Characterising DNA methylation in tissue and liquid samples from patients with renal tumours

Sabrina Helena Rossi

Department of Oncology

Faculty of Clinical Medicine

University of Cambridge

This dissertation is submitted for the degree of *Doctor of Philosophy*

Gonville & Caius College

February 2022

Declarations

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text (Chapter 4 outlines all collaborations and contributions). The thesis does not exceed the prescribed word limit (60,000 words excluding references, figure and table legends).

Sabrina H. Rossi
February 2022

Thesis Summary

Characterising DNA methylation in tissue & liquid samples from patients with renal tumours

Sabrina Helena Rossi

The incidence of renal cell carcinoma (RCC) and small renal masses (SRMs), defined as <4cm in diameter, is increasing dramatically. SRMs encompass a variety of potential diagnoses, including benign and malignant tumours, the most common of which is clear cell RCC (ccRCC). Current methods are unable to confidently distinguish pathological subtypes of SRMs, meaning patients with benign tumours are undergoing unnecessary invasive surgery. In addition, there are difficulties risk stratifying patients with ccRCC and predicting outcomes. Genomic alterations, such as mutational analysis, may have a role in RCC diagnosis and prognostication, but are unlikely to be sufficient alone due to low recurrence rates and significant intra-tumoral heterogeneity (ITH), which limit detection. Changes in DNA methylation are abundant and often early events in tumorigenesis, which may overcome these challenges as potential tumour markers both in tissue and liquid biopsies.

To address the aforementioned diagnostic challenge, I characterised DNA methylation and gene expression in tissue from patients with benign and malignant renal tumours to elucidate similarities and differences between tumour subtypes. Subsequently, DNA methylation data were combined on over 1200 tissue samples and these were used to train and test MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours), a machine learning model to predict common pathological subtypes of renal tumours. MethylBoostER was externally validated on four independent publicly available datasets (N=518), demonstrating a high accuracy (receiver operating characteristic area under the curve; AUC >0.90). MethylBoostER predicted consistent classification of multi-region samples from the same patient in 90% of individuals, suggesting ITH does not limit model applicability in a biopsy setting.

Subsequently, I undertook a systematic evaluation of methylation heterogeneity in ccRCC, exploring associations with clinical/prognostic parameters and highlighting implications for biomarker selection. I evaluated multi-region tissue samples (N=135) from ccRCC patients (N=18) and assessed heterogeneity between patients, within a patient and within a sample. Inter-patient heterogeneity dominated over intra-tumoural heterogeneity. My analysis represents the first evaluation of epipolymorphism, a measure of methylation heterogeneity within a sample, in ccRCC. Significant differential epipolymorphism was noted in ccRCC versus normal kidney at the promoter region of genes known to be implicated in kidney cancer and this finding was externally validated in an independent cohort (N=71). Although changes in epipolymorphism are believed to be a stochastic process, my results suggest that disordered methylation may accumulate in functionally relevant loci which are known to contribute to ccRCC tumorigenesis.

Circulating tumour DNA (ctDNA) represents a promising target for non-invasive liquid biopsy in both diagnostic and prognostic applications. Mutational analyses of ctDNA have produced disappointing detection rates in ccRCC, possibly hampered by low ctDNA levels and high mutational ITH. I therefore performed targeted methylation analysis of ctDNA using a novel method- Nimbus (Non-destructive Integration of Methylation to Boost Underlying Signals). Targeted analysis of hypomethylated regions in plasma ctDNA distinguished ccRCC from cancer-free controls with an AUC of 0.96 and produced superior detection rates compared to mutational analysis (93% vs 50%). My results suggest that tumour signal may be enriched in post-biopsy fluid (proximal sample) compared to plasma (distal sample), a strategy that could be useful in patients with SRMs to complement the current diagnostic pathway and overcome low concentrations of plasma ctDNA.

In summary, I comprehensively characterise DNA methylation in tissue and liquid samples derived from patients with renal tumours. In the future, DNA methylation analysis of renal tumour biopsy tissue and/or liquid biopsy samples could enable improved diagnosis of patients with SRMs as well as facilitating prognostic stratification.

Dedicato a μαμά e papà

Acknowledgements

I would like to thank my wonderful and very supportive supervisors: Dr Charlie Massie and Prof Grant D. Stewart. A special '*grazie mille'* to Charlie, from the bottom of my heart. Grant, you have offered me invaluable guidance for so many years, throughout this long journey.
Thank you to Dr Roland Schwarz, Victoria Dombrowe, Izzy Newsham and Dr Shamith Samarajiwa – your role was crucial and I couldn't have done this without you. I am very grateful for the help received from all my collaborators (detailed in Chapter 4), and in particular Sara Pita and Radek Lach for their bioinformatics expertise. Thank you to the entire Massie Group for their support and unwavering encouragement (Robert Hanson, Henno Martin, Ermira Lleshi, Anne Babbage, Gahee Park), the team at the Cancer Research UK Cambridge Institute (Wing Kit Leung, Chris Smith, Wendy Cooper) and Addenbrooke's Hospital (Harvey Dev, Tom Mitchell, Anne Warren). I would like to acknowledge Cancer Research UK for funding my doctoral research fellowship and The Evelyn Trust for providing additional funding. Thank you to the patients who provided samples, the entire biobanking team at Addenbrooke's Hospital, and the members of the core facilities (Hutchison MRC, CMDL and CRUK CI).

Publications arising from this work

The material presented in this dissertation is associated with a number of manuscripts, publications and grants, as outlined below.

I performed a Research Priority Setting Initiative to determine key unmet needs in kidney cancer research, and addressed these in my thesis [1, 2].

The background section presented in Chapter 2 is adapted from two book chapters [3, 4]. The background also contains information from two systematic reviews [5, 6]. The results presented in Chapter 5 form a manuscript which has now been accepted for publication (Science Advances, accepted July 2022) [7]. The full text can be found in Appendix 1. Two further manuscripts are planned containing results from Chapter 6 and Chapter 7. Data from Chapter 6 contributed to a further publication [8]. Based on preliminary results, I obtained a competitive grant from The Evelyn Trust (£42,000), which has allowed me to fund the analysis presented in Chapter 7.

Due to the coronavirus outbreak in Spring 2020, I paused my PhD and returned to full time clinical practice in the NHS so that I could help with the response to the pandemic. I was re-deployed from urology registrar to 'covid medical SHO' for 3 months. During this time, I recruited patients to the RECOVERY trial and participated in two additional research collaboratives, leading to several publications [9-14].

Publication References:

1. Rossi SH, Blick C, Handforth C, Brown JE, Stewart GD, Renal Cancer Gap Analysis Collaborative. Essential Research Priorities in Renal Cancer: A Modified Delphi Consensus Statement. Eur Urol Focus. 2020;6(5):991-8.
2. Rossi SH, Fielding A, Blick C, Handforth C, Brown JE, Stewart GD. Setting Research Priorities in Partnership with Patients to Provide Patient-centred Urological Cancer Care. Eur Urol. 2019;75(6):891-3.
3. Rossi SH, Stewart GD. Renal Cancer. In: Pang K, Osman NI, Catto J, Chapple C, editors. Basic Urological Sciences. 1 ed: CRC Press; 2021.
4. Rossi SH, Stewart GD. Epidemiology and screening for renal cancer. In: Anderson CJ, Patel HRH, editors. Renal Cancer: Current and Future Innovations. 1 ed: Springer Nature; 2022.
5. Usher-Smith JA, Li L, Roberts L, Harrison H, Rossi SH, Sharp SJ, et al. Risk models for recurrence and survival after kidney cancer: a systematic review. BJU Int. 2021. *In press.*
6. Flitcroft JG, Verheyen J, Vemulkar T, Welbourne EN, Rossi SH, Welsh SJ, et al. Early detection of kidney cancer using urinary proteins: a truly non-invasive strategy. BJU Int. 2022;129(3):290-303.
7. Rossi SH*, Newsham I* et al. Accurate detection of benign and malignant renal tumour subtypes with MethylBoostER: an epigenetic marker driven learning framework. Science Advances. 2022. *Accepted.*
8. Sciacovelli M, Dugourd A, Jimenez LV, Yang M… Rossi SH et al. Nitrogen partitioning between branched-chain amino acids and urea cycle enzymes sustains renal cancer progression. bioRxiv. 2021:2021.09.17.460635.
9. RECOVERY Trial Group, Horby P et al. Dexamethasone in Hospitalized Patients with Covid-19. The New England journal of medicine. 2021;384(8):693-704.
10. GlobalSurg Collaborative. SARS-CoV-2 infection and venous thromboembolism after surgery: an international prospective cohort study. Anaesthesia. 2022;77(1):28-39.
11. GlobalSurg Collaborative. Effects of pre-operative isolation on postoperative pulmonary complications after elective surgery: an international prospective cohort study. Anaesthesia. 2021;76(11):1454-64.
12. Covidsurg Collaborative. SARS-CoV-2 vaccination modelling for safe surgery to save lives: data from an international prospective cohort study. Br J Surg. 2021;108(9):1056-63.
13. GlobalSurg Collaborative. Timing of surgery following SARS-CoV-2 infection: an international prospective cohort study. Anaesthesia. 2021;76(6):748-58.
14. Bergamaschi L et al. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. Immunity. 2021;54(6):1257-75 e8.

Table of contents

List of figures

List of tables

List of abbreviations

450k array	Infinium® HumanMethylation450 BeadChip (Illumina)
APITH	Average pairwise ITH index
AS	Active surveillance
AUC	Area under the curve
BH	Benjamini Hochberg
BMI	Body mass index
Bp	Base pairs
BS	Bisulphite sequencing
ccRCC	Clear cell RCC
cfDNA	Cell free DNA
cfMeDIP–seq	Cell-free methylated DNA immunoprecipitation & high-throughput sequencing
CHH and CHG	Methylation in a non-CpG context
chRCC	Chromophobe RCC
CIMP	CpG island methylator phenotype
CMDL	Cambridge Cancer Molecular Diagnostics Laboratory
CpG	Cytosine and guanine dinucleotides
CRUK CI	Cancer Research UK Cambridge Institute
CT	Computed tomography
ctDNA	Circulating tumour DNA
DCT	Distal convoluted tubule
DEG	Differentially expressed genes
DMC	Differentially methylated cytosines
DMR	Differentially methylated regions
DNMT	DNA methyltransferase enzymes
EAU	European Association of Urology
e-locus	Epigenetic locus (four adjacent CpGs in a single sequencing read)
EMT	Epithelial to mesenchymal transition
Epic-seq	TruSeq Methyl Capture EPIC library preparation and sequencing (Illumina)
gDNA	Genomic DNA
GEO	Gene Expression Omnibus
GSEA	Gene set enrichment analysis
HIF	Hypoxia inducible factor
IQR	Inter-quartile range
Indels	Insertions and deletions
INVAR TAPAS	Integration of Variant Reads Tailored Panel Sequencing
ITH	Intra-tumoural heterogeneity
LMC	Latent methylation component
MAF	Mutant allele fraction
MCC	Matthews correlation coefficient
MethylBoostER	Methylation and XGBoost for Evaluation of Renal tumours
mRCC	Metastatic renal cell carcinoma
MRD	Minimal residual disease
Nimbus	Non-destructive Integration of Methylation to Boost Underlying Signals
OS	Overall survival
PCA	Principal component analysis
PCAR	Predicted to chronological age ratio
PCR	Polymerase Chain Reaction
PCT	Proximal convoluted tubule

pRCC	Papillary RCC
PFS	Progression free survival
RCC	Renal cell carcinoma
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic curve
SCNA	Somatic copy number alterations
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SRM	Small renal mass
TCGA	The Cancer Genome Atlas
TCGA KIRC	The Cancer Genome Atlas Kidney ccRCC
TET	Ten-eleven translocation enzyme
VHL	Von Hippel Lindau
WES	Whole exome sequencing
WGBS	Whole genome bisulphite sequencing
XGBoost	Extreme gradient boosting classifier model

Chapter 1 Overview

Renal cell carcinoma (RCC), also known as kidney cancer, is the 6th most common cancer in men and 10th most common cancer in women in the UK [15, 16]. RCC, defined as an adenocarcinoma arising from the renal tubular epithelial cells, comprises >90% of adult renal cancer. The incidence of RCC has been rising steadily in the UK, with incidence rates increasing faster than most other malignancies [17]. Despite advances in detection and management, mortality rates have only minimally improved [18, 19], meaning there is a need to focus research efforts on interventions which are most likely to benefit patients. Therefore, we developed an international collaboration amongst clinicians, scientists and patients, 'The Renal Gap Analysis Collaborative', to identify research priorities in RCC using a transparent, rigorous methodology and a patient centred approach. I led this initiative, with support from the Collaborative Steering Committee, which resulted in the identification of 14 key research gaps published as a Platinum Priority editorial in European Urology [1, 2]. Three of these research priorities are the need for: 1/ improved characterisation of small renal masses (SRMs); 2/ biomarkers that may be applied to renal biopsies to evaluate and overcome molecular heterogeneity; and 3/ biomarkers to enable risk stratification and aid management decisions in localised RCC. This thesis aims to characterise DNA methylation in kidney tissue and liquid samples from patients with malignant and benign renal tumours, to explore and address these key research questions.

In summary, Chapter 2 provides the background for this thesis, contextualising existing knowledge regarding the current clinical pathways, genetic and epigenetic landscape in RCC and the rationale for DNA methylation markers. Chapter 3 synthesizes this information and justifies my aims and objectives. The materials and methods, including experimental procedures and statistical analysis, are summarised in Chapter 4, as these are shared methods used across all subsequent chapters. Three results chapters are presented (Chapters 5 to 7), which cover the following topics: differentiating pathological subtypes of renal tumours, characterising methylation heterogeneity in ccRCC tissue and circulating tumour DNA (ctDNA) detection in liquid samples, respectively. Chapter 8 consists of an integrated discussion and conclusion, emphasising future plans. All the work mentioned in this thesis was undertaken by me, unless specifically mentioned in the text (Chapter 4 delineates collaborator contributions in detail).

Chapter 2 Background

2.1 The clinical landscape of renal cancer

The following section provides the background to the two main clinical questions addressed in my thesis, namely diagnostic challenges associated with differentiating small renal masses and difficulties predicting prognostic outcomes in patients with ccRCC.

2.1.1 Small renal masses represent a diagnostic challenge

The incidence of RCC has increased by 47% over 10 years [20], making it one of the fastest accelerating cancers and this is projected to rise further in future. This dramatic increase is attributed to both a true rise in the disease (which may be related to rising prevalence of risk factors such as age, obesity, hypertension and diabetes), but also due to increasing incidental detection during imaging for other complaints [19, 20]. Use of computed tomography (CT) has surged in recent decades due to technological advances (enabling superior resolution, reduced scanning times and lower radiation dose), widening availability and reducing costs [21, 22]. As the use and sensitivity of CT imaging increases, so has the incidental detection of small renal masses (SRMs), defined as <4cm in diameter. Welch et al report that 43% of individuals aged 65–85 years on Medicare in the USA undergo either a chest or abdominal CT over a 5-year period and renal lesions are amongst the most common incidental findings [22, 23]. The number of CTs performed was significantly correlated with the number of nephrectomies: it is estimated that performing 1000 additional CTs is associated with 4 additional nephrectomies [22]. SRM is a broad term which encompasses a mixture of potential diagnoses: including clear cell (ccRCC), papillary (pRCC), chromophobe (chRCC) RCC, or benign disease, such as oncocytoma or angiomyolipoma (AML). Each of these pathological subtypes has characteristic genetic and molecular features, such that it is argued that RCC is not a single disease [24].

ccRCC, derived from the proximal convoluted tubule in the renal cortex, accounts for approximately 75% of RCC (Figure 2.1) [25]. Microscopically, cells contain cytoplasm with a high lipid and glycogen content, which dissolves upon histopathological processing to create the distinctive clear appearance that gives rise to the name [26]. A minority of cases demonstrate eosinophilic, granular cytoplasm, and these may be mistaken for oncocytoma on renal biopsy. pRCC, also derived from the proximal convoluted tubule, accounts for 10-15% of cases of RCC and can be subdivided into types 1

and 2 (Figure 2.1) [25]. chRCC, which is derived from the distal nephron, represents 5% of RCC cases. Histological subtype is a major predictor of survival (Figure 2.1). The best prognosis is noted in chRCC and type 1 pRCC, with progressively worsening survival outcomes in type 2 pRCC, ccRCC and a subset of pRCC named CpG island methylator phenotype (CIMP) [24]. Oncocytomas are benign tumours derived from the intercalated cells of the distal nephron, which may be misdiagnosed as RCC (Figure 2.1). Indeed, it is the most common benign diagnosis found on nephrectomy [26, 27]. Microscopically, oncocytomas are characterised by eosinophilic cytoplasm and abundance of mitochondria.

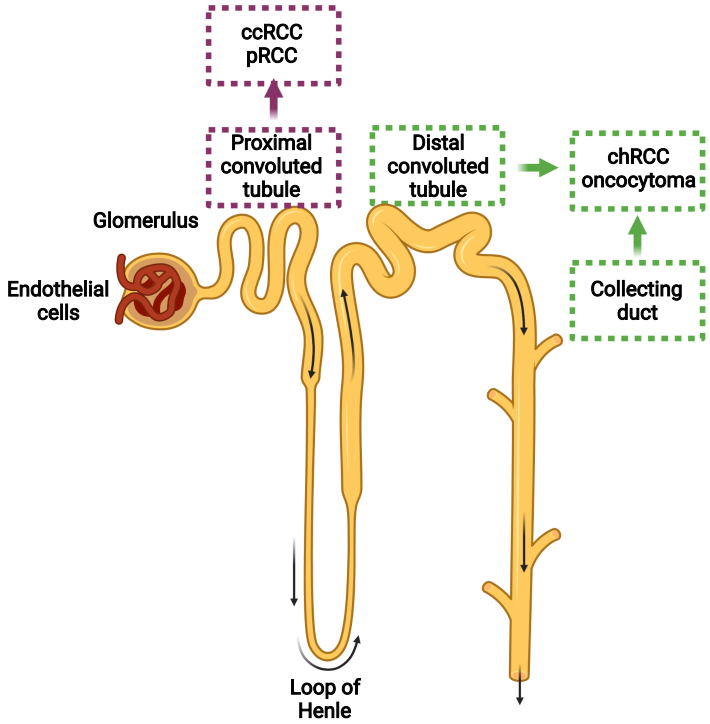


Figure 2.1: Kidney tubule anatomy and putative cell of origin for different tumour types

Patients with SRMs may be offered a number of management options depending on the most likely diagnosis, fitness levels and individual choice [27]. Options include nephrectomy (which is usually partial but may be radical), ablative therapy or active surveillance (AS). Local ablative therapy (which includes radiofrequency and cryo- ablation) is minimally invasive and can be used in patients with SRMs who are poor surgical candidates, though repeated treatments may be necessary [28]. Patients with benign or low-grade disease, or those with limited life expectancy, could benefit from AS. However, current imaging techniques and renal biopsy are unable to distinguish benign from malignant disease with confidence [29]. Contrast enhanced CT, ultrasound and magnetic resonance imaging may be used to aid diagnosis. The presence of enhancement, i.e. a change of ≥ 15 Hounsfield Units (HU) before and after contrast administration on CT, is considered the most important criterion for the differentiation of malignant from benign SRMs [27]. Enhancement delineates vascularity, can help distinguish cystic from solid lesions, as well as demonstrating specific patterns for each pathological subtype (for example oncocytomas classically exhibit a central stellate scar and segmental enhancement inversion) [29]. However, enhancement results in only 60% sensitivity and 73% specificity to discriminate benign lesions [30]. There is an increasing likelihood of malignancy and higher grade with increasing lesion size, however few imaging features are discriminatory, highlighting the need for additional specific markers [31].

Renal core biopsy is currently recommended by the European Association of Urology (EAU) Guidelines for individuals with indeterminate SRMs following imaging. In addition, biopsy may be used to obtain a tissue diagnosis prior to ablative therapy, AS or inclusion in a clinical trial [27]. The procedure, done under local anaesthetic, is now considered very safe, although there are known risks: pain, bleeding, infection/sepsis or accidental damage to adjacent structures (e.g. pneumothorax or abdominal organ injury) [32]. Of these, the most common complication is bleeding, which tends to be self-limiting (e.g. perinephric haematoma or visible haematuria in ~10% cases), but may require blood transfusion (~1% cases) or intervention (0.3% cases) [32, 33]. Reports of tumour seeding along the needle tract, although rare, have been published [34] and contribute to caution in the adoption of this procedure. Use of a co-axial needle is recommended as this may mitigate both seeding and bleeding by minimising punctures. Renal biopsy is a very helpful diagnostic tool but can be inconclusive due to a number of reasons. There may be difficulty in accessing the tumour due to anatomical location or small lesion size, leading to limited sampling, or presence of necrosis (biopsy is non-diagnostic in ~10% of cases) [35]. Importantly, pathologists face difficulties assigning tumour type from small tissue biopsy samples, in particular, differentiating oncocytomas from the eosinophilic variant of chRCC and ccRCC. A meta-analysis demonstrated one

in four renal biopsies reported as oncocytoma are found to be RCC following surgical excision [36]. Hybrid oncocytic/chromophobe tumours (HOCT), which fortunately are rare, include areas of both chRCC and oncocytoma. There are also difficulties differentiating chRCC (which tends to have low aggressive potential) from ccRCC (which is generally associated with a worse prognosis) [24]. Therefore, erring on the side of caution, patients with SRMs are often offered surgery and AS is under-utilised [37]. As a result, approximately 20%-30% of SRMs removed at surgery are found to be benign post-operatively [38, 39]. This means that a significant number of patients are undergoing surgery for a benign condition, with associated post-operative risks of morbidity and mortality, and long-term impact on renal function. The rates of post-operative complications following minimally invasive surgery are: blood transfusion (5%), re-operation (2-5%), respiratory complications (1-7%) and even death (4%) [40]. Although complication rates are reducing with advancements in techniques and technologies, the risk of severe complications in elderly patients is 6% [41]. Increasing the use of AS (which has been shown to be non-inferior to primary intervention) especially in patients with comorbidities who may have a limited life expectancy, could reduce over-treatment [42]. This needs to be balanced with the risks of a missed or delayed diagnosis of aggressive RCC. The majority of SRMs are slow growing and have low metastatic potential, though identifying aggressive disease remains a challenge (growth rates of SRMs are discussed more in detail below). Getting the diagnosis right is crucial and additional diagnostic tools are needed to improve the current pathway.

In addition to diagnostic challenges relating to SRMs (i.e. differentiation of pathological subtypes), there are also complexities relating to patient risk stratification and prognosis. Determinants of aggressive clinical course include: pathological subtype (where pRCC type 2 and ccRCC have the worse prognosis and chRCC and pRCC type 1 have the best), higher lesion grade, size, stage and tumour growth rate [43]. Unfortunately, these major determinants of prognosis in patients with SRMs cannot be reliably determined with current standard of care methods (imaging and biopsy). ccRCC is characterised by intra-tumoural heterogeneity (ITH) meaning renal biopsy may be limited by tumour sampling [44]. In up to 25% of cases, malignant RCC have coexisting areas of both low and high tumour grade [45, 46]. Stage may also be misclassified as the sensitivity and specificity of CT for perinephric extension, venous invasion, metastatic adenopathy and organ invasion is 46, 78, 83 and 60%, respectively, and 98, 96, 88 and 100%, respectively [29]. 10-40% of SRMs demonstrate evidence of local invasion, meaning the tumour is <4cm in diameter on imaging but ≥T3 on pathological staging determined post-operatively [47], though fortunately more recent studies have estimated this rate to be <10% [48, 49]. In addition, MRI may aid assessment of vena cava invasion

and improve the staging accuracy of CT [29]. 3-12% of SRMs will either present with concurrent metastases or will develop metastases at a later date [47]. SRMs on AS show very different rates of growth and clinical progression. Up to one-third of SRMs exhibit aggressive potential (rapid growth >0.5cm/year or doubling time <12 months), with the remainder growing slowly or remaining stable in size [50, 51]. Recently a growing number of observational studies are being performed which are increasing our understanding of the natural history of the disease [42, 52]. However, there is a lack of validated scores that will differentiate malignant SRMs that will progress rapidly on AS compared to those that have a more indolent course. Linear growth rate has been proposed as a marker for aggressiveness, but this has recently been challenged, as it did not correlate with overall outcome, and similar average growth rates were observed for benign and malignant (low and high grade) SRMs [53, 54]. Therefore, there is a risk for both overdiagnosis and overtreatment of indolent masses, as well as undertreatment of aggressive disease, highlighting the need for better diagnostic and prognostic markers.

2.1.2 Differentiating indolent from aggressive ccRCC represents a prognostic challenge

Accurately predicting outcomes in patients with localised ccRCC remains a research priority [1]. In patients with non-metastatic ccRCC undergoing surgery with curative intent, 30% of individuals will subsequently develop a recurrence [55]. An improved ability to predict risk of recurrence would enable tailored patient counselling, individualised post-operative follow-up and potentially neoadjuvant or adjuvant treatment [56]. Appropriate post-operative surveillance (using imaging or ideally a biomarker test) is particularly crucial to identify recurrence early, whilst minimizing the burden of follow-up on the patients' quality of life and impact on health care resources. Although tyrosine kinase inhibitors in the adjuvant setting have been largely disappointing [57], early data on immunotherapies appear to be more promising. The KEYNOTE-564 trial has recently shown an improvement in two-year disease-free survival in intermediate-high and high-risk patients with ccRCC randomised to adjuvant pembrolizumab monotherapy compared to placebo [58]. Longer follow-up and impact on overall survival (OS) are pending. These considerations have prompted the development of clinical and biomarker predictive models for localised ccRCC.

Table 2.1: Commonly used prognostic risk scores for non-metastatic RCC

Adapted from [5, 27].

Risk score	Factors evaluated	Risk classification
Leibovich score [59]	Points awarded for: • T classification (pT1a: 0, pT1b: 1, pT2:3, pT3-4: 4 points) • N classification (pNx/N0: 0, pN+: 2 points) • Tumour size (< 10 cm: 0, ≥10 cm: 1 point) • Grade (G1-2: 0, G3: 1, G4: 3 points) • Tumour necrosis (absent: 0, present: 1 point)	• Low risk: 0-2 points • Intermediate risk: 3-5 points • High risk: ≥6 points
University of California Los Angeles Integrated Staging System (UISS) [60]	• Eastern Cooperative Oncology Group performance status (ECOG PS) • T classification • N classification (N+ classified as metastatic) • Grade (G)	• Low risk: T1N0M0 G1–2, ECOG PS 0 • High risk: T3N0M0 G2–4, ECOG PS ≥1 OR T4N0M0 • Intermediate risk: Any other N0M0
Stage, Size, Grade and Necrosis score (SSIGN) [61]	• Stage (T, N, M) • Tumour size • Grade • Necrosis	• Increasing score (score 0-15) associated with worse prognosis
Karakiewicz score [62]	• T stage • N stage • M stage • Tumour size • Fuhrman grade • Symptom classification	• Nomogram giving continuous quantification of risk
Sorbellini score [63]	• TNM stage 2002 • Tumour size • Fuhrman grade • Necrosis • Vascular invasion • Clinical presentation (incidental asymptomatic, local symptoms or systemic symptoms)	• Nomogram giving continuous quantification of risk
Grade, Age, Nodes and Tumour score (GRANT) [64]	• Age > 60 years • T classification = T3b, pT3c or pT4 • N classification = pN1 • Fuhrman grade = G3 or G4	• Favourable risk: 0-1 factors • Unfavourable risk: ≥2 factors

A number of anatomical, histological and clinical prognostic parameters have been identified in the literature. Anatomical factors include tumour size, local invasion (venous or collecting system, adrenal, Gerota's perinephric fascia) and stage. Histological factors consist of tumour grade, pathological subtype, sarcomatoid or rhabdoid differentiation, necrosis, microvascular and lymphovascular invasion. Clinical factors (such as symptoms and performance status) are used in metastatic but not non-metastatic RCC [56]. Serum biochemistry and haematology tests which have been shown to have prognostic potential include: calcium, albumin, LDH, C-reactive protein, haemoglobin, platelet count and neutrophil-to-lymphocyte ratio [5, 19]. However, although these are routinely available blood tests, they are not recommended as prognostic markers in clinical practice [27]. A number of prognostic scores have been developed integrating different combinations of the above-named factors in order to predict outcomes in localised disease, including the well-known Leibovich score and the University of California Los Angeles Integrated Staging System (UISS) (Table 2.1) [56, 65]. Although the EAU guidelines advocate the use of prognostic scores, no specific score is recommended as none have been demonstrated to be superior [27]. We recently performed a comprehensive, systematic review of prognostic models in RCC [5]. The main finding of the review is that there is no clear single 'best' model, however the SSIGN (cited in the ESMO guidance) and the UISS model (cited in both ESMO and EAU guidelines) perform relatively poorly. Three other models (Sorbellini, Karakiewicz and Leibovich) performed highly in all three outcomes evaluated (recurrence free survival, cancer specific survival and OS), suggesting these may be more appropriate alternatives. However, clinical models are still limited in their ability to predict recurrence, particularly in intermediate risk patients, which has driven the search for predictive biomarkers in this setting. Though hundreds of studies have been published evaluating prognostic biomarkers in ccRCC (the most relevant are described in section 2.3), none have been adopted in clinical practice [66-69].

2.2 The genetic landscape of renal cancer

Although >95% of cases are sporadic, a number of hereditary RCC syndromes have been identified, which have contributed to our overall understanding of the disease process (Table 2.2). Large scale initiatives, such as The Cancer Genome Atlas (TCGA), have made significant advances elucidating the genetic landscape of sporadic RCC [24, 70]. It is hoped that this knowledge may translate to improvements in biomarker development.

RCC is characterised by a high prevalence of distinct patterns of somatic copy number alterations (SCNA), which are associated with each of the pathological subtypes. Recurrent SCNAs most often involve whole chromosome or chromosome arms [70]. ccRCC is characterised by loss of chromosome 3p which results in loss of heterozygosity (LOH) of RCC driver genes that are also recurrently mutated in ccRCC (*VHL, PBRM1*, *BAP1* and *SETD2*) [71]. Mitchell et al identified complex structural rearrangements associated with LOH at 3p and gains at 5q, which may occur in adolescence through a chromothripsis event decades before ccRCC develops [71]. pRCC type 1 is characterised by gains in chromosome 7 and 17, whilst a poor prognosis subset of pRCC type 2 is associated with high levels of genome-wide aneuploidy and chromosome 9p loss (deleting *CDKN2A*) [72]. chRCC is associated with loss of whole chromosomes including chromosome 1, 2, 6, 10, 13 and 17, with imbalanced chromosome duplication (ICD) associating with metastatic chRCC [24, 73, 74]. In addition to SCNAs associated with RCC subtypes, loss of 9p and 14q have been identified as hallmark driver events in tumour metastasis and together with genome wide aneuploidy and ICD have been linked with worse survival [75]. Genomic structural variation in SRMs may therefore provide some diagnostic and prognostic information.

In addition, RCC is characterised by a relatively low number of somatic mutations (~1 single nucleotide variant/Mbp). Recurrently mutated driver genes have been identified in RCC pathological subtypes, and a subset of these have prognostic value [24]. Over 90% of patients with ccRCC harbour an alteration in the *VHL* gene. This may occur through loss of the short arm of chromosome 3, mutations or promoter hypermethylation [76]. Indeed, inactivation of *VHL* is a driver event which occurs in the trunk of the phylogenetic tree of ccRCC evolution [77]. Other significantly mutated genes in ccRCC include genes involved in the *SWI/SNF* chromatin remodelling complex (e.g. *PBRM1*, also located within the LOH region on 3p) and *PI(3)K/AKT/MTOR* signalling pathway (Table 2.3). Mutations in the *MET* proto-oncogene are enriched in pRCC and *TP53* mutations are more frequent events in chRCC, but rare ccRCC cases also harbour *MET* mutations and *TP53* mutations are found in all RCC subtypes [70]. Similarly, many of the driver mutations that are enriched in ccRCC are also found in other RCC subtypes (e.g. *VHL* and *PBRM1* are mutated in a subset of pRCC and chRCC) [70]. Overall, the enrichment of specific driver gene mutations in subtypes of RCC highlight the potential diagnostic information associated with these genomic alterations, but as outlined above, these features alone are not sufficient to classify RCC subtypes. Several genomic features have been associated with aggressive disease and metastatic phenotypes, including *TP53* mutations and *PI(3)K/AKT/MTOR* signalling pathway (e.g. *PTEN*) [70, 73], suggesting that genomic mutations may provide additional prognostic information.

Table 2.2: Genetic syndromes which predispose to hereditary RCC

A number of hereditary kidney cancer syndromes have been identified, which aid our overall understanding of sporadic disease [26, 78-81].

Syndrome	Estimated incidence in the general population	Genetics	Lifetime risk of RCC	Renal manifestations	Extrarenal manifestations
Von Hippel Lindau (VHL) Disease	1:30,000 to 1:35,000	Autosomal dominant *VHL* gene on chromosome 3p25-26	50-70%	ccRCC Early age at onset Bilateral, multifocal Renal cysts	Retinal angioma Haemangioblastoma of brainstem, cerebellum or spinal cord, Phaeochromocytoma Renal, pancreatic and epididymal cysts Inner ear tumours
Hereditary papillary RCC syndrome (HPRCC)	Unknown	Autosomal dominant *c-MET* gene on chromosome 7q31	90%	Type 1 pRCC Bilateral, multifocal	No tumours in other organs
Hereditary leiomyomatosis and RCC (HLRCC)	1:200,000	Autosomal dominant *FH* gene on chromosome 1q42-43	15-20%	Most commonly type 2 pRCC, although collecting duct RCC has also been reported Often very aggressive	Cutaneous and uterine leiomyomas
Succinate dehydrogenase RCC (SD RCC)	Unknown	Autosomal dominant SD genes: *SDHB* (most commonly), *SDHA, SDHC, SDHD*	10-15%	Specific type of RCC	Paragangliomas Gastrointestinal stromal tumours
Birt-Hogg-Dubé Syndrome	1:200,000	Autosomal dominant *FCLN* gene on chromosome 17.p.11.2	10-30%	Chromophobe RCC, oncoytomas and hybrid oncocytic-chromophobe tumours ccRCC and other subtypes have also been observed	Cutaneous fibrofolliculomas Lung cysts and spontaneous pneumothorax
Tuberous sclerosis	1:6,000 to 1:10,000	Autosomal dominant *TSC1* gene on chromosome 9p34 or *TSC2* gene on chromosome 16p13.3	1-5%	Renal cysts AML ccRCC	Epilepsy, learning difficulties and adenoma sebaceum

Table 2.3: Mutations and somatic copy number aberrations in ccRCC

For somatic copy number aberrations (SCNA), data regarding affected genes is also supplied. Numbers in brackets refer to the frequency of the events. Turajlic et al [75] suggested that the frequencies of events may be underestimated in studies which only sample a single tumour region (compared to multi-region studies) as some of these events may be subclonal. Data obtained from [70, 75, 76, 82].

Mutations	Somatic copy number aberrations

• Frameshift/point mutations in *VHL* or *VHL* complex genes (e.g. *TCEB1*) are noted in 70%-80% of patients and are clonal
• Mutations of chromatin modifying genes:
  ○ 30-50% *PBRM1* (2nd most commonly mutated gene in ccRCC), is a subunit of the SWI/SNF chromatic remodelling complex. This is often an early event, preceding mutations in *SETD2* and *PI3K* pathway genes.
  ○ 10-30% *SETD2* = histone methyltransferase
  ○ 5-15% *BAP1*, a histone deubiquitinase, is associated with a high number of SCNA, confirming its role in chromosome stability. It has prognostic potential and is associated with poor survival
  ○ 5-15% *KDM5C, ARID1A, SMARCA4*
• *TERT* mutations (6-14%)
• *CSMD3* mutations
• Alterations in the *SWI/SNF* complex (including *PBRM1* as mentioned above)
• *PTEN* gene and *PI3K/AKT/mTOR* signalling pathway
  ○ This includes *MTOR* gene
  ○ Other genes involved in the pathway: *TSC1/TSC2/PIK3CA*
• *TP53* gene (2-10%) and *TP53* pathway
• *CDKN2A* mutations, which are prognostic
	• Heterozygous loss of the short arm of chromosome 3 is the most common event (90%)→ *VHL, PBRM1, BAP1* and *SETD2* (the 4 most commonly mutated genes) are all located here
	• Second most common aberration: 5q gain (60%). Co-occurrent loss of 3p and gain of 5q is noted in 36% of ccRCC patients
	  ○ *FGFR4* gains seen in ~50% of ccRCC, due to gains in 5q
	• 14q loss (40%), the third most common aberration, leads to the loss of *HIF1-α*
	• Arm level or focal losses on chromosomes 1p, 3p, 4q, 6q, 8p, 9p and 14q
	  ○ 1p loss → *NEGR1* tumour suppressor
	  ○ 6p loss → *QKI*
	  ○ 9p loss → *CDKN2A, PTPRD*
	  ○ 14q → *NRXN3*
	• Gains on chromosomes 1q, 2q, 5q, 7q, 8q, 12p, and 20q
	  ○ 1q gain → *MDM4* (*TP53* regulator)
	  ○ 8q gain → *MYC*

2.3 Genetic heterogeneity and tumour evolution in ccRCC

In the 1970s, Nowell first coined the Darwinian concept of tumour evolution [83]. This theory was lent further support over subsequent decades through the observation that tumours display high genomic instability, a feature that is noted as one of the hallmarks of cancer [84]. This genetic variation facilitates tumour adaptation to selective pressures, leading to the emergence of distinct tumour subclones and, in contexts where space and nutrients are not limiting, results in tumour heterogeneity [85]. In tumour phylogenetic analysis, genetic changes which are observed in all tumour cells are considered clonal, driver/truncal events in the phylogenetic tree, whereas less prevalent genetic changes are defined as subclonal, branched events.

