clear cell Renal Cell Carcinoma:
Biomarkers and Networks

Sakshi Gulati

University College London
and
Cancer Research UK London Research Institute
PhD Supervisor: Dr Paul A Bates

A thesis submitted for the degree of
Doctor of Philosophy
University College London
August 2015

Declaration

I Sakshi Gulati confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

clear cell Renal Cell Carcinoma (ccRCC) is the most prevalent subtype of Kidney Cancer and is the eighth most common cancer in the United Kingdom. Historically, the disease has been characterised by the biallelic loss of VHL gene and loss-of-heterozygosity of chromosome 3p. Inactivation of the VHL gene leads to constitutive up-regulation of the HIF family of transcription factors, thereby leading to a hypoxia response transcription signature. Recent studies have led to the identification of recurrent mutations in genes involved in chromatin remodelling and histone methylation. Increasing evidence has also been presented to show genetic intratumour heterogeneity (ITH) for this disease. These observations have led to important questions regarding disease etiology and the impact of ITH on disease biology as well as prognosis. This thesis investigates high throughput "omics" datasets and a comprehensive integrative analysis is performed of the genetic changes and transcriptome expression levels for ccRCC. Computational methods ranging from survival statistics, analyses of co-alteration and mutual exclusivity patterns for genetic alterations, gene expression analyses, to network algorithms have been used as part of this work to elucidate both ccRCC biology and pathology. To validate biomarkers, which could provide independent prognostic information in the clinic, published ccRCC prognostic biomarkers are investigated in an independent patient cohort published by the Cancer Genome Atlas (TCGA). The ITH of the most promising marker is then investigated in a multiregion biopsy dataset to guide biomarker optimisation. Furthermore, the functional consequences of cancer gene mutations as well as copy number events are interpreted by integrating them with gene expression data and by employing state of the art computational network algorithms.

Acknowledgements

A very wise friend once said to me, that by the time you reach this (submission) point, you will want to thank each and every person in this world. Truer words have never been said. This space will not be enough to thank all the people who have helped me achieve my goals and be the person I am today, however please do know that I am thinking of you.

To start with, I would like to thank my examiners Prof Christine Orengo and Dr Pietro Lio̓ for agreeing to review this thesis, and allowing me to defend it.

I would also like to thank Cancer Research UK LRI and University College London for funding my PhD as well as providing an amazing environment for my research. I would like to extend deep gratitude to the endless supporters, who donate to CRUK and enable it to be the excellent research institute that it is. All members of the academic staff especially Dr Sally Leevers, Sabina Ebbols and Andrew Brown for their constant guidance and help no matter how little my issues. My thesis committee, Prof Charles Swanton and Dr Neil McDonald for their input and insightful perspective through the years. For inspiring me to do better and encouraging me in each of our encounters. Charlie for being an amazing second supervisor and giving me the necessary push through the good and the bad times.

I also want to express my gratitude to my supervisor, Dr Paul Bates, for giving me the opportunity to undertake this Ph.D. I am indebted to him for his support, our long discussions, and his openness in letting me pursue my mind. For his understanding through my mistakes and for his patience whenever I panicked. I would also like to thank Dr Marco Gerlinger, who has been a fundamental part of my Ph.D. I would like to thank him for introducing me to the field of biomarkers, for encouraging me to pursue newer fields and all in all helping me learn a lot. I would also like to extend my thanks to all other members of Prof Swanton's laboratory, especially Dr Pierre Martinez and Dr Samra Turajlic for being great collaborators.

I can't thank enough all members of the Biomolecular modelling laboratory, for not only accepting me into their folds, but also enabling me to achieve my potential both professionally and personally. I truly believe that I am better person today, and my gratitude goes to all of you. Dr Raphaël Chaliel for keeping all the computing equipment up and running, for installing endless R packages, for listening to me whine about the farm and for being himself. Dr Melda Tozluoğlu for being the amazing person she is, for helping me with everything, however big or small and for being an inspiration to be a stronger, better scientist as well as a better person. Melda, I wouldn't have survived without you. Dr Tammy Cheng for introducing me to Python programming, for stimulating discussions both within and outside the lab and for being there always. Dr Rudi Agius, Dr Mieczyslaw Torchala, Dr Robert Jenkins and Erik Pfeiffenberger for their endless support and for simply being there whenever I have needed. All of you for being the amazing friends you are.

Beyond the lab, there are some other people who have been elemental in this four year journey, which includes all the students who started with me in 2011. You all have been amazing and I hope I have helped you as much as you have helped me through the years. I would especially like to thank Molly Godfrey, for being a great friend and an awesome flatmate. Marianne Burbage who is an absolute darling and a dear friend; Nil Ege, Ichha Khanal and Aldona Chmielewska, who are too much like me and for being the most amazing friends I could ask for.

I would also like to thank my Masters supervisor Dr Thomas Schlitt, who encouraged me to follow my goals and has had endless patience with my ups and downs, throughout the years, not only during my masters but also the PhD.

Finally I will like to thank my family as well as a few friends who are part of my extended family. Nupur Agarwal, Geetu Dhillon, Naveen Yadav and Siddarth Khandelwal for being amazing friends and for encouraging me and picking me up whenever I was down. Vatsala Koul, without whom, I absolutely could not and would not survive; words will never be enough to express how grateful I am to call you my friend. My loving grandparents and my uncles and aunts, for their blessings, good wishes and endless support. A special thanks to my aunt, Maneesha Puri, better known as my guardian angel. My sister for being the royal pain she is, and lastly but most importantly my mother. For pushing me to be independent and stand on my own feet, for being supportive, loving, caring but at the same time open enough to let me be my own person. In all for being the amazing mother that she is.

The same wise friend recently also said, when you focus on doing a good job, and not worry about the consequences, you actually end up doing a better job. I hope my PhD and this thesis could be counted as one of those instances.

Two years ago, I realised, that my Mayan calendar ended with my PhD. So life as I knew it is coming to an end. Let the adventure begin!

Table of Contents

Table of Figures

List of Tables

Abbreviations

Amp	Amplification
BMR	Background Mutation Rate
BPs	Biological Processes
C.I.	Confidence Interval
ccRCC	clear cell Renal Cell Carcinoma
Chrom	Chromosome
CR analysis	Competing Risk analysis
CSS	Cancer Specific Survival
DE	Differential Expression
Del	Deletion
ECM	Extracellular Matrix
FC	Fold Change
FDR	False Discovery Rate
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HIF	Hypoxia Inducible Factor
HR	Hazard Ratio
ITH	Intratumour Heterogeneity
KM	Kaplan-Meier
log FC	logarithmic Fold Change
LOH	Loss of Heterozygosity
mRCC	metastatic Renal Cell Carcinoma
MRs	Master Regulators
MVA	Multivariate Analysis
n	number of cases
n_{ccA}	number of ccA cases
n_{ccB}	number of ccB cases
NGS	Next Generation Sequencing
NMF	Non-negative Matrix Factorisation
OR	Odds Ratio
ORA	Overrepresentation Analysis
PPIN	Protein-Protein Interaction Network
PPIs	Protein-Protein Interactions
RCC	Renal Cell Carcinoma
RNA-Seq	RNA Sequencing
RSEM	RNA-Seq by Expectation-Maximization
SCNAs	Somatic Copy Number Alterations
SNP	Single Nucleotide Polymorphisms
SSIGN	Stage, Size, Grade and Necrosis
TCGA	The Cancer Genome Atlas
wGII	weighted Genomic Instability Index

List of Publications

A chronological list of publications published in peer-reviewed journals during the PhD project:

- **Gulati S.**, Turajlic S., Larkin J., Bates PA and Swanton C. (2015). Relapse Models for clear cell renal carcinoma. The Lancet Oncology 16(8), e376-e378

- Kanu N., Grönroos E., Martinez P., Burrell RA., Yi Goh X., Bartkova J., Maya-Mendoza A., Mistrík M., Rowan AJ., Patel H., Rabinowitz A., East P., Wilson G., Santos CR., McGranahan N., **Gulati S.**, Gerlinger M., Birkbak NJ., Joshi T., Alexandrov LB., Stratton MR., Powles T., Matthews N., Bates PA., Stewart A., Szallasi Z., Larkin J., Bartek J., Swanton C. (2015). SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair. **Oncogene** doi: 10.1038/onc.2015.24

- Fisher R., Horswell S., Rowan A., Salm MP., de Bruin EC., **Gulati S.**, McGranahan N., Stares M., Gerlinger M., Varela I., Crockford A., Favero F., Quidville V., André F., Navas C., Grönroos E., Nicol D., Hazell S., Hrouda D., O Brien T., Matthews N., Phillimore B., Begum S., Rabinowitz A., Biggs J., Bates PA., McDonald NQ., Stamp G., Spencer-Dene B., Hsieh JJ., Xu J., Pickering L., Gore M., Larkin J., Swanton C. (2014). Development of synchronous VHL syndrome tumors reveals contingencies and constraints to tumor evolution. **Genome Biology** 15(8), 433.

- **Gulati, S.**, Martinez, P., Joshi, T., Birkbak, N.J., Santos C.R., Rowan, A.J., Pickering, L., Gore M., Larkin, J., Szallasi, Z., Bates P.A., Swanton, C., Gerlinger, M. (2014). Systematic Evaluation of the Prognostic Impact and Intratumour Heterogeneity of Clear Cell Renal Carcinoma Biomarkers. **European Urology** 66, 936-948

- Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Mattews, N., Santos, C.R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., **Gulati, S.,** Bates, P.A., Stamp, G., Pickering, L., Gore, M., Nicol, D.L., Hazell, S., Futreal, P.A., Stewart, A. & Swanton, C. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. **Nature Genetics** 46(3), 225-233.

- **Gulati, S.**, Cheng, T.M.K. & Bates, P.A. (2013). Cancer networks and beyond: interpreting mutations using the human interactome and protein structure. **Seminars in Cancer Cell Biology** 23(4), 219-226.

- Cheng, T.M.K*, **Gulati, S.***, Agius, R.* & Bates, P.A.* (2012). Understanding Cancer Mechanisms through Network Dynamics. **Briefings in Functional Genomics** 11(6), 543-560.

All authors contributed equally

Chapter 1. Introduction

The past few decades have witnessed some of the largest advances in diagnosing, understanding and thereby treating complex diseases. With the increasing average age of the population, the rate of incidence of diseases such as Alzheimer's and cancer have risen dramatically; however, there have been corresponding improvements in patient care and treatment. Combined with this is the advent of newer technologies such as next generation sequencing (NGS), including DNA and RNA sequencing, which facilitate an improved measure of such diseases at various levels. Computational methods have further enabled analyses of this dearth of data. Through the work presented in this thesis, I have attempted to unravel the biological mechanisms underlying clear cell Renal Cell Carcinoma (ccRCC) and underpin important prognostic markers for this cancer. In this chapter, a background of ccRCC, both in terms of biology and prognosis, as well as an introduction to the analytical methods used for this study are presented.

1.1 A thesis justified

1.1.1 Cancer and its hallmarks

Cancer is a complex multifaceted disease characterised essentially by eight properties, namely: sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction. Together the above properties are commonly known as the hallmarks of cancer (Hanahan and Weinberg, Hanahan and Weinberg). Although monoclonal in origin, cancers acquire numerous genetic and epigenetic changes facilitated by a background of genomic instability (Burrell et al.). Such changes are acquired both at onset as well as throughout the development of the disease and thus lead to the evolution of the cancer cell in a manner similar to Darwinian evolution.

Decades of research have enabled the characterisation of mutations in several genes, which are central in causing and progressing cancer (‘Cancer Genes’), and the pathways through which these genes may act. However, since cancer can be considered to be a consequence of malfunctions within complex cellular systems, our understanding of the disease is by no means complete (Hornberg et al.). Through recent advances in DNA sequencing technologies, knowledge of the genetic alterations in cancer is accumulating rapidly allowing the descriptive analysis of cancer genomes at unprecedented speeds and at nucleotide resolution. This has led to the discovery of multiple new cancer genes, which are recurrently mutated (Dalgliesh et al., Varela et al., Wood et al., The Cancer Genome Atlas Research Network, International Cancer Genome et al.)

However, a typical solid tumour harbours tens to hundreds of non-synonymous somatic mutations, and it is now apparent that this mutational landscape is highly heterogeneous. This heterogeneity is characterized by only a few commonly mutated genes in each cancer type, which have been referred to as ‘mountains’ on the mutational landscape, and a much larger number of infrequently mutated genes or ‘hills’, which are only found in a small percentage of tumours of a particular type (Wood et al.). Furthermore, there are multiple levels of heterogeneity; inter-patient, intra-patient and intratumour. Inter-patient heterogeneity exists at the level where even within cancer of the same type, patients exhibit differences in terms of both biology and prognosis, for example, in breast cancer, based on gene expression, patients can be classified into at least four broad subtypes, namely basal (triple negative breast cancer (TNBC)), HER2+, and luminal A and B subtypes. Luminal A and B are estrogen positive cancers, with luminal A having the best prognosis. True to its name, HER2+ has overexpression of the HER2 growth enhancing protein. These are slow growing tumours and respond well to treatment. The Basal or TNBC subtype is triple negative for estrogen, progestin receptors and HER2. This is the most aggressive subtype, and is unresponsive to treatment. Extending on this, intra-patient heterogeneity is explained by the differences in the primary and the metastatic tumour sites; these include morphological and genetic differences, and differences in terms of tumour aggressiveness and proliferation. Intratumour heterogeneity is then explained by difference between regions within the same tumour mass (Gerlinger et al., Gerlinger et al.) (Figure 1.1).

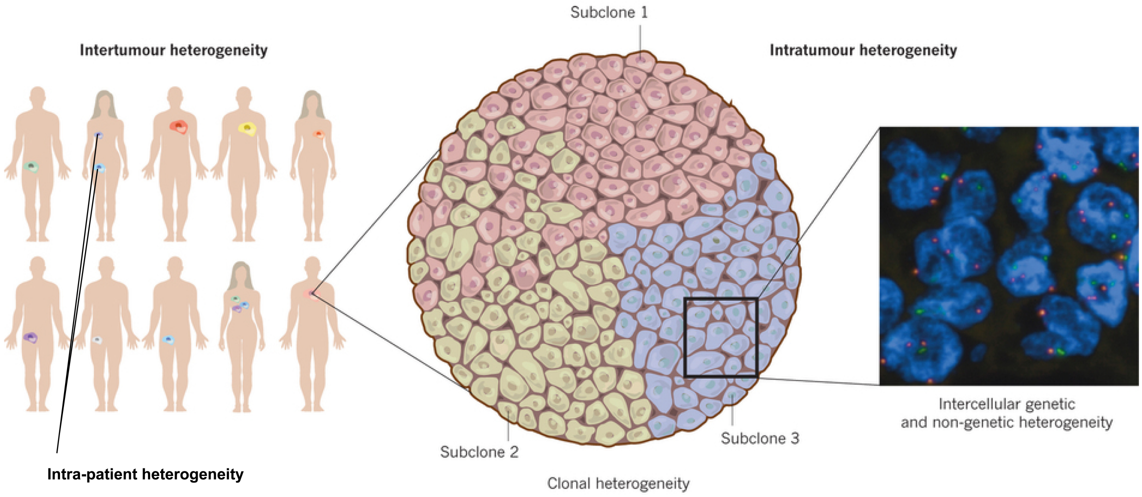

Intertumour heterogeneity
Subclone 1
Intratumour heterogeneity
Intercellular genetic
and non-genetic heterogeneity
Subclone 3
Subclone 2
Intra-patient heterogeneity
Clonal heterogeneity

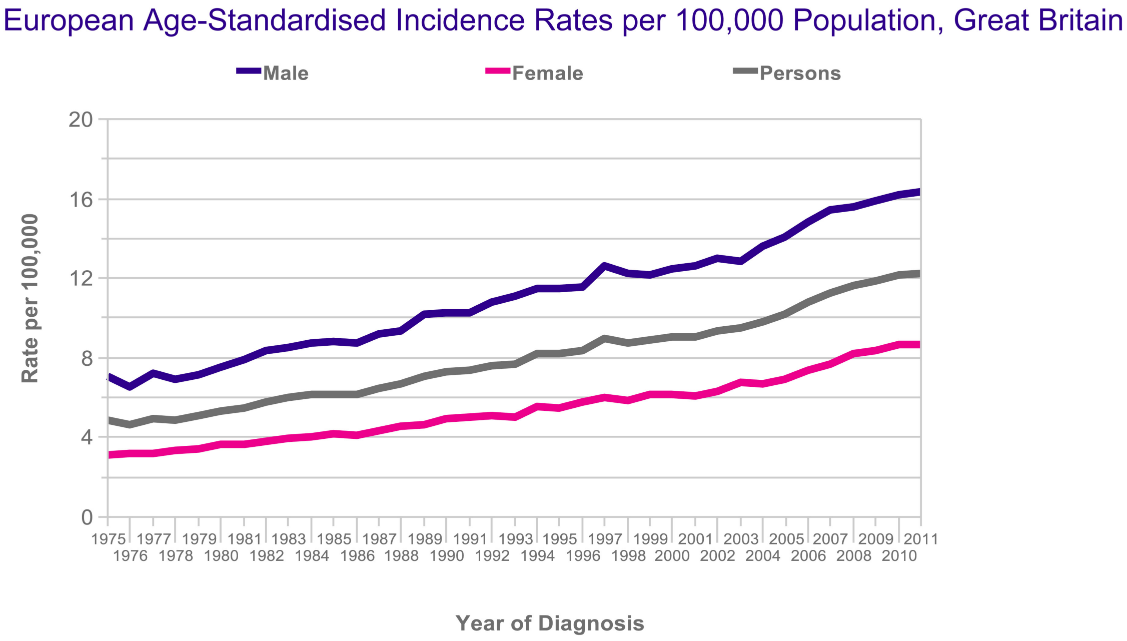
Figure 1.1: The different levels of tumour heterogeneity
This figure depicts tumour heterogeneity at different levels. At the top level, phenotypic and genotypic diversity between patients within the same cancer type is referred to as Intertumour or Inter-patient heterogeneity in the population. Within the same patient, we can observe heterogeneity between tumours from different sites, example primary and metastatic sites, which is referred to as Intra-patient heterogeneity and ultimately heterogeneity can be seen at the intratumour level. Multiple sites within the same tumour biopsy can show variations both at the genetic and non-genetic levels. Figure modified with permission from ((Burrell et al.), Nature Publishing Group).

Although the existence of high levels of intra-tumour heterogeneity now appears to be well established, little is known as to the exact advantages this provides tumours (Stratton). However, it is postulated that heterogeneous mutations may provide tumour cells with specific advantages. Such advantages at the cellular level may increase the fitness of tumours under specific environmental conditions, leading to cancer progression, drug resistance and eventually patient death. The big question facing the research community is how can the high volume of primary information, collected at both the genetic and phenotypic levels, be integrated to help cancer patients?

To an extent, recent developments in the high-throughput technologies have lessened the difficulties in monitoring the systemic changes occurring during cancer cell progression. Computational algorithms have become indispensable for the integration of different types of ‘omics’ datasets, such as sequencing, mRNA expression and protein interaction data. Moreover, since cellular systems are perturbed both during the onset and development of cancer, and the behavioural change of tumour cells usually involves a broad range of dynamic variations, computational approaches developed for network analysis are becoming especially useful for providing insights into the mechanism behind tumour development and metastasis. These ideas are embodied in this work.

1.1.2 clear cell Renal Cell Carcinoma

Renal Cell Carcinoma is by far the most common form of kidney cancer, with about 9 in 10 cases of kidney cancer being renal cell carcinoma (Motzer et al.). It is comprised of a set of different histologies, out of which the clear cell subtype (ccRCC) is most prevalent (60%-80%), followed by papillary and chromophobe subtypes (Kovacs et al., Thoenes et al.). Characteristically, ccRCCs are defined as cancer cells with clear cytoplasms and nested clusters of cells surrounded by dense endothelial networks (Jonasch et al.). ccRCC is one of the 10 most common cancers in both men and women (Rini and Atkins). ccRCC incidence has increased progressively in the past 30 years (Figure 1.2), which could be in part attributed to development in diagnostic techniques; however, there has been little corresponding improvement in survival (Brannon et al.). As yet, surgery for localised disease is the only curative therapy. Treatment of metastatic disease is even more challenging; with 5-year survival rates for patients with metastatic disease being less than 10% (Motzer et al.).



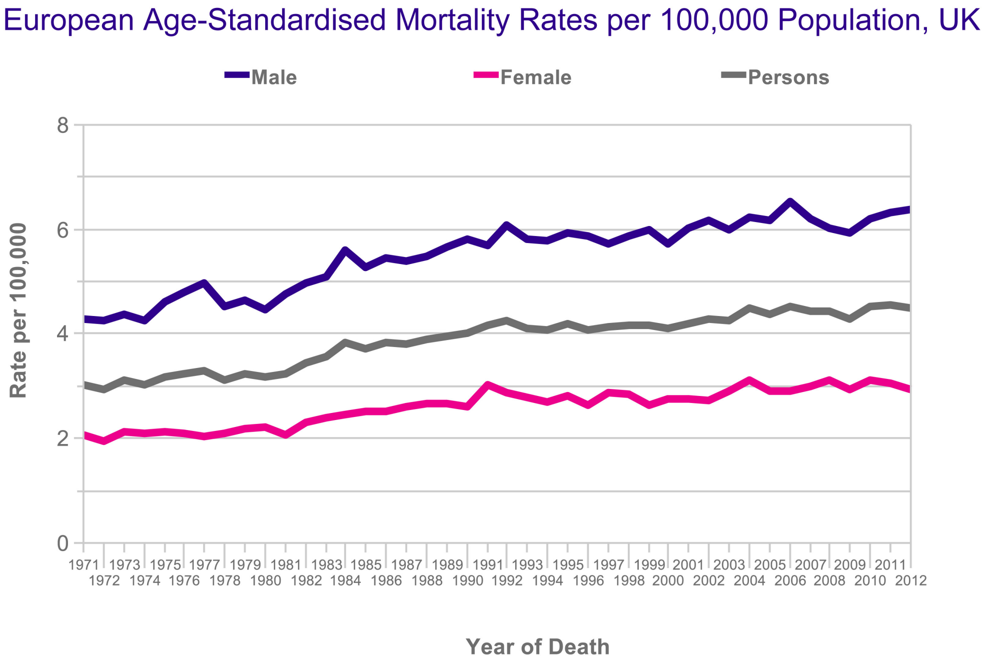


Figure 1.2: ccRCC incidence and mortality within the UK
Graphs depicting the incidence and mortality rates for ccRCC, during the 1970's-2012 in the UK for both males and females. Data presented by Cancer Research UK and adapted from the CRUK website.

Historically, ccRCC has been characterised by biallelic inactivation of the von Hippel Landau (VHL) gene, which can be found in approximately 85% of ccRCCs (An and Rettig). VHL is located on chromosome 3p25 and its biallelic inactivation occurs *via* loss of chromosome 3p along with either somatic mutations in the VHL gene or promoter hypermethylation (An and Rettig, Gossage et al.). Inactivation of the VHL gene leads to stabilisation of the hypoxia response pathway, which in turn leads to increased levels of tumour angiogenesis. Recent work has also shown the inactivation of other tumour suppressor genes such as PBRM1, SETD2, KDM5C and BAP1 (Dalgliesh et al., Guo et al., van Haaften et al., Varela et al., The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.). Intriguingly all these genes function as histone and/or chromatin modifiers. Thus ccRCC seems to be a metabolically and epigenetically controlled cancer (The Cancer Genome Atlas Research Network). In addition to the loss of chromosome 3p, other aberrations have also been reported as recurrent events in ccRCC (Klatte et al., Kroeger et al., Gunawan et al.): gains of 5q (Gunawan et al.), 8q (Klatte et al., Monzon et al.), 12 (Elfving et al.) and losses of 8p (Elfving et al.), 9p (Sanjmyatav et al., Klatte et al., La Rochelle et al., Moch et al., Brunelli et al.) and 14q (Kroeger et al., Monzon et al.). Moreover, along with this inter-patient heterogeneity, recent work has shown substantial genetic intratumour heterogeneity in ccRCC through exome sequencing of several regions from the same tumour as well as somatic copy number alterations (SCNAs) profiling (Gerlinger et al., Gerlinger et al., Martinez et al.); all of which is likely to influence clinical outcome, and provides a possible explanation for the slow progress in development of effective therapies for ccRCC.

To summarise, much work has been done on studying ccRCC in terms of sequencing (Dalgliesh et al., Guo et al., van Haaften et al., Varela et al., The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.), copy number analysis (Beroukhim et al.), analysing gene expression data to classify patients into different subgroups (Brannon et al., Zhao et al., Beleut et al.) and studying patient biology and survival patterns amongst patients (Jones et al., Vasselli et al., Zhao et al., Zhao et al.). This has led to the identification of key genetic events; however their biological effects and their impact on cancer cell fitness are largely unknown (with the exception perhaps of the VHL gene). Although incremental work over the past two decades has increased our knowledge of the disease, there is still the need to understand the mechanisms that not only lead to induction of the cancer but also allow it to proliferate and resist treatment. The implications of previous findings are further limited due to the extensive amounts of both inter-patient and intratumour heterogeneity. These observations mandate further research in the field of ccRCC biology. Large-scale tumour profiling efforts from the Cancer Genome Atlas (TCGA, (The Cancer Genome Atlas Research Network)) and International Cancer Genome Consortium (ICGC (International Cancer Genome et al.)), have made it possible to analyse larger cohorts of patients and unravel tumour mechanisms. A detailed analysis of ccRCC dataset profiled by the TCGA forms the base of the work presented here.

In this thesis, the aim is to present a comprehensive and integrated molecular analyses, analysing somatic mutation, copy number alteration as well as transcriptomics data, to shed light on key driver events for cancer progression and evolution, find prognostic biomarkers and elucidate biological mechanisms underlying ccRCC by interpreting the analyses of TCGA dataset. Moreover, the conclusions drawn from this cohort are related to an in-house multiregion ccRCC dataset (Swanton Laboratory) to gain insights into what drives ITH in ccRCC by defining the variable phenotypes established through genetic ITH.

1.2 clear cell Renal Cell Carcinoma: what is known and where the field stands

1.2.1 Genetics

While mutations in the VHL gene (both somatic or germline) and loss of heterozygosity (LOH) of chromosome 3p is observed in over 90% of ccRCC cases, thereby marking these two events as the major players of this disease as well as the necessary precursors, recent studies have identified somatic mutations in other genes including PBRM1, BAP1 and SETD2 as well as recurrent somatic copy number alterations. In this section, the VHL gene axis of ccRCC biology is reviewed followed by a review of the other recurrent ccRCC associated alterations.

The VHL gene and Hypoxia inducible factors (HIFs)

The VHL gene is a tumour suppressor that lies on chromosome 3p25. It was first characterised in 1993 (Latif et al.), which led to the identification of families with VHL disease. The VHL disease is a hereditary disorder, which predisposes patients to various benign and malignant neoplasms including ccRCC. Typically patients with VHL disease inherit a germline-mutated copy of the VHL gene and subsequently acquire a somatic alteration or loss of the second VHL allele (Latif et al.). Soon after, VHL was also seen to be the main driver event in sporadic ccRCC cohorts as well, where loss of function was taking place by either somatic mutation or promoter hypermethylation in up to 80-90% of the cohort (Gnarra et al., Nickerson et al., Shuin et al., Duan et al., Pause et al., Kibel et al., The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.). Thus the VHL gene is a classic example of a 'two-hit' tumour suppressor gene, where one copy of the gene is lost *via* deletion or chromosomal arm loss and the second copy is inactivated *via* somatic mutation or hypermethylation (Brugarolas, Linehan et al.).

The VHL gene product (pVHL) has various roles, the most important being its function as the substrate recognition module of an E3 ubiquitin ligase complex. This complex targets the hypoxia inducible factor (HIFα) and facilitates its oxygen-dependent ubiquitination. This has been well characterised in relationship to ccRCC (Keefe et al.). When the VHL gene is activated, it targets the HIFs for polyubiquitinilation and degradation. However, when the VHL gene is inactivated, HIFs are stabilised and the cell is under the condition of hypoxia. Under hypoxic conditions, HIF translocates to the nucleus and up-regulates a variety of genes including vascular endothelial growth factor (VEGF) and Erythropoietin (EPO) (Gordan and Simon), which enable the cell to adapt to hypoxic conditions. The up-regulation of VEGF accounts for the highly vascular nature of this disease (Brannon and Rathmell). Furthermore, HIF also targets various metabolism related genes, which mediates the global metabolic shift of ccRCCs (Keefe et al.). Major targets include Glut1, which is a glucose transporter and rate-limiting enzymes involved in glycolysis (hexokinase, phosphofructokinase, lactate dehydrogenase) (Osthus et al., Kim et al., Semenza). HIF translocation also leads to expression of pyruvate dehydrogenase kinases, thereby suppressing the entry of pyruvate into the citric acid cycle (Kim et al.) (Figure 1.3).

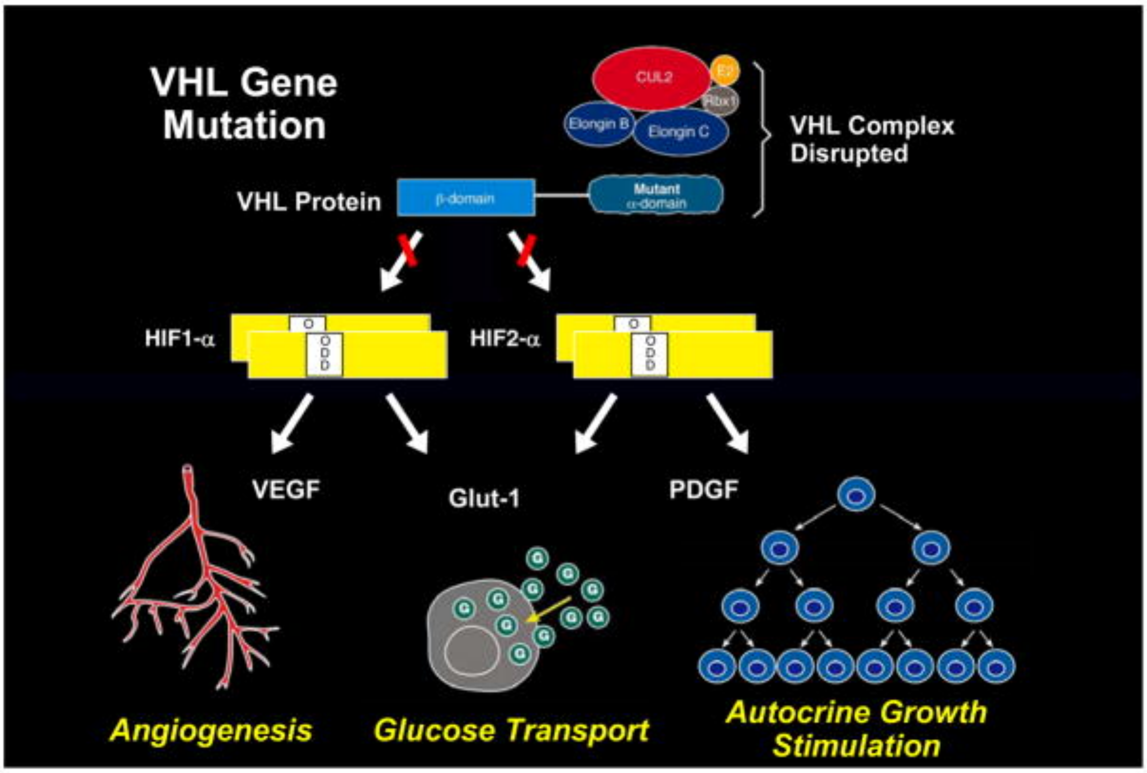


Figure 1.3: Regulation of HIF *via* the VHL protein.
In the absence of pVHL, the VHL protein – Elongin complex is disrupted, which then cannot target the degradation of HIF. Accumulation of HIF in turn leads to activation of downstream targets such as VEGF, PDGF and Glut1. Figure reproduced with permission from (Linehan et al.).

However, recent work has demonstrated that despite large-scale VHL gene loss and correlation with ccRCC, HIF deregulation is not uniform in all patients. It has been postulated that different mutation types may contribute differently to HIF1α and HIF2α regulation (the two most prominent members of the HIF family) (Lee et al., Rathmell et al.). Further work has shown that depending on whether tumours are expressing both HIF1 and HIF2 (H1H2) or only HIF2 (H2), differences can be seen in terms of c-myc transcription factor activity (Gordan et al.) and tumour metabolism. Further, Dalgliesh et al. (Dalgliesh et al.) suggest that H2 tumours may have certain nonsense mutations that puts them under selective pressure to lose HIF1 expression.

The estimates for VHL gene mutations in sporadic ccRCC cohorts vary greatly between studies (Yoshida et al., Banks et al., Gnarra et al., Brauch et al., The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.), with as high as

90% loss being reported (Nickerson et al.). While the importance of VHL gene mutation as a precursor event for ccRCC is well established, its efficacy as a biomarker still remains questionable (Brannon and Rathmell, Lee et al.).

Somatic mutations profile

The advent of next generation sequencing technologies facilitated the possibility of sequencing larger cohorts of patients to look for other majorly mutated genes in ccRCC. In 2009, van Haaften et al. (van Haaften et al.), performed sequencing of the genes involved in histone methylation in different human cancers (n=1390) including 419 ccRCCs. This study proposed the histone lysine demethylase gene UTX/KDM6A as a novel ccRCC gene. UTX is located on chromosome Xp11.2 and encodes a histone H3 lysine 27 (H3K27) demethylase. Studies in model organisms, have suggested that the UTX gene plays a role in cell cycle progression, affecting proliferation as well as cell fate (Wang et al., Herz et al.). Moreover, reintroducing UTX into cancer cells showed decrease in cell proliferation further supporting its role in ccRCC tumourigenesis (van Haaften et al.).

Following this, another study (Dalgliesh et al.), presented results from the selected sequencing of coding exons of 3544 genes in 101 ccRCC cases. They followed up 60 genes in another 246 cases of ccRCCs. Their results not only supported the potential role of UTX/KDM6A in ccRCC tumourigenesis but further highlighted recurrent mutations in major chromatin and histone modifying genes including SET Domain Containing 2 (SETD2), Lysine (K)-Specific Demethylase 5C (JARID1C/KDM5C) and Lysine (K)-Specific Methyltransferase 2D (MLL2) genes. All these genes have been seen to be mutated in many other cancers such as pancreatic, prostate and breast cancers (Rydzanicz et al.). Another gene, Neurofibromin 2 (NF2), which is a regulator of multiple receptor tyrosine kinases and activates pathways such as Ras/Raf/MEK/ERK, PI3K/AKT and mTORC1, was also seen to be recurrently mutated in this cohort. NF2 has been shown to be a potential tumour suppressor involved in mitogenic signalling and cell proliferation (Zhou and Hanemann, 2012). Independent studies have confirmed recurrent mutations in SETD2 (Duns et al.) and JARID1C/KDM5C genes (Guo et al.). Down-regulation of JARID1C has further been shown to promote tumour growth in xenograft models (Niu et al.). Recent work by our collaborators has suggested

SETD2 to be promoting tumourigenesis through potentially suppressing replication stress and coordinating DNA damage repair pathways (Kanu et al.).

Subsequently, Varela et al. (Varela et al.), performed exome sequencing for 7 ccRCC cases, identifying recurrent mutations in the Polybromo 1 (PBRM1) gene. This observation was followed up in a mixed RCC cohort of 257 cases, 227 of which were ccRCC. The PBRM1 gene was seen to be mutated in 41% of the cases in this cohort. PBRM1 encodes the BAF180 protein, which is part of the SWI/SNF chromatin-remodelling complex (Reisman et al.). This complex is involved in multiple processes such as replication, transcription (Burrows et al.), DNA repair, cell proliferation (Burrows et al., Xia et al.) and chromosome stability (Xue et al., Vries et al.). The observation that PBRM1 is the second most frequently mutated gene in ccRCC has since been verified in various independent study cohorts (Duns et al., Pena-Llopis et al., Kapur et al.). While mutations in PBRM1 have been shown to have better patient prognosis than those with mutations in the BAP1 gene (Kapur et al.), loss of PBRM1 protein expression nevertheless correlates with advanced tumour stage, low tumour grade and a relatively poor patient prognosis (Pawlowski et al.).

Guo et al. (Guo et al.) performed whole exome sequencing of 10 ccRCC samples, which identified 282 somatic mutations in over 234 genes. They further queried for genes identified in cosmic and cancer gene census to be associated with ccRCC and combined them with their identified genes, to compile a list of ~1100 genes which were screened for in 88 cases. This identified 23 genes as frequently mutated in ccRCC, confirming previous finding of frequent mutations in VHL, PBRM1, SETD2, KDM5C and additionally identified genes associated in the ubiquitin mediated proteolysis pathway (UMPP) to be mutated in ccRCC. These genes included BAP1, SYNE2, SPTBN4, RYR1, AHNAK, ZNF804A, TSC1, SHANK1, LRRK2, FMN2, FAM11B and ASB15. Further targeted sequencing of 135 genes involved in UMPP showed mutation in 50% of the analysed tumours in at least one of these genes, including VHL (27%), BAP1 (8%), CUL7 (3%) and BTRC (2%). The BAP1 gene has since then been shown to be of particular interest in independent ccRCC cohorts (Duns et al., Pena-Llopis et al., Kapur et al., The Cancer Genome Atlas Research Network) as well as other cancers (Rydzanicz et al.). It encodes a ubiquitin carboxy-terminal hydrolase/BRCA1-associated protein (BAP1), and is implicated in DNA damage response, cell cycle regulation and cell growth (Matsuoka et al., Yu et al.).

All the above studies were either targeted towards selected genes or were done on smaller cohorts. However in 2013, two studies were published showing whole exome sequencing (WES) data for a large cohort of ccRCC cases. The first study (Sato et al.), performed genome and exome sequencing; whole exome sequencing was done for ~110 cases. The top significantly mutated genes were VHL, PBRM1, BAP1 and SETD2. Furthermore, they also identified recurrent mutations in the Transcription Elongation Factor B (SIII), Polypeptide 1 (TCEB1) gene, which leads to HIF accumulation by disrupting the binding of the C-VHL gene product with elongin.

The second study was published by the TCGA (The Cancer Genome Atlas Research Network), which comprises the biggest study cohort for ccRCC cases to date. Over 400 cases were undertaken for WES, clinical follow-up, SNP-array analysis and RNA-Sequencing (RNA-Seq). This analysis underlined recurrent mutations in the chromatin machinery in ccRCC and underscored ccRCC being a metabolic and chromatin regulation controlled cancer.

The most recent study (Scelo et al.), includes a cohort of ~100 ccRCC cases of either European or British origin. Whole genome sequencing was performed for 94 cases out of which 25 were from the United Kingdom. Non-synonymous mutations were detected in 583 genes and the top 5 most frequent genes identified included VHL, PBRM1, SETD2, BAP1 and Zinc finger homeobox 4 (ZFHX4) gene. This study was important in highlighting that ccRCC mechanisms may differ in diverse populations. This has been discussed in more detail in Chapter 3.

A few striking features emerged from all these studies. The first is the emergence of recurrent mutations of chromatin and histone modifying genes, namely PBRM1, SETD2 and KDM5C (Figure 1.4). The second remarkable feature is that PBRM1, SETD2 and BAP1 all share proximity to the 3p25 locus which is where the VHL gene also resides. Most mutations in these genes are loss of function mutations and this combined with the fact that 3p LOH occurs in over 90% of ccRCC cases, results in the complete inactivation of these genes when mutated. While all these genes show recurrence in terms of being mutated in various ccRCC cohorts from different origins, their importance in terms of cancer progression and prognosis still remains to be shown.

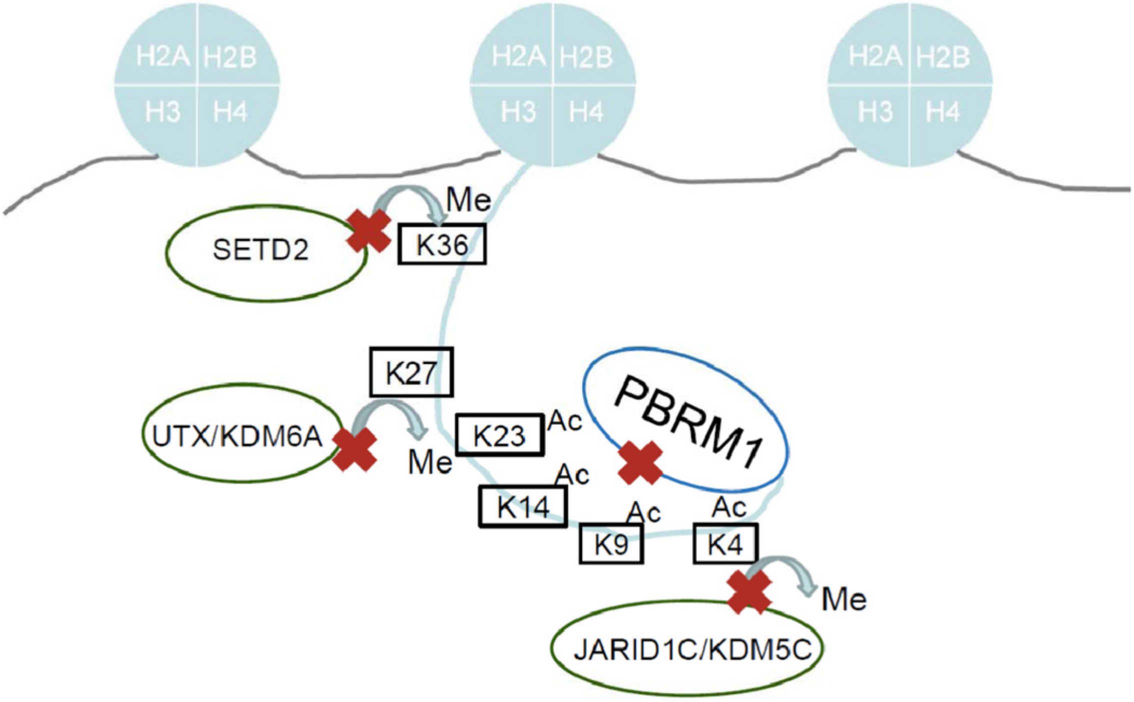


Figure 1.4: Major chromatin regulators in ccRCC
Major chromatin regulators altered by somatic mutations and somatic copy number alterations (SCNAs) in ccRCC are depicted. SETD2 is involved in trimethylation of H3K36, UTX/KDM6A is a H3K27 demethylase, KDM5C is a H3K4 demethylase and PBRM1 is part of the SWI/SNF complex. Figure reproduced with permission from ((Jonasch et al.), American Association for Cancer Research).

Chromosomal aberrations

Chromosomal aberrations include both numerical changes (duplications or deletions) of parts of or whole chromosome as well as structural rearrangements of chromosomes.

Besides somatic mutations described above and 3p LOH, there are multiple other genetic aberrations which maybe supplementing ccRCC from development to progression. Molecular cytogenetic methods such as fluorescent in-situ hybridisation (FISH), array based comparative genomic hybridisation (CGH) and more recently, single nucleotide polymorphism (SNP) based arrays have led to advancements in detecting recurring chromosomal aberrations for ccRCC at higher resolution and at genome wide levels (Rydzanicz et al.).

Loss of chromosome 3p spanning the VHL locus is an example of a hallmark somatic copy number alteration (SCNA) in ccRCC and has been described in various studies (Junker et al., 2003, Chen et al., Klatte et al.). Interestingly, two large-scale studies have described 3p LOH to be associated with a relatively good patient prognosis. In the first study (Klatte et al.), cytogenetic analysis was used to look for recurrent gains and losses in ccRCC and they assessed Cancer Specific Survival (CSS) to analyse the association of these aberrations with patient prognosis. Another study (Kroeger et al.) compared patients with loss of 3p or with loss of 14q and patients with loss of both these chromosomes with each other. Here too they observed that while patients with loss of only 3p had better prognosis, patients with loss of 14q had poorer survival in comparison and patients with loss of both had the worst prognosis.

The second most recurrent SCNA for ccRCC is gain of chromosome 5q (Gunawan et al., Junker et al., 2003, The Cancer Genome Atlas Research Network). In the TCGA publication (The Cancer Genome Atlas Research Network), using focal amplifications, the authors narrowed the amplification down to 5q35 focal region, with 60 putative target genes including genes involved in histone modification (EZH2), stress response (STC2), cell adhesion and migration (VCAN). All these genes are over-expressed in ccRCC. In addition, using cytogenetic analysis Gunawan et al. (Gunawan et al.) described the focal region 5q31-qter to be associated with CSS, where patients with 5q gain had relatively better survival than those without.

Multiple groups have discussed the loss of chromosome 9p as an important recurrent SCNA in ccRCC and its association with poor prognosis has been observed in various independent study cohorts (Moch et al., Sanjmyatav et al., Klatte et al., La Rochelle et al., Brunelli et al.). All these studies described loss of the whole p-arm of chromosome 9, apart from (Sanjmyatav et al.), where loss of two particular focal regions namely 9p21.3p24.1 and 9q32q33.1 was described to be associated with poor prognosis.

Besides deletions of chromosomes 3p and 9p, Klatte et al. (Klatte et al.), also described other recurrent gains and losses. Losses of chromosome 4p and 14q were each associated with poor patient prognosis.

Sanjmyatav et al. described some other SCNAs (Sanjmyatav et al.) detected using CGH arrays and Fluorescent in-situ hybridisation (FISH). They studied CSS in patients and observed gain of 7q36.3 region to be associated with poor patient survival. Gain of focal region 20q11.21q13.32 was also identified to be associated with poor patient prognosis.

Gain of the q-arm of chromosome 8 has been described by two groups to be associated with poor patient survival. Klatte et al. (Klatte et al.), used cytogenetics to identify recurrently altered chromosomes in ccRCC, and then studied CSS to identify 8q gain to be associated with poor prognosis. The second group (Monzon et al.) used SNP arrays to identify 8q gains in ccRCC patient, reasserting the above observation. They studied overall survival for their patient cohort. In the same study, the authors described loss of chromosome 14q to also be associated with poor patient prognosis.

Other ccRCC recurrent SCNAs in ccRCC include deletion of chromosomes 1p, 4q (Beroukhim et al., Girgis et al.), 6q (Toma et al.), 8p (Chen et al., Toma et al., Girgis et al.), and amplification of chromosomes 7q (Beroukhim et al., Girgis et al.) and 12q (Girgis et al.).

To further our understanding of associations of SCNAs with cancer, the algorithm GISTIC was developed (Beroukhim et al., Mermel et al.). GISTIC identifies genes targeted by SCNAs that drive cancer growth. The algorithm estimates the background rates for each category and defines the boundaries of SCNA regions, by separating SCNA profiles into underlying arm-level and focal alterations. Beroukhim et al. (Beroukhim et al.) used GISTIC on a cohort of 90 ccRCCs and identified gains of 1q, 2q, 5q, 7q, 8q,12p and 20q and losses of 1p, 3p, 4q, 6q,8p,9p and 14q of to be associated with ccRCC.

All SCNAs discussed above have potential tumour suppressors and oncogenes located on them, and further studies that analyse the effect of these SCNAs at the gene expression and functional levels should provide insights into ccRCC tumourigenesis.

1.2.2 Transcriptomics

Gene expression profiling is a very powerful tool when studying cellular phenotypes. Indeed multiple studies analysing mRNA expression, have been able to elucidate gene associated with ccRCC at the phenotypic level. Using microarray data analyses, a study (Tun et al.) has shown genes associated with three major pathways, namely loss of normal renal function, down-regulated metabolism and immune activation to be reflective of the ccRCCs. These findings were verified in another independent study (Zhou et al.). A recent study using RNA-Sequencing data from 537 patients profiled by TCGA, identified 186 differentially expressed genes after multiple testing correction with |log fold change | > 5 (Yang et al.). Moreover, expression levels for multiple genes have also been shown to associate with patient outcome. For example, high levels of CD31, EDNRB and TSPAN7 have shown to correlate with better prognosis (Wuttig et al.). A three-gene signature based on the expression levels VCAM1, EDNRB and RGS5 has been shown to predict ccRCC prognosis effectively and higher levels of these genes too correlate with better prognosis (Yao et al.).

Furthermore, similar to other cancer types, clustering of patients based on transcriptional signatures has revealed multiple molecular subgroups in ccRCC. Such analyses are done primarily with two objectives in mind. The first being to be able to classify patients into discernible subgroups and a second closely related objective is to be able to distinguish between these subgroups, not only in terms of biological signatures but also in terms of patient prognosis. Most such studies start as unsupervised clustering analyses (Zhao et al., Kosari et al., Vasselli et al., Takahashi et al., Brannon et al., Beleut et al.). These studies use gene expression analysis to form an understanding of the dataset such as, number of subgroups and genes differentially regulated between subgroups. Subsequently, a superimposed question is asked, such as how to maximise patient prognosis, enabling a more supervised analysis to be performed. There have been two landmark studies showing the existence of at least two major subgroups in ccRCC with differences in biology as well as prognosis (Zhao et al., Brannon et al.). These studies are discussed in further details in Section 1.2.7.

1.2.3 DNA methylation, microRNA profiling, Tissue Microarrays (TMAs), and plasma serum protein analysis

Other than those discussed above, other methodologies are available to detect alterations and biomarkers for ccRCC. While these methodologies do not form part of the work covered in this thesis, they are briefly outlined below.

DNA methylation

DNA methylation is a cellular mechanism frequent utilised by cells for epigenetic silencing of genes. For ccRCCs, methylation of the VHL promoter region is seen frequently as an alternative method of silencing the gene instead of somatic mutations (Herman et al., Clifford et al.). Promoter methylation of the DLEC1 tumour suppressor gene was shown to be associated with advanced tumour stage and grade (Zhang et al.). Significant correlation was also observed between methylation of SCUBE3 and increased risk of death or relapse for RCC (Morris et al.). In the TCGA publication (The Cancer Genome Atlas Research Network), the authors identified 289 genes epigenetically silenced in at least 5% of cases. For example, the UQCRH gene was observed to be hypermethylated in about 36% of the tumours, a gene previously been proposed to be a tumour suppressor gene.

MicroRNA expression

MicroRNAs (miRNA) are 21-23 nucleotide long segments of single-strand non-coding RNAs. They have been implicated in tumour development as well as progression. Since a single miRNA targets the expression of many genes, aberrant expression of miRNAs can be an effective mechanism for epigenetic regulation. A number of studies have been undertaken to ascertain the ability of miRNAs to distinguish RCCs from normal cells or between RCC histologies (Chow et al., Huang et al., Juan et al., Petillo et al.). Two studies were undertaken to assess miRNAs associated with metastasis, and their analysis suggested that miR-10b, miR-29a and miR-30a characterise metastatic potential (Junker et al.). Recent work has shown the overexpression of miR-210 in ccRCCs and while miR-210 positive patients had a higher chance of disease recurrence and shorter overall survival in univariate analysis, the statistical significance of this classification was lost when adjusted for tumour size and stage (Samaan et al.).

Tissue microarrays

Tumour protein expression levels can be efficiently assessed using Tissue microarrays (TMAs). TMAs from 800 ccRCCs were analysed for genes in pathways reported to be controlled by VHL and PTEN genes. Improved prognosis was observed for tumours that stained positive for p7 and CAIX (stage T2 and T3 tumours) (Dahinden et al.). In another study of 308 ccRCC patients, high nuclear HIF2α was shown to be associated with smaller tumour size and lower Fuhrman grades while high cytoplasmic HIF2α correlated with lymph node and distant metastasis as well as higher Fuhrman grades (Kroeger et al.).

Plasma serum proteins

Due to the ease of sampling, plasma serum proteins form an attractive category of biomarkers. Compared to tumour biopsies, a blood test is relatively simpler, as well as less invasive for the patients. Potential biomarkers for response to the drug sunitinib in metastatic RCC (mRCC) patients have been identified. Those responding had low levels of TNFα (tumour necrosis factor α) and MMP9 (matrix mettalloproteinaise-9) (Perez-Gracia et al.). Tumour response has also been seen to be correlated with changes in serum levels of VEGF, sVEGFR-s and sVEGFR-3 levels (Deprimo et al.). In another study, sVEGFR-3 and VEGF-C serum levels were shown to be associated with longer progression free survival and response rate in bevacizumab-refractory mRCCs (Rini et al.). While these serum proteins have been shown to have potential predictive power, they all still need to be validated in independent cohorts.

1.2.4 ccRRCs and the PIK3CA/mTOR pathway

Beside the VHL/HIF pathway, the PIK3CA/mTOR pathway is also a prominent in ccRCCs, indeed it is classified as the second major pathway recurrently altered in ccRCC. The mammalian target of rapamycin (mTOR) protein has been identified as part of two complexes (mTORC1 and mTORC2) in humans (Oosterwijk et al.). mTORC1 has been shown to play a role in cell growth regulation and metabolism (Jonasch et al.). mTORC1 has been shown to suppress autophagy (He and Klionsky, Sancak et al.) and regulating mitochondrial functions (Blagosklonny and Hall), but its most important role is to promote protein translation which is mediated through phosphorylation of S6K and the eukaryotic initiation factor 4E-binding protein 1 (Ma and Blenis). As suggested by its name, mTORC1 is inhibited by rapamycin and its analogues and it controls the regulation of HIF1 (Zhong et al., Brugarolas et al., Thomas et al.). Upstream regulation of the mTOR pathway is controlled by the PTEN gene. Mutations in PTEN are rare in ccRCC (The Cancer Genome Atlas Research Network); however, deregulation of the PIK3CA/mTOR pathway can be seen in 17-28% of the cases if mutations across different genes in pathway as well as ITH are taken into consideration (The Cancer Genome Atlas Research Network, Gerlinger et al.).

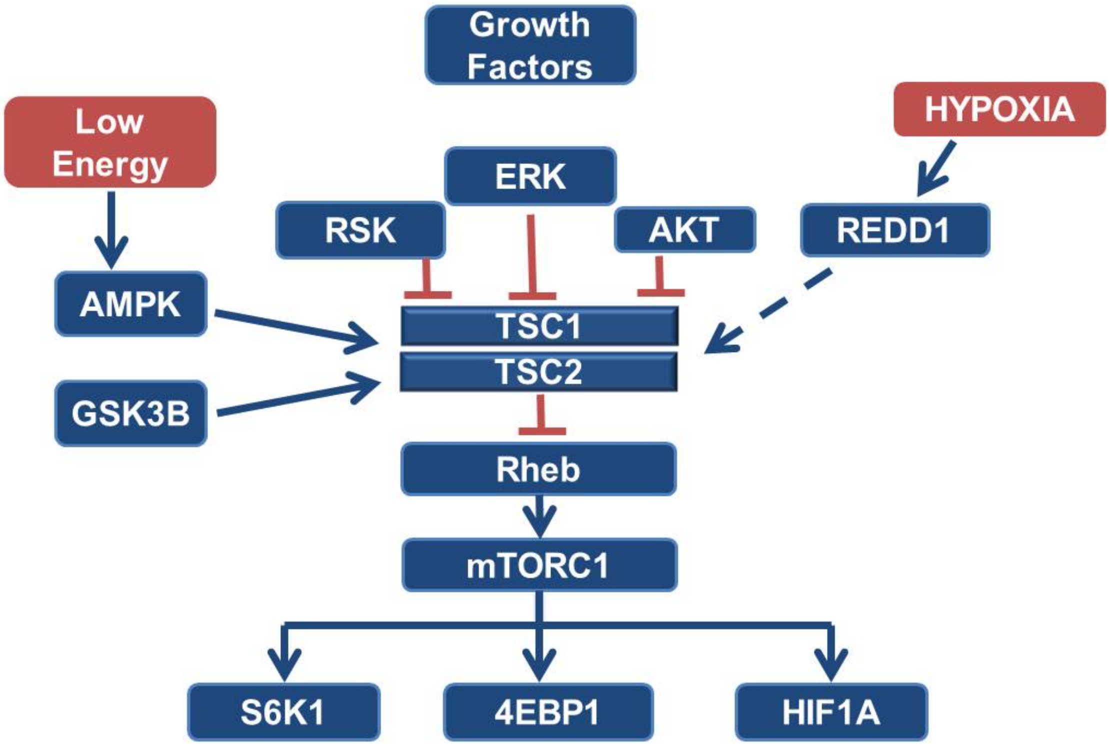


Figure 1.5: the mTORC1 pathway in ccRCC
Figure depicts the HIF and mTORC pathway connections. HIF expression is blocked under hypoxia in a TSC1/TSC2 and REDD1 dependent manner.

1.2.5 Intratumour heterogeneity

So far, this introduction has concentrated on the evidence of extensive genetic and phenotypic heterogeneity at the inter-patient level. Recent work in multiple cancers has also revealed existence of this heterogeneity within individual tumour samples, so called intra-tumour heterogeneity (ITH) (Gerlinger et al., Gerlinger et al., Nik-Zainal et al., Shah et al., Ding et al., Yachida et al., Sottoriva et al., Thirlwell et al., Campbell et al., Bashashati et al., Navin et al.). ITH is being increasingly perceived as a major challenge for the implementation of personalised cancer medicine. Branched cancer evolution analogous to Darwinian evolution, may lead to multiple distinct clones which may co-exist in a tumour mass and result in varying degree of ITH (Navin and Hicks).

Using single biopsies to sample such heterogeneous tumours may lead to undervaluing the extent of this heterogeneity. The analysis of multiple tumour regions from individual ccRCCs has identified substantial ITH, indicating that cells within a single tumour do not all share the same mutations, rather subgroups of tumour cells differ in their mutation spectrum (Gerlinger et al., Gerlinger et al.). The authors identified extensive genetic intra-tumour heterogeneity by whole exome sequencing of multiple regions i.e. multiple biopsies from the same tumour. According to these studies, approximately 50% of the mutations found in any one biopsy are not shared with other biopsies from the same patient. They showed that the ancestral relationship between the mutations in different regions from a patient could be represented in the form of a phylogenetic tree. The trunk of such a tree represents the clonal or ubiquitous mutations for a single tumour, i.e. mutations that are present in all malignant cells assessed, for example VHL (Figure 1.6). The branches represent shared or private mutations to a particular region i.e. the subclonal mutations. The mutational timing can be approximated from such a tree on the basis of the distance between subclonal mutations and the mutations is the most recent common ancestor; thus truncal mutations occur before the mutations on the branches. Spatially separated subclones harbouring distinct driver mutations and somatic copy number aberrations (SCNAs) were present within primary tumours and between primary tumours and metastases (Gerlinger et al., Gerlinger et al., Martinez et al.). Phylogenetic reconstruction revealed branched evolution, demonstrating that multiple subclones were evolving simultaneously within individual tumours. Assessment of a validated prognostic gene expression signature (Brannon et al.) showed expression of the good prognosis ccA signature, or poor prognosis ccB signature, in different tumour regions within the same patient (Gerlinger et al.).

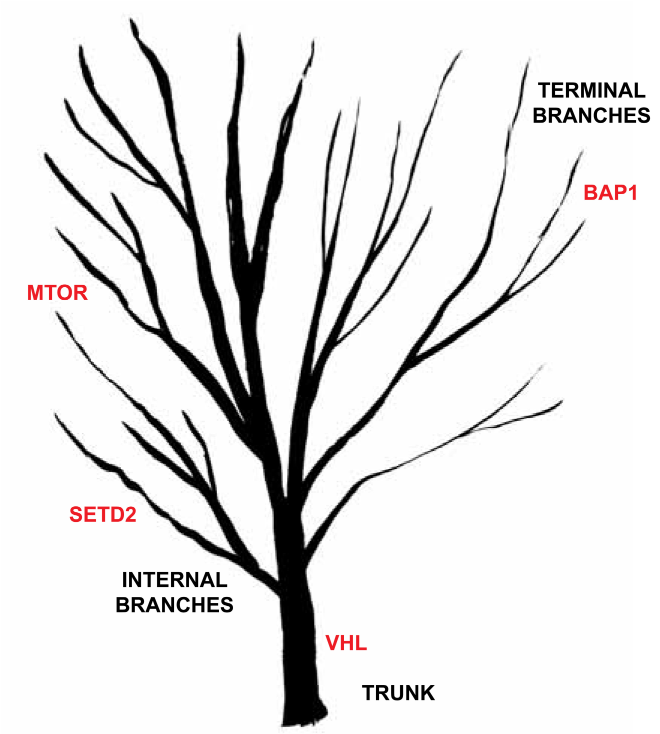


Figure 1.6: Branched evolution in ccRCC development
Figure depicts a tree to show the branched evolution as observed by sequencing multiple regions from individual tumour biopsies (Gerlinger et al.). Examples are shown for mutations in genes (red), representing clonal and subclonal mutations.

Furthermore, some of the major somatic mutations and SCNAs discussed above such as mutations in SETD2 and BAP1 genes and deletions of chromosome 9p and 14q were observed to be subclonal events (Gerlinger et al.). These studies have shown that while the clonal or ubiquitous events in ccRCC are consistent (VHL, 3p loss), there is still diversity in terms of patient outcomes. These observations support the hypothesis discussed previously, that subclonal events play a major role in tumour progression, and may increase the fitness of tumours under specific environmental conditions, leading to cancer progression, drug resistance and eventually patient death.

While quantitative analyses on mutation timing are beyond the scope of the work presented here, the qualitative impact of ITH, at both the biological and prognostic levels for ccRCC, are described in a number of places in this thesis.

1.2.6 Staging, prognosis and management of disease

Tumour stage describes the progress of the tumour cells. The American Joint Committee on Cancer (AJCC) tumour node metastasis (TNM) staging system has been the most widely adopted system for ccRCC. It classifies tumours with a combined stage between I-IV using three values. T gives the size of the primary tumour and extent of invasion, N describes if the tumour has spread to regional lymph nodes and M is indicative of distant metastasis. The exact method of combination of these three factors to give an overall stage has been explained in detail in Figure 2.1.

