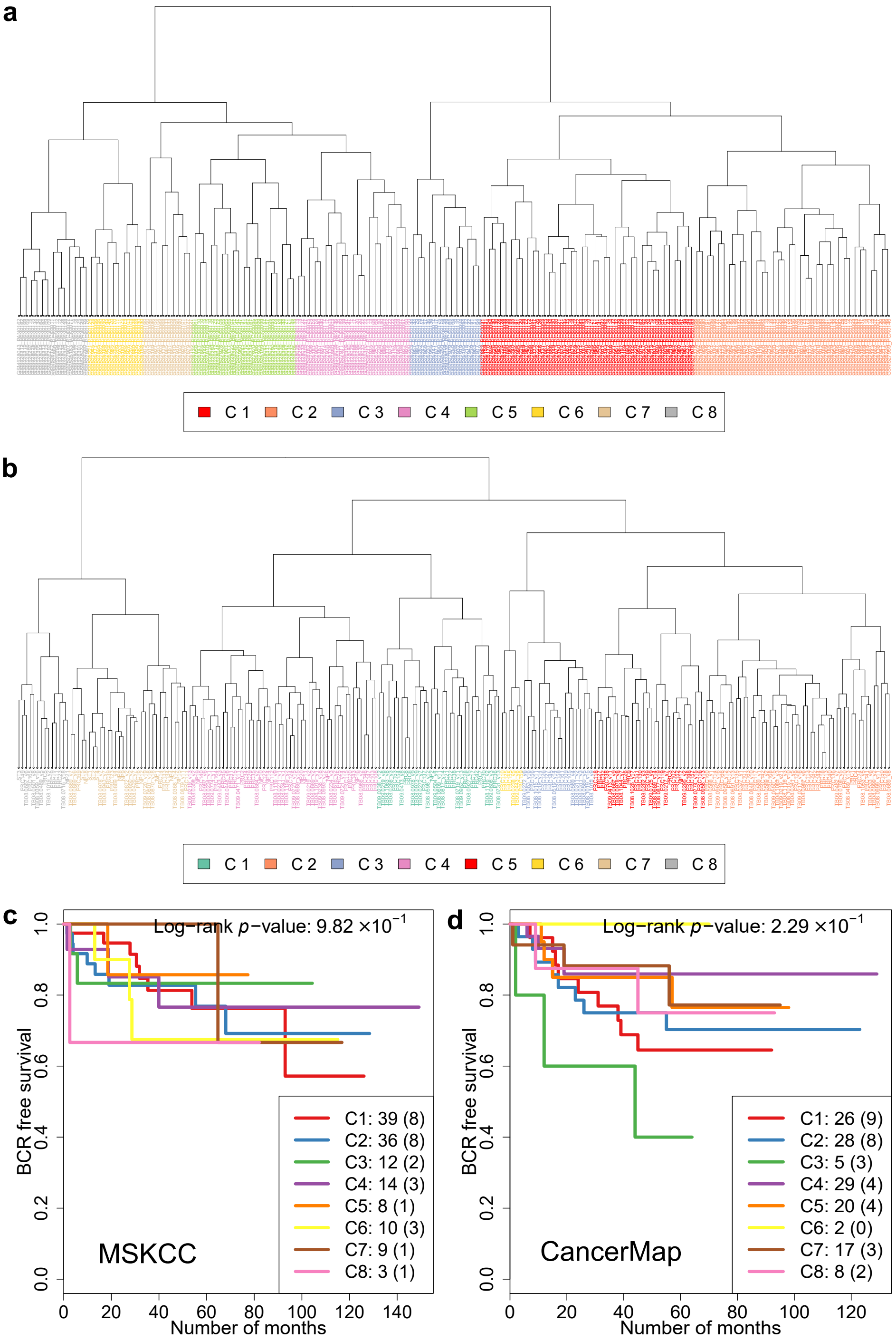

Figure B.10 Hierarchical clustering on the MSKCC (a) and CancerMap (b) datasets and the associated KM plots (c,d). In the top two panels, the colours represent the groups resulted from cutting the dendrogram into 8 groups. The KM plots correspond to the groups depicted in a) and b).