Sottoriva et al suggest a model of tumour evolution in which stochastic accumulation of mutations in the absence of selection results in early events making the largest contribution to intra-tumoural heterogeneity (ITH) [86]. This 'Big Bang' theory of tumour evolution suggests that early mutations lead to a variegated tumour cell population, and subsequently these persist in the form of heterogeneity. A corollary of this theory is that certain tumours are 'born to be bad', rather than acquiring aggressive potential through late acquisition of mutations and selection. This theory has implications for efforts to predict tumour behaviour and suggests that markers of poor prognosis may be detectable at early stages of tumour development.

ccRCC is characterised by a high degree of genetic ITH, with >60% of somatic mutations not detectable across all multi-regions sampled [44]. On average seven multi-region samples are needed to detect ≥75% of genetic variants [87]. Multi-region ccRCC analysis identified clonal aberrations in the VHL gene that were present in 90% of tumour samples, consistent with VHL mutations being early driver events [87, 88]. A subset of ccRCCs also demonstrated convergent evolution, meaning that different subclones evolved independently but converged with aberrations in the *VHL* gene. In addition, ccRCCs demonstrated convergent evolution in a number of other driver genes (i.e. *SETD2, BAP1* and *PTEN*), resulting in parallel outgrowth of subclones and further ITH [87].

Turajlic et al demonstrated that the pattern of ITH is associated with metastatic potential and survival, although the relationship is complex [75]. Low ITH (i.e. early fixation of driver events) was associated with rapid progression, whereas high ITH (i.e. >10 of subclonal drivers and highly branched evolution) was associated with attenuated progression [75]. This may reflect the outgrowth of a dominant clone that has outcompeted other clones, so may have a more aggressive malignant phenotype (and may therefore correlate with worse prognosis). Conversely, high ITH

index (compared to low ITH index) was significantly associated with higher tumour size, grade, stage and reduced progression free survival (PFS) [75]. Tumours with *VHL* monodriver events had the best PFS, whereas multiple clonal drivers had reduced PFS. This suggests that having a diverse clonal mixture would provide some evolutionary advantages (e.g. to facilitate bypass of selection bottlenecks). In addition, patients with both low ITH and low genome instability had attenuated metastatic progression, whereas patients with low ITH and high genome instability had the most rapid progression. This suggests that ITH may have further prognostic value in addition to specific genomic alterations, though ITH needs to be considered in the context of genomic instability rather than in isolation. Furthermore, genetic analysis of tumour thrombi showed that these are >90% similar to primary tumours, which may be a reflection of early selective sweeps or rapid progression. Analysis of metastases demonstrates these are more homogeneous than primary RCCs, consistent with early dominant clones that were 'born to be bad' or 'evolutionary bottlenecking' [75, 86]. A better understanding of tumour evolution is key to understanding how ccRCCs develop and progress, and to inform biomarker selection and interpretation.

It has been hypothesized that genetic ITH in ccRCC may hamper the identification and validation of prognostic risk scores. Hundreds of studies have been published evaluating RNA and proteins as prognostic markers in ccRCC [66]. Some of the best-known models include those by Rini et al [89], the cell cycle proliferation 'CCP' score [90], S3 score [91] and ClearCode34 [92, 93]. Gulati et al performed a systematic comparison of 28 different prognostic biomarkers/signatures from the literature [68]. Although the authors were able to validate 61% (17/28) of markers in univariate analysis, tumour stage and ClearCode34 ccB signatures were the only independent prognostic predictors in multivariate analysis. ClearCode34 represents a refined list of 34 genes, from an original study by Brannon et al [94]. In 2010, Brannon et al evaluated genome-wide gene transcription using the Agilent 44K messenger RNA microarray. They identified 110 genes (120 Agilent probes), through Logical Analysis of Data (LAD model), that discriminate between ccA and ccB subtypes [94]. This model was subsequently externally validated in six publicly available gene-expression datasets (N=480 tumours) [95]. ccA tumours demonstrated increased expression of genes associated with angiogenesis and hypoxia signalling, and had better prognosis. ccB tumours were characterised by overexpression of genes associated with growth and differentiation, including mitosis, growth factor and epithelial-to-mesenchymal transition (*TGF-b* and *Wnt* signalling). The larger list of genes identified by Brannon et al was refined into ClearCode34, which consists of a 34 gene classifier (24 genes ccA; 10 genes for ccB) [92, 93] using a centroid-based classification algorithm. ClearCode34 has gained popularity as it is relatively inexpensive and has been externally

validated in numerous studies. The validity of this model is not limited to localised disease. In fact, ClearCode34 can differentiate between good and poor prognosis in patients with metastatic RCC and slightly improve the prognostic power of the International Metastatic Renal Cell Carcinoma Database Consortium (IMDC) model in multivariable analysis [96]. However, genetic ITH confounds the use of genomic predictive models [44]. Indeed, ITH has been noted in 80% of patients evaluated using ClearCode34, meaning multi-region samples from the same patient were classified as high and low risk [68]. Similarly, ITH was noted in 60% of patients evaluated using the S3 score [91]. This demonstrates that ITH may hamper risk-score validation studies and use in clinical practice. Evaluating ITH may elucidate biological insights into tumour evolution. An ideal prognostic marker would be clonal, displaying low ITH, to maximise chances of detection and consistent tumour classification.

2.4 DNA methylation

Genomic markers may have a role in RCC diagnosis and prognosis, but are unlikely to be sufficient alone due to low recurrence rates and heterogeneity (e.g. few mutations, not present in all patients nor all multi-region samples). DNA methylation may overcome these challenges as a potential tumour marker.

2.4.1 Relevance of DNA methylation

DNA methylation in eukaryotic cells predominantly involves the addition of a methyl group to the fifth carbon in the cytosine ring, resulting in 5-methylcytosine. This epigenetic modification has a broad range of impacts on DNA pitch, secondary effects on chromatin structure, effects on binding of transcription factors, silencing of retrotransposons and effects on gene expression, without changing the underlying genomic sequence. DNA methylation also mediates gene imprinting and X chromosome inactivation. In mammals, DNA methylation most often occurs in the context of adjacent cytosine and guanine dinucleotides (CpGs), a self-complementary DNA sequence. Clusters of CpGs in the genome are referred to as CpG islands, and there are approximately 29,000 CpG islands in the human genome [97]. CpG islands are flanked by CpG shores and shelves. The following section discusses mechanisms which mediate DNA methylation modifications, temporal and spatial changes which occur in human normal and malignant cells.

DNA methylation is dynamically regulated through the addition and removal of a methyl group by DNA methyltransferases (DNMT) and the ten-eleven translocation (TET) demethylase enzymes

respectively [98]. Patterns are heritable and maintained through cell division cycles thanks to DNMT1, which copies DNA methylation from the parental strand to the daughter strand during DNA replication [99]. De-novo methylation is mediated by DNMT3 enzymes (DNMT3A and DNMT3B) in cells which are not actively proliferating [98]. De-novo methylation occurs during embryological development and cell differentiation and occurs preferentially at certain sites more than others. DNA methylation tends to be a stable process, however passive (i.e. replication dependent) or active (i.e. replication independent) de-methylation may occur. The former refers to failure to preserve methylation during cell division (i.e. methylation is not maintained). TET enzymes mediate active de-methylation through stepwise oxidation of methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Subsequently, 5fC and 5caC are excised by thymine DNA glycosylase (TDG) and unmethylated cytosines are inserted via base excision repair (BER) [100]. TET activity may be modified by chromatin accessibility, TET mutations and changes in availability of substrates and co-factors (such as Fe(II), oxygen and alpha ketoglutarate). For example, mutations in genes involved in the Krebs cycle (such as *IDH1, IDH2, SDH* or *FH*) may lead to altered alpha ketoglutarate levels and subsequent hypermethylation. In addition, methylated CpGs may be spontaneously de-aminated to thymine, which explains why there are less CpGs in the genome than expected by chance.

Epigenetic changes occur in a regulated fashion throughout the life history of a cell. De-methylation occurs during embryological development, followed by somatic de novo methylation. Methylation will determine early cell identity and fate, and changes are largely conserved. Thus, the vast majority of CpGs are stable and specific to a cell type, whilst some sites may change with ageing, environmental exposure, or the development of disease (including cancer). Conserved, stable patterns of methylation are cell lineage specific and can thus be used to identify the cell of origin [101]. DNA methylation changes with age and this is referred to as the 'epigenetic clock', similarly to clock-like mutational signatures. The mechanism of this is incompletely understood but may be secondary to exposure to environmental stressors and hazards, as well as errors in copying DNA methylation during replication [99]. A number of studies utilise methylation at CpG sites to estimate epigenetic age, which is highly correlated with chronological age, and is associated with increased cancer risk [98, 102]. In normal cells, CpG islands frequently occur at highly repetitive DNA sequences (these tend to be methylated) and near gene promoters (these tend to be unmethylated) [103, 104]. Overall, 70-80% of CpGs are methylated in normal cells [105]. Conversely, in cancer global DNA hypomethylation as well as local hypermethylation of gene promoter regions have been noted. Genome-wide hypomethylation has been linked to de-repression of retrotransposons which

may drive genomic structural variation [106]. Hypermethylation of promoters is associated with transcriptional silencing of tumour suppressor genes. Conversely, DNA hypermethylation is observed in the body of genes which are actively transcribed [105]. DNA methylation alterations in cancer are considered to be common, early and stable events, making these attractive diagnostic targets for tissue and liquid biopsies. In addition, further methylation changes may occur with cancer progression/aggressive disease, meaning there is scope for a prognostic application.

Spatial and temporal genomic analysis of DNA methylation patterns has revealed critical insights into tumorigenesis. In normal cells, adjacent CpGs demonstrate similar methylation levels (i.e. they are co-methylated) and co-methylation is determined by genomic distance along a chromosome [107]. This is believed to be secondary to the progressivity of DNMT and TET and coordinated function of these enzymes in a region-specific manner [108]. Methylation haplotype blocks are large areas of the genome which show highly concordant, cell-type specific methylation patterns in cells, which have been used to determine the cell of origin in ctDNA analysis [107]. These large blocks of concordant methylation have also been noted in cancer cells (and tend to be hypomethylated). However, certain regions may exhibit highly disordered methylation, meaning adjacent CpGs have different methylation status.

There is a growing interest in evaluating disordered methylation as it may increase our understanding of tumourigenesis and may serve as a cancer biomarker. Methylation heterogeneity has been observed in both normal and cancerous tissue, suggesting that a stochastic process may cause changes in individual CpGs, which may confer a selective advantage and drive cell development and tumorigenesis [109]. Disordered methylation has been shown to be associated with gene expression independently of average methylation within a locus [108, 110, 111]. Furthermore, methylation disorder is associated with worse prognostic outcomes in diffuse B-cell lymphoma and therefore has potential clinical relevance [112]. In this thesis, I will be evaluating methylation heterogeneity within a sample (in tumour versus normal) using the concept of epipolymorphism of epialleles. A single sequencing read enables us to evaluate the DNA methylation pattern which comes from one individual cell (i.e. an epiallele). An epiallele refers to the DNA methylation pattern seen in adjacent CpGs at an epigenetic locus (e-locus), and these changes are believed to be inherited together. It is therefore possible to evaluate the methylation pattern within a read, and between different reads derived from the same sample. If we evaluate epigenetic loci (e-loci) which consist of N adjacent CpGs, each CpG may be either methylated or unmethylated, therefore there are 2^N possible combinations of epialleles. Epipolymorphism is defined as 'the

probability that two epialleles randomly sampled from the locus differ from each other' [113]. Increased epipolymorphism recapitulates disordered methylation in a read (implying dynamic changes in DNA methylation) and disordered methylation between reads (implying epigenetic diversity from different tumour subclones and cells which compose the bulk tissue sample). Therefore, this has been proposed as a useful method to assess overall epigenetic heterogeneity within a sample. Although epipolymorphism has been evaluated in several blood born cancers and has demonstrated clinically relevant results, this has yet to be investigated in RCC.

2.4.2 DNA methylation in renal cancer

Due to the relatively low number of genetic mutations observed in RCC, there is a growing interest in DNA methylation as a tumour biomarker, as these are early events in tumorigenesis and highly recurrent sets of changes [114]. Herein, the existing evidence is reviewed regarding DNA methylation changes as diagnostic and prognostic markers in renal tumours, and current evidence regarding DNA methylation heterogeneity.

DNA methylation changes have been proposed as potential diagnostic markers to distinguish pathological subtypes of renal tumours [115, 116]. Many of the differences observed between tumour subtypes have been attributed to their cell of origin (pRCC and ccRCC derive from the proximal convoluted tube *vs* chRCC and oncocytoma from the distal nephron) [74, 117]. Compared to adjacent normal tissue, ccRCC is associated with global hypomethylation as well as a large number of hypermethylated sites. Both chRCC and oncocytoma are characterised by a high degree of hypomethylation, whilst conversely pRCC is characterised by hypermethylation [117, 118]. A relatively small number of studies have been published evaluating methylation markers to distinguish tumour types, and results are hindered by poor methodological study quality. A recently published systematic review searched >2500 articles and identified 12 studies assessing diagnostic DNA methylation biomarkers in renal tumour subtypes [115]. The studies identified in the review utilised well known tumour suppressor genes which display methylation changes in renal tumours, including *VHL*, *RASSF1A*, *CDKN2A*, *APC*, *TIMP3*, *IGFBP3*, *DAPK1*, *CDH1*, *SFRP* genes, *MGMT*, *DKK3* and *WIF1*. Many of the studies had small sample sizes (<200), used a small number of markers and lacked external validation. For example, Pires Luis et al evaluated a three gene promoter methylation panel in a cohort of 129 tumour and normal kidney samples. Using *OXR1* and *MST1R* in combination differentiated ccRCC from the other renal tumours with an AUC of 0.939, and *HOXA9*

discriminated between the latter with an AUC of 0.756 [119]. External validation was not performed, and promoter methylation was not found to be associated with cancer specific survival. Perhaps the most rigorous methodology was observed in a study by Chopra et al [120] and a study by Brennan et al [121]. Chopra et al developed a DNA methylation classifier based on 59 CpGs (2 for angiomyolipoma, 9 for oncocytoma, 11 normal kidney, 13 ccRCC, 14 pRCC and 10 chRCC) using data from a training cohort of 429 samples. The classifier was externally validated on 272 ex-vivo biopsies from 100 renal nephrectomy specimens (71% malignant) and was able to predict the correct pathological subtype in 85% of cases. Subsequently, Brennan et al developed a classifier based on 79 CpGs to distinguish oncocytoma and chRCC, which achieved an AUC of 0.87 in external validation (N=102 samples) [121]. Though promising, it is likely that a higher model accuracy would be necessary prior to adoption in clinical practice. In summary, the existing evidence suggests that DNA methylation changes could be useful biomarkers to differentiate tumour subtypes in a diagnostic setting and that further research into this topic, using rigorous methodology and larger sample sizes, is warranted.

As with many cancer types, ccRCC is characterised by global hypomethylation and promoter hypermethylation in hundreds of genes [114, 122]. Silencing of the *VHL* gene via promoter hypermethylation has been observed in approximately 15% of individuals with ccRCC [123], while promoter methylation of the *CDKN2A* tumour suppressor gene occurs in ~4% of all RCC subtypes and is linked to poor prognosis [70]. Promoter hypermethylation is also noted at a number of other renowned tumour suppressor genes such as *TP53*, *MHL1*, *CDH1*, *APC*, *UQCRH* and *RASSF1A*, with accompanying reductions in gene expression [70, 122]. Furthermore, several key tumorigenic pathways are dysregulated by promoter hypermethylation in RCC, including: the *Wnt/β-catenin* pathway (including *SFRP*, *DKK*, *IGFBP*, *HGF*, *WIF1* and *MET* genes), *TGFβ* signalling (*GATA-3*, *GREM-1*, and *SMAD-6*), pathways regulating cell adhesion and epithelial-to-mesenchymal transition (EMT), cell cycle (*RASSF1*, *KILLIN*), apoptosis (*APAF1*), genomic stability (*MGMT*) and angiogenesis (*TIMP3*) [124]. Hypermethylation is also noted at genes involved in cell metabolism and homeostasis, such as *VHL*, *SDHB*, *FLCN*, *SLC16A3*, *CDO1* and *AMPK* signalling pathway [122]. These are the same genes and pathways which are also disrupted by mutations in RCC.

DNA methylation changes may also serve as potential prognostic markers in ccRCC [24, 103]. A higher proportion of promoter hypermethylation has been associated with higher tumour grade and stage [70]. A number of literature reviews, systematic reviews and meta-analyses have been published evaluating the prognostic potential of methylation markers in ccRCC [67, 103, 125]. A

review by Joosten et al identified methylation of the promoter region of 9 genes (*BNC1*, *SCUBE3*, *GATA5*, *SFRP1*, *GREM1*, *RASSF1A*, *PCDH8*, *LAD1* and *NEFH*) as having prognostic potential in more than one study [67]. Unfortunately, prognostic studies suffer from similar pitfalls as those mentioned for diagnostic methylation studies, namely heterogeneous methods, small sample sizes, lack of external validation and reproducibility. Wei et al developed a five-CpG-based classifier which could be applied to nephrectomy specimens to predict overall survival in ccRCC patients following surgery [126]. The score was calculated using a LASSO Cox regression model based on methylation at 5 CpGs in the promoter region of *PITX1*, *FOXE3*, *TWF2*, *EHBP1L1* and *RIN1*. The median score in the cohort was used as a cut-off and patients with a risk score above the median value tended to have worse survival. The authors validated their score in three independent datasets (one of which was TCGA) and demonstrated reproducible results. However, Evelönn et al failed to validate the Wei model in their cohort of non-metastatic patients [127]. It is postulated that evaluating global methylation changes and large panels of markers may prove to be superior to individual CpGs to predict prognosis reliably and reproducibly. The CpG island methylator phenotype (CIMP) is one such example, where CIMP-positive tumours are characterised by genome-wide hypermethylation in CpG islands and poor prognosis. CIMP has been most extensively studied in colorectal cancer [128], but has been demonstrated to have prognostic value in a pan-cancer setting [129-131]. Arai et al performed the first study evaluating CIMP in ccRCC tissue (N=109) [132]. The authors identified a small subset of ccRCC samples (13%, N=14/109) which were CIMP-positive and these were associated with aggressive clinical parameters: higher tumour size, grade, stage, presence of necrosis, tumour thrombus, invasion of vessels and renal pelvis. The prognostic value of CIMP has been validated in independent cohorts of ccRCC samples [24, 133]. In addition, TCGA demonstrated that CIMP-positive tumours, characterised by hypermethylation in 1532 CpGs and poor prognosis, are present across different pathological subtypes of renal tumours (ccRCC, chRCC and pRCC) [24].

Although genetic heterogeneity has been extensively studied, there are very few data regarding methylation heterogeneity in ccRCC. This is an important biological question that will contribute to our understanding of tumour evolution and will also inform the selection and interpretation of RCC methylation markers. A genome wide study performed in 30 human cell lines and tissue types using whole genome bisulphite sequencing (WGBS) suggested that DNA methylation variability between tissues exceeds variability within tissues by one order of magnitude [134]. To our knowledge, there are only three studies which assessed methylation ITH in ccRCC [127, 135, 136]. For example, Evelönn et al aimed to explore prognostic methylation markers in ccRCC [127]. Their study also included data on 5 patients with multi-region samples: 2 or 3 tumour samples were available for

each patient, however matched normal data were not presented. Although methylation ITH was not the primary aim of their analysis, they showed that methylation patterns were similar in multi-region samples derived from the same patient (R^2 correlations 0.97 to 0.99). Similarly, Winter et al [136] evaluated multi-region samples from 3 patients (2 primary tissue samples each) and confirmed similar correlations between multi-region tumour samples from the same patient (R^2 >0.96). Stewart et al [135] demonstrated that in patients with mRCC, renal biopsy samples and multi-region nephrectomy samples (following sunitinib therapy) cluster together. None of the three studies evaluated sequence level methylation data, none evaluated >3 samples per patient and none attempted to reconstruct methylation phylogenies. Therefore, my thesis aims to address these points and systematically characterise methylation heterogeneity between patients, within a patient and within each sample, using sequence level data.

2.4.3 Methods used to evaluate DNA methylation

A number of experimental methods are available to sequence DNA methylation data. The main challenge is that traditional Sanger based sequencing cannot distinguish between methylated and unmethylated cytosines. Therefore, several techniques have been developed to convert methylated and unmethylated cytosines prior to sequencing to enable differentiation, including bisulphite sequencing (BS) and enzymatic conversion, both of which are used in this thesis. Bisulphite conversion is widely considered as the gold standard methodology. The bisulphite conversion reaction de-aminates unmethylated cytosine residues to uracil, whilst methylated cytosines remain unaffected. Following Polymerase Chain Reaction (PCR) amplification, uracils are copied as thymines, meaning the resulting libraries consist of two families of dsDNA molecules (originating from Watson and Crick strands), with a high thymine to cytosine ratio (Figure 2.2). The main drawback of BS is it consists of a biochemically harsh method (low pH and high temperature) resulting in ~90% DNA degradation, loss of starting material and low complexity libraries. BS is therefore routinely used to evaluate surgical tissue samples which result in high DNA yields (for example nephrectomy specimens), however it remains challenging where there is low DNA input such as biopsy samples or circulating tumour DNA. In addition, BS is unable to differentiate between 5mC and 5hmC, though this can be achieved by performing additional oxidative bisulphite conversion (OXBS) and subtracting results from the two methods [137]. 5hmC represents a marker of de-methylation, as well as being regarded as a potential cancer biomarker itself, due to its role in tumorigenesis and regulation of gene expression [138, 139]. 5hmc is now recognised as a stable epigenetic mark, with tissue-specific changes noted, and reduced 5hmc is considered an early cancer marker, including in RCC [139, 140].

An alternative method to BS (which I used in Chapter 5 and Chapter 6), which results in less sample loss is enzymatic conversion using TET2 and APOBEC (used in Chapter 7). TET2 oxidises methylated cytosines, subsequently APOBEC deaminates unmethylated cytosines into uracil whilst sparing methylated cytosines due to the previous TET2 action (Figure 2.2). Several other methods are currently being investigated in the literature to sequence DNA methylation data, including alternative conversion methods (TET-assisted or chemical-assisted pyridine borane sequencing; TAPS or CAPS) and Oxford Nanopore sequencing (which does not require conversion), although these are outside the scope of this thesis [141, 142].

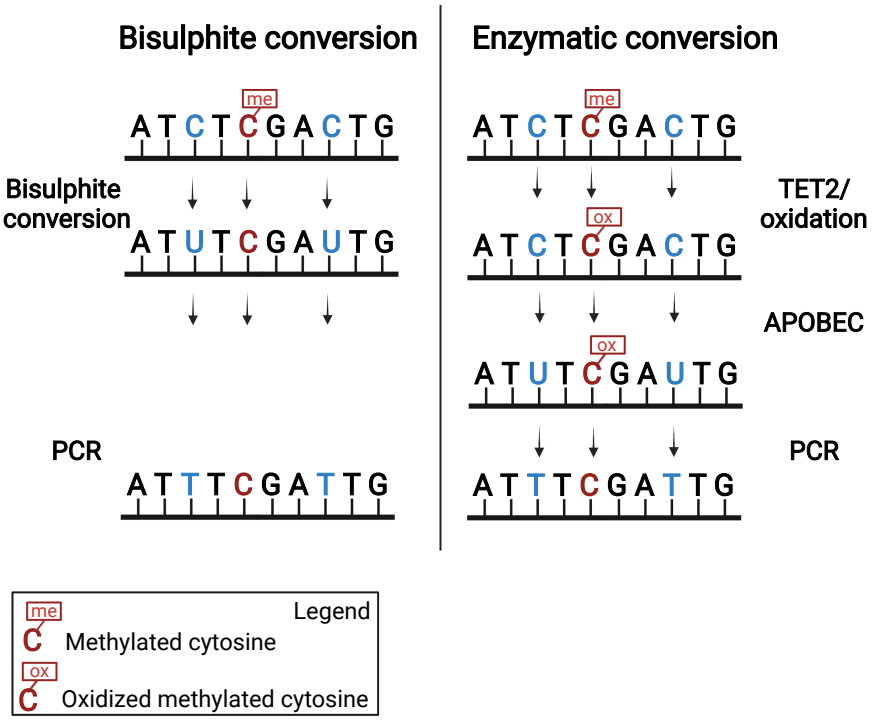


Figure 2.2: Bisulphite and enzymatic conversion

The bisulphite conversion reaction de-aminates unmethylated cytosine residues to uracil, whilst methylated cytosines remain unaffected. In the enzymatic conversion method, TET2 oxidises methylated cytosines, subsequently APOBEC deaminates unmethylated cytosines into uracil whilst sparing methylated cytosines due to the previous TET2 action. For both methods, during PCR amplification, uracil is converted to thymine resulting in a library with a high thymine to cytosine ratio.

Following conversion, experimental methods may consist of sequencing the entire genome (e.g. WGBS) or capturing regions of interest (e.g. arrays or Epic-seq). WGBS is prohibitively expensive and inefficient as 65% of sequencing reads do not contain any CpGs and are thus non-informative [134]. Array based technologies (such as the Illumina 27k, 450k and 850k arrays) are widely used in the literature, for example in The Cancer Genome Atlas. They are cost-effective but provide limited data (i.e. relatively small number of CpGs covered outside of CpG islands, no information regarding sequence context and are more error prone). Epic-seq, the method I used in Chapter 5 and Chapter 6, is a capture-based method which produces sequence level data on approximately 3 million CpGs. The main advantage is the ability to evaluate the entire sequence within a read, therefore assessing epipolymorphism and differentially methylated regions, rather than simply evaluating individual CpGs. A potential limitation is that Epic-seq and the 450k array cover different CpGs (Epic-seq covers >90% of the 450k targets, but the overlap between Epic-seq and 450k array is only ~13%), which may limit validation of findings across different platforms. The Nimbus method, which I used for cell free DNA methylation detection in liquid samples is another capture-based method (Chapter 7). I designed the capture panel by selecting informative regions to maximise effectiveness and reduce costs.

Several quality control metrics are used to demonstrate the effectiveness of DNA methylation sequencing methods. Following both bisulphite and enzymatic conversion, unmethylated cytosines are converted to thymines whilst methylated cytosines are read as cytosines. Therefore, the percentage methylation at a particular site is calculated by dividing the number of cytosines over the total number of cytosines and thymines at that site (i.e. number of cytosines / cytosines + thymines). The percentage methylation in non-CpG contexts (i.e. CHG and CHH methylation) is expected to be <1% in the human genome, therefore this is a useful metric of under-conversion in both bisulphite and enzymatic based sequencing methods [143, 144]. For example, in BS, under-conversion refers to incomplete de-amination of unmethylated cytosines, meaning they are read as cytosines (as if they were methylated), leading to false positive methylation calling. Under-conversion may occur due to incomplete dsDNA strand denaturation [143]. Over-conversion is more difficult to quantify in the absence of a known spike-in sequence. Other useful metrics for quality control include: mapping and duplication rates, coverage, on-target and off-target rates. Mapping is affected by the high thymine content post library preparation in BS, duplication rates and coverage are affected by DNA damage resulting in low yields, necessitating high PCR cycles. The on- and off- target rates enable quantification of the efficiency of the capture method. A caveat of DNA methylation analysis in bulk tissue is that the percentage methylation may be affected by copy number aberrations and tumour

purity (i.e. contamination by different cell types). Methylation levels may be evaluated at individual CpGs or regions consisting of adjacent CpGs in order to identify differentially methylated cytosines (DMCs) or differentially methylated regions (DMRs) respectively. In summary, this section introduces the experimental techniques and the analysis methods used in this thesis.

2.5 Cell free DNA as a novel tumour marker

As discussed in the previous sections, there is a real clinical need for the identification of biomarkers that differentiate benign from malignant SRMs and aggressive from indolent disease. A number of potential biomarker candidates have been investigated; however, none have been adopted in clinical practice [19, 67-69, 89, 92, 93, 126, 145]. Good practice guidelines and biomarker roadmaps have been published as a response, to encourage rigorous study design and methodology [146-148]. The ideal RCC biomarker would be sensitive and specific, tested non-invasively, using readily available and inexpensive methods. Epigenetic markers offer the benefits of high specificity and digital quantification, building on recent technological developments in quantitative genomics.

Circulating cell free DNA (cfDNA) has recently attracted interest from the translational research community due to its potential role as a non-invasive biomarker in multiple applications, including tumour detection. cfDNA consists of small fragments of DNA that originate from cells and can be found in bodily fluids such as plasma or urine, reflecting the cell-of-origin's genome and epigenome. cfDNA is believed to be released through cell apoptosis and necrosis, in addition to active secretion [149], although the dominant mechanism has not been characterised for different cell types in health and disease. cfDNA has a short half-life (0.25-2.5 hours) and clearance is believed to occur via: nuclease degradation, renal excretion into urine and degradation by macrophages in the liver and spleen [149]. The fragment size distribution of cfDNA molecules has a mode at 166bp, with additional peaks at multiples of 166bp, representing mono-, di- and tri-nucleosomes (DNA wrapped around a nucleosome plus linker DNA is 166bp). Fragment size is also believed to reflect the underlying biology of cfDNA release and nuclease degradation, with short fragments representing apoptosis or increased degradation and longer fragments secondary to necrosis or exosome release.

Tumour derived cfDNA (ctDNA) analysis may be used as a 'liquid biopsy' to sample the tumour genome and epigenome non-invasively and as such is a promising approach for cancer detection. Liquid biopsies have a myriad of applications, including: diagnosis, early detection, molecular profiling to tailor drug therapy, detection of minimal residual disease, monitoring treatment

response and disease progression. Furthermore, recent methodological advances and increased availability of next generation sequencing technologies mean that sequencing ctDNA in large cohorts for early detection is now more feasible [150]. These approaches offer high analytical sensitivity and specificity, at affordable prices, with the potential for widespread clinical application [149]. Evaluation of DNA methylation markers in ctDNA has a number of advantages over genomic markers. DNA methylation changes occur more frequently than gene mutations in RCC and therefore represent an attractive target for liquid biopsy assays. DNA methylation patterns are also cell-type specific and can determine the tumour cell-of-origin [151]. In addition, liquid biopsies using ctDNA represent particularly attractive avenues in RCC as this may overcome sampling bias due to mutational ITH seen in conventional tissue biopsies.

However, significant challenges exist for the detection of ctDNA, particularly in the context of cancer early detection (Figure 2.3). In patients with early-stage cancer the vast majority of cfDNA in the plasma (approximately 80%) and urine (median 52%; up to 93%) is derived from haematopoietic cell lysis, with only a low fraction derived from the tumour (often <1%) [152]. Levels of ctDNA may also be affected by exercise, the circadian rhythm and hydration status. Sensitivity is also limited by low absolute counts of ctDNA molecules. It has been estimated that there are approximately 2000 genome equivalents of cfDNA per millilitre of plasma from patients with early-stage cancer, therefore the theoretical limit of detection of a single mutation is less than 1 in a few thousand [153]. Bioinformatic simulations have demonstrated that the probability of detecting ctDNA increases by increasing the number of targets analysed at high sequencing coverage [154]. However, there are a limited number of recurrent somatic mutations in RCC that could be used as prospective targets for ctDNA analysis [24]. We therefore postulated that DNA methylation may be an ideal target for ctDNA detection, as methylation changes are abundant (several thousand markers) and often early events [154]. This prompted our laboratory to investigate targeted methylation analysis of ctDNA for cancer detection. In the last few years, a number of studies have been published evaluating this topic. In a head-to-head comparison of ctDNA mutational versus methylation analysis, DNA methylation provided superior detection rates and cell-of-origin localisation in a pan-cancer cohort [151]. GRAIL, a large biotech company, are currently evaluating their pan-cancer DNA methylation screening test, termed the Galleri test, in a large-scale trial within the NHS and are aiming to recruit 165,000 participants [150]. However, ctDNA detection rates using the Galleri test in patients with non-metastatic RCC were very low (sensitivity 0% in stage I, 25% in stage II and 50% in stage III) [151]. Levels of ctDNA vary between cancer types and reports consistently suggest that patients with RCC may have lower ctDNA levels than other malignancies [153, 155-157]. It is hypothesized that low ctDNA detection may be secondary to reduced secretion (reduced tumour

proliferation/shedding or low tumour vascularity), increased degradation or assays with low sensitivity. Low ctDNA detection in ccRCC is therefore surprising as these tumours are generally well vascularised and often characterised by necrosis [158]. Low detection rates in ccRCC remain unexplained, though a number of studies (described in more detail in Chapter 7) have shown more promising results [154, 159].

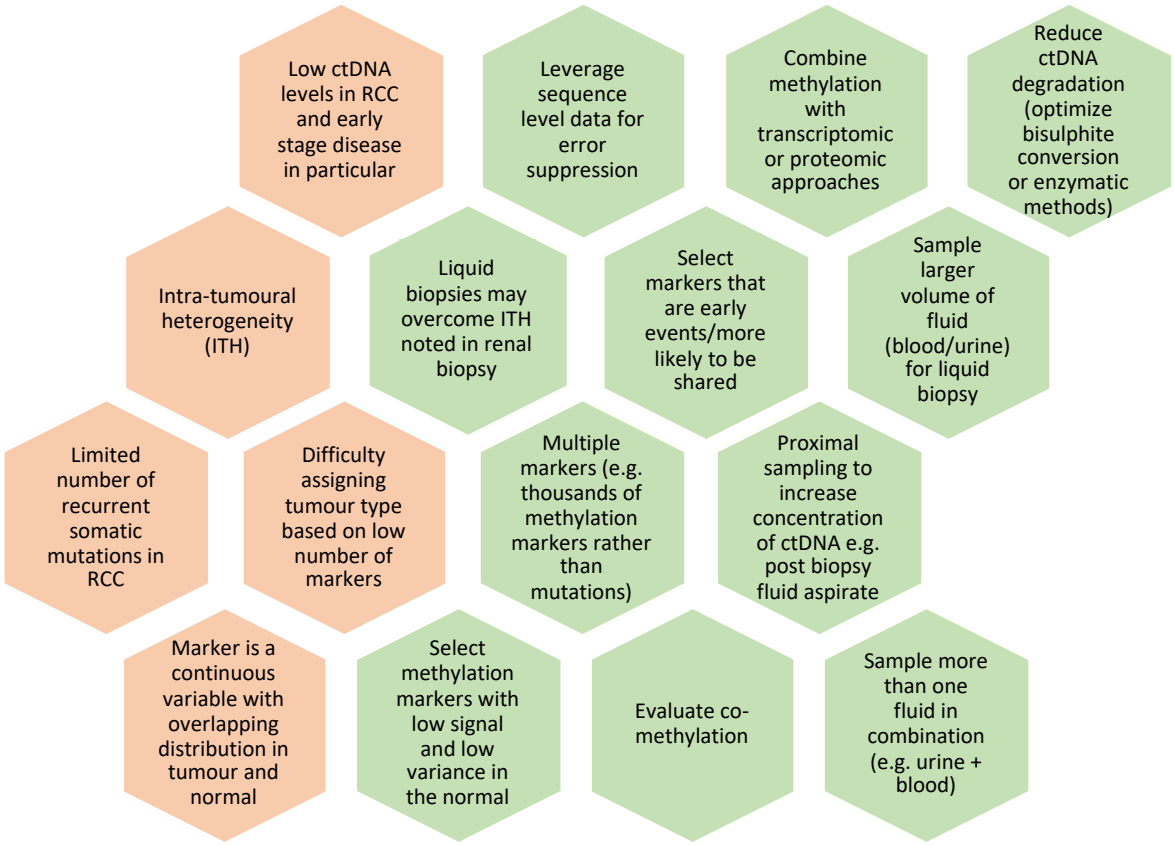


Figure 2.3: ctDNA detection in RCC: challenges and potential solutions

Potential challenges regarding biomarker development in RCC and in particular ctDNA detection (shown in red), along with strategies to overcome these (shown in green).