In terms of prognosis, Stage I patients have the best prognosis with 5 year survival rates of ~80-95%. The survival rates progressively worsen with stage, with Stage II patients having survival rates of ~80% and Stage III ~60%. Even with advances in targeted therapies, Stage IV patients have survival rates of just over two years (Jonasch et al.).

At the time of diagnosis, about three quarters of the ccRCC cases, present with localised disease and with no evidence of metastasis (Jonasch et al.). Partial (partial kidney removal) or radical (complete surgical removal) nephrectomy remains the gold standard for treatment of localised disease. Different clinical trials have discussed and compared the merits of both these methods and results favoured partial nephrectomy since it preserves kidney function.

1.2.7 Clinical prognosis and molecular biomarkers

As discussed in the previous section, the clinical behaviour of ccRCCs is highly variable, ranging from slow-growing localised tumours to aggressive metastatic disease. ccRCCs are resistant to both chemotherapy and radiotherapy, with surgery (nephrectomy) for localised disease, being the only suitable treatment. Prognostic markers in routine clinical use include tumour stage and histologic grade, necrosis and blood tests aimed at measuring levels of lactate dehydrogenase, haemoglobin, platelets, and calcium levels. Other markers include prior nephrectomy, symptoms, and performance status (Motzer et al., Heng et al., Tang et al., Sorbellini et al.). Multiple prognostic models and nomograms have been developed that evaluate and incorporate a combination of these factors. The Mayo clinic's SSIGN model is based on the stage, size, grade and the extent of necrosis (Ficarra et al., Frank et al.); while the University of California, Los Angeles integrated staging system (UISS) quantifies stage, tumour grade and performance status (Zisman et al., Han et al.). Another widely used model is the Leibovich score, which incorporates tumour size, stage, grade, necrosis and regional lymph node status to predict relapse of disease after radical nephrectomy (Leibovich et al.). A recent review (Lane and Kattan) compared the utility of these models. However, the accuracy of predictions for each individual patient remains limited.

Molecular prognostic markers are thus important to guide therapeutic intervention and follow-up strategies. Traditionally molecular biomarkers have been researched at the level of gene expression, for example, in breast cancer a gene expression panel to classify patients into prognostic groups is commercially available and has been adopted in clinical practice (Glas et al.). Multiple biomarker studies have been published for ccRCC, finding recurrent somatic mutations, SCNAs as well as gene expression signatures to be clinically associated with ccRCC. The first few clinical studies concentrated on finding association between the VHL gene mutations and patient prognosis. Two groups, both analysing loss of function (insertions/deletions and nonsense) mutations in the VHL gene, described patients with loss of function mutations to have a poor prognosis when compared to those who did not have these mutations (Schraml et al., Kim et al.). In addition, a third group (Yao et al.) studied VHL gene alterations, which included both mutation as well as hypermethylation events, and assessed the association of these with CSS, finding that patients with VHL gene alterations had better prognosis than those without. However, this association was seen only for Stage I-III cases. It has been shown that non-synonymous mutations in BAP1 gene are associated with poor survival in ccRCC when compared to patients with PBRM1 mutations (Kapur et al.). Further, other groups have validated the association of BAP1 gene with poor prognosis in independent cohorts (Hakimi et al., The Cancer Genome Atlas Research Network, Sato et al.). Mutations in other key genes such as the SETD2 gene have been described to be associated with poor patient prognosis (Hakimi et al., Sato et al.).

As described in section 1.2.1, several recurrent SCNAs have been observed to be associated with ccRCC prognosis. Some examples include deletion of 3p, 4p, 8p, 9p, 14q and 19 (Klatte et al., Kroeger et al., Elfving et al., Sanjmyatav et al., La Rochelle et al., Moch et al., Brunelli et al., Antonelli et al.), and amplification of 5q, 7q, 8q and 20q chromosomes (Gunawan et al., Sanjmyatav et al., Klatte et al., Monzon et al., Elfving et al.).

In terms of studying gene expression signatures, high expression levels of CD31, EDNRB and TSPAN7 have been described to be associated with better prognosis (Wuttig et al.). Using microarray data and hierarchical clustering Zhao et al. (Zhao et al.), showed the existence of two main subgroups of ccRCC differing in terms of prognosis. These two subgroups could be further divided into five subgroups differing in terms of gene expression. They also developed a panel of 259 genes that could be used to stratify patients into good and poor prognosis groups. Another landmark study in 2010 (Brannon et al.), showed that ccRCC patients could be divided into two molecular subtypes, ccA and ccB, where ccB patients had a worse prognosis as compared to ccA patients. This group also used microarray data and using logical analysis of data (LAD) analysis devised a panel of 120 probes (110 genes) to classify patients into ccA and ccB subgroups. A recent publication by Bostrom et al. (Bostrom et al.), showed the association of a TGFβ pathway expression signature with ccRCC prognosis. Here the authors showed that patients with higher TGFβ activity had a poorer survival prognosis than those with lower activity.

While all these publications have shown clinical association of various molecular signatures with patient prognosis, most associations have not been independently validated. Even those that have been validated have not entered clinical practice. Neither have these biomarkers been compared with each other to identify lead candidates for further development. This leads to the question as to which of these predictors are independently associated with patient prognosis and do they influence each other? Furthermore, as discussed in section 1.2.5, most of these alterations have been observed to be subclonal; thus ITH with spatially separated subclones can lead to sampling biases that may contribute to the lack of clinically qualified biomarkers in ccRCC. Such observations raise questions regarding how biomarker discovery strategies can be improved in heterogeneous tumours.

1.2.8 Therapy and targets

The increase in our understanding of the genetic factors underlying ccRCC, has translated to improvements in target identification and therapies. The understanding of VEGF and mTOR being central to ccRCC biology, led to the implementation of multiple antiangiogenic drugs (sunitinib, sorafenib, pazopanib, everolimus, and bevacizumab plus interferon-α) for ccRCC treatment (Junker et al.). VEGF targeted therapies produce a more robust Response Evaluation Criteria in Solid Tumours (RECIST) - response than cytokine therapy (Jonasch et al.). However, response rates vary between 10-50% depending upon the VEGF inhibitor used. mTOR-targeted therapy rates are reported to be more modest, although mTOR and VEGF-targeted therapies haven't been compared within the same cohorts (Jonasch et al.).

While the response rates for targeted therapies are impressive (Rini and Atkins), there are uncertainties regarding specific targeting of RCC cells (Huang et al., 2010a). Furthermore not only toxicity due to these drugs still remains an obvious concern (van der Veldt et al.), resistance mechanisms have also been observed to develop to both these targets (Huang et al., Rini and Atkins). Moreover, at least in part, it has been argued that both treatment response and toxicity maybe reflective of the underlying genetic makeup of the patients. Multiple studies have evaluated this (van Erp et al., van der Veldt et al., Garcia-Donas et al., Xu et al.) and the data suggest that there might be differential clinical benefits depending on the genotype of the patient and if such hypothesis could be confirmed, future studies could benefit and provide more tailored and personalised cancer treatments.

1.3 Survival modelling

As discussed in section 1.2.7, it is imperative to assess and identify biomarkers, either diagnostic or prognostic. The primary interest in such an analysis is to study the effect of a risk factor or treatment with respect to cancer progression. In survival modelling, the data is referred to as ‘time to event’ data. The objective here is to analyse the time that passes before an event occurs due to one or more covariates. This type of data has three main characteristics. Firstly, the dependent variable or response is defined as the waiting time until the occurrence of a well-defined event, for example, death. Secondly, there are certain observations, which are ‘censored’, i.e. there are certain cases in the data cohort, for whom the event of interest has not occurred at the time the data were analysed; or there is loss of information regarding these cases, the reason of which may be known or unknown. Lastly, the objective of such an analysis is to assess or to control the effect of predictors or explanatory variables on the waiting time. Computationally, there can be multiple ways to address the above, including but not limited to linear regression models, decision trees and support vector machines. However, due to the unique characteristics of this problem, the modelling of this data requires modelling of two specific functions namely the survival and hazard functions and sophisticated tests such as logrank tests (Bland and Altman) and Cox regression analysis (Cox). These methods have been developed specifically for this purpose and are the tests used as part of this thesis.

1.3.1 The Survival and Hazard functions

The two main functions that model survival data are the survival and hazard functions. The survival function, or the survival probability S(t), gives the probability of the survival of an individual from the original time to a future time t. This describes the survival experience of the cohort under consideration.

The hazard function h(t), however, is the probability of the individual to have the event at the given time t. It represents the instantaneous rate of event for an individual, who has also reached or survived to time t. It is different from the survival function in the respect that while the survival function estimates the probability of not having the event, the hazard focuses on the probability of the event occurrence (Clark et al.).

Mathematically the relationship between the hazard and survival functions can be represented as (Clark et al.):

$$h(t) = -\frac{d}{dt}[\log S(t)] \qquad \textbf{1.1}$$

1.3.2 Censoring

Censoring is an important constraint in survival analysis. In a survival analysis, not all patients reach the endpoint of interest till the end of the study period. Therefore 'survival' times are unknown for these individuals. This phenomenon is called censoring and such individuals are censored when analysing survival data. There are three cases in which a patient may be censored: 1) the patient did not reach the endpoint of interest by the end of the study period – for example, if death due to disease is the endpoint of interest, patients alive at the end of the study will fall into this category. 2) Loss of follow-up during the study period – this could happen due to various reasons such as the will of the patient to provide further information. 3) The patient experiences a different event than the endpoint of interest, which renders further follow up impossible – following from the previous example, in this case could be death due to causes other than the disease under consideration.

The above examples all come under what is known as 'right' censoring since the event (if it occurred) is beyond the timeline of the follow up period. However if there was a scenario where while we knew when the event occurs, the time from when it began is unknown, this falls under the category of 'left' censoring. Most survival data include right-censored cases, and for the purpose of this thesis, that is the censoring under use. Lastly, a very important consideration is that the censoring should be uninformative. This means that there is equal probability of the censored cases to have the event as there is of those cases that did have the event; i.e. the censoring does not carry any prognostic information (Clark et al.).

1.3.3 Methods for estimating and comparing survival times

A few typical errors in survival analysis include, counting only the event frequencies i.e. only assessing if the event occurred or not with no thought to the time to event or considering how long the patients were observed. Another mistake could be to exclude patients, for whom the event did not occur from the analysis, and lastly to assume that the time of censoring is equivalent to event time; i.e. no distinction is made whether the patient is recorded as ‘event’ or ‘censored’ (Zwiener et al.). Thus special methods are required to avoid these mistakes.

The Kaplan Meier (KM) method (Bland and Altman) is the most common method used to estimate the survival function. It is a non-parametric method, which estimates the survival probability from observed survival times taking both censored and uncensored observations into account. Mathematically, the KM estimator calculates the probability of survival at time point t from the probability of being alive at time point t-1, where t_0 = 0 and S(0) =1, i.e. at the beginning of time there is 100% survival probability.

However, often in studies assessing survival, there is a need to compare the survival of two or more groups; for example in a drug trial, the survival of patients on the drugs would be compared to the placebo group. While the KM estimator would assess the survival of each group, it does not provide a comparison. There are different tests available to perform these comparisons. Another important point of consideration is the effect of covariates. The survival of an individual may be the consequence of multiple factors, which can be assessed using multivariate models. In the following sections, both univariate tests and multivariate models are discussed.

1.3.3.1 Univariate analysis

The logrank test is the most common statistical test applied to test the significance of differences in survival between two or more groups. It tests the null hypothesis that there is no difference in survival between the two groups. The test is based on the same assumptions as the KM estimator, i.e. the censoring is uninformative, the events happened at the specified times and the probability of the event occurring for an individual is independent of when the individual was recruited for the study. An advantage of this test is that it makes no assumptions regarding the survival distributions or the shape of the survival curves (Bland and Altman). The null hypothesis here is that there is no difference in the survival between the groups under consideration. The exact method of assessment has been described in detail in Methods section 2.6.2.

The logrank test is considered to be a robust test for estimating differences in survival, when compared to other tests (Clark et al.), however, it cannot provide the effect size of the variable which is a definite limitation. A univariate Cox model (Cox) can be used here to compliment the logrank analysis and calculate the effect size or hazard due to the test variable. The Cox model is explained in greater detail in the section 2.6.5.

The competing risk (CR) analysis is another popular method of choice when estimating survival especially when competing events might be taking place. For example if 'death due to disease' is the event of interest, death due to other causes is a competing risk event, as either of these will prevent the other from happening (Satagopan et al., Putter et al.). In contrast to logrank test, which estimates the survival function of the variable under consideration, the CR analysis estimates the cumulative incidence function of a variable. This function calculates the proportion of patients at time t, which have had the event k, taking into account that patients can have other competing events.

1.3.3.2 Multivariate analysis

The Cox proportional hazards model (Cox) is the most commonly used multivariate model for analysing survival data. Here too the assumptions for censoring being right and is uninformative holds. It is a regression model, which defines the association between the occurrence of the event using the hazard function and a set of covariates. The Cox model estimates the hazard function based on a set of covariates, mathematically written as (Equation 1.2):

$$h(t) = h_0(t) exp(\beta x_1 + \beta x_2 + \beta x_3 + ... + \beta X_n) \quad \textbf{1.2}$$

Where β1, β2, β3 represent the coefficients of regression for each variable under consideration. The term h0 represents the baseline hazard and an important feature of the Cox model is that it estimates the baseline hazard non-parametrically giving the advantage of not assuming any underlying statistical distribution for the survival data (Bradburn et al.).

1.4 Computational approaches to understand cancer biology

1.4.1 Integrated genotype-phenotype analysis

The previous sections have discussed the technological advances that have occurred over recent years and that have led to the accumulation of large heterogenous cancer datasets within public repositories and the identification of a growing number of cancer associated genes. Of particular interest for the advancement of the field is the integration of genetic alteration data with the corresponding gene expression data, which holds merit by virtue of the fact that, if deregulation of expression levels of genes relative to the genetic altered genes can be observed, then this serves to provide a stronger signal for the interpretation of the genotype to phenotype relationships in the dataset.

At the most straightforward and simplistic level - combining gene expression data with somatic mutations has been shown to provide answers for elucidating the relationships between genotype and phenotype. Ragazzon et al. (Ragazzon et al.), performed a transcriptome analysis for Adrenocortical Cancers (ACC). They first identified two subgroups of patients by gene expression profiling, and showed that these subgroups had different survival outcomes. Subsequent integration of somatic mutation data with this classification of patients revealed that mutations in TP53 and CTNNB1 genes, which are the two most frequent mutations in ACC, seem to be mutually exclusive and occur only in the poor prognosis cases of ACC.

However, as pointed out previously, genome wide ‘omics’ analyses have revealed a staggering number of genetic alterations, and correlative approaches such as discussed in the previous paragraph, are no longer feasible. Thus, extending on the basic premise of genotype–phenotype relationships, newer methods need to be developed. The DriverNet algorithm (Bashashati et al.) was developed to identify driver mutations that affect and control the mRNA expression signatures of the disease of interest. This method creates a bipartite graph, where on one side are the genetic alterations and on the other is the gene expression network of the genes showing significant deregulation. An edge is drawn from one side to the other, if gene g_i is mutated in the left partition; gene g_j is deregulated in the right partition and there are known interactions between g_i and g_j. Then a greedy algorithm is applied so as to explain as many changes on the right using as few genes possible from the left partition, nominating these genes as driver alterations. Stochastic sampling for null distributions and statistical tests are then applied to filter the driver lists.

Another method, Conexic (Akavia et al.), combines copy number alterations and gene expression data to identify SCNA drivers of cancer progression using Bayesian modelling approaches. It uses a score guided approach to devise a combination of drivers that may best explain the observed gene expression signature across all tumour samples, and then searches for the maximum scoring drivers in the amplified and deleted regions of the samples, to determine the driver SCNAs.

1.4.2 Network analysis

While the above work is commendable, such correlative analyses have still to reach their potential. The presence of high number of non-synonymous somatic mutations in individual tumours proves to be a major hurdle for correlative analyses between expression signatures and individual somatic mutations. A traditional gene or pathway centric approach, which individually evaluates the contribution of each gene or pathway alteration towards the overall cancer phenotype, is highly limited in large-scale ‘omics’ datasets for two reasons: the datasets are huge, consisting of thousands of individual measurements and simple correlative analyses are often limited because of multiple testing errors. Secondly, gene- or pathway centric approaches are usually limited when applied to poorly characterized genes. Although they may reveal associations, they usually don't allow identification of the functional relevance of alterations in poorly characterised genes. Nevertheless, connections between different genes and pathways can potentially be identified through integrative analysis of genome wide datasets, which requires the investigation of multimodal "omics" datasets such as DNA sequencing and mRNA expression data with computational and statistical modelling approaches.

Coupled to this is our increased ability to generate detailed interactome maps that help to enrich our knowledge of the biological implications of cancer mutations. As a result, network analysis approaches have become an invaluable tool to predict and interpret mutations that are associated with tumour survival and progression. A detailed review has also been reported on the applicability of using protein networks information for disease analysis studies (Ideker and Sharan). The authors discuss the applications of mapping human disease associated genes to protein interaction networks, and the subsequent boost in our understanding of human disease mechanisms.

Interaction networks may be generated with information from multiple levels; networks can be created based on genetic alterations in the disease of interest representative of the genotype of the disease. Alternatively, networks can be generated based on gene expression data, which are representative of the phenotype of the disease. Examples of both these approaches are discussed below.

Genetic networks are generally based on a background of protein-protein interaction networks, with an aim to understand cancer biology as well as predict novel cancer driver genes. These studies also aim to use such networks to explain inter-patient heterogeneity and the pathways that the specific cancer may be targeting. Such networks usually exploit two features in genetic alterations; firstly that major 'driver' events will have higher recurrence and secondly alterations in the same pathway would probably be mutually exclusive.

Protein-protein interaction networks are generated using physical interactions between proteins. Based on the premise of what you sow, so shall you reap; various sources of PPI data have been advocated in different studies. Physical PPIs are inferred through high throughput methods such as yeast two hybrid (Fields and Song), and tandem affinity purification-mass spectrometry (TAP-MS) (Rigaut et al.) (see review by Shoemaker and Panchenko (Shoemaker and Panchenko) for more methods). Multiple databases such as IntAct (Orchard et al.), HPRD (Keshava Prasad et al.), MINT (Licata et al.), dip (Ding et al.) and BioGrid (Chatr-Aryamontri et al.) collect and store information of such interactions from the published literature. Other databases may use text mining approaches to generate putative PPI data, for example the STRING database, which is a meta database which collects data from other primary databases and also predicts putative interactions by text mining (Szklarczyk et al.). Other approaches include the prediction of PPIs based on interactions in homologous proteins (Jonsson and Bates). Further, there are curated pathway databases that give directionality to interactions but also annotate them as belonging to specific functional pathways. Major examples include KEGG (Kanehisa et al., Kotera et al., Tanabe and Kanehisa), Biocarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and Reactome (Croft et al., Haw and Stein).

Prototypical studies which set the stage for integrating cancer genes were more focused on studying the global topological features of cancer genes; including centrality, inbetweenness of nodes, characteristic path length and shortest path length (Jonsson and Bates, Goh et al., Xia et al.) (Figure 1.7).

Studies have also been successful in characterizing disease states by combining gene expression, sequence predictions and literature-based analysis. To identify links between transcription factors (TFs), Tuck et al. (Tuck et al.), constructed a human transcriptional regulatory network, by combining co-expression data with transcription factor-gene regulatory relationships based on sequence predictions and a careful examination of the literature. This approach can potentially identify key TF-gene pairs that show differential activity between diseased and healthy states. Once a list of potential TFs has been obtained, it is possible to build links and generate networks from variations at the DNA level, then simulate by computational means, how these TFs are controlled and regulated in a cell.

However, the view that that most biological networks are composed of strongly connected functional modules (Barabasi and Oltvai) (Figure 1.7), and developments in graph clustering algorithms to identify cancer genes (Bader and Hogue, Reimand et al.), has led to a shift in the trends to study cancer related networks.

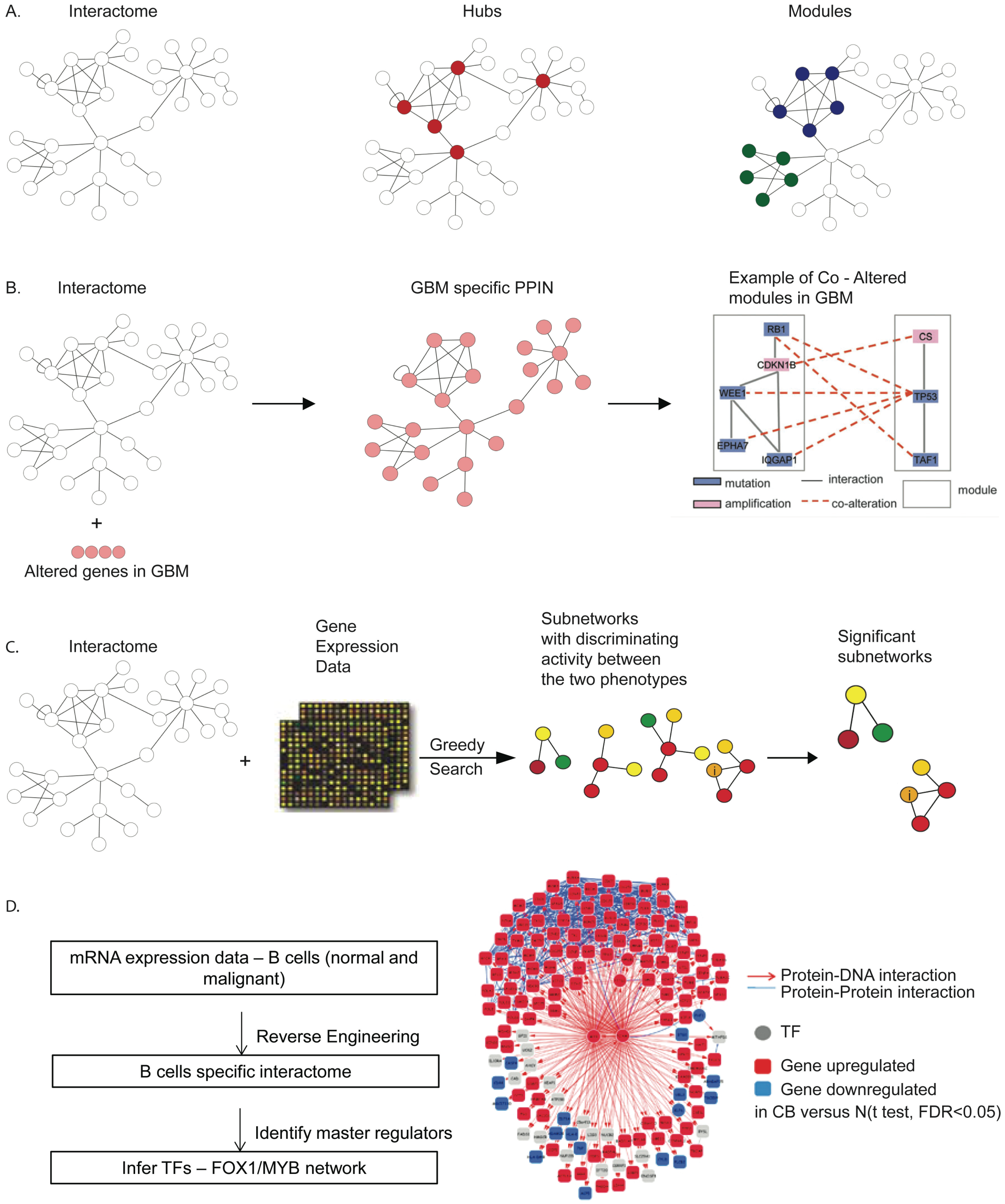


Figure 1.7: Network analysis based workflows to study disease biology
(A) Basic terminology and prototypical analysis as used in PPINs. (B) Methodology as applied by Gu et al. where the authors used a GBM specific PPIN to detect co-altered modules. The right most panel shows an example of such co-altered modules. (C) Methodology as used by Chuang et al. to classify breast cancer metastasis. Gene expression data was mapped onto the human interactome and a greedy search algorithm was applied to identify subnetworks showing maximum differentiating behaviour between the two phenotypes (metastatic and non-metastatic). (D) The left panel shows the main steps in the methodology as used in (Lefebvre et al.). First, the ARACNe algorithm to reverse engineer and generate a B cell specific interactome was applied, followed by the MARINa algorithm to infer master regulators (TFs) for this interactome. This figure is as presented in (Gulati et al.).

One of the most common approaches used to study cancer datasets has been to study characterised pathways/networks – such as signalling and metabolic networks as defined in databases such as KEGG (Kanehisa et al., Kotera et al.) and Reactome (Croft et al., Haw and Stein). Studies have indeed successfully used these databases to identify which curated pathways may be deregulated in specific cancers (Cancer Genome Atlas Research, Guichard et al.). However, it has been pointed out that concentrating only on pre-defined pathways is not sufficient, as many of the known human proteins have still not been assigned to any of the characterised pathways (Wu et al.). In this study, the authors generated a comprehensive PPIN by combining curated pathways with data from protein–protein interactions, gene expression, protein–domain interactions, gene ontology (GO) annotations and text-mined protein interactions. They employed this network and applied graph-clustering algorithms to identify network modules enriched for genes altered in glioblastoma multiforme (GBM). They found two network modules, one formed of genes whose products are localised in the cytoplasm and plasma membrane, and the other with gene products in the nucleus. The authors also found similar network patterns when they analysed data for breast, colorectal and pancreatic cancers.

Ciriello et al. (Ciriello et al.) have advocated that three major characteristics can be attributed to gene modules that may drive cancer progression. Firstly, these modules should be altered with high frequency, secondly genes in these modules belong to the same pathway or biological process and finally the genes exhibit patterns of mutual exclusivity across multiple patients. The authors designed an algorithm named ‘MEMo’ to detect such “candidate driver networks”. They have shown its applicability using the TCGA GBM and ovarian cancer datasets and claim, the affectivity of MEMo in suggesting genetic alterations, which have particularly strong selective effects when applied to any cancer dataset, and also are able to aid the design of drug combinations based on the rationale of ‘synthetic lethality’.

In contrast, another study has developed an algorithm named Dendrix (Vandin et al.), where they find driver modules within genetic alteration data using mutation patterns in patients only. The algorithm uses a Monte Carlo search with two guiding rules to search for driver networks. 1) The genetic alterations (genes) should have high coverage i.e. they should be occurring in a large number of patients and 2) the genes within the driver network should be mutually exclusive.

Another study has developed an algorithm called RME modules (recurrent and mutually exclusive modules), based on the principles of recurrence and mutual exclusivity (Miller et al.). A mutation matrix is generated for the cancer of interest from which a gene exclusivity network was created using the WINNOW tool, which is an online-learning linear threshold method that can effectively detect patterns of exclusivity in a noisy dataset. Using a greedy search algorithm, starting with each gene in the network as seed gene, RME modules were detected.

While the work discussed above has concentrated on detecting individual pathways and modules altered in cancer, a recent paper by Gu et al. (Gu et al.) has presented an approach to search for pairs of co-altered modules in glioblastoma (GBM) patients. They defined co-altered modules by two properties: firstly, each module is a group of genes that are strongly connected with each other with a frequency higher than random; secondly, alterations between module genes show a higher level of co-occurrence than random. They generated a GBM specific network, using genetic alteration profiles for GBM (both somatic mutations and genes altered due to copy number variations), and detected seven co-altered modules within this network. They found that genes occurring within these seven module pairs were significantly enriched with genes from both the F-census database and the Cancer Gene Census database. They also calculated that the average frequency of alteration of module genes was higher than the non-module genes (Figure 1.7).

In a similar fashion, an extension of the Dendrix algorithm, Multi-Dendrix (Leiserson et al.) has been reported. The aim here was again to simultaneously identify driver pathways *de novo* in somatic mutation data. Again, Multi-Dendrix, does not take physical PPIs into account and only relies on mutation patterns.

Another approach, based on tissue-specific expression data, has shown to provide useful indicators of breast cancer outcome (Taylor et al.). In this work, Taylor et al.

annotated a protein with greater than five interactions as a hub, and then divided them into two categories: intermodular and intramodular hubs, the former display a low correlation of co-expression with their interacting partners whereas the latter have high correlation of co-expression with their interaction partners. When they further preformed functional enrichment analysis on these hubs, they found that intramodular hubs have higher similarities with their partners than the intermodular. In addition, intermodular hubs were more associated with global network connectivity.

In recent years, the application of reverse engineering algorithms to generate tissue specific PPINs have gained increasing popularity. For example, an analysis of brain tumours identified two master regulators (transcription factors), namely C/EBP and STAT3, of the mesenchymal transformation pathway for these tumours; inactivation of these genes in mouse xenografts blocked tumour growth and development (Carro et al.). In a second study by the same group (Lefebvre et al.), an analysis of a B-cell interactome identified two genes, MYB and FOXM1, to be master regulators of proliferation in the germinal centre. These studies provided a paradigm for the applicability of interactome analysis for studying normal and pathogenic tissues (Figure 1.7).

Whether we study modules altered on their own or in a co-altered manner, a foremost objective in the field of cancer therapeutics is to extend their applicability as prognostic signatures. Two recent studies have used network analysis approaches to identify such prognostic signatures. The idea behind both studies being that it is not a single gene, but a network of genes that regulates cellular phenotypes, and hence using network analysis to study patient survival should provide better signatures for predicting prognosis. In the first study, Wu and Stein (Wu and Stein) developed a semi-supervised algorithm, which first identifies gene modules involved in disease independent of clinical status, and then applied the supervised principal component method (developed by Bair and Tibshirani (Bair and Tibshirani)) to identify clinically significant modules. When applied to breast cancer data, they found a signature consisting of 31-genes, which could be validated across five independent studies, and when applied to ovarian cancer, the algorithm identified a signature of 75-genes linked to patient survival.

In another study (Zhang et al.), networks were used to complement Cox regression analysis for studying patient survival in ovarian cancer. Their results show that by adding network information the accuracy of predicting survival outcome over using Cox regression on its own is improved. These results were shown to validate over three independent datasets.

Another important application of network analyses methods has been as effective tools to classify patients into meaningful subtypes. Combining gene expression profiles with PPINs (Chuang et al.), a successful framework was developed to differentiate between metastatic and non-metastatic breast cancers (Figure 1.7). According to the authors, using a network-based method has several advantages over differential expression analyses. This is because the resulting subnetworks not only provide models for molecular mechanisms underlying cancer, but also are more reproducible between different cohorts of patients. In addition, typical cancer mutations may or may not be detected through analysis of differential expression but they play a central role in PPINs by interconnecting many crucial genes.

Extending on the argument that network based signatures are much more reproducible than using single gene based signatures, in another study (Hofree et al.), somatic mutation profiles were used to stratify patients into subtypes by clustering patients with mutations in similar pathways/networks together. This method called network based stratification (NBS) was applied to cancer in different tissues, and in each case it was able to identify subtypes that were associated with clinical outcome.

In a different approach, Teschendorff and Severini (Teschendorff and Severini) have studied cancer metastasis by assessing network flux. They argue that integrating PPINs and gene expression data can not only overcome some of the inherent problems associated with microarrays, such as background noise, but also allows distinguishing between direct and indirect protein–protein interactions. They integrated PPI data with gene expression measurements, and by using local entropy measures, showed that the cancer metastasis phenotype displays an increased randomness of local information flux patterns. They conclude that using gene entropy measures on an integrated PPI and gene expression data set can be useful for identifying genes and pathways disrupted in one phenotype with respect to another.

All work discussed here, has been on static PPI networks, however there have been studies showing the application of dynamic networks such as Boolean networks and ordinary differential equations in cancer studies; these networks have been reviewed elsewhere (Cheng et al.), and were not covered as part of the work of this thesis.

1.5 Computational analysis in this thesis

There are two parallel tracks in this thesis and a computational framework has been developed to understand the molecular mechanisms and the prognostic biomarkers underpinning ccRCC. To understand ccRCC mechanisms (Chapter 3), key ccRCC genes were first validated in the TCGA cohort using integrated 'omics' datasets. A comprehensive human protein-protein interaction network (PPIN) was generated by collecting and combining protein-protein interaction (PPI) data from five primary databases namely, IntAct, BioGrid, HPRD, MINT and DIP. A ccRCC specific PPIN was obtained by combining the two datasets. Network properties of ccRCC were assessed and strongly connected subnetworks were detected using the MCODE (Bader and Hogue) algorithm. The Dendrix (Vandin et al.) algorithm was applied to detect *de novo* modules of genes altered in a mutually exclusive manner and the results were compared with the modules detected *via* the MCODE algorithm on the ccRCC specific PPIN. Co-altered modules were obtained by adapting a probabilistic model previously developed by Gu et al. (Gu et al.). Expression based drivers were assessed using the ARACNE (Margolin et al.) and MARINa algorithms (Lefebvre et al.). Lastly, genotype-phenotype relationships were explored using the DriverNet algorithm (Bashashati et al.).

To assess and validate prognostic markers, a statistical survival analyses framework was developed; where logrank and competing risk analyses were performed for univariate analyses and Cox regression was applied to identify independent prognostic markers for ccRCC (Chapter 5).

1.6 Thesis objectives and outline

As discussed in this chapter, recent large scale ‘omics’ studies have led to the identification of key alterations in ccRCC, shown extensive inter-tumour and intratumour heterogeneity, and the possible utility of molecular biomarkers as clinical prognosis tools. However, multiple questions too have arisen from these studies; which somatic mutations/SCNAs drive ccRCC progression, which pathways are driving the aggressive subtype, how is resistance to therapy developing, can molecular markers help to improve patient prognosis and last but not the least, how may ITH affects all of the above. In this thesis, I present two parallel tracks, where a comprehensive integrated analysis of the available ccRCC ‘omics’ datasets is performed, with the aim to better understand the biological mechanisms underpinning this cancer as well as evaluate prognostic biomarkers to improve prediction of patient prognosis. The remaining chapters of this thesis are organised as follows:

Chapter 2: Methods, presents the main methods and software applied in this work. This chapter introduces the two main datasets that were used to study ccRCC. The preliminary data processing before analyses is explained and all the methods and pipelines used in this thesis are detailed.

Chapter 3: The molecular landscape of ccRCC. In order to explore the biological landscape of the cancer, this first results chapter details the analyses of the TCGA dataset – somatic mutations, CNVs and gene expression analysis. Data from Sato et al. (Sato et al.) and Scelo et al. (Scelo et al.) is discussed and compared with the TCGA dataset (The Cancer Genome Atlas Research Network). All datasets are carefully explained using some preliminary analyses. First, each individual ‘omics’ dataset is analysed on its own, following which integrated data analyses are presented. Individual somatic aberrations are investigated, discussing putative drivers and the gene expression signatures potentially controlled by these aberrations. However, as cancer cells target pathways to mutate and not individual genes, this chapter details computational network analyses performed to explore: 1) network properties of ccRCC genes. 2) Patterns of co-occurrence or mutual exclusivity of genetic aberrations. 3) Patterns of co-alterations of networks/pathways.

Chapter 4: The quest for prognostic biomarkers. The above chapter investigates the biological factors contributing to variations in ccRCC. This chapter discusses the clinical impact of molecular signatures, including both genetic and transcriptomic makers for ccRCC, and further questions how ITH impacts the accuracy of these biomarkers. Here each identified molecular marker (n=28), is first assessed individually at the univariate level by 2 tests – logrank and competing risk analyses for association with Cancer Specific Survival (CSS). The validated biomarkers are then assessed in comparison to each other along with tumour stage and grade in a multivariate regression model. This model helped identify a molecular test, which adds further prognostic information to tumour stage. The effect of ITH of the identified biomarker is assessed in a multiregion biopsy cohort.

Chapter 5: Molecular drivers of the ccA/ccB signature. Having identified a molecular signature, which adds further prognostic information to tumour stage, and also shows reliable result despite ITH, this chapter explores this signature and the molecular drivers underlying these expression subtypes. The ultimate idea is to identify targets for adjuvant or immunotherapy for patients, which is as yet are lacking. Having observed clear differences between the ccA and ccB subgroups both at the genetic and transcriptomics levels individually, a more imperative objective is then to find the connections between these two levels and to identify putative drivers for both signatures, thereby identifying targets for therapy.

Chapter 6: Discussion. An evaluation of the work performed in this thesis and conclusions drawn. The level of, importance and implications of ITH for ccRCC are discussed further along with future avenues still to be explored in the ccRCC field.

Chapter 2. Methods

In this chapter, all methods used in the thesis are described in detail. Section 1 details the datasets analysed whilst section 2 covers the preliminary data processing. Sections 3-4 describe the algorithms used to query these datasets in order to understand ccRCC biology. At each stage of the analysis, various methods were considered and evaluated. The method fitting best to the needs of each analysis was applied with typically the method providing the greater functionality being chosen. Sections 5-7 cover the statistical tests used for the prognostic analyses. Section 8 details the tests used to identify and analyse differentially regulated genes. Finally, section 9 describes the statistical tests used at various places throughout the thesis.

2.1 Datasets

There are two main datasets used for the analyses presented in this thesis; the data published by The Cancer Genome Atlas (TCGA) and a multiregion profiling dataset published by our collaborators.

2.1.1 TCGA dataset

The TCGA consortium has molecularly profiled over 400 ccRCC cases for somatic mutations, SCNA, RNA-Seq and clinical data (The Cancer Genome Atlas Research Network). Samples were collected from patients newly diagnosed with ccRCC, undergoing partial or complete nephrectomy and received no prior treatment including chemo- or radiotherapy. There was no bias for sample collection against any surgical stage or histologic grade and staging was performed according to the American Joint Committee on Cancer (AJCC) staging system. According to this system, a tumour tissue can be classified as Stage I-IV, based on the size and extent of the primary tumour, the spread to neighbouring lymph nodes and the presence of distant metastasis (Figure 2.1). Stage I cancers are the least advanced while Stage IV are the most advanced with potential metastatic spread.

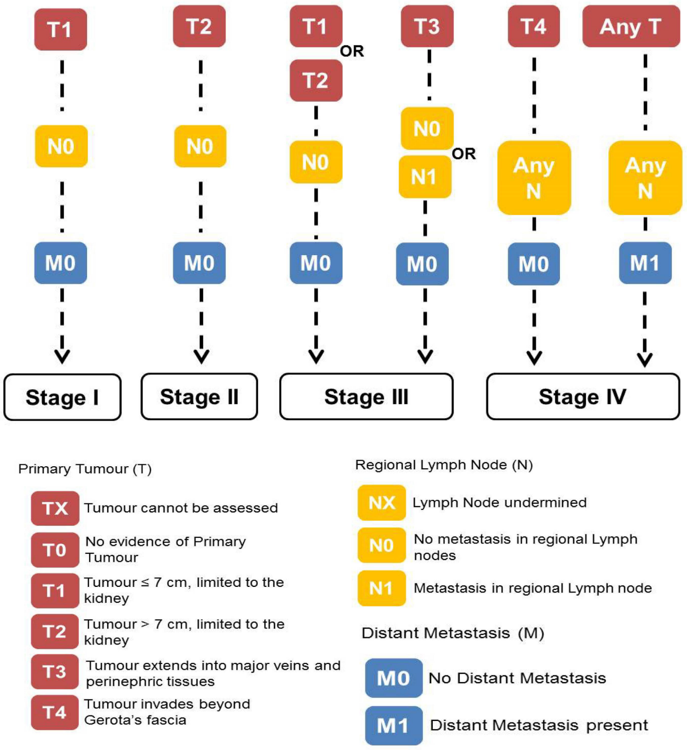

T1
N0
M0
Stage I
T2
N0
M0
Stage II
T1
OR
T2
N0
M0
T3
N0
OR
N1
M0
Stage III
T4
Any N
M0
Any T
Any N
M1
Stage IV
Primary Tumour (T)
TX Tumour cannot be assessed
T0 No evidence of Primary Tumour
T1 Tumour ≤ 7 cm, limited to the kidney
T2 Tumour > 7 cm, limited to the kidney
T3 Tumour extends into major veins and perinephric tissues
T4 Tumour invades beyond Gerota's fascia
Regional Lymph Node (N)
NX Lymph Node undermined
N0 No metastasis in regional Lymph nodes
N1 Metastasis in regional Lymph node
Distant Metastasis (M)
M0 No Distant Metastasis
M1 Distant Metastasis present

Figure 2.1: The AJCC Staging system
Figure explains the assignment of tumour stage I-IV based on the AJCC Staging system, taking into account the size of the primary tumour (T), the extent of necrosis (N) and the existence of metastasis (M).

Histologic grade describes the microscopic appearance of the cancer cells. For kidney cancer, the Fuhrman grading system is used. There are four grades under this system; the higher the grade, the more abnormal the cancer cells look. Similar to the tumour stage, Grade 1 tumours are the most ‘normal’ in appearance and least likely to have spread whereas Grade 4 tumours would be most likely to have spread to metastatic sites.

Normal tissue specimens, where available, comprised either of blood components, adjacent normal tissue from > 2cm away from tumour or previously extracted germline DNA from blood or non-malignant tissue. The contributing tissue source sites were Catholic Health Initiative - Penrose St. Francis Health Services, Catholic Health Initiative - St. Joseph's Medical Centre Cancer Institute, Christiana Care Health Services, Inc., Cureline, Inc., Fox Chase Cancer Centre, Harvard University, International Genomics Consortium, Mayo Clinic, MD Anderson, MSKCC, National Cancer Institute Urologic Oncology Branch, University of North Carolina and University of Pittsburgh.

Data from TCGA has further been divided and referred into multiple datasets; the 'original' dataset comprises data for somatic mutations (n=417), SCNAs (n=450), RNA-Seq (n=469) and clinical data (n=446); the downloaded time stamps were up to and including June 2013. A 'core' dataset of 350 cases was derived from this original dataset, which comprises cases for whom all of the above four information components is available. A further extension of some of these dataset components has been collated for use in certain places; time stamp for downloaded data corresponds to dates since January 2014, and this is referred to as the 'extended' dataset.

2.1.2 Multiregion biopsy dataset

Our collaborators in Prof. Swanton's laboratory have profiled multiple regions from each tumour biopsy for ten ccRCC patients (Gerlinger et al., Gerlinger et al.), comprising of stage T2 (n=2), T3 (n = 7) and T4 (n = 1) cases. Eight out of the ten cases had metastatic disease (stage IV tumours) whilst two (RMH008 and RK26) were stage II patients. Three cases had no treatment prior to nephrectomy, while six (EV001, EV002, EV003, EV005, EV006 and EV007) received everolimus treatment, which is a mTOR inhibitor and one case (RMH002) received sunitib treatment, which is an anti-angiogenic drug. Biopsies from perinephric metastasis (M1) and from a chest wall metastasis (M2a and M2b) were available for EV001. For EV002, a biopsy from a metastasis obtained at the time of disease progression on everolimus treatment was available (M2) (Gerlinger et al.). For EV006, lymph node metastasis (LN1a and LN1b) and a tumour thrombus from the renal vein of RMH004 (VT) were also available (Gerlinger et al.).

Microarray data was generated with Affymetrix Gene 1.0 arrays. For the analyses performed in this thesis both microarray and DNA Sequencing dataset are available for a cohort of 63 regions from the ten patients (GSE31610 and GSE53000) with a minimum of at least four regions per patient.

2.2 Data processing

2.2.1 Somatic mutations

Somatic mutation data was collected from the supplementary material of the TCGA ccRCC publication (The Cancer Genome Atlas Research Network). A mutation was considered to be non-synonymous (non-syn) depending on the value in the 'Variant classification' column (examples include 'indels', 'missense' and 'nonsense' for non-syn mutations and 'silent', 'intron' and '3'UTR' for silent mutations), and was assigned to the mutant patient subgroup for each gene. Mutation frequencies for all genes were calculated based on non-syn mutations only unless otherwise specified.

2.2.2 Copy number data and SCNA profiles

The raw copy number profiles (original dataset) were downloaded and processed by Pierre Martinez in our collaborating laboratory for the original dataset. The aroma R package (CRMA v2, CalMaTe "v1" algorithm & TumorBoost) (Bengtsson et al., Bengtsson et al., Ortiz-Estevez et al.) was used to obtain logR and BAF values from SNP array data that was generated on Affymetrix Genome-Wide SNP Array 6.0 platform by the TCGA, using normal samples as references. Sex chromosomes were excluded from the analysis. The Allele-specific copy number analysis of tumours (ASCAT) algorithm was applied to all 450 samples to obtain copy number profiles (Van Loo et al.) as described in (Martinez et al.). The SCNA data was converted to cytoband level data using the cytoband coordinates retrieved from the UCSC Genome Browser database (http://genome.ucsc.edu/)

(Meyer et al., 2013). For each cytoband a weighted average copy number was obtained, and deletions and amplifications were defined as copy numbers deviating from the ploidy, as estimated by ASCAT, by more than 0.6, similar to the original ASCAT publication (Van Loo et al.).

2.2.3 RNA-Sequencing data

For each gene, raw RSEM (RNA-Seq by Expectation-Maximization) RNA-Seq counts as well as normalised counts, which had been normalised to the upper quartile counts by TCGA consortium, were downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) on 18th September 2012. RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. The files downloaded from TCGA included raw counts per gene per patient and a separate file contained normalised count per gene per patient.

Raw counts were used where differential regulation analysis was performed (Sections 3.3.3.1 and 5.3.1.1). Normalised counts were used, after log2 transformation, for performing clustering analysis (Sections 3.3.3.4 and 5.3). Only genes for which normalized RSEM counts were above 30 in at least 80% of the samples were included in further analyses.

2.2.4 Microarray data

Samples for mRNA expression profiling were collected in Prof. Swanton's laboratory. RNA was extracted from all tumour specimens and corresponding normal kidney specimens where available and quantified on an Agilent Bioanlalyser. Expression profiling was done using the Affymetrix HuGene-1_0-st-v1 platform by an external service provider, for 63 tumour regions and for 6 samples from normal kidney from which sufficient high-quality RNA was available, and data was deposited in the in the GEO repository (GSE31232 and GSE53000). Samples were normalized using the oligo R package and the RMA algorithm in R.

2.2.5 Ploidy

Ploidy is a measure of the number of chromosomes in a cell. Normal cells are diploid - that is they have two sets of each of the 23 chromosomes. Cancer cells can be diploid too but generally tend to have higher ploidy. Aneuploid cells are those that have either too many or too few chromosomes. Aneuploidy cancer cells may be more aggressive than diploid cancer cells. Ploidy estimates for each sample are obtained from ASCAT (Van Loo et al.). The algorithm calculates ploidy as the average total copy number for each sample.

2.2.6 Weighted Genomic Instability Index (wGII)

The wGII (Burrell et al.) score gives a numerical score between 0 and 1 and is the measure of the genomic instability of an individual's genome. It is computed by first calculating for each chromosome of the sample, the proportion of bases whose copy number deviates from the ploidy value of the sample as given by ASCAT by more than 0.6. The sample wGII score is then calculated as sum of the chromosomal scores divided by the number of analysed chromosomes (n=22). This data was obtained from our collaborating laboratory.

2.3 Significantly mutated genes and SCNAs

2.3.1 MutsigCV

To determine significantly mutated genes i.e. which are mutated more often than by chance, the MutSig (Mutational significance) algorithm (Lawrence et al., 2013) was used.

This algorithm was implemented using the standalone package provided by the Broad Institute. A list of mutations identified through DNA sequencing is provided as input to the MutSig algorithm, along with information regarding the coverage of sequencing. The algorithm then models the background mutation processes that may be in play during tumour progression. It then assesses the mutations in each of the input gene and identifies genes, which are mutated more often than random, given the background mutation model.

Several versions of MutSig have been developed since its inception, differing in the manner in which they determine the background mutation rate (BMR); MutSigCV being the most recent and the most sophisticated to date. While MutSig1.0 assumed a constant BMR to be applicable throughout the genome, MutSig1.5, estimated the BMR from assessing the silent mutations in each gene as well as the estimated expression levels of the gene. The TCGA consortium used Mutsig1.5 in their publication. MutSig2.0 added two new measures, namely clustering of mutations in hotspots in a gene and the functional impact of each mutation.

After testing a number of prototype versions, MutSigCV was developed. The "CV" stands for "covariates". Here the BMR is estimated by taking into account information on 'neighbouring' genes in the covariate space. Genes selected to be in the neighbouring space are chosen on the basis of similar genomic properties to the gene under consideration. MutSigCV was applied to the original (n=417) as well as the extended (n=549) datasets and results were compared to those published in the TCGA publication (The Cancer Genome Atlas Research Network).

2.3.2 GISTIC

The results from a SNP array analysis, subsequently fed into the ASCAT algorithm provide a numerical copy number for each analysed cytoband. While this data can be used to assess which cytobands are amplified or deleted (as described in Section 2.2.2), identifying which of these regions maybe associated with the disease of interest presents a non-trivial problem. The GISTIC algorithm (Mermel et al.) was designed to identify regions of the genome, which are significantly amplified or deleted across a set of samples in a given phenotype (disease state). The algorithm works by assigning a so-called G-score to each sample that takes into account the amplitude of its aberration (amplification/deletion) as well as the frequency of aberration across all samples. Using this G-score, False Discovery

Rate (FDR) q-values are calculated for each aberrant region; regions with q-values below a user-defined threshold are considered significant.

GISTIC 2.0 package (Mermel et al.) available on the Broad Institute' GenePattern portal (Reich et al.) was run on SCNA data from TCGA to determine amplifications and deletions of interest in our cohort. The required inputs to the package include a segmentation data file, which provided copy number details for all cytobands for all samples (n=555). It has six columns, namely sample id, chromosome, start position, end position, number of markers for this region and a segmented copy number (CN) which is the log2 ratio of the CN with respect to the sample ploidy were provided. A markers file was also provided which gives the cytoband (markers) names and positions of the cytobands in the original dataset before segmentation. The reference genome used was hg19, the threshold for defining amplifications and deletions was set at the default value of log2 ratio = 0.1; all other parameters were set to their default values.

For the output, GISTIC provides an all lesions file, which gives details of all the significant regions, identified by the algorithm. For each significant region, a “peak region” is identified, which represents the part of the aberrant region, which has the highest amplitude, and frequency of alteration. Additionally, a “wide peak” region is also determined in a leave-one-out manner to allow for errors in the identification of the boundaries in a single sample. These wide peaks are considered to be more robust for identifying the more likely gene targets in the region.

2.4 Network analyses algorithms

As discussed in section 1.4.2 of the Introduction, network analyses methods provide a great medium to study cancer biology. Various state-of-the-art algorithms were either directly applied or adapted for the work presented in this thesis. This section describes these algorithms. In brief, for detection of strongly connected subnetworks in an undirected PPIN, the MCODE algorithm was applied. This method was chosen due to its availability as a computationally inexpensive package in Cytoscape. To compare the results obtained from analysis of the PPINs, the Dendrix algorithm was chosen as it provides a good contrast in terms of obtaining clusters of genes without using background knowledge of its interaction partners. To detect co-altered modules, an adaptation of a probabilistic algorithm by Gu et al. (Gu) was developed. At the time this study was performed, it was one of the few algorithms developed in this area and provided a good foundation for the analysis presented.

At the gene expression level, to detect master regulators, ARACNE and MARINa algorithms were applied. These are well-established algorithms that are widely used in the field. Lastly, to assess if algorithms extending the paradigm of co-relative analyses to test genotype-phenotype relationships could provide additional information, the DriverNet algorithm was applied.

2.4.1 Constructing a Protein-Protein Interaction Network (PPIN)

A protein-protein interactions (PPIs) database was created by downloading and combining data from five databases, namely IntAct (Orchard et al.), MINT (Licata et al.), DIP (Xenarios et al.), HPRD (Keshava Prasad et al.) and BioGRID (Chatr-Aryamontri et al.) (all downloaded before September 2013). These primary databases collect potential PPIs from the scientific literature; these PPIs have been experimentally verified using methodologies such as yeast two-hybrid, mass spectrometry and co-immunoprecipitation. MySQL was used to generate the combined database of interactions. Each row of the database corresponded to one protein-protein interaction and along with the protein-id information was extracted on the method of detection, the source of interaction (given as PubMed id), and the database from which it was extracted. Interactions were extracted from this database for ccRCC altered genes to generate a ccRCC specific PPIN.

2.4.2 Detecting strongly connected subnetworks

Module or subnetwork detection was performed using the Molecular Complex Detection (MCODE) (Bader and Hogue, 2003) algorithm. The algorithm uses vertex weighting by local neighbourhood density and traverses outwards from a locally dense seed protein to isolate the dense regions according to given parameters. There are several advantages of this algorithm; first it has a directed mode that allows fine-tuning clusters of interest without considering the rest of the network, secondly it allows examination of cluster interconnectivity, which is relevant for protein networks. Lastly, it is claimed not to be affected by a known high rate of false positives in data from high-throughput interaction techniques, which are generally applied to generate protein-protein interaction data.

2.4.3 Dendrix

To examine mutual exclusivity patterns of mutations independent of protein-protein interaction information the Dendrix algorithm was applied. Dendrix (De novo Driver Exclusivity) (Vandin et al., 2012) is an algorithm for discovery of mutated driver pathways in cancer using only mutation data. The algorithm applies a Monte Carlo search to find sets of genes mutations in which exhibit both high coverage and mutual exclusivity in the analysed samples.

2.4.4 Co-altered modules

To explore co-alteration of pathways, a probabilistic model (Gu et al., 2013) was applied. The algorithm was coded in the programming language Python. It takes into account the likelihood of co-occurrence of genetic alterations in patients and combines it with a network search algorithm to identify co-altered modules in a background gene interaction network. The ccRCC specific PPIN generated above was used as the background. The probability of co-occurrence of genes was calculated using Equation 2.1:

$$P_{(G1,G2)} = 1 - \sum_{k=0}^{a-1} \frac{\binom{a+b}{k}\binom{c+d}{a+c-k}}{\binom{n}{a+c}} \qquad \textbf{2.1}$$

Where G1, G2 are the modules, n is the number of all samples, a is number of samples with alterations in both modules, b is the number of samples with alterations only in G1, c is the number of samples with alterations only in G2 and d is the number of samples with alterations in none of the modules.

The score for each module pair is then calculated as the negative logarithm of the probability (Equation 2.2):

$$S_{between}(G1, G2) = -\ln[P_{(G1,G2)}] \tag{2.2}$$

2.4.5 Detecting master regulators at the Gene expression levels: MARINa algorithm

The MARINa (Master Regulator Inference Algorithm) (Lefebvre et al., 2010) algorithm was applied to find gene expression based regulators. First, a gene expression network needed to be generated for which the ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) algorithm (Margolin et al., 2006) was used. ARACNE reverse engineers a network from gene expression data using an information theoretic approach, to calculate mutual information between gene pairs. This approach has been described as effective to eliminate majority of indirect interactions, which are typically inferred by pairwise analysis. The ARACNE standalone implementation was downloaded and used to generate the network. The input to the tool is a gene expression matrix with each row representing a gene and columns correspond to patients. An adjacency matrix is retrieved as output with a probability of interaction between gene pairs in each cell.

The R package ‘viper’ was used to implement the MARINa algorithm. The adjacency matrix obtained from the ARACNE algorithm along with the original gene expression matrix is provided as input. The algorithm is designed to infer transcription factors (TFs) controlling the transition between the two phenotypes – ‘normal’ and ‘tumour’ in this work, as well as the maintenance of the latter phenotype.

2.4.6 DriverNet

Extending on the genotype-phenotype relationship detection, the DriverNet algorithm was also applied (Bashashati et al.). This algorithm integrates genetic alterations (mutation and copy number variation data) and transcriptome data (gene expression data) to predict functional important driver genes in cancer. This algorithm makes a bipartite graph with the genetic alterations on one side of the graph and the corresponding gene expression data on the other side (as discussed in Introduction section 1.4.1). It then uses an influence graph, which is a gene-gene interaction network derived from pathway data. A greedy algorithm is then applied to find the possible driver genes (based on frequency of alteration), which may be pushing the gene expression values of the connected genes to some extreme values.

2.5 Classification of patients into prognostic groups

In Chapter 4, a number of variables were assessed for association with ccRCC prognosis. This section describes how patients were classified into subgroups for each variable under consideration.

2.5.1 Somatic mutations

Somatic mutations in five genes, namely VHL, PBRM1, SETD2, BAP1 and TP53 were analysed. CSS was assessed for patients with tumours harbouring a non-synonymous mutation in the gene vs. patients with tumours without the mutation. For VHL, association with survival was also tested under the specific constrains of only being in stage I-III tumours and for those with loss of function mutations only (defined as frameshift and nonsense mutations), as compared to all other patients.

2.5.2 SCNAs

A total of 14 candidate SCNAs were assessed for association with CSS. CSS was compared between patients with tumours harbouring a specific SCNA vs. those with tumours without these SCNAs. Amplification or deletion of ≥50% of a chromosome arm or of both arms of a chromosome was considered to be equivalent to an arm level alteration as described (The Cancer Genome Atlas Research Network) or to a whole chromosome aberration respectively. For Chrom22 deletion (identified as a candidate prognostic biomarker), the SNP array data did not include any probes for the Chrom22 p-arm, thus deletion of Chrom22q was used as a substitute measure.

2.5.3 Gene expression RNA-Seq data

Log2 transformed expression data was used to divide the cohort into 2 groups at median values for CD31 and EDNRB expression levels and at 33rd percentile value for TSPAN7 expression levels (Wuttig et al.).

For all cluster-based analyses performed in this thesis, the Non-negative matrix factorisation (NMF) clustering (Brunet et al.) method has been used. Consensus NMF clustering uses the principle of dimensionality reduction, using non-negative matrix factorisation, to find a set of metagenes from the given gene list; it then uses these metagenes to perform the clustering of samples. The metagenes are defined as a positive linear combination of the genes for which expression has been provided in the dataset submitted as input. It repeatedly runs the algorithm against variations of the gene expression data and creates a consensus matrix to assess the stability of the resulting clusters.

The Consensus NMF clustering package provided on the Broad Institute' GenePattern portal (Reich et al.) was used to implement this method. The input to the algorithm is the gene expression matrix, where rows are genes and columns represent patients and thereby each cell represents the expression for the gene for the respective patient. Log2 normalised RSEM counts were provided as input for this analysis. The cluster number (k) range was set between $2 \leq k \leq 10$. The number of clustering's to be tested was set to 50 and 1000 iterations per clustering were performed for each k value.

Expression data for genes in each identified gene expression signatures (Kosari et al., Zhao et al., Lane et al., Brannon et al., Beleut et al.) was submitted for consensus NMF clustering analysis. Gene expression was available for 26 out of 35 genes (74%) from (Kosari et al.), 220 out of 259 genes (85%) from (Zhao et al.), 36 out of 44 (82%) genes from (Lane et al.), 103 out of 110 (94%) genes from (Brannon et al.) and 37 out of 48 (77%, Cluster B vs. A/C) and 21 out of 23 (91%, Cluster A vs. C) genes from the two gene panels from (Beleut et al.) respectively. The cluster number range was predefined from two to ten. Each clustering run returned a cophenetic correlation coefficient, which measures the stability of cluster assignments, as well as consensus clustering maps. Based on both these results, the optimal numbers of clusters for each gene expression panel were identified. For each signature, the same number of clusters was considered to be optimal as had been identified in the original publications.

For the TGFβ pathway signature (Bostrom et al.), a TGFβ activity score for each sample was defined as follows. RNA-Seq counts were available for 145/157 TGFβ regulated genes, the log2 normalised RSEM expression counts for these genes were multiplied by either 1 or -1, depending on their expected regulation by TGFβ. These values were then averaged to give a relative TGFβ score for each sample. Using the median score of all samples as the cut-off, patients were divided into TGFβ low activity and TGFβ high activity cohorts as previously described (Bostrom et al.).

2.5.4 Analysis of multi-region biopsy data: classification of tumour regions into ccA and ccB prognostic groups

Expression data was available for 107 out of 110 genes from the ccA/ccB signature (Brannon et al.). This data was used to classify the 63 tumour regions into either ccA/ccB expression subgroups by applying consensus NMF clustering analysis for a predefined number of clusters from two to ten. The cophenetic coefficient was highest for two clusters. Clustering was also performed using the Clearcode34 panel (Brooks et al.) and the same cluster assignments were obtained for 61 out of 63 (97%) regions.

2.6 Clinical statistics for biomarker assessment

One of the main questions asked in this thesis and presented in Chapter 4, is querying suitable biomarkers for predicting patient prognosis. To this end, the analyses query the effect of a biomarker on the patient cohort. A number of statistical measures in clinical studies were used, which are described here. All tests were implemented in R v 3.0.1 (R Development Core Team) using the packages 'survival' (Therneau and Grambsch, Therneau), 'cmprsk' (https://cran.r-project.org/web/packages/cmprsk/index.html) and randomforestSRC (Ishwaran and Kogalur, Ishwaran et al.).

2.6.1 The Kaplan-Meier estimate

The Kaplan-Meier (KM) estimate (Kaplan and Meier) is the most prominent test used to estimate the effect of a variable on the survival of patients. It measures the length of time patients lived after a predefined point in time. For example, it may be used to estimate the length of life of patients after receiving a particular treatment. For cancer studies, the end point of interest is either relapse of disease or death due to disease. Mathematically, the KM estimate measures the probability of surviving a given length of time while considering time in many small intervals (Equation 2.3).

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) \qquad \textbf{2.3}$$

Where S(t) is the probability of survival at time t, and F(t) represents the cumulative frequency distribution of the random variable T.