MSKCC

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	1.2915	0.49035	3.401	$6.05 \cdot 10^{-1}$
Gleason: ≤/> 7	5.7754	2.26426	14.731	$2.42 \cdot 10^{-4}$
PSA: ≤/> 10	1.2323	0.52448	2.896	$6.32 \cdot 10^{-1}$
Stage < 35 weeks: T2/T3-T4	5.6326	1.77791	17.845	$3.3 \cdot 10^{-3}$
Stage ≥ 35 weeks: T2/T3-T4	0.5174	0.06022	4.445	$5.48 \cdot 10^{-1}$

CancerMap

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	2.5107	1.1998	5.254	$1.45 \cdot 10^{-2}$
Gleason: ≤/> 7	5.3242	2.0583	13.773	$5.63 \cdot 10^{-4}$
Stage: T1-T2/T3-T4	1.8066	0.8611	3.79	$1.18 \cdot 10^{-1}$
PSA: ≤/> 10	1.1959	0.5389	2.654	$6.6 \cdot 10^{-1}$

Stephenson

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	3.3779	1.7102	6.672	$4.56 \cdot 10^{-4}$
Stage: T1-T2/T3-T4	1.5375	0.7791	3.034	$2.15 \cdot 10^{-1}$
Gleason: ≤/> 7	2.2833	1.1183	4.662	$2.34 \cdot 10^{-2}$
PSA: ≤/> 10	1.842	0.931	3.645	$7.94 \cdot 10^{-2}$

CamCap

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	2.8712	1.672	4.93	$1.31 \cdot 10^{-4}$
Stage: T1-T2/T3-T4	2.0584	1.162	3.648	$1.34 \cdot 10^{-2}$
Gleason: ≤/> 7	4.7835	2.587	8.845	$6.03 \cdot 10^{-7}$
PSA: ≤/> 10	1.9649	1.162	3.323	$1.17 \cdot 10^{-2}$

TCGA

Covariate	Hazard Ratio	CI lower 0.95	CI upper 0.95	P-value
Non-DESNT/DESNT	2.1145	1.0944	4.086	$2.59 \cdot 10^{-2}$
Gleason: ≤/> 3+4	1.3959	0.6379	3.055	$4.04 \cdot 10^{-1}$
Stage: T1-T2/T3-T4	3.8147	1.3074	11.131	$1.43 \cdot 10^{-2}$

Table B.8 Summary of the extended Cox PH model using the RF DESNT membership, Gleason score ($\leq$ / $>$ 7), PSA ($\leq$ / $>$ 10) and stage (T1-T2/T3-T4) as predictors of recurrence.

Probe ID	Gene Symbol	Dist TSS	Promoter	*P*-value	*P*-adjust
cg00813162	*ACTN1*	920	FALSE	$5.45 \cdot 10^{-35}$	$6.12 \cdot 10^{-33}$