A number of strategies can be adopted to increase ctDNA detection rates. Leveraging sequence level data (for error suppression) and multiple markers (e.g. thousands of methylation DMRs, potentially combined with transcriptomic or proteomic approaches) may further increase detection rates in low burden disease. Tumour derived ctDNA may also be enriched via size selection (either in vitro or in silico), leveraging the observation that tumour derived ctDNA is shorter than cfDNA [160]. Fragmentomic analysis may also enable ctDNA enrichment, for example by analysing ctDNA end fragments and nucleosomal footprinting [161, 162], as these methods can determine the cell of origin. Sampling error may also be reduced by evaluating more than one fluid (e.g. both plasma and urine), a larger volume of fluid (e.g. 10ml instead of 2ml plasma; 100-200ml instead of 2ml urine) or sampling at multiple time points. Although these approaches are generally feasible, it is important to minimise the burden on the patient and overall costs. Furthermore, the abundance of ctDNA could be increased by seeking 'alternative' sampling methods, such as proximal sampling (Figure 2.3).

Proximal sampling, defined as 'non-tissue sampling of the cancer-bearing organ as close to the organ at risk as possible', has been proposed as a method to increase detection in many cancer types [163]. Sampling closer to the organ of interest is hypothesized to be associated with higher levels of ctDNA (as well as other cancer markers) which may be low in more distal samples such as peripheral venous blood, voided urine or stool. Proximal samples may have varying degrees of proximity to the tissue of origin and levels of invasiveness. Samples which have been evaluated for gynaecological malignancies include ovarian cyst fluid, uterine cavity lavage, cervical smears and vaginal tampons [163]. In other cancer types, these include cerebrospinal fluid for neurological cancers, and for lung cancer: breath, saliva, sputum, bronchial lavage and pleural fluid [163]. Circulating tumour cells (CTCs), though not the focus of this thesis, demonstrate the value of proximal sampling. CTCs in the peripheral blood stream are rare, however CTC detection is increased in proximal blood (i.e. such as the pulmonary vein for lung cancer or hepatic vein for hepatocellular carcinoma) and may be clinically useful [164]. CTCs were detected from pulmonary vein samples in 48% (48/100) of lung cancer patients in the TRACERx study, and detection was an independent predictor of recurrence after adjusting for tumour stage [165]. Although CTCs have been investigated in RCC, studies have provided inconsistent results (and proximal sampling has yet to be evaluated) [166]. Perhaps one of the most promising proximal tests for ctDNA is PapSEEK, which aims to detect endometrial and ovarian cancers by evaluating aneuploidy and mutations in 18 genes in cervical fluid [167]. The assay produced improved cancer detection rates with more proximal samples (i.e. use of a Tao brush for intrauterine sampling vs Pap brush for endocervical sampling). In addition, cancer detection rates were higher when evaluating both peripheral blood and cervical fluid, than evaluating one sample

type alone (i.e. in some patients, ctDNA was detected in one sample type and not the other). This suggests multi-sample approaches may be a promising strategy. This data is consistent with results from Smith et al, who evaluated ctDNA detection rates in patients with ccRCC using mutational analysis [158]. Smith et al demonstrated that evaluating both plasma and urine ctDNA improves detection rates over assessing one sample type alone (i.e. once again, in some patients, ctDNA was detected in one sample type and not the other). At present it is unclear if this observation is simply due to the limit of detection or whether this is secondary to tumour biology (for example increased ctDNA degradation/clearance in certain sample types compared to others). Evaluating multiple marker types (such as ctDNA and protein in combination) has been shown to increase detection rates further [168]. To our knowledge, very limited research has been published assessing the value of proximal sampling in renal cancer. A handful of studies evaluated CA9 levels in renal cyst fluid to differentiate benign from malignant cysts [169], however no studies have been performed evaluating ctDNA in proximal body fluids. Therefore, one of the approaches investigated in this thesis is the use of proximal sampling to increase ctDNA detection rates in patients with RCC (Chapter 7). Taken together, these data suggest the value of exploring DNA methylation markers in cfDNA derived from patients with RCC.

Chapter 3 Rationale and thesis aims

In summary, I have highlighted two complementary clinical research priorities: (a) to improve the characterisation/diagnosis of SRMs and (b) differentiate indolent versus aggressive RCC. Though a number of genomic biomarker studies have been developed to evaluate diagnosis and prognosis in renal tumours, these are hampered by genetic intra-tumoural heterogeneity (ITH). DNA methylation markers are abundant, often early events in tumorigenesis, which are specific to the cell of origin. Very little is known regarding methylation ITH, though it is postulated this may be less pronounced than for genomic aberrations. An improved characterisation of tissue heterogeneity would enable us to identify methylation changes which are shared in the majority of multi-region samples, are more likely to represent early/stem events in RCC development and therefore more likely to be clinically useful markers in both renal and liquid biopsies. ctDNA detection represents a promising non-invasive liquid biopsy in both these diagnostic and prognostic clinical settings. In the former, liquid biopsy may be preferable to invasive renal biopsy, whilst in the latter, ctDNA may allow monitoring of minimal residual disease post-operatively, risk stratification and early detection of recurrence. RCC is characterised by a low mutational burden, which hampers mutational analysis of ctDNA and leads to low detection rates, particularly in earlier stage disease, which may be overcome by evaluating thousands of DNA methylation markers. Therefore, in this thesis I aimed to characterise DNA methylation patterns in patients with renal tumours to offer insights into disease biology and translate these findings into clinically relevant biomarkers (in renal and liquid biopsies) which may address the two clinical questions, namely improved diagnosis and prognosis.

In Chapter 5, I characterised DNA methylation in tissue obtained from patients with common pathological subtypes of malignant and benign renal tumours and adjacent normal kidney. I integrated DNA methylation data with information regarding gene expression to elucidate biological similarities and differences amongst pathological subtypes. Lastly, a machine learning model was developed to diagnose subtypes of renal tumours in a clinical setting, and I explored the potential impact of methylation ITH on model output. In Chapter 6, I assessed methylation ITH in tissue obtained from patients with ccRCC, evaluating implications regarding tumour evolution and highlighting learning points for biomarker selection. Chapter 7 evaluated DNA methylation in ctDNA derived from patients with ccRCC and controls who are cancer-free. I postulated that ctDNA may be enriched in proximal samples, a strategy which may be useful in patients with SRMs, as they are expected to have low plasma ctDNA levels. I therefore quantified ctDNA detection rates using methylation versus mutational analysis and compared distal versus proximal samples.

The hypotheses explored in this thesis are:

1. DNA methylation in renal tissue may help differentiate benign and malignant renal tumours
2. DNA methylation heterogeneity in renal tissue may be associated with clinical and prognostic factors in patients with ccRCC
3. Targeted methylation analysis may enable detection of ctDNA in liquid samples from patients with ccRCC

As such, the overarching thesis aims are:

1. Map DNA methylation profiles in tissue from patients with benign and malignant renal tumours to identify markers that may improve differentiation of small renal masses in a diagnostic setting
2. Characterise DNA methylation heterogeneity in tissue from patients with ccRCC
3. Explore ctDNA detection using a targeted methylation approach in patients with ccRCC versus controls

Chapter 4 Materials and methods

This chapter describes materials and methods. All work was conducted by me, unless specified, and all collaborator contributions are acknowledged in this chapter.

4.1 Samples

4.1.1 Patient nephrectomy tissue samples

A cohort of patients with benign and malignant renal tumours was identified from two biobanking studies at Addenbrooke's Hospital: 'Discovery and analysis of novel biomarkers in urological diseases' (DIAMOND; REC ID 03/018) and 'A Translational research Approach to development of optimal Renal cancer Treatments In Surgical and systemic Therapy patients' (ARTIST; REC ID 20/EE/0200). Ethical approval and patient consent were obtained. All participants were assigned anonymous IDs. For tissue analysis, samples (tumour and adjacent normal) were obtained from patients undergoing curative or cytoreductive nephrectomy between 2010 and 2018. Tissue samples were collected by the pathology team at Addenbrooke's Hospital (as described below), embedded in OCT (optimal cutting temperature) compound, sectioned and stored at -80°C. Subsequently, I received fresh frozen tissue specimens directly from the Tissue Bank.

In a subset of patients, multi-region tumour samples were collected by the pathologist along with adjacent normal kidney tissue (Figure 4.1). For samples taken post 2016, 'true' multi-region samples were collected from nephrectomy specimens using a 6mm core biopsy puncher and tumour maps delineating the location of multi-region sampling were available. For samples prior to 2016, multiple slices of renal tissue were obtained using a scalpel, however unfortunately maps were not available. In addition, several patients had large tissue slices that were subdivided into samples 'a' and 'b' representing distinct tissue samples from the same area of the kidney, but only a few millimetres apart. This is useful as it enables us to assess the similarity of samples which are spatially very close and are therefore expected to be similar.

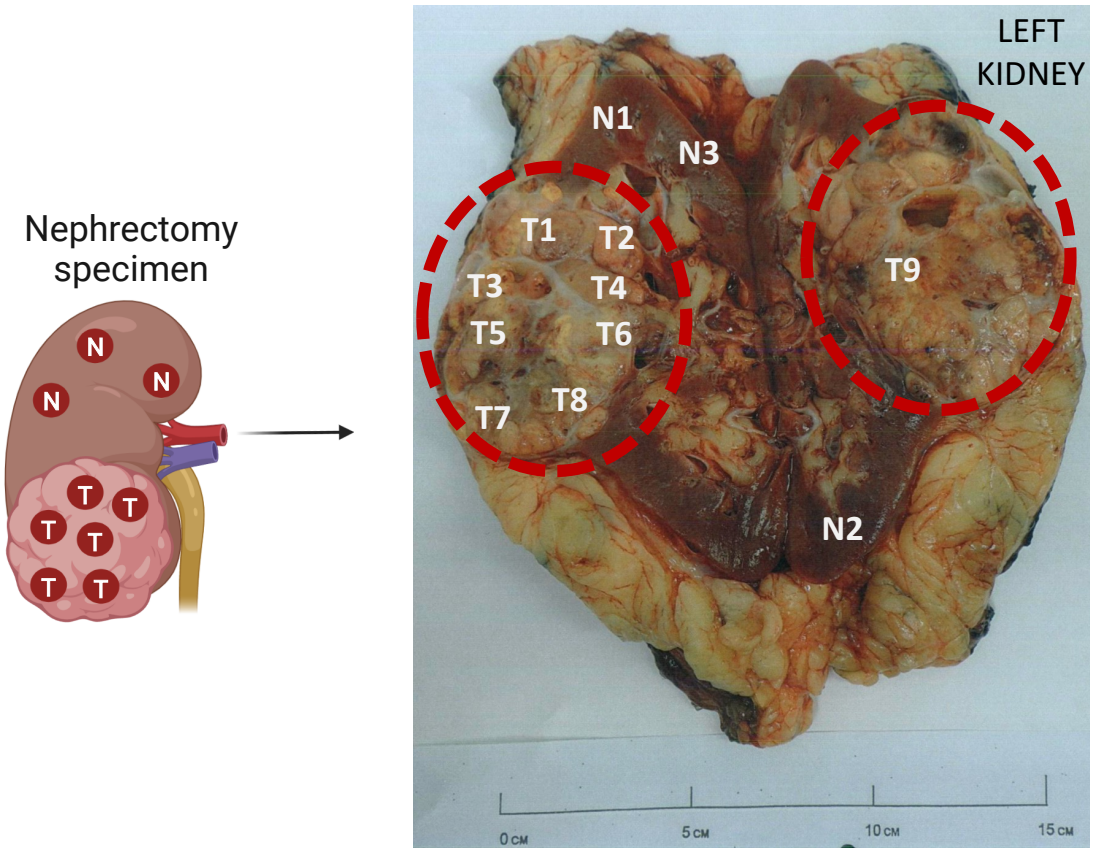


Figure 4.1: Multi-region kidney tissue sampling

The figure on the left is a cartoon schematic demonstrating multi-region sampling in a nephrectomy specimen, whilst the figure on the right demonstrates how this was done in practice. Thus, multiple tumour and normal kidney samples were collected for each patient by the pathology team at Addenbrooke's Hospital and I received fresh frozen tissue specimens.

4.1.2 Patient liquid samples

Plasma samples were obtained from patients with and without ccRCC, who were enrolled in the DIAMOND (REC ID 03/018) and ARTIST (REC ID 20/EE/0200) Biobanks at Addenbrooke's Hospital. Patients without ccRCC will be referred to as 'controls' hereafter. The choice of controls was purely based on sample availability. Controls consisted of men with raised prostate specific antigen undergoing clinical investigation for prostate cancer. Patients underwent prostate fusion biopsies (MRI planned, ultrasound guided) and no evidence of prostate cancer was found. The absence of renal and prostate cancer was confirmed on clinical follow up.

4.1.2.1 Feasibility study on proximal versus distal sampling

I organised and led a small study evaluating the feasibility of collecting and analysing proximal and distal samples. Matched samples were collected from 11 patients undergoing diagnostic renal biopsy, including tissue core biopsy, post-biopsy fluid (proximal sample; defined below) and plasma (distal sample). All patients were enrolled in the DIAMOND (REC ID 03/018) or ARTIST (REC ID 20/EE/0200) Biobanks at Addenbrooke's Hospital. Core biopsies were undertaken by a radiologist using ultrasound guidance. European Association of Urology (EAU) guidelines recommend that renal biopsies are taken using an 18-gauge needle via a co-axial sheath, which allows multiple biopsies to be taken from the same puncture site [27]. Following biopsy sampling, the co-axial sheath was aspirated using a 10ml syringe to obtain blood-stained fluid. This will be referred to as post-biopsy fluid for the remainder of the thesis. In cases where <1ml of post-biopsy fluid was obtained, sterile saline (variable volumes) was used to 'flush' the syringe to reduce the amount of post-biopsy fluid which would have otherwise remained in the syringe and have been discarded.

4.1.2.2 Liquid sample processing

Whole blood was collected and processed to obtain plasma, using different methods depending on the date of collection (due to changes in the DIAMOND Biobank standard operating procedure). For samples obtained prior to April 2016, 8ml of blood was collected into an EDTA tube and centrifuged at 2700g for 20 minutes, within 1 hour of collection. The plasma was aliquoted into cryotubes ('single-spun plasma'), while the buffy coat was stored in sterile 2ml microfuge tubes. After April 2016, 2 x 6ml of whole blood was collected into EDTA tubes, and centrifuged at 1600g for 10 minutes, within 1 hour of collection. Subsequently, 1ml plasma was aliquoted into 2ml RNase-free microfuge tubes, and these were spun at 14,000g for 10 minutes. The supernatant was then

transferred into 2ml sterile microfuge tubes and stored at -80˚C ('double-spun plasma'). The first, slower spin aims to separate plasma, whereas the second, faster centrifugation step aims to remove material from lysed cells [170].

For 11 patients recruited to the proximal sampling study, whole blood (2 x 6ml) and post-biopsy fluid (variable amount) were stored in Streck tubes. Streck tubes, a type of cell-stabilizing blood collection tubes, contain a proprietary preservative which stabilizes cell free DNA (cfDNA) for up to 2 weeks at room temperature [171]. This aims to limit cell lysis, therefore reducing contamination of cell free DNA from genomic DNA, and inhibit nuclease mediated degradation of cfDNA. Samples stored in Streck tubes were processed within 72 hours using the same 'double-spin' protocol as plasma samples obtained after April 2016. I performed liquid sample processing for patients recruited to the proximal sampling study, and all other specimens were processed by the DIAMOND Biobanking team at Addenbrooke's Hospital. A systematic comparison of blood collection tubes and processing protocols has previously been performed by the Rosenfeld group, Cancer Research UK Cambridge Institute (CRUK CI), using samples obtained from cancer patients [170]. Equivalent circulating tumour DNA yields (measured as mutant copies per ml of plasma) were achieved using cell-stabilizing blood collection tubes and EDTA tubes, provided the latter were processed within one hour of collection (as was the case in my study). Unpublished work conducted by Dr Chris Smith (Rosenfeld Group, CRUK CI) demonstrated that detection rates of circulating tumour DNA in patients with RCC were not significantly different using 'single-spun' versus 'double-spun' processing protocols.

4.1.2.3 Clinical data

Clinical annotation, including patient details and tumour pathological characteristics, was performed by retrospective review of prospectively maintained hospital records. Unfortunately, clinical data were only available for a subset of patients included in the analysis at the time of writing (although further data have been requested). The following parameters were obtained from the nephrectomy pathology report: tumour size (maximal diameter in cm), Fuhrman grade, pathological stage and presence of necrosis. Cell proliferation was estimated by Dr Anne Warren, histopathologist at Addenbrooke's Hospital, by evaluating ki67 immunohistochemical staining in formalin fixed paraffin embedded tissue slides, as previously described [158]. An anti-Ki67 monoclonal antibody (MIB-1 clone at 1:100 dilution; DAKO Agilent Technologies LDA) was used and staining was assessed in 20 individual high-power fields (>6000 tumour cells in total) per patient, at 400x magnification. For each

of the 20 regions, ki67 staining was graded manually as 0%, 1%, 10%, 30%, 75%, and 100% of positive cells, and these were converted into a score of 0, 1, 2, 3, 4, and 5 respectively. The 'ki67 sum' score was derived by adding the individual values for all 20 regions, thus obtaining a score from 0 to 100 [158]. Patient details included age, sex and body mass index (BMI) at the time of sampling. Leibovich score and recurrence data were available for a subset of patients (≥4 years follow up for every patient). Recurrence was defined as local relapse at the nephrectomy site or new metastases on CT imaging on clinical follow up.

4.1.3 Cell lines

Cell line genomic DNA (gDNA) and cell line supernatant were collected to represent a model system of gDNA in kidney tissue and cfDNA respectively. DNA was obtained from the following cell lines: HK2 (representing normal kidney), 786-O and 786-M1A (representing ccRCC tumours). Cell line choice was determined by sample availability. HK2 is an immortalized epithelial cell line derived from the proximal tubule of the cortex of a normal adult human kidney [172]. 786-O is one of the top three most cited and most well characterised RCC cell lines. The cell line was originally derived from a 58 year old Caucasian man with primary ccRCC and widespread metastases, and is characterised by a homozygous mutation in the VHL gene [173]. The 786-M1A cell line is a first-generation metastatic derivative of 786-O, representing an aggressive phenotype compared to the parental cell line [174, 175]. 786-O cells were injected into the tail vein of mice and cells isolated from rapidly growing lung metastases yielded the 786-M1A cell line. These cells have a 100x predilection for lung colonization and demonstrate both epithelioid and sarcomatoid differentiation [174, 175].

gDNA from two HK2 cell clones (clones A and B) was kindly provided by Dr Christina Schmidt (Frezza Group, Hutchison MRC Institute) (ATCC cat. No. CRL-219), whereas gDNA from 786-O and 786-M1A cells was provided by Dr Paulo Rodrigues (Vanharanta Group, Hutchison MRC Institute). HK2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Life Technology cat. no. 41966) supplemented with 10% v/v fetal bovine serum (FBS) in an incubator at 37°C with 5% CO2. 786-O cells and 786-M1A were cultured in DMEM/F12 supplemented with B27 (Invitrogen), streptomycin (µg/ml), and EGF/FGF (Peprotech 10 ng/mL). Cell line supernatant from HK2 and 786-M1A cell lines was provided by Dr Marco Sciacovelli (Frezza Group, Hutchison MRC Institute). The cells were grown in a T75 flask in RPMI and 10% FBS in an incubator at 37°C with 5% CO_2. The medium was collected when cells were subconfluent and centrifuged at 500g for 10 min to remove dead cells. Following

this, the cell line supernatant was double spun using the same protocol as for plasma samples obtained after April 2016 (i.e. 1600g for 10 minutes followed by 14,000g for 10 minutes).

All cell lines were authenticated using short-tandem repeat genetic profiling. The CRUK CI Cell Services Core Facility performed genotyping and data analysis using the Applied Biosystems Gene Mapper 5 software. The percentage match to the reference profile was 93% for the HK2 cell line, 100% for 786-O cells and 100% for 786-M1A.

4.2 Experimental methods

4.2.1 Nucleic acid extraction from tissue and liquid samples

Nucleic acid extraction from tissue was performed by different researchers, using different methods, depending on sample availability, as outlined below. This means that although DNA was available for all samples, RNA was only available for a minority of these. DNA from some patients was kindly provided by Dr Chris Smith (Rosenfeld Group, Cancer Research UK Cambridge Institute) and Mr Tom Mitchell (Mitchell Group, Wellcome Trust Sanger Institute). For these cases, DNA was extracted from a small section of frozen tissue (approximately 20mg), using the commercially available DNeasy Blood & Tissue kit (QIAGEN) according to the manufacturers protocol. For the remainder of patients, DNA and RNA were extracted using the AllPrep DNA/RNA Mini Kit (QIAGEN) according to the manufacturers protocol, by either myself, Anne Babbage (research associate, Massie Group) or the Cambridge Cancer Molecular Diagnostics Laboratory (CMDL). I quantified DNA using the Qubit™ 4 fluorometer (ThermoFisher Scientific). Where gDNA concentration was below the required threshold for subsequent analysis (<9ng/µl), I used the SpeedVac Vacuum Concentrator (ThermoFisher Scientific) to increase DNA concentration.

Cell free DNA (cfDNA) was extracted from plasma, post-biopsy fluid and cell line supernatant (2ml, 3ml, 4ml, 8ml or 10ml depending on volumes available) using the Qiasymphony DSP Circulating DNA mini kit at the CMDL. Where there was insufficient sample volume available for extraction, samples were topped up to the extraction volume using PBS. The SpeedVac Vacuum Concentrator was not used on cfDNA. I quantified cfDNA using the Qubit™ 4 fluorometer (ThermoFisher Scientific) and Tapestation (Agilent).

4.2.2 Epic-seq library preparation and sequencing

I used the TruSeq Methyl Capture EPIC Library Preparation Kit (Illumina), hereby referred to as Epic-seq, to evaluate methylation in gDNA obtained from tissue. I sheared gDNA samples (10ng/µl, 500ng total) using the S220 Focused-ultrasonicator (Covaris) to generate dsDNA fragments. Samples were sheared for 280 seconds using the following shearing settings: 175W peak incident power, 10% duty factor, 200 cycles per burst. The D1000 ScreenTape System (Agilent) was used to ensure >60% of DNA fragments were between 100 and 300bp long, with a mean fragment size of 180-200bp. The Epic-seq library preparation was performed using the manufacturers protocol. This consists of a capture-based method targeting ~3 million CpGs. Four samples were multiplexed in each capture reaction using sample indexing adaptors. The protocol involves hybridization of biotin-tagged probes to gDNA followed by capture using streptavidin beads (two hybridization-capture steps) followed by bisulphite conversion at 54°C for two hours. Twelve samples were pooled for sequencing on the HiSeq4000 Illumina Sequencing platform (single end 150bp read; 20% PhiX) using two lanes per library pool. I performed two technical replicates for cell line data (gDNA derived from HK2 cells) and evaluated CpG methylation. I focused my analysis on CpGs which achieved ≥10x minimum coverage and demonstrated that the Pearson correlation coefficient between technical replicates was 0.97, suggesting assay reproducibility.

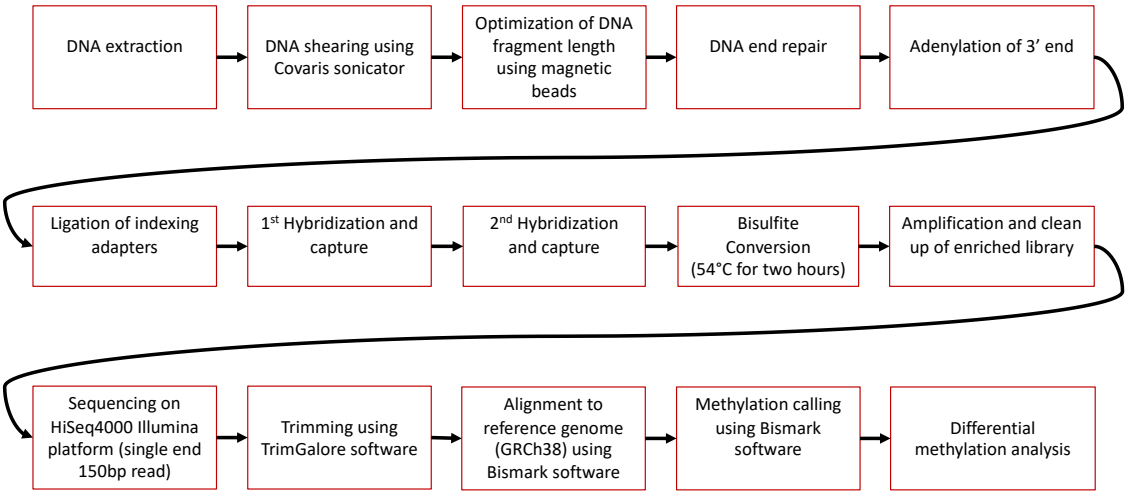


Figure 4.2: Epic-seq experimental methods

Epic-seq experimental methods are shown. Data processing is described in detail in section 4.4.1.

4.2.3 Nimbus library preparation and sequencing

I analysed cfDNA obtained from liquid samples (plasma, post-biopsy fluid and cell line supernatant) using the 'Non-destructive Integration of Methylation to Boost Underlying Signals' (Nimbus; Figure 4.3). The Nimbus experimental method and pipeline are currently under review for a patent; therefore the exact details cannot be disclosed (confidential protocol). In brief, enzymatic conversion was performed using the NEB Next Enzymatic Methyl-seq Kit (New England BioLabs). Following this, single stranded library preparation was performed and subsequently between 8 and 10 samples were pooled into a single hybridization capture reaction (using dual combinatorial indexing) to capture thousands of methylation markers. For the majority of samples, I performed all steps of the protocol (conversion, library preparation and capture). Due to limited sample availability, for control samples, converted libraries were obtained from Dr Gahee Park (postdoctoral research associate, Massie Group). I subsequently performed targeted capture and sequencing. The informative differentially methylated regions from my tissue-based analysis were used to create a panel covering 5801 regions of interest (i.e. panel size ~3.6 million base pairs). Section 4.8.1 describes how these capture targets were selected.

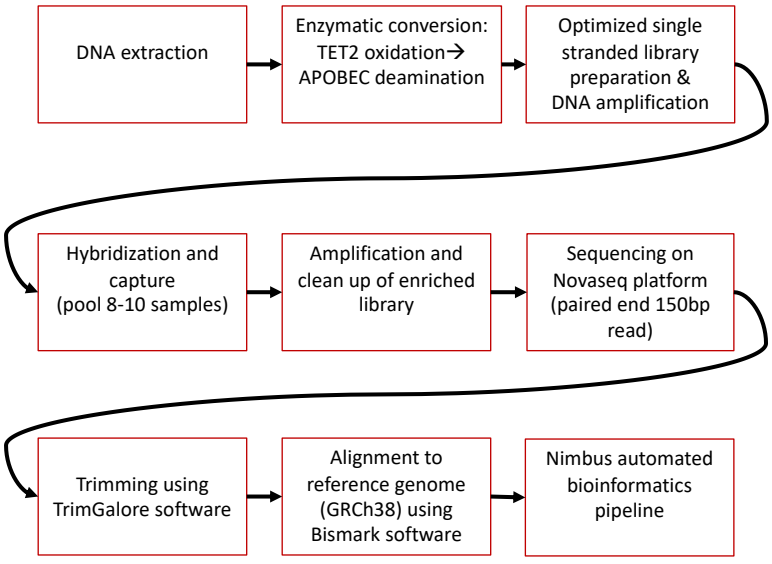


Figure 4.3: Nimbus experimental methods

Nimbus experimental methods are shown. Data processing is described in detail in section 4.4.1.

Although the Nimbus protocol has been developed for cfDNA analysis, it may also be applied to gDNA from nephrectomy specimens or core renal biopsies. Prior to running the protocol, I sheared gDNA samples using the S220 Focused-ultrasonicator (Covaris) to generate dsDNA fragments with an average insert size of 240–290 bp. Samples were sheared for 100-130 seconds using the following shearing settings: 140W peak incident power, 10% duty factor, 200 cycles per burst. The DNA input for Nimbus varied depending on sample type (50ng for cell line experiments, 50ng for human gDNA and 10ng for human cfDNA), due to the low concentration of cfDNA. For both gDNA and cfDNA, sequencing was performed using the NovaSeq Illumina Sequencing platform (paired end, 150bp) with 20% PhiX spike-in.

4.2.4 Whole exome library preparation and sequencing

In a subset of patients with ccRCC used for methylation analysis, whole exome sequencing (WES) of multi-region tumour tissue, normal kidney tissue samples and/or germline buffy coat DNA had previously been performed by Dr Christopher Smith at the Rosenfeld Group, CRUK CI, as previously described [158]. In brief, 50ng of gDNA were fragmented using the S220 Focused-ultrasonicator (Covaris). Library preparation was performed using the Thruplex DNA-Seq protocol (Rubicon Genomics; 5 PCR cycles). Next, exomes were captured using the TruSeq Exome Capture protocol (Illumina) according to the manufacturers protocol. Libraries were amplified (8 PCR cycles) and subsequently sequenced on the Hiseq4000 platform (Illumina). To increase the total number of samples available for my analysis, WES was performed for an additional set of samples. WES was undertaken by the genomics core at the CRUK CI, using Nextera™ Flex for Enrichment protocol (Illumina) according to the manufacturers protocol. Sequencing was performed using NovaSeq (paired end, 150bp, 55 samples on two lanes of S4).

4.2.5 RNA-seq library preparation and sequencing

RNA was available for a subset of multi-region samples (N=47) from patients with ccRCC used for methylation and WES analysis. I evaluated RNA integrity (RIN), measured on a scale from 0 to 10, using the Tapestation (Agilent). Unfortunately, the RNA was low quality (RIN values: median 5.8, minimum 1.4, maximum 8.8). Delays in sample freezing, processing and freeze/thaw cycles may have contributed to RNA degradation. RNA-seq was performed using the Illumina TruSeq stranded Total RNA kit on 225ng of RNA, according to the manufacturers protocol, by the genomics core at the CRUK CI. For library preparation, the location of the tumour and normal samples was

randomised on the plate to remove variability/batch effect related to the plate. In brief, ribosomal RNA (rRNA) was depleted using biotinylated, target-specific oligos combined with Ribo-Zero rRNA removal beads, to ensure only messenger RNA (mRNA) was left. Following this, the samples were fragmented for 2 minutes (due to the relatively low quality RNA). RNA was then copied to DNA using reverse transcriptase, the library was prepared and amplified (15 PCR cycles). Libraries were sequenced using NovaSeq (paired end, 50bp, 47 samples on 3 lanes of Novaseq SP) aiming to achieve approximately 10 million reads per sample.

4.3 Publicly available datasets

I searched the Gene Expression Omnibus (GEO) database to identify publicly available datasets containing both methylation and gene expression data for the four pathological subtypes of renal tumours. Unfortunately, no such data were available (TCGA contains ccRCC, pRCC and chRCC samples but no oncocytomas). Therefore, separate methylation and gene expression datasets were searched. A few studies were available in GEO reporting Affymetrix gene expression data for the four pathological subtypes, and the dataset with the largest sample size was selected for inclusion in this thesis [176].

Furthermore, I obtained additional publicly available DNA methylation and gene expression datasets from a number of sources, as shown in Table 4.1. TCGA data for kidney cancer tissue samples (ccRCC, pRCC, chRCC and adjacent normal) were obtained via the 'TCGAbiolinks' package v2.20.1 in R [177]. In TCGA, methylation data were assessed using the Illumina Infinium Human DNA Methylation 450k platform (450k array). Pre-processed beta values were downloaded for samples, along with clinical and sample characteristics. TCGA data were also evaluated using the following online interactive tools: 'TCGA Wanderer' [178] and 'cBioPortal' [179]. The Jones, Wei and Evelönn datasets were downloaded directly from GEO using the accession codes shown in Table 4.1. The Chopra data were downloaded from the Open Science Framework (repository OSF.IO/Y8BH2) and the Brennan data were obtained directly from the study authors. For deconvolution analysis (section 4.6.1), I obtained reference methylomes for various cell types (Table 4.2) from GEO and the ENCODE project [180]. In order to obtain reference methylomes for kidney cell lines, I performed Epic-seq on HK2 and 786-O cells (as described in section 4.2.2). Some of the publicly available datasets used previous genome builds (as shown in Table 4.1 and Table 4.2), therefore I converted these to hg38 using the 'liftover' function in the 'rtracklayer' package v1.52.1 in R [181]. Section 4.4.4 describes the methods I used to overlap methylation datasets from different platforms.

Table 4.1: List of publicly available datasets used in the analysis

The table summarises the data source (including GEO accession number and/or full reference), data type and use in the thesis.

Data	Sample type	Data type/Platform	Source	Use
TCGA [24]	Tissue from ccRCC, pRCC, chRCC and normal kidney	DNA Methylation (450k platform) and gene expression (RNA-seq)	TCGAbiolinks [177], TCGA Wanderer [178] and cBioPortal [179]	Chapter 5 & Chapter 6
Chopra et al [120]	Tissue ccRCC, pRCC, chRCC and normal kidney	DNA methylation (450k platform)	Open Science Framework (repository OSF.IO/Y8BH2)	Chapter 5
Brennan et al [121]	Tissue from ccRCC, chRCC and normal kidney	DNA methylation (450k platform)	Data obtained directly from study authors	Chapter 5
Wei et al [126]	Tissue from ccRCC and normal kidney	DNA methylation (450k platform)	GEO accession: GSE61441	Chapter 5
Evelönn et al [127]	Tissue from ccRCC and normal kidney	DNA methylation (450k platform)	GEO accession: GSE113501	Chapter 5
Jones et al [176]	Tissue from ccRCC, pRCC, chRCC and normal kidney	Gene expression (Affymetrix HGU-133A)	GEO accession: GSE15641	Chapter 5

Table 4.2: Reference methylomes used for methylation deconvolution analysis

Data	Sample type	Data type/ Platform	Source
Salas et al [182]	Immune cells	850k array	GEO accession number: GSE110555 Sample identifiers: GSM2998022 = NK cell GSM2998024 = B cell GSM2998030 = neutrophil GSM2998032 = CD8+ T cell GSM2998039 = monocyte GSM2998048 = CD4+ T cell
ENCODE [180]	Immune cells	WGBS (hg38) Minimum coverage 5x	ENCODE project, sample identifiers: ENCFF649RBS = myeloid progenitor ENCFF689TNG = monocyte ENCFF703XLD = B cell ENCFF953DKC = T cell
ENCODE [180]	Adipose cell	WGBS (hg38) Minimum coverage 5x	ENCODE project, sample identifiers: ENCFF528JMA = adipose cell
ENCODE [180]	Skin fibroblast (Primary cell)	WGBS (hg38) Minimum coverage 5x	ENCODE project, sample identifiers: GM23248 ENCFF752NXS = fibroblast replicate 1 ENCFF116DGM= fibroblast replicate2
Generated by me	HK2 & 786-O cell lines	Epic-seq (hg38) Minimum coverage 10x	Data generated by me

Abbreviations: WGBS = whole genome bisulphite sequencing

4.4 Data analysis: general methods

4.4.1 Data processing

For methylation (tissue and liquid samples), sequencing data were processed by Sara Pita (research assistant, Massie Group) using an in-house pipeline. Sequenced data were trimmed (TrimGalore v0.4.4) and aligned to the bisulphite converted human reference genome (GRCh38/hg38) using Bismark (v0.22.1) and Bowtie2 (v2.4). For gDNA derived from tissue and prepared with Epic-seq, duplicate reads were maintained, whereas for cfDNA prepared using the Nimbus protocol, duplicate reads were removed. De-duplication was performed in the latter to enable an estimate of the number of unique reads, as the Nimbus analysis pipeline calculates the number of tumour-derived (unique) reads in cfDNA. Methylation calling was performed using the Bismark suite of tools (v0.22.1). Trimming and alignment reports were compiled using MultiQC (v1.7), and further analysis was performed by myself. I evaluated quality control metrics using Picard Tools (default settings) [183] and the output from MultiQC.