However, certain points need to be considered here; apart from the cases which may or may not reach endpoint (death), there will be cases for which follow-up will be lost due to known or unknown reasons, such as death due to other unrecorded causes, loss of contact or unwillingness to maintain contact. In such cases, the patients are 'censored' at the point of last contact. Three main assumptions form the basis of the KM estimate; first, that at any point in time, survival probability is similar between cases that are censored and those that continue to be followed up. Secondly, the chances of survival of any patient are independent of the time of recruitment of the patient into the study cohort. Finally, the event happens at the time recorded in the study.

2.6.2 Logrank test

The logrank test (Clark et al.) is one of the most routinely used measures to assess the effect of either a drug or a prognostic biomarker in clinical studies. While the KM estimate discussed above is the best measure to estimate survival probability, it is not capable of comparing multiple survival curves and assess if they are significantly different.

Figure 2.2 clearly shows two survival curves to be different; however, the objective here is to assess the statistical significance of such differences. The logrank test is a widely used statistic to achieve this objective. It tests the null hypothesis that there is no difference between the KM curves for the different populations under consideration i.e. the probability of an event (here a 'death' event) is not different between the populations at any given point of time (Bland and Altman).

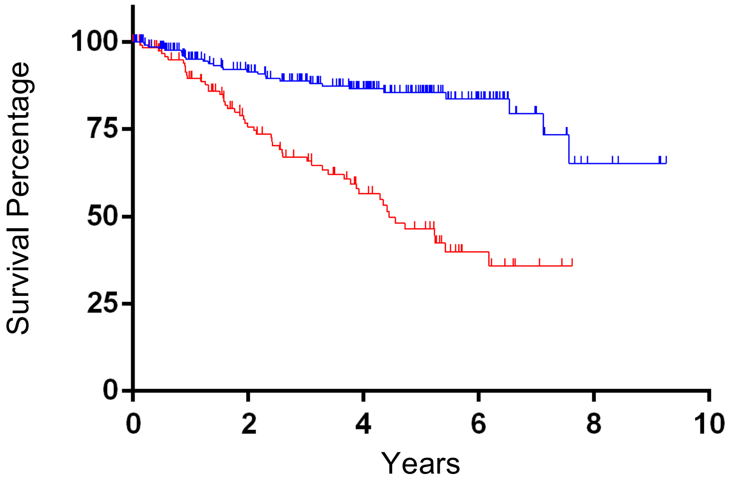


Figure 2.2: Kaplan-Meier survival curves
Figure depicts example survival curves to be compared and the need for a statistical test to distinguish between them

A definitive advantage of using the logrank test in this setting is the fact that it is based on the same assumptions as the KM estimate; i.e. censoring is unrelated to prognosis, the probability of survival is independent of the time of recruitment of the subjects and the events took place at the recorded time points (Bland and Altman).

Logrank test was performed using the survdiff() function from the package ‘survival’ in R to estimate the p-value of significance for differences between the KM curves.

2.6.3 Hazard ratio

In a survival analysis, the hazard ratio (HR) refers to the rate of risk (death) due to the explanatory variable with respect to a reference state. It is often used in clinical trials to measure the survival at any point of time for patients in a particular group (e.g. treatment) with respect to the placebo group. A HR = 1 means that there is no difference in the survival of the two groups being compared. HR > 1 means the risk of death is higher in the considered group, while HR < 0 means the patients in the reference group are at a greater risk of death.

Since the logrank test is a test of significance only, the effect of a variable using this test cannot be measured. For this purpose, a Cox proportional hazard model is applied. This model assumes that the hazard or risk of a factor is constant throughout the study period and calculates the hazard for a variable using the formula shown in equation 2.4; mathematically, if O_a and O_b are the observed number of events in the two groups a and b under consideration, and E_a and E_b are the expected number of events, the HR can be then calculated as:

$$HR = \frac{O_a/E_a}{O_b/E_b} \qquad \textbf{2.4}$$

The null hypothesis here will be that there is no difference between the HRs of the two groups. The coxph() function from the package ‘survival’ in R (R Development Core Team) was used to estimate the HRs and 95% C.I. for each variable.

2.6.4 Competing risk analysis

While the logrank is a widely used test, a limitation of the test is that the censoring is uninformative i.e. it cannot take into consideration the reasons for censoring. For example, for the work presented in this thesis, death due to cancer is the event of interest; thus for the purposes of the logrank test, all other patients, whether alive or dead due to other causes, were treated as censored at the time point where the information about them was last recorded. This may lead to overestimation of the effect of a biomarker. To avoid such overestimation, competing risk (CR) analysis was performed for all variables that were observed to be significant in the logrank analysis.

A CR analysis evaluates the cumulative incidence of the variable under consideration (Equation 2.5). It also takes into account the death of patients due to causes other than cancer. The cumulative incidence of the variable k represents the proportion of patients at a time t who have died from cause k, while accounting for the fact that patients can die from other causes.

$$C_k(t) = \int_0^t h_k(u|X)S(u)\mathrm{d}u \qquad \mathbf{2.5}$$

Where h_k represents the cause-specific hazard, X is the vector of covariates and S is the overall survival function. The CR test was implemented using the cuminc() function from the ‘cmprsk’ package in R and used to estimate the p-value of significance.

2.6.5 Multivariate Cox regression analysis

To assess the independence of the variables that were validated in the univariate analyses (logrank and CR analyses), and to question whether any of these variables added further information to the clinical variables in use, a multivariate Cox regression analysis was performed. A Cox model (Cox) calculates the hazard at any time point t as a function of the baseline hazard and the coefficient of regression β of each variable X in the model:

$$h(t) = h_0(t)exp(\beta x_1 + \beta x_2 + \beta x_3 + ... + \beta X_n) \quad \textbf{2.6}$$

The backwards-stepwise selection model was used. All variables, which had a p-value ≤ 0.05 in the univariate analyses, were added to the model along with the clinical variables Tumour stage and Fuhrman grade. An initial p-value was generated for all variables in the model, following which the variable with the worst (highest) p-value was iteratively removed from the model, until all variables in the model had a p-value ≤ 0.05. The hazard ratio, 95% confidence interval (C.I.) and p-value was noted for all the significant variables, whilst for all non-significant variables, the hazard ratio, 95% confidence interval (C.I.) and the p-value was generated at the step it was removed. The coxph() function from the package 'survival' in R was used to implement the backwards-stepwise regression model.

2.7 Supervised learning algorithms

2.7.1 Recursive partitioning – classification trees

Decision tree based algorithms provide a useful extension to survival analysis methods. These algorithms can be applied to build stratification models where patients are classified using the statistical tests described above to guide the splitting process, into distinguishable groups. Recursive partitioning for single decision trees and Random Forests for cross-validated stratification modelling are two of the most commonly applied and robust methods for classification and regression problems and were therefore used in this study.

Recursive Partitioning (RP) (Banerjee et al.) was used to build a stratified patient prognosis model. This is a tree-based analysis, which can be used to classify patients into different cohorts based on given input parameters. The response variable in this work was a survival object i.e. time to death and a R implementation of RP using the function 'ctree()' in the package 'party' (Hothorn et al.) was used. Briefly, the algorithm tests the null hypothesis for independence between the explanatory variables and the response. It stops if the hypothesis cannot be rejected; otherwise, it selects the input variable with strongest association to response. It then implements a binary split based on this variable. This step is recursively repeated. For survival response, the split is based on maximising the likelihood ratio of survival. The input to this function was a matrix where each row represented a patient and columns represented the input parameters along with the days to death/last follow up of patients and dead/alive status of the patients. Unlike Cox analysis, RP is adept in uncovering variables that may be largely operative within a specific patient subgroup but may have minimal effect or none in other patient subgroups.

2.7.2 Random forest

The random forest (RF) (Breiman) method consists of a supervised learning algorithm that is commonly used for both classification and regression problems. It was used in this thesis to implement a classification model and to test the significance of the variables most important to distinguish between the two classifications. It uses an ensemble of trees to decide the classifications where each tree is generated on the principle of recursive partitioning. For classification problems, the final prediction is a majority vote of all the trained decision trees. The 'Random' aspect of the RF algorithm is related to the way it builds each decision tree. For a training set of N samples, sampling with replacement is performed and two thirds of this sample is used as the training set for a given decision tree in the forest. The other one third (termed as the oob (out-of-bag) data), is used to get an unbiased estimate of the test error and for variable importance measures. The second randomization involved in the RF's decision trees is that at each node, not all features are available for making a split. Rather random samples of 'mtry' features are chosen at each node and the best split is chosen amongst them. An important aspect of the RF is that the test error is reduced with more accurate and less correlated decision trees. Part of the randomization procedures employed in the tree building are in fact aimed at introducing variability in the hope of achieving low correlation between decision trees. The mtry parameter is therefore central to the RF method. Given a powerful descriptor in the set of features, for high mtry values, it is more likely that this descriptor would be chosen in the random sample and subsequently used at the node split. Therefore this descriptor would dominate most of the trees, resulting in highly accurate trees but with low correlation. If the mtry parameter is set too low, then the powerful descriptor might be missed out from most of the trees. The RF would then consist of low correlation trees but with low accuracy. Though this parameter is the one to which the RF is most sensitive, it has a broad range of optimal values (Breiman).

RF Variable Importance Measure: Once the random forest has been built and the oob error estimate for each tree recorded, the importance of each feature to the prediction is measured as follows. For each feature m, all of its values are randomly permuted and the oob examples are fed through the trees with m randomly permuted. The importance score of feature m is the different between the original oob error estimates, and the new ones with m permuted. The importance score is then normalized by the standard deviation of these differences across all trees. Large values imply more important features.

Random forest was implemented using the R package 'randomforestSRC'. The function rfsrc() was used to build trees with all default parameters and selecting a 1000 trees to build the random forest. The Variable importance was then determined using the function vimp(), on the results obtained from the random forest prediction.

2.8 Differential expression analyses

Testing for differential expression between phenotypes (tumour vs. normal) enables the identification of the genes or pathways that may help define the phenotype. Differential expression analyses were performed at two levels; first to identify top deregulated genes (either up- or down-regulated) in ccRCC patients when compared to normal kidney samples. These genes were then assessed using overrepresentation analyses. In the second pipeline, Gene Set Enrichment analysis was performed using the complete set of genes for which RNA-Seq data was available. The details and differences between the two methods are discussed below.

2.8.1 Differentially regulated genes

To identify genes most significantly deregulated in ccRCC, differential expression (DE) analysis was performed for tumour samples over normal samples (Chapter 3) and between ccRCC subgroups (Chapter 5) using the edgeR (Robinson et al.) package in R.

The input to the package included two files; the first was a data matrix where the columns represented patients (j) and the genes represented rows (i) and thus every cell (ij) gave the raw RNA-Seq RSEM count for that gene for the respective patient. The second file was a phenotype file, which defined the patient's subgroup (tumour, normal or ccRCC subgroup). This file was then used in the package to specify which subgroups should be used to make the comparisons.

The edge R package is based on negative binomial distribution for count data. The algorithm estimates gene-wise dispersions and uses an empirical Bayes process to shrink this dispersion towards a consensus value. The differential expression is then assessed by an adaptation of Fisher's exact test for over-dispersed count data.

The output from the final DE estimation function gives three values for each gene that was assessed; the log fold change (FC), the log counts per million (CPM) and the p-value. These p-values were corrected using the False discovery rate (FDR) correction using the p.adjust() function in R. Using a fold change cut-off of $|\log FC| \geq 2.5$ and a FDR $q \leq 0.05$, the final list of differentially regulated genes was obtained.

2.8.2 Overrepresentation analyses

A Gene ontology (GO) or pathway overrepresentation (ORA) or enrichment analysis refers to an analysis which tests how significantly overrepresented are certain GO terms or pathways in a list of genes than if a similar list was chosen at random. Essentially each gene in the list is assigned to a term (GO term or pathway), and all genes belonging to the same term are then collected together;

this collection is compared with a random gene list of the same size to assign a p-value of significance to the enrichment of this term.

For this purpose, both a hit list (from the given gene list) and a population list (random list from all genes available) are compiled, and then the aim is to assess the significance of the difference between these two lists. The most common approach to test this statistically is by using the hypergeometric test (or its variants such as Fisher's exact test) to calculate the probability of seeing at least a particular number of genes containing the biological term of interest in the gene list. ORA analysis has been implemented in this thesis using tools available from MSigDB (Liberzon, Liberzon et al.) and genego portal (Thomson Reuters, https://portal.genego.com/).

For MSIGDB, for the pathway ORA, KEGG (Kanehisa et al., Kotera et al., Tanabe and Kanehisa), Biocarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and Reactome (Croft et al., Haw and Stein) databases were used to compile the background population lists. In the case of GO enrichment, only the biological processes (BPs) terms were used to compile the background. The top 100 pathways overrepresented in the input list (as compiled by DE analysis) were returned.

For the genego portal, for GO ORA, only the BP category as above was used. For the pathway ORA, pathway annotations from KEGG, Biocarta, Reactome were used. All pathways or GO processes which passed the FDR $q \leq 0.05$ were returned.

It was thought prudent to consider two sources, firstly for the sake of validation, and secondly since a longer/more extensive list was required and MSigDB only returns the top 100 hits; therefore, the genego portal was also used.

2.8.3 Gene set enrichment analysis

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., Mootha et al.) is a computational method that was developed to determine whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states, for example tumour samples and normal kidney samples in the work presented here. It is different from the ORA described above in the manner that a GSEA uses the complete list of genes assessed, irrespective of FCs and p-values, to find significance of overlap between deregulated genes and the population gene sets.

The stand-alone version of GSEA tool provided by the Broad Institute was used to run the analysis. For the purposes of this thesis, pre-ranked GSEA was performed. This means that edgeR was run, and using the log FC from the output, the complete list of genes was subsequently ranked. This ranked list was provided as input for GSEA.

The GSEA output has two lists; one is a list of pathways which are up-regulated i.e. show more positive FCs than seen by chance and the other is the list of down-regulated pathways. Using FDR q ≤ 0.05, a list of significant pathways for each comparison was compiled.

2.9 Statistical measures

2.9.1 Multidimensional Scaling

Multidimensional Scaling (MDS) is a method to visualise the similarities and differences between individual samples in a dataset. The R implementation of MDS, cmdscale() was used to visualise the differences between tumour and normal samples in the RNA-Seq original dataset (n=469). A distance matrix was provided as input, which was calculated as the Euclidean distances between the cases based on the gene expression values.

2.9.2 Wilcoxon test

The Wilcoxon (Wilcox) test is a non-parametric test, which can be used when comparing the distribution of repeated measurements of a numerical variable between two groups such as disease vs. normal. It tests the difference between the median observations between the two groups where the null hypothesis is defined as being no difference between the medians. It is used as an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed. It has been used at various points in this thesis to assess the differences between patient subgroups for various numerical variables. The wilcox.test() function in R was used to implement this test and to estimate a p-value of significance.

2.9.3 Fisher’s test

The Fisher's exact test is used to test the independence of two nominal variables. It estimates whether the proportions of one variable are significantly different depending on the values of the other variable. It has the definitive advantage over other tests, such as chi-square, since it can be used for smaller sample sizes as well. It has been used in this thesis to test the relationships of occurrence between mutations, SCNAs, patient subgroups and genetic alterations. The fisher.test() function in R was used to perform the Fisher’s exact test.

Chapter 3. The molecular landscape of ccRCC

3.1 Introduction

As discussed in the Introduction section 1.2, despite recent advances in the detection of recurrent mutations and somatic copy number alterations (SCNAs) in ccRCC, our understanding of the pathogenesis of the disease is still limited. While nephrectomy has shown to be curative for localised disease, relapse is not infrequent. ccRCC is highly resistant to chemotherapy and radiotherapy and therapeutic success is limited. To develop suitable therapies, understanding the disease mechanisms is imperative. Recent large-scale sequencing and expression profiling efforts provide the opportunity to comprehensively investigate genotype-phenotype correlations. This should improve our understanding of ccRCC biology and potentially shed lights on what drives ITH. Researchers at The Cancer genome atlas (TCGA) have comprehensively measured and provided genomic, epigenetic, expression as well as clinical data for over 400 ccRCC patients. This dataset is based on single biopsies, and thus they overlook ITH; however, they have large numbers of patients, allowing for correlative analysis.

In this chapter, with the aim to better understand ccRCC biology, somatic mutation, SCNA and RNA-Sequencing (RNA-Seq) data sets, molecularly profiled by the TCGA (The Cancer Genome Atlas Research Network) are investigated and a series of integrated analyses performed upon them. Wherever possible, comparisons were also made with other datasets identified in the literature (Sato et al., Scelo et al.), enabling some key insights into ccRCC mechanisms to be drawn.

In the TCGA publication on ccRCC (The Cancer Genome Atlas Research Network), a number of key disease associated factors were highlighted including mutations in oxygen sensing (VHL) and chromatin modifying genes (for example PBRM1), down-regulation of metabolism related pathways and recurrent mutations in the PIK3CA/MTOR pathway. Here these findings are expanded upon by making use of the availability of multiple "omics" datasets thereby providing the opportunity to perform integrated analyses of genetic and transcriptomic data to understand some of the functional consequences of the genetic mutations observed.

Heterogeneity in cancer, both inter-patient and intratumour, were discussed in sections 1.2.5 and 1.4 of the Introduction. This presents a hurdle for quantitative correlative analyses between expression signatures and individual somatic mutations. The big question now, is how can this high volume of primary information, collected at both the genetic and phenotypic levels, be integrated to help cancer patients? Computational network algorithms have become increasingly popular as useful tools for the integration and interpretation of these complex datasets to study cancer mechanisms. In this work, advantage is taken of key concepts and state-of-the-art algorithms, which are then applied to both genetic and transcriptomic ccRCC datasets.

3.2 Methods

All the methods and algorithms applied in this chapter are briefly explained in the following sections; references to more detailed descriptions, Chapter 2, are provided in each section.

3.2.1 Data processing

Somatic mutations

Somatic mutation data for the original dataset was obtained from the supplementary material of the TCGA publication for ccRCC (The Cancer Genome Atlas Research Network). A gene was considered to be mutated based on the classification in the ‘VARIANT TYPE’ column. Mutation frequencies were calculated based on non-synonymous mutations only.

Gene expression data

For both the core and extended datasets RNA Sequencing (RNA-Seq) data was downloaded from the TCGA data portal. Both raw counts generated by the RSEM method and normalised RSEM counts, normalised to the upper quartile by TCGA, were downloaded from the data portal (https://tcga-data.nci.nih.gov/tcga/) on 18 September 2012. Raw and normalised datasets have been used for different analysis in this chapter as described in the relevant sections. Normalised counts were log2 transformed before further analyses. In either case, only genes, for which the counts (raw and normalised resp.) were above 30 in at least 80% of the samples, were included in all analyses.

3.2.2 Significantly mutated genes

MutsigCV (Lawrence et al.) was run to assess significant mutations from the TCGA mutation matrix. The input data to MutSigCV is the list of mutations in all the samples for which DNA sequencing is available. It builds a model of the background mutation processes, and analyses the mutations of each gene to identify genes that were mutated more often than expected by chance, given the background model. (Methods section 2.3.1)

3.2.3 SCNAs significantly associated with ccRCC

To determine which SCNAs are associated with ccRCC biology, GISTIC (Mermel et al.) was run on the copy number calls from ASCAT (as described in Methods section 2.2.2). For this analysis an extended cohort of 555 cases for which SCNA data was available was used. This analysis was done in collaboration with Dr Peter Van Loo. GISTIC requires a segmented copy number profiles file, where each line represents the copy number of a particular chromosomal region (represented in coordinates) for each patient. Another file mapping these coordinates to specific regions is also provided (Methods section 2.3.2).

3.2.4 Multidimensional scaling

Multidimensional scaling (MDS) was applied on the gene expression RNA-Seq data to determine and visualise the differences in expression between normal and tumour samples. The R implementation of MDS, the function cmdscale() was used and a distance matrix was provided as input to estimate similarities and differences. (Methods section 2.9.1)

3.2.5 Differential regulation analysis

The edgeR (Robinson et al.) package in R (R Development Core Team) was used to find differentially regulated genes. The edgeR output provides three different values for each gene in the input, namely log FC, log CPM, and the p-value for the significance of deregulation. All p-values can then be corrected for multiple testing using the function p.adjust() in R (R Development Core Team). A list of significantly differentially regulated genes can then be generated using p-value and/or fold change cut-offs. (Methods section 2.8.1)

3.2.6 NMF clustering

The top 1500 genes showing the maximum variation (high standard deviation) across all tumour samples were used to perform NMF clustering to find the optimal number of subgroups for ccRCC (Methods section 2.5.3).

3.2.7 Overrepresentation analyses

Using MSigDB (Liberzon et al., Liberzon) and the genego portal (Thomson Reuters, https://portal.genego.com/), GO and pathway overrepresentation analyses were performed for the genes showing high levels of deregulation between tumour vs. normal kidney samples. For the pathway overrepresentation analyses, only pathways defined in KEGG, Reactome and Biocarta were used (Methods section 2.8.2).

3.2.8 Gene set enrichment analysis

The Gene Set Enrichment Analysis (GSEA) (Subramanian et al.) algorithm was applied to gene expression data to determine pathways/gene sets deregulated in ccRCC. This method tries to determine whether an a priori defined set of genes shows statistically significant concordant differences between two biological states; here ccRCC samples vs. normal kidney (Methods section 2.8.3).

3.2.9 Statistical analyses

Fisher's exact test was used to estimate co-occurrence or mutual exclusivity of genetic alterations relative to each other. All analyses were performed in R (R Development Core Team) version 3.0.1.

3.2.10 Generating a ccRCC specific protein-protein interaction network

A protein-protein interaction database was created by downloading and combining data from five databases, namely IntAct (Orchard et al.), MINT (Licata et al.), DIP (Xenarios et al.), HPRD (Keshava Prasad et al.) and BioGRID (Chatr-Aryamontri et al.). These primary databases collect potential PPIs from the scientific literature, which have been experimentally verified using methodologies such as yeast two-hybrid, mass spectrometry and co-immunoprecipitation. Interactions for genes altered in ccRCC were extracted from this database to create a ccRCC protein-protein interaction network (PPIN) (Methods section 2.4.1).

3.2.11 Detecting sub-networks

Module or subnetwork detection was performed using the Molecular Complex Detection (MCODE). This algorithm gives a weight to each node and uses neighbourhood density to grow a subnetwork from each tested seed node (Methods section 2.4.2).

3.2.12 Dendrix

Dendrix (De novo Driver Exclusivity) (Vandin et al.) is an algorithm for discovery of mutated driver pathways in cancer using only mutation data. It finds sets of genes, domains, or nucleotides whose mutations exhibit both high coverage and high exclusivity in the analysed samples (Methods section 2.4.3).

3.2.13 Co-altered modules

To identify co-altered modules likely to drive the growth of ccRCC, a probabilistic model (Gu et al.) was implemented, which takes into account the likelihood of co-occurrence of genetic alterations in patients, and combines it with a network search algorithm to identify co-altered modules in a given gene interaction network (Methods section 2.4.4).

3.2.14 ARACNE and MARINa

To generate a gene expression based network and find its regulators, the ARACNE (Margolin et al., Basso et al.) algorithm was applied. This algorithm reverse engineers a network by calculating mutual information between associated genes. Log2 normalised RSEM counts for RNA-Seq data were provided as input. The output network from ARACNE, provided as an adjacency matrix, was used as input to the MARINa (Lefebvre et al.) algorithm, which estimates the master regulators of a gene expression network. The MARINa algorithm was implemented in R using the ‘viper’ package (Methods section 2.4.5).

3.2.15 DriverNet

The DriverNet algorithm was used to assess genotype-phenotype relationships. This algorithm uses bipartite graphs and a background gene-gene interaction network to find cancer drivers at the genetic level that may explain the corresponding gene expression level changes (Methods section 2.4.6).

3.3 Results

3.3.1 Somatic mutations

In the original dataset, somatic mutation data was available for 417 patients, spanning mutations in a total of 10401 genes. These included 2,389 insertions/deletions, 16,821 missense mutations, 1149 nonsense mutations and 6,383 silent mutations.

Figure 3.1 shows the frequency of mutations of all genes in this dataset. Characteristic of cancer, a long tail of mutation frequency is observed. VHL, PBRM1, SETD2, BAP1 and JARID1C/KDM5C were observed to be the top five most recurrently mutated genes. However, taking only the frequency of mutation events is no longer considered to be the most appropriate method when assessing the most important cancer drivers (Lawrence et al.). It is imperative to take into account other factors, such as the background mutation frequency rate and gene size.

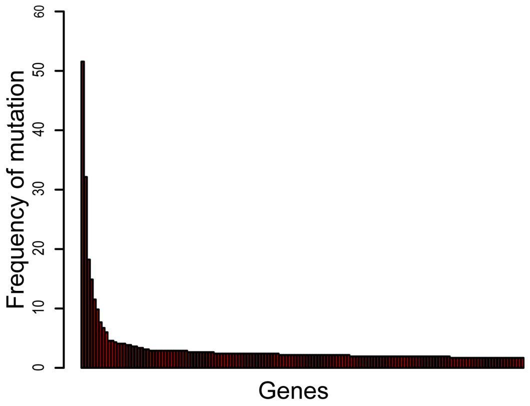
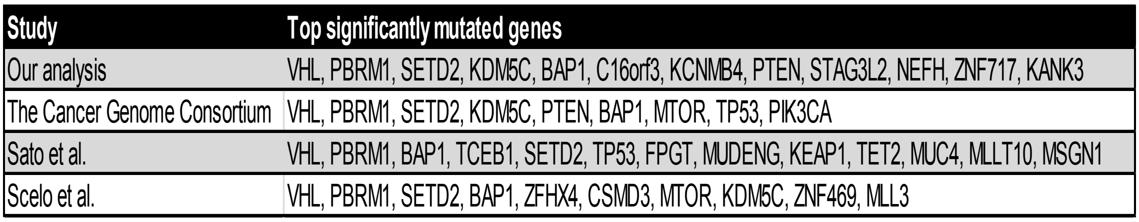


Figure 3.1: Somatic mutations in ccRCC
A Frequency bar chart for the top 100 genes with somatic mutations in ccRCC in the TCGA dataset. The most frequently mutated genes were VHL (53%), PBRM1 (33%), MUC4 (20%) SETD2 (12%), and BAP1 (10%). Beyond this Frequency of mutations was ≤ 10% for all genes.

MutSigCV (Lawrence et al.) was used to assess the most significantly mutated genes in this cohort. The TCGA ccRCC publication had used an older version of this same algorithm (see Methods section 2.3.1 for a comparison of the two versions). Comparing both these results as well as published literature, 11 genes namely VHL, PBRM1, SETD2, KDM5C, BAP1, MTOR, TP53, PTEN, PIK3CA, ARID1A and HMCN1 formed the top targets for further investigation. Other studies published at the same time (Sato et al.) and recently (Scelo et al.) observed 28 (Sato et al.) and 17 (Scelo et al.) genes to be significantly mutated in their study cohorts. Table 3.1 shows the genes assessed as significant in this analysis (FDR ≤ 0.05), and in the other studies (The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.).

Table 3.1: Significantly mutated genes in major ccRCC studies
This table shows the top significantly mutated genes obtained in this analysis in comparison to the 3 other major studies. As seen major genes such as VHL, PBRM1, SETD2 were observed in all analyses; however, each analyses showed additional genes such as STAG3L2, NEFH (this analysis) and MLL3 and CSMD3 (Scelo et al.)



Study	Top significantly mutated genes
Our analysis	VHL, PBRM1, SETD2, KDM5C, BAP1, C16orf3, KCNMB4, PTEN, STAG3L2, NEFH, ZNF717, KANK3
The Cancer Genome Consortium	VHL, PBRM1, SETD2, KDM5C, PTEN, BAP1, MTOR, TP53, PIK3CA
Sato et al.	VHL, PBRM1, BAP1, TCEB1, SETD2, TP53, FPGT, MUDENG, KEAP1, TET2, MUC4, MLLT10, MSGN1
Scelo et al.	VHL, PBRM1, SETD2, BAP1, ZFHX4, CSMD3, MTOR, KDM5C, ZNF469, MLL3

Interestingly VHL had a much higher frequency of mutations in the Scelo (Scelo et al.) cohort when compared to the TCGA (The Cancer Genome Atlas Research Network) and Sato (Sato et al.) datasets. The Scelo study was performed on a European cohort including patients from Czech Republic, Romania, Russia and the United Kingdom, whereas the TCGA cohort is a USA based dataset and the Sato study had patients from Asia. Differences in frequency were also observed for other genes; KDM5C showed lower frequencies of mutation in the Sato cohort, while MTOR showed higher frequency of recurrence in the Sato dataset. Furthermore, certain genes seen to be mutated with moderate to high frequencies in one study were not seen in the other (Figure 3.2). These data may indicate the role of race in ccRCC pathogenesis.

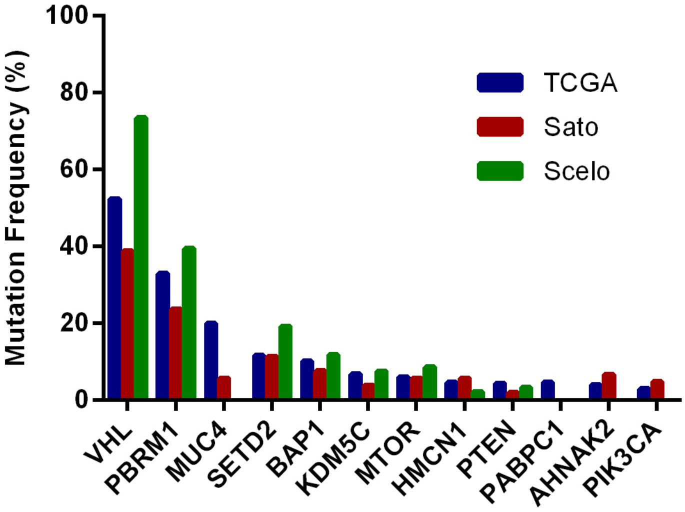


Figure 3.2: Comparison of frequency of mutations in the three data cohorts
Figure shows a bar chart depicting a comparison of the frequency of mutation for a few example genes in the three data cohorts (The Cancer Genome Atlas Research Network, Sato et al., Scelo et al.). As shown, mutations in the VHL gene were observed at much higher frequency in the Scelo cohort. Likewise, SETD2 gene was also mutated at higher frequency in the Scelo cohort. MUC4 and PIK3CA genes while observed to be mutated in both TCGA and Sato cohorts were not seen to be mutated in the Scelo cohort.

As discussed in section 1.2.4 of the Introduction, the PIK3CA/MTOR pathway has been shown to be an important pathway for ccRCC pathogenesis. Mutations in this pathway lead to an up-regulation in HIF1A levels, thereby giving a VHL independent route to HIF regulation (Zhong et al., Brugarolas et al., Thomas et al.). The MTOR gene was observed to be mutated in all three datasets at varying frequency; 6% (The Cancer Genome Atlas Research Network, Sato et al.), and 8.5% (Scelo et al.). When other genes in this pathway were assessed, the overall mutation frequency of the MTOR pathway associated genes was observed to be 17% in the TCGA dataset (Figure 3.3). Convergent evolution on the MTOR pathway by mutations in different genes of the pathway within the same patient has also been observed (Fisher et al.). These results highlight the fact, that assessing mutations in individual genes by themselves does not elucidate the complete picture mechanistically and pathways/networks should be assessed together. This idea is further explored from section 3.3.4 onwards.

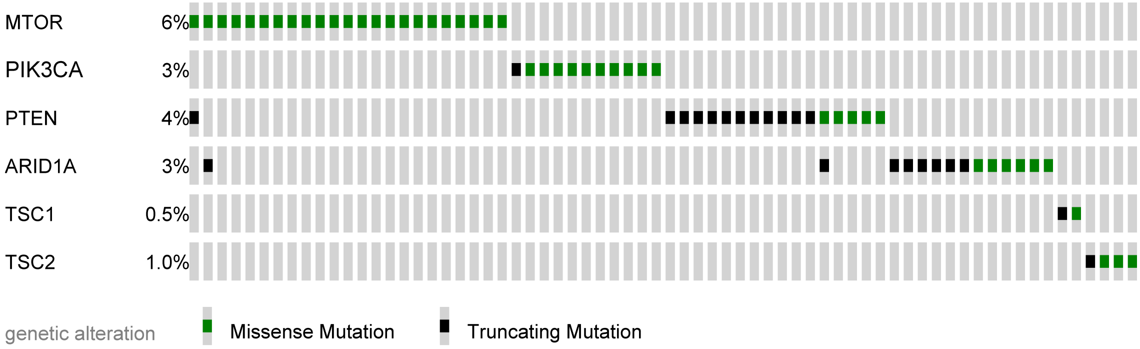


Figure 3.3: Somatic mutations in the MTOR/PIK3CA pathway
Figure shows a bar chart depicting somatic mutations occurring in the MTOR/PIK3CA pathway per sample in ccRCC. The MTOR gene is the most frequently mutated in this cohort, however, taken together, mutations in this pathway cover ~17% of cases. The figure was generated using the cbio portal tool (http://www.cbioportal.org/) and has been zoomed in and cropped to remove unaltered cases.

Next, to examine relationships between key mutations (n=11), co-occurrence and mutual exclusivity patterns were explored using Fisher's exact tests. Mutations in the VHL gene were observed to be co-occurring with mutations in the PBRM1 gene. PBRM1 mutations showed a tendency of co-occurrence with SETD2 mutations as well, while showing a tendency of mutual exclusivity with BAP1 mutations, which agrees with previous published data (Kapur et al.). Other genes were not seen to show any statistically significant associations.

Lastly, to assess the impact of intratumour heterogeneity (ITH), mutation frequencies were compared in the TCGA cohort (T2-T4, VHL mutated cases only) and the multiregion biopsy cohort (Gerlinger et al.). Apart from VHL and PBRM1, all other genes were observed to be subclonal events in the multiregion biopsy cohort. Most striking differences were observed for mutations in the TP53 gene, which was observed to be mutated in only 5% of the cases in TCGA as compared to 40% of the samples in the multiregion biopsy cohort (Table 3.2).

Table 3.2: Comparison of frequency of mutations in key genes for ccRCC between the TCGA and multiregion biopsy cohorts
The table compares the frequencies of mutations of key genes in ccRCC between the TCGA (n=102) and the multiregion biopsy cohorts. As can be seen while similar frequencies are observed when considering each multiregion biopsy as an individual sample (Columns 1 and 2 of the table). However, if all multiregion biopsies from each patient are considered as 1 sample (Column 3), much higher frequencies are observed for most genes, the most prominent example being TP53 and BAP1 genes.

Gene	Prevalence in TCGA samples (n=102)	Prevalence in all M-Seq regions (n=79 regions)	Prevalence in total M-Seq based cases (n=10)
PBRM1	42%	39%	60%
SETD2	18%	27%	30%
BAP1	21%	23%	40%
KDM5C	7%	11%	10%
TP53	5%	6%	40%
ATM	3%	4%	10%
ARID1A	6%	1%	10%
PTEN	5%	10%	20%
MTOR	9%	8%	10%
PIK3CA	3%	4%	20%
TSC2	2%	6%	10%

3.3.2 Somatic copy number alterations

After processing the raw copy number data into copy number calls using ASCAT, amplifications and deletions were determined as copy numbers deviating from the ploidy (estimated by ASCAT), by more than 0.6, similar to the original ASCAT publication (Van Loo et al.). Gistic (Mermel et al.) was applied to these copy number profiles to assess which of these cytobands may be significantly associated with ccRCC. Amplifications in 20 cytobands and deletions of 27 cytobands were observed to be significant in this cohort of 555 cases (Figure 3.4). These results were in concordance with those observed by the TCGA and previously published by others (Beroukhim et al.), though the exact focal positions differed from one dataset to the other (In the TCGA publication GISTIC was run on ~450 cases, whereas the data presented here is for the extended cohort of 550 case).

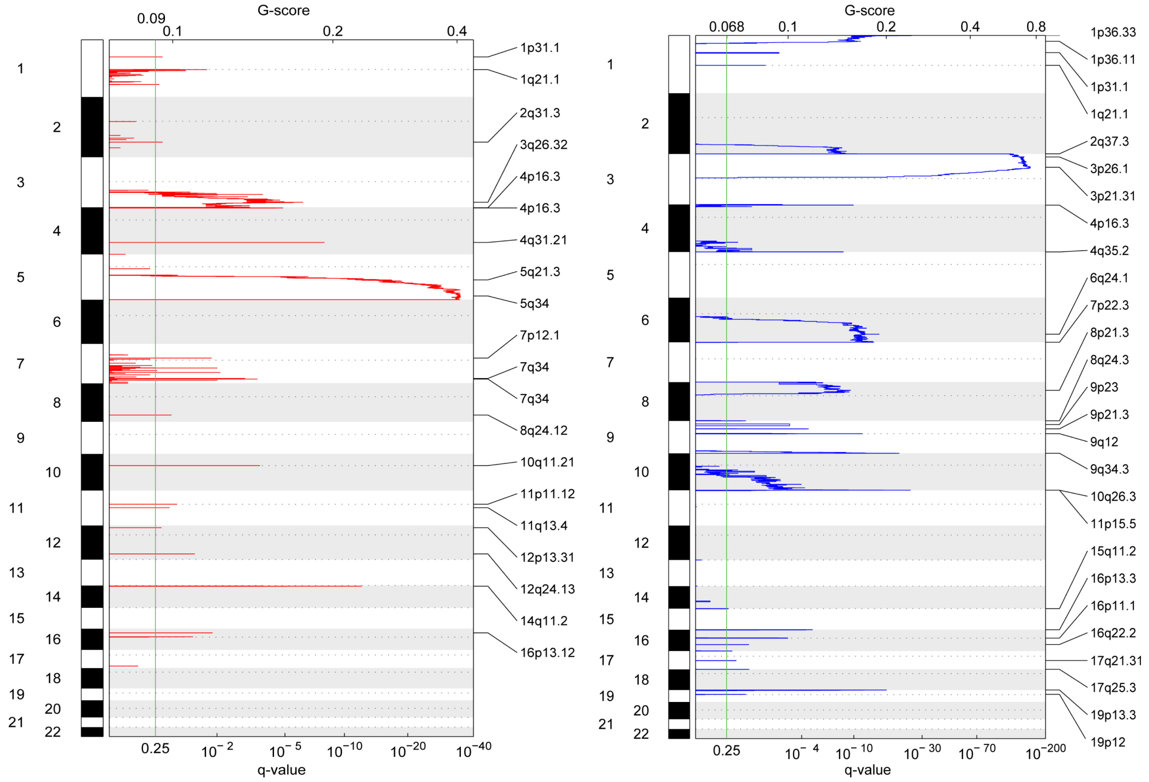


Figure 3.4: GISTIC analysis of copy-number changes in ccRCC tumours. The left panel of the figure depicts the significant amplifications (red) and the right panel depict the significant deletions (blue). The G-score represents the frequency average amplitude of the aberrations identified in the SNP arrays. False discovery rate q values, representing the statistical significance associated with these scores with correction for multiple testing, are displayed on the bottom axis of the figure. Regions with q values < 0.25 (green lines) were considered to be significantly altered. Chromosome positions are indicated along the vertical axis with centromere positions indicated by dotted lines. The locations of the peak regions of maximal copy-number change are annotated on the right.

To identify the cis-effects of these SCNAs, gene expression data was evaluated; to determine the genes showing the corresponding expression changes (up-regulation for amplifications and down-regulation for deletions). Table 3.3 details the frequency of these SCNAs and gives the examples of putative targets on these bands. Prominent kidney cancer genes such as ARID1A and SMARCA5 (chromatin regulators), and other cancer genes such as BRCA1, RET and BRAF were observed to be altered. This analysis requires further follow up with published literature as well as experimental validation to find evidence for the role of these putative target genes in ccRCC.

Table 3.3: Summary of SCNA drivers
This table shows the significant peaks (SCNAs) as observed in the GISTIC analysis. Column 3 gives the frequency of alteration of each peak. Column 4 and 5 give the number of putative target genes on each peak and some example genes respectively. Putative targets were assessed by considering the corresponding gene expression data.

Status	Peak	Frequency(%)	Number of putative targets	Examples
Amplification Peak 1	1p31.1	3	3	AK5, PTGFR, TTLL7
Amplification Peak 2	1q21.1	13	15	CHD1L, FAM72D
Amplification Peak 3	2q31.3	14	3	CERKL, CWC22
Amplification Peak 4	3q26.32	13	4	KCNMB2, KCNMB3
Amplification Peak 5	4p16.3	4	0	
Amplification Peak 6	4p16.3_1	6	16	FAM53A, POLN
Amplification Peak 7	4q31.21	5	5	SMARCA5, RNF150
Amplification Peak 8	5q21.3	43	3	PJA2, EFNA5
Amplification Peak 9	5q34	57	6	CCNG1, GABRB2
Amplification Peak 10	7p12.1	30	3	COBL, DDC
Amplification Peak 11	7q34	30	0	
Amplification Peak 12	7q34_1	31	23	BRAF, CASP2
Amplification Peak 13	8q24.12	13	4	MTBP, TAF2
Amplification Peak 14	10q11.21	6	2	RET, TMEM72
Amplification Peak 15	11p11.12	8	0	
Amplification Peak 16	11q13.4	7	4	P2RY2, P2RY6
Amplification Peak 17	12p13.31	20	41	CLEC2B, MLF2
Amplification Peak 18	12q24.13	20	11	TRAFD1, RPL6
Amplification Peak 19	14q11.2	9	0	
Amplification Peak 20	16p13.12	18	5	PARN, ERCC4
Deletion Peak 1	1p36.33	25	26	FAM132A, C1orf159
Deletion Peak 2	1p36.11	20	31	FCN3, ARID1A
Deletion Peak 3	1p31.1	14	14	SLC44A5, ACADM
Deletion Peak 4	1q21.1	10	8	NBPF15, FCGR1C
Deletion Peak 5	2q37.3	12	29	C2orf54, HDAC4
Deletion Peak 6	3p26.1	84	5	SUMF1
Deletion Peak 7	3p21.31	85	93	SETD2, AMT
Deletion Peak 8	4p16.3	14	27	CTBP1
Deletion Peak 9	4q35.2	16	3	KLKB1
Deletion Peak 10	6q24.1	27	8	TXLNB
Deletion Peak 11	7p22.3	6	1	LFNG
Deletion Peak 12	8p21.3	27	26	CHMP7, C8orf58
Deletion Peak 13	8q24.3	12	60	C8orf31
Deletion Peak 14	9p23	23	3	NFIB
Deletion Peak 15	9p21.3	24	5	KLHL9, MLLT3
Deletion Peak 16	9q12	27	0	
Deletion Peak 17	9q34.3	33	33	NOTCH1
Deletion Peak 18	10q26.3	23	11	PPP2R2D, BNIP3
Deletion Peak 19	11p15.5	9	1	IGF2
Deletion Peak 20	15q11.2	14	5	SNURF
Deletion Peak 21	16p13.3	10	9	CLDN9
Deletion Peak 22	16p11.1	7	0	
Deletion Peak 23	16q22.2	7	2	HYDIN
Deletion Peak 24	17q21.31	7	17	BRCA1, ADAM11
Deletion Peak 25	17q25.3	10	6	TSPAN10, AATK
Deletion Peak 26	19p13.3	15	9	PLIN4, PLIN5
Deletion Peak 27	19p12	6	1	ZNF208

3.3.3 Gene expression analyses

In the original dataset, RNA-Seq data was available for 469 tumour samples and 68 normal samples. Of these, 65 were matched normal and tumour samples, 3 were unmatched normal samples and 404 were unmatched tumour samples. Matched samples had sequencing data from both normal kidney and kidney cancer tissues in the same ccRCC patients, whereas unmatched samples only have sequencing data from either disease or no disease tissues.

To decide if it was suitable to compare all tumour samples to all normal samples, multidimensional scaling (MDS) was performed (Figure 3.5). As all tumour samples were sufficiently clustered together, and were clearly distinguishable from the normal samples, it was deemed acceptable to groups all tumour samples vs. all normal samples for all further analyses.

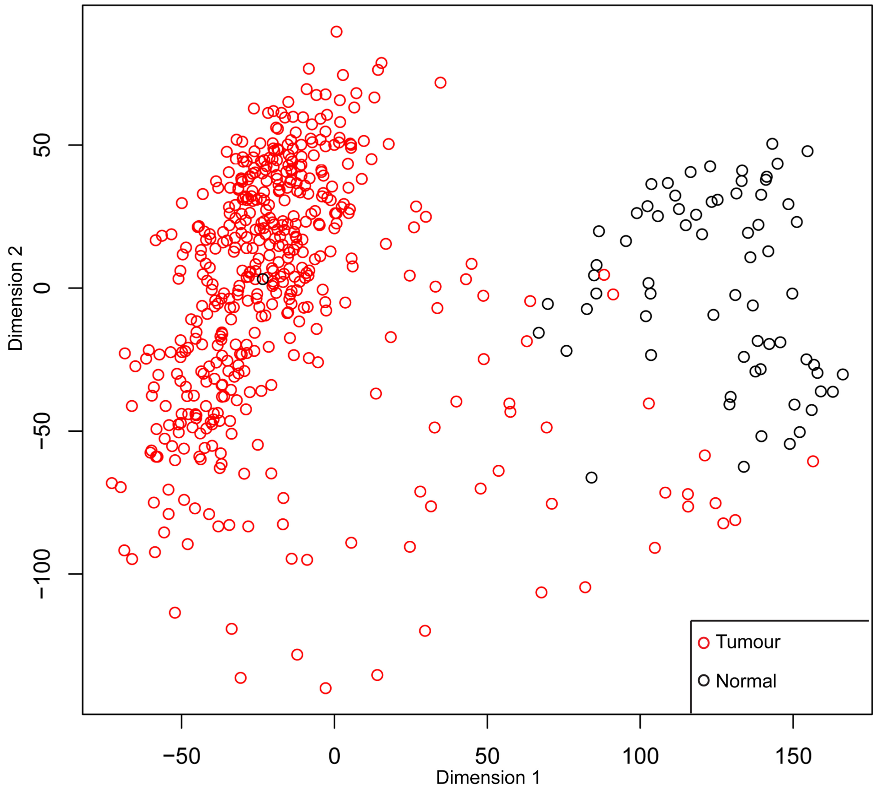


Figure 3.5: MDS plot for RNA-seq gene expression data for ccRCC
MDS plot comparing ccRCC tumour and normal tissue samples. 469 ccRCC tumour samples were available (red) along with 68 normal kidney tissue samples (black).

3.3.3.1 Differentially regulated genes

To obtain genes differentially regulated between the normal and tumour samples, the edgeR (Robinson et al.) package in R (R Development Core Team) was used. Using FDR q ≤ 0.05 and fold changes (FCs) of |log FC| ≥ 2.5 (equivalent to |FC| =5.6) as selection criteria, a list of 867 genes was obtained as significantly deregulated in ccRCC when compared to normal kidney samples.

Genes showing the highest FCs included DOC2A, LPPR5, BIRC7, HP and MYEOV showing high up-regulation and UMOD, SCL12A1, DUSP9, KCNJ1 and KNG1 showing high down-regulation. The biological roles of these genes are discussed in the following sections.

3.3.3.2 GO and pathway enrichment for significantly deregulated genes

GO overrepresentation analyses showed enrichment for genes involved in signal transduction, immune system processes, developmental processes and cell surface receptor linked signal transduction, and apoptosis related genes as top terms (Appendix A). Pathway enrichment analyses showed similar results; specifically cytokine receptor interaction genes, transmembrane transport, immune system genes and metabolism related genes represented the top pathways (Appendix 0).

3.3.3.3 Gene Set Enrichment Analysis (GSEA)

While an overrepresentation analyses as performed above enabled highlighting the major processes under play in ccRCC patients, a gene set enrichment analysis (GSEA) was performed to get a deeper understanding of the deregulation of these processes. All genes for which RNA-Seq data was assessed were included. A pre-ranked GSEA was performed whereby a list of genes ranked according to FC was provided as input. The top pathways up-regulated in the ccRCC cohort included pathways belonging to immune cell regulation processes such as cytokine-cytokine receptor interaction and allograft rejection. The down-regulated pathways included genes belonging to metabolic pathways, namely valine, leucine, isoleucine degradation, propanoate and fatty acid metabolism.

3.3.3.4 NMF Clustering reveals the existence of at least two subgroups for ccRCC

NMF clustering using the top 1500 most variable genes revealed two major subgroups of ccRCC. The cophenetic coefficient was highest for k=2 and k=3, but then dropped rapidly. The consensus clustering matrices also supplemented this result. The two subtypes contained 301 (Group 1) and 168 (Group 2) samples (Figure 3.6).

This result is in concordance with recent work (Brannon et al.), which showed two subgroups for ccRCC, namely ccA and ccB subgroups. These two subgroups were shown to have a different prognosis, with ccA patients having a better survival than ccB. Using the classifier panel devised in the publication, 305 cases were classified as the ccA and 164 were classified as the ccB subgroup. 88% of the samples overlapped between Group 1 above with ccA classification and Group 2 with ccB, the mismatch for the rest can be explained with the possible existence of a third subgroup which has been discussed above. Since the ccA/ccB classification scheme has been shown to have prognostic significance, it has been used as the classification of choice for the purpose of this thesis. This work has been discussed in greater details in the results chapters, Chapter 4 and Chapter 5.

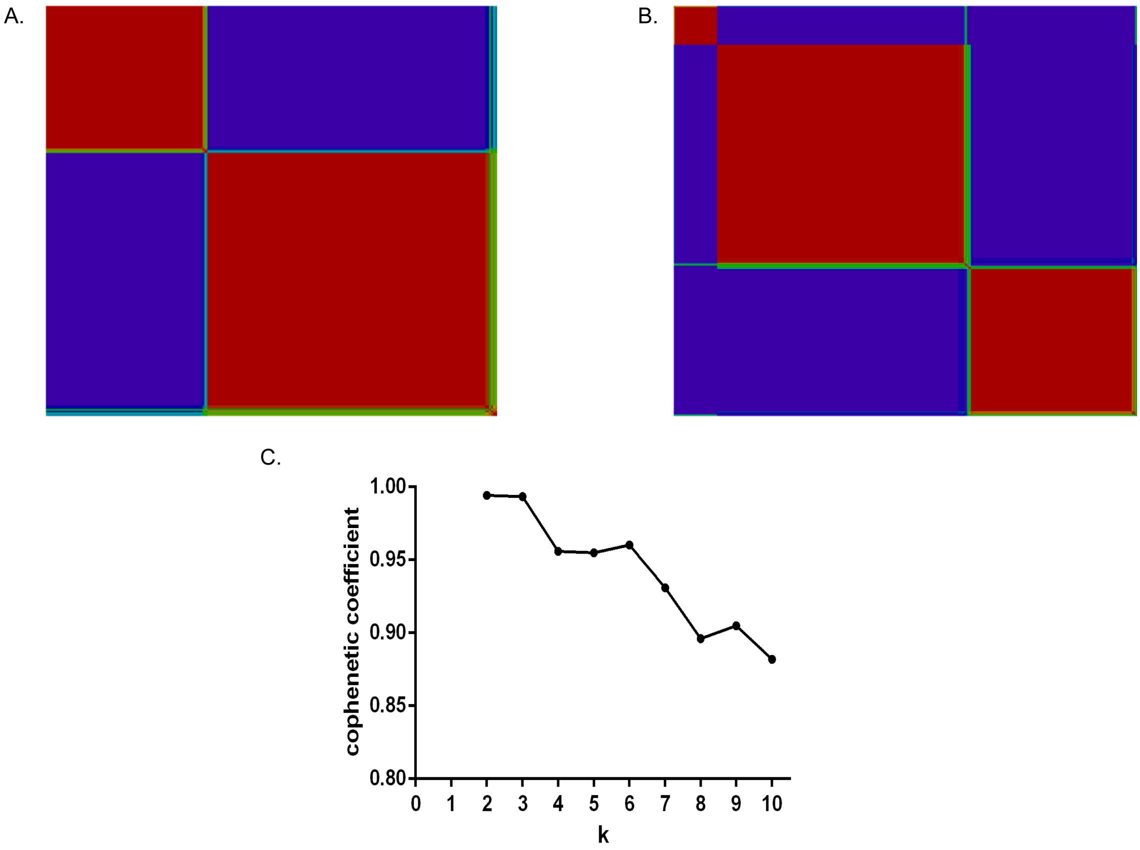


Figure 3.6: NMF clustering plot using the top 1500 most variable genes in the ccRCC cohort
Consensus NMF clustering heatmaps depicts the stability of consensus clustering assignment into clusters using the top 1500 most variable genes for ccRCC based on RNA-Seq data. A. Ordered Consensus map for k=2. B. Ordered Consensus map for k=3. C. Cophenetic coefficient plot for k=2 to k=10. The coefficient was highest for k=2 and k=3. Both the coefficient and the clustering maps suggested the existence of two ccRCC subgroups and potentially a small third subgroup of ccRCC tumours (top red square in panel C).

3.3.4 Network properties of ccRCC genes

To study the network properties of ccRCC genes, all genes with somatic mutations were considered. These genes were then mapped to the PPIN to study network properties.

Closely connected sub-networks (clusters) were determined for the ccRCC specific network using MCODE (Bader and Hogue). A total of 11 clusters were obtained with a degree cut-off of 4. Out of the 11 high confidence drivers, only five genes namely VHL, PBRM1, ARID1A, BAP1 and TP53 were observed to cluster. The VHL gene was part of a 29-gene cluster (Figure 3.7A). A total of 11 genes in this cluster are ribosomal protein encoding genes or are involved in RNA metabolism. A few others were implicated in immune system related pathways while 11 genes could not be assigned to any particular pathway.

PBRM1 and ARID1A are part of SWI/SNF complex. They were observed to be interacting with other genes in a larger complex of 147 genes (Figure 3.7B). A set of ~17 genes were implicated to be involved in the spliceosome and another large cluster of genes were identified to be immune system regulated genes. Multiple other chromatin modifiers were also observed to be part of this cluster. Both BAP1 and TP53 genes were observed as part of a large cluster of 352 genes. The genes from this cluster are involved in transcription, cell cycle and immune system related pathways.

Another cluster of 27 genes, which did not contain any apparent drivers (high or low confidence), was seen to be variably mutated in ~10% of the patient samples (Figure 3.8). These genes are ribosomal proteins involved in metabolism and translational processes. This observation might indicate that alternate pathways control multiple ccRCC mechanisms. Another smaller set of genes annotated as passengers, which were observed to be mutated in all together 5% of cases, seems to be part of a cluster (separate from the previous one) interacting with genes involved in ribosomal processes (the 42 gene cluster). While it is hard to comment on the relative significance of these modules, these data may indicate yet unthought-of pathways that may be playing a part in ccRCC biology. All 11 clusters are provided in appendix C.

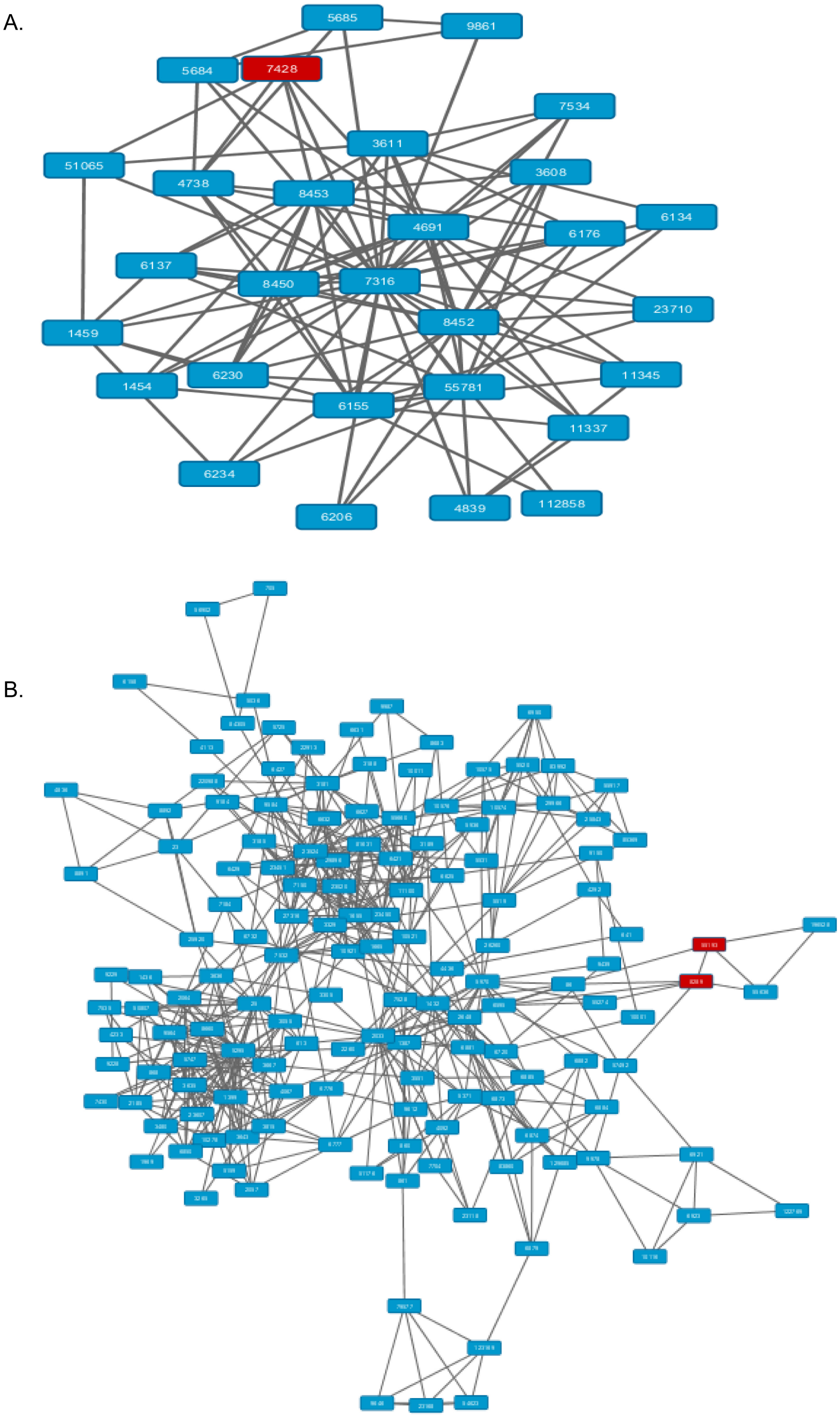


Figure 3.7: Strongly connected subnetworks within the ccRCC PPIN
Two example clusters (subnetworks) detected as part of the ccRCC specific PPIN. A. Depicts the 29-gene cluster which consisted of the VHL gene (red). B. Depicts the 147-gene cluster, which consisted of other chromatin modifiers such as ABL1 and CREBBP.

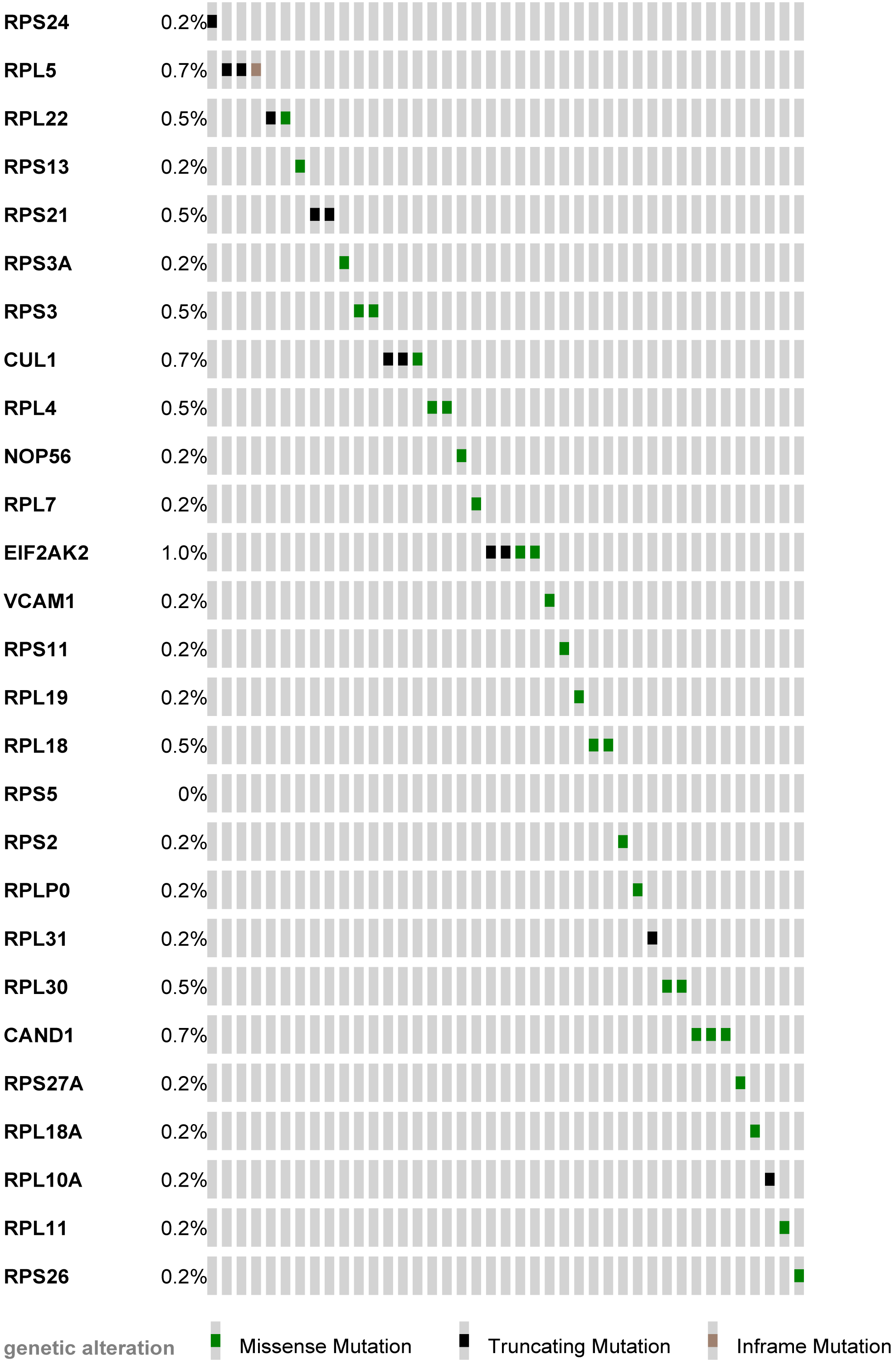


Figure 3.8: Mutations in a cluster consisting of ribosomal proteins
Figure depicts a bar chart showing somatic mutations occurring in genes encoding ribosomal proteins per sample in ccRCC. Taken together, mutations in this cluster have a total coverage of 10% of cases. The figure was generated using the cbio portal tool (http://www.cbioportal.org/) and has been zoomed in and cropped to remove unaltered cases.