cg27173151	*ATP2B4*	2415	FALSE	$2.28 \cdot 10^{-16}$	$1.6 \cdot 10^{-15}$
cg04148762	*ATP2B4*	2658	FALSE	$5.46 \cdot 10^{-14}$	$3 \cdot 10^{-13}$
cg17721249	*ATP2B4*	2679	FALSE	$4.06 \cdot 10^{-12}$	$1.88 \cdot 10^{-11}$
cg26253438	*ATP2B4*	2708	FALSE	$4.06 \cdot 10^{-12}$	$1.88 \cdot 10^{-11}$
cg15732851	*ATP2B4*	2846	FALSE	$1.13 \cdot 10^{-13}$	$6.1 \cdot 10^{-13}$
cg21397588	*ATP2B4*	3039	FALSE	$2.53 \cdot 10^{-18}$	$2.1 \cdot 10^{-17}$
cg02457623	*ATP2B4*	3174	FALSE	$7.48 \cdot 10^{-25}$	$1.23 \cdot 10^{-23}$
cg13808058	*ATP2B4*	9675	FALSE	$1.34 \cdot 10^{-26}$	$2.7 \cdot 10^{-25}$
cg13115222	*ATP2B4*	19093	FALSE	$1.44 \cdot 10^{-39}$	$4.05 \cdot 10^{-37}$
cg00319312	*ATP2B4*	19270	FALSE	$1.44 \cdot 10^{-39}$	$4.05 \cdot 10^{-37}$
cg22439651	*C7*	23845	FALSE	$3.7 \cdot 10^{-10}$	$1.51 \cdot 10^{-9}$
cg16453474	*C7*	72493	FALSE	$3.66 \cdot 10^{-9}$	$1.39 \cdot 10^{-8}$
cg06390484	*CALD1*	111107	FALSE	$1.83 \cdot 10^{-20}$	$1.95 \cdot 10^{-19}$
cg08698854	*CALD1*	111142	FALSE	$7.43 \cdot 10^{-22}$	$8.5 \cdot 10^{-21}$
cg07625383	*CALD1*	8344	FALSE	$1.31 \cdot 10^{-15}$	$8.7 \cdot 10^{-15}$
cg22313574	*CLU*	286	FALSE	$7.22 \cdot 10^{-28}$	$1.84 \cdot 10^{-26}$
cg14917244	*CLU*	266	FALSE	$7.22 \cdot 10^{-28}$	$1.84 \cdot 10^{-26}$
cg03296797	*CRISPLD2*	16479	FALSE	$1.14 \cdot 10^{-25}$	$2 \cdot 10^{-24}$
cg03476673	*CRISPLD2*	16616	FALSE	$1.14 \cdot 10^{-25}$	$2 \cdot 10^{-24}$
cg09967633	*CRISPLD2*	65207	FALSE	$7.28 \cdot 10^{-18}$	$5.79 \cdot 10^{-17}$
cg02455706	*CRISPLD2*	65264	FALSE	$7.28 \cdot 10^{-18}$	$5.79 \cdot 10^{-17}$
cg20623601	*EPAS1*	2302	FALSE	$7.5 \cdot 10^{-10}$	$3.02 \cdot 10^{-9}$
cg25124739	*EPAS1*	2557	FALSE	$1.04 \cdot 10^{-8}$	$3.75 \cdot 10^{-8}$
cg07072704	*FBLN1*	1017	FALSE	$6.75 \cdot 10^{-35}$	$6.89 \cdot 10^{-33}$
cg23497752	*FLNA*	4928	FALSE	$1.28 \cdot 10^{-10}$	$5.39 \cdot 10^{-10}$
cg02659086	*GSTP1*	-89	TRUE	$1.65 \cdot 10^{-14}$	$9.79 \cdot 10^{-14}$
cg06928838	*GSTP1*	424	FALSE	$6.9 \cdot 10^{-25}$	$1.15 \cdot 10^{-23}$
cg09038676	*GSTP1*	542	FALSE	$1.59 \cdot 10^{-28}$	$4.68 \cdot 10^{-27}$
cg11566244	*GSTP1*	720	FALSE	$6.3 \cdot 10^{-28}$	$1.68 \cdot 10^{-26}$
cg22224704	*GSTP1*	975	FALSE	$1.03 \cdot 10^{-36}$	$1.92 \cdot 10^{-34}$
cg23795217	*ITGA5*	1049	FALSE	$4.27 \cdot 10^{-27}$	$0.99 \cdot 10^{-25}$
cg03826594	*ITGA5*	964	FALSE	$2.61 \cdot 10^{-26}$	$5.05 \cdot 10^{-25}$
cg03640071	*JAM3*	81930	FALSE	$4.58 \cdot 10^{-28}$	$1.29 \cdot 10^{-26}$
cg16055185	*KCNMA1*	247059	FALSE	$3.38 \cdot 10^{-22}$	$4.17 \cdot 10^{-21}$
cg01858517	*KCNMA1*	992	FALSE	$1.08 \cdot 10^{-12}$	$5.32 \cdot 10^{-12}$
cg05479582	*KCNMA1*	952	FALSE	$1.08 \cdot 10^{-12}$	$5.32 \cdot 10^{-12}$
cg03354113	*KCNMA1*	783	FALSE	$1.08 \cdot 10^{-12}$	$5.32 \cdot 10^{-12}$
cg18660345	*MYL9*	-347	TRUE	$6.9 \cdot 10^{-31}$	$3.09 \cdot 10^{-29}$
cg05820491	*MYL9*	-292	TRUE	$6.9 \cdot 10^{-31}$	$3.09 \cdot 10^{-29}$
cg20669834	*MYLK*	112	FALSE	$9.65 \cdot 10^{-23}$	$1.27 \cdot 10^{-21}$
cg18731398	*MYLK*	5622	FALSE	$1.68 \cdot 10^{-32}$	$1.18 \cdot 10^{-30}$
cg07621385	*MYLK*	14613	FALSE	$1.31 \cdot 10^{-33}$	$1.23 \cdot 10^{-31}$
cg04376312	*MYLK*	663	FALSE	$3.14 \cdot 10^{-32}$	$2.07 \cdot 10^{-30}$
cg24242290	*PALLD*	112017	FALSE	$2.54 \cdot 10^{-26}$	$4.99 \cdot 10^{-25}$