For RNA-seq and WES, data processing was performed by Kamal Kishore at the bioinformatics core at the CRUK CI. For WES, sequenced data were aligned to the human reference genome (GRCh38/hg38) using bwa v0.7.17 (bwa-mem algorithm, using default settings). For RNA-seq, the results were aligned to the reference transcriptome using 'Salmon' v1.4.0 [184]. After alignment, all subsequent analysis was performed by me.

4.4.2 Data visualisation and statistical analysis

I performed data visualisation and analysis using the R statistical software (version 3.6.1 subsequently upgraded to 4.1.1) and created figure schematics using Biorender (Biorender.com). Unsupervised clustering was performed using principal component analysis (PCA) and/or dendrograms. Where multiple comparisons were performed (e.g. methylation and gene expression analysis of four pathological subtypes of renal tumours versus normal kidney), a Venn diagram was utilised to compare results. Heatmaps were used to depict DNA methylation and gene expression for multiple samples, using the 'ComplexHeatmap' package v2.8.0 [185]. In addition, locus plots were created to visualise DNA methylation levels for each sample, at each CpG within a given locus (reference genome hg38). For TCGA datasets, locus plots were evaluated using 'TCGA Wanderer' (reference genome hg19) [178].

Group differences were compared using Fisher's exact test (for categorical variables), Mann-Whitney or Wilcoxon Signed Rank Sum test (for non-parametric data) and Student's t test (for parametric data). For continuous variables, association was evaluated using Pearson's correlation coefficient and/or linear regression models to obtain an adjusted R^2, along with p values. P values were corrected for multiple testing using a Benjamini–Hochberg correction.

4.4.3 Differential methylation analysis in tissue

I generated DNA methylation data using Epic-seq, as described in section 4.2.2, and sequencing reads were processed as described in section 4.4.1. All subsequent analyses were performed by me, unless specified. The percentage methylation at a given locus was obtained by counting the number of methylated cytosines divided by the total number of reads (i.e. number of Cs/number of Ts+Cs). Methylation levels therefore range between 0 and 1, and this is often referred to as the beta value. CpGs located on the sex chromosomes were omitted to remove gender bias. In addition, CpGs located at the site of C/T and G/A SNPs were removed as these cannot be distinguished from differential DNA methylation in single read data [186]. Data were included in downstream analyses if a depth of ≥10x coverage was achieved, to reduce the risk of false positive calling. This level of minimum coverage allows a 10% methylation difference to be called (i.e. 1 out of 10 reads). For comparisons of pathological subtypes of renal tumours (pRCC, chRCC and oncocytoma) versus normal kidney tissue (section 4.5.1), CpGs were considered if ≥10x coverage was achieved in all samples. However, for ccRCC, CpGs were considered if ≥10x coverage was achieved in at least 90% of samples. Less stringent criteria were used for the latter to avoid a few low-quality samples impacting the overall number of CpGs assessed.

I performed differentially methylated cytosine (DMC) analysis at individual CpGs using the 'methylKit' package v1.12.0 in R [187]. This package uses logistic regression to compare CpGs between two groups (e.g. tumour *vs* normal), with p-value adjustment for multiple testing. Patient ID was used as a covariate in logistic regression to adjust for multi-region sampling. Significant DMCs were defined as ≥25% absolute methylation difference between groups (a commonly used cut-point in the literature) and q value <0.01 [187]. I performed differentially methylated region (DMR) analysis to identify regions that were differentially methylated using the 'dmrseq' package v1.6.0 [188] in R. Significant DMRs were defined as having a q value of <0.01 [188].

4.4.4 Overlapping Epic-seq and 450k array datasets

The Epic-seq method that I used to evaluate methylation for my samples generates sequence level data on approximately 3 million CpGs; in contrast with the approximately 450,000 CpGs included in the Illumina 450k array. A number of analyses in this thesis required an external validation set from publicly available sources (e.g. TCGA). In order to combine data from the two methods, Epic-seq methylation values within ±50bp of the 450k probes were averaged, as adjacent CpGs tend to be co-methylated [107]. This has previously been demonstrated to be a valid approach; 8 technical replicates (fresh frozen tissue) assessed on both Illumina Epic-seq and the 450k array obtained a correlation $\geq$0.96 using this method (data from Sara Pita, research assistant, Massie Group; manuscript under preparation). After combining datasets, I removed CpG probes found in two blacklists [189, 190] on the 450k array. These consist of CpG probes that either map to multiple regions, are located at repeat regions, on sex chromosomes or at the site of C/T and G/A SNPs.

4.4.5 Annotation and enrichment analysis

I annotated CpGs of interest to the human reference genome (hg38) to determine their location relative to CpG islands and shores, as well as annotating to the nearest proximal gene using the 'methylKit' v1.12.0 and 'ChIPseeker' v1.22.1 packages respectively [187, 191]. The hg38 annotation source was obtained from the 'EnsDb.Hsapiens.v86' package v2.99.0 in R [192]. CpGs which were within 1.5kb of the promoter region were then selected for enrichment analysis. Gene set enrichment analysis and ontology (including disease ontology, biological processes, molecular functions) were performed using 'clusterProfiler' v3.14.3 in R [193], where the background set was defined as the total number of features evaluated.

4.4.6 Gene expression using Affymetrix HG-U133A

I downloaded and analysed publicly available Affymetrix gene expression data (Accession number: GSE15641) [176]. The study authors performed transcriptional profiling of kidney tumours using Affymetrix HGU-133A chips according to the manufacturers protocol, for 84 fresh-frozen samples, including 32 ccRCC, 11 pRCC, 6 chRCC, 12 oncocytoma and 23 normal kidney. MAS5 normalised data were provided by the authors for 22283 probes, and the remainder of the analysis was performed by myself in R. I log transformed data (base 2) and removed probes with low expression, as previously described [194]. In order to define probes with a low expression, I created a histogram to evaluate median intensities and identified a suitable threshold (set as log2 value expression <5). Following

this, I excluded 265 probes below this threshold. In the Affymetrix platform, transcripts are represented by multiple probes, therefore signal at multimapping probes was averaged using the 'avereps' function in the 'Limma' package v3.42.2 [195]. After processing and filtering, gene expression data were available for 12606 genes. I used the 'Limma' package to identify significant differentially expressed genes (DEGs) for each pathological subtype compared to normal tissue, adjusting p values for multiple testing. 'Limma' fits a linear model with contrasts for disease subtypes compared to normal kidney tissue and uses a parametric empirical Bayes approach [195]. ANOVA is used to determine differences between all the groups and pairwise comparisons are performed for groups of interest, with p value adjustment for multiple testing (Benjamini Hochberg correction). Significant DEGs were defined as having an absolute log2 fold change >1 and an adjusted p value <0.01. Significant DEGs were ranked based on the greatest expression difference (i.e. highest absolute log2 fold change in all subtypes combined), to elucidate shared differences between renal tumours and normal tissue. In addition, significant DEGs were ranked based on the greatest absolute differences amongst renal tumours, to identify the most pronounced dissimilarities between tumour types.

4.4.7 Gene expression using RNA-seq

RNA-seq data were generated for 47 fresh-frozen ccRCC and normal kidney tissue samples as described in section 4.2.5. The sequencing alignment step was performed by Kemal Kishore at the CRUK CI bioinformatics core, and all subsequent analyses were performed by me. After alignment to the human transcriptome with 'Salmon' v1.4.0 [184], I converted transcriptome level count data to gene level data using 'tximport' v1.14.2 in R [196]. Genes with low expression (counts <5) were subsequently removed. 'DESeq2' v1.32.0 was utilised to evaluate differentially expressed genes in ccRCC versus normal kidney, using non-normalised count level data [197]. Importantly, Patient ID was used as a covariate to account for multi-region sampling, and p values were adjusted for multiple testing. RNA-seq data were also used to determine the ClearCode34 prognostic risk score for each sample, as described by the study authors [92, 93]. In brief, data were available for 31 out of the 34 ClearCode34 genes. First data were median centred, then log transformed (log2 + 0.1) and visualised in a heatmap with unsupervised hierarchical clustering.

4.5 Differentiating pathological subtypes of renal tumours and normal kidney

4.5.1 Characterising DNA methylation & gene expression in pathological subtypes of renal tumours

I generated DNA methylation data for 326 tissue samples using Epic-seq (139 ccRCC, 27 pRCC, 27 chRCC, 24 oncocytomas and 109 adjacent normal kidney; section 4.2.2). Subsequently, I performed differential methylation analysis amongst normal samples, and between each pathological subtype versus normal kidney. DMC analysis was performed using 'methylKit' v1.12.0 [187], as described in section 4.4.3. Furthermore, I obtained Affymetrix HGU-133A gene expression data from Jones et al for 84 tissue samples (32 ccRCC, 11 pRCC, 6 chRCC, 12 oncocytoma and 23 normal kidney) [176]. I determined DEGs for each pathological subtype versus normal kidney, as described in section 4.4.6.

In order to identify epigenetically regulated genes, I overlapped the significant DMCs from my methylation analysis with significant DEGs. This approach (i.e. overlapping two separate datasets) was taken due to the absence of a comprehensive dataset containing matched methylation and gene expression data for all pathological subtypes of renal tumours. First, I annotated DMCs from my analysis to the nearest gene, and those located within 1.5kb of the transcription start site (TSS) were selected and overlapped with gene expression data. Gene set enrichment and ontology were performed for genes which demonstrated a negative correlation between promoter methylation and expression (as these are most likely to be functionally relevant), as described in section 4.4.5. Genes were ranked by the number of significant DMCs within the promoter, as well as based on greatest differential methylation and gene expression. The latter was performed (rather than ranking by p-value) to quantify the magnitude of the effect on methylation and gene expression. The top-ranking genes were reported in tables and selected genes were discussed in detail in the text. Thus, I identified genes which may be epigenetically regulated, for each subtype. In order to validate whether these genes were epigenetically regulated in an external dataset, I assessed TCGA data for ccRCC, pRCC and chRCC (no data available for oncocytoma). I obtained matched methylation and gene expression data and scatterplots of methylation versus gene expression from 'TCGA Wanderer' [178] and 'cBioPortal' [179]. The association between methylation at CpG probes and gene expression was assessed using Pearson's correlation coefficient.

For ccRCC vs normal kidney, I also externally validated the epigenetically regulated genes in a separate cohort. I generated matched data (Epic-seq and RNA-seq) for a subset of ccRCC and normal kidney samples (N=47) (see sections 4.2.2 and 4.4.7). I evaluated promoter methylation and gene expression for these samples, to externally validate the top-ranking genes identified in my earlier analysis. A variance stabilising transformation (VST) was applied to gene expression data to ensure that the variance is approximately the same across different mean values (i.e. the data are homoscedastic). For these samples I was able to directly quantify the association between promoter methylation and gene expression using Pearson's correlation coefficient.

4.5.2 Machine learning model to predict pathological subtypes of renal tumours

The following work was performed in collaboration with Izzy Newsham, to develop MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours). First, I will describe the methods used to create the model, subsequently I will delineate author contributions.

MethylBoostER is an extreme gradient boosting (XGboost) machine learning model which uses DNA methylation data to classify tissue samples into one of five pathological subtypes. Figure 5.14 in Chapter 5 represents a graphical summary. The model was developed using a testing/training set, and externally validated on four independent datasets. For the training/testing set, samples were combined from three sources, my Epic-seq data (N=319), TCGA (N=872) [24] and Chopra Training set (N=37) [120]. The probes were processed and filtered as described in section 4.4.4. Furthermore, CpGs were removed if >5% of data were missing (or if data were missing for all samples from one dataset), obtaining a total of 158,670 CpG probes. Four publicly available datasets were used for the external validation: Chopra validation (N=245), Brennan (N=37), Wei (N=92) and Evelönn (N=144) [120, 121, 126, 127]. The same 158,670 CpG probes used in the training/testing set were selected for all four datasets in the external validation. Percentage methylation (i.e. beta values) at these 158,670 CpG probes were converted to M values. M values represent a homoscedastic transformation of beta values (β), using the formula shown below. For β values of 0 and 1, this would result in M values of positive and negative infinity. In these cases, it is customary to set the maximum and minimum M values as the maximum and minimum finite values within the dataset.

$$M = log_2\left(\frac{\beta}{1-\beta}\right)$$

MethylBoostER was developed using four-fold nested cross validation on the training/testing set, with integrated hyperparameter optimisation (see Appendix 1). The multi-class model predicts one of five classes (ccRCC, pRCC, chRCC, oncocytoma and normal kidney). The dataset was randomly split into training and testing sets (75:25 split) four times, whilst maintaining all multi-region samples from the same patient in the same set, to avoid data leakage. This ensures that the model was not tested and trained on samples from the same patient, which would result in overfitting. Nested cross validation consists of further splitting this training set (75:25) to enable hyperparameter tuning. The following hyperparameters were explored: number of trees, maximum tree depth, learning rate, L1 and L2 regularization terms. This nested cross validation method enables the identification of hyperparameters that maximise the Matthews correlation coefficient (MCC) score, prior to testing on an unseen set of data (i.e. the 25% of data not used to select the hyperparameters) to avoid over-fitting. The MCC is defined as the correlation coefficient between actual and predicted values, and is a more reliable metric than accuracy when classes are imbalanced [198]. Using this method, it is therefore possible to report the results of the entire training/testing set (since all samples will have been evaluated in one of the four testing sets during cross validation). Samples were assigned weights to account for imbalances in class sizes (referred to as 'class weights'); for example, ccRCC and normal kidney are the most common classes. Weights were also assigned to account for the use of multi-region samples (referred to as 'patient weights'), to ensure that individuals with relatively more multi-region samples do not influence the model disproportionately compared to those with less samples. Overall sample weights were obtained by multiplying 'class weights' with 'patient weights.' The following model metrics were evaluated: accuracy, precision, recall, MCC, Receiver Operating Characteristic (ROC) curve and Precision-Recall curve (Table 4.3). For both curves, data were plotted for each class separately (i.e. for each class compared to the remainder) and the area under the curve (AUC) was derived.

Table 4.3: Metrics evaluated in MethylBoostER

Metric	Description
Accuracy	The proportion of correctly classified samples (i.e. correctly classified samples divided by the total number of samples).
Matthews correlation coefficient	The correlation coefficient between actual and predicted values.
Precision (i.e., positive predictive value)	The number of true positives given a positive test. Calculated by dividing true positives by the sum of true positives and false positives.
Recall (i.e., sensitivity)	The number of true positives given a positive diagnosis. Calculated by dividing true positives by the sum of true positives and false negatives.
Precision-Recall curve	Plot of precision versus recall (i.e. the positive predictive value over the true positive rate). Therefore, if model precision is 1, there are no false positives; and if recall is 1, there are no false negatives.
Receiver Operating Characteristic curve	Plot of true positive versus false positive rate.

The prediction probability was evaluated for each sample in the training set, as this represents a measure of the confidence of the prediction accuracy. High- and moderate-confidence predictions were therefore defined as prediction probabilities above and below the threshold *t* respectively. The optimal value of *t* was derived by plotting three metrics (the accuracy of high- and moderate-confidence predictions and the fraction of high-confidence predictions) over the testing set, at different values of *t*. Simple linear models were fitted for these (to smooth results), and *t* was selected to maximise the average of the three metrics. The value of *t* was independently validated in the four external validation sets. For high-confidence predictions, the MethylBoostER model outputs the most likely class, whereas in moderate-confidence predictions, the model will output the two most likely classes (referred to as first and second prediction hereafter).

I derived tumour purity using DNA methylation data via the 'InfiniumPurify' package v1.3.1 in R [199] (see section 4.6.1). Purity was compared in samples which were correctly predicted on the first prediction, second prediction and incorrectly predicted samples using the Wilcoxon Signed Rank Sum test with BH correction for multiple testing. The association between tumour purity and prediction probability was evaluated using Pearson's correlation coefficient. The accuracy achieved at different purity thresholds was also evaluated. Furthermore, model results were reported for multi-region samples derived from the same patient to evaluate the impact of intra-tumoral heterogeneity (ITH) on predictions.

Herein I summarise authorship contributions for analysis and manuscript preparation (Appendix 1). The analysis was performed in collaboration with Izzy Newsham and supervision was provided by Dr Shamith Samarajiwa (bioinformatics), Dr Charles Massie (DNA methylation and tumour biology) and Prof Grant Stewart (clinical application). The initial study idea was my own, as the diagnosis of SRMs is a key research question addressed in my thesis. Furthermore, I ensured that the analysis evaluated the impact of tumour purity and methylation ITH on model predictions, as these are clinical priorities. I generated the experimental data and provided expertise in DNA methylation and clinical applications, and Izzy Newsham provided advanced bioinformatics skills. Specifically, I performed the following tasks: generated experimental Epic-seq data, pre-processed Epic-seq data to obtain beta values, searched the literature to identify publicly available datasets for use in the model. Data merging/filtering and the machine learning model were undertaken in collaboration with Izzy Newsham. Specifically, Izzy Newsham wrote the code in Python, however decisions regarding model methods, structure and analysis were undertaken in collaboration with myself, through an iterative process. For example, the following decisions were developed in collaboration

with Izzy Newsham (and our supervisors): method to avoid data leakage for multi-region samples, weighting to mitigate patient and class bias and use of high- and moderate-confidence predictions to maximise clinical utility. All analyses evaluating the association between purity and model output were done solely by me. The evaluation of methylation ITH was undertaken collaboratively, whereas work regarding the model's clinical application was undertaken by me.

A number of collaborators were also involved in the manuscript (Appendix 1), as outlined below. Dr Kevin Brennan and Dr Olivier Gevaert (Gevaert Group, Stanford University, USA) provided sample data and bioinformatics support. Dr Thomas Mitchell (Wellcome Trust Sanger Institute) and Dr John Leppert (Stanford University) provided advice regarding clinical applications and reviewed the manuscript. Wing Kit Leung (Tavare Group, CRUK CI) and Dr Gahee Park (Massie Group, Hutchison MRC) offered laboratory supervision, while Dr Anne Warren (Pathology, Addenbrooke's Hospital) reviewed pathology slides. All collaborators read and approved the manuscript.

4.6 Tumour purity assessment and cell type deconvolution

4.6.1 Purity and deconvolution using DNA methylation data

I estimated tumour purity from methylation data using the 'InfiniumPurify' package v1.3.1 in R, as previously described [199]. Importantly, the package estimates tumour purity relative to contamination with normal (non-cancerous) tissue. The function compares methylation in normal and tumour tissue (taking into account the variance of methylation in tumour) and identifies informative differentially methylated CpG sites (iDMCs). iDMCs are then used to estimate purity using Gaussian kernel density [199].

Percentage methylation at CpGs for each sample (i.e. beta values) represent a mixture of methylation values from reads derived from different cell types. I performed cell type deconvolution from methylation data using the 'MeDeCom' package v1.0.0 in R [200, 201]. Good quality reference epigenome maps for ccRCC purified cell components are lacking (lack of available data and poor overlap between CpGs covered by Epic-seq). Therefore, I selected a 'reference-free' deconvolution method to perform an unconstrained analysis. First, feature selection was performed: DMCs were ranked based on the highest methylation variance in tumour tissue and the top 10% (N=10794 DMCs) were identified in order to select the most informative features. 'MeDeCom' is unable to handle missing data, therefore imputation was performed for missing CpGs using the k nearest

neighbour ('impute' package v1.60.0 in R) [202]. 'MeDeCom' uses regularized non-negative matrix factorization to decompose the DNA methylation matrix into two matrices: cell-type-specific latent methylation components (LMCs) and the proportion of LMCs in each sample [201]. LMCs represent the reference methylomes of unknown cell populations. The method was run over multiple iterations and the parameters K (i.e., the number of LMCs) and λ (i.e., regularization parameter) which minimize the cross-validation error were selected. In order to identify the potential cell type corresponding to each LMC, these were correlated with reference methylomes for known cell types. I performed Epic-seq to establish reference methylomes for the HK2 and 786-O cell lines, representing normal kidney proximal epithelium and ccRCC tumour respectively. The remainder of the cell type reference methylomes were obtained from the literature, as described in section 4.3 and Table 4.2. The LMCs were also correlated with purity estimates derived by 'InfiniumPurify' (DNA methylation), 'ESTIMATE' (RNA-seq) and 'ASCAT' (WES) (described in section 4.6.2). The Wilcoxon signed rank sum test was used to assess whether there was a significant difference in the LMC content by tumour stage (stage I-II vs III-IV), grade, Leibovich score (low versus intermediate/high) and recurrence status (no recurrence vs recurrence), whilst adjusting for multiple testing.

4.6.2 Purity and deconvolution using RNA-seq and WES data

I used RNA-seq data to estimate tumour purity using the 'Estimation of STromal and Immune cells in MAlignant Tumours using Expression data' (ESTIMATE) package v1.0.13 in R, as previously described [203]. 'ESTIMATE' determines tumour purity as a function of admixtures of immune and stromal cell components based on gene expression data. I compared immune and stromal cell components derived from 'ESTIMATE' for my dataset (N=47) and TCGA data, confirming my results were within the expected range. Tumour purity was estimated from WES data using 'ASCAT', by my collaborator Victoria Dombrowe (Schwarz Group, Max Delbrück Center, Berlin) [204]. Subsequently, I used Pearson's correlation coefficient to compare the association between purity calculated using WES, RNA-seq and Epic-seq.

In addition, I performed cell type deconvolution using RNA-seq data and 'CIBERSORTx' [205], via the 'Immunedeconv' package v2.0.4 in R [206]. Gene expression data were normalised (normalization method: transcripts per million, not log transformed) prior to decomposition. 'CIBERSORTx' enables the deconvolution of bulk RNA-seq data into 22 immune cell types, based on a reference transcriptome provided by the study authors (termed the 'LM22 signature matrix') [205].

4.7 Analysis of methylation heterogeneity in ccRCC tissue

I performed an analysis of methylation heterogeneity on three levels: between patients, within a patient and within a sample, as previously described [207]. Figure 6.1 in Chapter 6 represents a graphical summary. I performed all the methylation analyses using Epic-seq data generated from ccRCC and normal kidney tissue as described in section 4.2.2.

4.7.1 Heterogeneity between patients

Heterogeneity between patients was defined as an evaluation of methylation patterns in tumour samples from different patients with the same diagnosis (i.e. ccRCC). In order to evaluate methylation heterogeneity between patients, methylation beta values were obtained. All CpGs at SNP sites were removed (i.e. not just C/T and G/A SNPs) to ensure that the heterogeneity noted between patients was driven by methylation rather than SNPs. Principal component analysis was performed using data from all available CpGs (~1.1million CpGs) and sample clustering was evaluated. Subsequently the top-most variable CpG (i.e. CpGs with the highest variance in tumour samples) were selected for visualisation in a heatmap. This analysis was compared for the top 10,000 CpGs and the top 50,000 CpGs.

4.7.2 Heterogeneity within a patient

Heterogeneity within a patient was defined as an evaluation of methylation patterns amongst multi-region tumour samples obtained from one patient (i.e. multiple tumour and normal samples are taken from each individual ccRCC tumour). ITH was evaluated by assessing epigenetic age, the average pairwise ITH index (APITH) and phylogenetic trees derived from methylation and somatic copy number data (SCNA). For each of these parameters, I evaluated the association with clinical and prognostic factors. I assessed whether there was a significant difference by tumour stage (stage I-II vs III-IV), grade, tumour size, Leibovich score (low vs intermediate/high) and recurrence status (no recurrence vs recurrence), whilst adjusting for multiple testing.

4.7.2.1 DNA methylation age

The predicted DNA methylation age of each sample was calculated using Horvath's epigenetic clock, using publicly available code published by the study authors [102]. In order to demonstrate reliability of methods, first I evaluated methylation age for TCGA ccRCC tumour (N=325) and normal kidney (N=160) tissue samples. Subsequently, I repeated the analysis for my Epic-seq samples. Since Horvath's clock was developed using the 21k Illumina methylation array, the Epic-seq methylation data for my samples was overlapped with the 21k array by averaging methylation levels within ±100bp of Illumina CpG probes, as described in section 4.4.4. The 100bp threshold (rather than 50bp) was selected to increase the number of CpG probes for which data might be available. Epic-seq and TCGA data were combined into one data frame and for missing values, imputation was performed using k-nearest neighbours with the 'impute' package v1.60.0 in R (k=10, rowmax=0.25) [202]. The association between chronological age and predicted DNA methylation was evaluated for normal and tumour samples separately, using Pearson's correlation coefficient. The predicted to chronological age ratio (PCAR) was calculated by dividing DNA methylation age by real age, as previously described [208]. Accelerated ageing was defined as a PCAR ≥ 1.

4.7.2.2 Average Pairwise ITH Index (APITH)

Heterogeneity within a patient was evaluated by calculating the Average Pairwise ITH (APITH), using DNA methylation and copy number data respectively, as previously described [209]. For DNA methylation, I calculated the APITH for all CpGs as well as the top 5000 most variable CpGs, using the equation below [209]. Methylation beta values were obtained for tumour samples at the CpGs of interest, subsequently, the pairwise Euclidean distance was calculated and then the average was obtained. In the equation, k represents the total number of samples, and dij is defined as the pairwise Euclidean distance between two samples (i and j):

$$APITH = \frac{2}{k(k-1)} \sum_{1 \leq i < j \leq k} d_{ij}$$

Subsequently, I explored whether the methylation APITH index may be confounded by tumour purity. Tumour purity was obtained using WES or RNA-seq as described in section 4.6.2. I assessed the correlation between the APITH and the variance of the purity of tumour samples derived from the same patient (the latter represents a measure of the spread of purity values). In addition, methylation beta values were adjusted for purity using the 'Infiniumpurify' package in R [199], and

the methylation APITH score were calculated once again. I compared the APITH index derived from methylation data which were unadjusted versus adjusted for tumour purity, to assess whether this produced similar results.

The copy number APITH index was calculated by Dr Roland Schwarz (Max Delbrück Center in Berlin) by evaluating the percentage of the genome which is affected by private SCNA (using the 'ASCAT' package [204]) in each sample and the average pairwise distance between samples, as previously described [209]. I evaluated the correlation between the methylation APITH and the copy number APITH in my dataset using Pearson's correlation coefficient.

4.7.2.3 Phylogenies using DNA methylation and copy number data

Phylogenies were created using methylation data and SCNA data respectively. Patients were included in the analysis if matched methylation and SCNA data were available on ≥4 tumour samples (N=8 patients), to allow comparisons between phylogenetic and phylo-epigenetic tree topologies. I created phylo-epigenetic trees using the 'Ape' package v5.5, where DNA methylation beta values were treated as a continuous variable between 0 and 1 [210]. I selected the top 10% of CpGs with the highest variance in tumour samples and calculated the Euclidean distance matrix. Trees were subsequently inferred using the ordinary least squares minimum evolution algorithm [211], as previously described [212, 213]. Phylogenetic trees were created using SCNA data from WES using 'Minimum-Event Distance for Intra-tumour Copy-number Comparisons' (MEDICC2) by my collaborator Victoria Dombrowe (Schwarz Group, Max Delbrück Center, Berlin) [214, 215]. In brief, allele-specific copy number analysis was performed using 'ASCAT' [204], and subsequently these underwent reference phasing using 'Refphase' [216]. 'MEDICC2' calculates the pairwise minimum-event distance between samples, and these data are used to create phylogenetic trees using the neighbour joining algorithm [217]. I was provided with phylogenetic trees (in Newick format) and performed all subsequent analyses, including tree visualisation and comparisons. I used the Robinson-Fould measure to compare similarities between phylogenetic and phylo-epigenetic trees for each patient, using the 'TreeDist' package v2.2.0 [218]. In brief, the Robinson-Fould measure is derived by counting the number of unique splits which occur in each tree (i.e. splits which occur in one tree and not the other), and the overall metric is normalised (0 to 1 scale) to enable comparisons across trees. In this case a split is defined as a bipartition in a tree which separates two taxa. Phylogenetic and phylo-epigenetic trees were visualised using the 'plot' function in the 'Ape' package v5.5.

4.7.3 Heterogeneity within a sample

Methylation heterogeneity within a sample was assessed by calculating epipolymorphism. Epipolymorphism is defined as 'the probability that two epialleles randomly sampled from the locus differ from each other', where an epigenetic locus (e-locus) consists of four adjacent CpGs in a single sequencing read (i.e. a 150bp window) [113, 144]. Given 4 adjacent CpGs, there are 16 (i.e. 2^4) possible combinations of methylated and unmethylated cytosines, so there are 16 possible epialleles. Epipolymorphism values were calculated using the 'methclone' package [144], using the formula below [219]. In brief, the proportion of each methylation pattern (p) is squared, then all values are summed and subtracted from one. There are a total of 16 possible methylation patterns (i.e. epialleles) therefore in this case S=16.

$$Epipolymorphism = 1 - \sum_{i=1}^{S} p_i^2$$

Epipolymorphism values range between 0 (i.e., fully concordant methylation pattern) and approaching 1 (i.e. highest degree of heterogeneity) [113]. The following section justifies the thresholds/variables used in my analysis. Li et al previously explored the number of adjacent CpGs (up to 10 CpGs) used to define an e-locus and 4 CpGs was selected as this optimised the number of reads and epialleles [144]. Therefore, in my analysis I defined an e-locus as four adjacent CpGs in a single sequencing read. Next, I evaluated the number of e-loci obtained at differing thresholds of minimum coverage. For example, ≥10x coverage obtained 138,412 e-loci, whereas ≥20x coverage obtained 59,480 e-loci, meaning <50% of data compared to ≥10x. Therefore, e-loci were included in my analysis if methylation data were present in ≥75% of samples, at ≥10x coverage in order to increase the number of e-loci that were considered. E-loci located on sex chromosomes were excluded from the analysis. The 'epihet' package v1.2.0 was used to compare average epipolymorphism at each e-locus in two groups (for example ccRCC versus normal kidney) and to determine e-loci with significant differential epipolymorphism (defined as absolute epipolymorphism difference >0.1 and adjusted p value <0.01) [219]. The epipolymorphism difference cut-off of 0.1 represents a >10% difference in epipolymorphism values, and is the cut-off routinely used in the literature [219]. The following illustrative example puts this into context. Given 10 reads, if all reads had the same methylation pattern the epipolymorphism would be 0, whereas if one read had a different pattern, then epipolymorphism would be 0.18. Using the selected cut-off of >0.1, this difference would be called as significant (provided the adjusted p value was <0.01).

I evaluated epipolymorphism in a cohort of ccRCC versus normal kidney samples (N=135 samples) to identify e-loci with significant differential epipolymorphism. Significant e-loci were annotated to the nearest gene and GSEA was performed, as described in section 4.4.5. I externally validated my results by assessing differential epipolymorphism in an independent cohort of ccRCC and normal kidney samples (N=71 samples, Epic-seq generated by me). Subsequently, epipolymorphism was evaluated in HK2 (6 technical replicates) and 786-O cell lines (4 technical replicates), which represent a model system of ccRCC and normal renal proximal tubule epithelium respectively. Cell lines are 100% pure, thus enabling an evaluation of epipolymorphism which is not confounded by the presence of heterogeneous groups of cells (i.e. gDNA derived from cell lines represents a single cell type, whereas kidney tissue contains multiple cell types).

Average methylation was calculated at each e-locus using the 'epihet' package v1.2.0, by calculating the average methylation across 4 adjacent CpGs, across all reads at that locus. Significance was defined as absolute methylation difference >15%, and an adjusted p value < 0.01. Whilst for individual CpGs the commonly used threshold for differential methylation is >25%, a lower threshold was used for average methylation across a read as this includes four adjacent CpGs. The relationship between average methylation and epipolymorphism at each e-locus was evaluated graphically using a scatterplot. Furthermore, I evaluated whether average methylation and epipolymorphism may predict gene expression, as previously described [110]. Matched Epic-seq and RNA-seq data were obtained for a subset of ccRCC and normal kidney samples (N=47) (as described in section 4.2.2 and section 4.2.5 respectively). First, I evaluated a linear model predicting gene expression based on epipolymorphism, with a Benjamini-Hochberg (BH) correction for multiple testing. To ascertain the effect of epipolymorphism beyond methylation, I evaluated a linear model predicting gene expression based on methylation alone or methylation and epipolymorphism and compared the adjusted R^2 from the two models using a likelihood ratio test. The analysis was performed for individual e-loci, and where multiple e-loci were significant for one gene, the e-locus with the lowest BH adjusted p values was shown, along with the number of e-loci per gene. An adjusted p value of <0.05 was considered significant.

4.7.4 Homogeneously vs heterogeneously methylated CpGs

CpGs were defined as homogeneously and heterogeneously methylated as described by Hao et al [220] (Table 4.4). I performed the analysis for each individual ccRCC patient separately. CpGs were included in the analysis if a coverage ≥10x was achieved for all multi-region samples from one patient. First, CpGs were identified that distinguish tumour from normal (i.e. average methylation difference in tumour vs normal samples is ≥25%). The 25% threshold was selected as this a commonly used cut-point in the literature [187], and also the threshold used to call differences between tumour and normal in my analysis (see section 4.4.3). Subsequently, these CpGs were defined as homogeneously methylated if there was ≤15% methylation difference amongst any tumour samples within a patient; and heterogeneously methylated if there was ≥40% methylation difference amongst any tumour samples within a patient. CpGs that were recurrent in over one third of the patient cohort (i.e. heterogeneously or homogeneously methylated in over 6 out of 18 patients) were considered to be more likely to be clinically significant. Table 4.4 explains the rationale behind the thresholds used. Hao et al found that differing the choice of thresholds produced broadly similar results [220] (just more or less stringent numbers of CpGs called), and this was also observed when varying the thresholds in my dataset. It is recognised that differing tumour purity content of the samples analysed may confound this analysis. Therefore, the analysis was repeated after removal of low purity samples.

Table 4.4: Homogeneously and heterogeneously methylated CpGs, definition and rationale

Threshold definition	Justification
Homogeneously methylated CpGs CpGs distinguish tumour from normal (≥25%) and have *similar* methylation patterns in all tumour samples (≤15% methylation difference amongst any tumour samples within a patient).	• CpGs were identified that distinguish tumour from normal kidney (i.e. average methylation in normal samples is ≥25% different to average methylation in tumour samples, within a patient). • Two technical replicates were assessed on the Epic-seq platform (i.e. gDNA from the same human tissue sample was assessed twice). For each CpG, the absolute methylation difference between the two technical replicates was calculated. The third quartile was 10%, meaning that CpG methylation can vary by 10% in technical replicates. This is expected if coverage 10x is used, as one error can cause 10% variation in methylation values. Therefore, CpGs were considered homogeneously methylated provided the difference between two tumour samples was ≤15%.
Heterogeneously methylated CpGs CpGs distinguish tumour from normal (≥25%) and have *different* methylation patterns in tumour samples (≥40% methylation difference amongst any tumour samples within a patient).	• CpGs were identified that distinguish tumour from normal kidney (i.e. average methylation in normal samples is ≥25% different to average methylation in tumour samples, within a patient). • Given the 15% thresholds used for homogeneously methylated CpGs, 40% (i.e. 25% plus 15%) was used for heterogeneously methylated CpGs. In other words, we seek to identify CpGs that have a greater methylation difference between tumour samples, than between tumour and normal kidney.

4.8 Analysis of DNA methylation in liquid samples

Targeted methylation analysis in cfDNA was performed using Nimbus. Nimbus couples library preparation with an automated bio-informatics pipeline (Nimbus refers to both wet lab and dry lab experimental methods). The wet lab experimental methods are described in section 4.2.3. Below, I describe how the targeted methylation panel was selected (section 4.8.1) and how Nimbus scores were generated and analysed (section 4.8.2).

4.8.1 Selection of informative methylation marker panel in tissue to be used in Nimbus

I first sought to create a custom methylation marker panel specific to ccRCC which could be used for cfDNA detection using Nimbus. I performed genome-wide DNA methylation analysis in tissue to identify a panel of differentially methylated regions (DMRs) that can distinguish ccRCC from normal kidney and are therefore useful methylation markers for cfDNA analysis. I generated Epic-seq methylation data on 75 fresh frozen kidney tissue samples (53 ccRCC and 22 normal kidney), as described in section 4.2.2. For this analysis, this was subsequently referred to as the 'discovery cohort.' I used 'dmrseq' to determine DMRs that distinguish ccRCC from normal kidney (as described in section 4.4.3). Subsequently, these tissue-derived DMRs were refined to select those which are most likely to be informative in plasma cfDNA. Data were obtained from 32 healthy (cancer-free) controls from a previously published study [221]. DMRs were selected if there was a >60% methylation difference between ccRCC tissue and healthy control cfDNA samples (analysis performed by Dr Radoslaw Lach, Massie Group). These DMRs were used in the Nimbus capture panel.