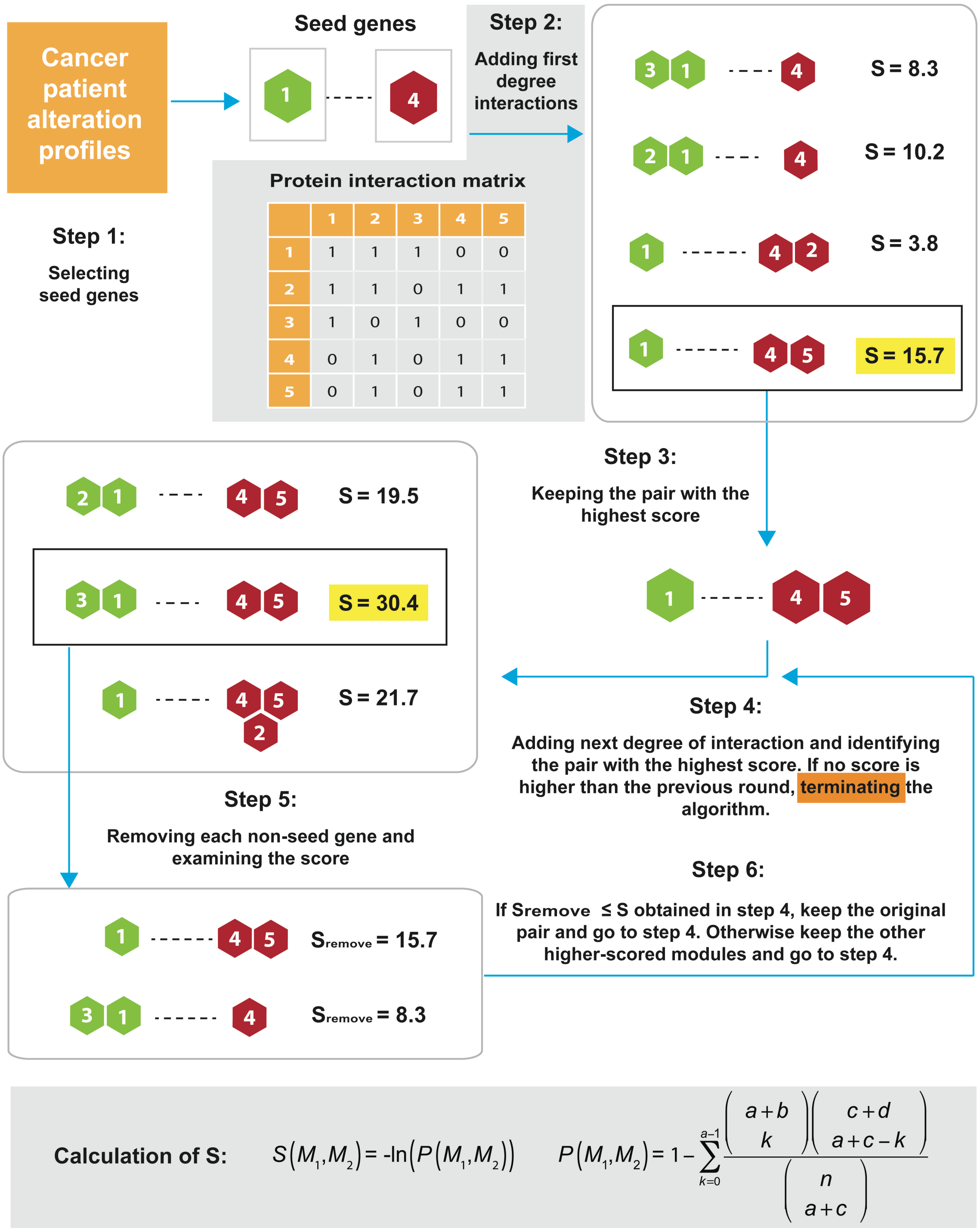
3.3.5 Determining ccRCC driver modules using mutual exclusivity patterns

The Dendrix (Vandin et al.) algorithm was applied to the mutation matrix of TCGA patient samples. Parameters were set similar to the original Dendrix publication. The algorithm was run for a range of set sizes k, for 2 ≤k ≤10. For each k, the algorithm was run 10000 times, starting with random seeds. Similar to the original publication, only sets that were sampled at frequency ≥ 1% were considered to be significant for each k. In the first instance, all genes mutated in at least three cases (total n=417) were considered as part of the analyses. Statistical significance was not observed for any sets with k ≥ 3. For k=2, five sets were sampled with a frequency ≥ 1%, (VHL and MUC4), (VHL and SETD2), (VHL and PBRM1), (VHL and PABPC1) and (VHL and KDM5C). Similar to the Dendrix publication, the mutual exclusivity of these gene pairs is not significant when tested using standard statistical tests. However, it may be postulated that these gene pairs might be biased due to the high coverage of the VHL gene. Therefore, the Dendrix algorithm was run again after removing the VHL gene; two sets, namely (MUC4 and PBRM1) and (BAP1 and PBRM1) were observed to be sampled with average frequencies of ~8% and ~2% respectively. While no direct interactions are known between either of PBRM1 with BAP1 or MUC4 genes, the role of mutations in the SWI/SNF (PBRM1 complex) and its impact on other pathways has been highlighted by the TCGA (The Cancer Genome Atlas Research Network). Furthermore, as discussed in sections 1.2.1 and 1.2.7, BAP1 and PBRM1 genes have been previously shown to be mutated in a mutually exclusive manner (Kapur et al.). The authors also showed differences in patient outcome for BAP1 and PBRM1 mutated cases. Moreover, a recent pan-cancer analysis also showed the mutual exclusivity between the SWI/SNF (PBRM1 complex) and the BAP1 complex in ccRCC in the TCGA cohort using the HotNet2 algorithm (Leiserson et al.).

Biologically, MUC4 (Mucin 4) gene is known to be associated with intestinal epithelial cell differentiation. It has been implicated in tumour progression by repression of apoptosis. Expression levels of MUC4 have been associated with various cancers (up-regulated in pancreatic (Ansari et al., Singh et al.) and down-regulated in breast (Cho et al.)). There are no known interactions between MUC4 and PBRM1 genes. However, for set size k=3 (after removing VHL), two sets namely, (MLL3, MUC4 and PBRM1) and (BAP1, MUC4 and PBRM1) were reported with an average frequency ~2.6% and 1.4% respectively. MLL3/KMT2C (Lysine (K)-Specific Methyltransferase) is a histone methyltransferase gene, a central component of MLL2/MML3 complex and is a coactivator complex of nuclear receptors, involved in transcriptional co-activation. It was also reported to be significantly mutated in Scelo et al. (Scelo et al.). While all these genes (MLL3, MUC4 and PBRM1) have different mechanisms of action, they all could potentially be modulating their functions by affecting the transcription machinery of the cancer cells. No sets were reported with frequency ≥ 1% above k=3.

3.3.6 Co-alteration patterns reveals chromatin modifiers as key players in ccRCC

Until now this analysis has focused on either driver genes or driver modules in isolation. Next, gene modules are explored that may be altered in a co-altered fashion. A probabilistic algorithm was adapted from the literature (Gu et al.) to test this hypothesis. The algorithm assigns each pair of co-altered modules a score, where a higher score indicates a higher likelihood for the module to be a putative signal for ccRCC (Methods section 2.4.4 and Figure 3.9). When the distribution of frequency of scores was assessed, a peak of coaltered modules with scores ≥ 30 (max score = 36) was observed. Co-altered modules were analysed in this peak area, which contained ~19000 module pairs. This analysis is ongoing work being pursued in collaboration with Dr Tammy MK Cheng and data presented here form part of the preliminary analysis.


Cancer patient alteration profiles
Seed genes
1
4
Step 1:
Selecting seed genes
Step 2:
Adding first degree interactions
Protein interaction matrix
3 1 ---- 4 S = 8.3
2 1 ---- 4 S = 10.2
1 ------ 4 2 S = 3.8
1 ------ 4 5 S = 15.7
Step 3:
Keeping the pair with the highest score
1 ------ 4 5
2 1 ---- 4 5 S = 19.5
3 1 ---- 4 5 S = 30.4
1 ----- 4 5 2 S = 21.7
Step 4:
Adding next degree of interaction and identifying the pair with the highest score. If no score is higher than the previous round, terminating the algorithm.
Step 5:
Removing each non-seed gene and examining the score
1 ------ 4 5 Sremove = 15.7
3 1 ------ 4 Sremove = 8.3
Step 6:
If Sremove ≤ S obtained in step 4, keep the original pair and go to step 4. Otherwise keep the other higher-scored modules and go to step 4.
Calculation of S:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	0	1	1
3	1	0	1	0	0
4	0	1	0	1	1
5	0	1	0	1	1

$$S(M_1,M_2) = -\ln\left(P(M_1,M_2)\right) \qquad P(M_1,M_2) = 1 - \sum_{k=0}^{a-1} \frac{\binom{a+b}{k}\binom{c+d}{a+c-k}}{\binom{n}{a+c}}$$

Figure 3.9: Schematic for detection of co-altered modules in ccRCC

Figure depicts the workflow used to determine modules of genes that maybe mutated in co-altered manner. In each iteration of the algorithm, the search is begun with a pair of random seed genes (mutations in ccRCC), and iteratively the closest interaction partner of each seed gene is added to the module. At each step the score is calculated as shown at the bottom of the figure. Genes whose addition to the module does not increase the score are removed. Apart from the seed genes, all other genes in the module are removed iteratively to get the maximum score possible. The search stops when the score stops increasing.

The module pairs with the most frequent gene occurrences, or sets of genes, were interpreted. Chromatin modifier genes especially PBRM1 and ARID1A were the most frequent gene set. Modules containing these genes were observed to be coaltered along with genes associated with a range of functions and pathways, for example, ubiquitin proteolysis, the TGFβ pathway, the cell cycle and the EGFR1 pathway. Further assessment showed that this gene set was most commonly altered with gene sets consisting of the VHL and SETD2 genes. While this is in line with the data presented here (section 3.3.1), it does contradict the results obtained from Dendrix when VHL was included in the analysis; however, as previously postulated, those results may have been biased due to high recurrence of VHL mutations. Nevertheless, further validation of these findings in independent cohorts will be required.

Other genes that were observed to be part of multiple module pairs were TP53, VHL, SETD2 and the MTOR pathway related genes (MTOR, PIK3CA and PTEN). While unsurprisingly modules containing the TP53 gene were co-altered with a whole range of other genes with various cellular functions, VHL showed a tendency of co-alteration with genes belonging to immune system related functions or signalling genes such as those belonging to WNT, NOTCH and MAPK pathways. Intriguingly, most other top hits were seen to be occurring in conjunction with one or more chromatin modifying genes (for example, KDM5C). As part of this first pass analyses, chromatin regulators were observed to be the key players in ccRCC biology; however, at the time of writing of this thesis, certain analysis were still missing. These include assessing the differences between genes, which were observed to be part of module pairs vs. those not part of the module pairs, accounting for bias if any introduced due to the background PPIN and lastly, correlating mutational patterns of module pairs with ccRCC aggressiveness still remains to be elucidated.

3.3.7 Detecting master regulators at the gene expression level

To assess regulators at the gene expression stage, the MARINa algorithm was used. First a ccRCC specific gene expression network was generated using the

ARACNE algorithm (section 2.4.5 and 3.2.14). ARACNE uses minimal information between nodes to reverse engineer a network from gene expression data. MARINa was then applied to this network to get the master regulators (MRs) of gene expression for ccRCC. A total of 8442 regulators were obtained, out of which 692 were significant at the FDR cut-off $q \leq 0.05$ after bootstrapping. Following this list with a shadow analysis, to remove false positive regulators, gave a list of 76 putative MRs. One of the top MRs in this was the epidermal growth factor (EGF) gene. EGF is implicated to be upstream of the PIK3CA/MTOR pathway and thus is involved in its regulation. Furthermore, not surprisingly, pathway enrichment on this list showed some genes belonging to metabolic pathways. However, while now a list of putative MRs at the gene expression level has been obtained much work is still required. This list needs to be comprehensively followed up with a literature search to gather data on these genes and their potential roles and how they may be enabling ccRCC mechanisms. Ultimately, experimental validation of the final list of MRs will be required.

3.3.8 Genotype-Phenotype relationships

Finally, a preliminary analysis to explore genotype-phenotype relations by integrating the genetic alterations to the gene expression data was performed. To find which genetic alterations may explain the gene expression changes observed, the DriverNet algorithm was applied. As explained in the Methods sections 2.4.6 and 3.2.15, DriverNet, generates a bipartite graph and an edge is drawn from the left (genetic alterations) to the right (gene expression matrix), if an alteration could explain the corresponding gene expression changes. Intriguingly, while DriverNet, came up with a list of 100 putative drivers, this list was composed of primarily rare mutations. While it consisted of key transcription factors such as TP53 (mutated in 2% of cases) and MYC (~0.5%), which do regulated numerous other genes, their coverage is low, which fails to explain what would be controlling the genes in the majority of the patient cohort. This avenue of investigation will require further exploration.

3.4 Summary

The analyses performed in this chapter revealed the key genes involved in ccRCC pathogenesis. At the genetic level, these analyses, along with others reported in the literature, confirmed the importance of somatic mutations in VHL, PBRM1, SETD2, BAP1, KDM5C and TP53. Analyses of SCNA data, confirmed previously results (amplification of 5q and deletion of 9p), and also indicated deletion of chromosomes 1p (ARID1A) and 6q as putative target events. Comparison of these events with multiregion biopsy data, revealed that apart from VHL mutations and loss of chromosome 3p, almost all other events are subclonal, and single biopsy approaches are not effective in detecting such alterations due to under-sampling (Gerlinger et al.). Gene expression analysis established the existence of two subgroups of ccRCC, which map to the published prognostic signature ccA/ccB (Brannon et al.).

Multiple network algorithms were employed to understand the pathways that genetic alterations are targeting as well as tackle inter-patient heterogeneity and in turn shed light on ITH. Using the MCODE algorithm, subnetworks/clusters of genes consisting of VHL, PBRM1 and ARID1A could be detected. In comparison, using the Dendrix algorithm to *de novo* identify new pathways, showed that in this dataset of ccRCC cases, adding PPI data is imperative to understand the mechanisms and that using mutational patterns alone are underpowered to do so.

Further, assessing co-altered pathways at the genetic level revealed the importance of chromatin modifier genes both in terms of coverage and also as important genes that are altered along with multiple other pathways. Moreover, VHL, PBRM1 and SETD2, which are the three most recurrently altered genes in ccRCC, appears to be co-altered within pathways; co-alteration within key pathways was observed for other gene sets. These results shed insight into ccRCC mechanisms. Further exploration of such analysis for larger cohorts, with higher frequency of events, may enable answering questions such as which combinations of co-alteration patterns may lead to the more aggressive disease phenotype, or which combination of protein functions within key pathways should be simultaneously targeted to gain maximum therapeutic benefit.

Importantly, most analysis pointed out the significance of PBRM1 and other members of the SWI/SNF machinery, including ARID1A, to be key players in ccRCC biology. The complex was observed to be part of a larger module consisting of 142 genes (MCODE). PBRM1 was observed to be significant along with BAP1 and MUC4 genes (Dendrix). Even in the co-altered module analysis, a high likelihood is assigned for multiple pathways to be co-altered with PBRM1, ARID1A and other chromatin modifying genes. Therefore, these analyses strongly emphasize the importance of chromatin regulators in ccRCC biology. While the utility of PBRM1 as a prognostic marker is still disputable (further discussed in Chapter 4 and see (Kapur et al., Gulati et al.)), these results are indicative of the importance of the SWI/SNF complex in ccRCC biology.

While an attempt was made to elucidate genotype-phenotype relationships using the DriverNet algorithm, significant results were not obtained. There could be multiple reasons for this. The analyses may have been underpowered due to low frequency of events at the genetic level. Furthermore, the gene list obtained from DriverNet was enriched for genes that were hub genes i.e. had high numbers of PPIs. It can be speculated that the results may be biased towards such genes and thus producing false positives. Lastly, the algorithm was developed for microarray data and no significant changes have been made to the underlying calculations for RNA-Seq data. There is a step within the algorithm where expression outliers are calculated from the gene expression matrix, it may be that the algorithm is under detecting deregulated genes due to inherent differences in the microarray and RNA-Seq read outs.

Finally, it should be noted that due to time limitations, for all network-based analyses only somatic mutation data was considered. Adding genes altered *via* SCNAs could potentially add more definitive information to the analyses performed.

Chapter 4. The quest for prognostic biomarkers

4.1 Introduction

So far it can be seen that, fitting to the overall picture of a typical cancer, ccRCC has a heterogeneous landscape with a few highly recurrent and numerous low frequency, somatic mutations as well as somatic copy number alterations (SCNAs). Moreover, in a number of studies, multiple gene expression based subtypes have been observed for ccRCC (Brannon et al., Zhao et al., Beleut et al.). As discussed in section 1.2.7 of the Introduction, outcome prediction for ccRCC greatly relies on clinical factors, such as tumour stage and tumour grade. Prognostic models have been formulated using these and other factors identified through routine clinical practice; such as the Mayo clinic, stage, size, grade and necrosis (SSIGN) score for predicting cancer specific survival (CSS) and the University of California Integrated Staging System (UISS), which includes TNM category, Fuhrman grade, and performance status to predict overall survival (OS). It is fit to presume that combining the existing clinical models to molecular biomarkers may help improve the accuracy of prognostic models. Through the years various research groups have found recurrent somatic mutations, SCNAs as well as gene expression signatures to be clinically associated with ccRCC. While a few of these events such as mutations in the BAP1 gene and deletion of chromosome 9p have been observed to be associated with patient prognosis in multiple studies (Kapur et al., Hakimi et al., Sanjmyatav et al., Klatte et al., La Rochelle et al.), most of these signatures have not been independently validated on different patient cohorts. Furthermore, the presence of a substantial amount of genetic intra-tumour heterogeneity in ccRCC, detected through exome sequencing (Gerlinger et al., Gerlinger et al.), as well as SNP array analysis (Martinez et al.), of several regions from the same tumour, has raised questions regarding the applicability of such signatures in clinical practice. Taken together, these caveats suggest that further research is required to validate all available signatures in larger and independent cohorts.

Availability of somatic mutation, SCNA, gene expression (RNA sequencing) and follow up data for over 400 ccRCC cases, published by The Cancer Genome Atlas consortium (TCGA, https://tcga-data.nci.nih.gov/tcga/), has enabled the direct comparison of the known ccRCC genomic predictors and provide the opportunity to systematically validate previously identified genetic and transcriptomic prognostic biomarkers in a large independent patient cohort.

Thus, the analyses explained in this chapter were devised with an aim to answer two particular questions; first, to validate and compare published ccRCC prognostic biomarkers in an independent patient cohort and secondly, to assess intratumour heterogeneity (ITH) of the most promising markers to guide biomarker optimisation.

4.2 Methods

The framework for the analysis in this chapter has been described in detail here. All statistics applied in this chapter are briefly outlined with references to more detailed descriptions in Chapter 2 provided in relevant places.

4.2.1 Literature Search

While putative driver events were identified in Chapter 3 (such as mutations in the VHL gene, BAP1 gene and SCNAs in Chrom 3p), due to the lack of availability of a second independent validation cohort, candidate prognostic markers to be assessed in this chapter were selected using an exhaustive literature search. Biomarkers were selected that had been previously shown to have distinctive prognostic association as apposed to just simply being identified with driver events.

The aim here was to validate these prognostic biomarkers using an independent cohort from the TCGA data set.

To compile an exhaustive list of possible biomarkers for validation, a systematic search of the PubMed and Google Scholar databases for publications describing genetic or transcriptomic prognostic biomarkers for RCC, was performed. The terms, renal cell carcinoma, biomarker, prognosis and survival were used as keywords, and the search restricted to combinations of these terms. Articles published before and until December 2013, and in the English language, were considered for further analyses. Studies had to be based on either exclusively clear cell histology or could be mixed cohorts with other histologies; studies exclusively based on non-clear cell histology were excluded. Additional literature cited in identified prognostic marker publications or recent review articles (Brannon and Rathmell, Jonasch et al., Tang et al., Eichelberg et al., Junker et al., Arsanious et al., Oosterwijk et al.) was also assessed.

As a final filter, the inclusion of follow-up data to show association with prognosis was deemed essential; studies that only showed an association with other poor prognosis clinical factors, such as tumour stage and grade were removed. Several publications investigating gene expression levels as potential prognostic biomarkers lacked information about how the identified genes can be applied to clinical samples in order to identify prognostically distinct subgroups. These were also excluded from further analysis. Using all these filters, 30 publications describing in all 32 RCC genetic or gene expression based prognostic biomarkers were identified in the literature search. However, four biomarkers were excluded from further analysis due to technical reasons. One biomarker (Yao et al.) was based on regression coefficients devised using microarray gene expression data. This could not to be applied to RNA-Seq data and was therefore excluded. The other three studies included multi-gene expression signatures, for which fewer than 70% of gene probes mapped to genes annotated in the TCGA RNA-Seq dataset, which was chosen as an arbitrarily justified cut-off to be able to reproduce the respective signature (Takahashi et al., Sultmann et al., Vasselli et al.).

4.2.2 Patient cohort

To study the above outlined objects, two cohorts were used. For the first part of the analysis, where an attempt to identify prognostic biomarkers was made, data published by TCGA was used. As described in Methods, sections 2.1.1, somatic mutation (n=417) and clinical data (n=446) were obtained for the same cohort from the supplementary material of the TCGA ccRCC publication (The Cancer Genome Atlas Research Network), SNP array (n=450) and RNA sequencing (RNA-Seq) data (n=469), and were downloaded (https://tcga-data.nci.nih.gov/tcga/) on 14th March 2012 and 18th September 2012, respectively. All of somatic mutations, SCNA, RNA-Seq and clinical data were available for a common cohort of 354 patients. However, follow-up data or tumour grade were missing for four patients, leaving 350 patients, which formed our study cohort (core dataset).

For the second part of the analysis, to assess ITH of the identified biomarkers, data published by our collaborating laboratory was used (Gerlinger et al., Gerlinger et al.). The multi-region gene expression datasets GSE31610 and GSE53000 were downloaded from the gene expression omnibus for the assessment of ITH. The dataset and preliminary processing of the microarray data is detailed in sections 2.1.2 and 2.2.4; clustering of regions was performed using NMF clustering, as explained in section 0.

4.2.3 Classification of patients into prognostic groups

For somatic mutations, patients were classified into prognostic groups as those having non-syn mutations in the gene versus those not having these mutations. For VHL two additional cases were considered; 1) Non-syn mutations for Stage I/II/III cases only and 2) Loss-of-function mutations only. For SCNAs, patients were compared as those with the specific SCNA versus those without. For gene expression based signatures, for the three individual genes, CD31, EDNRB and TSPAN7; the cut-offs as given in the reference publication were used. For gene panel based classifiers, NMF clustering was performed to classify patients into cohorts. Lastly, for the TGFβ signature, pathway activity score was calculated and patients were divided into two cohorts using median score value (Methods section 2.5).

4.2.4 Statistical Methods

Assessment of association with prognosis was done at both i) univariate, using the logrank test (Methods Section 2.6.2) and the competing risk analysis (Methods Section 2.6.4), and at the ii) multivariate level using the Cox regression analysis (Methods Section 2.6.5). Death due to ccRCC was the chosen endpoint of interest. For all analyses, patients with the field "Composite Vital Status" = "DECEASED" and "Composite Tumour Status" = "WITH TUMOR" were considered to be dead with clear cell renal cancer related causes, while those with "Composite Vital Status" = "DECEASED" and "Composite Tumour Status" = "TUMOR FREE" were considered to be dead due to other causes. Follow-up time was defined using the "Composite Days to Death" field in case of patient death, and "Composite Days to Last Contact" for patients alive at the end of study period. For the multivariate Cox regression analysis, a backwards-stepwise selection process was implemented. The selection step was repeated till all the variables left in the model had $p \leq 0.05$. Although it is hard to define a formal way to determine the number of parameters, which can be tested in multivariate analysis based on the death event rate, to the best of our knowledge, we should not have more than 'n' number of variables in the final model where n = total number of deaths from disease/10, which for our study would equal 8 variables (Zwiener et al.). Our final multivariate model after stepwise selection has only two variables, which is in accordance with this criterion.

Recursive partitioning (Methods section 2.7.1) was performed using the ctree() function in the 'party' package in R to generate a prognostication model. The logrank method was used to generate the p-values, and each node was split based on the logrank statistic and $p \leq 0.05$.

All statistical analyses were performed in R (v3.0.1) (R Development Core Team), using the packages 'survival' 'gplots', 'cmprsk' and 'party'. Survival graphs were generated with GraphPad Prism (v6.03).

4.3 Results

The median follow-up for the analysed patient cohort was 51 months. Clinical and pathological characteristics for the cohort are described in

Table 4.1 and were similar to the RCC cohorts from which the candidate biomarkers had been identified. All patients had undergone nephrectomy, which is the current line of treatment for kidney cancers and from which the samples for molecular analysis had been taken. After passing through the filters described in section 4.2.1, the literature search resulted in a total of 26 studies describing in all 28 prognostic biomarkers (Table 4.2).

Table 4.1: Patient and tumour characteristics for the data cohort
Table gives key clinical patient and tumour characteristics for the core dataset of 350 cases.

Variable	TCGA Cohort (n=350)
Age	
Median (IQR)	61 (52-70)
Gender	
Male	222 (63%)
Female	128 (37%)
Fuhrman Grade	
G1	4 (1%)
G2	145 (41%)
G3	146 (42%)
G4	55 (16%)
Clinical Stage	
Stage I	162 (46%)
Stage II	34 (10%)
Stage III	96 (27%)
Stage IV	58 (17%)
Primary Tumour Spread	
T1	166 (48%)
T2	40 (11%)
T3	139 (40%)
T4	5 (1%)
Metastatic Spread	
M0	293 (84%)
M1	57 (16%)
Lymphnode Spread	
N0	168 (48%)
N1	8 (2%)
NX (Undetermined)	174 (50%)
Median Follow-up	51 months
Total number of deaths	121
Number of deaths from ccRCC	80

Table 4.2: Literature Search Results
This table gives a summary of all the studies, biomarkers from which were considered part of this analysis, along with a reference of each study.

Variable	Prognosis	Analysis	Cohort Size* (n)	Reference
		Somatic Mutations		
VHL (loss of function+ mutations)	Poor (OS/PFS)	Sequencing	56	Kim et al.
VHL (loss of function+ mutations)	Poor (CSS)	Sequencing	83	Schraml et al.
VHL (somatic mutations)	Better (CSS/CFS)	Sequencing	134	Yao et al.
PBRM1	Better (OS)	Sequencing	145 + 327	Kapur et al.
BAP1	Poor (OS)	Sequencing	145 + 327	Kapur et al.
BAP1	Poor (CSS)	Sequencing	188 + 421	Hakimi et al.
BAP1	Poor (OS)	Sequencing	>400	The Cancer Genome Atlas Research Network
BAP1	Poor (OS)	Sequencing	240	Sato et al.
SETD2	Poor (CSS)	Sequencing	188 + 421	Hakimi et al.
SETD2	Poor (CFS)	Sequencing	240	Sato et al.
TP53	Poor (CSS)	Sequencing	416	Kandoth et al.

Details on the cohort size of the original study, along with the method of detection of the biomarker as well as the endpoint of interest (CSS or overall survival) are shown.

Table 4.2 continued

Variable	Prognosis	Analysis	Cohort Size* (n)	Reference
Somatic Copy Number Variations				
5q31-qter (5q focal) Amplification	Better (CSS)	Cytogenetics	104	Gunawan et al.
7q36.2 (7q focal) Amplification	Poor (CSS)	array CGH, FISH	53	Sanjmyatav et al.
8q Amplification	Poor (CSS)	Cytogenetics	336	Klatte et al.
8q Amplification	Poor (OS)	SNP array	85	Monzon et al.
12 Amplification	Poor (RFS)	Cytogenetics	50	Elfving et al.
20q11.21q13.32 (20q focal) Amplification	Poor (CSS)	array CGH, FISH	53	Sanjmyatav et al.
20 Amplification	Poor (RFS)	Cytogenetics	50	Elfving et al.
3p Deletion	Better (CSS)	Cytogenetics	246	Klatte et al.
3p Deletion	Better (CSS)	Cytogenetics	288	Kroeger et al.
4p Deletion	Poor (CSS)	Cytogenetics	246	Klatte et al.
8p Deletion	Poor (RFS)	Cytogenetics	50	Elfving et al.
9p21.3p24.1 (9p focal) Deletion	Poor (CSS)	CGH, FISH	53	Sanjmyatav et al.
9p Deletion	Poor (CSS)	Cytogenetics	246	Klatte et al.
9p Deletion	Poor (CSS/ RFS)	Cytogenetics, FISH	703	La Rochelle et al.
9p Deletion	Poor (RFS)	CGH	37	Moch et al.
9p Deletion	Poor (CSS)	FISH	73	Brunelli et al.
14q Deletion	Poor (CSS)	Cytogenetics	246	Klatte et al.
14q Deletion	Poor (CSS)	Cytogenetics	288	Kroeger et al.
14q Deletion	Poor (OS/RFS)	SNP array	85	Monzon et al.
19 Deletion	Poor (CSS)	Cytogenetics	131	Antonelli et al.
22 Deletion	Poor (CSS)	Cytogenetics	131	Antonelli et al.

Table 4.2 continued

Variable	Prognosis	Analysis	hort Size* (n)	Reference
	Gene Expression Analysis			
CD31, EDNRB and TSPAN7 expression levels	Higher expression levels of each are better	mRNA arrays	24	Wuttig et al.
Aggressive and non-aggressive ccRCCs classified using 35 g	Aggressive worse than non-aggressive (CS	mRNA arrays	66	Kosari et al.
Two gene expression clusters classified using 259 genes	Cluster 2 worse than Cluster 1 (CSS)	mRNA arrays	177	Zhao et al.
Indolent and aggressive ccRCC classified using 44 genes	Aggressive worse than indolent	cDNA arrays	38	Lane et al.
ccA/ccB subgroup classified using 110 genes	ccB worse than ccA (CSS)	mRNA arrays	48 + 177	Brannon et al.
Cluster A,B and C classified using 48 (B vs. A/C) and 23 (A vs.	Cluster A better than B and C, C being poo	mRNA arrays	176	Beleut et al.
TGFβ signature -scored with a panel of 157 TGFβ genes	Poor for higher expression (CSS)	mRNA arrays	176	Boström et al.

NOTE: *The cohort size in this table signifies the number of cases for which follow up data was available. + Loss of function mutation was defined as frameshift or nonsense mutations.

4.3.1 Univariate analyses

In order to validate the identified prognostic biomarkers, univariate analyses were performed using two different tests. All identified biomarkers were first tested using the logrank test, and all biomarkers assessed to be significant in logrank test, were then re-validated using the competing risk analysis.

4.3.1.1 logrank test

The logrank test assesses the significance of the difference in the survival distribution of samples under two or more conditions. The working of the logrank test is explained in more detail in section 2.6.2. This test was used to validate all identified prognostic signatures along with tumour stage and Fuhrman grade. The analysis started with assessing clinical factors, which have established association with prognosis. As expected, higher tumour stage and grade were significantly associated with poor CSS (Figure 4.1 and Table 4.3). Other established clinical prognostic variables such as blood test results, performance status or necrosis were not available for all patients and hence were not evaluated at this point.

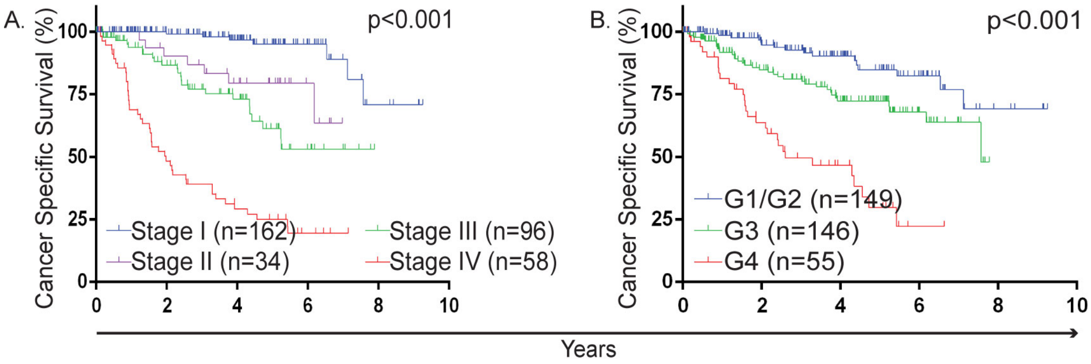


Figure 4.1: Kaplan Meier survival estimates for cancer specific survival determined for key clinical variables
A. Depicts the Kaplan Meier (KM) survival curves based on tumour stage. Tumour Stage I has the best survival while stage IV has the worst survival. B. Depicts the KM curves based on Fuhrman grade. Due to the low number of G1 cases (n=2), they have been included along with G2 cases, with these cases having the best prognosis and G4 the worst.

Somatic mutations

Somatic mutations in five tumour suppressor genes have been described to have prognostic associations (Kim et al., Schraml et al., Yao et al., Kapur et al., Hakimi et al., Sato et al., Kandoth et al.). While for genes, PBRM1, BAP1, SETD2 and TP53, association has been observed for non-synonymous mutations with CSS; for the VHL gene, association has been shown for non-synonymous mutations in one case for stage I-III grade tumours only (Yao et al.), while in two other studies, loss-of-function mutations (frameshift and nonsense mutations) were shown to be associated with prognosis (Kim et al., Schraml et al.). Significance was assessed in accordance with each study, however only non-synonymous mutations in the BAP1 (HR 1.94, p=0.022) and TP53 (HR 5.09, p<0.001) tumour suppressor genes validated as predictors of poor CSS (Figure 4.2 and Table 4.3).

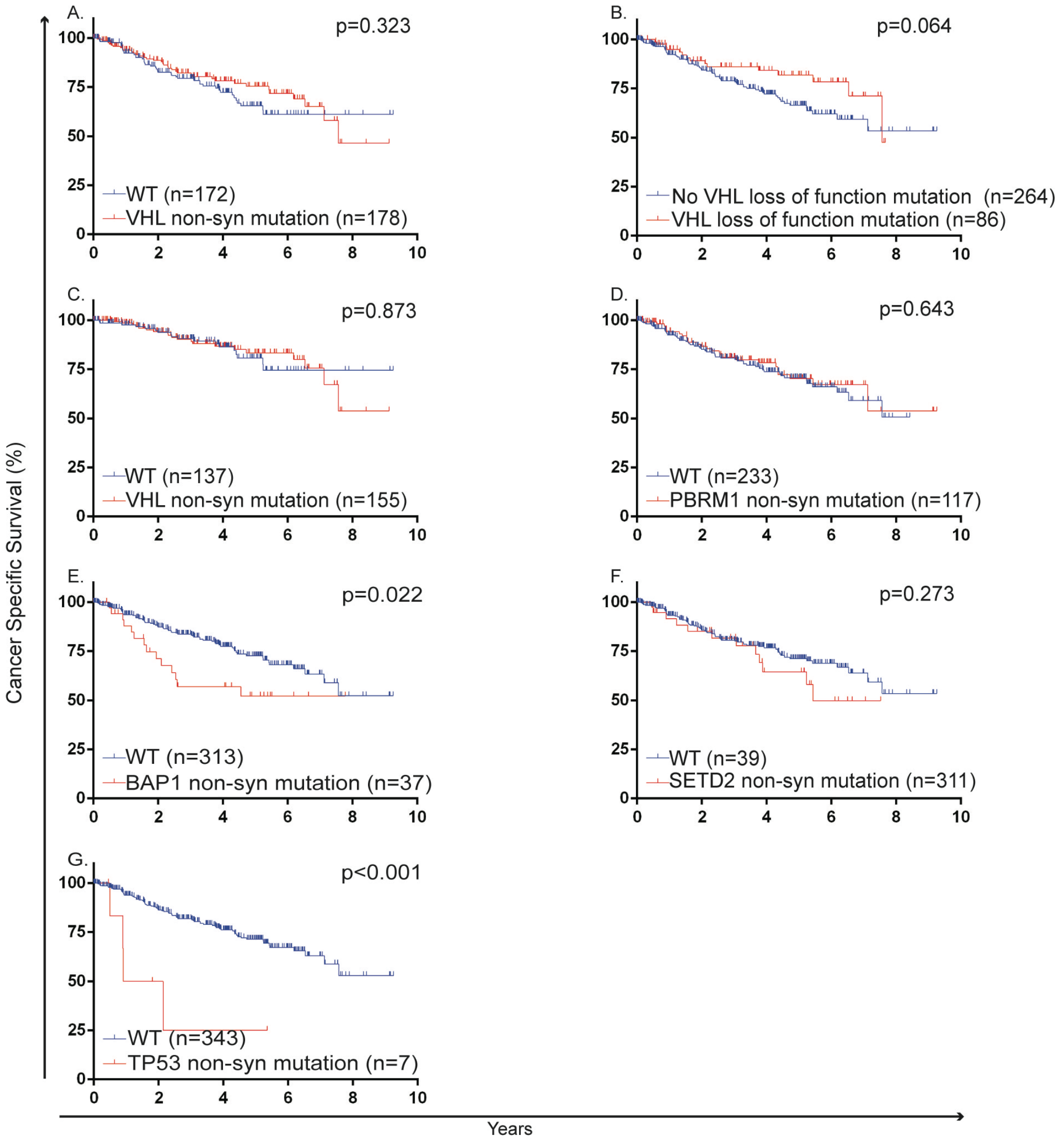


Figure 4.2: Kaplan Meier survival estimates for cancer specific survival determined for somatic mutations

A. VHL non-synonymous (non-syn) mutations (all cases), B. VHL loss-of-function mutations, C. VHL non-syn mutations (Stage I-III), D. PBRM1 non-syn mutations, E. BAP1 non-syn mutation status, F. SETD2 non-syn mutations, G. TP53 non-syn mutation status

WT = wild type

Somatic copy number alterations

A total of 14 copy number alteration events were identified as candidate biomarkers, including four focal SCNAs (Gunawan et al., Sanjmyatav et al.), six arm level alterations (Klatte et al., Monzon et al., Klatte et al., Kroeger et al., Elfving et al., La Rochelle et al., Moch et al., Brunelli et al.) and four whole chromosome alterations (Elfving et al., Antonelli et al.). Several of these SCNAs have been identified by cytogenetic and other low-resolution analyses. To facilitate comparison, copy number profiles generated from high resolution SNP array data, obtained from TCGA, were converted into lower resolution cytoband level data. Amplification or deletion of ≥50% of a chromosome arm, or of both arms of a chromosome, was considered to be equivalent to an arm level alteration, or to a whole chromosome aberration, respectively (The Cancer Genome Atlas Research Network).

Nine out of the 14 SCNAs validated to be associated with prognosis and interestingly all showed association with poor prognosis. Chromosome 8q (Chrom8q) amplification (HR 2.70, $p<0.001$), Chrom12 amplification (HR 1.74, $p=0.034$), Chrom20 focal amplification (HR 2.44, $p<0.001$), Chrom20 amplification (HR 2.37, $p<0.001$), Chrom4p deletion (HR 1.97, $p=0.019$), Chrom9p focal deletion (HR 2.33, $p<0.001$), Chrom9p deletion (HR 2.56, $p<0.001$), Chrom19 deletion (HR 3.25, $p=0.034$) and Chrom22q deletion (HR 2.23, $p=0.012$) were significantly associated with poor CSS. The remaining five SCNA markers failed validation (Figure 4.3 and Table 4.3).

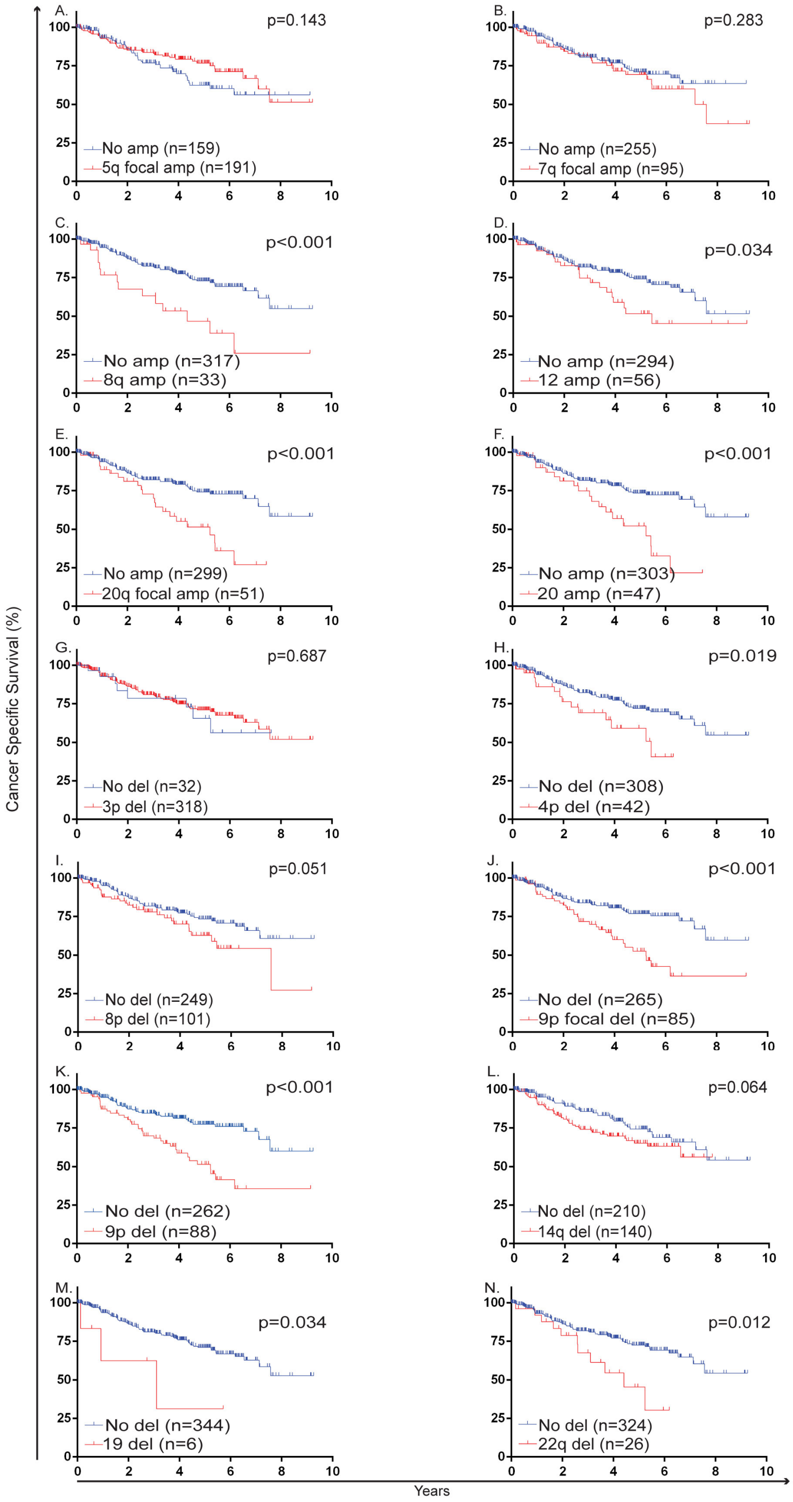


Figure 4.3: Kaplan Meier survival estimates for cancer specific survival for somatic copy number alterations

A. Chrom5q focal amplification (amp) status, B. Chrom 7q focal amp status, C. Chrom8q amp status, D. Chrom12 amp status, E. Chrom20q focal amp status, F. Chrom20 amp status, G. Chrom3p deletion (del) status, H. Chrom4p del status, I. Chrom8p del status, J. Chrom9p focal del status, K. Chrom9p del status, L. Chrom14q del status, M. Chrom19 del status, N. Chrom22q del status.

Gene expression analysis

Nine gene expression biomarkers were identified, which included gene expression levels of three individual genes namely EDNRB, CD21 and TSPAN7, five gene expression panel based classifiers, and one signature based on TGFβ pathway activity. Eight out of the nine signatures validated. EDNRB and TSPAN7 gene-expression above defined cut-offs (Wuttig et al.) correlated with improved CSS (HR 0.29, p <0.001 and HR 0.37, p<0.001 respectively); however, CD31 overexpression was not significant. NMF clustering was applied for each multi-gene expression signature (Brannon et al., Kosari et al., Lane et al., Zhao et al., Beleut et al.) in order to identify samples with distinct expression profiles (consensus clustering maps can be found in Appendix D). All prognostic gene expression signatures validated: the aggressive subgroup defined by Kosari (Kosari et al.) had worse CSS than the non-aggressive subgroup (HR 2.85, p<0.001); the Zhao (Zhao et al.) poor prognosis Cluster 2 had worse CSS than Cluster 1 (HR 5.26, p<0.001); the aggressive subgroup defined by Lane (Lane et al.) showed worse CSS than the indolent subgroup (HR 4.21, p<0.001); the Brannon (Brannon et al.) poor prognosis ccB subgroup (HR 4.90, p<0.001) had worse CSS than the ccA subgroup. Based on Beleut (Beleut et al.), CSS was significantly worse for patients in the poor prognosis Clusters C (HR 2.21, p=0.034) and B (HR 2.46, p=0.002) than for those in Cluster A; although CSS of Clusters B and C showed no significant difference. Bostrom's (Bostrom et al.) poor risk subgroup with high TGFβ score had worse CSS than the subgroup with a low score (HR 1.98, p=0.003) (Figure 4.4 and Table 4.3).

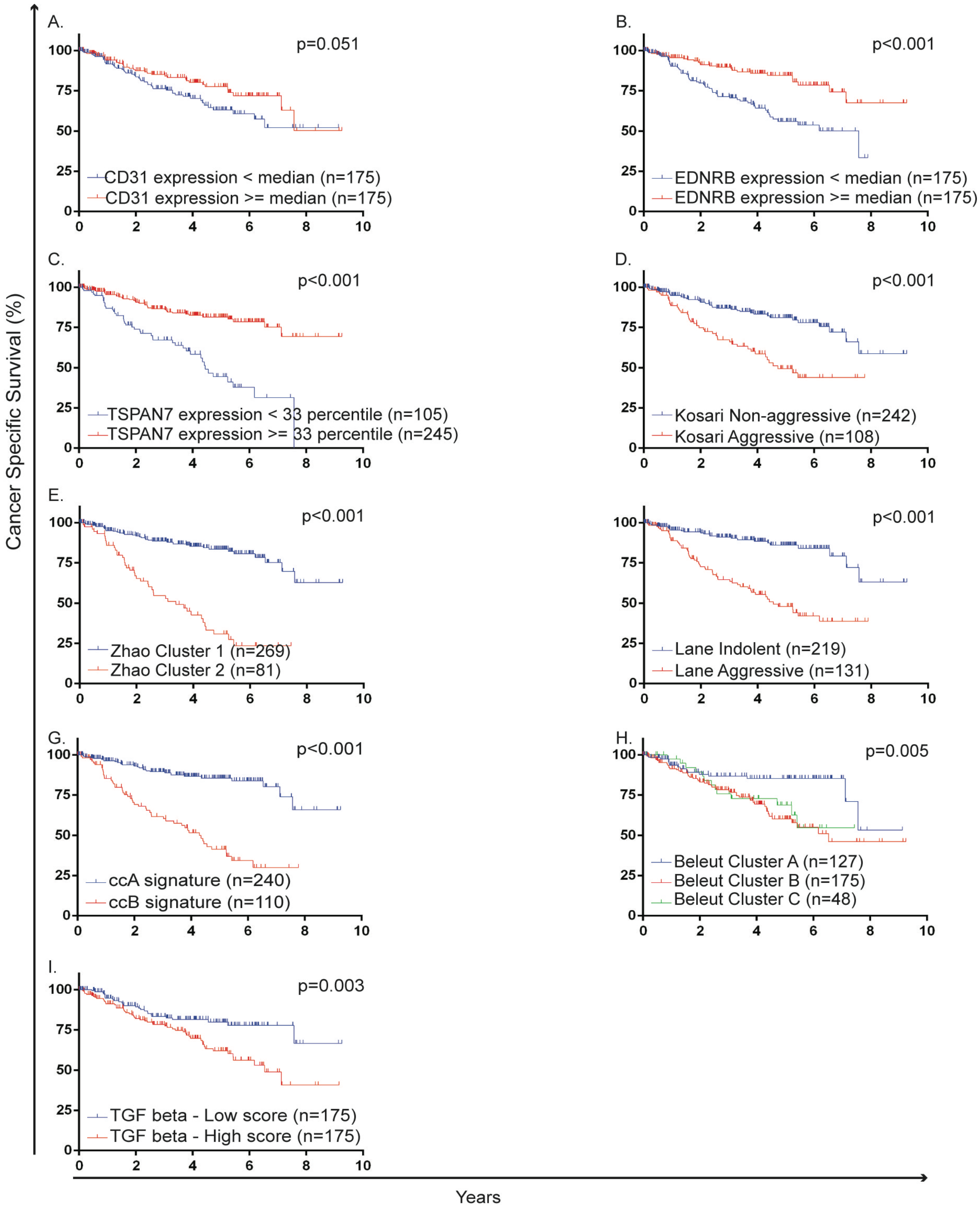


Figure 4.4: Kaplan Meier survival estimates for cancer specific survival for clinical gene expression based signatures

A. EDNRB expression levels, B. TSPAN7 expression levels, C. Gene expression subgroup of patients – Kosari signature, D. Gene expression subgroup of patients – Zhao signature, E. Gene expression subgroup of patients – Lane signature, F. Gene expression subgroup of patients – ccA/ccB, G. Gene expression subgroup of patients – Beleut signature, and H. Gene expression subgroup of patients according to TGFβ activity score.

Table 4.3: Results from logrank analysis
Table gives the results as obtained by logrank analysis for each of the 28 assessed biomarkers. For each biomarker, the number of cases, the hazard ratio (HR) calculated by univariate Cox analysis and the p-value of significance calculated by logrank analysis is provided.

Variable	Number of cases (n=350)	HR (95% C.I.)	p-value
Clinical and Pathological Characteristics			
Stage II vs. Stage I	34 (10%)	4.45 (1.55 – 12.77)	0.006
Stage III vs. Stage I	96 (27%)	7.34 (3.16 – 17.08)	<0.001
Stage IV vs. Stage I	58 (17%)	25.24 (11.26 – 56.71)	<0.001
G3 vs. G1/G2	146 (42%)	2.35 (1.30 – 4.26)	0.005
G4 vs. G1/G2	55 (16%)	7.43 (3.99 – 13.81)	<0.001
Somatic Mutations			
VHL loss of function mutation	86 (24.5%)	0.59 (0.34 – 1.04)	0.064
VHL non-syn mutation (all cases)	178 (51%)	0.80 (0.51 – 1.25)	0.323
VHL non-syn mutations (stage I-III cases)	155/292 (53%)	0.95 (0.50 – 1.80)	0.873
PBRM1 non-syn mutation	117 (33%)	0.90 (0.56 – 1.43)	0.643
BAP1 non-syn mutation	37 (10.5%)	1.94 (1.08 – 3.45)	0.022
SETD2 non-syn mutation	39 (11%)	1.41 (0.76 – 2.60)	0.273
TP53 non-syn mutation	7 (2%)	5.09 (1.85 – 14.00)	<0.001
Copy Number Variations			
5q focal Amplification	191 (54.5%)	0.72 (0.47 – 1.12)	0.143
7q focal Amplification	95 (27%)	1.29 (0.81 – 2.05)	0.283
8q Amplification	33 (9%)	2.70 (1.52 – 4.81)	<0.001
12 Amplification	56 (16%)	1.74 (1.04 – 2.91)	0.034
20q focal Amplification	51 (15%)	2.44 (1.49 – 3.99)	<0.001
20 Amplification	47 (13%)	2.37 (1.41 – 3.97)	<0.001
3p Deletion	318 (91%)	0.86 (0.41 – 1.79)	0.687
4p Deletion	42 (12%)	1.97 (1.10 – 3.52)	0.019
8p Deletion	101 (29%)	1.58 (0.99 – 2.50)	0.051
9p focal Deletion	85 (24%)	2.33 (1.49 – 3.64)	<0.001
9p Deletion	88 (25%)	2.56 (1.64 – 3.99)	<0.001
14q Deletion	140 (40%)	1.51 (0.97 – 2.35)	0.064
19 Deletion	6 (1.7%)	3.25 (1.02 – 10.32)	0.034
22q Deletion	26 (7%)	2.23 (1.18 – 4.23)	0.012

Table 4.3 continued

Variable	Number of cases (n=350)	HR (95% C.I.)	p-value
Expression Analysis			
CD31 expression			
< median	175 (50%)	0.64 (0.41 – 1.01)	0.051
≥ median	175 (50%)		
EDNRB expression			
< median	175 (50%)	0.37 (0.23 – 0.59)	<0.001
≥ median	175 (50%)		
TSPAN7 expression			
< 33 percentile	105 (30%)	0.29 (0.18 – 0.45)	<0.001
≥ 33 percentile	245 (70%)		
Kosari signature			
Non - aggressive	242 (69%)	2.85 (1.84 – 4.43)	<0.001
Aggressive	108 (31%)		
Zhao signature			
Cluster 1 (good)	269 (77%)	5.26 (3.37 – 8.22)	<0.001
Cluster 2 (poor)	81 (23%)		
Lane signature			
Indolent	219 (63%)	4.21 (2.62 – 6.77)	<0.001
Aggressive	131 (37%)		
ccA/ccB status			
ccA	240 (69%)	4.90 (3.09 – 7.76)	<0.001
ccB	110 (31%)		
Beulet signature			
Cluster A	127(36%)	1.00 (Ref)	0.009
Cluster B	175 (50%)	2.27 (1.31 – 3.96)	
Cluster C	48 (14%)	2.30 (1.13 – 4.66)	
TGFβ signature			
Low expression score	175 (50%)	1.98 (1.23 – 3.16)	0.003
High expression score	175 (50%)		

4.3.1.2 Competing risk analysis

While a logrank test is commonly used when assessing the association of an event with patient prognosis, as discussed in section 2.6.4, a competing risk (CR) analysis assesses the cumulative incidence of an event and takes into account death due to other causes. Thus, a CR analysis is less likely to over fit the significance of an event. Keeping this in mind, all 19 of the 28 molecular biomarkers, which were observed to be significantly associated (p≤0.05) with CSS in the logrank test, were re-assessed using competing risk analysis. As shown in Table 4.4, 17 out of the 19 assessed biomarkers showed significant association with poor prognosis and only non-synonymous mutations in the BAP1 gene and Chrom19 deletion failed to validate.

Table 4.4: Competing risk analysis
Table gives the p-value of significance for the 19 biomarkers which were re-assessed in a competing risk analysis. 17 out of the 19 biomarkers validated in this analysis.

Variable	p-value
BAP1 non-syn mutation	0.072
TP53 non-syn mutation	0.006
8q Amplification	<0.001
12 Amplification	0.047
20q focal Amplification	<0.001
20 Amplification	0.001
4p Deletion	0.028
9p focal Deletion	<0.001
9p Deletion	<0.001
19 Deletion	0.081
22q Deletion	0.016
EDNRB >= median	<0.001
TSPAN7 >= 33%	<0.001
Kosari signature : aggressive	<0.001
Zhao signature: poor subgroup	<0.001
Lane signature: aggressive	<0.001
ccA/ccB subgroup status	<0.001
Beulet signature	0.015
TGFβ signature: high expression	0.003

4.3.2 Multivariate Cox regression analysis

At this stage, 17 of the 28 identified candidate biomarkers, using two univariate analyses, could be validated. Of further interest, is how independent these biomarkers were in comparison to each other, and if any of them were able to added prognostic information to established clinical factors. To assess this, a multivariate analysis was performed, containing these validated biomarkers along with established clinical variables. As there were overlapping SCNA events that passed validation in the univariate analysis, to avoid redundancy, Chrom9p focal deletion and Chrom20 whole arm amplification were excluded on the basis of their lower hazard ratios as compared to the overlapping Chrom9p arm level deletions and Chrom20 focal amplifications. Two instances of Cox regression were performed; in the first instance, the 17 biomarkers remaining after removing the above two, which had been validated in the logrank analysis, were included together with tumour stage and grade into the multivariate analysis (MVA). Tumour stage, the ccA/ccB gene expression signature and Chrom19 deletions were the only independent predictors of CSS (Table 4.5). In the second instance of the analysis, the two markers (BAP1 mutations, Chrom19 deletions), which had not been significant in the competing risk analysis, were excluded and only tumour stage and the ccA/ccB signature remained significant in the MVA (Table 4.5). Taking both these results into consideration, along with the small number of six tumours showing Chrom19 deletions, the ccB signature was the lead candidate for further assessment. For all non-significant variables, the hazard ratio, 95% confidence interval (C.I.), and a p-value, was generated at the step it was removed (Appendix E).

Table 4.5: Multivariate Cox Regression analysis
Table shows the results as obtained in both the iterations of the backwards-stepwise regression analysis. The left side of the table gives the HR and p-value for the three significant variables remaining at the end when all 19 variables were considered. The right side of the table shows the results when the analysis was performed considering only the 17 variables which validated both in the logrank and competing risk analysis.

Variable	Including BAP1 mutations and Chrom19 deletion		Excluding BAP1 mutations and Chrom19 deletion	
	Hazard Ratio (C.I.)	p-value	Hazard Ratio (C.I.)	p-value
Tumour stage				
Stage I	1.00 (Ref)		1.00 (Ref)	
Stage II	3.48 (1.20 – 10.06)	0.022	3.40 (1.18 – 9.82)	0.024
Stage III	4.61 (1.93 – 11.00)	<0.001	4.86 (2.05 – 11.55)	<0.001
Stage IV	18.01 (7.89 – 41.12)	<0.001	17.77 (7.79 – 40.53)	<0.001
Chromosome 19 deletion	4.18 (1.27 – 13.69)	0.018	-	-
ccA/ccB status				
ccA status	1.00 (Ref)	<0.001	1.00 (Ref)	<0.001
ccB status	2.99 (1.87 – 4.80)		2.95 (1.84 – 4.72)	

The ccB signature was consistently associated with a worse prognosis in patients with stage I (HR >10, p<0.001), stage II/III (HR 3.03, p=0.003) and stage IV ccRCCs (HR 2.15, p=0.015) (Figure 4.5). A total of 135 patients with stage I tumours expressing the ccA signature, demonstrated particularly good outcomes with no cancer specific deaths for over 6 years.

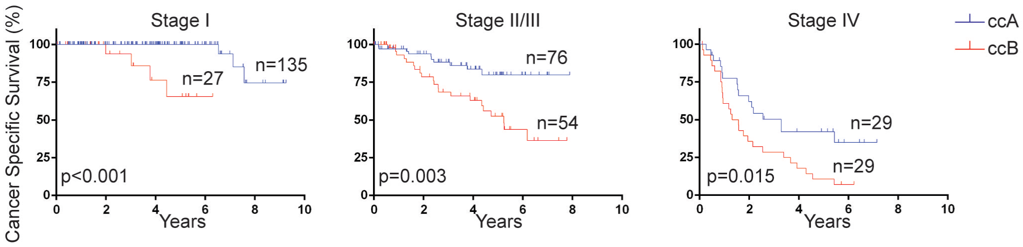


Figure 4.5: Kaplan Meier survival estimates for cancer specific survival for ccA/ccB split by tumour stage
KM curves depicting that even if patient cohorts are divided according to stage, for each stage wise cohort, significant differences in survival (all logrank p <0.05) are observed based on the ccA/ccB subgroup of the patients.

A further point to note here is that in section 3.2.6 of chapter 3, we had seen that results using the ccA/ccB gene panel of 110 genes correlate well with the results of using the much larger panel of 1500 genes when attempting to classify patients into subgroups.

In February 2014, after completion of the literature search, a newer prognostic signature named ClearCode34, which is based on the ccA/ccB signature, was published (Brooks et al.). This signature is based on the expression of 34 genes to classify patients into the ccA and ccB subgroups. As the ccA/ccB signature proved to be a lead candidate throughout all the analyses, this signature was also tested for its applicability. ClearCode34 was significant in univariate analysis, and together with tumour stage in MVA, if the ccA/ccB signature was omitted (Appendix F). Although the HR for ClearCode34 in the MVA was lower (HR=2.23) than that of the ccA/ccB signature (HR=2.95), the implementation of this 34-gene signature may be easier in clinical practice than the 110-gene ccA/ccB signature. However, for the purposes of this work, the cluster assignments obtained using the full 110-gene panel were used for all further analyses.