cg12768523	*PALLD*	184456	FALSE	$1.02 \cdot 10^{-26}$	$2.12 \cdot 10^{-25}$
cg25054754	*PALLD*	1172	FALSE	$0.99 \cdot 10^{-10}$	$4.27 \cdot 10^{-10}$
cg13573928	*PALLD*	1378	FALSE	$0.99 \cdot 10^{-10}$	$4.27 \cdot 10^{-10}$
cg15948536	*PALLD*	16936	FALSE	$2.81 \cdot 10^{-24}$	$4.37 \cdot 10^{-23}$
cg18389639	*PDLIM1*	1294	FALSE	$2.64 \cdot 10^{-29}$	$8.99 \cdot 10^{-28}$
cg11682697	*PPAP2B*	52884	FALSE	$1.71 \cdot 10^{-28}$	$4.92 \cdot 10^{-27}$
cg17221758	*RBPMS*	1297	FALSE	$2.72 \cdot 10^{-15}$	$1.74 \cdot 10^{-14}$
cg00447833	*RBPMS*	1316	FALSE	$2.72 \cdot 10^{-15}$	$1.74 \cdot 10^{-14}$
cg13456241	*RBPMS*	12979	FALSE	$8.21 \cdot 10^{-29}$	$2.56 \cdot 10^{-27}$
cg12167489	*RBPMS*	48545	FALSE	$2.39 \cdot 10^{-11}$	$1.06 \cdot 10^{-10}$
cg24128292	*RBPMS*	84620	FALSE	$6.82 \cdot 10^{-32}$	$4.25 \cdot 10^{-30}$
cg18318006	*SORBS1*	12682	FALSE	$3.1 \cdot 10^{-23}$	$4.46 \cdot 10^{-22}$
cg10741308	*SORBS1*	6557	FALSE	$1.75 \cdot 10^{-33}$	$1.31 \cdot 10^{-31}$
cg02370232	*SORBS1*	6464	FALSE	$1.75 \cdot 10^{-33}$	$1.31 \cdot 10^{-31}$
cg06282596	*SORBS1*	6350	FALSE	$1.75 \cdot 10^{-33}$	$1.31 \cdot 10^{-31}$
cg09072216	*SPG20*	1301	FALSE	$8.29 \cdot 10^{-18}$	$6.41 \cdot 10^{-17}$
cg10558887	*SPG20*	1236	FALSE	$8.29 \cdot 10^{-18}$	$6.41 \cdot 10^{-17}$
cg20691205	*SPG20*	971	FALSE	$1.35 \cdot 10^{-14}$	$8.16 \cdot 10^{-14}$
cg01404317	*SPG20*	826	FALSE	$2.38 \cdot 10^{-14}$	$1.36 \cdot 10^{-13}$
cg18755783	*SPG20*	740	FALSE	$2.38 \cdot 10^{-14}$	$1.36 \cdot 10^{-13}$
cg00947032	*SPG20*	685	FALSE	$1.7 \cdot 10^{-14}$	$1.01 \cdot 10^{-13}$
cg00049475	*SVIL*	164	FALSE	$7.52 \cdot 10^{-18}$	$5.9 \cdot 10^{-17}$
cg04678141	*SVIL*	100809	FALSE	$7.52 \cdot 10^{-18}$	$5.9 \cdot 10^{-17}$
cg03241461	*SVIL*	100471	FALSE	$1.11 \cdot 10^{-29}$	$4.22 \cdot 10^{-28}$
cg24212268	*SVIL*	88580	FALSE	$1.45 \cdot 10^{-20}$	$1.56 \cdot 10^{-19}$
cg06197966	*SVIL*	76405	FALSE	$2.48 \cdot 10^{-35}$	$3.09 \cdot 10^{-33}$
cg13324103	*SVIL*	76301	FALSE	$2.48 \cdot 10^{-35}$	$3.09 \cdot 10^{-33}$
cg09287650	*SVIL*	43513	FALSE	$4.12 \cdot 10^{-21}$	$4.57 \cdot 10^{-20}$
cg23721586	*TGFBR3*	130071	FALSE	$1.56 \cdot 10^{-12}$	$7.59 \cdot 10^{-12}$
cg25769732	*TGFBR3*	31656	FALSE	$3.72 \cdot 10^{-16}$	$2.56 \cdot 10^{-15}$
cg03323067	*TNS1*	880	FALSE	$9.13 \cdot 10^{-16}$	$6.24 \cdot 10^{-15}$
cg11936410	*TPM1*	4488	FALSE	$1.25 \cdot 10^{-12}$	$6.11 \cdot 10^{-12}$