Furthermore, I explored these DMRs in tissue to confirm that these may be appropriate markers to take forward into subsequent analysis. I compared the number of CpGs contained within each DMR, for hyper and hypomethylated regions respectively. A representative DMR was visualised in tissue using the 'dmrseq' package v1.6.0 in R [188]. Subsequently, I assessed DNA methylation at these DMRs in the discovery cohort, and an additional independent cohort of samples. The latter, termed the 'validation cohort', consisted of 159 ccRCC and normal kidney tissue samples. I generated Epic-seq data for these samples (as described in section 4.2.2) and obtained methylation values for the DMRs. First, I performed principal component analysis (PCA) for the discovery cohort. Next, I projected the validation cohort into the PCA space of the first dataset. Both datasets were visualised using a PCA plot.

4.8.2 Targeted methylation analysis in liquid samples

Targeted DNA methylation analysis was performed using Nimbus, for cell line and human samples. I derived quality control metrics from 'MultiQC' and 'Picard tools' (see section 4.4.1). A Bland Altman plot was used to compare a gDNA human sample run using Nimbus and Epic-seq respectively (for CpGs which achieved ≥10x coverage using both methods). The analysis was performed using the 'BlandAltmanLeh' package v0.3,1 in R [222]. Correlation between technical replicates was assessed using Pearson's correlation coefficient (for on-target CpGs which achieved ≥10x coverage).

First, Nimbus was run on cfDNA derived from plasma from patients with and without ccRCC (N=67). Nimbus scores were generated by Dr Radoslaw Lach (postdoctoral research associate, Massie Group) for each sample using the Nimbus analysis pipeline, for DMRs which are hypermethylated and hypomethylated in ccRCC respectively. All subsequent analyses were performed by me. I evaluated the ability of the Nimbus score to differentiate ccRCC from control samples. I calculated the ROC curve and selected the Nimbus score which maximized sensitivity and specificity, using the 'pROC' package [223]. Furthermore, I sought to compare cfDNA detection rates using methylation (Nimbus) versus mutational analysis (INVAR-TAPAS). For a subset of cfDNA plasma samples from ccRCC patients evaluated by Nimbus (N=14), mutational analysis of cfDNA was performed by Dr Chris Smith (Rosenfeld Group, CRUK CI) using the 'INtegration of Variant Reads-Tailored Panel Sequencing' (INVAR-TAPAS) pipeline. In brief, nephrectomy samples were sequenced using WES (as described in section 4.2.4) to identify patient-specific mutations. Custom panels were created to detect cfDNA in plasma collected prior to nephrectomy. Dr Smith provided me with cfDNA detection rates and mutant allele fraction (MAF) estimates for each patient, from his recent publication [158]. I compared detection rates using Nimbus versus INVAR-TAPAS using a chi squared test for proportions. I also evaluated the association between Nimbus scores, MAF estimated by INVAR-TAPAS and clinical parameters. For continuous variables, Pearson's correlation coefficient was used, whereas for non-parametric categorical variables, the Wilcoxon Rank Sum Test was used. All p values were corrected for multiple testing (using the BH method).

Lastly, I performed a feasibility study evaluating Nimbus scores in cfDNA derived from post-biopsy fluid (N=11), plasma (N=11) and gDNA derived from renal biopsy tissue (N=8). Sample and collection details are found in section 4.1.2.1. Nimbus scores were generated for cfDNA derived from matched samples by Dr Lach, using the automated analysis pipeline and hypomethylated DMRs. I compared Nimbus scores in matched samples using a Wilcoxon Rank Sum Test for paired data (adjusting for multiple testing) and evaluated the association between Nimbus scores and clinical parameters.

Chapter 5 DNA methylation in tissue from common pathological subtypes of malignant and benign renal tumours

5.1 Brief introduction

Small renal masses (SRMs) represent a diverse cohort of potential diagnoses, including malignant renal cell carcinoma and benign tumours (such as oncocytoma or angiomyolipoma). Despite the recent drive to increase rates of renal biopsy, biopsies remain underutilised in certain centres, typically due to inadequate service provision/local expertise, fear of complications or perceived limited impact on treatment choice [27]. When biopsies are performed, pathologists may struggle to differentiate subtypes, meaning biopsy may be inconclusive or non-diagnostic (especially if limited tissue is sampled or there is evidence of necrosis) [35, 45]. In particular, it may be difficult to distinguish benign oncocytomas from the eosinophilic variant of chRCC and ccRCC, or to distinguish ccRCC from chRCC. There are also challenges within RCC subtypes, for example ccRCC is characterised by intra-tumoral heterogeneity, and often tumour grade is variable depending on the region which is biopsied [45, 46]. Pathological subtype and grade are major determinants of prognosis and therefore treatment choice in patients with SRMs, thus the limitations of biopsy histology may in certain cases result in either under- or over-treatment (the latter being more common). In summary, current methods are unable to confidently determine tumour pathology and grade in all patients. Thus approximately 20%-30% of patients with SRMs are found to have benign disease post-operatively, meaning they underwent unnecessary surgery, with associated morbidity and potential long-term effects on renal function [38, 39]. There is a drive to reduce overdiagnosis and overtreatment across the healthcare spectrum, including the BMJ's 'Too much medicine' campaign, especially in the early detection of cancer [224, 225]. Precision medicine, using molecular classification in combination with histopathology assessment, may be an approach to reduce overtreatment and I hypothesize that DNA methylation markers could be used to achieve this.

A better understanding of the molecular characteristics of renal tumour pathological subtypes may facilitate improved diagnostic and management strategies. DNA methylation changes are abundant, genome-wide, early events in renal tumorigenesis, and are specific to the cell of origin [74, 101, 103, 115, 226]. I therefore hypothesized that a comprehensive characterisation of DNA methylation in different pathological subtypes of renal tumours would improve our understanding of the disease and consequently our ability to diagnose patients with SRMs. The analysis in this chapter sets out to

explore these possibilities. Although several potential diagnoses exist, my research focuses on the three most common malignant subtypes (ccRCC, pRCC and chRCC) and the most common benign disease (oncocytoma). Previous studies evaluating DNA methylation and gene expression noted similarities between ccRCC and pRCC, and chRCC and oncocytoma, and these have been attributed to their common cell of origin [74, 117, 227, 228]. Renal cell carcinomas arise from epithelial cells within the kidney tubule, with ccRCC and pRCC deriving from the proximal convoluted tubule (PCT), whilst chRCC and oncocytoma derive from the intercalated cells of the distal nephron (Figure 2.1). Considerable research has been performed elucidating the cell of origin of ccRCC and pRCC, with PT1 cells (an epithelial cell subtype derived from the PCT) having been identified as the most likely progenitor [25]. Conversely, chRCC and oncocytoma remain less well characterised with some debate regarding whether they are derived from the collecting duct or distal convoluted tubule (DCT); the former being more likely based on recent single cell analyses [25, 228]. Although a number of studies are present in the literature which focus on gene expression (including studies evaluating bulk RNA-seq and single cell RNA-seq), only a handful of studies focus on characterising DNA methylation in the different pathological subtypes of renal tumours and these studies do not integrate DNA methylation with gene expression. Therefore, in the first part of this chapter, I aim to evaluate DNA methylation and gene expression to characterise similarities and differences between pathological subtypes.

As already highlighted, the main clinical question, which was recently recognised as a research priority, is to distinguish pathological subtypes of renal masses to improve the diagnostic pathway [1]. Several predictive models have been developed to tackle this question (described in more detail in the Discussion section of this chapter), however none have been adopted in clinical practice, reflecting the lack of adequate validation of these models. In the second part of the chapter, I therefore aim to create and validate MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours), a machine learning model based on DNA methylation that can be applied to tissue samples to differentiate pathological subtypes. It is envisioned that MethylBoostER could be used on renal biopsy samples, to predict the most likely pathological subtype, providing a more confident pre-surgical diagnosis to guide treatment decision making. I also explore the role of intra-tumoral heterogeneity and tumour purity on the output of MethylBoostER, as these real-world challenges may be barriers towards adoption in clinical practice.

5.2 Chapter aims

1) Characterise DNA methylation in the most common pathological subtypes of renal tumours (ccRCC, pRCC, chRCC and oncocytoma) comparing these to normal kidney
2) Evaluate gene expression in different pathological subtypes of renal tumours (ccRCC, pRCC, chRCC and oncocytoma) comparing these to normal kidney to highlight similarities and differences amongst subtypes
3) Integrate methylation data with gene expression data in different pathological subtypes of renal tumours to evaluate the functional relevance of methylation changes
4) Evaluate known methylation markers of pathological subtypes from the literature in my cohort of samples
5) Develop a model to classify pathological subtypes of renal tumours using methylation data, with the aim of differentiating pathological subtypes in a diagnostic setting

5.3 Results

This chapter can broadly be categorised into two sections. The first part (section 5.3.1) focuses on characterising methylation and gene expression to improve our understanding of similarities and differences between tumour subtypes. The second section (section 5.3.2) develops a machine learning model to predict pathological subtypes of renal tumours in a diagnostic setting. Details regarding samples, experimental methods and data analyses are found in the Methods (Chapter 4).

5.3.1 Methylation and gene expression in pathological subtypes of renal tumours *vs* normal kidney

The aim of this section was to explore similarities and differences between pathological subtypes of renal tumours using DNA methylation and gene expression. Here, I provide an outline of my analysis, samples used and justify the rationale (Figure 5.1). In summary, I characterised patterns of DNA methylation (section 5.3.1.1) in pathological subtypes of renal tumours, subsequently I evaluated gene expression (section 5.3.1.2), and lastly methylation and gene expression were integrated to determine methylation changes which may have a functional relevance (section 5.3.1.3) (Figure 5.1).

In this analysis, I obtained fresh-frozen kidney tissue samples from a cohort of patients with different pathological subtypes of renal tumours undergoing nephrectomy (N= 326 samples) and generated

methylation values using Epic-seq (see Methods section 4.2.2 for details). I analysed these data, comparing each pathological subtype (ccRCC, pRCC, chRCC and oncocytoma) to normal kidney to characterise DNA methylation changes (section 5.3.1.1). For a small subset of ccRCC and normal tissue specimens, I generated matched RNA-seq (N=47 samples); though unfortunately this was not possible for the other pathological subtypes. Therefore, I obtained publicly available [176] Affymetrix gene expression data for tissue samples (N=84) and compared each pathological subtype (ccRCC, pRCC, chRCC and oncocytoma) to normal kidney, to explore differential gene expression (section 5.3.1.2; Figure 5.1). I then integrated my Epic-seq data with Affymetrix gene expression to assess which methylation markers might be associated with transcriptional changes in each pathological subtype (ccRCC, pRCC, chRCC and oncocytoma) (section 5.3.1.3). In order to validate my findings, I evaluated matched Epic-seq and RNA-seq in the cohort of 47 ccRCC and normal kidney samples that I generated, and explored the quantitative relationship between methylation and gene expression. In addition, for ccRCC, pRCC and chRCC, I validated the association between methylation and gene expression by analysing samples from The Cancer Genome Atlas (TCGA) [24]. No TCGA data exist for oncocytoma, therefore it was not possible to validate my results in this subtype. Whilst some analysis is performed comparing ccRCC versus normal kidney in this chapter, further analysis is found in Chapter 6, which focuses on ccRCC markers, evolution and heterogeneity.

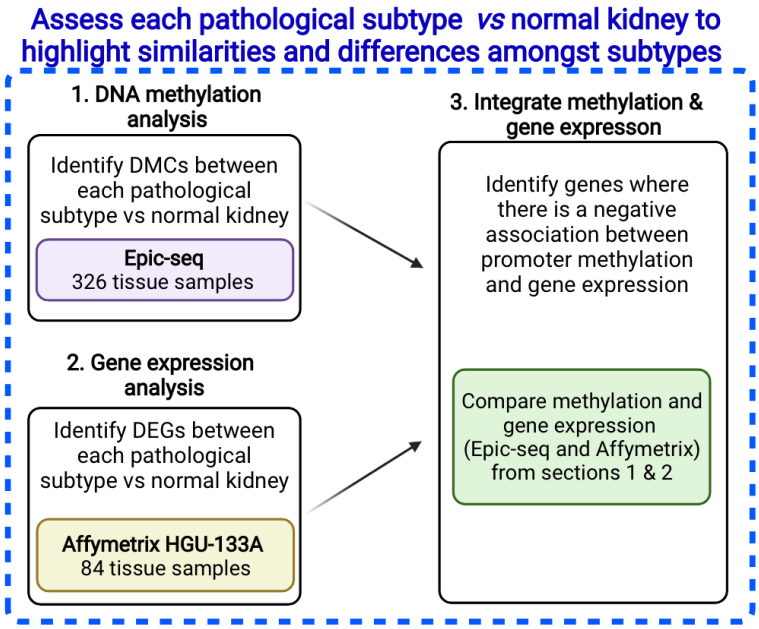


Figure 5.1: Analysis overview.

In this chapter, first I evaluate DMCs (differentially methylated cytosines) in each subtype vs normal kidney (section 5.3.1.1), then I evaluate DEGs (differentially expressed genes) in each subtypes vs normal kidney (section 5.3.1.2). Lastly, I integrate methylation and gene expression data to highlight which methylation changes may have a functional relevance (section 5.3.1.3).

5.3.1.1 Differential methylation in pathological subtypes of renal tumours compared to normal kidney

In order to characterise global methylation patterns in pathological subtypes of renal tumours, I obtained kidney tissue samples from patients with malignant and benign renal tumours (ccRCC, pRCC, chRCC and oncocytomas) undergoing partial or radical nephrectomy. Tissue was sampled from the tumour and adjacent normal kidney, and Epic-seq was performed (see Methods sections 4.1.1 and 4.2.2 for details). I generated methylation data for 326 samples, including 109 adjacent normal kidney, 139 ccRCC, 27 pRCC, 27 chRCC and 24 oncocytomas. Unsupervised hierarchical clustering based on all the methylation data (~2.3 million CpGs), demonstrated that, as expected, chRCC and oncocytomas cluster together, whereas pRCC and ccRCC cluster together (Figure 5.2A). This is in keeping with the shared cell of origin for these pathological subtypes [117]. Interestingly, oncocytoma and chRCC cluster more closely with normal samples. Furthermore, these are the two subtypes that seem most difficult to separate based on genome-wide methylation patterns (Figure 5.2A-B), which also reflects the predominant clinical challenge. In the principal component analysis (PCA), normal samples clustered very closely together and there was less variability between normal samples than between samples from any other subtypes; with the most heterogeneity noted in ccRCC (Figure 5.2B). Methylation intra-tumoral heterogeneity (ITH) in ccRCC is explored more in detail in the next chapter (Chapter 6). It is evident that a subset of type 2 pRCC samples cluster together, and are separate from other subtypes (as shown in the upper right corner of the PCA plot; Figure 5.2B). I hypothesized these samples may have a CpG island methylator phenotype (CIMP) based on previous reports [24], and this was confirmed upon further investigation in the following section (Figure 5.3B-C).

In a more detailed analysis of these data, I first sought to evaluate normal samples (N=109), to assess any evidence of field-effects or other pathology-specific alterations in normal adjacent kidney tissue. In brief, these normal samples were derived from patients with different diagnoses (i.e. ccRCC, pRCC, chRCC and oncocytoma) and they represent normal kidney macroscopically distant from the tumour site. Figure 5.2C demonstrates a dendrogram of normal samples, with each colour indicating the diagnosis of the patient's adjacent tumour. Evidently, samples do not cluster by adjacent tumour type. Next, I evaluated methylation differences in normal tissue samples derived from patients with ccRCC (N=70) and patients with chRCC (N=14), as these were the two most abundant classes representing two different cells of origin. Comparing normal tissue samples from patients with ccRCC versus chRCC identified only one CpG which was called as significantly different (chr20:58,088,274). This CpG is not in the promoter region of any gene and may reflect a type-II error, given the large number of CpGs assessed (>2 million). This suggests there is no methylation

difference between the two groups. Since the methylomes of normal samples were very similar regardless of concurrent tumour type, for all subsequent analyses normal samples were combined together as one class. This approach has also been taken by other studies in the literature [121].

Figure 5.2: DNA methylation in pathological subtypes of renal tumours and normal kidney

See Figure on the following page

Panel A and B- Dendrogram and principal component analysis (PCA) of all tissue samples, demonstrating that in unsupervised analysis based on all CpGs, normal kidney samples cluster together whereas tumour subtypes cluster by cell of origin (ccRCC cluster with pRCC and chRCC cluster with oncocytoma). Panel C- PCA of normal tissue samples, where each colour represents the diagnosis of the patient's adjacent tumour. Indeed, normal samples were derived from patients with different diagnoses (i.e. ccRCC, pRCC, chRCC and oncocytoma) and they represent normal kidney samples located macroscopically away from the tumour site. Normal samples do not cluster by adjacent tumour type. Panel D- Venn diagram showing DMCs (differentially methylated cytosines) which differentiate each pathological subtype from normal kidney. The greatest similarity between subtypes was noted between chRCC and oncocytoma; and between ccRCC and pRCC, in keeping with the known shared cell of origin. Panel E- Hyper and hypomethylated DMCs for each pathological subtypes vs normal tissue are annotated. Hypermethylated DMCs are distributed more often at the promoter site, compared to hypomethylated DMCs, which are distributed more often in the gene body or distal intergenic region.

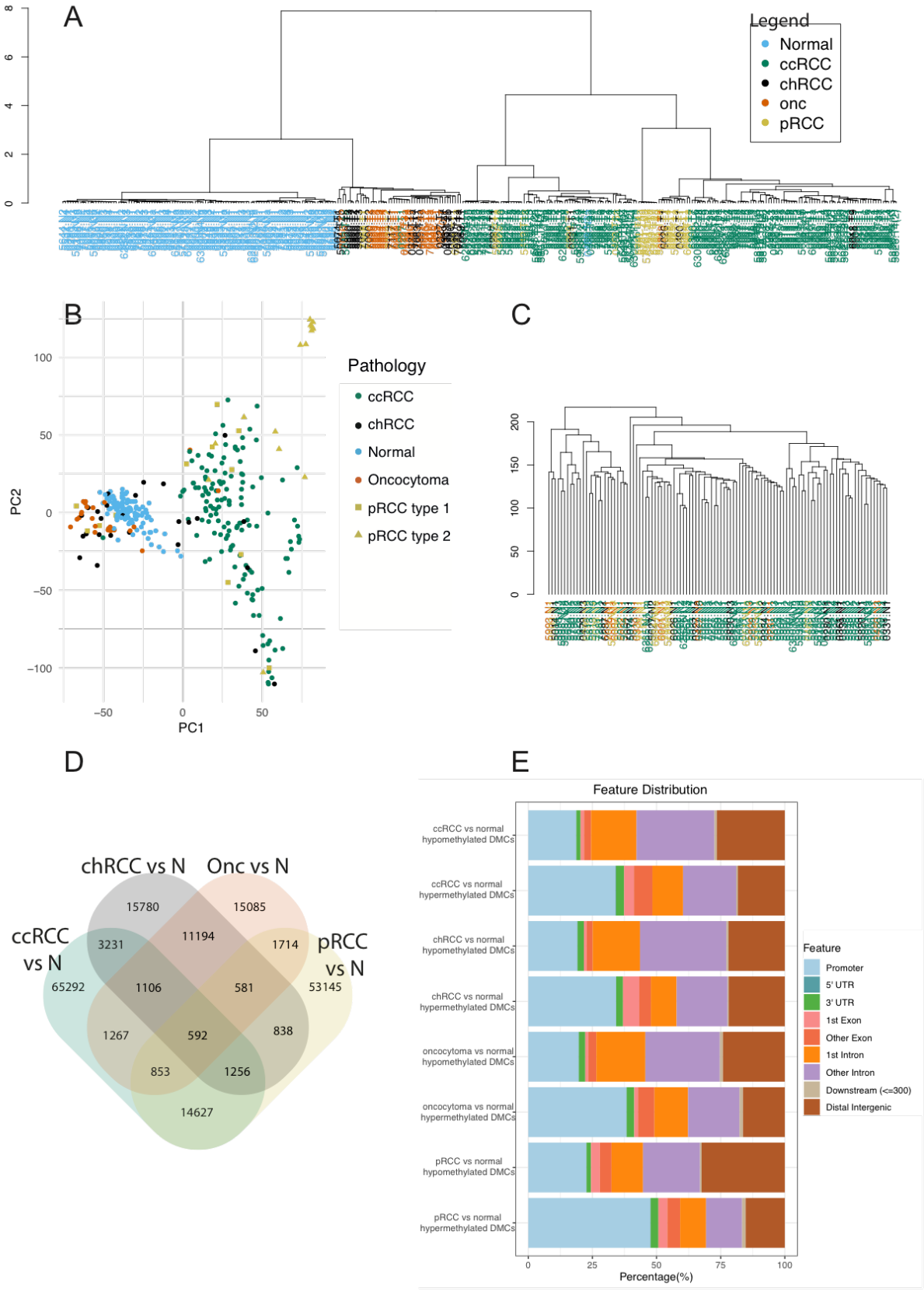


Figure 5.2: DNA methylation in pathological subtypes of renal tumours and normal kidney.

Subsequently, I performed differential DNA methylation analysis to determine global methylation patterns which characterise individual subtypes of renal tumours, highlighting similarities and differences. Methylation was compared for each subtype versus normal kidney tissue, as previously described (section 4.4.3) [121]. Table 5.1 highlights the number of significant differentially methylated cytosines (DMCs) for each comparison, after removal of CpGs located on sex chromosomes and at SNP sites (q value <0.01, methylation difference >25%). In ccRCC, there were more hypomethylated than hypermethylated CpGs compared to normal kidney (approximately 57,000 vs 33,000 DMCs; Table 5.1). Both oncocytoma and chRCC were characterised by pronounced hypomethylation relative to normal tissue. In chRCC there were nearly ten-fold more hypo than hypermethylated sites (31,802 hypomethylated vs 3231 hypermethylated), whereas pRCC was characterised by pronounced hypermethylation compared to normal tissue (Table 5.1). The significant DMCs from each subtype were overlapped to enable a comparison (Venn diagram in Figure 5.2D). The greatest similarity between subtypes was noted between chRCC and oncocytoma, with over 11,000 shared DMCs; and between ccRCC and pRCC, with over 14,000 shared DMCs (Figure 5.2D). The similarity between subtypes has been noted in the literature and is believed to reflect their common cell of origin [117]. Fewer DMCs were identified for chRCC (~35,000) and oncocytoma (~33,000), than ccRCC (~90,000) and pRCC (~75,000), suggesting the former might be more similar to normal kidney (Table 5.1). Alternatively, this could suggest that methylation may play a lesser role in mediating tumorigenesis in chRCC and oncocytoma, relative to ccRCC and pRCC. To explore this concept further, in subsequent analyses (below and in section 5.3.1.3), I performed annotation and gene set enrichment of the DMCs to define pathways which may be epigenetically regulated, for each subtype.

Table 5.1: DMCs identified in pairwise comparisons between each pathological subtype and normal tissue

Differentially methylated cytosines (DMCs) are subdivided into hyper and hypomethylated CpGs based on methylation at the tumour site relative to normal kidney tissue. Significant DMCs are defined as DMCs with an absolute methylation difference >25% and a q value <0.01. The total number of CpGs evaluated varies for each comparison, as described in section 4.4.3.

	ccRCC vs normal	pRCC vs normal	chRCC vs normal	Oncocytoma vs normal
Total CpGs evaluated	2,850,419	1,543,085	1,527,799	1,500,195
Hypermethylated DMCs	33,399	57,832	3231	5610
Hypomethylated DMCs	57,334	17,187	31,802	27,304
Sum of significant DMCs	90,733	75,019	35,033	32,914

I performed DMC annotation, as location in the genome can offer insights into function (for example promoter hypermethylation is associated with transcriptional repression) [103, 104]. In all of the tumour versus normal comparisons, hypermethylated DMCs were distributed more often at gene promoter sites, compared to hypomethylated DMCs, which were distributed more often in gene bodies or distal intergenic regions (Figure 5.2E). This is in line with the literature, which suggests that cancer is characterised by local hypermethylation at gene promoters (often associated with transcriptional silencing) and genome-wide hypomethylation [229]. For example, in pRCC versus normal kidney, 48% of hypermethylated DMCs were in gene promoter regions compared to only 23% of hypomethylated DMCs. In addition, hypermethylated DMCs in pRCC vs normal kidney were more often in the promoter region relative to the other subtypes (48% in pRCC compared to 34%, 34% and 38% in ccRCC, chRCC and oncocytoma, respectively). Promoter hypermethylation is associated with an oncogenic phenotype [133, 230], so it may be that hypermethylation plays a greater role in pRCC tumorigenesis than other subtypes. This observation prompted me to evaluate CIMP in my dataset.

A landmark study found that a subset (5-12%) of pRCC type-2 tumours are characterised by CIMP and this is associated with poor overall survival [72]. Interestingly, a subset of samples characterised by CIMP and poor prognosis have been found in all tumour pathological subtypes (including chRCC, which is usually characterised by hypomethylation) [24]. Arai et al identified 16 CpGs which are hypermethylated and characteristic of the aggressive CIMP phenotype [132]. Methylation values were available for 13 of these CpGs in my dataset and were used to cluster tumour samples in order to explore the CIMP phenotype across RCC subtypes (Figure 5.3A). It is evident that there are two major branches in the dendrogram: the first branch (on the left) contains a subset of pRCC type-2 and ccRCC samples which are characterised by hypermethylation, whereas the second branch is characterised by relative hypomethylation, with all normal samples tending to cluster together on the far right. Re-evaluating the PCA from Figure 5.2B, it is evident that the pRCC type-2 and ccRCC samples which are characterised by hypermethylation in the CIMP CpGs are also the outliers in the PCA (Figure 5.3B-C). This suggests that the 13 CpGs identified by Arai et al are representative of a global hypermethylation phenotype, which sets these tumours apart. In my dataset, all the pRCC type-2 samples identified as CIMP positive had promoter hypermethylation of the *CDKN2A* gene (Figure 5.3D). In TCGA, *CDKN2A* alterations (focal loss of 9p21, mutation or promoter hypermethylation) were observed in 100% of CIMP pRCC, as well as 5-25% of other pathological subtypes [24], and *CDKN2A* silencing was associated with worse overall survival [24, 72]. Previous studies have suggested that approximately half of patients with CIMP have germline or somatic *FH*

mutations, which leads to impaired TET activity (an enzyme responsible for de-methylation) which may explain the observed dominant pattern of promoter hypermethylation [231]. Unfortunately, no mutational data nor recurrence data were available in my cohort (the only recurrence data available was for ccRCC but not the other subtypes), so it was not possible for me to assess CIMP and prognosis across subtypes. For the remainder of this chapter, pRCC type-1 and type-2 were combined into one class to increase sample sizes, however it is recognised that the two subtypes may have distinct molecular characteristics and prognostic outcomes. Future work could evaluate subtypes of pRCC and also characterise associations between DNA methylation and prognosis in different pathological subtypes of malignant renal tumours, given larger numbers of samples and detailed follow-up data. In summary, this section highlights global methylation patterns that characterise each pathological subtype. Integrating these methylation changes with gene expression data is key to determine which of these DMCs are biologically meaningful.

Figure 5.3: CpG island methylator phenotype (CIMP) in tissue samples

See Figure on the following page

Panel A- Heatmap showing methylation levels in CpGs which are known to be hypermethylated in CIMP positive tissues. In the heatmap, samples are shown as columns and CpG probes are shown as rows. A subset of pRCC type-2 and ccRCC samples (left branch of the dendrogram) are clearly characterised by hypermethylation at these sites, and are therefore referred to as 'CIMP positive'. Panel B- Principal component analysis (PCA) of all tissue samples, based on all CpGs. This figure is the same as Figure 5.2B. Panel C- PCA of all tissue samples, based on all CpGs, highlighting CIMP positive (CIMP +) and CIMP negative (CIMP -) tumours. The CIMP positive samples (i.e. subset of pRCC type-2 and ccRCC) which were characterised by promoter hypermethylation in Panel A, are clear outliers on the PCA, suggesting these are different to the rest of the samples. Panel D- Methylation levels at the promoter region of the *CDKN2A* gene for normal kidney samples and pRCC type-2 samples identified as CIMP positive. All the pRCC type-2 CIMP positive samples demonstrate promoter hypermethylation.

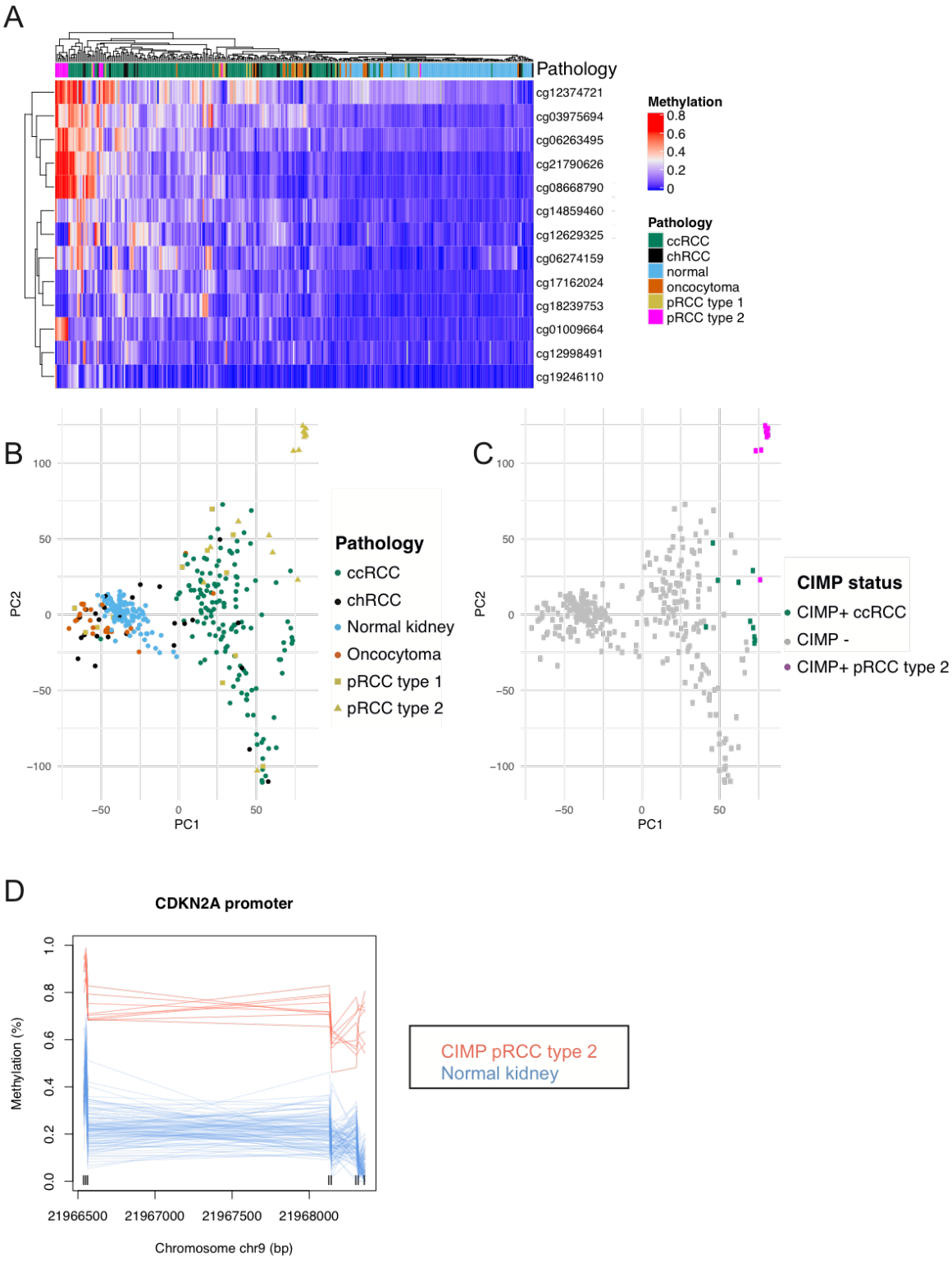


Figure 5.3: CpG island methylator phenotype (CIMP) in tissue samples

5.3.1.2 Differential gene expression in pathological subtypes of renal tumours compared to normal kidney

In order to explore patterns of gene expression that characterise pathological subtypes of renal tumours and to identify genes regulated by recurrent epigenetic alterations, I analysed publicly available Affymetrix gene expression data (GEO accession number: GSE15641) [176]. In the absence of paired DNA methylation and RNA expression data, I sought to overlay recurrent epigenetic changes with recurrent differential expression in order to map functional epigenetic alterations in renal tumour subtypes. In this section, I explore differentially expressed genes (DEGs) between tumour subtypes, subsequently I integrate gene expression and methylation in section 5.3.1.3.

In this publicly available dataset, transcriptional profiling of kidney tumours was performed using Affymetrix HGU-133A chips, for 84 fresh-frozen samples, including 32 ccRCC, 11 pRCC, 6 chRCC, 12 oncocytoma and 23 normal kidney. Normalised data were provided by the authors for 22,283 probes (see Methods section 4.4.6 for details). After processing and filtering, I obtained gene expression data for 12,606 genes. The PCA plot based on gene expression values for all genes demonstrated that chRCC and oncocytomas clustered closely together and these were closer to normal, whereas ccRCC and pRCC clustered together (Figure 5.4A). This resembles very closely the PCA produced using Epic-seq DNA methylation data (Figure 5.2B).
I identified significant DEGs for each pathological subtype compared to normal tissue, using a linear model with contrasts for disease subtypes compared to normal kidney tissue (see Methods section 4.4.6 for details) [195]. Table 5.2 summarises the number of significant genes for each pairwise comparison (adjusted p value <0.01; absolute log2 fold change > 1) and the Venn diagram depicts the overlap between the subtypes (Figure 5.4B). Notably, chRCC is the subtype with the least differences compared to normal tissue (Table 5.2), and this pattern was also observed in the methylation analysis (Table 5.1). Indeed, the molecular similarity of chRCC to normal kidney tissue has also been noted in the literature [121].

Table 5.2: Results of differential gene expression analysis

The number of significant differentially expressed genes (DEGs) is shown (q value <0.01 and absolute log2 fold change >1).

	Significant DEGs
ccRCC vs normal	1558 DEGs (675 increased expression and 883 decreased expression)
pRCC vs normal	1326 (388 increased expression and 856 decreased expression)
chRCC vs normal	1007 (252 increased expression and 755 decreased expression)
Oncocytoma vs normal	1174 (398 increased expression and 776 decreased expression)

Similarities and differences in gene expression in the four subtypes were assessed, to identify shared and subtype-specific expression profiles. As expected, the Venn diagram demonstrates that subtypes derived from the same cell of origin share more DEGs (Figure 5.4B). However, there are also similarities amongst the four subtypes, with 187 genes differentially expressed in all subtypes compared to normal tissue (Figure 5.4B). These DEGs were ranked based on the greatest expression difference (i.e. highest absolute log2 fold change in all subtypes), to elucidate common differences between renal tumours and normal tissue. Amongst the top ranked genes, increased expression of genes involved in metabolic reprogramming was noted: *SLC16A3* and *SLC38A1* (a monocarboxylate and glutamine transporter, respectively), *APOC1* (apolipoprotein involved in HDL and LDL metabolism), *ALDOA* (glycolytic enzyme which regulates adaptation to hypoxia) and *PKM* (pyruvate kinase glycolytic enzyme). These alterations in expression of key metabolic regulators are consistent with the metabolic reprogramming observed in all four renal tumour types, although the underlying metabolic alterations are distinct in each renal tumour subtype. The Warburg effect, defined by increased anaerobic glycolysis, is a key feature of ccRCC and pRCC [232] and is indeed a hallmark of cancer [233]. In the literature, pRCC is characterised by disturbances in TCA cycle genes (such as *FH* mutations) and CIMP pRCCs are characterised by increased glycolysis [72, 234], whereas oncocytomas and chRCC (particularly the eosinophilic subtype) are characterised by mitochondrial defects [235]. These divergent routes to metabolic dysfunction may be reflected in distinct clinical phenotypes, but the shared expression changes in metabolite transporters and rate-limiting enzymes may reflect shared phenotypes arising from the cell of origin or adaptations common to all renal tumour subtypes. Further work will be required to dissect the origins and functional importance of these shared metabolic expression changes that could have diagnostic and/or therapeutic implications.