4.3.3 Recursive partitioning reiterates the importance of ccA/ccB subgroup status

The multivariate Cox regression model developed in this analysis enabled a consensus set of prognostic markers to be selected in an unbiased manner. A tree based recursive partitioning was applied, using these markers, namely, tumour stage and ccA/ccB subgroup status, to propose a risk stratification model. This analysis highlighted key points; firstly patients with tumour stage I and expressing the ccA expression signature represented the low risk groups, with no deaths for over 6 years within this subgroup. Secondly, while the ccA subgroup showed significantly different survival between stages I and stage II/III cases, this was not true for the ccB subgroup. Finally, even within the Stage IV cases, for cases expressing the ccA subgroup signature, the median survival was almost twice that of those in the ccB subgroup (Figure 4.6). This analysis reconfirmed the results observed by Cox regression analysis and the stage-wise logrank test for ccA/ccB subgroup status (Figure 4.5 and Table 4.5).

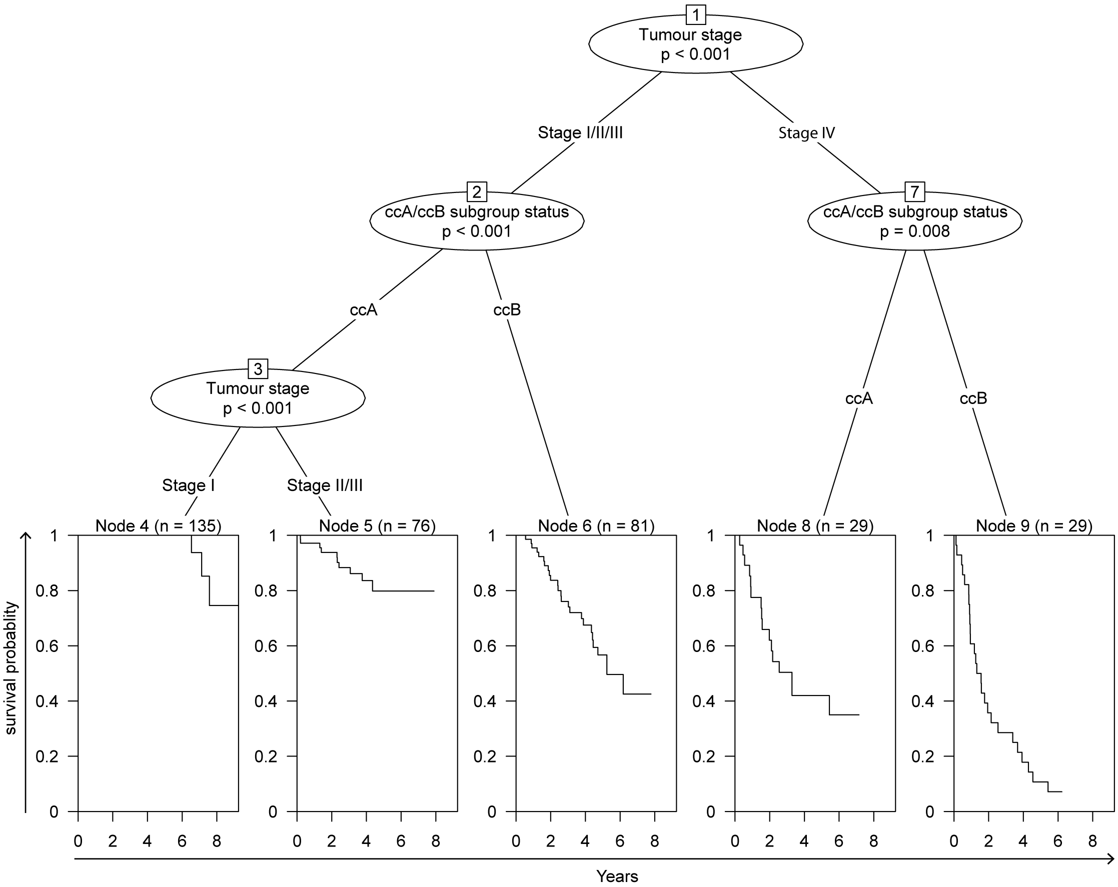


Figure 4.6: A prognostic model based on tumour stage and the ccA/ccB subgroup status of patients

Risk stratification based on recursive partitioning using only tumour stage and the ccA/ccB subgroup status of patients. Each node is split based on logrank p ≤ 0.05. The bottom panel of the figure shows the Kaplan Meier curve for each risk subgroup.

4.3.4 Comparison of the ccA/ccB signature with other prognostic scoring schemes

The SSIGN prognostic scoring system is a validated and a widely used scoring measure used to predict ccRCC prognosis and is based on stage, grade and tumour necrosis (Frank et al., Ficarra et al., Zigeuner et al.). As tumour necrosis data was missing for 16 of the 350 cases in our cohort, this measure was not included in our main analyses; however it was compared with the ccA/ccB signature in a separate analysis. In the multivariate setting, the ccA/ccB signature was significant when adjusted for the SSIGN score. CSS of patients whose tumours displayed the ccA or ccB signature were significantly different in three out of five validated SSIGN score categories as seen in logrank tests (Ficarra et al., Zigeuner et al.) (Table 4.6 and Figure 4.7).

Table 4.6: Multivariate analysis with the SSIGN score and ccA/ccB subgroup status
Table shows the results of Multivariate Cox analysis considering both the SSIGN score and the ccA/ccB subgroup of the patients. As shown, even when adjusted for the SSIGN score, the ccA/ccB subgroup status still remains significant predictor of CSS.

Variable	Hazard Ratio (C.I.)	p-value
SSIGN Score		
SSIGN 0-2	1.00 (Ref)	
SSIGN 3-4	2.69 (0.64 – 11.29)	0.175
SSIGN 5-6	8.28 (2.28 – 30.06)	0.001
SSIGN 7-9	13.23 (3.92 – 44.61)	<0.001
SSIGN ≥ 10	34.73 (10.29 – 117.20)	<0.001
ccA subgroup ccB subgroup	2.24 (1.38 – 3.64)	0.001

The ccA/ccB signature could not be compared with other clinical nomograms, such as haemoglobin, LDH, ECOG performance and UISS score (Motzer et al., Heng et al., Zisman et al., Sorbellini et al.) as essential parameters were not available for the majority of patients in the TCGA cohort.

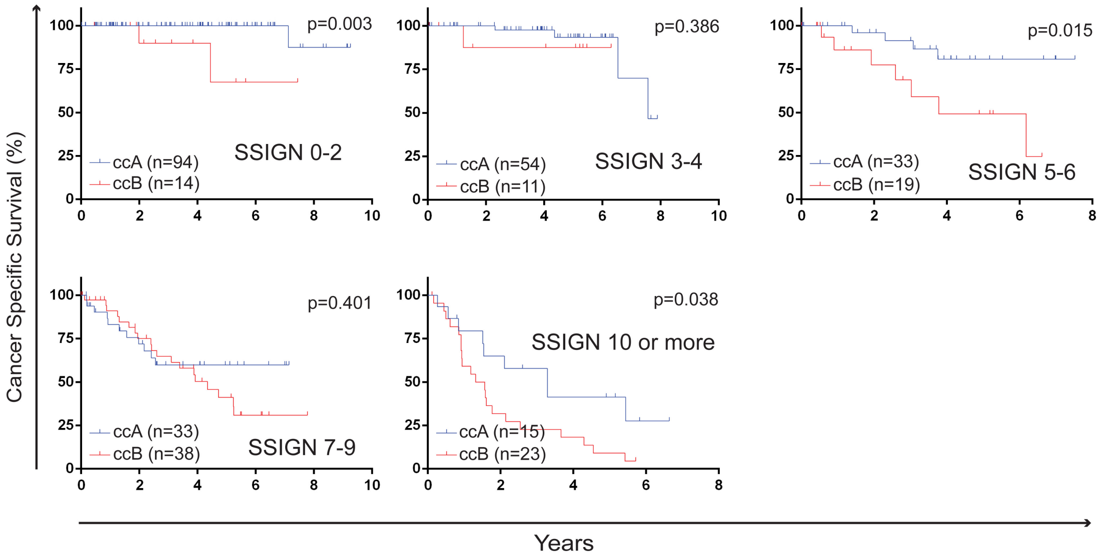


Figure 4.7: Kaplan-Meier cancer specific survival estimates for ccA/ccB subgroups split by SSIGN score classes (n=334)
Cases were split by SSIGN score categories (Ficarra et al., Zigeuner et al.) and Kaplan-Meier estimates were reassessed based on ccA/ccB subgroup status for each category. The subgroup status showed significant association with CSS in three of the five assessed categories.

4.3.5 ITH of the ccA/ccB signature

In the Introduction, section 1.2.5, the evidence of extensive levels of Intratumour heterogeneity in ccRCC was discussed (Gerlinger et al., Gerlinger et al.). Also, in (Gerlinger et al.), it was shown that the ccA and the ccB signature were present simultaneously within an individual ccRCC. The results discussed so far in this chapter, were based on single-biopsy data, and at this stage of the analysis, it was important to evaluate whether ITH can lead to ccA and ccB signatures being displayed within a single tumour. To this end, previously analysed published gene expression data was reinvestigated (Gerlinger et al., Gerlinger et al.). 63 tumour regions from 10 stage II-IV ccRCCs were classified as ccA/ccB using the gene expression panel (Appendix G), and the results mapped onto the phylogenetic trees previously published for these tumours (Gerlinger et al.) (Figure 4.8). ccA and ccB signatures were observed to occur within the same tumour in eight out of the 10 cases and only two tumours homogenously expressed the ccA signature. This signifies the need to sample multiple tumour regions in order to reliably detect poor prognostic clones.

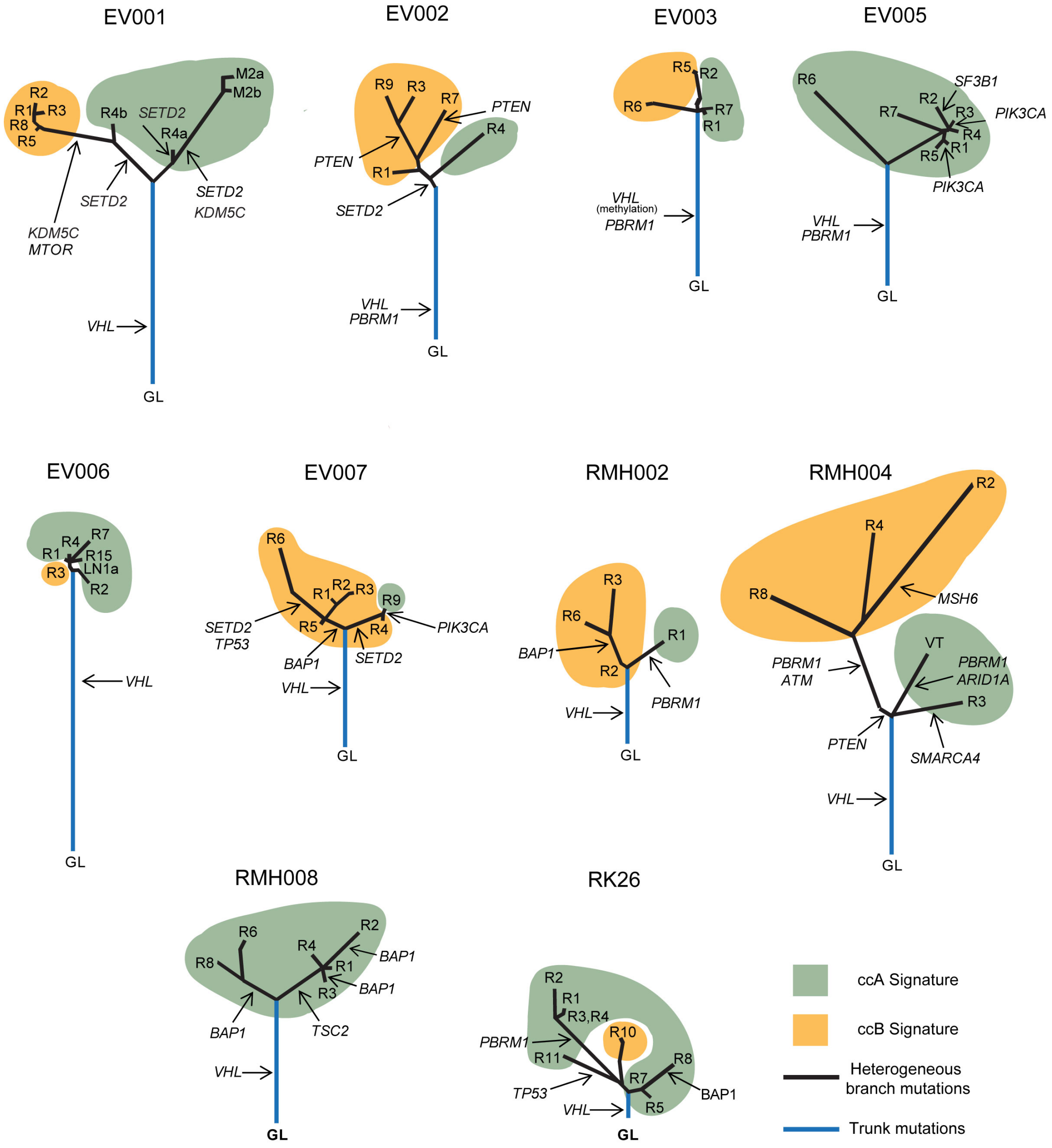


Figure 4.8: Heterogeneity analysis of ccA/ccB expression profiles
ccA and ccB profiles detected by consensus NMF clustering in a multi-region analysis dataset from 10 ccRCCs, which were mapped onto the phylogenetic trees of these tumours (adapted with permission from (Gerlinger et al.)). Regional gene expression signatures were assigned to the dominant clones detected within the region and the minority clones detected in some regions in the original publication have been omitted. This figure is as presented in (Gulati et al.).

4.4 Summary

In this biomarker study, performed in an independent validation cohort, out of 28 previously published genetic and transcriptomic prognostic ccRCC markers, 17 validated in logrank and competing risk analyses as predictors of CSS. However, only the ccB gene expression signature, along with tumour stage, was significant in the MVA. Taken together, this analyses suggested that the ccA/ccB gene expression signature outperforms other transcriptomic and genetic biomarkers for the prediction of ccRCC CSS and that it adds prognostic information to tumour stage and to the SSIGN prognostic model. This signature could be particularly relevant for the profiling of stage I ccRCCs where the detection of the ccA signature was associated with an excellent prognosis. Stage I ccA tumours may only require minimal follow-up whereas ccB tumours may benefit from more stringent surveillance and may therefore be good candidates for adjuvant therapy trials.

Evaluation of the ccA/ccB signature across multiple tumour regions from each of 10 stage II-IV ccRCCs demonstrated heterogeneous expression patterns with ccA and ccB signatures co-existing in 8/10 cases. ITH within spatial separated subclones, that may harbour distinct transcriptomic profiles, demonstrates that single tumour biopsies are unlikely to reveal a complete picture of the landscape of even the best current binary classification ccRCC biomarkers. These observations highlighted the need for multi-region profiling of larger cohorts, which could help define how to integrate heterogeneity assessments into biomarker predictions and subsequently improve the accuracy of the ccA/ccB signature.

This study had a few limitations. Firstly, several candidate markers, which failed validation in univariate or multivariate analysis, such as Chrom19 deletion, Chrom8q amplification and BAP1 and TP53 mutations have low prevalence (≤10%); therefore this study may be underpowered to definitively assess the role of these markers. Secondly, biomarkers based on immunohistochemistry could not be assessed due to the lack of protein expression data for the validation cohort.

Chapter 5. Molecular drivers of the ccA/ccB signature

5.1 Introduction

The analyses in the previous chapter have shown that the ccA/ccB gene expression signature was the only biomarker, which validated as an independent predictor of patient prognosis even when adjusted for clinical established variables. Furthermore, although ITH was observed for this signature; in a single biopsy per patient setting, this signature was observed to successfully classify poor prognosis patients. While, questions regarding the impact of ITH on the accuracy of these predictions still remain unanswered, an important question is to ascertain what drives the poor prognosis ccB subgroup and how is it different from the ccA subgroup. This is imperative, since due to a marked absence of effective adjuvant strategies, prognostic ccRCC markers are of limited clinical utility. Thus in this chapter, an attempt is made to deconvolve the molecular mechanisms behind these expression signatures.

Previously, in the original publication for this signature, the authors revealed that genes overexpressed in samples with the ccA signature are enriched for genes implicated in angiogenesis, fatty acid-, organic acid- and pyruvate metabolism. Whereas genes overexpressed in samples displaying the ccB signature are enriched for cell differentiation, epithelial to mesenchymal transition, mitotic cell cycle control, response to wounding and TGFβ and Wnt signalling pathway regulation (Brannon et al.). In this chapter, these signatures were further explored. Three major objectives were set:

1. To determine the pathways deregulated in both ccA and ccB subgroups.
2. To elucidate drivers of ccA and ccB subgroups at the genetic level.
3. Finding driver networks and regulators for the ccA/ccB signature by performing genotype-phenotype analysis.

5.2 Methods

The methods used for the analyses in this chapter are briefly outlined below; for further details, a reference to the appropriate section is provided at the end of each section.

5.2.1 Patient cohort

In the analyses performed in this chapter, the original dataset of 469 ccRCC patients for which RNA-Seq was available for each patient from TCGA was used. Depending upon the requirements of the analysis, either raw count generated by the RSEM method or normalised RSEM counts to the upper quartile normal counts of the TCGA, were used. This has been clarified in the relevant sections. Normalised counts were log2 transformed before further analyses. In either case, only genes, for which the counts (raw and normalised resp.) were above 30 in at least 80% of the samples, were included in the analyses.

The ccA/ccB subgroups were determined for all 469 cases using supervised NMF clustering; however, when comparing with other variables the number of cases may be different depending upon data availability of the variable under consideration. Case counts (n) have been given in all such places.

5.2.2 NMF clustering

Expression data was available for 103/110 genes in gene expression signature used to validate biomarkers for ccRCC (Brannon et al.), and was submitted for consensus NMF clustering analysis (Brunet et al.) to the Broad Institute's GenePattern server (Reich et al.). The cluster number (k) range was predefined from two to 10. Each clustering run returned a cophenetic correlation coefficient that measures the stability of cluster assignments as well as a consensus clustering maps for the respective k value. Based on both these data, the optimal number of clusters could be identified. (Refer methods section 2.5.3)

5.2.3 Statistical analysis

Differential regulation analysis was performed using the edgeR (Robinson et al., 2010) package in R (R Development Core Team, 2013). (Refer methods section 2.8.1).Over representation analysis for gene ontology and pathways was performed using the MSigDB (Liberzon, Liberzon et al.) and genego portal (Thomson Reuters, https://portal.genego.com/). (Refer methods section 2.8.2). Gene Set Enrichment Analysis was performed using the standalone tool from the Broad Institute using the curated pathways dataset as background (Subramanian et al., 2005). (Refer methods section 2.8.3). To evaluate the differences in the occurrences of genetic alterations in the ccA/ccB subgroups, Fisher's exact test was performed in R to calculate the odds ratio and the p-value of significance.

5.2.4 Weighted Genomic Instability Index (wGII)

The weighted genomic instability index (wGII) (Burrell et al.), is a measure of the overall copy number alterations within a tumour genome, and thus provides a score for the level of genomic perturbations within tumour samples; a numeric score between 0 and 1 is returned where a wGII ≥ 0.2 is considered to be genomically unstable (Lee et al.). The returned numbers are a weighted average of the deviation from the proportion of bases on each chromosome away from the sample ploidy. (Refer methods section 2.2.6)

5.2.5 Machine learning: random forest

In order to find the more important features to classify the ccA (n=240) and ccB (n=110) subgroupings the Random Forest (RF) machine learning protocol, in R, was used; the package randomForestSRC (Ishwaran et al., Ishwaran and Kogalur). In this implementation, the number of trees was set to 1000 for each iteration; all other parameters were set to their default values. (Refer methods section 2.7.2). After training, the random forest feature importance values were invoked (Breiman). This gives a ranked list of the features, which were most important for accurate classification, and hence features which can discriminate best between ccA and ccB.

5.3 Results

The ccA/ccB subgroup status was ascertained for the 469 cases for which RNA-Seq data were available. To meet the objectives set out in the introduction of this chapter, firstly a pathway enrichment analyses was performed using the gene expression data to find out major pathways controlling the ccA and ccB subgroups. Following this the putative genetic drivers of the ccA/ccB subgroups were ascertained through enrichment analyses using Fisher's exact tests. Subsequently, genotype – phenotype relationships were elucidated for these subgroups.

Consensus NMF clustering led to the classification of 305 cases belonging to the ccA subgroup and 164 cases belonging to the ccB subgroup (Figure 5.1). Clinically, the ccA subgroup tended to be more Stage I/II than Stage III/IV, while the ccB subgroup tended to consist of more from the higher stages (n=417, OR=3.43, p<0.001). There were no statistical differences in the age of patients in the two subgroups (n=417, p=0.348). In the full cohort of patients, for whom both SCNA and RNA-Seq data were available, chromosome 3p was observed to be more enriched in the ccA subgroup (n=422, OR=0.31, p<0.001). In contrast to the primary study (Brannon et al.), in this cohort of cases there were higher odds of ccA patients having a non-synonymous VHL mutations than the ccB patients (n=390, OR=0.48, p<0.001). Median survival for ccA was not yet reached at the end of the study period where as for ccB subgroup this was 4.45 years (n=415, n_{ccA}=281, n_{ccB}=134).

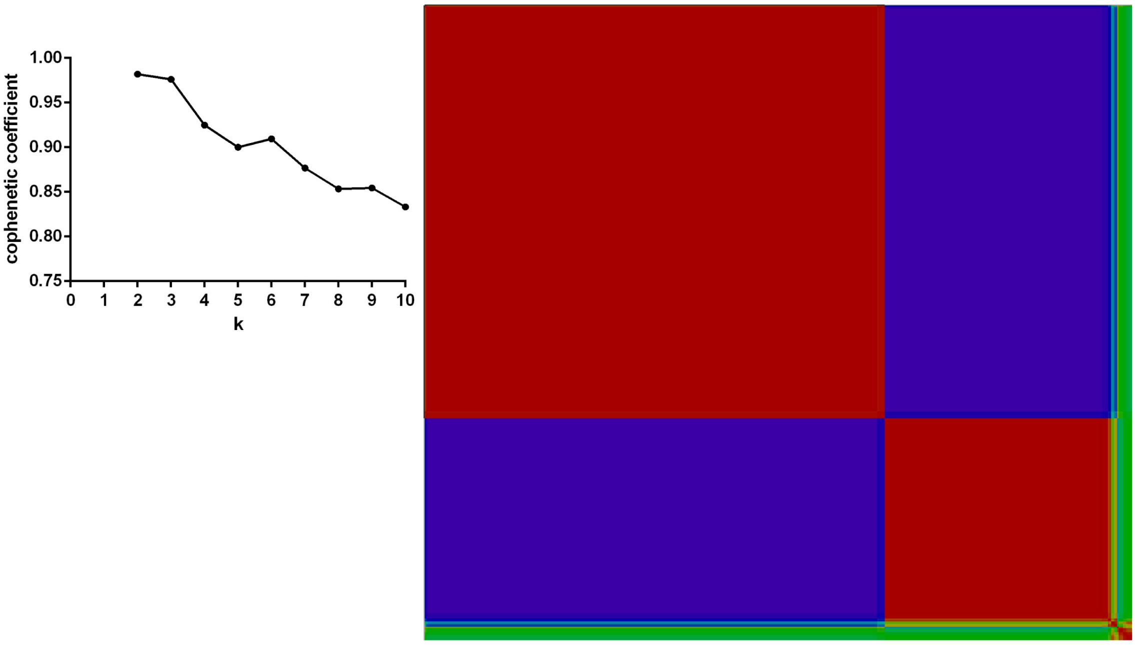


Figure 5.1: NMF clustering results for the ccA/ccB gene panel
A. Cophenetic correlation coefficient. The cophenetic coefficient depicts the strength of clustering at different values of k. The coefficient was highest for k=2. B. Consensus clustering matrix for k=2.

5.3.1 Differential regulation analysis

Differential gene expression analysis was performed using the edgeR (Robinson et al.) package, for three separate comparisons. Genes deregulated in ccA cases when compared to normal kidney samples in the first instance, second comparison was between ccB and normal cases and the last analysis was performed by comparing ccA cases to ccB. The edgeR output provides three different values for each gene in the input namely log fold changes (log FC), log CPM, and the p-value. A list of significantly differentially regulated genes can then be generated using p-value and/or fold change cut-offs. However, another approach is to rank the input list of genes using either the fold changes or p-values and then run a Gene set enrichment analysis (GSEA). Both analyses were performed for all pairwise comparisons.

5.3.1.1 Differentially regulated genes

To generate a list of significantly differentially regulated genes, all p-values were first corrected for multiple testing. The final list of differentially expressed genes was compiled using a statistical FDR q-value ≤ 0.05 and a FC cut-off of |log FC| ≥ 2.5 (equivalent to |FC| =5.6) as selection criteria, for each comparison.

When comparing ccA and ccB gene expression levels to normal kidney cell expression levels, some genes were identified to be deregulated in both subgroups. This was not unexpected since both are ccRCC subtypes. After removing these genes, 235 genes were obtained to be uniquely deregulated in the ccA subgroup, and 539 genes were deregulated in the ccB subgroup, only.

Gene ontology (GO) term enrichment and pathway over-representation analysis on these genes were performed using two sources, namely MSigDB (Broad Institute (Liberzon, Liberzon et al.)) and the genego portal (Thomson Reuters, https://portal.genego.com/).

MSigDB was used to obtain the top 100 pathways and the top 100 GO Biological Processes (BPs), enriched respectively in the ccA vs. Normal samples (Appendix H) and ccB vs. Normal samples (Appendix I). In the pathway enrichment analysis only 13 pathways were considered to be enriched for the ccA subgroup using the FDR q-value ≤ 0.05 as cut off, whereas the ccB subgroup had 100 pathways enriched of out of which only five were also observed in the ccA subgroup. The five pathways which were present in both lists, included genes involved in developmental biology, PDGF signalling, homeostasis, axon guidance and peptide ligand-binding receptors. ccA samples were enriched for genes involved in glycine, serine and threonine metabolism, GPCR signalling and transmembrane transport. Whereas, genes involved in mitotic cell cycle and cell cycle checkpoints, immune signalling, immune response activation genes, regulation of the immune system were enriched specifically in the ccB subgroup.

For the GO enrichment analysis, 44 BPs were seen to be enriched for from the ccA list of differentially expressed genes, whereas a total of 100 BPs were obtained from the ccB list, 12 of which were shared with ccA. The ccA subgroup showed an enrichment for genes involved in genes involved in cellular transport, G protein coupled signalling while the ccB subgroup showed similar results to the pathway enrichment with significant enrichment of mitotic cell cycle, cell cycle regulation and checkpoint control genes, apoptosis control and immune regulation. The results from genego portal did not add any significant new pathways to the above results.

As a final check, differential expression was also tested specifically between the ccB subgroup vs. the ccA subgroup. 144 genes were obtained to be deregulated with FDR q ≤ 0.05 and |log fc| ≥ 2.5. Enrichment analyses for this list showed differential regulation of genes involved in organ development, immune response, response to external and internal stimuli and apoptosis pathways. This comparison highlighted the differential regulation of these pathways within ccRCC subgroups, and may indicate their role in the pathogenesis of the poor prognosis ccB subgroup.

5.3.1.2 Gene set enrichment analysis (GSEA)

Furthermore, gene set enrichment analyses was also performed for all the three comparisons under consideration. When both of the ccA and ccB subgroups were compared to the Normal samples, using the FDR q-value ≤0.05 as cut off, 90 significantly enriched pathways were obtained for the ccA subgroup and 150 pathways were enriched for in the ccB subgroup. As previously observed (Brannon et al.) the metabolic pathways showed a down-regulation in both cohorts. In the ccA subgroup, up-regulation was seen for phagosome and myogenesis pathways while genes belonging to the oxidative phosphorylation pathway and signalling by ERBB4, and amino acid synthesis pathways, showed a down-regulation. In contrast, in the ccB subgroup, significant up-regulation was seen primarily for genes involved in extracellular matrix (ECM) reorganisation, ECM-receptor interaction, mitotic cell cycle and cell cycle check point genes, while genes belonging to histidine metabolism, pyruvate metabolism, glycolysis, glucose transport and transmembrane transport pathways in general, were seen to be significantly down-regulated.

These results suggest that the ccA subgroup is primarily controlled by metabolic pathways, which is representative of the metabolic nature of ccRCC. However, the ccB subgroup has additional deregulation of immune signalling pathways and cell cycle checkpoint genes, which may be contributing to the aggressive nature of this subgroup.

5.3.2 Molecular drivers of the ccA and ccB subgroups

To determine if the molecular drivers of the ccA/ccB expression subgroups could be associated with the genetic factors assessed in the prognostic analyses described in Chapter 4, the following analysis was performed.

All the genetic prognostic factors that were found in the literature search (n=17*, duplicate entries of Chrom9p and Chrom20 were removed based on lower (HRs)), were categorised as those that validated in the log-rank tests (n=9) and those that failed to validate (n=8). The ccB expression signature was then investigated to see if it might reflect the transcriptomic impact of the poor risk genetic alterations, which were significant in logrank analysis but failed in the multivariate analysis. For this analysis, the cohort of all 350 cases, as devised in Chapter 4, was used. Seven out of the nine poor prognosis genetic alterations (BAP1 and TP53 mutations; Chrom8q, Chrom12 and Chrom20q focal amplifications; Chrom9p and Chrom22q deletions) were significantly enriched ($p<0.05$) in the ccB subgroup (Figure 5.2). In contrast, on repeating the analyses for the eight candidate genetic markers that had failed univariate validation, only two were found to be enriched in ccB samples (Figure 5.3).

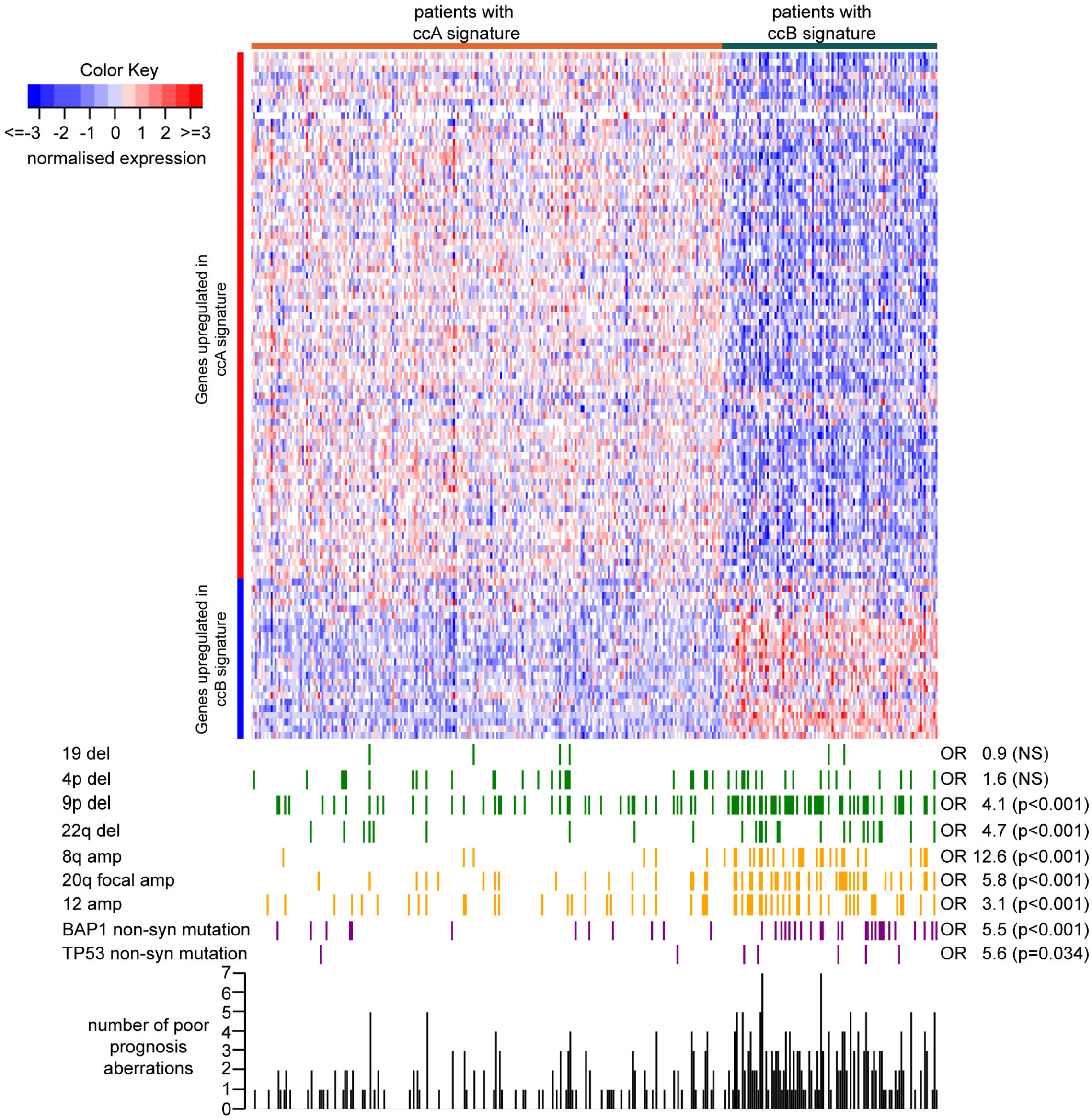


Figure 5.2: Enrichment analysis for the poor prognosis genetic events in the ccA/ccB subgroups

The top part of the figure depicts a heatmap, showing the gene expression of the 103/110 gene panel (Brannon et al.). The ccA subgroup is represented on the left and the ccB subgroup on the right. The bars below the heatmap depict the occurrence of the genetic events in each patient. The odds ratio of the occurrence of these events in the ccB subgroup with respect to the ccA subgroup is given on the right side, along with a p-value of significance for the odds (Fisher's exact test). The barchart at the bottom of the figure depicts the total number of these events per patient. The highest number of these events occurring in a single patient is seven, with both of these cases belonging to the ccB subgroup. This figure is as presented in (Gulati et al.).

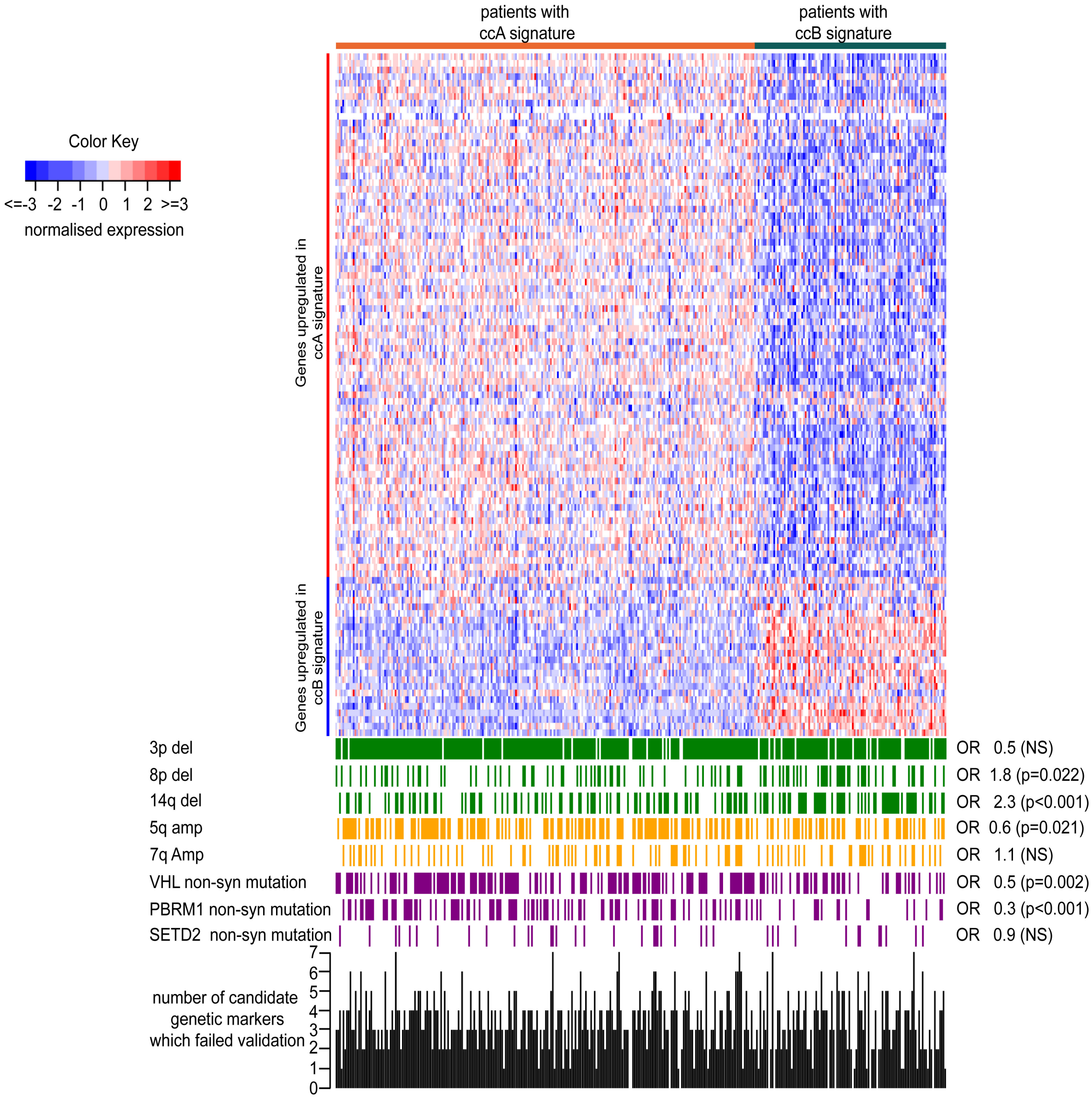


Figure 5.3: Enrichment analysis for the genetic events, which failed validation in the ccA/ccB subgroups

The top part of the figure depicts a heatmap, showing the gene expression of the 103/110 gene panel (Brannon et al.). The ccA subgroup is represented on the left and the ccB subgroup on the right. The bars below the heatmap depict the occurrence of the genetic events in each patient. The odds ratio of the occurrence of these events in the ccB subgroup with respect to the ccA subgroup is given on the right side, along with a p-value of significance for the odds (Fisher's exact test). The barchart at the bottom of the figure depicts the total number of these events per patient. The highest number of these events occurring in a single patient is seven, with both of these cases belonging to the ccB subgroup. This figure is as presented in (Gulati et al.).

Further assessment of these aberrations showed that about 72% of the ccB samples had at least one of the seven enriched aberrations in contrast to only 30% of ccA samples (Figure 5.4A). Both, the maximum and the median number of the poor prognosis aberrations per sample were higher in the ccB group than in the ccA group (Figure 5.4A and 5.4B). However, when the distribution of aberrations which failed validation in the prognostic analysis was compared, the median number of these aberrations between ccA and ccB samples was not statistically different (Figure 5.4C and 5.4D).

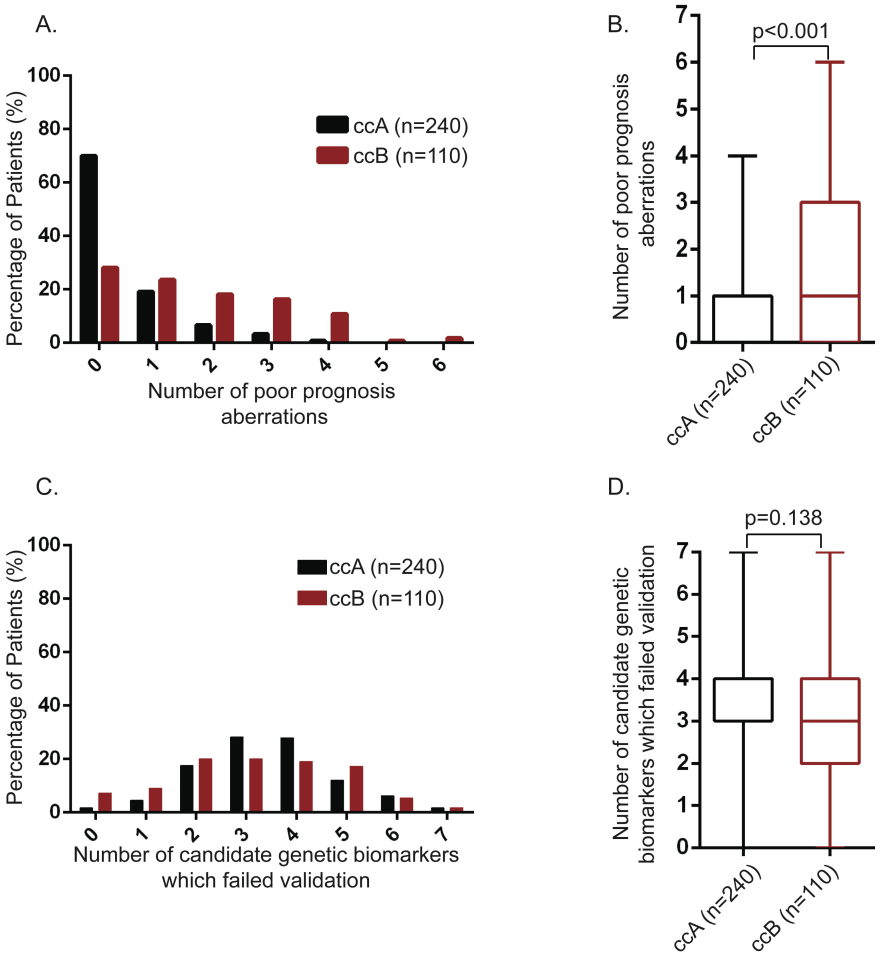


Figure 5.4: Comparison of genetic markers in the ccA/ccB subgroups
A. Comparison of the number of poor prognosis genetic aberrations per sample between ccA and ccB subgroups. Only aberrations, which are enriched in the ccB subgroup, were considered. B. Box and whisker plot comparing median number of poor prognosis genetic aberrations between samples assigned to the ccA and the ccB group. C. Comparison of the number of number of genetic aberrations, which did not pass univariate validation per sample between ccA and ccB subgroups. D. Boxplot and whisker plot showing the median number of genetic aberrations, which did not pass univariate validation between ccA and ccB subgroups.

Chromosomal instability is known to foster the acquisition of SCNAs and has been associated with poor prognosis in several cancers (McGranahan et al., 2012). To reveal whether enrichment of chromosomal aberrations in ccB was a result of increased chromosomal instability, the weighted Genomic Instability Index (wGII), a measure of overall copy number aberrations, was calculated for each sample (wGII ≥ 0.2 is considered unstable (Lee et al.)). The ccB samples had significantly higher wGII scores when compared to ccA samples (p<0.001, Figure 5.5A). However, the mutation load was not significantly different between the two cohorts (p>0.05, Figure 5.5B and 5.5C). Based on these results, it appears possible that the aggressive ccB phenotype is partially driven by several poor prognosis genetic alterations, co-occurring within these samples, which may be permitted by a cancer genomic background of elevated chromosomal instability.

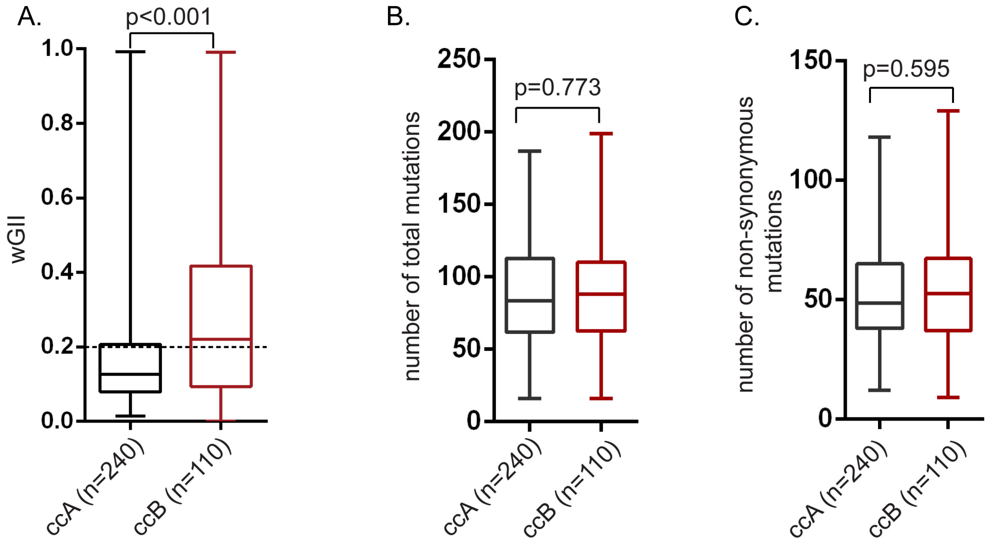


Figure 5.5: Comparison of genomic measures in the ccA/ccB subgroups
Box and whisker plots comparing genomic factors between the ccA/ccB subgroups. A. Comparison of wGII between the two cohorts where wGII ≥ 0.2 is deemed to be genomically unstable; ccB patients were observed to be more genomically unstable. B. and C. compare the total mutation load and the number of non-syn mutations between the two subgroups respectively. No statistical differences were seen between the two cohorts.

5.3.3 Random forests elucidate the most important determinants of the ccB subgroup

At this point, it was decided to test the hypothesis, that while high genomic instability fosters the development of the aggressive ccB subgroup, it is not the absolute determinant of the aggressiveness of this subgroup. Since an important part of the random forest classification method is to report the weights of each factor contributing to the classification, this method was chosen to test the hypothesis. Multiple iterations were performed with different sets of events as variables to first determine the most accurate set of variables, which could distinguish between the ccA and ccB subgroups (Table 5.1). Most accurate (least error rate) classification was achieved using all the variables under consideration in the classifier (Iteration 6, Table 5.1). The variables from this iteration were ranked to find the most important variables that were able to distinguish between the ccA and ccB subgroups (Figure 5.6). In decreasing order of importance, Chrom8q, Chrom20q and Chrom5q amplification status, the BAP1 non-synonymous mutations and deletion of Chrom9p were observed to be the most important variables for distinguishing between the ccA and ccB subgroups; followed by genomic instability. It should be noted here, none of the variable sets achieved high accuracy on cross validation, especially when predicting the ccB subgroup; however, since the aim of analysis was not to build a classifier but to determine the factors more likely to be important for distinguishing between the subgroup. It may be said with some confidence that while a genomically unstable background fosters the aggressive subtype, other unknown factors, along with the ones discussed in this chapter, are important for the establishment and progression of the cancer

Table 5.1: Random Forest Iterations
Table gives the results as obtained by multiple iterations of the random forest analysis.

Iteration	Variable set	Variables	Overall error rate	ccA error rate	ccB error rate
1	Clinical	Tumour Stage and Fuhrman Grade	25%	7%	66%
2	Mutations	VHL, PBRM1, SETD2, BAP1 and TP53 non-syn mutation status	26%	6%	70%
3	SCNAs	Amplifications of 5q, 7q, 8q, 12 and 20 and deletions of 3p, 4p, 8p,9p, 14q, 19, 22q	26%	13%	54%
4	Variables significant in log-rank analysis and wGII	wGII , BAP1 , TP53 , 8q , 12 , 20q ,4p, 9p, 19 , 22q	29%	15%	61%
5	Variables which did not validate in log-rank analysis and wGII	wGII , VHL , PBRM1 , SETD2 , 5q ,7q, 3p, 8p , 14q	27%	15%	52%
6	All genetic variables (n=17) and wGII	wGII , VHL , PBRM1 , SETD2 , BAP1 , TP53 , 8q , 12 , 20q , 4p , 9p , 19 , 22q , 5q , 7q , 3p , 8p , 14q	25%	15%	46%

Each row represents 1 iteration of the analysis, and the details for the variables considered and error rates: overall error rate and error rate in specifically predicting the ccA and ccB subgroups.

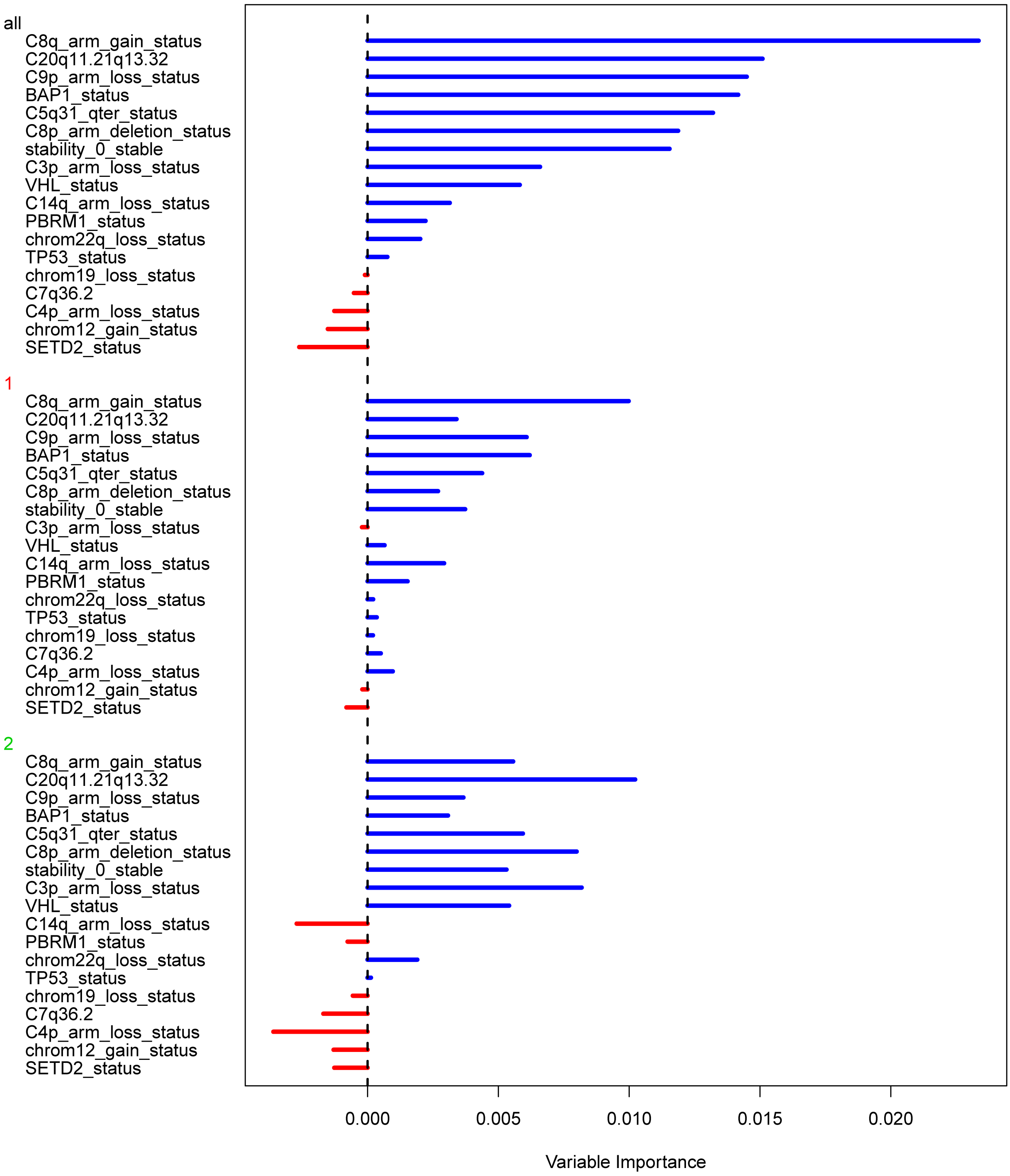


Figure 5.6: Variable importance calculated from Random forest analysis
The bar charts depict the importance of various factors when distinguishing between the ccA/ccB subgroups. The top bar chart depicts the importance of each variable when predicting the overall classification, while the remaining two depict the importance of each variable when specifically differentiating the ccA subgroup (1) and the ccB subgroup (2). Positive importance is shown as blue bars whereas negative is shown as red bars.

5.3.4 Genotype – phenotype relationships

In the above analyses, both genetic and transcriptomic have ascertained that there are clear differences in the ccA and ccB subgroups. This led to the question of finding what genetic alterations may be controlling the transcriptomic regulation of these two signatures, identifying drivers and thereby suggesting targets for therapy.

Interestingly, results from employing the random forest method also indicated that taking into account only the above 17 genetic events, along with wGII, is not sufficient to accurately distinguish between the ccA and ccB phenotypes. However, investigating genotype - phenotype relationships in a very simplistic manner, simply by comparing the co-occurrence of events relative to the ccA/ccB subgroups, some putative driver events can be identified.

Apart from BAP1, Chrom8q, Chrom20q and Chrom9p, all have genes involved in the ubiquitin mediated proteolysis pathway (UMPP), all of which have higher odds of alteration in the ccB subgroup; which may explain the higher deregulation of the UMPP in this subgroup of ccRCC cases. Further, Chrom8q and Chrom20q also have genes involved in the ECM receptor interaction pathway, which is deregulated in the ccB subtype.

5.4 Summary

In this chapter the ccA/ccB signature is explored in detail, with the objective of explaining the mechanisms that distinguish the aggressive ccB subgroup from ccA. The analyses confirmed previously known observations but also add further information and understanding. As discussed in the introduction of this chapter, previous work had revealed that genes overexpressed in samples with the ccA signature are enriched for genes implicated in angiogenesis, fatty acid-, organic acid- and pyruvate metabolism. Genes overexpressed in samples displaying the ccB signature are enriched for cell differentiation, epithelial to mesenchymal transition, mitotic cell cycle, response to wounding and TGFβ and Wnt signalling genes (Brannon et al.). In the differential gene expression analysis, while the ccA subgroups showed a deregulation of genes involved in metabolic pathways, in concordance with previous results, the ccB group in addition showed significant deregulation in immune regulation and cell cycle checkpoint pathways.

Seven out of nine specific genetic alterations, which were validated in univariate analysis, were shown to be enriched in ccB samples with 72% of samples harbouring at least one and up to six of these alterations per patient. These genetic changes were only found in 30% of the ccA samples with a maximum of four aberrations per sample. Thus, the ccB signature may reflect the transcriptomic impact of these poor prognosis alterations, but more than one alteration may be necessary to establish this phenotype and as yet unknown alterations are also likely to contribute. We had also pointed out that prognostic markers are of limited clinical utility in ccRCC due to the current absence of effective adjuvant strategies. Further study of the interplay of these genetic aberrations and the pathways deregulated in the ccB signature are clearly necessary in order to reveal the mechanisms and biological implications of the ccB phenotype. Such insights could eventually foster the development of specific therapeutic approaches for poor prognosis ccRCCs.

Chromosomal instability indices (wGII) were higher in ccB than in ccA samples; however; the mutational load was not statistically different between these two cohorts. Random forest analysis also ranked individual alterations higher than wGII, when distinguishing between the two subgroups. These data suggest that chromosomal instability may catalyse the evolution of the ccB phenotype by providing the permissive heterogeneous genomic background, from which these genetic alterations can be selected, but it may not be the sole driver of the aggressive subtype. These results are hypothesis generating and will require further study.

In a simple genotype-phenotype analysis, putative drivers of the ccB signature can be identified; however, a complete picture of specific drivers of the ccA/ccB subgroups could not be discerned. There are a number of confounding factors. Firstly, there may be other elements, such as epigenetic alterations and methylation patterns, contributing to the aggressive subtype, which have not been considered as part of this overall analysis. Secondly, intratumour heterogeneity is not taken into consideration and this is likely to play an important role, not least because some tumours may actually consist of both a ccA and ccB cellular phenotype. Finally, as exemplified by the relatively high error rates associated with the ccA/ccB classification when employing the random forest methodology, it appears various factors such as cohort sizes and the accuracy of some of the calculated features, is not yet sufficient to attempt a full and definitive classification for the drivers and dynamic network attributes associated with each cancer subgroup.

Chapter 6. Epilogue

In this thesis, I have reported on ccRCC in terms of its biology and prognosis using multi ‘omics’ datasets. Primary aims were to evaluate the catalogue of genetic alterations, understand the molecular mechanisms driving ccRCC evolution, come up with better molecular markers to improve prognostication and even to predict novel therapeutic avenues. Chapter 3 covered the biology led analyses. Chapter 4 covered the prognostic analyses, with both these chapters culminating in a description of the molecular mechanisms of the lead prognostic marker, namely the ccA/ccB gene expression subgroup, described in Chapter 5.

Today, NGS technologies permit analyses at nucleotide resolution for both protein coding regions (exomes) and whole genomes. Through deep coverage, it is possible to detect mutations that occur in even a small population of cells, allowing the subclonal architecture of tumours to be inferred (Nik-Zainal et al.). Over the past few years, this has led to the identification of multiple new ccRCC genes, increasing our knowledge of the disease. Our research (Chapter 3) as well as well as that of others (The Cancer Genome Atlas Research Network, Gerlinger et al.), have established loss of chromosome 3p and biallelic mutations in the VHL gene as the two key events in ccRCC. Moreover, mutations in major chromatin modifiers such as PBRM1, SETD2 and KDM5C were also verified as key recurrent events. In terms of SCNAs, amplification of chromosomes 5q, 8q, 12 and 20q as well as deletion of chromosomes 9p, 8p and 14q were observed to be potential driver events. At the gene expression level, we were able to confirm the existence of at least two subgroups for ccRCC (Brannon et al.). Further gene expression analysis showed ccRCC to be primarily a metabolism driven cancer and that the more aggressive subgroup has higher deregulation of immune signalling pathways (Chapters 3 and 5).

However, inter-patient as well as intratumour heterogeneity, especially taking into account the fact that most genetic events in ccRCC are rare, raises important questions as to the pathogenesis of the cancer. Inter-patient heterogeneity has been thought to be the most common reason for the diversity in patient outcomes even between tumours with the same histology, stage and grade (Gerlinger et al.). Nevertheless, computational network analysis algorithms have been successfully utilised to study cancer and disease genome previously and therefore provided the platform to explore these rare events in ccRCC. In Chapter 3, multiple state-of-the-art algorithms were used, leading to the finding of multiple chromatin regulators as the major players in ccRCC biology. While this has been observed by others (The Cancer Genome Atlas Research Network), our analysis on co-altered modules further emphasised their role in ccRCC pathogenesis.

Furthermore, in Chapter 3, using the ARACNE and MARINa algorithms, putative drivers at the gene expression level were identified. This provided an indication of some important genes controlling the gene expression of ccRCCs; however, further analysis of these genes is warranted. One important exercise could be to compare if any of these drivers are regulated in other cancer types. This may provide important clues as to the mechanism of actions of these genes and also validate the analysis presented in this thesis.

Compounding this inter-patient heterogeneity further, is the intratumour heterogeneity, whereby molecular characteristics vary within individual tumour biopsies. ITH complicates the precise molecular profiling of tumours and thereby constitutes a major hurdle in identifying optimal patient therapy. Our capabilities to detect and characterise ITH have improved significantly as the sequencing technologies no longer present the biggest limitation (Gerlinger et al.). Nevertheless, the challenge now is to develop optimal sampling technologies that will enable the identification of somatic alterations across different regions from the same tumour and associated metastasis, which would enable effective modelling of the dynamics of tumour evolution.

Cancer evolution has always been depicted as a linear evolution over time, where successive accumulation of alterations provides corresponding increases in fitness (Gerlinger et al.). However, work by and with our colleagues has refuted this claim and shown branched evolution patterns for 10 ccRCCs that were analysed through multiregion biopsies (Gerlinger et al., Gerlinger et al.). Moreover, analyses presented in this thesis on assessment of the heterogeneity of the ccA/ccB gene expression signature, has shown heterogeneity for this signature within individual tumour regions (Gulati et al.).

In terms of prognostic biomarkers, as shown in Chapter 5, 17 out of the 28 published genetic and transcriptomic prognostic ccRCC markers could be validated in logrank and competing risk analysis as predictors of CSS for an independent validation cohort. Of those, only the ccB gene expression signature was significant in MVA. Tumour stage was the only other independent predictor of CSS in MVA. Importantly, the ccA signature identified patients with Stage I ccRCCs who had an excellent prognosis with no cancer specific deaths over more than 6 years of follow up. The ccA/ccB signature was also significant in MVA with the established SSIGN prediction model, demonstrating that this molecular marker can add additional information to one of the best currently available predictors based on clinical and pathological information. Thus, the ccA/ccB signature could refine personalized follow up strategies or stratification into adjuvant therapy trials. The novel ClearCode34 signature is based on the ccA/ccB signature but can be assessed from 34 instead of 110 genes. The performance of this new marker was slightly inferior but it may nevertheless be valuable as clinical adoption may be easier (Gulati et al.).

Alternatively, as discussed in section 1.4.2 of the Introduction, methods combining network analysis with cox regression analysis (NetCox) can be employed to build more robust prognostic signatures. NetCox based strategies may prove to be an effective pipeline to overcome the problems associated with rare genetic alterations.

It also provides a method to combine different transcriptomic and epigenetic data types such as somatic mutations, SCNAs, methylation data, and RNA-Seq data.

Therapeutically speaking, several actionable driver genetic markers were shown to be subclonal, for example mutations in the MTOR gene (Gerlinger et al.), thereby raises questions as to the suitability of these markers as therapeutic targets. It has been suggested that targeting the alterations on the trunk (ubiquitous alterations) may be an effective clinical strategy (Yap et al.); however, thus far identified definitive ubiquitous events in ccRCC are limited to mutations in VHL gene and loss of chromosome 3p. Targeted therapies against VEGF, which is downstream to VHL mutations, have been previously tested, but as discussed in the Introduction, these are no longer thought to be the most effective route for treatment.