Table B.9 Differentially methylated probes between RF DESNT and RF non-DESNT. The third column contains the distance (in base pairs) of the probe from the transcription start site. The fourth column indicates if the probes is within the region of the gene promoter.

Gene	χ^2 ***p*-val**	**FDR adj.** χ^2 ***p*-value**
PTEN	$3.14 \cdot 10^{-1}$	1
SPOP	$8.75 \cdot 10^{-1}$	1
TP53	$5.43 \cdot 10^{-2}$	1
ATM	$8.44 \cdot 10^{-1}$	1
CHD1	1	1
FOXA1	$1.31 \cdot 10^{-1}$	1
KMT2C	$3.94 \cdot 10^{-1}$	1
PIK3CA	$9.83 \cdot 10^{-1}$	1
BRAF	$2.22 \cdot 10^{-1}$	1
KMT2D	$3.9 \cdot 10^{-1}$	1
BRCA2	$4.44 \cdot 10^{-1}$	1
CTNNB1	$6.6 \cdot 10^{-1}$	1
MED12	1	1
ZMYM3	$9.83 \cdot 10^{-1}$	1
AKT1	1	1
CDK12	1	1
CDKN1B	$5.48 \cdot 10^{-1}$	1
HRAS	1	1
IDH1	$7.48 \cdot 10^{-1}$	1
RB1	1	1

Table B.10 Correlations between mutations and RF DESNT membership.

Gene	χ^2 ***p*-val**	**FDR adj.** χ^2 ***p*-value**
PTEN	$2.56 \cdot 10^{-1}$	$4.58 \cdot 10^{-1}$
TP53	$9.72 \cdot 10^{-2}$	$2.19 \cdot 10^{-1}$
CHD1	$8.15 \cdot 10^{-2}$	$2.19 \cdot 10^{-1}$
BRCA2	$9.72 \cdot 10^{-2}$	$2.19 \cdot 10^{-1}$
CDKN1B	1	1
RB1	$5.55 \cdot 10^{-1}$	$7.14 \cdot 10^{-1}$
CDK12	$9.72 \cdot 10^{-2}$	$2.19 \cdot 10^{-1}$
FANCD2	1	1
SPOPL	$3.05 \cdot 10^{-1}$	$4.58 \cdot 10^{-1}$

Table B.11 Correlations between homozygous deletions and RF DESNT membership.

Gene	χ^2 ***p*-val**	**FDR adj. χ^2 *p*-value**
SPOP	$8.75 \cdot 10^{-1}$	1
ATM	$8.44 \cdot 10^{-1}$	1
FOXA1	$1.31 \cdot 10^{-1}$	$4.9 \cdot 10^{-1}$
KMT2C	$3.94 \cdot 10^{-1}$	$7.87 \cdot 10^{-1}$
PIK3CA	$9.83 \cdot 10^{-1}$	1
BRAF	$2.22 \cdot 10^{-1}$	$6.11 \cdot 10^{-1}$
KMT2D	$3.9 \cdot 10^{-1}$	$7.87 \cdot 10^{-1}$
CTNNB1	$6.6 \cdot 10^{-1}$	1
MED12	1	1
ZMYM3	$9.83 \cdot 10^{-1}$	1
AKT1	1	1
HRAS	1	1
IDH1	$7.48 \cdot 10^{-1}$	1
PTEN	$1.42 \cdot 10^{-1}$	$4.9 \cdot 10^{-1}$
TP53	$\mathbf{3.84 \cdot 10^{-3}}$	$8.45 \cdot 10^{-2}$
CHD1	$9.14 \cdot 10^{-2}$	$4.9 \cdot 10^{-1}$
BRCA2	$\mathbf{2.07 \cdot 10^{-2}}$	$2.28 \cdot 10^{-1}$
CDK12	$\mathbf{4.19 \cdot 10^{-2}}$	$3.07 \cdot 10^{-1}$
CDKN1B	$4.95 \cdot 10^{-1}$	$9.08 \cdot 10^{-1}$
RB1	$1.56 \cdot 10^{-1}$	$4.9 \cdot 10^{-1}$
FANCD2	1	1
SPOPL	$3.05 \cdot 10^{-1}$	$7.47 \cdot 10^{-1}$