Compared to normal kidney, all four renal tumour subtypes demonstrated down-regulation of genes implicated in renal development, possibly reflecting a loss of terminal differentiation in renal tumours (Figure 5.4B). These down-regulated genes included markers of distal convoluted tubule cells (DCT: *UMOD*, *SLC12A1* and *CALB1*) and proximal convoluted tubule markers (PCT: *ASS1*, *PAH*, *ALDOB*, *BBOX1*, *SLC22A8* and *KNG1*). The gene *UMOD* has been independently found to be amongst the top DEGs for renal tumours [236]. In particular, *SLC12A1* downregulation was more pronounced in ccRCC and pRCC (DCT marker); whereas *ASS1* and *BBOX1* were more downregulated in oncocytoma and chRCC (PCT markers) in keeping with the known cell of origin of these tumour types (Figure 5.4B). Having explored DEGs which are similar amongst subtypes, I then focused on DEGs which might be subtype-specific.

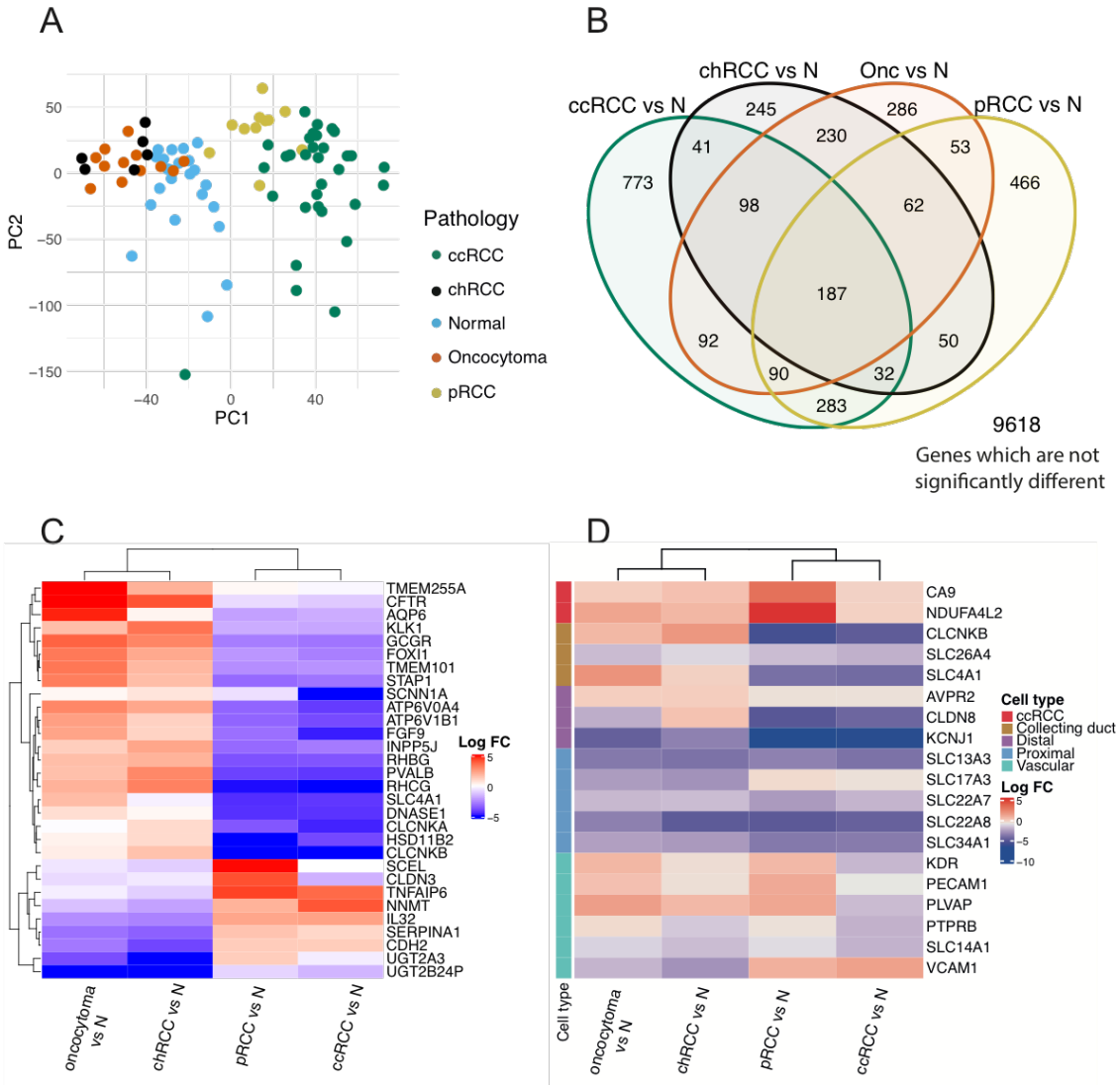


Figure 5.4: Gene expression in pathological subtypes of renal tumours and normal kidney.

Panel A- Principal component analysis of all tissue samples, using data from all available genes. In unsupervised analysis, tumour subtypes cluster by cell of origin (ccRCC cluster with pRCC and chRCC cluster with oncocytoma). Panel B- Venn diagram showing DEGs (differentially expressed genes) which differentiate each pathological subtype from normal kidney. The greatest similarity between subtypes was noted between chRCC and oncocytoma; and between ccRCC and pRCC. Panel C- Heatmap demonstrating the top genes with the highest differential expression between pathological subtypes of renal tumours and normal kidney. The log2 fold change (abbreviated to log FC in the diagram) is shown (for each subtype compared to normal) and genes are ranked based on highest absolute log2 fold change (top 30 genes are shown). Panel D- Heatmap demonstrating differential expression between pathological subtypes of renal tumours and normal kidney, for genes which are known to be cell-type specific markers based on the literature [25].

Differentially expressed genes with the greatest differences between subtypes were selected for clustering analysis (Figure 5.4C), to highlight markers that best discriminate tumour subtypes. One of the most notable findings, is that the DEGs included cell-type-specific markers, thus reflecting the tumours' cell of origin. Oncocytoma and chRCC were characterised by increased expression of *KLK1* and *FOXI1*, whereas ccRCC and pRCC displayed reduced expression (Figure 5.4C). *KLK1* is a member of the kallikrein family of serine proteases, a gene family that includes the prostate cancer marker *KLK3* (also known as prostate specific antigen). In the kidney, *KLK1* regulates vasodilation and ion reabsorption, and has anti-tumour effects (including regulating angiogenesis, cell growth, proliferation and remodelling of the extracellular matrix) [237, 238]. Similarly to my findings, TCGA data has previously shown reduced *KLK1* expression in ccRCC and pRCC (as well as in breast, thyroid and uterine cancers), with increased expression in chRCC [239]. *KLK1* is localised to the DCT [237], which may explain these differences, and as such has been proposed as a marker to differentiate RCC subtypes [239].

In my data, *SLC4A1* was upregulated in oncocytoma (downregulated in ccRCC and pRCC, no change in chRCC), whereas chRCC had upregulation of *CLCNKB* (downregulation in ccRCC and pRCC, no change in oncocytoma; Figure 5.4C). These two genes are markers of intercalated cells of the collecting duct, as is *FOXI1*. There were also notable differences in genes that are part of the *FOXI1* regulated transcriptional network (*CLCNKA*, *RHBG*, and *RHCG*, *ATP6V0A4* and *ATP6V1B1*), with upregulation in distal nephron cancers and downregulation in proximal nephron cancers (Figure 5.4C). The *FOXI1* transcription factor is involved in the differentiation of intercalated cells in the distal nephron and regulation of renal ATP proton pumps [240]. These findings have been independently reported in several studies in the literature [227, 228, 240], providing further confirmation of the validity of the DEG analysis presented here.

In oncocytoma and chRCC, I observed an increase in the expression of *FGF9* and *TMEM255A*, changes which were absent in other subtypes (Figure 5.4C). *TMEM255A* expression has previously been used in a 44-gene model to predict subtypes of renal tumours [241]. TMEMs are a group of transmembrane proteins, a number of which have shown kidney cell-type-specific staining [242]. Fibroblast growth factor signalling (including *FGF9*) has been shown to be over-expressed in oncocytoma vs normal tissue in several studies [235]. Oncocytoma was characterised by increased expression of *AQP6* (reduced expression in ccRCC and pRCC) and indeed this has been proposed as a potential marker for this subtype [243]. Taken together, these findings suggest that differential expression in pathological subtypes of renal tumours at least in part reflects their cell of origin.

In addition to these cell-type markers, notable differences were seen in genes which are known to play a role in tumorigenesis. ccRCC and pRCC had increased expression of *NNMT*, *CDH2* and *TNFAIP6* compared to normal tissue, whereas these were downregulated in both chRCC and oncocytoma (Figure 5.4C). *NNMT* activates the *PI3K/Akt/SP1/MMP-2* pathway and promotes cell invasion in ccRCC, with knockdown of *NNMT* expression inhibiting ccRCC growth and metastasis in murine models [244]. *CDH2* is a marker of epithelial to mesenchymal transition (EMT), whereas *TNFAIP6* (TNF alpha induced protein 6) is a *HIF* target gene implicated in ccRCC progression [245]. *SCEL* and *CLDN3* were overexpressed in pRCC alone, with either under-expression or no change in the other subtypes (Figure 5.4C). Sciellin (encoded by *SCEL*) mediates EMT and has been identified as a marker for pRCC [246]. TCGA data demonstrated significantly higher expression of *SCEL* in pRCC versus normal tissue, with no evidence of increased expression in ccRCC and chRCC. The differential expression of *CLDN3* in pRCC tumours observed in my analysis is supported by other studies [247]. Claudins are a group of transmembrane tight junction proteins which are differentially expressed in various anatomical parts of the renal tubule, therefore regulating segment-specific kidney epithelial permeability to solutes, which is key to renal physiology [248]. As tight junctions, they also play a key role in cell adhesion, cell signalling pathways involving growth and differentiation, and have been implicated in a variety of cancers [249]. These subtype-specific differences have biological relevance and could also be harnessed as useful diagnostic markers.

Since exploratory analysis of DEGs identified distinct expression patterns at several established markers of cell ontogeny (Figure 5.4C), this was systematically assessed using renal cell lineage markers. The significant DEGs derived from my analysis of pathological subtypes were overlapped with a curated set of kidney lineage-specific biomarkers [25]. Once again, oncocytomas clustered with chRCC and ccRCC clustered with pRCC (Figure 5.4D). As expected, ccRCC had high expression of *CA9* and *NDUFA4L2* compared to normal tissue and other subtypes of tumours; these are known markers associated with ccRCC. ccRCC and pRCC had lower expression of genes associated with DCT cells (e.g. *KCNJ1*, *CLDN8*) and collecting duct (e.g. *CLCNKB*, *SLC4A1*), and increased expression of *VCAM1*. In contrast, oncocytoma and chRCC had reduced expression of markers associated with PCT cells (e.g. *SLC17A3*) and vascular cells (e.g. *VCAM1*). Overall, this targeted analysis is consistent with a strong influence of cell-of-origin on DEGs in renal tumour subtypes. The concordance with previous studies provides validation of these gene expression data and supports the integrated analysis with DNA methylation changes in the next section.

5.3.1.3 Integration of methylation and gene expression data to identify functional epigenetic alterations

A key biological question is whether the recurrent DNA methylation changes observed in renal tumour subtypes are functionally relevant or are merely epigenetic markers that arise through disease evolution. Although I had matched Epic-seq and RNA-seq for a subset of ccRCC versus normal kidney samples (N=47), I did not have matched data for the other three subtypes (due to sample availability). To my knowledge, there are no available datasets that include both methylation and gene expression for the same samples for all pathological subtypes of renal tumours (for example TCGA lacks data on oncocytomas). Therefore, I overlapped the significant DMCs from my methylation analysis from section 5.3.1.1 with genes which are differentially expressed in each subtype compared to normal in the Affymetrix data from section 5.3.1.2. I validated my findings regarding ccRCC versus normal kidney using the matched Epic-seq and RNA-seq data that I generated, as well as TCGA. For pRCC and chRCC, I validated my findings using TCGA data. Figure 5.5 provides a schematic overview of this analysis.

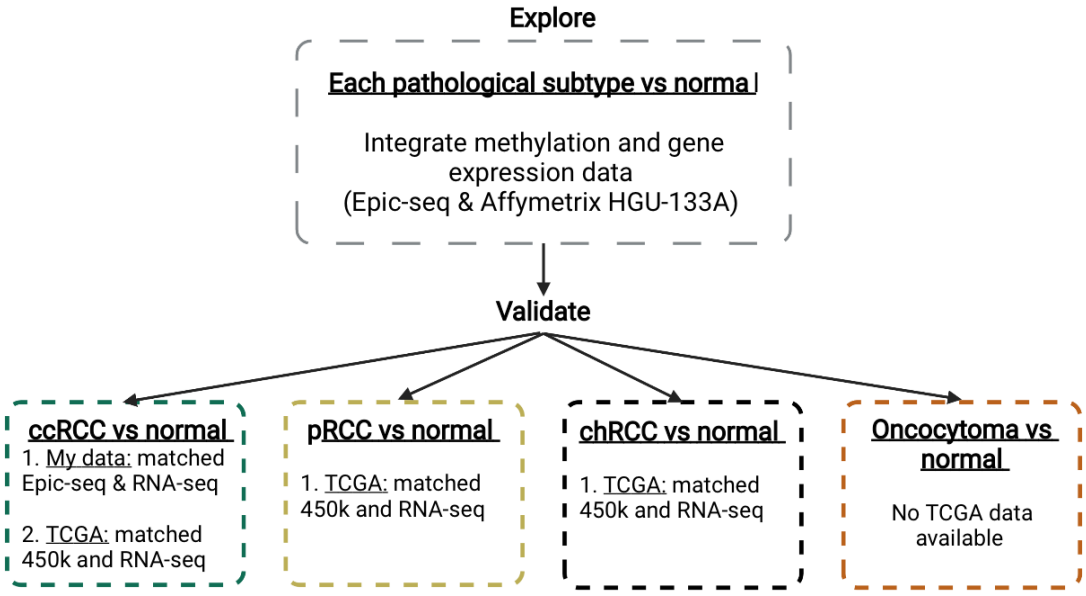


Figure 5.5: Overview of analysis integrating methylation and gene expression data

I overlapped the significant DMCs from my methylation analysis from section 5.3.1.1 with genes which are differentially expressed in each subtype compared to normal in the Affymetrix data from section 5.3.1.2. I validated my findings regarding ccRCC versus normal kidney using the matched Epic-seq and RNA-seq data that I generated, as well as TCGA. For pRCC and chRCC, I validated my findings using TCGA data (no TCGA data were available for oncocytoma).

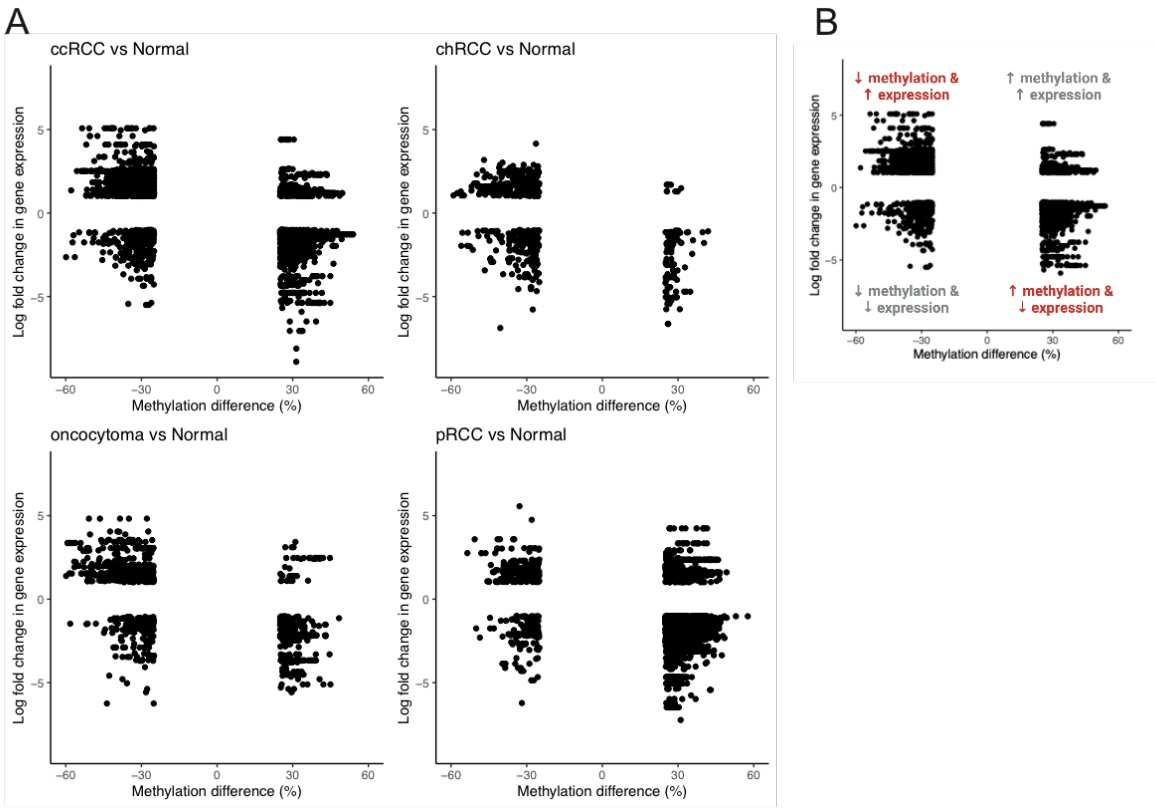


Figure 5.6: Overlap of methylation and gene expression data

Panel A -Each pathological subtype (ccRCC, pRCC, chRCC and oncocytoma) was compared to normal kidney. The x-axis represents differential methylation for DMCs located within the promoter region and the y-axis represents differential gene expression (log2 fold change). Panel B- Schematic showing a scatterplot of differential methylation versus gene expression (log2 fold change), as in Panel A, with annotations. Genes which demonstrate a negative correlation between promoter methylation and expression (red label) are more likely to be epigenetically regulated.

My aim was to explore which DMCs may modulate gene expression, to highlight pathways which may be epigenetically regulated for each tumour type. DMCs were annotated to the nearest gene, and those located within 1.5kb of the transcription start site (TSS) were selected and were overlapped with gene expression data from section 5.3.1.2. My analysis focused on a core set of DMCs at the TSS of genes which are hypermethylated and down-regulated or are hypomethylated and up-regulated (i.e. negative correlation between promoter methylation and expression). These selection criteria were used to identify genes that may be directly regulated by promoter methylation, aiming to enrich for epigenetic changes that are functionally relevant to disease biology. For each pathological subtype, the proportion of overall DMCs located in the promoter in which there was hypomethylation/increased gene expression or hypermethylation/reduced gene expression was calculated. This was: 73% for ccRCC vs normal, 80% for pRCC vs normal, 64% for chRCC vs normal and 75% for oncocytoma vs normal (Table 5.3, Figure 5.6). The highest proportion

was noted in pRCC and lowest was noted in chRCC. This is consistent with results in section 5.3.1.1 which identified a high number of hypermethylated DMCs located in gene promoters in pRCC compared to normal kidney. Subsequently, I performed gene ontology (GO) and gene set enrichment analysis (GSEA) for the core set of epigenetically regulated genes (i.e. those with promoter hypomethylation/increased gene expression and promoter hypermethylation/reduced gene expression).

Table 5.3: Significant DMCs were overlapped with significant DEGs

The proportion of DMCs located in the promoter region which have a negative correlation between methylation and gene expression are shown (i.e. hypomethylation with increased gene expression and hypermethylation with reduced gene expression). There are more DMCs than DEGs, since there may be methylation changes in several consecutive DMCs in the promoter region of a gene.

	ccRCC vs normal kidney	pRCC vs normal kidney	chRCC vs normal kidney	Oncocytoma vs normal kidney
Total overlap between DMCs and DEGs	2765 DMCs overlap with 578 unique DEGs	2904 DMCs overlap with 415 unique DEGs	527 DMCs overlap with 220 unique DEGs	974 DMCs overlap with 283 unique DEGs
Promoter hypomethylation and increased expression	1155 DMCs at 242 unique DEGs	185 DMCs at 54 unique DEGs	242 DMCs at 75 unique DEGs	420 DMCs at 109 unique DEGs
Promoter hypermethylation and reduced expression	876 DMCs at 196 unique DEGs	2125 DMCs at 277 unique DEGs	94 DMCs at 43 unique DEGs	309 DMCs at 98 unique DEGs
Percentage of all DMCs in the promoter having a negative correlation between methylation and gene expression	73% (N=2031/2765)	80% (N=2310/2904)	64% (N=336/527)	75% (N=729/974)

In summary, GO and GSEA suggested that in all four pathological subtypes, methylation may regulate genes involved in kidney development, cell differentiation, cell adhesion and membrane transport (Figure 5.7 and Figure 5.8). The most notable finding was that genes involved in embryological kidney formation were identified in all four subtypes, which suggests that methylation of key developmental genes may regulate cell differentiation and tumorigenesis and therefore explain characteristic patterns associated with the tumours' cell of origin. Additionally, all four pathological subtypes were associated with genes involved in excretion and transport of molecules across the membrane. These are key kidney functions and transporters are expressed on the epithelial membrane in a cell-type specific manner, therefore once again these may be related to cell differentiation. All four subtypes were also involved in pathways associated with cell adhesion,

migration and proliferation, as well as cell-signalling associated with tumorigenesis (such as Notch and tyrosine kinase pathways). As expected, ccRCC was characterised by a high number of immune pathways, and was the only subtype which identified T-cell activation among the epigenetically regulated differentially expressed genes. Indeed, ccRCC is considered 'immune-hot,' reflected in the sensitivity of ccRCC to immune-therapy [228]. Thus, my data suggests that immune activation and immune cell signalling in ccRCC could in part be mediated by changes in DNA methylation. Epigenetic target genes in both ccRCC and pRCC were enriched for pathways associated with response to hypoxia and angiogenesis, which have previously been highlighted in the literature as playing a key role in tumorigenesis in renal cancer [133, 236, 250, 251]. Future work is required to elucidate the functional relevance of the shared and unique epigenome-associated pathways, through deep molecular profiling and perturbation studies. Taken together, these results suggest that there is epigenetic reprogramming of cell differentiation and cell signalling cascades involved in tumorigenesis in all four renal tumour subtypes, which is likely to reflect convergent epigenetic patterning relating to renal epithelial cell origins.

Subsequently, differential gene expression and differential methylation (along with the number of significant DMCs within the promoter) were compared for each pathological subtype in turn, to map subtype-specific alterations (Figure 5.5). Genes which contain a larger number of significant DMCs within their promoter region may be more likely to be regulated by methylation. Therefore, genes were ranked by the number of significant DMCs, as well as based on greatest differential methylation and gene expression (Table 5.4, Table 5.6, Table 5.8, Table 5.9). In the following sections, genes of interest were selected to be discussed more in detail, as illustrative examples, where interesting differences were noted between pathological subtypes. For the malignant subtypes (i.e. excluding benign oncocytoma), results for candidate epigenetically regulated genes were validated in TCGA datasets.

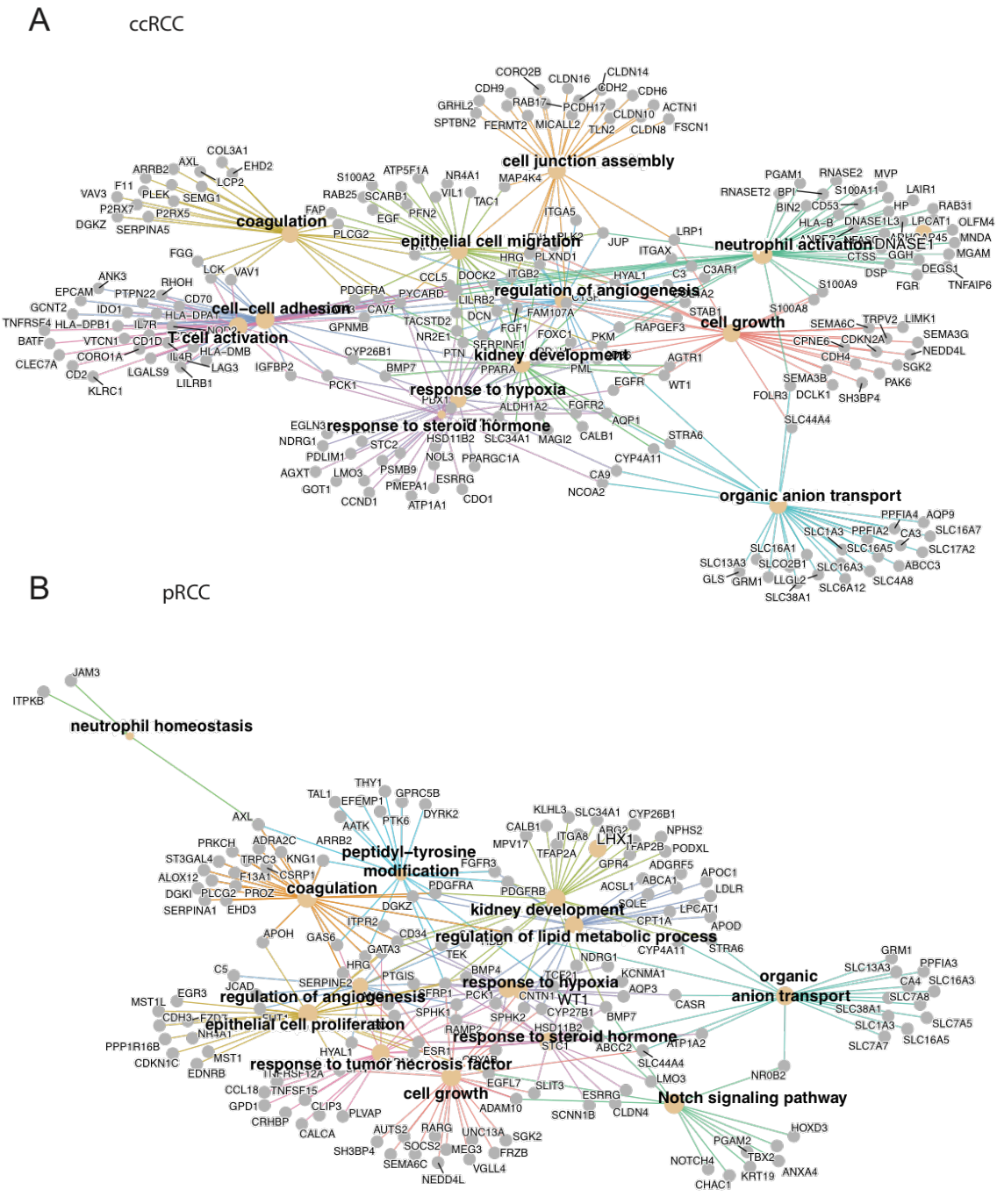


Figure 5.7: Gene ontology analysis for ccRCC and pRCC

Gene ontology for ccRCC (Panel A) and pRCC (Panel B) for genes which were noted to have a negative association between promoter methylation and gene expression.

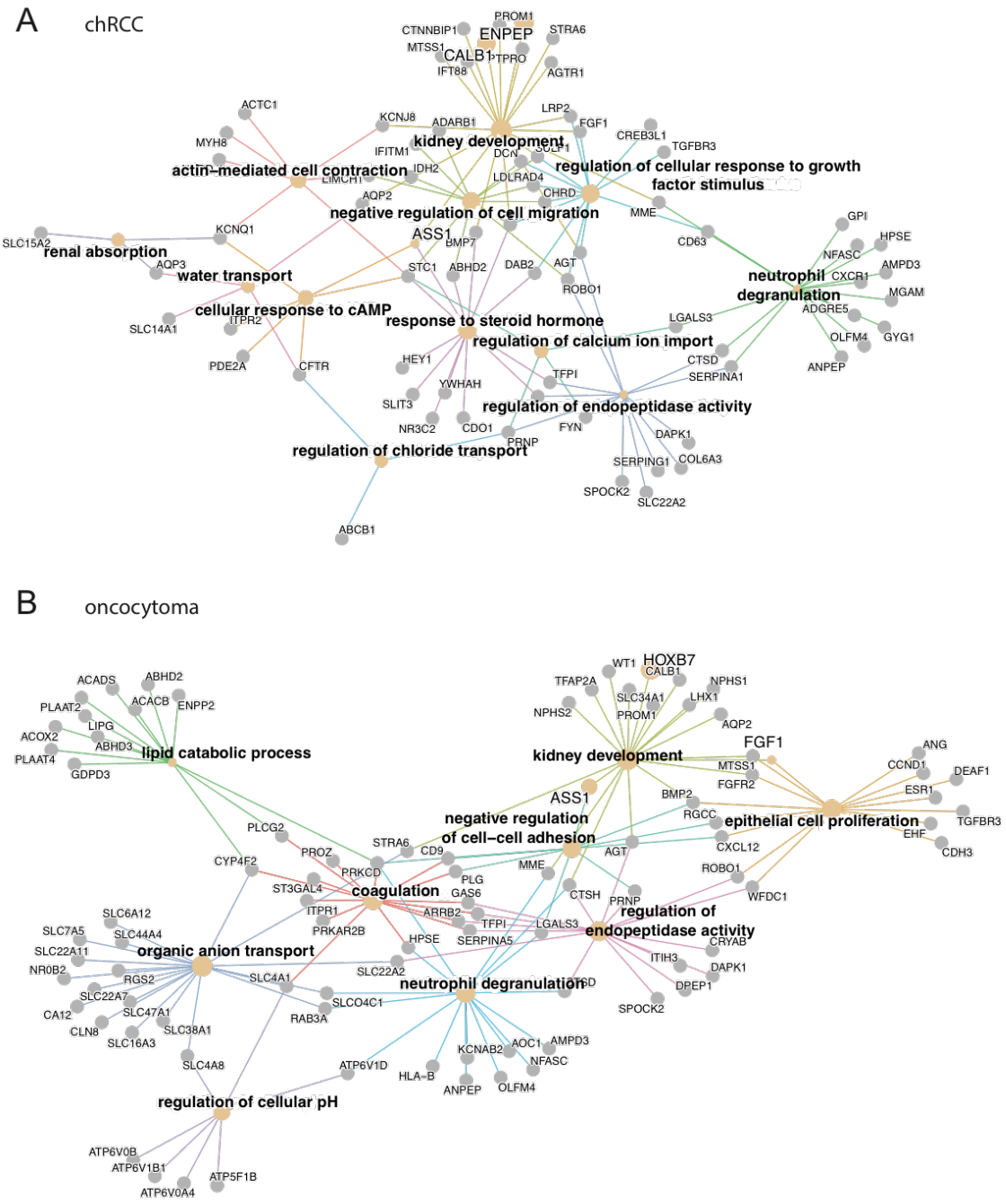


Figure 5.8: Gene ontology analysis for chRCC and oncocytoma

Gene ontology analysis for chRCC (Panel A) and oncocytoma (Panel B) for genes which were noted to have a negative association between promoter methylation and gene expression.

5.3.1.3.1 ccRCC vs normal kidney

In order to identify epigenetically regulated genes in ccRCC, I evaluated the top genes with the greatest differential methylation and differential expression in ccRCC compared to normal kidney (Table 5.4). I sought to validate these genes (from Table 5.4) in my cohort of 47 samples (ccRCC vs normal kidney), in which I generated matched DNA methylation and RNA-seq data (Table 5.5). Lastly, I validated my findings further by evaluating methylation and gene expression in ccRCC samples from TCGA (Table 5.5), and by comparing these against other subtypes (pRCC and chRCC).

Table 5.4: Top-ranking genes which may be epigenetically regulated in ccRCC

For ccRCC versus normal kidney, significant DMCs in the promoter region and significant DEGs were overlapped. The table demonstrates the top ten genes which had the largest number of significant DMCs within their promoter. In addition, the table summarises the top ten genes with the greatest differential methylation and differential gene expression (log2 fold change; FC).

Genes with the largest number of significant DMCs in their promoter region								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
GATA3	34	30%	-4.8		*SLC16A3*	68	-39%	2.5
BCAM	31	45%	-1.3		*ELMO1*	21	-29%	1.4
SLC5A2	30	30%	-2.7		*ADCY7*	20	-29%	1.2
NAV2	26	36%	-1.6		*NDRG1*	20	-31%	1.2
PBX1	20	30%	-2.4		*SHMT2*	20	-42%	1.8
SCNN1A	20	38%	-5.4		*ARHGAP45*	19	-27%	2.6
WT1	20	28%	-2.8		*BIN2*	19	-31%	1.8
ARL4D	17	34%	-1.5		*NDUFA4L2*	18	-38%	5.1
CRYBG2	15	29%	-1.1		*LILRB1*	17	-30%	1.7
MPPED2	13	29%	-2.5		*RIN3*	17	-29 %	1.1
Genes with largest methylation difference and largest differential gene expression								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
SCNN1A	20	38%	-5.4		*FABP6*	5	-44%	4.6
CALB1	3	35%	-6.5		*TGFBI*	6	-39%	3.4
CLCNKB	1	34%	-5.9		*NDUFA4L2*	18	-38%	5.1
RAB25	11	37%	-3.8		*NNMT*	15	-36%	4.1
KCNJ1	4	32%	-7.1		*TACC3*	1	-37%	3.0
CLDN8	2	33%	-5.3		*SLC16A3*	68	-39%	2.5
EPCAM	2	37%	-3.0		*DOCK2*	2	-37%	2.6
ATP6V0A4	6	33%	-4.2		*BIRC3*	2	-44%	2.1
SLC12A1	1	31%	-8.9		*CA9*	5	-36%	3.1
ALDOB	1	31%	-8.1		*AQP9*	6	-36%	2.6

Table 5.5: Validation of putative epigenetically regulated genes in ccRCC

The association between promoter methylation and gene expression in ccRCC and normal kidney is shown in my data and in TCGA. My data consists of a cohort of 47 ccRCC and normal kidney samples with matched Epic-seq and RNA-seq.

	My data		TCGA		
Gene	Correlation	P value	CpG probe ID	Correlation	P value
CA9	-0.90	5.43e-18	cg20610181	-0.77	1.30e-65
NDUFA4L2	-0.91	1.02e-18	cg08163918	-0.80	7.91e-75
SCNN1A	-0.89	5.53e-17	cg16048383	-0.77	4.90e-65
CALB1	-0.81	4.77e-12	cg18335796	-0.45	6.68e-18
CLCNKB	-0.85	1.27e-13	cg21660130	-0.63	3.47e-37
KCNJ1	-0.42	0.004	cg13267718	-0.63	2.34e-37
CLDN8	-0.81	1.14e-11		No data	No data
ATP6V0A4	-0.83	4.76e-13	cg02811197	-0.73	2.28e-54
SLC12A1	-0.76	6.09e-10	cg23705224	-0.42	1.48e-15
EPCAM	-0.66	4.46e-07	cg16076328	-0.41	1.65e-14
RAB25	-0.96	7.17e-27	cg19580810	-0.76	1.73e-62
AQP9	-0.48	0.0006	cg11098259	-0.35	6.54e-11
SLC16A3	-0.94	3.00e-22	cg19284277	-0.71	2.54e-50
NNMT	-0.91	4.56e-19	cg14520913	-0.62	1.61e-35
TGFBI	-0.68	1.84e-07	cg21583694	-0.81	4.40e-77

Interestingly, the top-most hypermethylated and repressed genes in ccRCC vs normal kidney were markers of the distal nephron (*SCNN1A, CALB1, CLCNKB, KCNJ1, CLDN8, ATP6V0A4, SLC12A1*) (Table 5.4). This was confirmed in both my validation dataset and TCGA (Table 5.5). Genes which represent nephron markers had a very strong, significant inverse correlation between promoter methylation and gene expression, with correlation values < -0.75 (except *KCNJ1* and *AQP9* which had a moderate inverse correlation) (Table 5.5). Two markers of the distal nephron (*CLCNKB* and *SCNN1A*) were visualised as illustrative examples. There was evidence of promoter hypermethylation and reduced expression in ccRCC vs normal kidney in my dataset, which was confirmed in TCGA (Figure 5.9A-B). Comparing different subtypes of renal cancers in TCGA, it was evident that these genes demonstrated relative hypomethylation and increased expression in chRCC, with hypermethylation and reduced expression in ccRCC and pRCC (Figure 5.9C-D). Once again, this suggests methylation may play a role in regulating lineage specific markers in ccRCC, in keeping with the results of section 5.3.1.3.