The work described here (Chapter 6), along with the work of others (Brannon et al.), has also shown that the aggressive ccB subtype is associated with genes involved in vascular, immune response, inflammation, cell cycle progression, and proliferation pathways. It has been suggested that this may explain the ineffectiveness of existing targeted treatment strategies as they mainly tackle angiogenesis pathways; hence these treatments might preferentially target vascularised tumours and are in turn ineffective in the treatment of the highly aggressive hypoxic renal cell carcinomas (Kroeger et al.). Likewise, targeting multiple alterations from different subclones might also provide more effective treatment strategies and help in improving outcomes. Furthermore, immune therapies which are independent of the heterogeneity of single target genes have been suggested to have potential as they may overcome ITH (McGranahan and Swanton). Indeed, before the emergence of antiangiogenic targeted therapies, immune based therapies were the method of choice for metastatic ccRCC care. Curative responses had been observed for IL-2 based therapies, whereas only delayed progression has been shown for targeted therapies (Fyfe et al., McDermott et al., Yang et al.). Recently, there has been a resurrection of immune therapies for ccRCC in the form of immune checkpoint inhibitors, which have shown good results for melanoma and are now in clinical trials and show promising early stage results (Naidoo et al., Topalian et al., Yang et al.). A recent review also comprehensively assesses the progress of PD1/PDL1 inhibitors for Urologic cancers (Carosella et al.).

Further extending on the topic of prognostication models, recently, Rini and colleagues (Rini et al.) investigated the association between outcome post-nephrectomy for ccRCC patients and the expression levels of 732 genes in a cohort of 942 cases. They selected 11 best performing genes that represent key ccRCC pathways, combining them with 5 reference genes to develop a recurrence score. They then validated this score in a cohort of 626 cases.

To address ITH, Rini et al. (Rini et al.) focused on 8 cases. They used two representative formalin-fixed paraffin-embedded (FFPE) blocks for each case and sampled 3 sections from each block. They concluded that little or no intratumour heterogeneity was associated with their score. However, in our analysis of more extensively sampled, albeit more advanced-stage tumours (Gerlinger et al.), we observe pervasive intratumour variability of expression of the 11 genes from the Rini score (Figure 6.1). Thus, our analysis does not support the authors' conclusion that this transcriptomic signature is a truncal event that can be fully captured in a single biopsy approach (Gulati et al.).

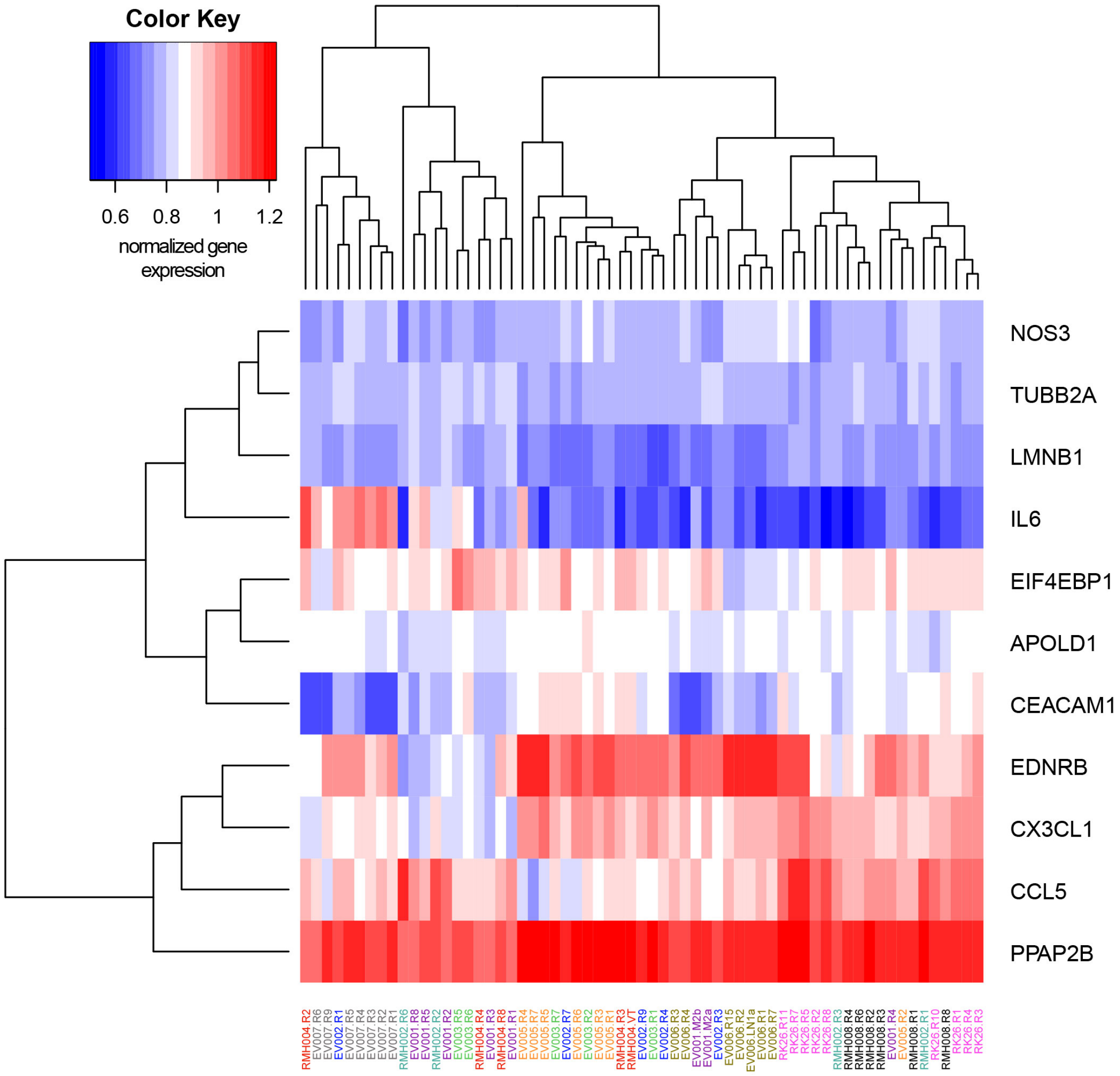


Figure 6.1: Heatmap based on hierarchical clustering of multiple regions from 10 tumours (Gerlinger et al.) based on the 11 cancer related genes from the Rini score (Rini et al.).
Columns represent tumour regions (n=63) and rows correspond to the genes (n=11). Regions derived from the same tumour (coloured identically) do not cluster together, demonstrating intratumour heterogeneity with respect to the expression of the 11 genes. Specifically, based on EDNRB expression Region 1 (R1) from patient EV002 shows down-regulation while the remaining regions show varying degrees of up-regulation; based on IL6 expression Region 2 (R2) from patient RMH004 shows strong up-regulation while all other regions show down-regulation.

Besides our work on the ccA/ccB gene expression signature validating as the only biomarker adding prognostic value over and above the clinical parameters available for the TCGA cohort (Gulati et al.), ClearCode34, as discussed above, which is based on the ccA/ccB signature, has been shown to be a significant predictor of RFS (Brooks et al.). While the Rini score is based on the expression of 16 genes, ClearCode34 is based on a 34-gene signature. Both the Rini score and the ClearCode34 model have been independently validated and add value to recurrence predictions. However, while the Rini score has been developed based on RT-PCR data, the ccA/ccB signature was developed based on microarray data, and validated using RNA-Sequencing data. There are inherent differences in these methodologies both in terms of the read-out and expression normalisation. Therefore, it would be worthwhile to compare both signatures in the same cohort, profiled by the same technique. This should enable an assessment of the most robust signature and its clinical applicability.

The scoring scheme devised by Rini et al. is commendable. Whilst we have shown the existence of heterogeneity for both these signatures, it is yet to be established as to which of them would be more robust against the background of ITH. Caution is therefore recommended when commenting on the relative contribution of tumour heterogeneity to prognostication models.

These data suggest some interesting avenues for research. Despite ITH, the ccB signature out performs every other candidate biomarker in this analysis. It is currently unknown whether a tumour with a small ccB component has a similarly poor prognosis to an identical size tumour, which is dominated by the ccB signature. If the absolute size of the poor risk clone, irrespective of the entire tumour population, is the most critical parameter, then ITH may be less problematic in small tumours as the chance of analytical techniques sampling the high risk cell population would be high. However, detection of a poor risk ccB clone in larger tumours may be more difficult unless the entire tumour is sampled or dominated by the ccB signature. These considerations demonstrate that insights into the impact of ITH on clinical outcomes are limited, raising important questions regarding the clinical interpretation of subclonal abundance and how heterogeneous tumours can be better profiled for biomarker discovery and precision medicine.

All the analyses presented in this thesis underline that the challenges lying ahead of us are linked to sampling technologies and how they can improve both prognostic models as well as understanding how tumour clonal heterogeneity impacts upon clinical outcome. How cancer subclones compete, adapt, and evolve through the disease course in relation to therapy, is an area of unmet clinical and scientific need. Multiple samples from the same tumour are imperative for the determination of the most aggressive molecular signature or subclone within the tumour. A major effort has been launched in this direction known as the Lung TRACERx (TRAcking non-small cell lung Cancer Evolution through therapy [Rx], ClinicalTrials.gov number, NCT01888601), which is a prospective study in primary non-small cell lung cancer (NSCLC). This study, aims to define the genomic landscape of NSCLC and to understand the impact of ITH on therapeutic and survival outcome through multiregion and longitudinal tumour sampling and sequencing (Jamal-Hanjani et al.). While this methodology was initially proposed for the TRACERx trial for Non-small cell lung cancer, our colleagues are now implementing this for ccRCC as well (Soultati et al.).

While efforts such as TRACERx are commendable, longitudinal sampling for solid tumours presents greater problems both in terms of finances and discomfort for the patient. Non-invasive sampling methods such as circulating tumour cells (CTCs) and circulating tumour DNA (ctDNA) are increasingly becoming popular as proxy measures for tumour biopsies. While solid tumour biopsies remain the gold standard for tumour characterisation, the evolving demands of precision medicine require the development of more real time assays which can enable tumour evolution studies (Mateo et al.). CTCs shed from tumour cells into the blood stream are extremely rare (Allard et al.). Metastatic tumours are more likely to have higher counts of CTCs in the blood stream (Tanaka et al.), and primary tumours undergoing treatment have also shown the existence of CTCs, reflective of the probability of recurrence of disease (Hofman et al.).

Studies have been successful in showing the correlation between the number of CTCs in the blood stream and patient prognosis for metastatic disease for prostrate, breast and colorectal cancer (Cristofanilli et al., de Bono et al., Cohen et al.); promising results are also being seen in other cancer studies, for example, lung (Krebs et al., Hiltermann et al.), melanoma (Rao et al.), head and neck (Nichols et al.) and pancreatic cancers (Han et al.). Furthermore, CTCs have been shown to be promising biomarkers for early stage diseases in colorectal (Iinuma et al.) and breast cancers (Rack et al.), establishing that enumeration of CTCs is a powerful prognostic tool.

Moreover, profiling of DNA and RNA from CTCs can enable extensive longitudinal studies evaluating the molecular landscape of the cancer, intratumour heterogeneity and the response of the cancer top therapy. Indeed in multiple cancers, targeted approaches have revealed tumour specific SCNAs (Shaw et al.), mutations driving drug response (Diaz et al.). Whole exome and genome sequencing has revealed the clonal structure of the primary tumour (Chan et al.), variant selection by therapy (Murtaza et al., Dawson et al.) and de-novo genomic rearrangements (Leary et al.).

While questions regarding the extent to which ctDNA is representative of tumour DNA present some obvious caveats, the benefits in terms of sampling and ease of analysis and the obvious associations seen with prognosis warrant further studies to evaluate and develop the clinical applicability of CTCs and ctDNA. Both the above mentioned longitudinal trials are incorporating the serial collection of blood samples in their assessments and are working on evaluating the utility of ctDNA to study disease burden and progression. Such studies will provide unparalleled platforms to study cancer mechanisms and offer insights into further advancing cancer therapy and personalised treatment opportunities.

Appendix

A. GO overrepresentation analysis: Tumour vs. Normal

Gene Set Name	Genes in Gene Set (K)	Genes in Overlap (k)	k/K	p-value	FDR q-value
SIGNAL TRANSDUCTION	1634	116	0.071	6.30E-36	5.20E-33
IMMUNE SYSTEM PROCESS	332	47	0.1416	1.99E-27	8.21E-25
ESTABLISHMENT OF LOCALIZATION	870	68	0.0782	1.48E-23	4.07E-21
ANATOMICAL STRUCTURE DEVELOPMENT	1013	72	0.0711	1.76E-22	3.63E-20
RESPONSE TO EXTERNAL STIMULUS	312	38	0.1218	4.34E-20	7.17E-18
IMMUNE RESPONSE	235	33	0.1404	1.55E-19	2.04E-17
DEFENSE RESPONSE	270	35	0.1296	1.73E-19	2.04E-17
SYSTEM PROCESS	563	48	0.0853	1.82E-18	1.88E-16
MULTICELLULAR ORGANISMAL DEVELOPMENT	1049	64	0.061	8.33E-17	7.64E-15
SYSTEM DEVELOPMENT	861	57	0.0662	1.24E-16	1.02E-14
CELL CELL SIGNALING	404	38	0.0941	2.83E-16	2.12E-14
CELL SURFACE RECEPTOR LINKED SIGNAL TRANSDUCTION	641	48	0.0749	3.12E-16	2.14E-14
ION TRANSPORT	185	26	0.1405	1.08E-15	6.83E-14
APOPTOSIS GO	431	38	0.0882	2.33E-15	1.37E-13
PROGRAMMED CELL DEATH	432	38	0.088	2.51E-15	1.38E-13
CELL DEVELOPMENT	577	44	0.0763	2.87E-15	1.48E-13
REGULATION OF DEVELOPMENTAL PROCESS	440	38	0.0864	4.54E-15	2.20E-13
TRANSPORT	795	51	0.0642	1.79E-14	8.21E-13
ION HOMEOSTASIS	129	21	0.1628	3.06E-14	1.33E-12
CELLULAR HOMEOSTASIS	147	22	0.1497	4.52E-14	1.87E-12
POSITIVE REGULATION OF BIOLOGICAL PROCESS	709	47	0.0663	5.94E-14	2.33E-12
CELL PROLIFERATION GO 0008283	513	39	0.076	1.21E-13	4.56E-12
REGULATION OF APOPTOSIS	341	31	0.0909	3.92E-13	1.40E-11
REGULATION OF PROGRAMMED CELL DEATH	342	31	0.0906	4.23E-13	1.46E-11

Gene Set Name	Genes in Gene Set (K)	Genes in Overlap (k)	k/K	p-value	FDR q-value
REGULATION OF BIOLOGICAL QUALITY	419	34	0.0811	7.40E-13	2.44E-11
BEHAVIOR	153	21	0.1373	9.66E-13	3.07E-11
EXCRETION	36	12	0.3333	1.12E-12	3.43E-11
RESPONSE TO WOUNDING	190	23	0.1211	1.19E-12	3.52E-11
CHEMICAL HOMEOSTASIS	155	21	0.1355	1.25E-12	3.56E-11
CATION HOMEOSTASIS	109	18	0.1651	1.62E-12	4.46E-11
POSITIVE REGULATION OF CELLULAR PROCESS	668	43	0.0644	1.87E-12	4.98E-11
SECRETION	178	22	0.1236	2.41E-12	6.22E-11
REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	151	20	0.1325	6.56E-12	1.64E-10
HOMEOSTATIC PROCESS	207	23	0.1111	7.20E-12	1.75E-10
CELLULAR CATION HOMEOSTASIS	106	17	0.1604	1.10E-11	2.58E-10
INFLAMMATORY RESPONSE	129	18	0.1395	3.04E-11	6.97E-10
ANATOMICAL STRUCTURE MORPHOGENESIS	376	29	0.0771	1.21E-10	2.69E-09
CELL ACTIVATION	77	14	0.1818	1.27E-10	2.75E-09
ORGAN DEVELOPMENT	571	36	0.063	2.13E-10	4.50E-09
LOCOMOTORY BEHAVIOR	95	15	0.1579	2.24E-10	4.63E-09
T CELL ACTIVATION	44	11	0.25	3.28E-10	6.60E-09
METAL ION TRANSPORT	117	16	0.1368	5.14E-10	9.86E-09
NEGATIVE REGULATION OF BIOLOGICAL PROCESS	677	39	0.0576	5.14E-10	9.86E-09
CELLULAR DEFENSE RESPONSE	58	12	0.2069	5.53E-10	1.04E-08
REGULATION OF CELL PROLIFERATION	308	25	0.0812	7.98E-10	1.46E-08
POSITIVE REGULATION OF DEVELOPMENTAL PROCESS	218	21	0.0963	8.35E-10	1.50E-08
LYMPHOCYTE ACTIVATION	61	12	0.1967	1.03E-09	1.80E-08
RESPONSE TO CHEMICAL STIMULUS	314	25	0.0796	1.19E-09	2.04E-08

Gene Set Name	Genes in Gene Set (K)	Genes in Overlap (k)	k/K	p-value	FDR q-value
NEGATIVE REGULATION OF CELLULAR PROCESS	646	37	0.0573	1.67E-09	2.81E-08
CATION TRANSPORT	147	17	0.1156	2.09E-09	3.45E-08
SODIUM ION TRANSPORT	22	8	0.3636	3.16E-09	5.02E-08
ANION TRANSPORT	31	9	0.2903	3.17E-09	5.02E-08
LEUKOCYTE ACTIVATION	69	12	0.1739	4.55E-09	7.08E-08
RESPONSE TO STRESS	508	31	0.061	8.15E-09	1.25E-07
NERVOUS SYSTEM DEVELOPMENT	385	26	0.0675	1.72E-08	2.58E-07
MONOVALENT INORGANIC CATION TRANSPORT	94	13	0.1383	1.87E-08	2.75E-07
G PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY	342	24	0.0702	2.94E-08	4.25E-07
INTRACELLULAR SIGNALING CASCADE	667	35	0.0525	4.08E-08	5.80E-07
NEGATIVE REGULATION OF DEVELOPMENTAL PROCESS	197	17	0.0863	1.67E-07	2.33E-06
REGULATION OF LYMPHOCYTE ACTIVATION	35	8	0.2286	1.88E-07	2.59E-06
NEUROLOGICAL SYSTEM PROCESS	379	24	0.0633	1.99E-07	2.69E-06
NEGATIVE REGULATION OF CELL PROLIFERATION	156	15	0.0962	2.18E-07	2.90E-06
REGULATION OF IMMUNE SYSTEM PROCESS	67	10	0.1493	3.95E-07	5.17E-06
CELL CELL ADHESION	86	11	0.1279	5.20E-07	6.70E-06
SENSORY PERCEPTION	190	15	0.0789	2.68E-06	3.40E-05
POSITIVE REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	66	9	0.1364	3.28E-06	4.10E-05
SECOND MESSENGER MEDIATED SIGNALING	153	13	0.085	5.52E-06	6.79E-05
STEROID METABOLIC PROCESS	71	9	0.1268	6.08E-06	7.37E-05
REGULATION OF BODY FLUID LEVELS	57	8	0.1404	9.27E-06	1.11E-04
REGULATION OF T CELL ACTIVATION	28	6	0.2143	9.99E-06	1.18E-04
MULTI ORGANISM PROCESS	165	13	0.0788	1.25E-05	1.46E-04
REGULATION OF CELL DIFFERENTIATION	60	8	0.1333	1.37E-05	1.57E-04

Gene Set Name	Genes in Gene Set (K)	Genes in Overlap (k)	k/K	p-value	FDR q-value
RESPONSE TO BIOTIC STIMULUS	120	11	0.0917	1.40E-05	1.59E-04
INORGANIC ANION TRANSPORT	18	5	0.2778	1.44E-05	1.59E-04
DIGESTION	44	7	0.1591	1.45E-05	1.59E-04
ORGAN MORPHOGENESIS	144	12	0.0833	1.53E-05	1.66E-04
POSITIVE REGULATION OF CELL PROLIFERATION	149	12	0.0805	2.15E-05	2.31E-04
NEGATIVE REGULATION OF APOPTOSIS	150	12	0.08	2.30E-05	2.43E-04
CELLULAR LIPID METABOLIC PROCESS	255	16	0.0627	2.33E-05	2.43E-04
NEGATIVE REGULATION OF PROGRAMMED CELL DEATH	151	12	0.0795	2.46E-05	2.54E-04
REGULATION OF PHOSPHORYLATION	49	7	0.1429	3.00E-05	3.05E-04
REPRODUCTION	265	16	0.0604	3.71E-05	3.73E-04
REGULATION OF CELL ADHESION	37	6	0.1622	5.36E-05	5.32E-04
REGULATION OF CATALYTIC ACTIVITY	276	16	0.058	6.02E-05	5.91E-04
LEUKOCYTE DIFFERENTIATION	38	6	0.1579	6.26E-05	6.08E-04
HEMOPOIESIS	75	8	0.1067	7.08E-05	6.79E-04
HEMOPOIETIC OR LYMPHOID ORGAN DEVELOPMENT	77	8	0.1039	8.55E-05	8.11E-04
REGULATION OF MOLECULAR FUNCTION	324	17	0.0525	1.22E-04	1.12E-03
SKELETAL DEVELOPMENT	103	9	0.0874	1.22E-04	1.12E-03
IMMUNE SYSTEM DEVELOPMENT	81	8	0.0988	1.23E-04	1.12E-03
LIPID METABOLIC PROCESS	325	17	0.0523	1.26E-04	1.14E-03
NITROGEN COMPOUND METABOLIC PROCESS	155	11	0.071	1.47E-04	1.32E-03
REPRODUCTIVE PROCESS	162	11	0.0679	2.16E-04	1.92E-03
ANGIOGENESIS	48	6	0.125	2.38E-04	2.07E-03
HEMOSTASIS	48	6	0.125	2.38E-04	2.07E-03
NEGATIVE REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	32	5	0.1562	2.74E-04	2.36E-03
NEGATIVE REGULATION OF CELL ADHESION	18	4	0.2222	2.80E-04	2.38E-03
AMINE METABOLIC PROCESS	141	10	0.0709	2.92E-04	2.46E-03
POSITIVE REGULATION OF IMMUNE SYSTEM PROCESS	51	6	0.1176	3.34E-04	2.78E-03
ANTI APOPTOSIS	118	9	0.0763	3.40E-04	2.81E-03

B. Pathway overrepresentation analysis: Tumour vs. Normal

Gene Set Name	# Genes in Gene Set (K	# Genes in Overlap (k	k/K	p-value	FDR q-value
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION	267	41	0.154	1.77E-25	1.90E-22
REACTOME TRANSMEMBRANE TRANSPORT OF SMALL MOLECULES	413	46	0.111	2.05E-22	1.10E-19
REACTOME SLC MEDIATED TRANSMEMBRANE TRANSPORT	241	33	0.137	3.42E-19	1.23E-16
REACTOME GPCR LIGAND BINDING	408	40	0.098	1.10E-17	2.95E-15
REACTOME IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON LYMPHOID CELL	70	19	0.271	2.19E-17	4.73E-15
REACTOME CLASS A1 RHODOPSIN LIKE RECEPTORS	305	32	0.105	2.98E-15	5.35E-13
REACTOME PEPTIDE LIGAND BINDING RECEPTORS	188	25	0.133	1.41E-14	2.01E-12
REACTOME IMMUNE SYSTEM	933	56	0.06	1.50E-14	2.01E-12
REACTOME G ALPHA I SIGNALLING EVENTS	195	25	0.128	3.32E-14	3.98E-12
BIOCARTA NO2IL12 PATHWAY	17	10	0.588	7.09E-14	7.63E-12
REACTOME ADAPTIVE IMMUNE SYSTEM	539	40	0.074	1.28E-13	1.26E-11
BIOCARTA CTL PATHWAY	15	9	0.6	1.02E-12	9.18E-11
KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY	137	19	0.139	9.64E-12	7.98E-10
REACTOME HEMOSTASIS	466	34	0.073	1.41E-11	1.06E-09
REACTOME SIGNALING BY GPCR	920	50	0.054	1.47E-11	1.06E-09
REACTOME GPCR DOWNSTREAM SIGNALING	805	46	0.057	1.86E-11	1.25E-09
KEGG COMPLEMENT AND COAGULATION CASCADES	69	14	0.203	2.67E-11	1.69E-09
REACTOME CHEMOKINE RECEPTORS BIND CHEMOKINES	57	13	0.228	2.83E-11	1.69E-09
BIOCARTA TCYTOTOXIC PATHWAY	14	8	0.571	3.38E-11	1.92E-09
KEGG CELL ADHESION MOLECULES CAMS	134	18	0.134	5.81E-11	3.13E-09
BIOCARTA IL12 PATHWAY	23	9	0.391	1.46E-10	7.51E-09
KEGG CHEMOKINE SIGNALING PATHWAY	190	20	0.105	4.40E-10	2.15E-08
REACTOME TRANSPORT OF GLUCOSE AND OTHER SUGARS BILE SALTS AND ORGANIC ACIDS METAL IONS AND AMINE (	89	14	0.157	9.40E-10	4.40E-08
KEGG T CELL RECEPTOR SIGNALING PATHWAY	108	15	0.139	1.43E-09	6.41E-08

Gene Set Name	# Genes in Gene Set (K	# Genes in Overlap (k	k/K	p-value	FDR q-value
REACTOME TRANSPORT OF INORGANIC CATIONS ANIONS AND AMINO ACIDS OLIGOPEPTIDES	94	14	0.149	1.97E-09	8.50E-08
BIOCARTA THELPER PATHWAY	14	7	0.5	2.09E-09	8.67E-08
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	272	22	0.081	8.73E-09	3.48E-07
REACTOME CELL SURFACE INTERACTIONS AT THE VASCULAR WALL	91	13	0.143	1.25E-08	4.81E-07
REACTOME NEURONAL SYSTEM	279	22	0.079	1.39E-08	5.15E-07
BIOCARTA TCAPOPTOSIS PATHWAY	11	6	0.546	1.60E-08	5.74E-07
BIOCARTA CTLA4 PATHWAY	21	7	0.333	6.34E-08	2.20E-06
REACTOME COSTIMULATION BY THE CD28 FAMILY	63	10	0.159	2.17E-07	7.32E-06
REACTOME REGULATION OF INSULIN LIKE GROWTH FACTOR IGF ACTIVITY BY INSULIN LIKE GROWTH FACTOR BINDING F	16	6	0.375	2.56E-07	8.36E-06
BIOCARTA IL17 PATHWAY	17	6	0.353	3.90E-07	1.23E-05
BIOCARTA NKT PATHWAY	29	7	0.241	7.48E-07	2.30E-05
KEGG ALDOSTERONE REGULATED SODIUM REABSORPTION	42	8	0.191	8.43E-07	2.52E-05
KEGG FOCAL ADHESION	201	16	0.08	1.13E-06	3.28E-05
KEGG TYPE I DIABETES MELLITUS	44	8	0.182	1.23E-06	3.47E-05
REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE	32	7	0.219	1.54E-06	4.25E-05
REACTOME PLATELET ACTIVATION SIGNALING AND AGGREGATION	208	16	0.077	1.77E-06	4.73E-05
REACTOME G ALPHA Q SIGNALLING EVENTS	184	15	0.082	1.80E-06	4.73E-05
REACTOME TRANSMISSION ACROSS CHEMICAL SYNAPSES	186	15	0.081	2.06E-06	5.28E-05
BIOCARTA TCRA PATHWAY	13	5	0.385	2.34E-06	5.86E-05
BIOCARTA INTRINSIC PATHWAY	23	6	0.261	2.89E-06	7.01E-05
KEGG PRIMARY IMMUNODEFICIENCY	35	7	0.2	2.93E-06	7.01E-05
BIOCARTA CLASSIC PATHWAY	14	5	0.357	3.58E-06	8.21E-05
REACTOME TRANSLOCATION OF ZAP 70 TO IMMUNOLOGICAL SYNAPSE	14	5	0.357	3.58E-06	8.21E-05
BIOCARTA CSK PATHWAY	24	6	0.25	3.80E-06	8.52E-05

Gene Set Name	# Genes in Gene Set (K	# Genes in Overlap (k	k/K	p-value	FDR q-value
KEGG STARCH AND SUCROSE METABOLISM	52	8	0.154	4.57E-06	1.01E-04
REACTOME AXON GUIDANCE	251	17	0.068	4.79E-06	1.03E-04
KEGG HEMATOPOIETIC CELL LINEAGE	88	10	0.114	5.09E-06	1.07E-04
REACTOME GASTRIN CREB SIGNALLING PATHWAY VIA PKC AND MAPK	205	15	0.073	6.79E-06	1.41E-04
REACTOME GENERATION OF SECOND MESSENGER MOLECULES	27	6	0.222	7.97E-06	1.62E-04
BIOCARTA LAIR PATHWAY	17	5	0.294	1.06E-05	2.07E-04
REACTOME INTRINSIC PATHWAY	17	5	0.294	1.06E-05	2.07E-04
REACTOME PD1 SIGNALING	18	5	0.278	1.44E-05	2.77E-04
REACTOME INTEGRIN CELL SURFACE INTERACTIONS	79	9	0.114	1.47E-05	2.79E-04
REACTOME G ALPHA S SIGNALLING EVENTS	121	11	0.091	1.52E-05	2.82E-04
REACTOME DEVELOPMENTAL BIOLOGY	396	21	0.053	1.74E-05	3.14E-04
KEGG GLYCOLYSIS GLUCONEOGENESIS	62	8	0.129	1.75E-05	3.14E-04
REACTOME GPVI MEDIATED ACTIVATION CASCADE	31	6	0.194	1.86E-05	3.29E-04
BIOCARTA COMP PATHWAY	19	5	0.263	1.93E-05	3.30E-04
BIOCARTA STATHMIN PATHWAY	19	5	0.263	1.93E-05	3.30E-04
KEGG CALCIUM SIGNALING PATHWAY	178	13	0.073	2.80E-05	4.72E-04
BIOCARTA TCR PATHWAY	49	7	0.143	3.00E-05	4.96E-04
BIOCARTA TOB1 PATHWAY	21	5	0.238	3.27E-05	5.34E-04
REACTOME BILE SALT AND ORGANIC ANION SLC TRANSPORTERS	11	4	0.364	3.35E-05	5.38E-04
REACTOME NEUROTRANSMITTER RECEPTOR BINDING AND DOWNSTREAM TRANSMISSION IN THE POSTSYNAPTIC CEL	137	11	0.08	4.84E-05	7.61E-04
BIOCARTA FIBRINOLYSIS PATHWAY	12	4	0.333	4.95E-05	7.61E-04
REACTOME TANDEM PORE DOMAIN POTASSIUM CHANNELS	12	4	0.333	4.95E-05	7.61E-04
KEGG REGULATION OF ACTIN CYTOSKELETON	216	14	0.065	5.22E-05	7.92E-04
REACTOME BIOLOGICAL OXIDATIONS	139	11	0.079	5.52E-05	8.26E-04

Gene Set Name	# Genes in Gene Set (K	# Genes in Overlap (k	k/K	p-value	FDR q-value
REACTOME TCR SIGNALING	54	7	0.13	5.71E-05	8.42E-04
KEGG ALLOGRAFT REJECTION	38	6	0.158	6.26E-05	9.11E-04
REACTOME POTASSIUM CHANNELS	98	9	0.092	8.28E-05	1.19E-03
KEGG BLADDER CANCER	42	6	0.143	1.12E-04	1.56E-03
KEGG GRAFT VERSUS HOST DISEASE	42	6	0.143	1.12E-04	1.56E-03
KEGG ENDOCYTOSIS	183	12	0.066	1.57E-04	2.16E-03
KEGG ECM RECEPTOR INTERACTION	84	8	0.095	1.58E-04	2.16E-03
REACTOME INITIAL TRIGGERING OF COMPLEMENT	16	4	0.25	1.72E-04	2.28E-03
REACTOME PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS	16	4	0.25	1.72E-04	2.28E-03
REACTOME L1CAM INTERACTIONS	86	8	0.093	1.87E-04	2.45E-03
REACTOME EXTRACELLULAR MATRIX ORGANIZATION	87	8	0.092	2.02E-04	2.63E-03
REACTOME RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	89	8	0.09	2.37E-04	3.04E-03
REACTOME COMPLEMENT CASCADE	32	5	0.156	2.74E-04	3.43E-03
KEGG PPAR SIGNALING PATHWAY	69	7	0.101	2.74E-04	3.43E-03
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	9	0.076	3.40E-04	4.21E-03
REACTOME METABOLISM OF AMINO ACIDS AND DERIVATIVES	200	12	0.06	3.56E-04	4.32E-03
KEGG DRUG METABOLISM CYTOCHROME P450	72	7	0.097	3.57E-04	4.32E-03
REACTOME ION TRANSPORT BY P TYPE ATPASES	34	5	0.147	3.67E-04	4.40E-03
KEGG AUTOIMMUNE THYROID DISEASE	53	6	0.113	4.13E-04	4.88E-03
KEGG PRION DISEASES	35	5	0.143	4.22E-04	4.94E-03
BIOCARTA AMI PATHWAY	20	4	0.2	4.31E-04	4.99E-03
KEGG MAPK SIGNALING PATHWAY	267	14	0.052	4.71E-04	5.40E-03
KEGG STEROID HORMONE BIOSYNTHESIS	55	6	0.109	5.05E-04	5.73E-03
REACTOME DOWNSTREAM TCR SIGNALING	37	5	0.135	5.50E-04	6.17E-03
KEGG CYTOSOLIC DNA SENSING PATHWAY	56	6	0.107	5.57E-04	6.18E-03
KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY	102	8	0.078	5.97E-04	6.56E-03
REACTOME GABA B RECEPTOR ACTIVATION	38	5	0.132	6.24E-04	6.79E-03
BIOCARTA DC PATHWAY	22	4	0.182	6.32E-04	6.81E-03

C. MCODE Clusters

Cluster	Genes
Cluster 1	RPS24,RPL5,RPL22,RPS13,RPS21,RPS3A,RPS3,CUL1,RPL4,NOP56,RPL7,EIF2AK2,VCAM1,RPS11,RPL19,RPL18,RPS5,RPS2,RPLP0,RPL31,RPL30,CAND1,RPS27A,RPL18A,RPL10A,RPL11,RPS26
Cluster 2	RPLP0P6,FN1,RPS19,RPL23A,RPSA,RPS23,ESR1,RPL23,RPS4X,RPS16,RPL14,RPS6,RPL37A,SLC25A5,EEF1A1,ITGA4,RPL6,CSNK1A1,RPS8,ILF3,COPS5,RPL8,UBL4A,RPS7,RPL10L,RPS15A,RPL24,CUL5,RPS14,CDK2,RPL3,RPL17,FBL,RPL15,PAN2,RPS20,RPL21,RPL9,NHP2L1,RPL12,HNRNPM,RPL7A
Cluster 3	PSMB6,GNL3,PSMD14,PSMB2,ADRBK1,MED4,MED29,PSMD4,ANAPC5,MED13,MED10,CDK8,PSMD12,STK25,PSMD8,DHX9,PSMC4,TRAF6,EIF6,RPS6KA1,RPS29,MAP1LC3A,MYBBP1A,PSMA6,MAP3K14,ICAM1,RPS27,HNRNPK,RPS17,RPL32,FAU,PIK3R2,MED12,UCHL5,PSMA7,RPS9,ANAPC7,RPL36,CDC23,RAD23A,PABPC1,SND1,PSMD2,RPS10,FKBP8,ZC3H13,PSMD7,PSMB7,PSMC6,IGSF8,ESR2,PSMD11,PSMD13,PSMA5,PSMC5,MKI67IP,PSMD1,PSMB3,PSMB1,CDK19,PSMC3,PSMA1,MED24,PSMB5,USP14,PSMA2,MED14,PSMD3,PSMC1,MED16,MED30,MED25,MED26,MED28,HNRNPD,NOS2,PRKAA1,PSMC2,SIRT7,HSP90AB1,MED19,PHKG2,MED9,TUFM,MED18
Cluster 4	MRE11A,UBE2D1,DCAF11,IGF2BP3,MED17,TUBA4A,CCT8,RTCB,COPS7B,EEF2,DDB2,ERBB3,RAD50,NFKB2,SYNCRIP,CDC27,HNRNPR,COPS2,RPL27A,EIF4A3,SF3B2,CBL,NPM1,RPL29,DDX1,RPS18,ANAPC11,CD81,RPL35,RPS15,CDC26,ACTB,SOS1,TRIM33,FBXO5,DCAF8,BUB1B,APC2,ADRM1,ANAPC16,MED21,RPLP2,RPL38,CDC16,DCUN1D1,FZR1,KRT2,SRSF5,CAMK1,TRRAP,STAU1,ANAPC4,EEF2K,OXSR1,STK4,CRK,TADA2A,TAF15,PRKACB,PTTG1,FBXW7,RAD21,GPS1,PLCG1,MED15,PRPF19,EPAS1,DDA1,RFWD2,SRSF1,ARRB2,EIF3CL,MDC1,UBE2C,TSR1,HNRNPU,ANAPC2,GNB2L1,HNRNPL,ERBB4,HNRNPH1,EFTUD2,COPS8,COPS7A,SH3KBP1,HNRNPA1,FTSJ3,COPS4,COPS3,TP53BP1,NOP58,CDC20,CCNB1,MED27,RFC1,PSMD5,SNRPD3,U2AF2,ANAPC1,MOV10,MED8,SMURF1,ERCC8,CDC5L,MED1,DDB1,PRKACA,FLT1,SF3A1,n/a,RRS1,EBNA1BP2,UBD,MED11,HNRNPA0,PSMB4,LYAR,EWSR1,KRT10
Cluster 5	CUL4B,PSMA3,RPS28,RPS12,RPL13,PSMA4,TP53RK,RPS25,CSNK1E,CUL2,GABARAPL1,RPL27,VHL,RPLP1,NEDD8,CSNK2A2,GABARAPL2,GABARAP,NOP2,ILK,RPS27L,UBC,RPL10,RIOK2,ILF2,NCL,PSMD6,YWHAZ,CUL3

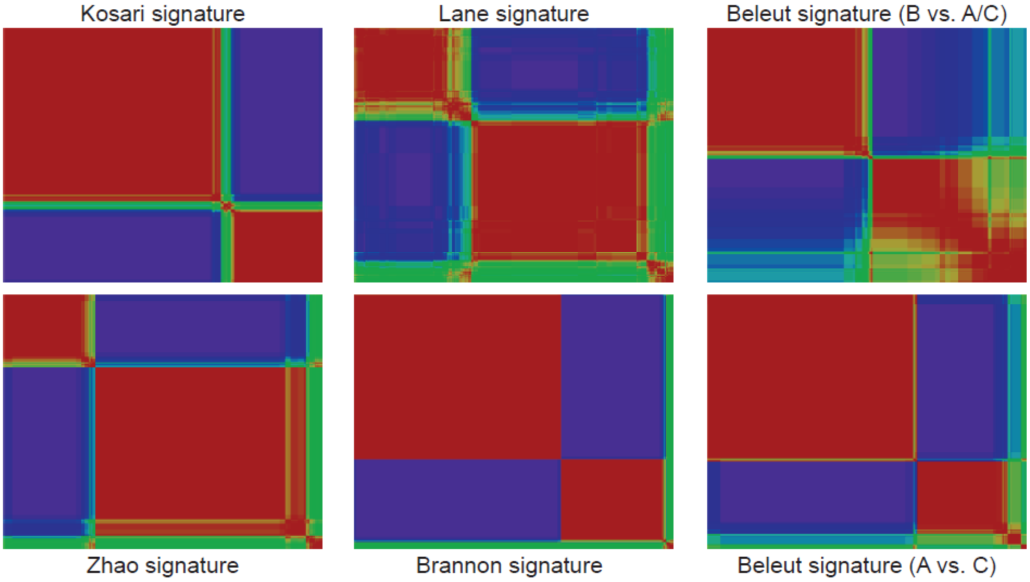
Cluster 6	TAF10,WIBG,HNRNPF,ABCF1,STAT5A,PPP4C,PDGFRB,SNRPD1,TAF13,HNRNPUL1,CCT2,DHX15,CCT4,SRSF2,CCT7,SNRNP70,DDX17,IRS2,SRSF4,TAF2,CD2AP,SF3B3,SRPK1,RELA,STRN3,STRIP1,PHF10,TAF8,TCP1,ERBB2,TAF7,TCEB2,LRR1,TCEB1,IKBKB,SMARCA2,INPP5D,ARID1B,YY1,MED6,ARID1A,MED23,HRAS,IGF1R,INSR,PTK2,ACTL6A,YWHAG,SRRM2,ZAP70,HSPA1L,SFPQ,EZR,HCK,EPOR,TAF3,SNRPA1,HNRNPA2B1,EIF2B3,PTK2B,PML,ARID2,PIK3R1,CHD7,SYK,PBRM1,STAT5B,TOP1,PTBP1,CRKL,MAP3K7,MAP1LC3B,KIT,SREBF1,SNRNP200,EP300,LEF1,NCOR2,SF3B1,CBLB,FGFR1,RUNX1,CTR9,SRSF9,HNRNPH3,TAB2,TAF4,PRPF40A,BCAR1,RPL28,EFS,FBXO25,CTTNBP2NL,NMT1,TRA2A,NELFB,PAF1,HSP90B1,HNRNPH2,PNO1,BCR,SRA1,DLGAP2,RBMX,HNRNPDL,CTDP1,ASAP1,RTF1,CREBBP,RBX1,EPHA2,LEO1,PPP2R2A,MLH1,CDC73,RNPS1,CTTNBP2,MOB4,BLM,PA2G4,KAT2A,IRS1,BUB3,MET,BYSL,PPP2R1B,ABL1,CSF1R,HNRNPA3,MAGEB2,RUNX2,DLGAP1,SNRPC,EIF2B2,RBM4,FEM1B,TAF11,DDX5,SMAD7,HSPD1,RBM39,MAPK14,RALY,ZBTB16,INPPL1,LYN,MSH2
Cluster 7	TRAPPC8,TRAPPC2,TRAPPC3L,TRAPPC10,TRAPPC11,TRAPPC9,TRAPPC3,TRAPPC4,TRAPPC12
Cluster 8	COG4,COG6,COG8,COG1,COG3,COG2

Cluster 9	ERC1,ELL2,TMED9,TXN,GTF2H2,VAV2,CSF3R,CCT3,AP3M1,PPP2R5A,TRIM63,AP3S1,STK11,PTPRE,KIAA1279,GHR,MLLT4,DOCK5,AP3S2,BIRC5,GAB1,STAT2,EZH2,ATXN2,ANK1,ZNF259,COPG1,SPEN,KCNA2,SPC25,KCNA1,USP8,SOCS7,RTN4,ALB,PTPRS,NDUFV1,ARPC5L,HOOK3,NFYA,CCNA2,TNRC6C,GRB7,KRT3,CREM,MDH1,SYN1,MAP2K5,CASP8AP2,HECTD3,RYK,SUMO3,WHSC1,DSG1,COG5,SEC24D,PTPRA,AKTIP,HSPB1,TBPL1,RBL1,GTF3C5,DNAJC7,SHMT1,ZC3H15,TAB1,IL4R,AGL,HIST1H4A,IL2RG,EIF3B,TSEN34,MIS12,PDGFRA,PAWR,MIER1,KHDRBS1,ATF4,AARSD1,NKX2-1,NSL1,CD247,TOR1AIP1,RNF8,GINS3,MORF4L1,ARAP1,NDUFS8,COMMD1,KIAA1598,PIAS2,BCL11B,ITGB3,PTPN18,SERBP1,RASA1,TAF1B,LAT2,CCP110,SCNN1A,XRCC6,ARNTL,KANK2,PAG1,EGF,UFC1,SLC25A3,USP9X,TRMT112,SSSCA1,NEURL4,RAB3GAP1,MAP3K5,EIF3J,ARHGEF7,TAF1L,KRT73,AHR,DCD,HIST1H1C,MYCBP,EIF5,EIF4G2,TFDP2,DAPK3,TARDBP,APPL1,PRKCE,CBX5,GTF2B,TEK,TRIM37,ZNF217,PRPF4,ACTR3B,DCC,TPM3,SEC24A,ATP5B,CEP290,SH3GL1,CEP76,EEF1B2,SUV39H2,CHERP,CD22,TERF2,EHMT2,CSTF3,CXCR4,LIG4,TTC8,CEP97,EGFR,CCNT1,THOC1,PTGES3,E2F1,BBS5,GTF2E1,ICT1,GTF3C2,IRAK2,PDHX,BBIP1,SLC1A5,LRSAM1,USP5,HIST4H4,DOK2,PPP2R5D,SET,CHEK1,TOPBP1,TOPORS,SART3,CALU,MRPS15,PDLIM5,GNE,NCKIPSD,PRDX6,CBX1,PIAS3,SNX6,RPL34,ALDOB,USP22,PRMT5,SCNN1G,AP1M2,NDC80,RQCD1,CNOT3,CNOT2,LAP3,CNOT1,NUF2,GSS,OGFR,BECN1,KDM4A,GTF2H3,EIF2S1,SRCAP,TRAT1,STUB1,MLLT1,CBFA2T2,GTF2H4,PRMT3,GRB2,PAX3,CD2BP2,ARPC2,PNMA1,BBS12,TUT1,GTF2H5,FES,TP53,TIAM1,LCP2,RBM7,ARPC3,PHC2,ERCC2,RAB3GAP2,KDM5B,IQCB1,SRSF11,DNMT1,BRCA1,AFF1,HOOK2,VARS,BAP1,CBFA2T3,BRD8,SAE1,SKAP2,EPM2AIP1,USP25,SLC7A11,NEDD4L,MYH10,PARD6A,KRT9,ITK,SUPT3H,BMPR1B,TNFAIP3,NINL,SPC24,SPP1,TDG,HP1BP3,MCM10,IL1R1,COG7,NR4A1,TPI1,UBQLN2,AURKA,IRAK4,LAGE3,COPB2,KDM5A,KHDRBS2,KCNA3,UBXN7,ADAM15,MYD88,TBCB,PPARD,MECP2,SERTAD1,ARPC1A,CCND2,ARCN1,COPG2,TAF9B,ADRB2,CDK4,ATF7IP,SPTAN1,ARHGAP32,PEBP1,TSC1,KAT5,PELI1,SOCS3,BBS4,MSX1,SLC9A2,ATRX,HDAC3,CTBP1,EXOC7,WIZ,DDX39B,CDC34,MEPE,SKIL,KMT2A,SMARCA5,MERTK,AIFM1,OGDH,USP10,EIF4G3,PPP5C,NCOA2,GRAP2,CLOCK,FLNC,TAF9,FLNB,AP2M1,SP100,ZMIZ1,UBASH3A,DNTTIP2,GTF3C4,SH3GLB2,POLR3C,TNK2,CDYL,RPS6KA5,BBS1,FHOD1,BBS7,NUDCD3,PSMG1,PRKD1,RANBP3,APEX1,ERCC5,THOC3,PABPN1,FARSB,ANKRD28,DFFA,MAPK8IP1,ERC2,MRPS27,TRAF5

Cluster 10	LUC7L3,EIF3E,EIF4A2,HSPBP1,DIS3,RANGAP1,PPME1,NXF1,SRSF3,OGT,LEPRE1,ALAD,TNFRSF1A,OSGEP,AGFG1,BAZ1B,INO80E,EIF3L,G3BP2,FADD,TRADD,SRRM1,EPS15,RBM14,SMARCD2,ARPC4,ALYREF,NCOA3,EXOSC9,ORC3,RIPK1,RIPK2,SIRT2,TNFRSF10B,PPP2R4,HIST2H2BE,AP2B1,HIST3H3,AR,VCP,CDC7,LPP,NAP1L4,SMARCB1,JAK2,PRMT1,SAP30,RRN3,LARS,PINX1,RNF11,TONSL,TPD52L2,TAF1A,MCRS1,TAF1C,DYRK2,TFPT,EPRS,VPRBP,NR3C1,ETS1,ORC6,PPP2CB,EIF1B,SRPK2,SIN3B,PPP2R2D,EXOSC6,MMS22L,NFRKB,ACTR8,PANK4,DBNL,INO80B,TAF1D,H3F3A,U2AF1,PDIA6,INO80C,CCT6A,AARS,RBM25,EFTUD1,EXOSC5,ACTR5,AICDA,PDS5A,WAPAL,SKIV2L2,WFDC5,HDAC9,CDCA5,STAG2,ORC5,EIF3A,PCBP1,EXOSC1,TAF12,NR2C1,KDM1A,EXOSC7

Cluster 11	TNKS2,FAF1,CHM,GSTK1,XPO7,SEPHS1,SLC9A3R2,BFAR,RAB8B,MGA,EDC3,COPB1,NUDCD2,NUB1,DUSP1,KRT15,COL1A1,C12orf10,FKBP4,RIC8A,SCPEP1,ID3,GBF1,GPAA1,DDX4,SSU72,PRPH,PPP2R5B,CCNB2,EIF2B1,CAV1,BAI1,GINS2,SIK2,RORC,HAUS2,BHLHE40,CASP8,HAUS4,SPSB2,PIGK,HAUS8,HMGXB3,RCOR3,LMNB1,KPNA2,S100A7,MGRN1,HAUS3,POLD1,NT5C2,HAUS5,LAMTOR2,HAUS6,ELN,POLD2,IKZF4,ALDH7A1,CLN3,UNC45A,TRPV4,FKBP10,KPNA4,HIRIP3,B4GALT1,AGO3,STIP1,TGFB3,XRCC1,SMARCAD1,VRK2,PCK2,RB1,UBE2Q2,LAMTOR5,TXNRD1,CRYZ,MDM4,PIH1D1,WHSC1L1,MRPL24,DHX8,AATF,PIN1,GRM5,SLC25A6,SFN,MKRN3,GMNN,ENG,FERMT2,POLI,TBL1XR1,UBE2U,PKP2,KANSL1,FUBP1,NF1,YWHAB,MEF2A,ATF3,MSH3,KLC2,DDAH2,MAPT,CCNH,TELO2,TNF,PTPRU,TTI1,EIF3D,CCNG2,JUNB,IMPDH2,MARCH7,ITGA5,NUP210,PCGF3,TRIM74,RUVBL1,ATR,APP,ARNT,UTP14A,MRPL4,NOSIP,ACD,TUBB2A,ARHGAP5,EFNB2,SMC3,DEPTOR,TPBG,MLST8,MALT1,ERCC6,BRD4,MAP3K2,CS,FRK,CLASP1,SPSB1,GSR,EPHA3,INADL,KYNU,STK33,LRP6,AXIN1,RASL12,PPFIA1,PLD2,ATP2A2,TRAF7,HSPA4,MAPRE1,DDX42,ELAC2,TGFBR3,TUBB,PLIN3,ABI2,UBE2G2,UBE2B,MDM2,PRDM16,CDC37,PPFIA3,STK3,MRPL42,PPFIA2,KRT85,KCNJ4,PPFIBP1,EPB41L5,SEPT1,PIGU,PIGT,PTPRD,ACAT1,PFKP,UBOX5,IFT57,BMP7,RNF10,CDK7,RNF2,PPP2R2B,MRPS5,RING1,INHBA,TAGLN,TRAIP,GNAO1,CDK12,KIF3A,PRKD2,LIN7A,CASK,GOPC,ANKMY2,OS9,HOMER3,VDAC2,GET4,ATF2,OTUD7B,TNFRSF10D,ATP6V1E1,AGO4,TRIM23,TRIM65,DBN1,YWHAH,RHOQ,GLRX3,EFEMP2,NLRP2,CTBP2,PIGS,RAB7A,HSPB2,SDHB,SLU7,PFDN1,ARHGEF1,PFN1,TSSK6,CD46,HERC4,REL,WASF1,CD9,BCL2,CDKN1A,DNMBP,CAPNS1,SGK3,CNBP,HAUS1,SUGP1,RBBP6,ATG5,AIP,CARD9,LCK,ATIC,CFH,LDHA,SOCS6,USP21,PICALM,CASP2,RUFY1,CLK2,ATP6V1D,HIF1AN,PIAS4,CTNNBL1,UBE2D4,SUV420H2,SMAD3,UNK,ZFYVE19,KSR1,LTBR,NRD1,SMURF2,PPID,EIF2B5,NFIA,HDAC2,MARK3,PKN1,MKNK1,TNFSF14,C3,SP3,NUP155,CSK,CWC15,MRPS35,SULT1A1,ECE1,COPZ1,GJB1,NEFH,CRELD2,FXR2,THRA,CLU,RBBP8,RNF7,DCAF6,PARD6B,CHD4,MNDA,BCL2L11,PRDX3,LIMA1,PGK1,CTGF,PRKAG2,KCNJ12,SAMHD1,PFKL,FOXH1,DACT1,TRPC1,CFTR,PAFAH1B1,MAPK9,NLRP12,RANBP2,ACTN4,ZMYND11,KDM2A,RNF126,KIF3B,MYOZ1,SNX5,SMEK1,DHPS,ITGA2,DPP8,DDX39A,KATNB1,KIF2A,CLSTN1,CENPF,PHLDA1,HAX1,UBE2H,KPNA1,UGGT1,FRYL,ATP6V1B2,MCM3AP,ZWINT,BRAP,FOXM1,CCND3,WARS,PDCD6,CCAR1,GNL2,RAN,ADSL,CAD,M6PR,TUBA1A,CPT1A,BCOR,PTPRK

D. Ordered Consensus NMF clustering maps for k=2 for gene expression classifying panels included in the prognostic study



Ordered Consensus maps for k=2. Each heatmap depicts the stability of consensus clustering assignment for two clusters.

E. Multivariate analysis results - hazard ratios and p-values for all assessed variables ranked according to order of elimination.

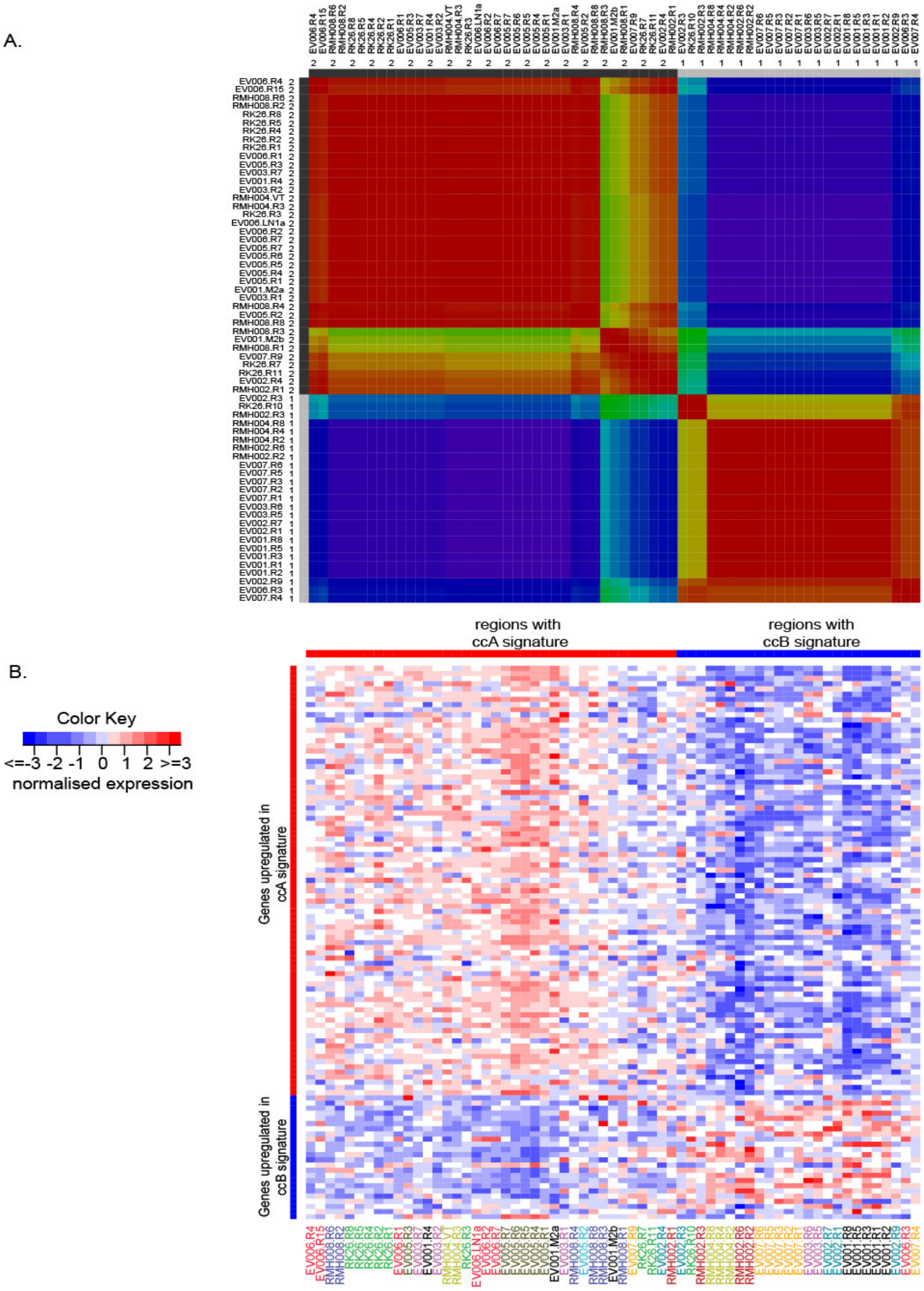
All variables which failed validation are highlighted in red and final significant variables are highlighted in green.

Variable	Hazard Ratio (C.I.)	p-value
EDNRB expression		
< median	1.00 (Ref)	0.972
≥ median	0.98 (0.44 – 2.23)	
Beulet signature		
Cluster A	1.00 (Ref)	
Cluster B	1.52 (0.78 – 2.96)	0.211
Cluster C	0.95 (0.40 – 2.30)	0.915
12 Amplification	1.00 (0.46 – 1.91)	0.882
BAP1 non-syn mutation	1.08 (0.56 – 2.09)	0.819
4p Deletion	1.13 (0.54 – 2.37)	0.737
Lane signature		
Indolent	1.00 (Ref)	0.748
Aggressive	1.13 (0.54 – 2.38)	
22q Deletion	1.24 (0. 58 – 2.67)	0.578
8q Amplification	1.27 (0.59 – 2.68)	0.536
TGFβ signature		
Low expression score	1.00 (Ref)	0.415
High expression score	1.25 (0.72 – 2.18)	
TP53 non-syn mutation	1.67 (0.54 – 5.19)	0.368
Furhmann Grade		
G1/G2	1.00 (Ref)	
G3	1.45 (0.77 – 2.70)	0.243
G4	1.87 (0.87 – 4.02)	0.107
9p Deletion	1.35 (0.82 – 2.23)	0.232
20q focal Amplification	0.69 (0.40 – 1.20)	0.194
Zhao signature		
Cluster 1 (good)	1.00 (Ref)	0.246
Cluster 2 (poor)	1.51 (0.75 – 3.00)	
Kosari signature		
Non - aggressive	1.00 (Ref)	0.137
Aggressive	0.62 (0.32 – 1.16)	
TSPAN7 expression		
< 33 percentile	1.00 (Ref)	0.341
≥ 33 percentile	0.76 (0.43 – 1.34)	
Tumour stage		
Stage I	1.00 (Ref)	
Stage II	3.48 (1.20 – 10.06)	0.022
Stage III	4.61 (1.93 – 11.00)	<0.001
Stage IV	18.01 (7.89 – 41.12)	<0.001
Chrom 19 deletion	4.18 (1.27 – 13.69)	0.018
ccA subgroup	1.00 (Ref)	
ccB subgroup	2.99 (1.87 – 4.80)	<0.001

F. Multivariate analysis with ClearCode34 signature

Variable	Hazard Ratio (C.I.)	p-value
Tumour stage		
Stage I	1.00 (Ref)	
Stage II	3.92 (1.36 – 11.32)	0.012
Stage III	4.86 (2.51 – 13.90)	<0.001
Stage IV	19.32 (8.44 – 44.21)	<0.001
ClearCode34		
ccA subgroup	1.00 (Ref)	<0.001
ccB subgroup	2.23 (1.39 – 3.60)	

G. Consensus NMF clustering analysis for multiregion biopsy dataset.



A. Consensus NMF clustering matrix for multi-region biopsy dataset for two clusters (obtained from http://genepattern.broadinstitute.org/), B. Heatmap shows consensus NMF clustering analysis for the multi-region biopsy dataset using gene expression data of 107 ccA/ccB signature genes. Tumour regions assigned to the ccA or ccB prognostic subgroups is indicated by coloured bars at the top of the heatmap.