Table B.12 Correlations between the combined mutation/homozygous status and RF DESNT membership.

Chr.	Start	End	Nr. Pb.	Gene	Dist TSS	Promoter
chr1	56992372	56992372	1	*PPAP2B*	52885	FALSE
chr1	92197531	92197531	1	*TGFBR3*	130072	FALSE
chr1	92295946	92295946	1	*TGFBR3*	31657	FALSE
chr1	203598330	203599089	7	*ATP2B4*	2415	FALSE
chr1	203605590	203605590	1	*ATP2B4*	9675	FALSE
chr1	203670963	203671140	2	*ATP2B4*	19093	FALSE
chr10	29923736	29924258	3	*SVIL*	0	TRUE
chr10	29936149	29948428	3	*SVIL*	76302	FALSE
chr10	29981216	29981216	1	*SVIL*	43514	FALSE
chr10	79150517	79150517	1	*KCNMA1*	247060	FALSE
chr10	79396584	79396793	3	*KCNMA1*	784	FALSE
chr10	97049610	97049610	1	*PDLIM1*	1295	FALSE
chr10	97169147	97175479	4	*SORBS1*	6351	FALSE
chr11	67350976	67350976	1	*GSTP1*	-90	TRUE
chr11	67351271	67352041	6	*GSTP1*	205	FALSE
chr11	134020750	134020750	1	*JAM3*	81930	FALSE
chr12	54811762	54812085	3	*ITGA5*	965	FALSE
chr13	36919344	36919960	6	*SPG20*	686	FALSE
chr14	69443362	69443362	1	*ACTN1*	921	FALSE
chr15	63345124	63345124	1	*TPM1*	4488	FALSE
chr16	84870066	84870203	2	*CRISPLD2*	16479	FALSE
chr16	84918794	84918851	2	*CRISPLD2*	65207	FALSE
chr2	46526843	46527098	2	*EPAS1*	2302	FALSE
chr2	218767655	218767655	1	*TNS1*	881	FALSE
chr20	35169380	35169594	3	*MYL9*	-293	TRUE
chr22	45899736	45899736	1	*FBLN1*	1017	FALSE
chr3	123339417	123339568	2	*MYLK*	0	TRUE
chr3	123414733	123414733	1	*MYLK*	5623	FALSE
chr3	123535716	123535716	1	*MYLK*	14614	FALSE
chr3	123602485	123602485	1	*MYLK*	664	FALSE
chr4	169664785	169664785	1	*PALLD*	112017	FALSE
chr4	169737224	169737224	1	*PALLD*	184456	FALSE
chr4	169754328	169754534	2	*PALLD*	1172	FALSE
chr4	169770092	169770092	1	*PALLD*	16936	FALSE
chr5	40933444	40982092	2	*C7*	23845	FALSE
chr7	134575145	134575524	5	*CALD1*	110981	FALSE
chr7	134626083	134626083	1	*CALD1*	8344	FALSE
chr8	27468981	27469186	3	*CLU*	82	FALSE
chr8	30243241	30243260	2	*RBPMS*	1297	FALSE
chr8	30254923	30254923	1	*RBPMS*	12979	FALSE
chr8	30290489	30290489	1	*RBPMS*	48545	FALSE
chr8	30419935	30419935	1	*RBPMS*	84620	FALSE
chrX	153598077	153598077	1	*FLNA*	4929	FALSE

Table B.13 Differentially methylated regions between RF DESNT and RF non-DESNT. The numbers in the third column represent the number of probes in the DMR. The fifth column contains the distance (in base pairs) of the probe from the transcription start site. The sixth column indicates if the probes is within the region of the promoter.