In addition, my analysis suggests that genes involved in cell adhesion, proliferation, invasion and tumorigenesis (such as *EPCAM, RAB25, NNMT* and *TGFBI*) may be epigenetically regulated in ccRCC and that there may be subtype-specific patterns. For example, in TCGA, ccRCC was characterised by hypermethylation and reduced expression of *EPCAM*, whereas chRCC was characterised by hypomethylation and overexpression (Figure 5.9E). Conversely, *TGFBI* was a clear example in which ccRCC revealed hypomethylation and increased expression, whereas chRCC showed hypermethylation and reduced expression (Figure 5.9F). As expected, in the majority of cases, pRCC resembled ccRCC. In summary, these consistent findings across different datasets confirm epigenetically regulated genes in ccRCC. Furthermore, these genes demonstrated distinct methylation and expression patterns in the different malignant subtypes of RCC, meaning they could potentially be used as subtype-specific markers.

Subsequently, *CA9* and *NDUFA4L2* were explored in detail as putative ccRCC-specific biomarkers which may be epigenetically regulated. For both genes, I found evidence of significant promoter hypomethylation in ccRCC compared to normal tissue (Figure 5.10A,C). There was a strong negative correlation between methylation and gene expression, both in my dataset and in TCGA, for both genes (Figure 5.10B,D; Table 5.5). Subsequently, I evaluated methylation and expression of these two genes in the other pathological subtypes of renal tumours. In my data (Epic-seq and Affymetrix), there was no significant difference in *CA9* nor *NDUF4AL2* promoter methylation nor gene expression for chRCC and oncocytoma relative to normal tissue. pRCC was characterised by significant *CA9* and *NDUFA4L2* promoter and gene body hypermethylation relative to normal kidney, however there was no change in gene expression. Taken together, my results suggested that *CA9* and *NDUF4AL2* may be epigenetically regulated in ccRCC but not in the other pathological subtypes of renal tumours. The existing literature suggests that over-expression of *CA9* in ccRCC is likely to be largely driven by *VHL* inactivation (by increasing the HIF1-α transcription factor), however, several studies demonstrate that methylation may play a role in facilitating this hypoxia signalling axis [252-254]. For example, cell line treatment with demethylating agent 5-Aza-2'-deoxycytidine (5-Aza-dc) induces *CA9* mRNA and protein expression [255-257]. My study provides further evidence to support the notion that methylation may regulate transcription of these genes in ccRCC. In summary, the methylation differences noted in the four pathological subtypes at the *CA9* and *NDUF4AL2* gene loci suggests that these could be valid markers to differentiate ccRCC, in addition to being candidate genes functionally regulated by DNA methylation in ccRCC.

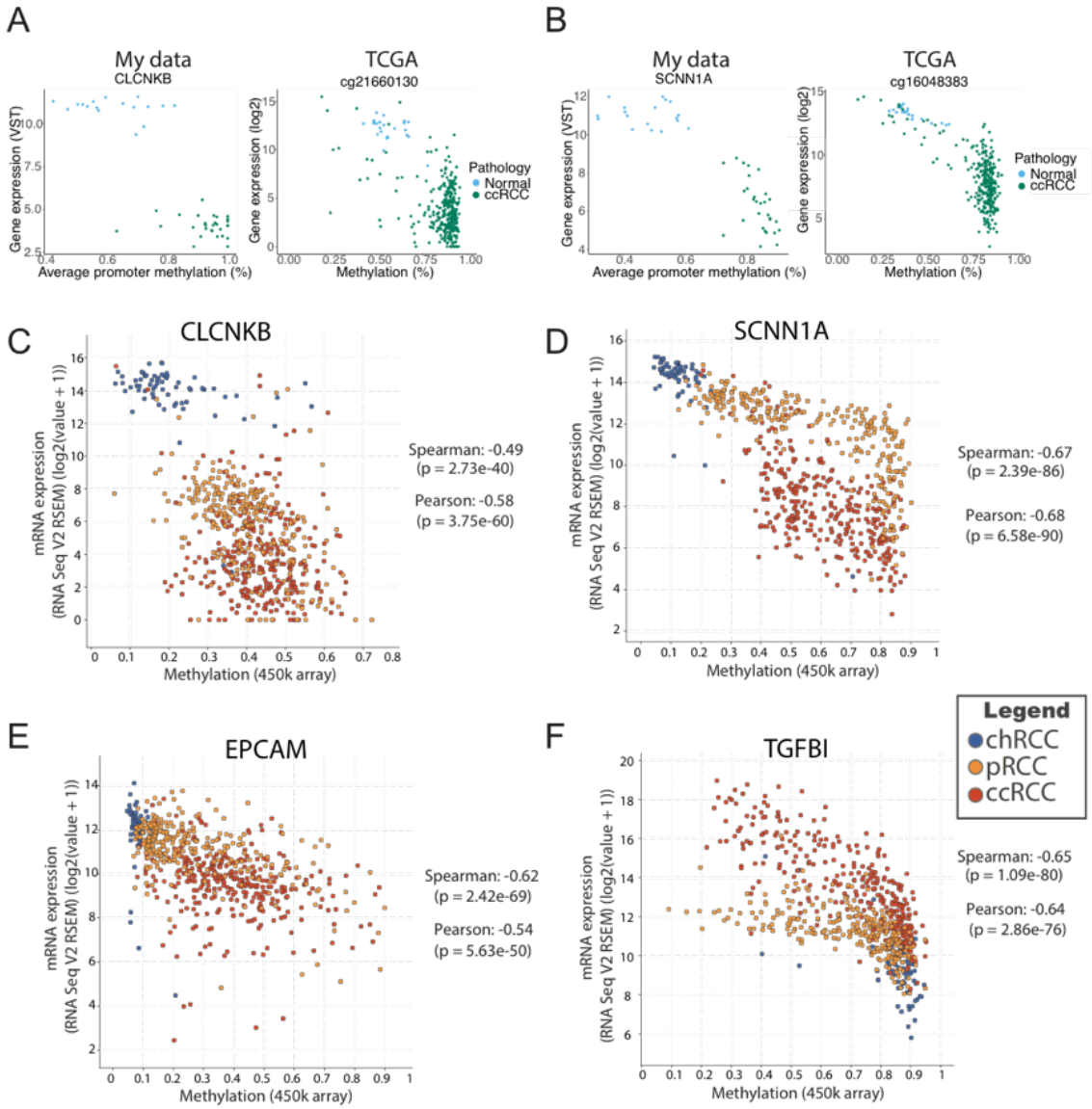


Figure 5.9: Methylation versus gene expression for selected genes, in ccRCC tissue samples and other subtypes.

Panel A and B- Promoter methylation versus gene expression for *CLCNKB* (Panel A) and *SCNN1A* (Panel B), in ccRCC and normal kidney tissue samples in my data (left) and TCGA (right). There is a strong negative correlation in both datasets. Panel C-F- Methylation versus gene expression is visualised for ccRCC, pRCC and chRCC samples derived from TCGA for *CLCNKB* (Panel C), *SCNN1A* (Panel D), *EPCAM* (Panel E) and *TGFBI* (Panel F). There is a strong negative correlation between methylation and genes expression observed in these genes, with clear subtype specific differences. The scatterplots in Panel C-F were adapted from cBioPortal [179].

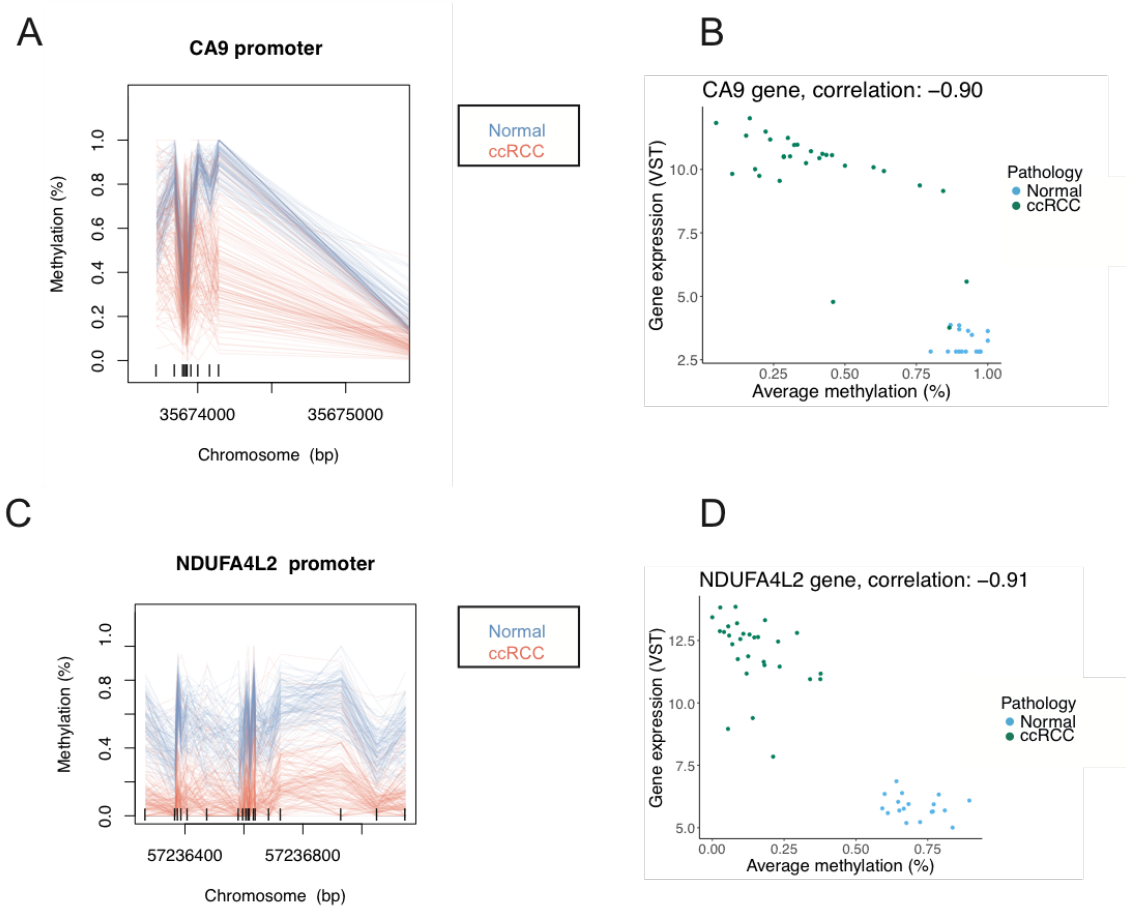


Figure 5.10: Methylation and gene expression of CA9 and NDUFA4L2, two ccRCC specific markers

Panel A- Methylation levels for ccRCC and normal kidney samples at the *CA9* gene promoter region in my data. There is clear promoter hypomethylation in ccRCC compared to normal kidney. Panel B- Methylation versus gene expression is visualised for the *CA9* gene for my data (47 ccRCC and normal kidney samples). There is a strong negative correlation between methylation and gene expression, with high expression/hypomethylation noted in ccRCC. Panel C- Methylation levels for ccRCC and normal kidney samples at the *NDUFA4L2* gene promoter region in my data. There is clear promoter hypomethylation in ccRCC compared to normal kidney. Panel D- Methylation versus gene expression is visualised for the *NDUFA4L2* gene for my data (47 ccRCC and normal kidney samples). Similarly to *CA9*, there is a strong negative correlation between methylation and gene expression, with high expression/hypomethylation noted in ccRCC.

5.3.1.3.2 pRCC vs normal kidney

Aiming to identify epigenetically regulated genes, I ranked genes with the greatest differential methylation and differential expression in pRCC compared to normal kidney (Table 5.6) and subsequently validated these candidate markers using TCGA (Table 5.7). Similar to the results of the analysis for ccRCC, genes related to kidney differentiation and membrane function were noted, including: membrane transporters (*SLC5A2*, *SLC44A4*), claudins (*CLDN1*, *CLDN10*), cell lineage markers (*PCK1*, a proximal tubule marker; *ATP6V0A4*, a distal nephron marker) and genes involved in renal embryogenesis (*WT1*, *HOXD10*, *LHX1* and *FZD7*). *CLDN1* has been previously proposed as a marker to differentiate pRCC from ccRCC, and expression is associated with improved survival in pRCC [249]. Additionally, genes involved in immune activation were also identified (including *IFI27*, *CCL18*, *IL32*, *C7*). Selected functional candidate genes are discussed in depth in the section below, to highlight potential targets of epigenetic regulation in pRCC and differences between pRCC and the other subtypes.

The *HOXD10* gene is part of the homeobox D family of transcription factor genes, which plays a role in EMT, proliferation and renal metanephric development, with strong expression noted in the normal adult urogenital tract [258, 259]. In my data, there was evidence of reduced expression in pRCC relative to normal kidney, with a strong negative correlation between methylation and gene expression (Table 5.6). In pRCC, promoter hypermethylation was noted spanning a relatively large region (57 significant DMCs), although a handful of outlier samples were noted (Figure 5.11A). These findings were confirmed in TCGA (Figure 5.11B and C, Table 5.7). TCGA data allowed a comparison between the three malignant pathological subtypes, highlighting increased expression of *HOXD10* (and relative promoter hypomethylation) in ccRCC and chRCC relative to pRCC (Figure 5.11D). Interestingly, ccRCC demonstrated different expression patterns compared to pRCC, despite the shared cell of origin. This is consistent with previous reports that show differential *HOXD10* gene expression in a tissue specific manner, with either oncogenic or tumour suppressor functions depending on the context [260, 261]. Although epigenetic regulation of *HOXD10* has been demonstrated in gastric cancer [262], to my knowledge, this is the first account to highlight differences between renal tumour subtypes and to link these with changes in promoter methylation.

A further notable example of putative epigenetic dysregulation of developmental genes in pRCC was *WT1*. *WT1* plays a key role in initiating kidney development by mediating EMT and outgrowth of the ureteric bud from the mesonephros [263] and is a known tumour suppressor gene (implicated in Wilms' tumour for example) [264]. My data showed promoter hypermethylation and associated

reduced expression in both pRCC and ccRCC relative to normal tissue, but no significant difference in the promoter of chRCC (Table 5.4 and Table 5.6), changes which could be harnessed as potential diagnostic markers.

Table 5.6: Top-ranking genes which may be epigenetically regulated in pRCC

For pRCC versus normal kidney, significant DMCs in the promoter region and significant DEGs were overlapped. The table demonstrates the top ten genes which had the largest number of significant DMCs within their promoter. In addition, the table summarises the top ten genes with the greatest differential methylation and differential gene expression (log2 fold change; FC).

Genes with the largest number of significant DMCs in their promoter region								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
LHX1	93	29%	-2.5		*HRH1*	18	-37%	1.6
HOXD10	57	31%	-2.8		*SPON2*	16	-36%	1.0
CHRD	51	32%	-1.3		*ALDH3B1*	8	-32%	1.4
WT1	51	29%	-4.7		*CAPG*	8	-35%	2.0
FZD7	47	35%	-1.1		*SLC44A4*	7	-30%	1.9
ESRRG	46	31%	-2.8		*IL32*	6	-32%	2.3
SLC5A2	39	33%	-2.4		*RHBDF2*	6	-32%	1.3
DGKI	34	32%	-1.2		*SQOR*	6	-40%	1.1
SNED1	34	32%	-1.5		*AHNAK2*	5	-41%	3.6
CLDN10	33	30%	-1.9		*ANG*	5	-34%	1.0
Genes with largest methylation difference and largest differential gene expression								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
HSD11B2	2	43%	-5.4		*AHNAK2*	5	-41%	3.6
ATP6V0A4	4	41%	-3.4		*CLU*	5	-48%	2.8
HPD	1	35%	-5.8		*IFI27*	5	-37%	3.0
ADGRF5	2	40%	-3.1		*CREB5*	2	-38%	2.2
C7	2	36%	-3.8		*CCL18*	1	-33%	5.6
TBX2	15	37%	-3.1		*CLDN1*	2	-38%	1.9
TEK	1	35%	-3.4		*CAPG*	8	-35%	2.0
MPPED2	14	36%	-3.1		*IL32*	6	-32%	2.3
PCK1	1	32%	-4.0		*EXOC7*	1	-35%	1.8
NES	11	40%	-2.3		*SEL1L3*	1	-33%	1.9

Table 5.7: Validation of putative epigenetically regulated genes in pRCC

The table shows the association between methylation and gene expression in pRCC and normal kidney in TCGA.

	TCGA		
Gene	CpG probe ID	Correlation	P value
WT1	cg20989480	-0.32	3.97e-06
HOXD10	cg03850256	-0.69	2.47e-30
LHX1	cg20950167	-0.53	1.95e-16
FZD7	cg04913005	0.26	1.25e-04
CLDN1	cg03601836	-0.56	1.36e-18
CLDN10	cg18470456	-0.34	5.77e-07
PCK1	cg03840472	-0.33	1.83e-06
AHNAK2	cg06799735	-0.79	3.10e-45
ESRRG	cg13242895	0.45	9.35e-12
TBX2	cg27005487	-0.75	4.12e-38
ESRRG	cg01432520	-0.43	1.30e-10

Evaluating up-regulated genes in which promoter hypomethylation was noted in pRCC, the *AHNAK2* gene was the top-most target in my dataset (Table 5.6). This finding was externally validated in TCGA pRCC data (Table 5.7, Figure 5.11E). This gene has been independently identified as having the highest correlation between gene expression and methylation in pRCC in the literature [265], confirming the validity of my analysis. In addition, I found promoter hypomethylation and reduced gene expression in ccRCC versus normal tissue in my matched RNA-seq and Epic-seq dataset, suggesting a similar epigenetic regulation is present in ccRCC (Figure 5.11F). TCGA data confirmed that whilst ccRCC and pRCC had similar hypomethylation and increased expression, in chRCC there appears to be a subset of cases with hypermethylation and reduced expression (Figure 5.11G) similar to the pattern observed in normal kidney tissue (Figure 5.11E). The *AHNAK2* gene codes for desmoyokin, which regulates *TGFβ/SMAD* signalling, and therefore plays a role in cell cycle progression, cell growth and migration [266]. Increased expression of *AHNAK2* (as is the case for ccRCC and pRCC) is associated with increased proliferation, EMT and tumorigenesis. In ccRCC, hypoxia upregulates *AHNAK2* expression via *HIF1-α* [267]. To my knowledge, this is the first study to compare *AHNAK2* in different pathological subtypes of renal tumours. It remains unclear why such a different pattern is seen in chRCC, compared to ccRCC and pRCC, warranting further investigation in future work with deeper profiling and spatially resolved molecular analysis.

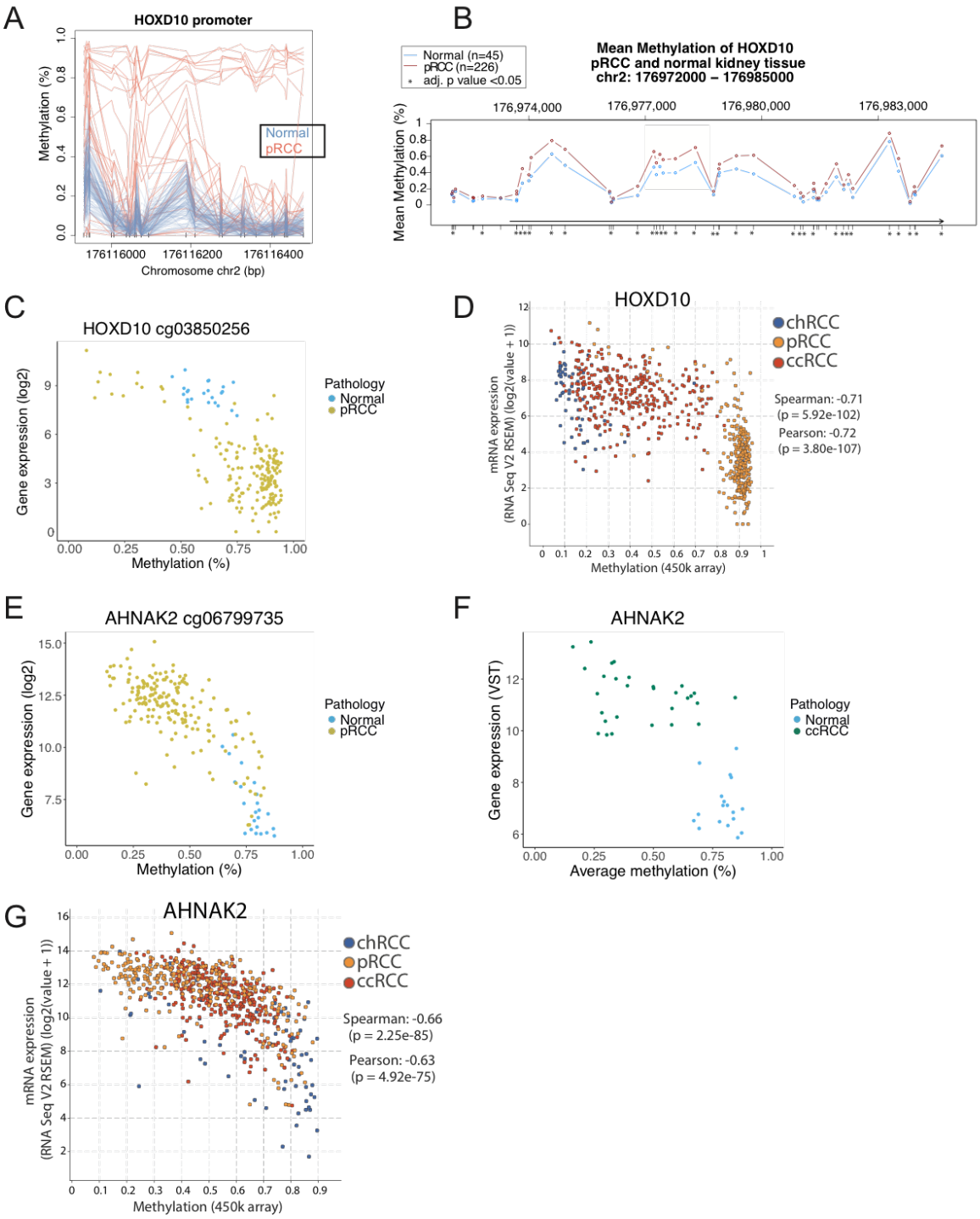


Figure 5.11: Methylation and gene expression for selected genes, in pRCC tissue samples and other subtypes

Panel A and B- Methylation in pRCC and normal kidney at the *HOXD10* promoter region, in my dataset (Panel A; hg38) and TCGA (Panel B; hg19). Panel C and D- Methylation vs gene expression for *HOXD10* for pRCC and normal kidney (Panel C) and for pRCC, ccRCC and chRCC (Panel D) derived from TCGA. Panel E and G: Methylation vs gene expression for *AHNAK2* for pRCC and normal kidney (Panel E) and the three pathological subtypes (Panel G) in TCGA. Panel F- Methylation vs gene expression for *AHNAK2* for my samples (I generated matched RNA-seq and Epic-seq for 47 ccRCC and normal kidney tissue samples). Panel B was obtained from TCGA Wanderer [178], and Panels D and G were adapted from cBioPortal [179].

5.3.1.3.3 chRCC vs normal kidney

Differential methylation and differential gene expression were explored in chRCC compared to normal kidney to identify epigenetically regulated genes in my dataset. Subsequently, I selected three of these genes to be explored in TCGA as potential candidate diagnostic markers. Table 5.8 summarises the top-ranking genes which had a negative correlation between gene expression and promoter methylation. In keeping with my previous findings (section 5.3.1.3), putative epigenetically regulated genes were enriched for kidney developmental markers and membrane transporters. For example, chRCC was characterised by hypomethylation and increased expression of genes associated with the collecting duct (*RHCG* and *ATP6V0A4*) and DCT (*PVALB*), both of which have been postulated as a potential cell of origin for chRCC [268]. My analysis identified three genes in which there were clear expression differences between chRCC, normal kidney and other pathological subtypes of renal tumours, and these are described in further detail below.

Comparing chRCC against normal tissue, my data showed clear promoter hypomethylation and associated increased expression for *LGALS3*, *NEDD4L* and *CTSD* genes. The same pattern was also noted in oncocytoma, for all three genes (Table 5.9). In addition, TCGA data confirmed there was increased expression with associated promoter hypomethylation in chRCC compared to ccRCC and pRCC for all three genes (Figure 5.12A-C). Therefore, there could be potential utility in using these markers to distinguish renal tumours derived from the PCT and DCT.

Both *CTSD* and *NEDD4L* regulate cell growth, as well as having specific functions within the renal tubule epithelium. The *CTSD* gene codes for cathepsin-D, a lysosomal protease, involved in regulating apoptosis and extracellular matrix degradation, as well as cell differentiation and growth via *PI3K-mTOR* signalling [269, 270]. The literature suggests that in the normal adult kidney, cathepsin-D expression is localised to the DCT and collecting ducts, which may explain the subtype specific differences in renal tumours [269]. *NEDD4L* is a E3 ubiquitin ligase enzyme which acts on target proteins and facilitates proteasome degradation. In the collecting duct, *NEDD4L* regulates water and sodium homeostasis by ubiquitination of AQP2 and the epithelial sodium channel (ENaC) [271]. In addition, it regulates several signalling pathways (including *TGF-β*, *Wnt*, *PI3K-AKT-MTOR* and *EGFR* signalling) and has been linked with tumorigenesis in several cancers [272]. My results suggests that these genes may be epigenetically regulated in renal cancer.

Table 5.8: Top-ranking genes which may be epigenetically regulated in chRCC.

For chRCC versus normal kidney, significant DMCs in the promoter region and significant DEGs were overlapped. The table demonstrates the top ten genes which had the largest number of significant DMCs within their promoter. In addition, the table summarises the top ten genes with the greatest differential methylation and differential gene expression (log2 fold change; FC).

Genes with the largest number of significant DMCs in their promoter region								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
SLC22A2	7	27%	-3.8		*LIMCH1*	16	-40%	1.3
CXCL14	6	30%	-4.7		*PLCG2*	16	-35%	1.8
ROBO1	6	29%	-1.8		*NEDD4L*	13	-31%	1.5
SLC66A2	5	29%	-1.2		*VAC14*	13	-44%	1.6
UPB1	5	28%	-5.0		*CA12*	10	-34%	1.1
GATM	4	27%	-3.1		*AMPD3*	8	-33%	1.6
RCAN1	4	27%	-1.4		*LGALS3*	8	-30%	2.4
SLC22A11	4	27%	-4.4		*SLC49A3*	8	-32%	2.5
CARHSP1	3	33%	-1.1		*ACSS3*	7	-47%	1.1
GPER1	3	26%	-2.2		*ATP6V0A4*	7	-29%	2.3
Genes with largest methylation difference and largest differential gene expression								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
CXCL14	6	30%	-4.7		*RHCG*	1	-47%	3.2
CALB1	2	28%	-5.8		*PVALB*	1	-40%	3
UPB1	5	28%	-5		*GDE1*	2	-48%	2.3
NAT8B	1	33%	-3.3		*BSG*	2	-39%	2.7
ADARB1	1	36%	-2.4		*BACE2*	1	-35%	2.8
SERPINA1	2	29%	-3.4		*CTSD*	7	-36%	2.5
DPYS	1	27%	-5.2		*ABHD3*	3	-40%	2.1
CDH2	2	27%	-4		*GPSM2*	2	-37%	2.3
GBA3	1	28%	-3.2		*PARVB*	3	-33%	2.9
NAT8	2	26%	-6.6		*CLPB*	2	-35%	2.2

LGALS3 codes for galectin-3, which regulates cell growth, adhesion, signalling and differentiation, and thus plays a role in ureteric bud formation during renal embryological development. In the normal adult kidney, galectin-3 is expressed in the distal tubules and intercalated cells of the collecting duct, from which chRCC is believed to be derived [273]. This could explain why increased levels of galectin-3 are seen in chRCC relative to renal tumours which are derived from the PCT (i.e. this is an epigenetically regulated lineage marker for DCT epithelial cells). An immunohistochemical study of renal tumours has confirmed strong galectin-3 expression in over 90% of oncocytomas and chRCC, but only 34% of ccRCCs and 13% of pRCCs; meaning this could have potential as a diagnostic marker [274]. Additionally, studies have shown that lower galectin-3 may have prognostic potential in renal cancer, by mediating decreased cell adhesion and increased invasion. Lower levels were associated with higher tumour grade, stage and worse overall survival [274, 275]. Genes which have both a diagnostic and prognostic utility may be useful markers in clinical practice as they would enable not only an accurate differentiation of SRMs, but also provide data regarding prognosis which could guide management options (e.g. active surveillance versus treatment).

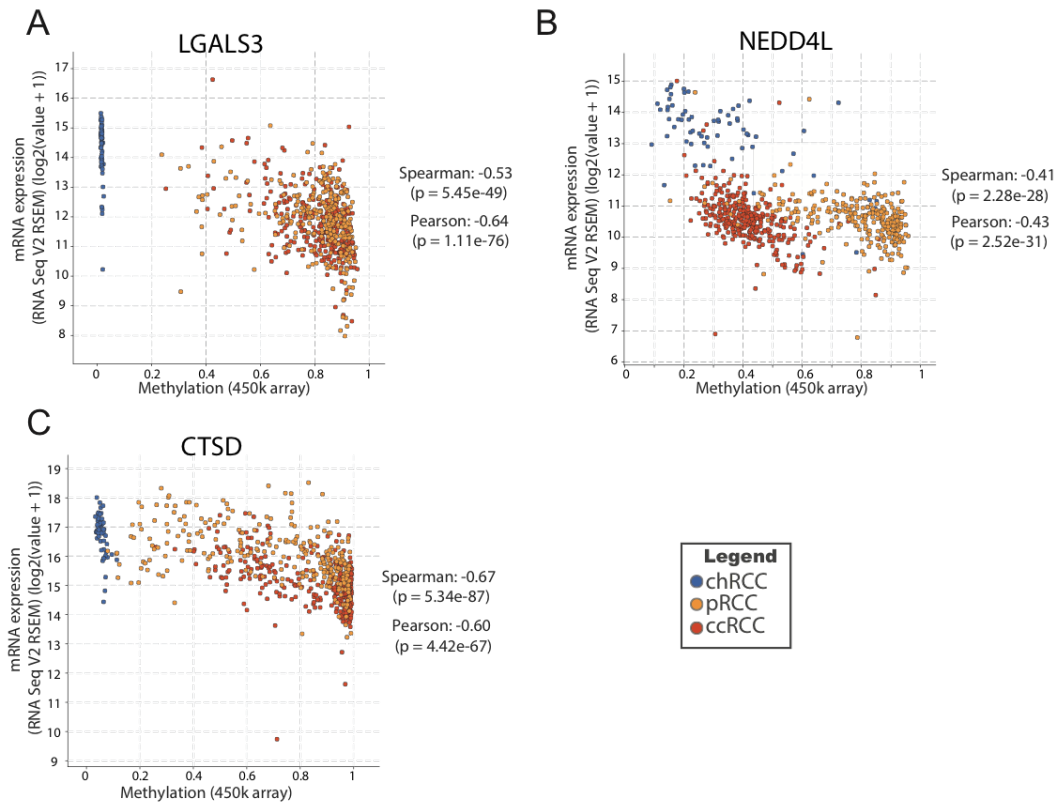


Figure 5.12: Methylation versus gene expression for selected genes, in chRCC, pRCC and ccRCC

The following genes are shown: *LGALS2* (Panel A), *NEDD4L* (Panel B) and *CTSD* (Panel C). All three figures are adapted from cBioPortal [179].

5.3.1.3.4 Oncocytoma vs normal kidney

Methylation and gene expression were evaluated in oncocytoma versus normal tissue, and some obvious similarities were noted with genes highlighted in the analysis of chRCC (Table 5.8 and Table 5.9). In both oncocytoma and chRCC, a number of genes demonstrated promoter hypomethylation and upregulated expression, including genes associated with the collecting duct (*ATP6V0A4* and *ATP6V1B1*), DCT (*KCNQ1*), solute carriers (*SLC38A1*), and transmembrane proteins (*TMEM101*), as well as genes involved in lipid metabolism (*ABHD2*, *ABHD3*), mitosis (*KNTC1*) and apoptosis (*CLPB*). In addition, both oncocytoma and chRCC demonstrated hypermethylation and reduced expression in PCT markers (*GATM*, *ASS1*), DCT markers (*CALB1*), and solute carriers (*SLC22A2*, *SLC66A2*, *SLC22A11*) relative to normal kidney. Unfortunately, it was not possible to validate these findings. Future work should focus on highlighting differences between chRCC and oncocytoma, which could be used to aid diagnosis.

Whilst I performed a detailed analysis of each subtype versus normal kidney, I was unable to evaluate all subtypes simultaneously using routinely available methods for methylation analysis. Indeed, the 'methylKit' package only enables two comparisons at a time and the 'dmrseq' package was unable to run as my large sample size made it too computationally intensive. This prompted me to collaborate with Izzy Newsham (section 5.3.2), leveraging her bio-informatics expertise to enable a simultaneous comparison of all subtypes.

In summary, in section 5.3.1.3, I identified methylation changes which are likely to regulate gene transcription in each of the pathological subtypes. I confirmed these findings by evaluating TCGA data, demonstrating the validity of my results. I highlighted pathways which may be regulated by DNA methylation as this could offer insights into how methylation mediates tumorigenesis and I explored candidate epigenetically regulated genes in greater detail. Given none of the markers demonstrate high specificity for a single subtype, it is likely that using a larger number of methylation changes in combination (rather than individual markers) may be more useful as a diagnostic test. This is therefore explored in the next section (5.3.2), where DNA methylation data were used to create a machine learning model to predict pathological subtypes of SRMs.

Table 5.9: Top-ranking genes which may be epigenetically regulated in oncocytoma

For oncocytoma versus normal kidney, significant DMCs in the promoter region and significant DEGs were overlapped. The table demonstrates the top ten genes which had the largest number of significant DMCs within their promoter. In addition, the table summarises the top ten genes with the greatest differential methylation and differential gene expression (log2 fold change; FC).

Genes with the largest number of significant DMCs in their promoter region								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
EPS8L1	31	32%	-2.2		*PLCG2*	22	-37%	3.1
SLC22A2	27	33%	-3.7		*NEDD4L*	18	-33%	1.9
PDLIM4	25	29%	-2.1		*FOXI1*	16	-55%	3.4
UPB1	21	30%	-4.5		*SLC20A1*	14	-42%	1.1
ANPEP	10	29%	-3.8		*TMEM101*	14	-36%	3.4
EPS8L2	9	27%	-1.5		*MYO10*	13	-37%	1.3
NR2F1	9	31%	-1.5		*SLC16A3*	12	-31%	1.7
BHMT2	7	33%	-3.3		*HOXB7*	11	-27%	1.1
GATM	6	30%	-4.3		*CD9*	10	-30%	1.2
IL32	6	30%	-2.5		*IGFBP1*	10	-38%	3.5
Genes with largest methylation difference and largest differential gene expression								
Promoter hypermethylation & reduced gene expression					Promoter hypomethylation & increased gene expression			
Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression		Gene	Number of DMCs	Average methylation difference	Log2 FC gene expression
HPD	5	37%	-5.1		*FOXI1*	16	-55%	3.4
CXCL14	2	36%	-4.8		*ARRB2*	2	-52%	2.9
SLC22A2	27	33%	-3.7		*GPSM2*	2	-43%	3.9
GALNT14	1	32%	-3.6		*ATP6V1B1*	1	-51%	2.5
BHMT2	7	33%	-3.3		*SNTB1*	1	-56%	2.2
UPB1	21	30%	-4.5		*CELF2*	3	-51%	2.2
GATM	6	30%	-4.3		*LIMS1*	2	-43%	2.9
SLC47A1	1	30%	-5.6		*AQP6*	5	-40%	4.8
PTH1R	1	32%	-2.9		*CLGN*	2	-43%	2.9
NAT8	2	29%	-5.4		*SFTPB*	1	-39%	4.0

5.3.2 MethylBoostER: Machine learning model to predict pathological subtypes of renal tumours and normal kidney tissue

5.3.2.1 Exploring methylation markers of different pathological subtypes from the literature

Differentiating pathological subtypes of renal tumours is a clinical priority [1]. First, I visualised methylation markers from the literature in my dataset, subsequently I collaborated with Izzy Newsham, PhD candidate from the Samarajiwa Research Group, to develop a machine learning model to predict pathological subtypes of renal tumours.