H. MSigDB Overrepresentation analysis for ccA vs. Normal Kidney

	Pathways				
Gene Set Name	# Genes in Gene Set (K)	# Genes in Overlap (k)	k/K	p-value	FDR q-value
REACTOME TRANSMEMBRANE TRANSPORT OF SMALL MOLECULES	413	11	0.0266	8.75E-06	9.42E-03
REACTOME SIGNALING BY PDGF	122	6	0.0492	3.67E-05	1.98E-02
REACTOME HEMOSTASIS	466	10	0.0215	1.35E-04	4.27E-02
REACTOME DEVELOPMENTAL BIOLOGY	396	9	0.0227	1.91E-04	4.27E-02
REACTOME INTERACTION BETWEEN L1 AND ANKYRINS	23	3	0.1304	2.03E-04	4.27E-02
REACTOME GPCR LIGAND BINDING	408	9	0.0221	2.38E-04	4.27E-02
REACTOME AXON GUIDANCE	251	7	0.0279	2.93E-04	4.46E-02
REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS	27	3	0.1111	3.31E-04	4.46E-02
REACTOME PEPTIDE LIGAND BINDING RECEPTORS	188	6	0.0319	3.91E-04	4.57E-02
REACTOME CA DEPENDENT EVENTS	30	3	0.1	4.55E-04	4.57E-02
KEGG GLYCINE SERINE AND THREONINE METABOLISM	31	3	0.0968	5.01E-04	4.57E-02
REACTOME NEURONAL SYSTEM	279	7	0.0251	5.50E-04	4.57E-02
REACTOME DAG AND IP3 SIGNALING	32	3	0.0938	5.51E-04	4.57E-02

GO Biological Process

Gene Set Name	# Genes in Gene Set (K)	# Genes in Overlap (k)	k/K	p-value	FDR q-value
ESTABLISHMENT OF LOCALIZATION	870	24	0.0276	1.86E-11	1.54E-08
SIGNAL TRANSDUCTION	1635	28	0.0171	1.70E-08	6.66E-06
SYSTEM PROCESS	563	16	0.0284	3.23E-08	6.66E-06
CELL SURFACE RECEPTOR LINKED SIGNAL TRANSDUCTION GO 0007166	641	17	0.0265	3.23E-08	6.66E-06
TRANSPORT	795	18	0.0226	1.34E-07	2.21E-05
G PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY	342	12	0.0351	1.91E-07	2.63E-05
SECRETION	178	9	0.0506	3.26E-07	3.84E-05
ANATOMICAL STRUCTURE DEVELOPMENT	1014	19	0.0187	1.04E-06	1.07E-04
SYSTEM DEVELOPMENT	861	17	0.0197	1.95E-06	1.79E-04
NEUROLOGICAL SYSTEM PROCESS	379	11	0.029	3.89E-06	3.21E-04
ION TRANSPORT	185	8	0.0432	4.76E-06	3.57E-04
MULTICELLULAR ORGANISMAL DEVELOPMENT	1049	18	0.0172	6.75E-06	4.64E-04
REGULATION OF BODY FLUID LEVELS	57	5	0.0877	1.02E-05	6.46E-04
SYNAPTIC TRANSMISSION	174	7	0.0402	3.00E-05	1.77E-03
TRANSMISSION OF NERVE IMPULSE	189	7	0.037	5.07E-05	2.79E-03
RESPONSE TO EXTERNAL STIMULUS	312	8	0.0256	1.93E-04	9.53E-03
IMMUNE RESPONSE	235	7	0.0298	1.96E-04	9.53E-03
REGULATION OF CELL DIFFERENTIATION	60	4	0.0667	2.39E-04	1.10E-02
IMMUNE SYSTEM PROCESS	332	8	0.0241	2.93E-04	1.27E-02
CELLULAR LOCALIZATION	371	8	0.0216	6.09E-04	2.51E-02
TISSUE DEVELOPMENT	138	5	0.0362	6.79E-04	2.67E-02
NERVOUS SYSTEM DEVELOPMENT	385	8	0.0208	7.74E-04	2.80E-02
EXCRETION	36	3	0.0833	7.82E-04	2.80E-02
BODY FLUID SECRETION	10	2	0.2	1.09E-03	3.76E-02
MONOVALENT INORGANIC CATION TRANSPORT	94	4	0.0426	1.31E-03	3.97E-02
BLOOD COAGULATION	43	3	0.0698	1.32E-03	3.97E-02
DEVELOPMENTAL GROWTH	11	2	0.1818	1.33E-03	3.97E-02
REGULATION OF BIOLOGICAL QUALITY	420	8	0.019	1.35E-03	3.97E-02
COAGULATION	44	3	0.0682	1.41E-03	4.01E-02
G PROTEIN SIGNALING COUPLED TO IP3 SECOND MESSENGERPHOSPHOLIPASE	45	3	0.0667	1.50E-03	4.13E-02
HEMOSTASIS	48	3	0.0625	1.81E-03	4.45E-02
PHOSPHOINOSITIDE MEDIATED SIGNALING	48	3	0.0625	1.81E-03	4.45E-02
REGULATION OF DEVELOPMENTAL PROCESS	441	8	0.0181	1.83E-03	4.45E-02
SKELETAL DEVELOPMENT	103	4	0.0388	1.83E-03	4.45E-02

I. MSigDB Overrepresentation analysis for ccB vs Normal Kidney

Pathways					
Gene Set Name	# Genes in Gene Set	# Genes in Overlap	k/K	p-value	FDR q-valu
REACTOME CELL CYCLE MITOTIC	325	30	0.092	2.82E-18	3.04E-15
REACTOME CELL CYCLE	421	31	0.074	4.61E-16	2.48E-13
REACTOME DNA REPLICATION	192	22	0.115	1.02E-15	3.39E-13
REACTOME MITOTIC M M G1 PHASES	172	21	0.122	1.26E-15	3.39E-13
REACTOME IMMUNE SYSTEM	933	44	0.047	4.63E-15	9.98E-13
REACTOME MITOTIC PROMETAPHASE	87	14	0.161	1.52E-12	2.72E-10
REACTOME PLATELET ACTIVATION SIGNALING AND AGGREGATION	208	18	0.087	4.91E-11	7.56E-09
REACTOME HEMOSTASIS	466	25	0.054	3.03E-10	4.08E-08
REACTOME EXTRACELLULAR MATRIX ORGANIZATION	87	12	0.138	3.88E-10	4.64E-08
REACTOME ADAPTIVE IMMUNE SYSTEM	539	26	0.048	1.28E-09	1.37E-07
REACTOME INTEGRIN CELL SURFACE INTERACTIONS	79	11	0.139	1.89E-09	1.85E-07
KEGG CELL CYCLE	128	13	0.102	3.42E-09	3.06E-07
KEGG ECM RECEPTOR INTERACTION	84	11	0.131	3.69E-09	3.06E-07
REACTOME GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	15	6	0.4	1.04E-08	7.44E-07
REACTOME P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	15	6	0.4	1.04E-08	7.44E-07
REACTOME COLLAGEN FORMATION	58	9	0.155	2.14E-08	1.44E-06
REACTOME GPVI MEDIATED ACTIVATION CASCADE	31	7	0.226	5.33E-08	3.38E-06
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION	267	16	0.06	1.08E-07	6.49E-06
BIOCARTA FIBRINOLYSIS PATHWAY	12	5	0.417	1.47E-07	8.33E-06
REACTOME PLATELET AGGREGATION PLUG FORMATION	36	7	0.194	1.61E-07	8.67E-06
REACTOME COMMON PATHWAY	14	5	0.357	3.64E-07	1.87E-05
REACTOME CYCLIN A B1 ASSOCIATED EVENTS DURING G2 M TRANSITION	15	5	0.333	5.41E-07	2.55E-05
REACTOME INTEGRIN ALPHAIIB BETA3 SIGNALING	27	6	0.222	5.45E-07	2.55E-05
KEGG FOCAL ADHESION	201	13	0.065	7.13E-07	3.20E-05
REACTOME RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	89	9	0.101	9.32E-07	4.01E-05
REACTOME IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON LYMPHOID CELL	70	8	0.114	1.48E-06	6.12E-05
KEGG FC GAMMA R MEDIATED PHAGOCYTOSIS	97	9	0.093	1.93E-06	7.71E-05
REACTOME G1 S SPECIFIC TRANSCRIPTION	19	5	0.263	2.02E-06	7.76E-05
KEGG B CELL RECEPTOR SIGNALING PATHWAY	75	8	0.107	2.51E-06	9.32E-05
BIOCARTA AMI PATHWAY	20	5	0.25	2.67E-06	9.57E-05
REACTOME TCR SIGNALING	54	7	0.13	2.86E-06	9.92E-05
KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY	137	10	0.073	4.67E-06	1.52E-04
REACTOME MITOTIC G1 G1 S PHASES	137	10	0.073	4.67E-06	1.52E-04
REACTOME NCAM1 INTERACTIONS	39	6	0.154	5.35E-06	1.65E-04
REACTOME INNATE IMMUNE SYSTEM	279	14	0.05	5.37E-06	1.65E-04
BIOCARTA INTRINSIC PATHWAY	23	5	0.217	5.62E-06	1.68E-04
KEGG GRAFT VERSUS HOST DISEASE	42	6	0.143	8.35E-06	2.43E-04
REACTOME CELL SURFACE INTERACTIONS AT THE VASCULAR WALL	91	8	0.088	1.08E-05	3.04E-04
KEGG TYPE I DIABETES MELLITUS	44	6	0.136	1.10E-05	3.04E-04
BIOCARTA EXTRINSIC PATHWAY	13	4	0.308	1.14E-05	3.08E-04
REACTOME GENERATION OF SECOND MESSENGER MOLECULES	27	5	0.185	1.30E-05	3.41E-04
KEGG CHEMOKINE SIGNALING PATHWAY	190	11	0.058	1.45E-05	3.72E-04
KEGG COMPLEMENT AND COAGULATION CASCADES	69	7	0.101	1.50E-05	3.76E-04
REACTOME ASSOCIATION OF LICENSING FACTORS WITH THE PRE REPLICATIVE COMPLEX	14	4	0.286	1.59E-05	3.88E-04
KEGG INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	48	6	0.125	1.84E-05	4.41E-04
KEGG CELL ADHESION MOLECULES CAMS	134	9	0.067	2.71E-05	6.35E-04
REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE	32	5	0.156	3.09E-05	7.08E-04
KEGG FC EPSILON RI SIGNALING PATHWAY	79	7	0.089	3.65E-05	8.19E-04
REACTOME AXON GUIDANCE	251	12	0.048	4.00E-05	8.79E-04
REACTOME MITOTIC G2 G2 M PHASES	81	7	0.086	4.29E-05	9.24E-04
REACTOME E2F MEDIATED REGULATION OF DNA REPLICATION	35	5	0.143	4.84E-05	1.01E-03

Pathways					
Gene Set Name	# Genes in Gene Set	# Genes in Overlap (	k/K	p-value	FDR q-valu
REACTOME G1 S TRANSITION	112	8	0.071	4.88E-05	1.01E-03
REACTOME G ALPHA Q SIGNALLING EVENTS	184	10	0.054	6.02E-05	1.22E-03
REACTOME TOLL RECEPTOR CASCADES	118	8	0.068	7.07E-05	1.41E-03
BIOCARTA NKCELLS PATHWAY	20	4	0.2	7.27E-05	1.41E-03
KEGG HEMATOPOIETIC CELL LINEAGE	88	7	0.08	7.31E-05	1.41E-03
REACTOME NCAM SIGNALING FOR NEURITE OUT GROWTH	64	6	0.094	9.63E-05	1.82E-03
REACTOME GASTRIN CREB SIGNALLING PATHWAY VIA PKC AND MAPK	205	10	0.049	1.47E-04	2.73E-03
REACTOME G2 M CHECKPOINTS	45	5	0.111	1.66E-04	3.03E-03
REACTOME G0 AND EARLY G1	25	4	0.16	1.81E-04	3.25E-03
REACTOME NUCLEOTIDE BINDING DOMAIN LEUCINE RICH REPEAT CONTAINING RECEPTOR NLR SIGNALING PATHWAYS	46	5	0.109	1.84E-04	3.25E-03
KEGG VIRAL MYOCARDITIS	73	6	0.082	2.00E-04	3.48E-03
KEGG REGULATION OF ACTIN CYTOSKELETON	216	10	0.046	2.24E-04	3.76E-03
KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	140	8	0.057	2.32E-04	3.76E-03
BIOCARTA BLYMPHOCYTE PATHWAY	11	3	0.273	2.34E-04	3.76E-03
REACTOME CD28 DEPENDENT VAV1 PATHWAY	11	3	0.273	2.34E-04	3.76E-03
REACTOME CDC6 ASSOCIATION WITH THE ORC ORIGIN COMPLEX	11	3	0.273	2.34E-04	3.76E-03
KEGG CALCIUM SIGNALING PATHWAY	178	9	0.051	2.40E-04	3.80E-03
REACTOME STRIATED MUSCLE CONTRACTION	27	4	0.148	2.47E-04	3.84E-03
REACTOME CLASS A1 RHODOPSIN LIKE RECEPTORS	305	12	0.039	2.49E-04	3.84E-03
REACTOME PLATELET ADHESION TO EXPOSED COLLAGEN	12	3	0.25	3.09E-04	4.68E-03
REACTOME M G1 TRANSITION	81	6	0.074	3.54E-04	5.29E-03
REACTOME PEPTIDE LIGAND BINDING RECEPTORS	188	9	0.048	3.59E-04	5.29E-03
KEGG OOCYTE MEIOSIS	114	7	0.061	3.67E-04	5.34E-03
KEGG ARGININE AND PROLINE METABOLISM	54	5	0.093	3.94E-04	5.66E-03
REACTOME ACTIVATION OF THE PRE REPLICATIVE COMPLEX	31	4	0.129	4.27E-04	6.05E-03
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	7	0.059	4.52E-04	6.32E-03
REACTOME REGULATION OF MITOTIC CELL CYCLE	85	6	0.071	4.59E-04	6.33E-03
KEGG PROGESTERONE MEDIATED OOCYTE MATURATION	86	6	0.07	4.88E-04	6.66E-03
BIOCARTA PLATELETAPP PATHWAY	14	3	0.214	5.02E-04	6.76E-03
REACTOME SIGNALING BY PDGF	122	7	0.057	5.52E-04	7.34E-03
REACTOME CELL CYCLE CHECKPOINTS	124	7	0.057	6.08E-04	7.99E-03
KEGG PRIMARY IMMUNODEFICIENCY	35	4	0.114	6.85E-04	8.89E-03
REACTOME REGULATION OF INSULIN LIKE GROWTH FACTOR IGF ACTIVITY BY INSULIN LIKE GROWTH FACTOR BINDING P	16	3	0.188	7.60E-04	9.74E-03
REACTOME DEVELOPMENTAL BIOLOGY	396	13	0.033	7.78E-04	9.86E-03
REACTOME INFLAMMASOMES	17	3	0.177	9.14E-04	1.14E-02
BIOCARTA IL2RB PATHWAY	38	4	0.105	9.40E-04	1.14E-02
KEGG ALLOGRAFT REJECTION	38	4	0.105	9.40E-04	1.14E-02
REACTOME ACTIVATION OF ATR IN RESPONSE TO REPLICATION STRESS	38	4	0.105	9.40E-04	1.14E-02
KEGG PPAR SIGNALING PATHWAY	69	5	0.073	1.22E-03	1.46E-02
REACTOME PHASE1 FUNCTIONALIZATION OF COMPOUNDS	70	5	0.071	1.30E-03	1.54E-02
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	272	10	0.037	1.33E-03	1.56E-02
KEGG DRUG METABOLISM CYTOCHROME P450	72	5	0.069	1.47E-03	1.67E-02
KEGG LEISHMANIA INFECTION	72	5	0.069	1.47E-03	1.67E-02
REACTOME APC C CDH1 MEDIATED DEGRADATION OF CDC20 AND OTHER APC C CDH1 TARGETED PROTEINS IN LATE M	72	5	0.069	1.47E-03	1.67E-02
REACTOME IL 3 5 AND GM CSF SIGNALING	43	4	0.093	1.50E-03	1.68E-02
BIOCARTA CTLA4 PATHWAY	21	3	0.143	1.73E-03	1.92E-02
KEGG JAK STAT SIGNALING PATHWAY	155	7	0.045	2.22E-03	2.41E-02
KEGG PROXIMAL TUBULE BICARBONATE RECLAMATION	23	3	0.13	2.26E-03	2.41E-02
REACTOME TAK1 ACTIVATES NFKB BY PHOSPHORYLATION AND ACTIVATION OF IKKS COMPLEX	23	3	0.13	2.26E-03	2.41E-02

GO Biological Processes					
Gene Set Name	# Genes in Gene Set	# Genes in Overlap (	k/K	p-value	FDR q-valu
IMMUNE SYSTEM PROCESS	332	36	0.108	4.57E-24	3.77E-21
IMMUNE RESPONSE	235	27	0.115	5.13E-19	2.11E-16
SIGNAL TRANSDUCTION	1635	64	0.039	1.91E-17	5.25E-15
REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	151	21	0.139	8.48E-17	1.75E-14
MITOTIC CELL CYCLE	153	21	0.137	1.12E-16	1.84E-14
CELL CYCLE PHASE	170	21	0.124	9.90E-16	1.36E-13
CELL CYCLE GO 0007049	315	26	0.083	7.57E-15	8.92E-13
CELL CYCLE PROCESS	193	21	0.109	1.31E-14	1.35E-12
M PHASE OF MITOTIC CELL CYCLE	85	15	0.177	5.82E-14	5.34E-12
CELL PROLIFERATION GO 0008283	513	31	0.06	9.79E-14	8.08E-12
M PHASE	114	16	0.14	3.47E-13	2.60E-11
MITOSIS	82	14	0.171	6.50E-13	4.47E-11
POSITIVE REGULATION OF BIOLOGICAL PROCESS	709	35	0.049	8.77E-13	5.57E-11
RESPONSE TO EXTERNAL STIMULUS	312	23	0.074	2.87E-12	1.69E-10
DEFENSE RESPONSE	270	21	0.078	9.27E-12	5.10E-10
REGULATION OF IMMUNE SYSTEM PROCESS	67	12	0.179	1.62E-11	8.33E-10
POSITIVE REGULATION OF CELLULAR PROCESS	668	31	0.046	8.43E-11	4.09E-09
REGULATION OF MITOSIS	41	9	0.22	8.38E-10	3.84E-08
ORGAN DEVELOPMENT	571	27	0.047	9.43E-10	4.09E-08
CELL ACTIVATION	77	11	0.143	1.43E-09	5.89E-08
LYMPHOCYTE ACTIVATION	61	10	0.164	2.00E-09	7.86E-08
NEGATIVE REGULATION OF BIOLOGICAL PROCESS	677	29	0.043	2.13E-09	7.99E-08
CELL SURFACE RECEPTOR LINKED SIGNAL TRANSDUCTION GO 0007166	641	28	0.044	2.65E-09	9.50E-08
REGULATION OF CELL CYCLE	182	15	0.082	3.93E-09	1.35E-07
LEUKOCYTE ACTIVATION	69	10	0.145	6.95E-09	2.29E-07
BIOPOLYMER METABOLIC PROCESS	1684	47	0.028	2.96E-08	9.40E-07
RESPONSE TO STRESS	508	23	0.045	3.67E-08	1.12E-06
APOPTOSIS GO	431	21	0.049	4.20E-08	1.24E-06
PROGRAMMED CELL DEATH	432	21	0.049	4.37E-08	1.24E-06
RESPONSE TO WOUNDING	190	14	0.074	5.35E-08	1.47E-06
NEGATIVE REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	32	7	0.219	6.75E-08	1.77E-06
POSITIVE REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	66	9	0.136	6.86E-08	1.77E-06
TISSUE DEVELOPMENT	138	12	0.087	7.84E-08	1.96E-06
MULTICELLULAR ORGANISMAL DEVELOPMENT	1049	34	0.032	8.90E-08	2.13E-06
CELL DEVELOPMENT	577	24	0.042	9.02E-08	2.13E-06
REGULATION OF CELL PROLIFERATION	308	17	0.055	1.42E-07	3.25E-06
NEGATIVE REGULATION OF CELLULAR PROCESS	646	25	0.039	1.90E-07	4.24E-06
REGULATION OF DEVELOPMENTAL PROCESS	441	20	0.045	2.77E-07	6.01E-06
SYSTEM DEVELOPMENT	861	29	0.034	3.67E-07	7.77E-06
ANATOMICAL STRUCTURE DEVELOPMENT	1014	32	0.032	3.84E-07	7.92E-06
REGULATION OF APOPTOSIS	341	17	0.05	5.92E-07	1.19E-05
REGULATION OF PROGRAMMED CELL DEATH	342	17	0.05	6.16E-07	1.21E-05
CALCIUM MEDIATED SIGNALING	16	5	0.313	7.80E-07	1.50E-05
PROTEIN METABOLIC PROCESS	1231	35	0.028	1.21E-06	2.27E-05
CELL CYCLE CHECKPOINT GO 0000075	48	7	0.146	1.26E-06	2.31E-05
REGULATION OF MOLECULAR FUNCTION	325	16	0.049	1.49E-06	2.67E-05
REGULATION OF IMMUNE RESPONSE	33	6	0.182	1.92E-06	3.38E-05
CYTOKINE PRODUCTION	73	8	0.11	2.04E-06	3.51E-05
REGULATION OF CELLULAR METABOLIC PROCESS	787	26	0.033	2.10E-06	3.53E-05

GO Biological Processes					
Gene Set Name	# Genes in Gene Set	# Genes in Overlap (	k/K	p-value	FDR q-valu
REGULATION OF METABOLIC PROCESS	799	26	0.033	2.75E-06	4.54E-05
CELLULAR PROTEIN METABOLIC PROCESS	1117	32	0.029	2.98E-06	4.82E-05
CELLULAR MACROMOLECULE METABOLIC PROCESS	1131	32	0.028	3.84E-06	6.10E-05
REGULATION OF RESPONSE TO STIMULUS	59	7	0.119	5.23E-06	8.15E-05
INTERPHASE OF MITOTIC CELL CYCLE	62	7	0.113	7.32E-06	1.12E-04
NUCLEOBASENUCLEOSIDENUCLEOTIDE AND NUCLEIC ACID METABOLIC PROCESS	1244	33	0.027	1.03E-05	1.54E-04
T CELL ACTIVATION	44	6	0.136	1.10E-05	1.62E-04
ESTABLISHMENT OF LOCALIZATION	870	26	0.03	1.23E-05	1.78E-04
INTERPHASE	68	7	0.103	1.36E-05	1.94E-04
INTERFERON GAMMA PRODUCTION	14	4	0.286	1.59E-05	2.18E-04
NEGATIVE REGULATION OF IMMUNE SYSTEM PROCESS	14	4	0.286	1.59E-05	2.18E-04
EPIDERMIS DEVELOPMENT	71	7	0.099	1.81E-05	2.45E-04
POSITIVE REGULATION OF IMMUNE RESPONSE	29	5	0.172	1.87E-05	2.49E-04
INFLAMMATORY RESPONSE	129	9	0.07	2.00E-05	2.62E-04
REGULATION OF CELLULAR PROTEIN METABOLIC PROCESS	162	10	0.062	2.04E-05	2.63E-04
POSITIVE REGULATION OF IMMUNE SYSTEM PROCESS	51	6	0.118	2.63E-05	3.33E-04
CHROMOSOME SEGREGATION	32	5	0.156	3.09E-05	3.80E-04
HUMORAL IMMUNE RESPONSE	32	5	0.156	3.09E-05	3.80E-04
REGULATION OF PROTEIN METABOLIC PROCESS	173	10	0.058	3.58E-05	4.34E-04
ACTIVATION OF IMMUNE RESPONSE	17	4	0.235	3.67E-05	4.39E-04
ECTODERM DEVELOPMENT	80	7	0.088	3.96E-05	4.67E-04
INTRACELLULAR SIGNALING CASCADE	668	21	0.031	4.07E-05	4.73E-04
REGULATION OF LYMPHOCYTE ACTIVATION	35	5	0.143	4.84E-05	5.55E-04
CELLULAR DEFENSE RESPONSE	58	6	0.103	5.51E-05	6.23E-04
B CELL ACTIVATION	20	4	0.2	7.27E-05	8.00E-04
REGULATION OF CYTOKINE BIOSYNTHETIC PROCESS	38	5	0.132	7.27E-05	8.00E-04
BIOPOLYMER MODIFICATION	650	20	0.031	8.24E-05	8.95E-04
CELL CELL SIGNALING	404	15	0.037	8.40E-05	9.00E-04
REGULATION OF SECRETION	40	5	0.125	9.36E-05	9.90E-04
REGULATION OF CATALYTIC ACTIVITY	277	12	0.043	1.02E-04	1.07E-03
CYTOKINE BIOSYNTHETIC PROCESS	41	5	0.122	1.06E-04	1.08E-03
POSITIVE REGULATION OF RESPONSE TO STIMULUS	41	5	0.122	1.06E-04	1.08E-03
MESODERM DEVELOPMENT	22	4	0.182	1.08E-04	1.08E-03
LIPID METABOLIC PROCESS	325	13	0.04	1.19E-04	1.17E-03
CYTOKINE METABOLIC PROCESS	42	5	0.119	1.19E-04	1.17E-03
REGULATION OF MITOTIC CELL CYCLE	23	4	0.174	1.29E-04	1.25E-03
ADAPTIVE IMMUNE RESPONSE GO 0002460	24	4	0.167	1.54E-04	1.45E-03
DNA INTEGRITY CHECKPOINT	24	4	0.167	1.54E-04	1.45E-03
POST TRANSLATIONAL PROTEIN MODIFICATION	476	16	0.034	1.55E-04	1.45E-03
PROTEIN MODIFICATION PROCESS	631	19	0.03	1.62E-04	1.50E-03
ADAPTIVE IMMUNE RESPONSE	25	4	0.16	1.81E-04	1.66E-03
REGULATION OF INTERFERON GAMMA BIOSYNTHETIC PROCESS	11	3	0.273	2.34E-04	2.10E-03
NEGATIVE REGULATION OF CELLULAR METABOLIC PROCESS	259	11	0.043	2.34E-04	2.10E-03
POSITIVE REGULATION OF DEVELOPMENTAL PROCESS	218	10	0.046	2.42E-04	2.14E-03
NEGATIVE REGULATION OF METABOLIC PROCESS	262	11	0.042	2.58E-04	2.26E-03
INTERFERON GAMMA BIOSYNTHETIC PROCESS	12	3	0.25	3.09E-04	2.68E-03
PHOSPHORYLATION	313	12	0.038	3.15E-04	2.71E-03
SYSTEM PROCESS	563	17	0.03	3.42E-04	2.91E-03
POSITIVE REGULATION OF CELL PROLIFERATION	149	8	0.054	3.53E-04	2.96E-03
REGULATION OF NUCLEOBASENUCLEOSIDENUCLEOTIDE AND NUCLEIC ACID METABOLIC PROCESS	618	18	0.029	3.55E-04	2.96E-03
G1 PHASE OF MITOTIC CELL CYCLE	13	3	0.231	3.98E-04	3.22E-03

Reference List

AKAVIA, U. D., LITVIN, O., KIM, J., SANCHEZ-GARCIA, F., KOTLIAR, D., CAUSTON, H. C., POCHANARD, P., MOZES, E., GARRAWAY, L. A. & PE'ER, D. 2010. An integrated approach to uncover drivers of cancer. *Cell,* 143**,** 1005-17.

ALLARD, W. J., MATERA, J., MILLER, M. C., REPOLLET, M., CONNELLY, M. C., RAO, C., TIBBE, A. G., UHR, J. W. & TERSTAPPEN, L. W. 2004. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin Cancer Res,* 10**,** 6897-904.

AN, J. & RETTIG, M. B. 2005. Mechanism of von Hippel-Lindau protein-mediated suppression of nuclear factor kappa B activity. *Mol Cell Biol,* 25**,** 7546-56.

ANSARI, D., UREY, C., GUNDEWAR, C., BAUDEN, M. P. & ANDERSSON, R. 2013. Comparison of MUC4 expression in primary pancreatic cancer and paired lymph node metastases. *Scand J Gastroenterol,* 48**,** 1183-7.

ANTONELLI, A., ARRIGHI, N., TARDANICO, R., BALZARINI, P., ZANOTELLI, T., CORTI, S., ZANI, D., COZZOLI, A., CUNICO, S. C. & SIMEONE, C. 2010. Prognostic value of cytogenetic analysis in clear cell renal carcinoma: a study on 131 patients with long-term follow-up. *Anticancer Res,* 30**,** 4705-9.

ARSANIOUS, A., BJARNASON, G. A. & YOUSEF, G. M. 2009. From bench to bedside: current and future applications of molecular profiling in renal cell carcinoma. *Mol Cancer,* 8**,** 20.

BADER, G. D. & HOGUE, C. W. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics,* 4**,** 2.

BAIR, E. & TIBSHIRANI, R. 2004. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol,* 2**,** E108.

BANERJEE, M., GEORGE, J., SONG, E. Y., ROY, A. & HRYNIUK, W. 2004. Tree-based model for breast cancer prognostication. *J Clin Oncol,* 22**,** 2567-75.

BANKS, R. E., TIRUKONDA, P., TAYLOR, C., HORNIGOLD, N., ASTUTI, D., COHEN, D., MAHER, E. R., STANLEY, A. J., HARNDEN, P., JOYCE, A., KNOWLES, M. & SELBY, P. J. 2006. Genetic and epigenetic analysis of von Hippel-Lindau (VHL) gene alterations and relationship with clinical variables in sporadic renal cancer. *Cancer Res,* 66**,** 2000-11.

BARABASI, A. L. & OLTVAI, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet,* 5**,** 101-13.

BASHASHATI, A., HA, G., TONE, A., DING, J., PRENTICE, L. M., ROTH, A., ROSNER, J., SHUMANSKY, K., KALLOGER, S., SENZ, J., YANG, W., MCCONECHY, M., MELNYK, N., ANGLESIO, M., LUK, M. T., TSE, K., ZENG, T., MOORE, R., ZHAO, Y., MARRA, M. A., GILKS, B., YIP, S., HUNTSMAN, D. G., MCALPINE, J. N. & SHAH, S. P. 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol,* 231**,** 21-34.

BASHASHATI, A., HAFFARI, G., DING, J., HA, G., LUI, K., ROSNER, J., HUNTSMAN, D. G., CALDAS, C., APARICIO, S. A. & SHAH, S. P. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol,* 13**,** R124.

BASSO, K., MARGOLIN, A. A., STOLOVITZKY, G., KLEIN, U., DALLA-FAVERA, R. & CALIFANO, A. 2005. Reverse engineering of regulatory networks in human B cells. *Nat Genet,* 37**,** 382-90.

BELEUT, M., ZIMMERMANN, P., BAUDIS, M., BRUNI, N., BUHLMANN, P., LAULE, O., LUU, V. D., GRUISSEM, W., SCHRAML, P. & MOCH, H. 2012. Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient outcome. *BMC Cancer,* 12**,** 310.

BENGTSSON, H., NEUVIAL, P. & SPEED, T. P. 2010. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics,* 11**,** 245.

BENGTSSON, H., WIRAPATI, P. & SPEED, T. P. 2009. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics,* 25**,** 2149-56.

BEROUKHIM, R., BRUNET, J. P., DI NAPOLI, A., MERTZ, K. D., SEELEY, A., PIRES, M. M., LINHART, D., WORRELL, R. A., MOCH, H., RUBIN, M. A., SELLERS, W. R., MEYERSON, M., LINEHAN, W. M., KAELIN, W. G., JR. & SIGNORETTI, S. 2009. Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res,* 69**,** 4674-81.

BEROUKHIM, R., GETZ, G., NGHIEMPHU, L., BARRETINA, J., HSUEH, T., LINHART, D., VIVANCO, I., LEE, J. C., HUANG, J. H., ALEXANDER, S., DU, J., KAU, T., THOMAS, R. K., SHAH, K., SOTO, H., PERNER, S., PRENSNER, J., DEBIASI, R. M., DEMICHELIS, F., HATTON, C., RUBIN, M. A., GARRAWAY, L. A., NELSON, S. F., LIAU, L., MISCHEL, P. S., CLOUGHESY, T. F., MEYERSON, M., GOLUB, T. A., LANDER, E. S., MELLINGHOFF, I. K. & SELLERS, W. R. 2007. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A,* 104**,** 20007-12.

BEROUKHIM, R., MERMEL, C. H., PORTER, D., WEI, G., RAYCHAUDHURI, S., DONOVAN, J., BARRETINA, J., BOEHM, J. S., DOBSON, J., URASHIMA, M., MC HENRY, K. T., PINCHBACK, R. M., LIGON, A. H., CHO, Y. J., HAERY, L., GREULICH, H., REICH, M., WINCKLER, W., LAWRENCE, M. S., WEIR, B. A., TANAKA, K. E., CHIANG, D. Y., BASS, A. J., LOO, A., HOFFMAN, C., PRENSNER, J., LIEFELD, T., GAO, Q., YECIES, D., SIGNORETTI, S., MAHER, E., KAYE, F. J., SASAKI, H., TEPPER, J. E., FLETCHER, J. A., TABERNERO, J., BASELGA, J., TSAO, M. S., DEMICHELIS, F., RUBIN, M. A., JANNE, P. A., DALY, M. J., NUCERA, C., LEVINE, R. L., EBERT, B. L., GABRIEL, S., RUSTGI, A. K., ANTONESCU, C. R., LADANYI, M., LETAI, A., GARRAWAY, L. A., LODA, M., BEER, D. G., TRUE, L. D., OKAMOTO, A., POMEROY, S. L., SINGER, S., GOLUB, T. R., LANDER, E. S., GETZ, G., SELLERS, W. R. & MEYERSON, M. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature,* 463**,** 899-905.

BLAGOSKLONNY, M. V. & HALL, M. N. 2009. Growth and aging: a common molecular mechanism. *Aging (Albany NY),* 1**,** 357-62.

BLAND, J. M. & ALTMAN, D. G. 1998. Survival probabilities (the Kaplan-Meier method). *BMJ,* 317**,** 1572.

BLAND, J. M. & ALTMAN, D. G. 2004. The logrank test. *BMJ,* 328**,** 1073.

BOSTROM, A. K., LINDGREN, D., JOHANSSON, M. E. & AXELSON, H. 2013. Effects of TGF-beta signaling in clear cell renal cell carcinoma cells. *Biochem Biophys Res Commun,* 435**,** 126-33.

BRADBURN, M. J., CLARK, T. G., LOVE, S. B. & ALTMAN, D. G. 2003. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer,* 89**,** 431-6.

BRANNON, A. R. & RATHMELL, W. K. 2010. Renal cell carcinoma: where will the state-of-the-art lead us? *Curr Oncol Rep,* 12**,** 193-201.

BRANNON, A. R., REDDY, A., SEILER, M., ARREOLA, A., MOORE, D. T., PRUTHI, R. S., WALLEN, E. M., NIELSEN, M. E., LIU, H., NATHANSON, K. L., LJUNGBERG, B., ZHAO, H., BROOKS, J. D., GANESAN, S., BHANOT, G. & RATHMELL, W. K. 2010. Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. *Genes Cancer,* 1**,** 152-163.

BRAUCH, H., WEIRICH, G., BRIEGER, J., GLAVAC, D., RODL, H., EICHINGER, M., FEURER, M., WEIDT, E., PURANAKANITSTHA, C., NEUHAUS, C., POMER, S., BRENNER, W., SCHIRMACHER, P., STORKEL, S., ROTTER, M., MASERA, A., GUGELER, N. & DECKER, H. J. 2000. VHL alterations in human clear cell renal cell carcinoma: association with advanced tumor stage and a novel hot spot mutation. *Cancer Res,* 60**,** 1942-8.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

BROOKS, S. A., BRANNON, A. R., PARKER, J. S., FISHER, J. C., SEN, O., KATTAN, M. W., HAKIMI, A. A., HSIEH, J. J., CHOUEIRI, T. K., TAMBOLI, P., MARANCHIE, J. K., HINDS, P., MILLER, C. R., NIELSEN, M. E. & RATHMELL, W. K. 2014. ClearCode34: A Prognostic Risk Predictor for Localized Clear Cell Renal Cell Carcinoma. *Eur Urol*.

BRUGAROLAS, J. 2014. Molecular genetics of clear-cell renal cell carcinoma. *J Clin Oncol,* 32**,** 1968-76.

BRUGAROLAS, J. B., VAZQUEZ, F., REDDY, A., SELLERS, W. R. & KAELIN, W. G., JR. 2003. TSC2 regulates VEGF through mTOR-dependent and -independent pathways. *Cancer Cell,* 4**,** 147-58.

BRUNELLI, M., ECCHER, A., GOBBO, S., FICARRA, V., NOVARA, G., COSSU-ROCCA, P., BONETTI, F., MENESTRINA, F., CHENG, L., EBLE, J. N. & MARTIGNONI, G. 2008. Loss of chromosome 9p is an independent prognostic factor in patients with clear cell renal cell carcinoma. *Mod Pathol,* 21**,** 1-6.

BRUNET, J. P., TAMAYO, P., GOLUB, T. R. & MESIROV, J. P. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A,* 101**,** 4164-9.

BURRELL, R. A., MCCLELLAND, S. E., ENDESFELDER, D., GROTH, P., WELLER, M. C., SHAIKH, N., DOMINGO, E., KANU, N., DEWHURST, S. M., GRONROOS, E., CHEW, S. K., ROWAN, A. J., SCHENK, A., SHEFFER, M., HOWELL, M., KSCHISCHO, M., BEHRENS, A., HELLEDAY, T., BARTEK, J., TOMLINSON, I. P. & SWANTON, C. 2013a. Replication stress links structural and numerical cancer chromosomal instability. *Nature,* 494**,** 492-6.

BURRELL, R. A., MCGRANAHAN, N., BARTEK, J. & SWANTON, C. 2013b. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature,* 501**,** 338-45.

BURROWS, A. E., SMOGORZEWSKA, A. & ELLEDGE, S. J. 2010. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. *Proc Natl Acad Sci U S A,* 107**,** 14280-5.

CAMPBELL, P. J., YACHIDA, S., MUDIE, L. J., STEPHENS, P. J., PLEASANCE, E. D., STEBBINGS, L. A., MORSBERGER, L. A., LATIMER, C., MCLAREN, S., LIN, M. L., MCBRIDE, D. J., VARELA, I., NIK-ZAINAL, S. A., LEROY, C., JIA, M., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., GRIFFIN, C. A., BURTON, J., SWERDLOW, H., QUAIL, M. A., STRATTON, M. R., IACOBUZIO-DONAHUE, C. & FUTREAL, P. A. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature,* 467**,** 1109-13.

CANCER GENOME ATLAS RESEARCH, N. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455**,** 1061-8.

CAROSELLA, E. D., PLOUSSARD, G., LEMAOULT, J. & DESGRANDCHAMPS, F. 2015. A Systematic Review of Immunotherapy in Urologic Cancer: Evolving Roles for Targeting of CTLA-4, PD-1/PD-L1, and HLA-G. *Eur Urol,* 68**,** 267-79.

CARRO, M. S., LIM, W. K., ALVAREZ, M. J., BOLLO, R. J., ZHAO, X., SNYDER, E. Y., SULMAN, E. P., ANNE, S. L., DOETSCH, F., COLMAN, H., LASORELLA, A., ALDAPE, K., CALIFANO, A. & IAVARONE, A. 2010. The transcriptional network for mesenchymal transformation of brain tumours. *Nature,* 463**,** 318-25.

CHAN, K. C., JIANG, P., ZHENG, Y. W., LIAO, G. J., SUN, H., WONG, J., SIU, S. S., CHAN, W. C., CHAN, S. L., CHAN, A. T., LAI, P. B., CHIU, R. W. & LO, Y. M. 2013. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem,* 59**,** 211-24.

CHATR-ARYAMONTRI, A., BREITKREUTZ, B. J., OUGHTRED, R., BOUCHER, L., HEINICKE, S., CHEN, D., STARK, C., BREITKREUTZ, A., KOLAS, N., O'DONNELL, L., REGULY, T., NIXON, J., RAMAGE, L., WINTER, A., SELLAM, A., CHANG, C., HIRSCHMAN, J., THEESFELD, C., RUST, J., LIVSTONE, M. S., DOLINSKI, K. & TYERS, M. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Res,* 43**,** D470-8.

CHEN, M., YE, Y., YANG, H., TAMBOLI, P., MATIN, S., TANNIR, N. M., WOOD, C. G., GU, J. & WU, X. 2009. Genome-wide profiling of chromosomal alterations in renal cell carcinoma using high-density single nucleotide polymorphism arrays. *Int J Cancer,* 125**,** 2342-8.

CHENG, T. M., GULATI, S., AGIUS, R. & BATES, P. A. 2012. Understanding cancer mechanisms through network dynamics. *Brief Funct Genomics,* 11**,** 543-60.

CHO, J. S., PARK, M. H., LEE, J. S. & YOON, J. H. 2015. Reduced MUC4 expression is a late event in breast carcinogenesis and is correlated with increased infiltration of immune cells as well as promoter hypermethylation in invasive breast carcinoma. *Appl Immunohistochem Mol Morphol,* 23**,** 44-53.

CHOW, T. F., YOUSSEF, Y. M., LIANIDOU, E., ROMASCHIN, A. D., HONEY, R. J., STEWART, R., PACE, K. T. & YOUSEF, G. M. 2010. Differential expression profiling of microRNAs and their potential involvement in renal cell carcinoma pathogenesis. *Clin Biochem,* 43**,** 150-8.

CHUANG, H. Y., LEE, E., LIU, Y. T., LEE, D. & IDEKER, T. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol,* 3**,** 140.

CIRIELLO, G., CERAMI, E., SANDER, C. & SCHULTZ, N. 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res,* 22**,** 398-406.

CLARK, T. G., BRADBURN, M. J., LOVE, S. B. & ALTMAN, D. G. 2003. Survival analysis part I: basic concepts and first analyses. *Br J Cancer,* 89**,** 232-8.

CLIFFORD, S. C., PROWSE, A. H., AFFARA, N. A., BUYS, C. H. & MAHER, E. R. 1998. Inactivation of the von Hippel-Lindau (VHL) tumour suppressor gene and allelic losses at chromosome arm 3p in primary renal cell carcinoma: evidence for a VHL-independent pathway in clear cell renal tumourigenesis. *Genes Chromosomes Cancer,* 22**,** 200-9.

COHEN, S. J., PUNT, C. J., IANNOTTI, N., SAIDMAN, B. H., SABBATH, K. D., GABRAIL, N. Y., PICUS, J., MORSE, M., MITCHELL, E., MILLER, M. C., DOYLE, G. V., TISSING, H., TERSTAPPEN, L. W. & MEROPOL, N. J. 2008. Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer. *J Clin Oncol,* 26**,** 3213-21.

COX, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series. Series B (Methodological),* 34**,** 187-220.

CRISTOFANILLI, M., BUDD, G. T., ELLIS, M. J., STOPECK, A., MATERA, J., MILLER, M. C., REUBEN, J. M., DOYLE, G. V., ALLARD, W. J., TERSTAPPEN, L. W. &

HAYES, D. F. 2004. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med,* 351**,** 781-91.

CROFT, D., O'KELLY, G., WU, G., HAW, R., GILLESPIE, M., MATTHEWS, L., CAUDY, M., GARAPATI, P., GOPINATH, G., JASSAL, B., JUPE, S., KALATSKAYA, I., MAHAJAN, S., MAY, B., NDEGWA, N., SCHMIDT, E., SHAMOVSKY, V., YUNG, C., BIRNEY, E., HERMJAKOB, H., D'EUSTACHIO, P. & STEIN, L. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res,* 39**,** D691-7.

DAHINDEN, C., INGOLD, B., WILD, P., BOYSEN, G., LUU, V. D., MONTANI, M., KRISTIANSEN, G., SULSER, T., BUHLMANN, P., MOCH, H. & SCHRAML, P. 2010. Mining tissue microarray data to uncover combinations of biomarker expression patterns that improve intermediate staging and grading of clear cell renal cell cancer. *Clin Cancer Res,* 16**,** 88-98.

DALGLIESH, G. L., FURGE, K., GREENMAN, C., CHEN, L., BIGNELL, G., BUTLER, A., DAVIES, H., EDKINS, S., HARDY, C., LATIMER, C., TEAGUE, J., ANDREWS, J., BARTHORPE, S., BEARE, D., BUCK, G., CAMPBELL, P. J., FORBES, S., JIA, M., JONES, D., KNOTT, H., KOK, C. Y., LAU, K. W., LEROY, C., LIN, M. L., MCBRIDE, D. J., MADDISON, M., MAGUIRE, S., MCLAY, K., MENZIES, A., MIRONENKO, T., MULDERRIG, L., MUDIE, L., O'MEARA, S., PLEASANCE, E., RAJASINGHAM, A., SHEPHERD, R., SMITH, R., STEBBINGS, L., STEPHENS, P., TANG, G., TARPEY, P. S., TURRELL, K., DYKEMA, K. J., KHOO, S. K., PETILLO, D., WONDERGEM, B., ANEMA, J., KAHNOSKI, R. J., TEH, B. T., STRATTON, M. R. & FUTREAL, P. A. 2010. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature,* 463**,** 360-3.

DAWSON, S. J., TSUI, D. W., MURTAZA, M., BIGGS, H., RUEDA, O. M., CHIN, S. F., DUNNING, M. J., GALE, D., FORSHEW, T., MAHLER-ARAUJO, B., RAJAN, S., HUMPHRAY, S., BECQ, J., HALSALL, D., WALLIS, M., BENTLEY, D., CALDAS, C. & ROSENFELD, N. 2013. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med,* 368**,** 1199-209.

DE BONO, J. S., SCHER, H. I., MONTGOMERY, R. B., PARKER, C., MILLER, M. C., TISSING, H., DOYLE, G. V., TERSTAPPEN, L. W., PIENTA, K. J. & RAGHAVAN, D. 2008. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clin Cancer Res,* 14**,** 6302-9.

DEPRIMO, S. E., BELLO, C. L., SMERAGLIA, J., BAUM, C. M., SPINELLA, D., RINI, B. I., MICHAELSON, M. D. & MOTZER, R. J. 2007. Circulating protein biomarkers of pharmacodynamic activity of sunitinib in patients with metastatic renal cell carcinoma: modulation of VEGF and VEGF-related proteins. *J Transl Med,* 5**,** 32.

DIAZ, L. A., JR., WILLIAMS, R. T., WU, J., KINDE, I., HECHT, J. R., BERLIN, J., ALLEN, B., BOZIC, I., REITER, J. G., NOWAK, M. A., KINZLER, K. W., OLINER, K. S. & VOGELSTEIN, B. 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature,* 486**,** 537-40.

DING, L., LEY, T. J., LARSON, D. E., MILLER, C. A., KOBOLDT, D. C., WELCH, J. S., RITCHEY, J. K., YOUNG, M. A., LAMPRECHT, T., MCLELLAN, M. D., MCMICHAEL, J. F., WALLIS, J. W., LU, C., SHEN, D., HARRIS, C. C., DOOLING, D. J., FULTON, R. S., FULTON, L. L., CHEN, K., SCHMIDT, H., KALICKI-VEIZER, J., MAGRINI, V. J., COOK, L., MCGRATH, S. D., VICKERY, T. L., WENDL, M. C., HEATH, S., WATSON, M. A., LINK, D. C., TOMASSON, M. H., SHANNON, W. D., PAYTON, J. E., KULKARNI, S., WESTERVELT, P., WALTER, M. J., GRAUBERT, T. A., MARDIS, E. R., WILSON, R. K. &

DIPERSIO, J. F. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature,* 481**,** 506-10.

DUAN, D. R., PAUSE, A., BURGESS, W. H., ASO, T., CHEN, D. Y., GARRETT, K. P., CONAWAY, R. C., CONAWAY, J. W., LINEHAN, W. M. & KLAUSNER, R. D. 1995. Inhibition of transcription elongation by the VHL tumor suppressor protein. *Science,* 269**,** 1402-6.

DUNS, G., HOFSTRA, R. M., SIETZEMA, J. G., HOLLEMA, H., VAN DUIVENBODE, I., KUIK, A., GIEZEN, C., JAN, O., BERGSMA, J. J., BIJNEN, H., VAN DER VLIES, P., VAN DEN BERG, E. & KOK, K. 2012. Targeted exome sequencing in clear cell renal cell carcinoma tumors suggests aberrant chromatin regulation as a crucial step in ccRCC development. *Hum Mutat,* 33**,** 1059-62.

DUNS, G., VAN DEN BERG, E., VAN DUIVENBODE, I., OSINGA, J., HOLLEMA, H., HOFSTRA, R. M. & KOK, K. 2010. Histone methyltransferase gene SETD2 is a novel tumor suppressor gene in clear cell renal cell carcinoma. *Cancer Res,* 70**,** 4287-91.

EICHELBERG, C., JUNKER, K., LJUNGBERG, B. & MOCH, H. 2009. Diagnostic and prognostic molecular markers for renal cell carcinoma: a critical appraisal of the current state of research and clinical applicability. *Eur Urol,* 55**,** 851-63.

ELFVING, P., MANDAHL, N., LUNDGREN, R., LIMON, J., BAK-JENSEN, E., FERNO, M., OLSSON, H. & MITELMAN, F. 1997. Prognostic implications of cytogenetic findings in kidney cancer. *Br J Urol,* 80**,** 698-706.

FICARRA, V., MARTIGNONI, G., LOHSE, C., NOVARA, G., PEA, M., CAVALLERI, S. & ARTIBANI, W. 2006. External validation of the Mayo Clinic Stage, Size, Grade and Necrosis (SSIGN) score to predict cancer specific survival using a European series of conventional renal cell carcinoma. *J Urol,* 175**,** 1235-9.

FIELDS, S. & SONG, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature,* 340**,** 245-6.

FISHER, R., HORSWELL, S., ROWAN, A., SALM, M. P., DE BRUIN, E. C., GULATI, S., MCGRANAHAN, N., STARES, M., GERLINGER, M., VARELA, I., CROCKFORD, A., FAVERO, F., QUIDVILLE, V., ANDRE, F., NAVAS, C., GRONROOS, E., NICOL, D., HAZELL, S., HROUDA, D., O'BRIEN, T., MATTHEWS, N., PHILLIMORE, B., BEGUM, S., RABINOWITZ, A., BIGGS, J., BATES, P. A., MCDONALD, N. Q., STAMP, G., SPENCER-DENE, B., HSIEH, J. J., XU, J., PICKERING, L., GORE, M., LARKIN, J. & SWANTON, C. 2014. Development of synchronous VHL syndrome tumors reveals contingencies and constraints to tumor evolution. *Genome Biol,* 15**,** 433.

FRANK, I., BLUTE, M. L., CHEVILLE, J. C., LOHSE, C. M., WEAVER, A. L. & ZINCKE, H. 2002. An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score. *J Urol,* 168**,** 2395-400.

FYFE, G., FISHER, R. I., ROSENBERG, S. A., SZNOL, M., PARKINSON, D. R. & LOUIE, A. C. 1995. Results of treatment of 255 patients with metastatic renal cell carcinoma who received high-dose recombinant interleukin-2 therapy. *J Clin Oncol,* 13**,** 688-96.

GARCIA-DONAS, J., ESTEBAN, E., LEANDRO-GARCIA, L. J., CASTELLANO, D. E., DEL ALBA, A. G., CLIMENT, M. A., ARRANZ, J. A., GALLARDO, E., PUENTE, J., BELLMUNT, J., MELLADO, B., MARTINEZ, E., MORENO, F., FONT, A., ROBLEDO, M. & RODRIGUEZ-ANTONA, C. 2011. Single nucleotide polymorphism associations with response and toxic effects in patients with advanced renal-cell carcinoma treated with first-line sunitinib: a multicentre, observational, prospective study. *Lancet Oncol,* 12**,** 1143-50.

GERLINGER, M., CATTO, J. W., ORNTOFT, T. F., REAL, F. X., ZWARTHOFF, E. C. & SWANTON, C. 2015. Intratumour heterogeneity in urologic cancers: from molecular evidence to clinical implications. *Eur Urol,* 67**,** 729-37.

GERLINGER, M., HORSWELL, S., LARKIN, J., ROWAN, A. J., SALM, M. P., VARELA, I., FISHER, R., MCGRANAHAN, N., MATTHEWS, N., SANTOS, C. R., MARTINEZ, P., PHILLIMORE, B., BEGUM, S., RABINOWITZ, A., SPENCER-DENE, B., GULATI, S., BATES, P. A., STAMP, G., PICKERING, L., GORE, M., NICOL, D. L., HAZELL, S., FUTREAL, P. A., STEWART, A. & SWANTON, C. 2014a. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet,* 46**,** 225-233.

GERLINGER, M., MCGRANAHAN, N., DEWHURST, S. M., BURRELL, R. A., TOMLINSON, I. & SWANTON, C. 2014b. Cancer: evolution within a lifetime. *Annu Rev Genet,* 48**,** 215-36.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. & SWANTON, C. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.

GIRGIS, A. H., IAKOVLEV, V. V., BEHESHTI, B., BAYANI, J., SQUIRE, J. A., BUI, A., MANKARUOS, M., YOUSSEF, Y., KHALIL, B., KHELLA, H., PASIC, M. & YOUSEF, G. M. 2012. Multilevel whole-genome analysis reveals candidate biomarkers in clear cell renal cell carcinoma. *Cancer Res,* 72**,** 5273-84.

GLAS, A. M., FLOORE, A., DELAHAYE, L. J., WITTEVEEN, A. T., POVER, R. C., BAKX, N., LAHTI-DOMENICI, J. S., BRUINSMA, T. J., WARMOES, M. O., BERNARDS, R., WESSELS, L. F. & VAN'T VEER, L. J. 2006. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics,* 7**,** 278.

GNARRA, J. R., TORY, K., WENG, Y., SCHMIDT, L., WEI, M. H., LI, H., LATIF, F., LIU, S., CHEN, F., DUH, F. M. & ET AL. 1994. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat Genet,* 7**,** 85-90.

GOH, K. I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. & BARABASI, A. L. 2007. The human disease network. *Proc Natl Acad Sci U S A,* 104**,** 8685-90.

GORDAN, J. D., LAL, P., DONDETI, V. R., LETRERO, R., PAREKH, K. N., OQUENDO, C. E., GREENBERG, R. A., FLAHERTY, K. T., RATHMELL, W. K., KEITH, B., SIMON, M. C. & NATHANSON, K. L. 2008. HIF-alpha effects on c-Myc distinguish two subtypes of sporadic VHL-deficient clear cell renal carcinoma. *Cancer Cell,* 14**,** 435-46.

GORDAN, J. D. & SIMON, M. C. 2007. Hypoxia-inducible factors: central regulators of the tumor phenotype. *Curr Opin Genet Dev,* 17**,** 71-7.

GOSSAGE, L., EISEN, T. & MAHER, E. R. 2015. VHL, the story of a tumour suppressor gene. *Nat Rev Cancer,* 15**,** 55-64.

GU, Y., WANG, H., QIN, Y., ZHANG, Y., ZHAO, W., QI, L., ZHANG, Y., WANG, C. & GUO, Z. 2013. Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma. *Mol Biosyst,* 9**,** 467-77.

GUICHARD, C., AMADDEO, G., IMBEAUD, S., LADEIRO, Y., PELLETIER, L., MAAD, I. B., CALDERARO, J., BIOULAC-SAGE, P., LETEXIER, M., DEGOS, F., CLEMENT, B., BALABAUD, C., CHEVET, E., LAURENT, A., COUCHY, G., LETOUZE, E., CALVO, F. & ZUCMAN-ROSSI, J. 2012. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet,* 44, 694-8.

GULATI, S., CHENG, T. M. & BATES, P. A. 2013. Cancer networks and beyond: interpreting mutations using the human interactome and protein structure. *Semin Cancer Biol,* 23, 219-26.

GULATI, S., MARTINEZ, P., JOSHI, T., BIRKBAK, N. J., SANTOS, C. R., ROWAN, A. J., PICKERING, L., GORE, M., LARKIN, J., SZALLASI, Z., BATES, P. A., SWANTON, C. & GERLINGER, M. 2014. Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. *Eur Urol,* 66, 936-48.

GULATI, S., TURAJLIC, S., LARKIN, J., BATES, P. A. & SWANTON, C. 2015. Relapse models for clear cell renal carcinoma. *Lancet Oncol,* 16, e376-8.

GUNAWAN, B., HUBER, W., HOLTRUP, M., VON HEYDEBRECK, A., EFFERTH, T., POUSTKA, A., RINGERT, R. H., JAKSE, G. & FUZESI, L. 2001. Prognostic impacts of cytogenetic findings in clear cell renal cell carcinoma: gain of 5q31-qter predicts a distinct clinical phenotype with favorable prognosis. *Cancer Res,* 61, 7731-8.

GUO, G., GUI, Y., GAO, S., TANG, A., HU, X., HUANG, Y., JIA, W., LI, Z., HE, M., SUN, L., SONG, P., SUN, X., ZHAO, X., YANG, S., LIANG, C., WAN, S., ZHOU, F., CHEN, C., ZHU, J., LI, X., JIAN, M., ZHOU, L., YE, R., HUANG, P., CHEN, J., JIANG, T., LIU, X., WANG, Y., ZOU, J., JIANG, Z., WU, R., WU, S., FAN, F., ZHANG, Z., LIU, L., YANG, R., LIU, X., WU, H., YIN, W., ZHAO, X., LIU, Y., PENG, H., JIANG, B., FENG, Q., LI, C., XIE, J., LU, J., KRISTIANSEN, K., LI, Y., ZHANG, X., LI, S., WANG, J., YANG, H., CAI, Z. & WANG, J. 2012. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat Genet,* 44, 17-9.

HAKIMI, A. A., OSTROVNAYA, I., REVA, B., SCHULTZ, N., CHEN, Y. B., GONEN, M., LIU, H., TAKEDA, S., VOSS, M. H., TICKOO, S. K., REUTER, V. E., RUSSO, P., CHENG, E. H., SANDER, C., MOTZER, R. J. & HSIEH, J. J. 2013. Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. *Clin Cancer Res,* 19, 3259-67.

HAN, K. R., BLEUMER, I., PANTUCK, A. J., KIM, H. L., DOREY, F. J., JANZEN, N. K., ZISMAN, A., DINNEY, C. P., WOOD, C. G., SWANSON, D. A., SAID, J. W., FIGLIN, R. A., MULDERS, P. F. & BELLDEGRUN, A. S. 2003. Validation of an integrated staging system toward improved prognostication of patients with localized renal cell carcinoma in an international population. *J Urol,* 170, 2221-4.

HAN, L., CHEN, W. & ZHAO, Q. 2014. Prognostic value of circulating tumor cells in patients with pancreatic cancer: a meta-analysis. *Tumour Biol,* 35, 2473-80.

HANAHAN, D. & WEINBERG, R. A. 2000. The hallmarks of cancer. *Cell,* 100, 57-70.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144, 646-74.

HAW, R. & STEIN, L. 2012. Using the reactome database. *Curr Protoc Bioinformatics,* Chapter 8, Unit8 7.

HE, C. & KLIONSKY, D. J. 2009. Regulation mechanisms and signaling pathways of autophagy. *Annu Rev Genet,* 43, 67-93.

HENG, D. Y., XIE, W., REGAN, M. M., WARREN, M. A., GOLSHAYAN, A. R., SAHI, C., EIGL, B. J., RUETHER, J. D., CHENG, T., NORTH, S., VENNER, P., KNOX, J. J., CHI, K. N., KOLLMANNSBERGER, C., MCDERMOTT, D. F., OH, W. K., ATKINS, M. B., BUKOWSKI, R. M., RINI, B. I. & CHOUEIRI, T. K. 2009. Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted agents: results from a large, multicenter study. *J Clin Oncol,* 27, 5794-9.

HERMAN, J. G., LATIF, F., WENG, Y., LERMAN, M. I., ZBAR, B., LIU, S., SAMID, D., DUAN, D. S., GNARRA, J. R., LINEHAN, W. M. & ET AL. 1994. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A,* 91**,** 9700-4.

HERZ, H. M., MADDEN, L. D., CHEN, Z., BOLDUC, C., BUFF, E., GUPTA, R., DAVULURI, R., SHILATIFARD, A., HARIHARAN, I. K. & BERGMANN, A. 2010. The H3K27me3 demethylase dUTX is a suppressor of Notch- and Rb-dependent tumors in Drosophila. *Mol Cell Biol,* 30**,** 2485-97.

HILTERMANN, T. J., PORE, M. M., VAN DEN BERG, A., TIMENS, W., BOEZEN, H. M., LIESKER, J. J., SCHOUWINK, J. H., WIJNANDS, W. J., KERNER, G. S., KRUYT, F. A., TISSING, H., TIBBE, A. G., TERSTAPPEN, L. W. & GROEN, H. J. 2012. Circulating tumor cells in small-cell lung cancer: a predictive and prognostic factor. *Ann Oncol,* 23**,** 2937-42.

HOFMAN, V., BONNETAUD, C., ILIE, M. I., VIELH, P., VIGNAUD, J. M., FLEJOU, J. F., LANTUEJOUL, S., PIATON, E., MOURAD, N., BUTORI, C., SELVA, E., POUDENX, M., SIBON, S., KELHEF, S., VENISSAC, N., JAIS, J. P., MOUROUX, J., MOLINA, T. J. & HOFMAN, P. 2011. Preoperative circulating tumor cell detection using the isolation by size of epithelial tumor cell method for patients with lung cancer is a new prognostic biomarker. *Clin Cancer Res,* 17**,** 827-35.

HOFREE, M., SHEN, J. P., CARTER, H., GROSS, A. & IDEKER, T. 2013. Network-based stratification of tumor mutations. *Nat Methods,* 10**,** 1108-15.

HORNBERG, J. J., BRUGGEMAN, F. J., WESTERHOFF, H. V. & LANKELMA, J. 2006. Cancer: a Systems Biology disease. *Biosystems,* 83**,** 81-90.

HOTHORN, T., HORNIK, K. & ZEILEIS, A. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics,* 15**,** 651--674.

HUANG, D., DING, Y., LI, Y., LUO, W. M., ZHANG, Z. F., SNIDER, J., VANDENBELDT, K., QIAN, C. N. & TEH, B. T. 2010a. Sunitinib acts primarily on tumor endothelium rather than tumor cells to inhibit the growth of renal cell carcinoma. *Cancer Res,* 70**,** 1053-62.

HUANG, D., DING, Y., ZHOU, M., RINI, B. I., PETILLO, D., QIAN, C. N., KAHNOSKI, R., FUTREAL, P. A., FURGE, K. A. & TEH, B. T. 2010b. Interleukin-8 mediates resistance to antiangiogenic agent sunitinib in renal cell carcinoma. *Cancer Res,* 70**,** 1063-71.