Gene	Probe	Methylation	Correlation
ACTN1	cg00813162	hyper	-0.65
ATP2B4	cg00319312	hypo	0.75
ATP2B4	cg02457623	hyper	-0.61
ATP2B4	cg04148762	hyper	-0.38
ATP2B4	cg13115222	hypo	0.70
ATP2B4	cg13808058	hyper	-0.58
ATP2B4	cg15732851	hyper	-0.38
ATP2B4	cg17721249	hyper	-0.40
ATP2B4	cg21397588	hyper	-0.56
ATP2B4	cg26253438	hyper	-0.49
ATP2B4	cg27173151	hyper	-0.57
C7	cg16453474	hypo	0.21
C7	cg22439651	hypo	0.28
CALD1	cg06390484	hyper	-0.47
CALD1	cg07625383	hyper	-0.52
CALD1	cg08698854	hyper	-0.49
CLU	cg14917244	hyper	-0.57
CLU	cg22313574	hyper	-0.58
CRISPLD2	cg02455706	hyper	-0.43
CRISPLD2	cg03296797	hyper	-0.52
CRISPLD2	cg03476673	hyper	-0.51
CRISPLD2	cg09967633	hyper	-0.47
EPAS1	cg20623601	hyper	-0.40
EPAS1	cg25124739	hyper	-0.36
FBLN1	cg07072704	hyper	-0.47
FLNA	cg23497752	hyper	-0.46
GSTP1	cg02659086	hyper	-0.72
GSTP1	cg06928838	hyper	-0.84
GSTP1	cg09038676	hyper	-0.84
GSTP1	cg11566244	hyper	-0.82
GSTP1	cg22224704	hyper	-0.87
ITGA5	cg03826594	hyper	-0.44
ITGA5	cg23795217	hyper	-0.50
JAM3	cg03640071	hyper	-0.67
KCNMA1	cg01858517	hyper	-0.54
KCNMA1	cg03354113	hyper	-0.42
KCNMA1	cg05479582	hyper	-0.53
KCNMA1	cg16055185	hyper	-0.53
MYL9	cg05820491	hyper	-0.50
MYL9	cg18660345	hyper	-0.58
MYLK	cg00465319	hyper	-0.63
MYLK	cg04376312	hyper	-0.65

MYLK	cg07621385	hyper	-0.68
MYLK	cg18731398	hyper	-0.67
MYLK	cg20669834	hyper	-0.50
PALLD	cg12768523	hyper	-0.61
PALLD	cg13573928	hyper	-0.55
PALLD	cg15948536	hyper	-0.61
PALLD	cg24242290	hypo	0.53
PALLD	cg25054754	hyper	-0.47
PDLIM1	cg18389639	hyper	-0.75
PPAP2B	cg11682697	hyper	-0.52
RBPMS	cg00447833	hyper	-0.55
RBPMS	cg12167489	hyper	-0.52
RBPMS	cg13456241	hyper	-0.61
RBPMS	cg17221758	hyper	-0.55
RBPMS	cg24128292	hyper	-0.67
SORBS1	cg02370232	hyper	-0.58
SORBS1	cg06282596	hyper	-0.55
SORBS1	cg10741308	hyper	-0.50
SORBS1	cg18318006	hyper	-0.49
SPG20	cg00947032	hyper	-0.70
SPG20	cg01404317	hyper	-0.72
SPG20	cg09072216	hyper	-0.76
SPG20	cg10558887	hyper	-0.74
SPG20	cg18755783	hyper	-0.74
SPG20	cg20691205	hyper	-0.74
SVIL	cg00049475	hyper	-0.54
SVIL	cg03241461	hyper	-0.57
SVIL	cg04678141	hyper	-0.51
SVIL	cg06197966	hyper	-0.62
SVIL	cg09287650	hyper	-0.52
SVIL	cg13324103	hyper	-0.61
SVIL	cg24212268	hyper	-0.59
TGFBR3	cg23721586	hyper	-0.37
TGFBR3	cg25769732	hyper	-0.48
TNS1	cg03323067	hyper	-0.50
TPM1	cg11936410	hypo	0.26

Table B.14 Pearson's correlation between the beta values of each differentially methylated probe mapping to the 45 genes in the LPD DESNT signature and the expression levels of the corresponding gene, across samples.