HUANG, Y., DAI, Y., YANG, J., CHEN, T., YIN, Y., TANG, M., HU, C. & ZHANG, L. 2009. Microarray analysis of microRNA expression in renal clear cell carcinoma. *Eur J Surg Oncol,* 35**,** 1119-23.

IDEKER, T. & SHARAN, R. 2008. Protein networks in disease. *Genome Res,* 18**,** 644-52.

IINUMA, H., WATANABE, T., MIMORI, K., ADACHI, M., HAYASHI, N., TAMURA, J., MATSUDA, K., FUKUSHIMA, R., OKINAGA, K., SASAKO, M. & MORI, M. 2011. Clinical significance of circulating tumor cells, including cancer stem-like cells, in peripheral blood for recurrence and prognosis in patients with Dukes' stage B and C colorectal cancer. *J Clin Oncol,* 29**,** 1547-55.

INTERNATIONAL CANCER GENOME, C., HUDSON, T. J., ANDERSON, W., ARTEZ, A., BARKER, A. D., BELL, C., BERNABE, R. R., BHAN, M. K., CALVO, F., EEROLA, I., GERHARD, D. S., GUTTMACHER, A., GUYER, M., HEMSLEY, F. M., JENNINGS, J. L., KERR, D., KLATT, P., KOLAR, P., KUSADA, J., LANE, D. P., LAPLACE, F., YOUYONG, L., NETTEKOVEN, G., OZENBERGER, B., PETERSON, J., RAO, T. S., REMACLE, J., SCHAFER, A. J., SHIBATA, T., STRATTON, M. R., VOCKLEY, J. G., WATANABE, K., YANG, H., YUEN, M. M., KNOPPERS, B. M., BOBROW, M., CAMBON-THOMSEN, A., DRESSLER, L.

G., DYKE, S. O., JOLY, Y., KATO, K., KENNEDY, K. L., NICOLAS, P., PARKER, M. J., RIAL-SEBBAG, E., ROMEO-CASABONA, C. M., SHAW, K. M., WALLACE, S., WIESNER, G. L., ZEPS, N., LICHTER, P., BIANKIN, A. V., CHABANNON, C., CHIN, L., CLEMENT, B., DE ALAVA, E., DEGOS, F., FERGUSON, M. L., GEARY, P., HAYES, D. N., HUDSON, T. J., JOHNS, A. L., KASPRZYK, A., NAKAGAWA, H., PENNY, R., PIRIS, M. A., SARIN, R., SCARPA, A., SHIBATA, T., VAN DE VIJVER, M., FUTREAL, P. A., ABURATANI, H., BAYES, M., BOTWELL, D. D., CAMPBELL, P. J., ESTIVILL, X., GERHARD, D. S., GRIMMOND, S. M., GUT, I., HIRST, M., LOPEZ-OTIN, C., MAJUMDER, P., MARRA, M., MCPHERSON, J. D., NAKAGAWA, H., NING, Z., PUENTE, X. S., RUAN, Y., SHIBATA, T., STRATTON, M. R., STUNNENBERG, H. G., SWERDLOW, H., VELCULESCU, V. E., WILSON, R. K., XUE, H. H., YANG, L., SPELLMAN, P. T., BADER, G. D., BOUTROS, P. C., CAMPBELL, P. J., et al. 2010. International network of cancer genome projects. *Nature,* 464**,** 993-8.

ISHWARAN, H. & KOGALUR, U. B. 2015. Random Forests for Survival, Regression and Classification (RF-SRC).

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. & LAUER, M. S. 2008. Random survival forests. *Ann. Appl. Statist.,* 2**,** 841--860.

JAMAL-HANJANI, M., HACKSHAW, A., NGAI, Y., SHAW, J., DIVE, C., QUEZADA, S., MIDDLETON, G., DE BRUIN, E., LE QUESNE, J., SHAFI, S., FALZON, M., HORSWELL, S., BLACKHALL, F., KHAN, I., JANES, S., NICOLSON, M., LAWRENCE, D., FORSTER, M., FENNELL, D., LEE, S. M., LESTER, J., KERR, K., MULLER, S., ILES, N., SMITH, S., MURUGAESU, N., MITTER, R., SALM, M., STUART, A., MATTHEWS, N., ADAMS, H., AHMAD, T., ATTANOOS, R., BENNETT, J., BIRKBAK, N. J., BOOTON, R., BRADY, G., BUCHAN, K., CAPITANO, A., CHETTY, M., COBBOLD, M., CROSBIE, P., DAVIES, H., DENISON, A., DJEARMAN, M., GOLDMAN, J., HASWELL, T., JOSEPH, L., KORNASZEWSKA, M., KREBS, M., LANGMAN, G., MACKENZIE, M., MILLAR, J., MORGAN, B., NAIDU, B., NONAKA, D., PEGGS, K., PRITCHARD, C., REMMEN, H., ROWAN, A., SHAH, R., SMITH, E., SUMMERS, Y., TAYLOR, M., VEERIAH, S., WALLER, D., WILCOX, B., WILCOX, M., WOOLHOUSE, I., MCGRANAHAN, N. & SWANTON, C. 2014. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol,* 12**,** e1001906.

JONASCH, E., FUTREAL, P. A., DAVIS, I. J., BAILEY, S. T., KIM, W. Y., BRUGAROLAS, J., GIACCIA, A. J., KURBAN, G., PAUSE, A., FRYDMAN, J., ZURITA, A. J., RINI, B. I., SHARMA, P., ATKINS, M. B., WALKER, C. L. & RATHMELL, W. K. 2012. State of the science: an update on renal cell carcinoma. *Mol Cancer Res,* 10**,** 859-80.

JONASCH, E., GAO, J. & RATHMELL, W. K. 2014. Renal cell carcinoma. *BMJ,* 349**,** g4797.

JONES, J., OTU, H., SPENTZOS, D., KOLIA, S., INAN, M., BEECKEN, W. D., FELLBAUM, C., GU, X., JOSEPH, M., PANTUCK, A. J., JONAS, D. & LIBERMANN, T. A. 2005. Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res,* 11**,** 5730-9.

JONSSON, P. F. & BATES, P. A. 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics,* 22**,** 2291-7.

JUAN, D., ALEXE, G., ANTES, T., LIU, H., MADABHUSHI, A., DELISI, C., GANESAN, S., BHANOT, G. & LIOU, L. S. 2010. Identification of a microRNA panel for clear-cell kidney cancer. *Urology,* 75**,** 835-41.

JUNKER, K., FICARRA, V., KWON, E. D., LEIBOVICH, B. C., THOMPSON, R. H. & OOSTERWIJK, E. 2013. Potential role of genetic markers in the management of kidney cancer. *Eur Urol,* 63**,** 333-40.

JUNKER, K., WEIRICH, G., AMIN, M. B., MORAVEK, P., HINDERMANN, W. & SCHUBERT, J. 2003. Genetic subtyping of renal cell carcinoma by comparative genomic hybridization. *Recent Results Cancer Res,* 162**,** 169-75.

KANDOTH, C., MCLELLAN, M. D., VANDIN, F., YE, K., NIU, B., LU, C., XIE, M., ZHANG, Q., MCMICHAEL, J. F., WYCZALKOWSKI, M. A., LEISERSON, M. D., MILLER, C. A., WELCH, J. S., WALTER, M. J., WENDL, M. C., LEY, T. J., WILSON, R. K., RAPHAEL, B. J. & DING, L. 2013. Mutational landscape and significance across 12 major cancer types. *Nature,* 502**,** 333-9.

KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M. & TANABE, M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res,* 40**,** D109-14.

KANU, N., GRONROOS, E., MARTINEZ, P., BURRELL, R. A., YI GOH, X., BARTKOVA, J., MAYA-MENDOZA, A., MISTRIK, M., ROWAN, A. J., PATEL, H., RABINOWITZ, A., EAST, P., WILSON, G., SANTOS, C. R., MCGRANAHAN, N., GULATI, S., GERLINGER, M., BIRKBAK, N. J., JOSHI, T., ALEXANDROV, L. B., STRATTON, M. R., POWLES, T., MATTHEWS, N., BATES, P. A., STEWART, A., SZALLASI, Z., LARKIN, J., BARTEK, J. & SWANTON, C. 2015. SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair. *Oncogene*.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *J Am Stat Assoc,* 53**,** 457–481.

KAPUR, P., PENA-LLOPIS, S., CHRISTIE, A., ZHREBKER, L., PAVIA-JIMENEZ, A., RATHMELL, W. K., XIE, X. J. & BRUGAROLAS, J. 2013. Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation. *Lancet Oncol,* 14**,** 159-67.

KEEFE, S. M., NATHANSON, K. L. & RATHMELL, W. K. 2013. The molecular biology of renal cell carcinoma. *Semin Oncol,* 40**,** 421-8.

KESHAVA PRASAD, T. S., GOEL, R., KANDASAMY, K., KEERTHIKUMAR, S., KUMAR, S., MATHIVANAN, S., TELIKICHERLA, D., RAJU, R., SHAFREEN, B., VENUGOPAL, A., BALAKRISHNAN, L., MARIMUTHU, A., BANERJEE, S., SOMANATHAN, D. S., SEBASTIAN, A., RANI, S., RAY, S., HARRYS KISHORE, C. J., KANTH, S., AHMED, M., KASHYAP, M. K., MOHMOOD, R., RAMACHANDRA, Y. L., KRISHNA, V., RAHIMAN, B. A., MOHAN, S., RANGANATHAN, P., RAMABADRAN, S., CHAERKADY, R. & PANDEY, A. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res,* 37**,** D767-72.

KIBEL, A., ILIOPOULOS, O., DECAPRIO, J. A. & KAELIN, W. G., JR. 1995. Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. *Science,* 269**,** 1444-6.

KIM, J. H., JUNG, C. W., CHO, Y. H., LEE, J., LEE, S. H., KIM, H. Y., PARK, J., PARK, J. O., KIM, K., KIM, W. S., PARK, Y. S., IM, Y. H., KANG, W. K. & PARK, K. 2005. Somatic VHL alteration and its impact on prognosis in patients with clear cell renal cell carcinoma. *Oncol Rep,* 13**,** 859-64.

KIM, J. W., TCHERNYSHYOV, I., SEMENZA, G. L. & DANG, C. V. 2006. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab,* 3**,** 177-85.

KLATTE, T., KROEGER, N., RAMPERSAUD, E. N., BIRKHAUSER, F. D., LOGAN, J. E., SONN, G., RISS, J., RAO, P. N., KABBINAVAR, F. F., BELLDEGRUN, A. S. & PANTUCK, A. J. 2012. Gain of chromosome 8q is associated with metastases and poor survival of patients with clear cell renal cell carcinoma. *Cancer,* 118**,** 5777-82.

KLATTE, T., RAO, P. N., DE MARTINO, M., LAROCHELLE, J., SHUCH, B., ZOMORODIAN, N., SAID, J., KABBINAVAR, F. F., BELLDEGRUN, A. S. &

PANTUCK, A. J. 2009. Cytogenetic profile predicts prognosis of patients with clear cell renal cell carcinoma. *J Clin Oncol,* 27**,** 746-53.

KOSARI, F., PARKER, A. S., KUBE, D. M., LOHSE, C. M., LEIBOVICH, B. C., BLUTE, M. L., CHEVILLE, J. C. & VASMATZIS, G. 2005. Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness. *Clin Cancer Res,* 11**,** 5128-39.

KOTERA, M., HIRAKAWA, M., TOKIMATSU, T., GOTO, S. & KANEHISA, M. 2012. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol Biol,* 802**,** 19-39.

KOVACS, G., AKHTAR, M., BECKWITH, B. J., BUGERT, P., COOPER, C. S., DELAHUNT, B., EBLE, J. N., FLEMING, S., LJUNGBERG, B., MEDEIROS, L. J., MOCH, H., REUTER, V. E., RITZ, E., ROOS, G., SCHMIDT, D., SRIGLEY, J. R., STORKEL, S., VAN DEN BERG, E. & ZBAR, B. 1997. The Heidelberg classification of renal cell tumours. *J Pathol,* 183**,** 131-3.

KREBS, M. G., SLOANE, R., PRIEST, L., LANCASHIRE, L., HOU, J. M., GREYSTOKE, A., WARD, T. H., FERRALDESCHI, R., HUGHES, A., CLACK, G., RANSON, M., DIVE, C. & BLACKHALL, F. H. 2011. Evaluation and prognostic significance of circulating tumor cells in patients with non-small-cell lung cancer. *J Clin Oncol,* 29**,** 1556-63.

KROEGER, N., KLATTE, T., CHAMIE, K., RAO, P. N., BIRKHAUSER, F. D., SONN, G. A., RISS, J., KABBINAVAR, F. F., BELLDEGRUN, A. S. & PANTUCK, A. J. 2013. Deletions of chromosomes 3p and 14q molecularly subclassify clear cell renal cell carcinoma. *Cancer,* 119**,** 1547-54.

KROEGER, N., SELIGSON, D. B., SIGNORETTI, S., YU, H., MAGYAR, C. E., HUANG, J., BELLDEGRUN, A. S. & PANTUCK, A. J. 2014. Poor prognosis and advanced clinicopathological features of clear cell renal cell carcinoma (ccRCC) are associated with cytoplasmic subcellular localisation of Hypoxia inducible factor-2alpha. *Eur J Cancer,* 50**,** 1531-40.

KROEGER, N., ZIMMERMANN, U., BURCHARDT, M. & PANTUCK, A. J. 2015. Prognostication in localised renal cell carcinoma. *Lancet Oncol,* 16**,** 603-4.

LA ROCHELLE, J., KLATTE, T., DASTANE, A., RAO, N., SELIGSON, D., SAID, J., SHUCH, B., ZOMORODIAN, N., KABBINAVAR, F., BELLDEGRUN, A. & PANTUCK, A. J. 2010. Chromosome 9p deletions identify an aggressive phenotype of clear cell renal cell carcinoma. *Cancer,* 116**,** 4696-702.

LANE, B. R. & KATTAN, M. W. 2008. Prognostic models and algorithms in renal cell carcinoma. *Urol Clin North Am,* 35**,** 613-25; vii.

LANE, B. R., LI, J., ZHOU, M., BABINEAU, D., FABER, P., NOVICK, A. C. & WILLIAMS, B. R. 2009. Differential expression in clear cell renal cell carcinoma identified by gene expression profiling. *J Urol,* 181**,** 849-60.

LATIF, F., TORY, K., GNARRA, J., YAO, M., DUH, F. M., ORCUTT, M. L., STACKHOUSE, T., KUZMIN, I., MODI, W., GEIL, L. & ET AL. 1993. Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science,* 260**,** 1317-20.

LAWRENCE, M. S., STOJANOV, P., POLAK, P., KRYUKOV, G. V., CIBULSKIS, K., SIVACHENKO, A., CARTER, S. L., STEWART, C., MERMEL, C. H., ROBERTS, S. A., KIEZUN, A., HAMMERMAN, P. S., MCKENNA, A., DRIER, Y., ZOU, L., RAMOS, A. H., PUGH, T. J., STRANSKY, N., HELMAN, E., KIM, J., SOUGNEZ, C., AMBROGIO, L., NICKERSON, E., SHEFLER, E., CORTES, M. L., AUCLAIR, D., SAKSENA, G., VOET, D., NOBLE, M., DICARA, D., LIN, P., LICHTENSTEIN, L., HEIMAN, D. I., FENNELL, T., IMIELINSKI, M., HERNANDEZ, B., HODIS, E., BACA, S., DULAK, A. M., LOHR, J., LANDAU, D. A., WU, C. J., MELENDEZ-ZAJGLA, J., HIDALGO-MIRANDA, A., KOREN, A., MCCARROLL, S. A., MORA, J., LEE, R. S., CROMPTON, B., ONOFRIO, R., PARKIN, M., WINCKLER, W., ARDLIE, K., GABRIEL, S. B., ROBERTS, C. W., BIEGEL, J. A., STEGMAIER, K., BASS, A. J., GARRAWAY, L. A., MEYERSON, M., GOLUB, T. R., GORDENIN, D. A., SUNYAEV, S., LANDER, E. S. & GETZ, G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-8.

LEARY, R. J., SAUSEN, M., KINDE, I., PAPADOPOULOS, N., CARPTEN, J. D., CRAIG, D., O'SHAUGHNESSY, J., KINZLER, K. W., PARMIGIANI, G., VOGELSTEIN, B., DIAZ, L. A., JR. & VELCULESCU, V. E. 2012. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med,* 4**,** 162ra154.

LEE, A. J., ENDESFELDER, D., ROWAN, A. J., WALTHER, A., BIRKBAK, N. J., FUTREAL, P. A., DOWNWARD, J., SZALLASI, Z., TOMLINSON, I. P., HOWELL, M., KSCHISCHO, M. & SWANTON, C. 2011. Chromosomal instability confers intrinsic multidrug resistance. *Cancer Res,* 71**,** 1858-70.

LEE, C. M., HICKEY, M. M., SANFORD, C. A., MCGUIRE, C. G., COWEY, C. L., SIMON, M. C. & RATHMELL, W. K. 2009. VHL Type 2B gene mutation moderates HIF dosage in vitro and in vivo. *Oncogene,* 28**,** 1694-705.

LEFEBVRE, C., RAJBHANDARI, P., ALVAREZ, M. J., BANDARU, P., LIM, W. K., SATO, M., WANG, K., SUMAZIN, P., KUSTAGI, M., BISIKIRSKA, B. C., BASSO, K., BELTRAO, P., KROGAN, N., GAUTIER, J., DALLA-FAVERA, R. & CALIFANO, A. 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol,* 6**,** 377.

LEIBOVICH, B. C., BLUTE, M. L., CHEVILLE, J. C., LOHSE, C. M., FRANK, I., KWON, E. D., WEAVER, A. L., PARKER, A. S. & ZINCKE, H. 2003. Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: a stratification tool for prospective clinical trials. *Cancer,* 97**,** 1663-71.

LEISERSON, M. D., BLOKH, D., SHARAN, R. & RAPHAEL, B. J. 2013. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol,* 9**,** e1003054.

LEISERSON, M. D., VANDIN, F., WU, H. T., DOBSON, J. R., ELDRIDGE, J. V., THOMAS, J. L., PAPOUTSAKI, A., KIM, Y., NIU, B., MCLELLAN, M., LAWRENCE, M. S., GONZALEZ-PEREZ, A., TAMBORERO, D., CHENG, Y., RYSLIK, G. A., LOPEZ-BIGAS, N., GETZ, G., DING, L. & RAPHAEL, B. J. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet,* 47**,** 106-14.

LIBERZON, A. 2014. A description of the Molecular Signatures Database (MSigDB) Web site. *Methods Mol Biol,* 1150**,** 153-60.

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDOTTIR, H., TAMAYO, P. & MESIROV, J. P. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics,* 27**,** 1739-40.

LICATA, L., BRIGANTI, L., PELUSO, D., PERFETTO, L., IANNUCCELLI, M., GALEOTA, E., SACCO, F., PALMA, A., NARDOZZA, A. P., SANTONICO, E., CASTAGNOLI, L. & CESARENI, G. 2012. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res,* 40**,** D857-61.

LINEHAN, W. M., SRINIVASAN, R. & SCHMIDT, L. S. 2010. The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol,* 7**,** 277-85.

MA, X. M. & BLENIS, J. 2009. Molecular mechanisms of mTOR-mediated translational control. *Nat Rev Mol Cell Biol,* 10**,** 307-18.

MARGOLIN, A. A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., DALLA FAVERA, R. & CALIFANO, A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* 7 Suppl 1**,** S7.

MARTINEZ, P., BIRKBAK, N. J., GERLINGER, M., MCGRANAHAN, N., BURRELL, R. A., ROWAN, A. J., JOSHI, T., FISHER, R., LARKIN, J., SZALLASI, Z. & SWANTON, C. 2013. Parallel evolution of tumour subclones mimics diversity between tumours. *J Pathol,* 230**,** 356-64.

MATEO, J., GERLINGER, M., RODRIGUES, D. N. & DE BONO, J. S. 2014. The promise of circulating tumor cell analysis in cancer management. *Genome Biol,* 15**,** 448.

MATSUOKA, S., BALLIF, B. A., SMOGORZEWSKA, A., MCDONALD, E. R., 3RD, HUROV, K. E., LUO, J., BAKALARSKI, C. E., ZHAO, Z., SOLIMINI, N., LERENTHAL, Y., SHILOH, Y., GYGI, S. P. & ELLEDGE, S. J. 2007. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science,* 316**,** 1160-6.

MCDERMOTT, D. F., REGAN, M. M., CLARK, J. I., FLAHERTY, L. E., WEISS, G. R., LOGAN, T. F., KIRKWOOD, J. M., GORDON, M. S., SOSMAN, J. A., ERNSTOFF, M. S., TRETTER, C. P., URBA, W. J., SMITH, J. W., MARGOLIN, K. A., MIER, J. W., GOLLOB, J. A., DUTCHER, J. P. & ATKINS, M. B. 2005. Randomized phase III trial of high-dose interleukin-2 versus subcutaneous interleukin-2 and interferon in patients with metastatic renal cell carcinoma. *J Clin Oncol,* 23**,** 133-41.

MCGRANAHAN, N., BURRELL, R. A., ENDESFELDER, D., NOVELLI, M. R. & SWANTON, C. 2012. Cancer chromosomal instability: therapeutic and diagnostic challenges. *EMBO Rep,* 13**,** 528-38.

MCGRANAHAN, N. & SWANTON, C. 2015. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell,* 27**,** 15-26.

MERMEL, C. H., SCHUMACHER, S. E., HILL, B., MEYERSON, M. L., BEROUKHIM, R. & GETZ, G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol,* 12**,** R41.

MEYER, L. R., ZWEIG, A. S., HINRICHS, A. S., KAROLCHIK, D., KUHN, R. M., WONG, M., SLOAN, C. A., ROSENBLOOM, K. R., ROE, G., RHEAD, B., RANEY, B. J., POHL, A., MALLADI, V. S., LI, C. H., LEE, B. T., LEARNED, K., KIRKUP, V., HSU, F., HEITNER, S., HARTE, R. A., HAEUSSLER, M., GURUVADOO, L., GOLDMAN, M., GIARDINE, B. M., FUJITA, P. A., DRESZER, T. R., DIEKHANS, M., CLINE, M. S., CLAWSON, H., BARBER, G. P., HAUSSLER, D. & KENT, W. J. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res,* 41**,** D64-9.

MILLER, C. A., SETTLE, S. H., SULMAN, E. P., ALDAPE, K. D. & MILOSAVLJEVIC, A. 2011. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics,* 4**,** 34.

MOCH, H., PRESTI, J. C., JR., SAUTER, G., BUCHHOLZ, N., JORDAN, P., MIHATSCH, M. J. & WALDMAN, F. M. 1996. Genetic aberrations detected by comparative genomic hybridization are associated with clinical outcome in renal cell carcinoma. *Cancer Res,* 56**,** 27-30.

MONZON, F. A., ALVAREZ, K., PETERSON, L., TRUONG, L., AMATO, R. J., HERNANDEZ-MCCLAIN, J., TANNIR, N., PARWANI, A. V. & JONASCH, E. 2011. Chromosome 14q loss defines a molecular subtype of clear-cell renal cell carcinoma associated with poor prognosis. *Mod Pathol,* 24**,** 1470-9.

MOOTHA, V. K., LINDGREN, C. M., ERIKSSON, K. F., SUBRAMANIAN, A., SIHAG, S., LEHAR, J., PUIGSERVER, P., CARLSSON, E., RIDDERSTRALE, M., LAURILA, E., HOUSTIS, N., DALY, M. J., PATTERSON, N., MESIROV, J. P., GOLUB, T. R., TAMAYO, P., SPIEGELMAN, B., LANDER, E. S., HIRSCHHORN, J. N., ALTSHULER, D. & GROOP, L. C. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet,* 34**,** 267-73.

MORRIS, M. R., RICKETTS, C. J., GENTLE, D., MCRONALD, F., CARLI, N., KHALILI, H., BROWN, M., KISHIDA, T., YAO, M., BANKS, R. E., CLARKE, N., LATIF, F. & MAHER, E. R. 2011. Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. *Oncogene,* 30**,** 1390-401.

MOTZER, R. J., MAZUMDAR, M., BACIK, J., BERG, W., AMSTERDAM, A. & FERRARA, J. 1999. Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *J Clin Oncol,* 17**,** 2530-40.

MURTAZA, M., DAWSON, S. J., TSUI, D. W., GALE, D., FORSHEW, T., PISKORZ, A. M., PARKINSON, C., CHIN, S. F., KINGSBURY, Z., WONG, A. S., MARASS, F., HUMPHRAY, S., HADFIELD, J., BENTLEY, D., CHIN, T. M., BRENTON, J. D., CALDAS, C. & ROSENFELD, N. 2013. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature,* 497**,** 108-12.

NAIDOO, J., PAGE, D. B. & WOLCHOK, J. D. 2014. Immune modulation for cancer therapy. *Br J Cancer,* 111**,** 2214-9.

NAVIN, N., KENDALL, J., TROGE, J., ANDREWS, P., RODGERS, L., MCINDOO, J., COOK, K., STEPANSKY, A., LEVY, D., ESPOSITO, D., MUTHUSWAMY, L., KRASNITZ, A., MCCOMBIE, W. R., HICKS, J. & WIGLER, M. 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

NAVIN, N. E. & HICKS, J. 2010. Tracing the tumor lineage. *Mol Oncol,* 4**,** 267-83.

NICHOLS, A. C., LOWES, L. E., SZETO, C. C., BASMAJI, J., DHALIWAL, S., CHAPESKIE, C., TODOROVIC, B., READ, N., VENKATESAN, V., HAMMOND, A., PALMA, D. A., WINQUIST, E., ERNST, S., FUNG, K., FRANKLIN, J. H., YOO, J., KOROPATNICK, J., MYMRYK, J. S., BARRETT, J. W. & ALLAN, A. L. 2012. Detection of circulating tumor cells in advanced head and neck cancer using the CellSearch system. *Head Neck,* 34**,** 1440-4.

NICKERSON, M. L., JAEGER, E., SHI, Y., DUROCHER, J. A., MAHURKAR, S., ZARIDZE, D., MATVEEV, V., JANOUT, V., KOLLAROVA, H., BENCKO, V., NAVRATILOVA, M., SZESZENIA-DABROWSKA, N., MATES, D., MUKERIA, A., HOLCATOVA, I., SCHMIDT, L. S., TORO, J. R., KARAMI, S., HUNG, R., GERARD, G. F., LINEHAN, W. M., MERINO, M., ZBAR, B., BOFFETTA, P., BRENNAN, P., ROTHMAN, N., CHOW, W. H., WALDMAN, F. M. & MOORE, L. E. 2008. Improved identification of von Hippel-Lindau gene alterations in clear cell renal tumors. *Clin Cancer Res,* 14**,** 4726-34.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M., SHLIEN, A., COOKE, S. L., HINTON, J., MENZIES, A., STEBBINGS, L. A., LEROY, C., JIA, M., RANCE, R., MUDIE, L. J., GAMBLE, S. J., STEPHENS, P. J., MCLAREN, S., TARPEY, P. S., PAPAEMMANUIL, E., DAVIES, H. R., VARELA, I., MCBRIDE, D. J., BIGNELL, G. R., LEUNG, K., BUTLER, A. P., TEAGUE, J. W., MARTIN, S., JONSSON, G., MARIANI, O., BOYAULT, S., MIRON, P., FATIMA, A., LANGEROD, A., APARICIO, S. A., TUTT, A., SIEUWERTS, A. M., BORG, A., THOMAS, G., SALOMON, A. V., RICHARDSON, A. L., BORRESEN-DALE, A. L., FUTREAL, P. A., STRATTON, M. R., CAMPBELL, P. J. & BREAST CANCER WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C. 2012. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

NIU, X., ZHANG, T., LIAO, L., ZHOU, L., LINDNER, D. J., ZHOU, M., RINI, B., YAN, Q. & YANG, H. 2012. The von Hippel-Lindau tumor suppressor protein regulates gene expression and tumor growth through histone demethylase JARID1C. *Oncogene,* 31**,** 776-86.

OOSTERWIJK, E., RATHMELL, W. K., JUNKER, K., BRANNON, A. R., POULIOT, F., FINLEY, D. S., MULDERS, P. F., KIRKALI, Z., UEMURA, H. & BELLDEGRUN, A. 2011. Basic research in kidney cancer. *Eur Urol,* 60**,** 622-33.

ORCHARD, S., AMMARI, M., ARANDA, B., BREUZA, L., BRIGANTI, L., BROACKES-CARTER, F., CAMPBELL, N. H., CHAVALI, G., CHEN, C., DEL-TORO, N., DUESBURY, M., DUMOUSSEAU, M., GALEOTA, E., HINZ, U., IANNUCCELLI, M., JAGANNATHAN, S., JIMENEZ, R., KHADAKE, J., LAGREID, A., LICATA, L., LOVERING, R. C., MELDAL, B., MELIDONI, A. N., MILAGROS, M., PELUSO, D., PERFETTO, L., PORRAS, P., RAGHUNATH, A., RICARD-BLUM, S., ROECHERT, B., STUTZ, A., TOGNOLLI, M., VAN ROEY, K., CESARENI, G. & HERMJAKOB, H. 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res,* 42**,** D358-63.

ORTIZ-ESTEVEZ, M., ARAMBURU, A., BENGTSSON, H., NEUVIAL, P. & RUBIO, A. 2012. CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation. *Bioinformatics,* 28**,** 1793-4.

OSTHUS, R. C., SHIM, H., KIM, S., LI, Q., REDDY, R., MUKHERJEE, M., XU, Y., WONSEY, D., LEE, L. A. & DANG, C. V. 2000. Deregulation of glucose transporter 1 and glycolytic gene expression by c-Myc. *J Biol Chem,* 275**,** 21797-800.

PAUSE, A., LEE, S., WORRELL, R. A., CHEN, D. Y., BURGESS, W. H., LINEHAN, W. M. & KLAUSNER, R. D. 1997. The von Hippel-Lindau tumor-suppressor gene product forms a stable complex with human CUL-2, a member of the Cdc53 family of proteins. *Proc Natl Acad Sci U S A,* 94**,** 2156-61.

PAWLOWSKI, R., MUHL, S. M., SULSER, T., KREK, W., MOCH, H. & SCHRAML, P. 2013. Loss of PBRM1 expression is associated with renal cell carcinoma progression. *Int J Cancer,* 132**,** E11-7.

PENA-LLOPIS, S., VEGA-RUBIN-DE-CELIS, S., LIAO, A., LENG, N., PAVIA-JIMENEZ, A., WANG, S., YAMASAKI, T., ZHREBKER, L., SIVANAND, S., SPENCE, P., KINCH, L., HAMBUCH, T., JAIN, S., LOTAN, Y., MARGULIS, V., SAGALOWSKY, A. I., SUMMEROUR, P. B., KABBANI, W., WONG, S. W., GRISHIN, N., LAURENT, M., XIE, X. J., HAUDENSCHILD, C. D., ROSS, M. T., BENTLEY, D. R., KAPUR, P. & BRUGAROLAS, J. 2012. BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet,* 44**,** 751-9.

PEREZ-GRACIA, J. L., PRIOR, C., GUILLEN-GRIMA, F., SEGURA, V., GONZALEZ, A., PANIZO, A., MELERO, I., GRANDE-PULIDO, E., GURPIDE, A., GIL-BAZO, I. & CALVO, A. 2009. Identification of TNF-alpha and MMP-9 as potential baseline predictive serum markers of sunitinib activity in patients with renal cell carcinoma using a human cytokine array. *Br J Cancer,* 101**,** 1876-83.

PETILLO, D., KORT, E. J., ANEMA, J., FURGE, K. A., YANG, X. J. & TEH, B. T. 2009. MicroRNA profiling of human kidney cancer subtypes. *Int J Oncol,* 35**,** 109-14.

PUTTER, H., FIOCCO, M. & GESKUS, R. B. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med,* 26**,** 2389-430.

R DEVELOPMENT CORE TEAM 2013. R: A Language and Environment for Statistical Computing.

RACK, B., SCHINDLBECK, C., JUCKSTOCK, J., ANDERGASSEN, U., HEPP, P., ZWINGERS, T., FRIEDL, T. W., LORENZ, R., TESCH, H., FASCHING, P. A., FEHM, T., SCHNEEWEISS, A., LICHTENEGGER, W., BECKMANN, M. W., FRIESE, K., PANTEL, K., JANNI, W. & GROUP, S. S. 2014. Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J Natl Cancer Inst,* 106.

RAGAZZON, B., LIBE, R., GAUJOUX, S., ASSIE, G., FRATTICCI, A., LAUNAY, P., CLAUSER, E., BERTAGNA, X., TISSIER, F., DE REYNIES, A. & BERTHERAT, J. 2010. Transcriptome analysis reveals that p53 and {beta}-catenin alterations occur in a group of aggressive adrenocortical cancers. *Cancer Res,* 70**,** 8276-81.

RAO, C., BUI, T., CONNELLY, M., DOYLE, G., KARYDIS, I., MIDDLETON, M. R., CLACK, G., MALONE, M., COUMANS, F. A. & TERSTAPPEN, L. W. 2011. Circulating melanoma cells and survival in metastatic melanoma. *Int J Oncol,* 38**,** 755-60.

RATHMELL, W. K., HICKEY, M. M., BEZMAN, N. A., CHMIELECKI, C. A., CARRAWAY, N. C. & SIMON, M. C. 2004. In vitro and in vivo models analyzing von Hippel-Lindau disease-specific mutations. *Cancer Res,* 64**,** 8595-603.

REICH, M., LIEFELD, T., GOULD, J., LERNER, J., TAMAYO, P. & MESIROV, J. P. 2006. GenePattern 2.0. *Nat Genet,* 38**,** 500-1.

REIMAND, J., TOOMING, L., PETERSON, H., ADLER, P. & VILO, J. 2008. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res,* 36**,** W452-9.

REISMAN, D., GLAROS, S. & THOMPSON, E. A. 2009. The SWI/SNF complex and cancer. *Oncogene,* 28**,** 1653-68.

RIGAUT, G., SHEVCHENKO, A., RUTZ, B., WILM, M., MANN, M. & SERAPHIN, B. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol,* 17**,** 1030-2.

RINI, B., GODDARD, A., KNEZEVIC, D., MADDALA, T., ZHOU, M., AYDIN, H., CAMPBELL, S., ELSON, P., KOSCIELNY, S., LOPATIN, M., SVEDMAN, C., MARTINI, J. F., WILLIAMS, J. A., VERKARRE, V., RADULESCU, C., NEUZILLET, Y., HEMMERLE, I., TIMSIT, M. O., TSIATIS, A. C., BONHAM, M., LEBRET, T., MEJEAN, A. & ESCUDIER, B. 2015. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. *Lancet Oncol*.

RINI, B. I. & ATKINS, M. B. 2009. Resistance to targeted therapy in renal-cell carcinoma. *Lancet Oncol,* 10**,** 992-1000.

RINI, B. I., MICHAELSON, M. D., ROSENBERG, J. E., BUKOWSKI, R. M., SOSMAN, J. A., STADLER, W. M., HUTSON, T. E., MARGOLIN, K., HARMON, C. S., DEPRIMO, S. E., KIM, S. T., CHEN, I. & GEORGE, D. J. 2008. Antitumor activity and biomarker analysis of sunitinib in patients with bevacizumab-refractory metastatic renal cell carcinoma. *J Clin Oncol,* 26**,** 3743-8.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

RYDZANICZ, M., WRZESINSKI, T., BLUYSSEN, H. A. & WESOLY, J. 2013. Genomics and epigenomics of clear cell renal cell carcinoma: recent developments and potential applications. *Cancer Lett,* 341**,** 111-26.

SAMAAN, S., KHELLA, H. W., GIRGIS, A., SCORILAS, A., LIANIDOU, E., GABRIL, M., KRYLOV, S. N., JEWETT, M., BJARNASON, G. A., EL-SAID, H. & YOUSEF, G. M. 2015. miR-210 is a prognostic marker in clear cell renal cell carcinoma. *J Mol Diagn,* 17**,** 136-44.

SANCAK, Y., PETERSON, T. R., SHAUL, Y. D., LINDQUIST, R. A., THOREEN, C. C., BAR-PELED, L. & SABATINI, D. M. 2008. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science,* 320**,** 1496-501.

SANJMYATAV, J., JUNKER, K., MATTHES, S., MUEHR, M., SAVA, D., STERNAL, M., WESSENDORF, S., KREUZ, M., GAJDA, M., WUNDERLICH, H. & SCHWAENEN, C. 2011. Identification of genomic alterations associated with metastasis and cancer specific survival in clear cell renal cell carcinoma. *J Urol,* 186**,** 2078-83.

SATAGOPAN, J. M., BEN-PORAT, L., BERWICK, M., ROBSON, M., KUTLER, D. & AUERBACH, A. D. 2004. A note on competing risks in survival data analysis. *Br J Cancer,* 91**,** 1229-35.

SATO, Y., YOSHIZATO, T., SHIRAISHI, Y., MAEKAWA, S., OKUNO, Y., KAMURA, T., SHIMAMURA, T., SATO-OTSUBO, A., NAGAE, G., SUZUKI, H., NAGATA, Y., YOSHIDA, K., KON, A., SUZUKI, Y., CHIBA, K., TANAKA, H., NIIDA, A., FUJIMOTO, A., TSUNODA, T., MORIKAWA, T., MAEDA, D., KUME, H., SUGANO, S., FUKAYAMA, M., ABURATANI, H., SANADA, M., MIYANO, S., HOMMA, Y. & OGAWA, S. 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet,* 45**,** 860-7.

SCELO, G., RIAZALHOSSEINI, Y., GREGER, L., LETOURNEAU, L., GONZALEZ-PORTA, M., WOZNIAK, M. B., BOURGEY, M., HARNDEN, P., EGEVAD, L., JACKSON, S. M., KARIMZADEH, M., ARSENEAULT, M., LEPAGE, P., HOW-KIT, A., DAUNAY, A., RENAULT, V., BLANCHE, H., TUBACHER, E., SEHMOUN, J., VIKSNA, J., CELMS, E., OPMANIS, M., ZARINS, A., VASUDEV, N. S., SEYWRIGHT, M., ABEDI-ARDEKANI, B., CARREIRA, C., SELBY, P. J., CARTLEDGE, J. J., BYRNES, G., ZAVADIL, J., SU, J., HOLCATOVA, I., BRISUDA, A., ZARIDZE, D., MOUKERIA, A., FORETOVA, L., NAVRATILOVA, M., MATES, D., JINGA, V., ARTEMOV, A., NEDOLUZHKO, A., MAZUR, A., RASTORGUEV, S., BOULYGINA, E., HEATH, S., GUT, M., BIHOREAU, M. T., LECHNER, D., FOGLIO, M., GUT, I. G., SKRYABIN, K., PROKHORTCHOUK, E., CAMBON-THOMSEN, A., RUNG, J., BOURQUE, G., BRENNAN, P., TOST, J., BANKS, R. E., BRAZMA, A. & LATHROP, G. M. 2014. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun,* 5**,** 5135.

SCHRAML, P., STRUCKMANN, K., HATZ, F., SONNET, S., KULLY, C., GASSER, T., SAUTER, G., MIHATSCH, M. J. & MOCH, H. 2002. VHL mutations and their correlation with tumour cell proliferation, microvessel density, and patient prognosis in clear cell renal cell carcinoma. *J Pathol,* 196**,** 186-93.

SEMENZA, G. L. 2007. HIF-1 mediates the Warburg effect in clear cell renal carcinoma. *J Bioenerg Biomembr,* 39**,** 231-4.

SHAH, S. P., ROTH, A., GOYA, R., OLOUMI, A., HA, G., ZHAO, Y., TURASHVILI, G., DING, J., TSE, K., HAFFARI, G., BASHASHATI, A., PRENTICE, L. M., KHATTRA, J., BURLEIGH, A., YAP, D., BERNARD, V., MCPHERSON, A., SHUMANSKY, K., CRISAN, A., GIULIANY, R., HERAVI-MOUSSAVI, A., ROSNER, J., LAI, D., BIROL, I., VARHOL, R., TAM, A., DHALLA, N., ZENG, T., MA, K., CHAN, S. K., GRIFFITH, M., MORADIAN, A., CHENG, S. W., MORIN, G. B., WATSON, P., GELMON, K., CHIA, S., CHIN, S. F., CURTIS, C., RUEDA, O. M., PHAROAH, P. D., DAMARAJU, S., MACKEY, J., HOON, K., HARKINS, T., TADIGOTLA, V., SIGAROUDINIA, M., GASCARD, P., TLSTY, T., COSTELLO, J. F., MEYER, I. M., EAVES, C. J., WASSERMAN, W. W., JONES, S., HUNTSMAN, D., HIRST, M., CALDAS, C., MARRA, M. A. & APARICIO, S. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature,* 486**,** 395-9.

SHAW, J. A., PAGE, K., BLIGHE, K., HAVA, N., GUTTERY, D., WARD, B., BROWN, J., RUANGPRATHEEP, C., STEBBING, J., PAYNE, R., PALMIERI, C., CLEATOR, S., WALKER, R. A. & COOMBES, R. C. 2012. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. *Genome Res,* 22**,** 220-31.

SHOEMAKER, B. A. & PANCHENKO, A. R. 2007. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol,* 3**,** e42.

SHUIN, T., KONDO, K., TORIGOE, S., KISHIDA, T., KUBOTA, Y., HOSAKA, M., NAGASHIMA, Y., KITAMURA, H., LATIF, F., ZBAR, B. & ET AL. 1994. Frequent somatic mutations and loss of heterozygosity of the von Hippel-Lindau tumor suppressor gene in primary human renal cell carcinomas. *Cancer Res,* 54**,** 2852-5.

SINGH, A. P., MONIAUX, N., CHAUHAN, S. C., MEZA, J. L. & BATRA, S. K. 2004. Inhibition of MUC4 expression suppresses pancreatic tumor cell growth and metastasis. *Cancer Res,* 64**,** 622-30.

SORBELLINI, M., KATTAN, M. W., SNYDER, M. E., REUTER, V., MOTZER, R., GOETZL, M., MCKIERNAN, J. & RUSSO, P. 2005. A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. *J Urol,* 173**,** 48-51.

SOTTORIVA, A., SPITERI, I., PICCIRILLO, S. G., TOULOUMIS, A., COLLINS, V. P., MARIONI, J. C., CURTIS, C., WATTS, C. & TAVARE, S. 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A,* 110**,** 4009-14.

SOULTATI, A., STARES, M., SWANTON, C., LARKIN, J. & TURAJLIC, S. 2015. How should clinicians address intratumour heterogeneity in clear cell renal cell carcinoma? *Curr Opin Urol*.

STRATTON, M. R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science,* 331**,** 1553-8.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SULTMANN, H., VON HEYDEBRECK, A., HUBER, W., KUNER, R., BUNESS, A., VOGT, M., GUNAWAN, B., VINGRON, M., FUZESI, L. & POUSTKA, A. 2005. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. *Clin Cancer Res,* 11**,** 646-55.

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P., KUHN, M., BORK, P., JENSEN, L. J. & VON MERING, C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res,* 43**,** D447-52.

TAKAHASHI, M., RHODES, D. R., FURGE, K. A., KANAYAMA, H., KAGAWA, S., HAAB, B. B. & TEH, B. T. 2001. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. *Proc Natl Acad Sci U S A,* 98**,** 9754-9.

TANABE, M. & KANEHISA, M. 2012. Using the KEGG database resource. *Curr Protoc Bioinformatics,* Chapter 1**,** Unit1 12.

TANAKA, F., YONEDA, K., KONDO, N., HASHIMOTO, M., TAKUWA, T., MATSUMOTO, S., OKUMURA, Y., RAHMAN, S., TSUBOTA, N., TSUJIMURA, T., KURIBAYASHI, K., FUKUOKA, K., NAKANO, T. & HASEGAWA, S. 2009. Circulating tumor cell as a diagnostic marker in primary lung cancer. *Clin Cancer Res,* 15**,** 6980-6.

TANG, P. A., VICKERS, M. M. & HENG, D. Y. 2011. Clinical and molecular prognostic factors in renal cell carcinoma: what we know so far. *Hematol Oncol Clin North Am,* 25**,** 871-91.

TAYLOR, I. W., LINDING, R., WARDE-FARLEY, D., LIU, Y., PESQUITA, C., FARIA, D., BULL, S., PAWSON, T., MORRIS, Q. & WRANA, J. L. 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol,* 27**,** 199-204.

TESCHENDORFF, A. E. & SEVERINI, S. 2010. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst Biol,* 4**,** 104.

THE CANCER GENOME ATLAS RESEARCH NETWORK 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature,* 499**,** 43-9.

THERNEAU, T. M. 2014. A Package for Survival Analysis in S.

THERNEAU, T. M. & GRAMBSCH, P. M. 2000. *Modeling Survival Data: Extending the {C}ox Model*, Springer.

THIRLWELL, C., WILL, O. C., DOMINGO, E., GRAHAM, T. A., MCDONALD, S. A., OUKRIF, D., JEFFREY, R., GORMAN, M., RODRIGUEZ-JUSTO, M., CHIN-ALEONG, J., CLARK, S. K., NOVELLI, M. R., JANKOWSKI, J. A., WRIGHT, N. A., TOMLINSON, I. P. & LEEDHAM, S. J. 2010. Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology,* 138**,** 1441-54, 1454 e1-7.

THOENES, W., STORKEL, S. & RUMPELT, H. J. 1986. Histopathology and classification of renal cell tumors (adenomas, oncocytomas and carcinomas). The basic cytological and histopathological elements and their use for diagnostics. *Pathol Res Pract,* 181**,** 125-43.

THOMAS, G. V., TRAN, C., MELLINGHOFF, I. K., WELSBIE, D. S., CHAN, E., FUEGER, B., CZERNIN, J. & SAWYERS, C. L. 2006. Hypoxia-inducible factor determines sensitivity to inhibitors of mTOR in kidney cancer. *Nat Med,* 12**,** 122-7.

TOMA, M. I., GROSSER, M., HERR, A., AUST, D. E., MEYE, A., HOEFLING, C., FUESSEL, S., WUTTIG, D., WIRTH, M. P. & BARETTON, G. B. 2008. Loss of heterozygosity and copy number abnormality in clear cell renal cell carcinoma discovered by high-density affymetrix 10K single nucleotide polymorphism mapping array. *Neoplasia,* 10**,** 634-42.

TOPALIAN, S. L., HODI, F. S., BRAHMER, J. R., GETTINGER, S. N., SMITH, D. C., MCDERMOTT, D. F., POWDERLY, J. D., CARVAJAL, R. D., SOSMAN, J. A., ATKINS, M. B., LEMING, P. D., SPIGEL, D. R., ANTONIA, S. J., HORN, L., DRAKE, C. G., PARDOLL, D. M., CHEN, L., SHARFMAN, W. H., ANDERS, R. A., TAUBE, J. M., MCMILLER, T. L., XU, H., KORMAN, A. J., JURE-KUNKEL, M., AGRAWAL, S., MCDONALD, D., KOLLIA, G. D., GUPTA, A., WIGGINTON, J. M. & SZNOL, M. 2012. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med,* 366**,** 2443-54.

TUCK, D. P., KLUGER, H. M. & KLUGER, Y. 2006. Characterizing disease states from topological properties of transcriptional regulatory networks. *BMC Bioinformatics,* 7**,** 236.

TUN, H. W., MARLOW, L. A., VON ROEMELING, C. A., COOPER, S. J., KREINEST, P., WU, K., LUXON, B. A., SINHA, M., ANASTASIADIS, P. Z. & COPLAND, J. A. 2010. Pathway signature and cellular differentiation in clear cell renal cell carcinoma. *PLoS One,* 5**,** e10696.

VAN DER VELDT, A. A., VROLING, L., DE HAAS, R. R., KOOLWIJK, P., VAN DEN EERTWEGH, A. J., HAANEN, J. B., VAN HINSBERGH, V. W., BROXTERMAN, H. J. & BOVEN, E. 2012. Sunitinib-induced changes in circulating endothelial cell-related proteins in patients with metastatic renal cell cancer. *Int J Cancer,* 131**,** E484-93.

VAN ERP, N. P., EECHOUTE, K., VAN DER VELDT, A. A., HAANEN, J. B., REYNERS, A. K., MATHIJSSEN, R. H., BOVEN, E., VAN DER STRAATEN, T., BAAK-PABLO, R. F., WESSELS, J. A., GUCHELAAR, H. J. & GELDERBLOM,

H. 2009. Pharmacogenetic pathway analysis for determination of sunitinib-induced toxicity. *J Clin Oncol,* 27**,** 4406-12.

VAN HAAFTEN, G., DALGLIESH, G. L., DAVIES, H., CHEN, L., BIGNELL, G., GREENMAN, C., EDKINS, S., HARDY, C., O'MEARA, S., TEAGUE, J., BUTLER, A., HINTON, J., LATIMER, C., ANDREWS, J., BARTHORPE, S., BEARE, D., BUCK, G., CAMPBELL, P. J., COLE, J., FORBES, S., JIA, M., JONES, D., KOK, C. Y., LEROY, C., LIN, M. L., MCBRIDE, D. J., MADDISON, M., MAQUIRE, S., MCLAY, K., MENZIES, A., MIRONENKO, T., MULDERRIG, L., MUDIE, L., PLEASANCE, E., SHEPHERD, R., SMITH, R., STEBBINGS, L., STEPHENS, P., TANG, G., TARPEY, P. S., TURNER, R., TURRELL, K., VARIAN, J., WEST, S., WIDAA, S., WRAY, P., COLLINS, V. P., ICHIMURA, K., LAW, S., WONG, J., YUEN, S. T., LEUNG, S. Y., TONON, G., DEPINHO, R. A., TAI, Y. T., ANDERSON, K. C., KAHNOSKI, R. J., MASSIE, A., KHOO, S. K., TEH, B. T., STRATTON, M. R. & FUTREAL, P. A. 2009. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet,* 41**,** 521-3.

VAN LOO, P., NORDGARD, S. H., LINGJAERDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C. M., BORRESEN-DALE, A. L. & KRISTENSEN, V. N. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

VANDIN, F., UPFAL, E. & RAPHAEL, B. J. 2012. De novo discovery of mutated driver pathways in cancer. *Genome Res,* 22**,** 375-85.

VARELA, I., TARPEY, P., RAINE, K., HUANG, D., ONG, C. K., STEPHENS, P., DAVIES, H., JONES, D., LIN, M. L., TEAGUE, J., BIGNELL, G., BUTLER, A., CHO, J., DALGLIESH, G. L., GALAPPATHTHIGE, D., GREENMAN, C., HARDY, C., JIA, M., LATIMER, C., LAU, K. W., MARSHALL, J., MCLAREN, S., MENZIES, A., MUDIE, L., STEBBINGS, L., LARGAESPADA, D. A., WESSELS, L. F., RICHARD, S., KAHNOSKI, R. J., ANEMA, J., TUVESON, D. A., PEREZ-MANCERA, P. A., MUSTONEN, V., FISCHER, A., ADAMS, D. J., RUST, A., CHAN-ON, W., SUBIMERB, C., DYKEMA, K., FURGE, K., CAMPBELL, P. J., TEH, B. T., STRATTON, M. R. & FUTREAL, P. A. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature,* 469**,** 539-42.

VASSELLI, J. R., SHIH, J. H., IYENGAR, S. R., MARANCHIE, J., RISS, J., WORRELL, R., TORRES-CABALA, C., TABIOS, R., MARIOTTI, A., STEARMAN, R., MERINO, M., WALTHER, M. M., SIMON, R., KLAUSNER, R. D. & LINEHAN, W. M. 2003. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc Natl Acad Sci U S A,* 100**,** 6958-63.

VRIES, R. G., BEZROOKOVE, V., ZUIJDERDUIJN, L. M., KIA, S. K., HOUWELING, A., ORUETXEBARRIA, I., RAAP, A. K. & VERRIJZER, C. P. 2005. Cancer-associated mutations in chromatin remodeler hSNF5 promote chromosomal instability by compromising the mitotic checkpoint. *Genes Dev,* 19**,** 665-70.

WANG, J. K., TSAI, M. C., POULIN, G., ADLER, A. S., CHEN, S., LIU, H., SHI, Y. & CHANG, H. Y. 2010. The histone demethylase UTX enables RB-dependent cell fate control. *Genes Dev,* 24**,** 327-32.

WOOD, L. D., PARSONS, D. W., JONES, S., LIN, J., SJOBLOM, T., LEARY, R. J., SHEN, D., BOCA, S. M., BARBER, T., PTAK, J., SILLIMAN, N., SZABO, S., DEZSO, Z., USTYANKSKY, V., NIKOLSKAYA, T., NIKOLSKY, Y., KARCHIN, R., WILSON, P. A., KAMINKER, J. S., ZHANG, Z., CROSHAW, R., WILLIS, J., DAWSON, D., SHIPITSIN, M., WILLSON, J. K., SUKUMAR, S., POLYAK, K., PARK, B. H., PETHIYAGODA, C. L., PANT, P. V., BALLINGER, D. G., SPARKS, A. B., HARTIGAN, J., SMITH, D. R., SUH, E., PAPADOPOULOS, N., BUCKHAULTS, P., MARKOWITZ, S. D., PARMIGIANI, G., KINZLER, K. W., VELCULESCU, V. E. & VOGELSTEIN, B. 2007. The genomic landscapes of human breast and colorectal cancers. *Science,* 318, 1108-13.

WU, G., FENG, X. & STEIN, L. 2010. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol,* 11, R53.

WU, G. & STEIN, L. 2012. A network module-based method for identifying cancer prognostic signatures. *Genome Biol,* 13, R112.

WUTTIG, D., ZASTROW, S., FUSSEL, S., TOMA, M. I., MEINHARDT, M., KALMAN, K., JUNKER, K., SANJMYATAV, J., BOLL, K., HACKERMULLER, J., ROLLE, A., GRIMM, M. O. & WIRTH, M. P. 2012. CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases. *Int J Cancer,* 131, E693-704.

XENARIOS, I., SALWINSKI, L., DUAN, X. J., HIGNEY, P., KIM, S. M. & EISENBERG, D. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res,* 30, 303-5.

XIA, J., SUN, J., JIA, P. & ZHAO, Z. 2011. Do cancer proteins really interact strongly in the human protein-protein interaction network? *Comput Biol Chem,* 35, 121-5.

XIA, W., NAGASE, S., MONTIA, A. G., KALACHIKOV, S. M., KENIRY, M., SU, T., MEMEO, L., HIBSHOOSH, H. & PARSONS, R. 2008. BAF180 is a critical regulator of p21 induction and a tumor suppressor mutated in breast cancer. *Cancer Res,* 68, 1667-74.

XU, C. F., BING, N. X., BALL, H. A., RAJAGOPALAN, D., STERNBERG, C. N., HUTSON, T. E., DE SOUZA, P., XUE, Z. G., MCCANN, L., KING, K. S., RAGONE, L. J., WHITTAKER, J. C., SPRAGGS, C. F., CARDON, L. R., MOOSER, V. E. & PANDITE, L. N. 2011. Pazopanib efficacy in renal cell carcinoma: evidence for predictive genetic markers in angiogenesis-related and exposure-related genes. *J Clin Oncol,* 29, 2557-64.

XUE, Y., CANMAN, J. C., LEE, C. S., NIE, Z., YANG, D., MORENO, G. T., YOUNG, M. K., SALMON, E. D. & WANG, W. 2000. The human SWI/SNF-B chromatin-remodeling complex is related to yeast rsc and localizes at kinetochores of mitotic chromosomes. *Proc Natl Acad Sci U S A,* 97, 13015-20.

YACHIDA, S., JONES, S., BOZIC, I., ANTAL, T., LEARY, R., FU, B., KAMIYAMA, M., HRUBAN, R. H., ESHLEMAN, J. R., NOWAK, M. A., VELCULESCU, V. E., KINZLER, K. W., VOGELSTEIN, B. & IACOBUZIO-DONAHUE, C. A. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467, 1114-7.

YANG, J. C., HUGHES, M., KAMMULA, U., ROYAL, R., SHERRY, R. M., TOPALIAN, S. L., SURI, K. B., LEVY, C., ALLEN, T., MAVROUKAKIS, S., LOWY, I., WHITE, D. E. & ROSENBERG, S. A. 2007. Ipilimumab (anti-CTLA4 antibody) causes regression of metastatic renal cell cancer associated with enteritis and hypophysitis. *J Immunother,* 30, 825-30.

YANG, W., YOSHIGOE, K., QIN, X., LIU, J. S., YANG, J. Y., NIEMIERKO, A., DENG, Y., LIU, Y., DUNKER, A., CHEN, Z., WANG, L., XU, D., ARABNIA, H. R., TONG, W. & YANG, M. 2014. Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics,* 15 Suppl 17, S2.

YAO, M., HUANG, Y., SHIOI, K., HATTORI, K., MURAKAMI, T., SANO, F., BABA, M., KONDO, K., NAKAIGAWA, N., KISHIDA, T., NAGASHIMA, Y., YAMADA-OKABE, H. & KUBOTA, Y. 2008. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. *Int J Cancer,* 123, 1126-32.

YAO, M., YOSHIDA, M., KISHIDA, T., NAKAIGAWA, N., BABA, M., KOBAYASHI, K., MIURA, T., MORIYAMA, M., NAGASHIMA, Y., NAKATANI, Y., KUBOTA, Y. & KONDO, K. 2002. VHL tumor suppressor gene alterations associated with good prognosis in sporadic clear-cell renal carcinoma. *J Natl Cancer Inst,* 94**,** 1569-75.

YAP, T. A., GERLINGER, M., FUTREAL, P. A., PUSZTAI, L. & SWANTON, C. 2012. Intratumor heterogeneity: seeing the wood for the trees. *Sci Transl Med,* 4**,** 127ps10.

YOSHIDA, M., YAO, M., ISHIKAWA, I., KISHIDA, T., NAGASHIMA, Y., KONDO, K., NAKAIGAWA, N. & HOSAKA, M. 2002. Somatic von Hippel-Lindau disease gene mutation in clear-cell renal carcinomas associated with end-stage renal disease/acquired cystic disease of the kidney. *Genes Chromosomes Cancer,* 35**,** 359-64.

YU, H., MASHTALIR, N., DAOU, S., HAMMOND-MARTEL, I., ROSS, J., SUI, G., HART, G. W., RAUSCHER, F. J., 3RD, DROBETSKY, E., MILOT, E., SHI, Y. & AFFAR EL, B. 2010. The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. *Mol Cell Biol,* 30**,** 5071-85.

ZHANG, J., WU, L. Y., ZHANG, X. S. & ZHANG, S. 2014. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics,* 15**,** 271.

ZHANG, Q., YING, J., LI, J., FAN, Y., POON, F. F., NG, K. M., TAO, Q. & JIN, J. 2010. Aberrant promoter methylation of DLEC1, a critical 3p22 tumor suppressor for renal cell carcinoma, is associated with more advanced tumor stage. *J Urol,* 184**,** 731-7.

ZHAO, H., LJUNGBERG, B., GRANKVIST, K., RASMUSON, T., TIBSHIRANI, R. & BROOKS, J. D. 2006. Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med,* 3**,** e13.

ZHAO, H., ZONGMING, M., TIBSHIRANI, R., HIGGINS, J. P., LJUNGBERG, B. & BROOKS, J. D. 2009. Alteration of gene expression signatures of cortical differentiation and wound response in lethal clear cell renal cell carcinomas. *PLoS One,* 4**,** e6039.

ZHONG, H., CHILES, K., FELDSER, D., LAUGHNER, E., HANRAHAN, C., GEORGESCU, M. M., SIMONS, J. W. & SEMENZA, G. L. 2000. Modulation of hypoxia-inducible factor 1alpha expression by the epidermal growth factor/phosphatidylinositol 3-kinase/PTEN/AKT/FRAP pathway in human prostate cancer cells: implications for tumor angiogenesis and therapeutics. *Cancer Res,* 60**,** 1541-5.

ZHOU, L., CHEN, J., LI, Z., LI, X., HU, X., HUANG, Y., ZHAO, X., LIANG, C., WANG, Y., SUN, L., SHI, M., XU, X., SHEN, F., CHEN, M., HAN, Z., PENG, Z., ZHAI, Q., CHEN, J., ZHANG, Z., YANG, R., YE, J., GUAN, Z., YANG, H., GUI, Y., WANG, J., CAI, Z. & ZHANG, X. 2010. Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. *PLoS One,* 5**,** e15224.

ZHOU, L. & HANEMANN, C. O. 2012. Merlin, a multi-suppressor from cell membrane to the nucleus. *FEBS Lett,* 586**,** 1403-8.

ZIGEUNER, R., HUTTERER, G., CHROMECKI, T., IMAMOVIC, A., KAMPEL-KETTNER, K., REHAK, P., LANGNER, C. & PUMMER, K. 2010. External validation of the Mayo Clinic stage, size, grade, and necrosis (SSIGN) score for clear-cell renal cell carcinoma in a single European centre applying routine pathology. *Eur Urol,* 57**,** 102-9.

ZISMAN, A., PANTUCK, A. J., DOREY, F., SAID, J. W., SHVARTS, O., QUINTANA, D., GITLITZ, B. J., DEKERNION, J. B., FIGLIN, R. A. & BELLDEGRUN, A. S. 2001.

Improved prognostication of renal cell carcinoma using an integrated staging system. *J Clin Oncol,* 19**,** 1649-57.

ZWIENER, I., BLETTNER, M. & HOMMEL, G. 2011. Survival analysis: part 15 of a series on evaluation of scientific publications. *Dtsch Arztebl Int,* 108**,** 163-9.