A small number of studies are available in the literature which use DNA methylation to predict pathological subtypes of renal tumours. Dr Brennan et al recently published a model which uses Prediction Analysis of Microarrays (PAM) and data from 79 CpG probes to distinguish normal kidney, oncocytoma and chRCC [121]. The model was developed using 450k array methylation data. Of these 79 CpGs, 45 CpGs were available in my dataset and these are visualised in a heatmap (Figure 5.13A). Based on clustering alone, the majority of samples from the same pathological subtype cluster together, however many samples cluster away from their subtype. Oncocytoma and chRCC samples cluster together closely. It is evident that some CpGs appear to be more informative than others- for example cg11983867 and cg00394316 appear to be hypomethylated in virtually all samples despite being presumed chRCC markers. I contacted Dr Brennan (Gevaert Group, Stanford University, USA) and enquired whether they would be able to externally validate their model on my dataset. The study authors reported that unfortunately this is not possible as the PAM model cannot tolerate missing data and requires imputation of missing CpGs; and this would not be appropriate in my dataset which is missing 43% (45/79) of required probes.

Chopra et al created a model to distinguish renal tumour subtypes based on 57 CpG probes using 450k array methylation data [120]. In the testing set, the model predicted the correct pathological subtype in only 58% of oncocytomas and 64% of pRCC. Results were more favourable when predicted output was classified as malignant versus non-malignant, rather than into the different subtypes (100% of oncocytomas were predicted as non-malignant and 86% of pRCC were predicted as malignant). Of the 59 CpGs, data were available for 41 CpGs in my dataset (Figure 5.13B). Based on hierarchical clustering alone, samples broadly cluster by subtype: normal cluster very closely together, ccRCC cluster more closely with pRCC and oncocytoma and chRCC are the most similar. Given both Brennan and Chopra models are based on <100 CpGs, it is difficult to apply the models

on external datasets which may contain missing data or be processed using a different methylation platform (e.g. Epic-seq vs 450k array) since this results in a large proportion of missing probes. This limitation could be overcome by including a larger number of CpGs (i.e. 1000 rather than 100) and creating a model which tolerates missing data without the requirement for imputation. Subsequent analysis will aim to create such a model to distinguish subtypes of renal tumours using my methylation data.

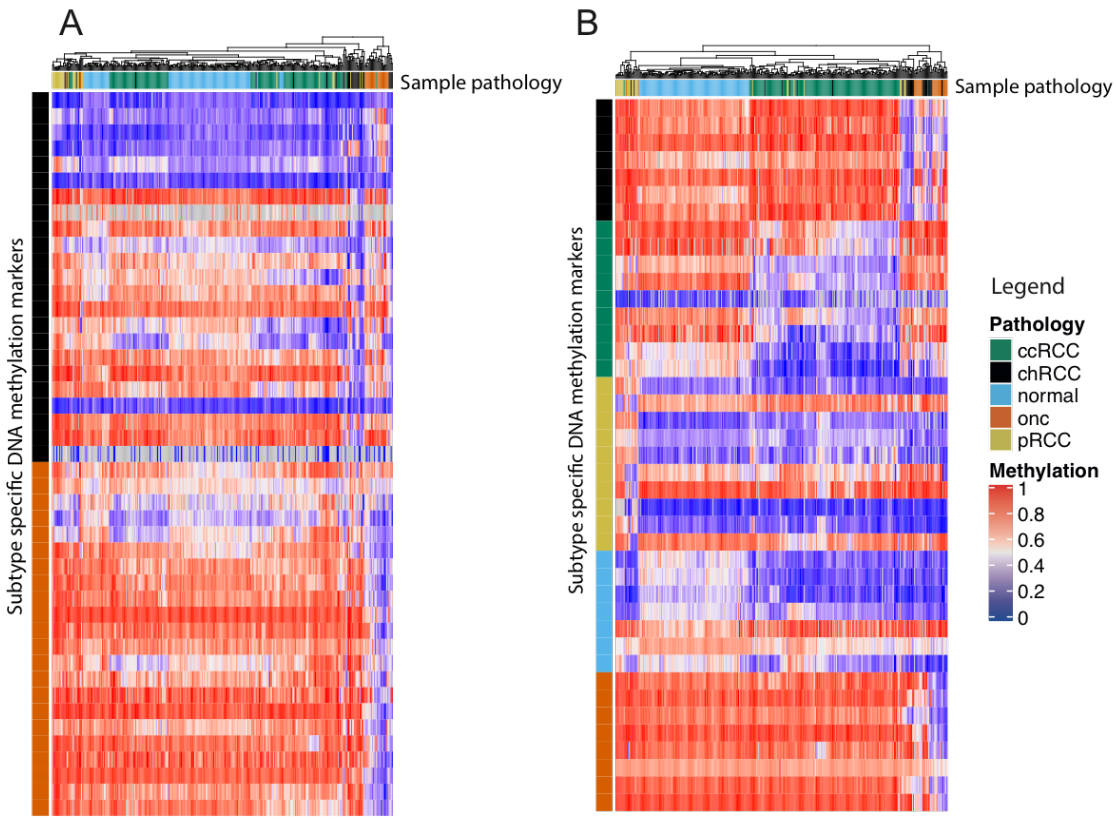


Figure 5.13: CpGs identified in the literature which aim to differentiate pathological subtypes

Heatmap demonstrating methylation levels for my Epic-seq samples (ccRCC, pRCC, chRCC, oncocytoma and normal kidney) at CpGs that differentiate pathological subtypes of renal tumours in previous studies. Tissue samples are shown as columns and CpGs are shown as rows. The coloured horizontal bar demonstrates the pathological diagnosis of each sample. The coloured vertical bar on the left demonstrates subtype-specific CpG markers from the literature. Based on hierarchical clustering alone (Panel A and B), samples broadly cluster by subtype: normal cluster very closely together, ccRCC cluster more closely with pRCC and oncocytoma and chRCC are the most similar. Panel A- CpGs identified by Brennan et al [121] which are meant to distinguish oncocytoma from chRCC. Panel B- CpGs identified by Chopra et al [120] which are meant to distinguish different pathological subtypes.

5.3.2.2 MethylBoostER: a machine learning model using DNA methylation to differentiate pathological subtypes of renal tumours

The following work is the result of a collaboration with Izzy Newsham (Samarajiwa Group; see section 4.5.2). In brief, we developed MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours), a machine learning model which uses DNA methylation to classify tissue samples into normal kidney or common pathological subtypes of malignant and benign renal tumours (ccRCC, pRCC, chRCC and oncocytoma). Figure 5.14 summarises the workflow and samples used. I combined my Epic-seq tissue data along with two publicly available datasets (TCGA and Chopra et al) and this was used as the training and testing set [24, 120]. Subsequently, MethylBoostER was externally validated on four independent datasets [120, 121, 126, 127]. A strength of the model is that we evaluate performance on multi-region samples, demonstrating methylation heterogeneity does not limit model applicability. In addition, I assessed tumour purity and explored the impact on model predictions, as low purity biopsy samples represent a real-world clinical challenge.

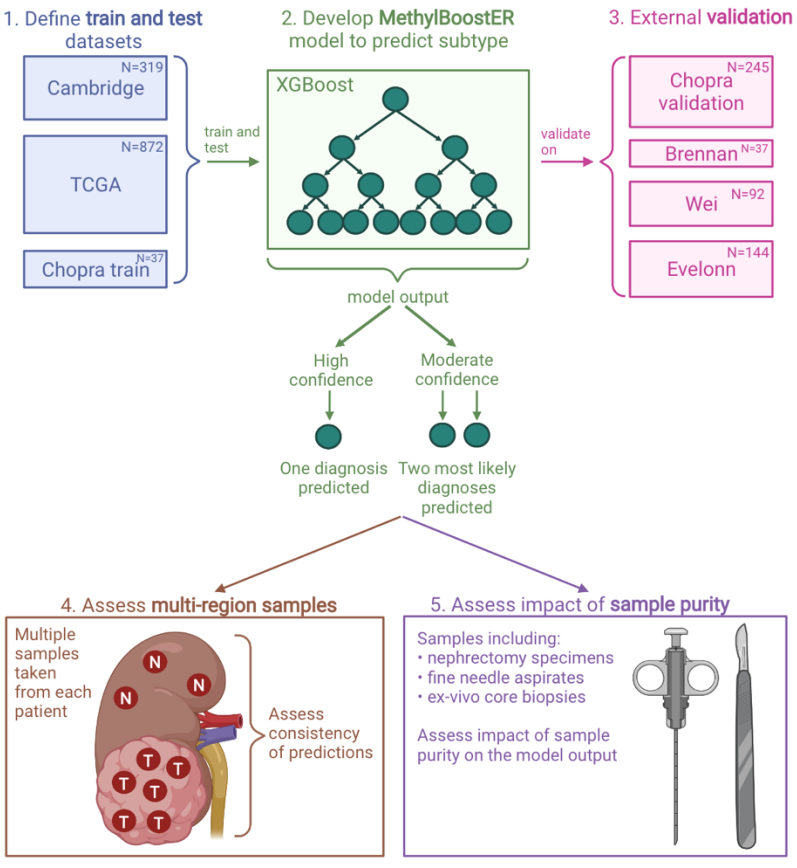


Figure 5.14: MethylBoostER analysis and sample overview

Methylation data were combined from three sources and these were used to train and test a machine learning model, called MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours). Subsequently, MethylBoostER was externally validated on four independent datasets.

5.3.2.2.1 Model development and results of the training and testing set

For the training and testing set, samples were combined from three sources, my Epic-seq data (N=319), TCGA (N=872)[24] and Chopra Training set (N=37)[120] (Figure 5.15A). For the remainder of the chapter, my Epic-seq data will be referred to as Cambridge data, in keeping with the nomenclature used in the manuscript submitted for publication (Appendix 1). As described in Methods section 4.5.2, data were available for 158,670 CpG probes for 1228 samples. The UMAP demonstrates that when using methylation values from all 158,670 CpG probes, samples cluster first by dataset, then by subtype (Figure 5.15B). This is not entirely unexpected due to the different data sources and platforms used. This combined dataset was split into training and testing sets (75:25 split) using four-fold nested cross-validation. An extreme gradient boosting (XGboost) machine learning model was developed to classify samples into one of five subtypes or 'classes': normal kidney, ccRCC, pRCC, chRCC and oncocytoma. The resulting model was named MethylBoostER. The 1331 features selected by MethylBoostER are visualised in Figure 5.15D, demonstrating that these separate samples by pathological subtype rather than by data source (as in Figure 5.15B). This suggests that the features selected by the model are not dataset specific but are features that distinguish each class in all three datasets.

First, I evaluated the model's performance in the testing set. The following model metrics were assessed: accuracy, precision and recall curves, Receiver Operating Characteristics (ROC) curves and confusion matrices (Figure 5.15C). Each of these metrics are defined in Methods section 4.5.2. In brief, the confusion matrix summarises the samples' true label and the label predicted by the model, for each pathological subtype. Four models were created since four-fold cross validation was performed, and the results of each of these models in the testing set is shown (Figure 5.16A-H). In all models, the ROC AUC was 1 for normal samples, and in 3 out of 4 models the ROC AUC was 1 for oncocytoma, demonstrating these were the best performing classes. In model 4, the ROC AUCs were: 1 for normal kidney, 1 for oncocytoma, 0.992 for ccRCC, 0.992 for pRCC and 0.988 for chRCC (Figure 5.16G-H). As this was the best performing model, it was selected to be taken forward into external validation in all subsequent analyses (results for other models are available in Appendix 1).

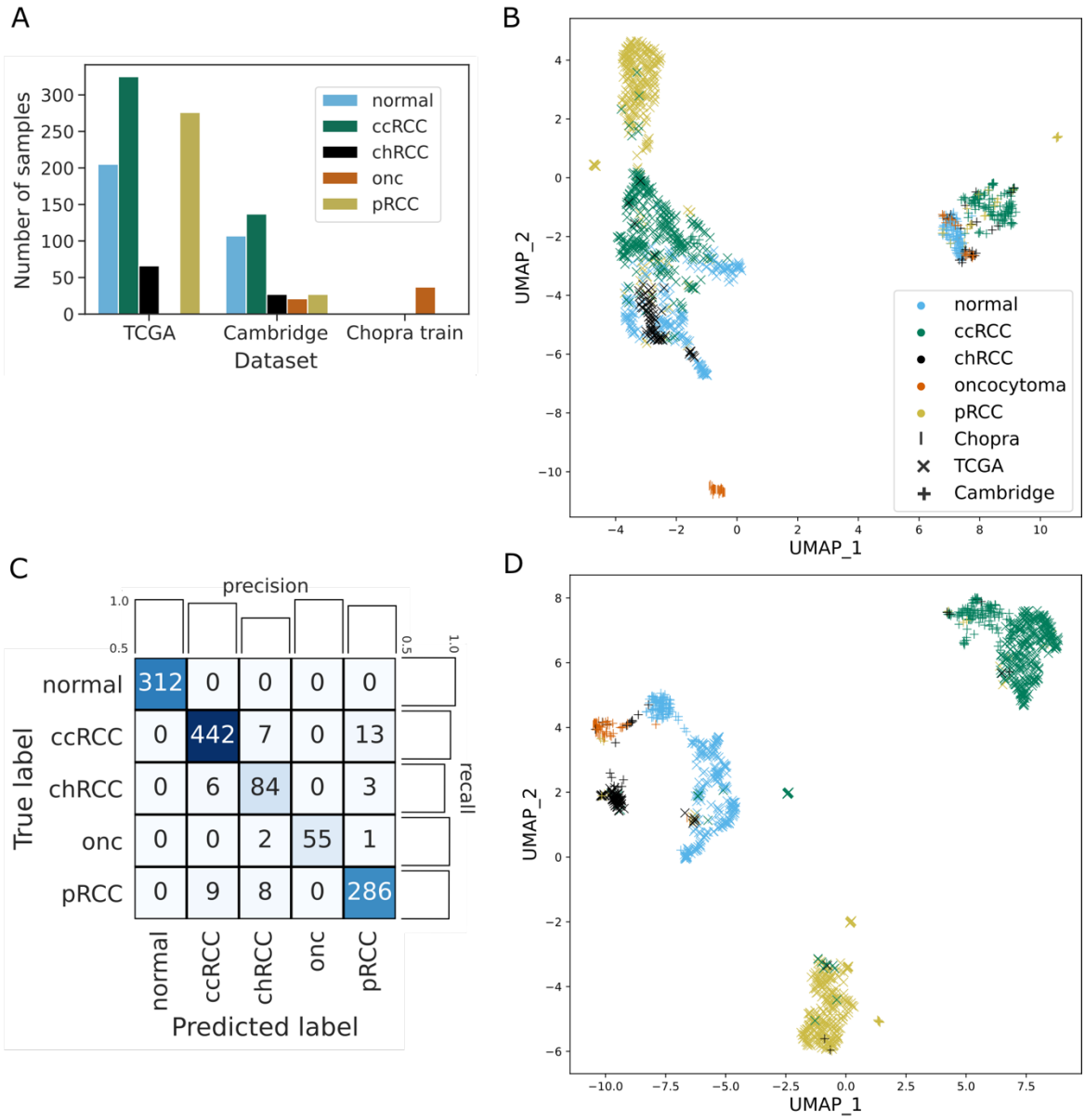


Figure 5.15: Data characteristics and MethylBoostER performance in the testing set

Panel A- Bar plot summarising the number of samples in the training/testing set, split by pathological subtype, for each dataset. Panel B- UMAP of the training/testing dataset, using all input features (158,670 CpG probes). Here samples broadly cluster by dataset. Panel C- Confusion matrix demonstrating performance in the testing set, with precision and recall bars. Panel D- UMAP of the training/testing set, using only features learnt by MethylBoostER (Model number 4; 1331 probes). Here samples broadly cluster by subtype rather than by dataset. Figures 5.15-5.25 were created in collaboration with Izzy Newsham as described in section 4.5.2.

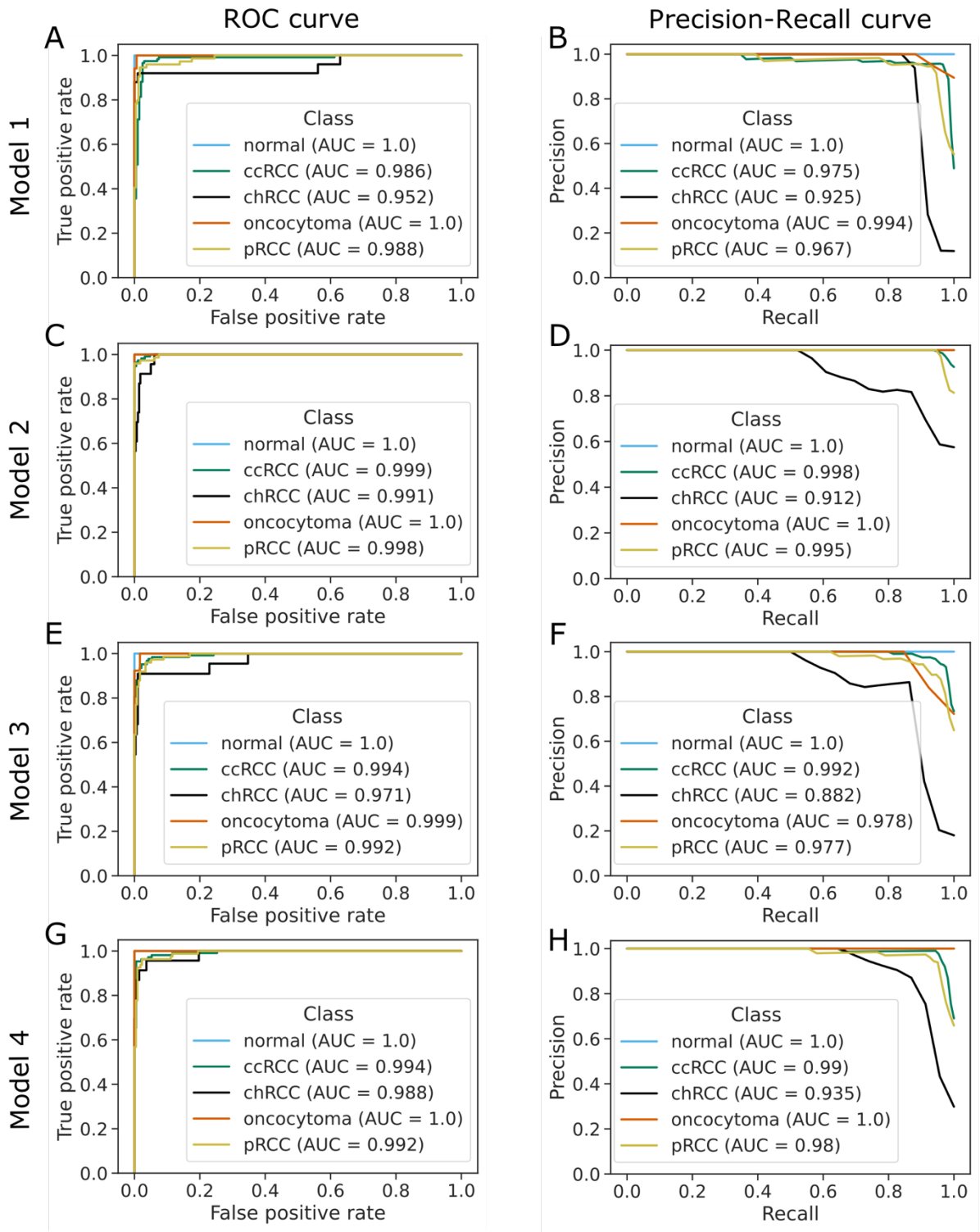


Figure 5.16: ROC and precision-recall curves over the testing set, split by pathological subtype

Four models were created since four-fold cross validation was performed, and the results of each of these is shown (Panels A-H). The area under the curve (AUC) is shown for each subtype.

I assessed the model's performance for each subtype in turn, to determine whether any subtype was misclassified more commonly than the others. chRCC was the worst performing class, and both the confusion matrix (Figure 5.15C) and the precision-recall curves confirmed this, as evidently the ROC curve for chRCC drops off steeply compared to other subtypes (Figure 5.16). Out of 1228 samples evaluated, 96% (1179/1228) were on the diagonal of the confusion matrix (Figure 5.15C), indicating they were correctly predicted (Accuracy: 96%). The most common misclassifications included ccRCC being predicted as pRCC (N=13) or chRCC (N=7), pRCC being predicted as ccRCC (N=9) or chRCC (N=8) or chRCC being predicted as ccRCC (N=6) (Figure 5.15C). Arguably, all three malignant subtypes would be managed with active treatment (rather than surveillance) therefore mitigating the consequence of these misclassifications. Difficulties in discriminating between ccRCC and pRCC may be attributed to their shared cell of origin, and this is consistent with my results shown in section 5.3.1.

Problems accurately differentiating between ccRCC and chRCC on histopathology are a well-known challenge. In TCGA, 15 cases were initially classified as ccRCC on histopathological slide review, however, these were later re-reviewed by specialist uro-histopathologists and reclassified as chRCC [24, 74]. Eight of these samples from TCGA were included in our testing/training dataset, and MethylBoostER classified five of these as chRCC and three as ccRCC. The former suggests that our model can correctly classify five chRCC samples better than a general pathologist, and more akin to a specialist uro-histopathologist. We explored the methylation and gene expression profiles of TCGA samples (Figure 5.17A-B) and demonstrated that the three samples which our model classifies as ccRCC, cluster more closely with ccRCC than chRCC based on both methylation and gene expression data (TCGA participant IDs: 4821, 4688 and 4696). We hypothesized that these samples may indeed be ccRCC (i.e. the first classification was correct rather than the re-classification). These findings highlight existing challenges in diagnosing subtypes using standard of care histopathology methods and emphasize the need to produce accurate predictive models. Another challenge is that predictive simulations are trained on datasets in which the true diagnosis is based on histopathology, and if this is incorrect, it may bias the model. Importantly, although some oncocytoma samples were classified as malignant disease, no malignant samples were classed as benign oncocytoma or normal tissue (Figure 5.15C). Given current clinical practice (which errs on the side of caution and removes SRMs which are later found to be benign), it would be acceptable to remove an oncocytoma in case this might be malignant, but it would not be acceptable to confuse a malignant mass as benign, as this would risk the cancer spreading.

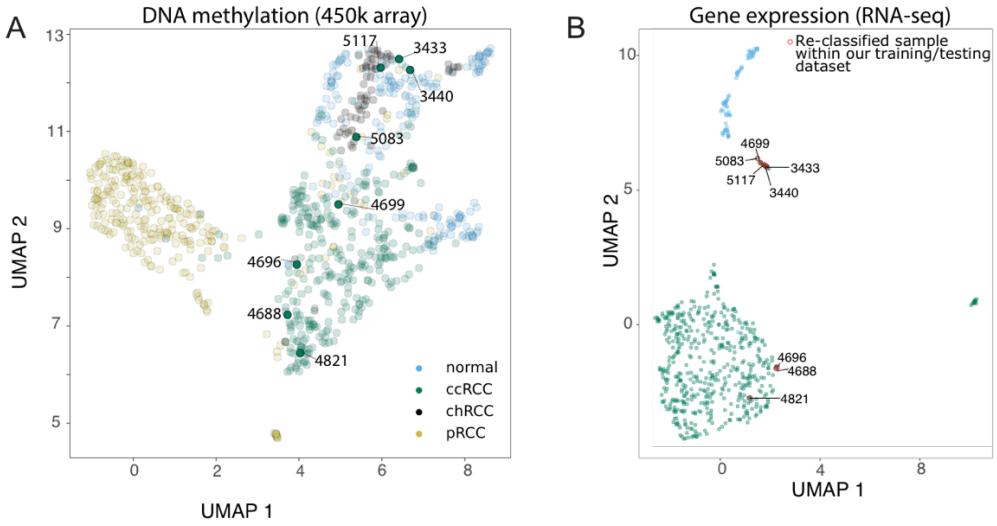


Figure 5.17: UMAP of all TCGA samples in the training/testing dataset

Panel A and B- UMAP of TCGA samples using methylation data (Panel A) and gene expression data (Panel B). There were 8 samples that were originally classified as ccRCC and then re-classified as chRCC following review by a uro-histopathologist, and these samples are labelled with their TCGA participant ID. Three of these samples (TCGA participant IDs 4821, 4688 and 4696) are classified as ccRCC by MethylBoostER, and appear to cluster more closely with ccRCC than chRCC for both methylation (Panel A) and gene expression (Panel B). This suggests these 3 samples may indeed by ccRCC (i.e. the first classification may be correct and the re-classification may be incorrect).

5.3.2.2.2 High- and moderate-confidence predictions make model outputs clinically more informative

It is recognised that a clinician would not use the MethylBoostER class prediction in isolation, but would integrate the model output with clinical, histopathological and radiological parameters to guide patient-centred care. In addition, a clinician requires an estimate of MethylBoostER's 'confidence' in the predicted output, which would allow adequate weight to be placed on the model result as well as other available clinical information. As such, we developed the concept of high- and moderate-confidence predictions. Given a particular sample, MethylBoostER will output predicted probabilities for that sample belonging to each of the five classes (normal kidney and renal tumour subtypes). For example, given sample X, MethylBoostER may predict that the sample is ccRCC (probability=0.90), chRCC (probability=0), pRCC (probability=0.10), oncocytoma (probability=0) and normal kidney (probability=0), which would lead the clinician to conclude that there is a high chance the diagnosis is ccRCC. However, in this case, producing all outputs is unnecessary as the probability of the diagnosis being ccRCC is very high; whereas if the probability was lower, the clinician would benefit from further information. In the testing set, the vast majority of cases (87%) had a predicted output probability >0.95, however in a number of cases this was lower, with probabilities as low as

0.26 (Figure 5.18A). High- and moderate-confidence predictions aim to address this. If MethylBoostER's predicted probability is greater than a certain threshold *t*, this is deemed a high-confidence prediction and the most likely output is provided. If MethylBoostER's predicted probability is below the threshold *t*, this is deemed a moderate-confidence prediction and the top two most likely outputs are provided (referred to as first and second predictions for the remainder of the analysis). The probability threshold selected was *t* = 0.85 (Figure 5.18B), as this maximises the accuracy of both high- and moderate-confidence predictions and the fraction of high-confidence predictions over the testing set. In the testing set, the average accuracy of high-confidence predictions was 0.982, whilst for moderate-confidence predictions this was 0.871 (where the prediction was treated as correct if the first or second prediction was correct). The use of high- and moderate-confidence predictions is a strength of the model as it allows the clinician to make an assessment of the probability that MethylBoostER is correct (i.e. the strength of the confidence of the prediction) and use this alongside clinical data, thus increasing the chances of the model becoming integrated in clinical practice (Figure 5.18).

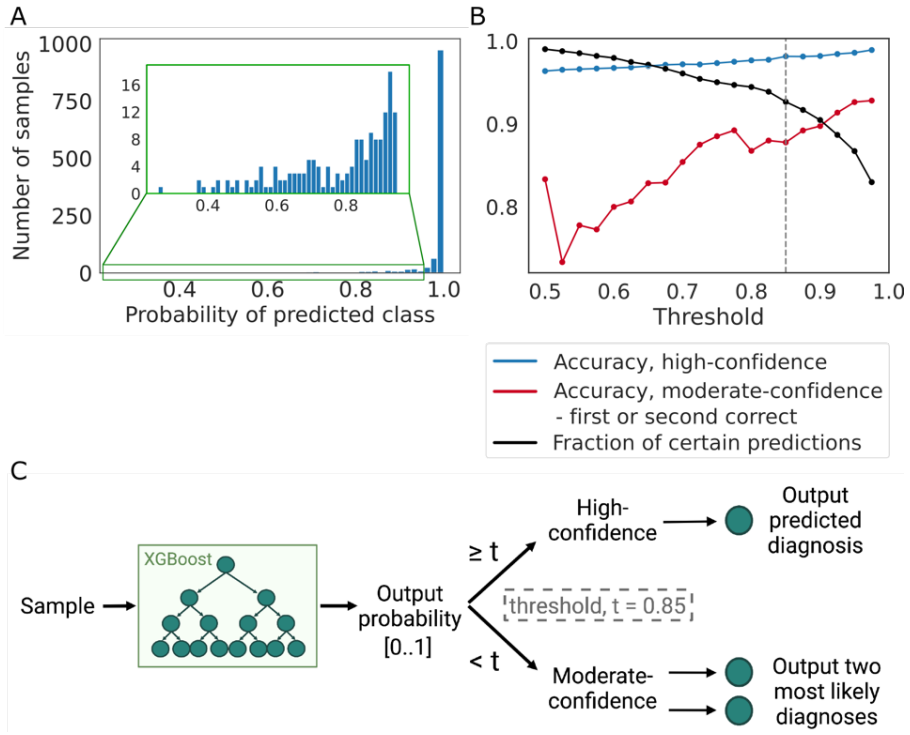


Figure 5.18: High and moderate confidence predictions in the testing set

Panel A- Histogram of the probability of the predicted class, for the testing set (N=1228). Panel B- Line plot demonstrating the accuracy of high and moderate confidence predictions and the fraction of certain predictions, for each threshold *t* (0-1). The threshold *t* = 0.85 was selected as this achieved the maximum accuracy and fraction of certain predictions (vertical dotted line). Panel C- Schematic overview of how high and moderate confidence predictions are incorporated into MethylBoostER. High confidence predictions (probability of the predicted class > 0.85) will output one class, whereas moderate confidence predictions will output the two most likely classes.

5.3.2.2.3 External validation in four independent datasets

MethylBoostER was externally validated using four independent datasets: Chopra validation (N=245), Brennan (N=37), Wei (N=92) and Evelönn (N=144) (Figure 5.19A) [120, 121, 126, 127]. First MethylBoostER was applied to these datasets without taking into account high- and moderate-confidence predictions, meaning only the first prediction was used. The accuracy was 0.824 for Chopra validation, 0.703 for Brennan, 0.875 for Wei and 0.894 for Evelönn (Figure 5.19B). Importantly, although some oncocytomas were misclassified as malignant, no malignant tumours were misclassified as oncocytoma (in keeping with results in the training set, section 5.3.2.2.1). Figure 5.20A-D illustrates ROC curves for each dataset. For all subtypes and all datasets, ROC AUCs were >0.90, meaning that the model can be generalised to different datasets. Normal kidney and ccRCC had the highest ROC AUC (>0.96 in all four external validation datasets), with the lowest values noted in chRCC and oncocytoma. Subsequently, I evaluated the utility of high- and moderate-confidence predictions in the external validation datasets.

Predicted probabilities were used to split outputs into high- and moderate-confidence predictions, which led to increased overall performance. The accuracy for each of these datasets is shown in Figure 5.19C, and as expected, accuracy in high-confidence predictions was high (>0.90 for all datasets). For moderate-confidence predictions, accuracy in the first prediction alone was <0.65 in all datasets, and this increased to >0.70 by taking into account second predictions, suggesting this may be a useful strategy. Figure 5.20A-D illustrates confusion matrices (splitting high- and moderate-confidence predictions) for each dataset. Subsequently, I evaluated commonly misclassified samples. In the Brennan dataset, the worse performing class was chRCC. Conversely, chRCC was the best performing class in the Chopra validation set (the worse performing class was oncocytoma). The lower model performance in chRCC and oncocytoma may be due to the comparatively low number of chRCC and oncocytoma in the training, testing and validation sets, relative to ccRCC and normal kidney samples (which reflects disease prevalence). Furthermore, as I have shown in section 5.3.1.1, chRCC and oncocytoma are the two most difficult subtypes to differentiate due to their shared cell of origin, and this is confirmed by the existing literature [117, 121]. Indeed, differentiating chRCC and oncocytoma is the main challenge in clinical practice. I noted that in the Chopra validation dataset, oncocytoma samples had low tumour purity, and I hypothesized that this may be driving the low prediction accuracy in this subtype. Indeed, in the Chopra validation dataset, 12 oncocytomas were predicted as normal kidney, and all but two of these sample had a tumour purity < 0.50. Oncocytomas may be predicted as normal due to their low tumour purity, or do to their benign phenotype. This prompted me to evaluate the impact of tumour purity on MethylBoostER.

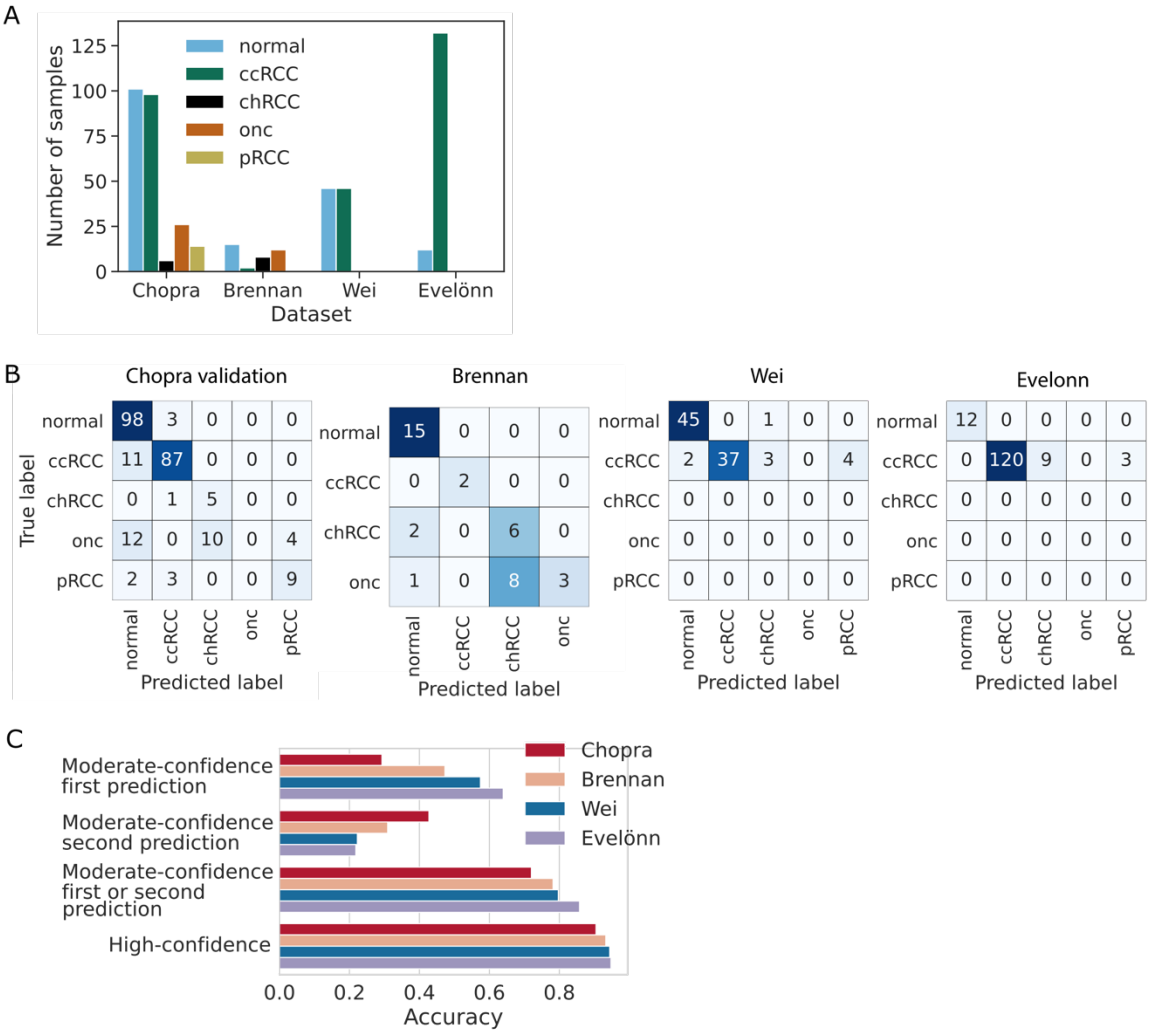


Figure 5.19: MethylBoostER external validation on four independent datasets

Panel A- Bar plot summarising the number of samples in each validation dataset, split by pathological subtype. Panel B- Confusion matrices demonstrating performance in each dataset (high- and moderate-confidence predictions are not considered). Panel C- Accuracy for high- and moderate-confidence predictions, for each dataset. For moderate-confidence predictions, accuracies are shown for the following three scenarios: the first prediction is correct, the second prediction is correct, the first or second prediction are correct.

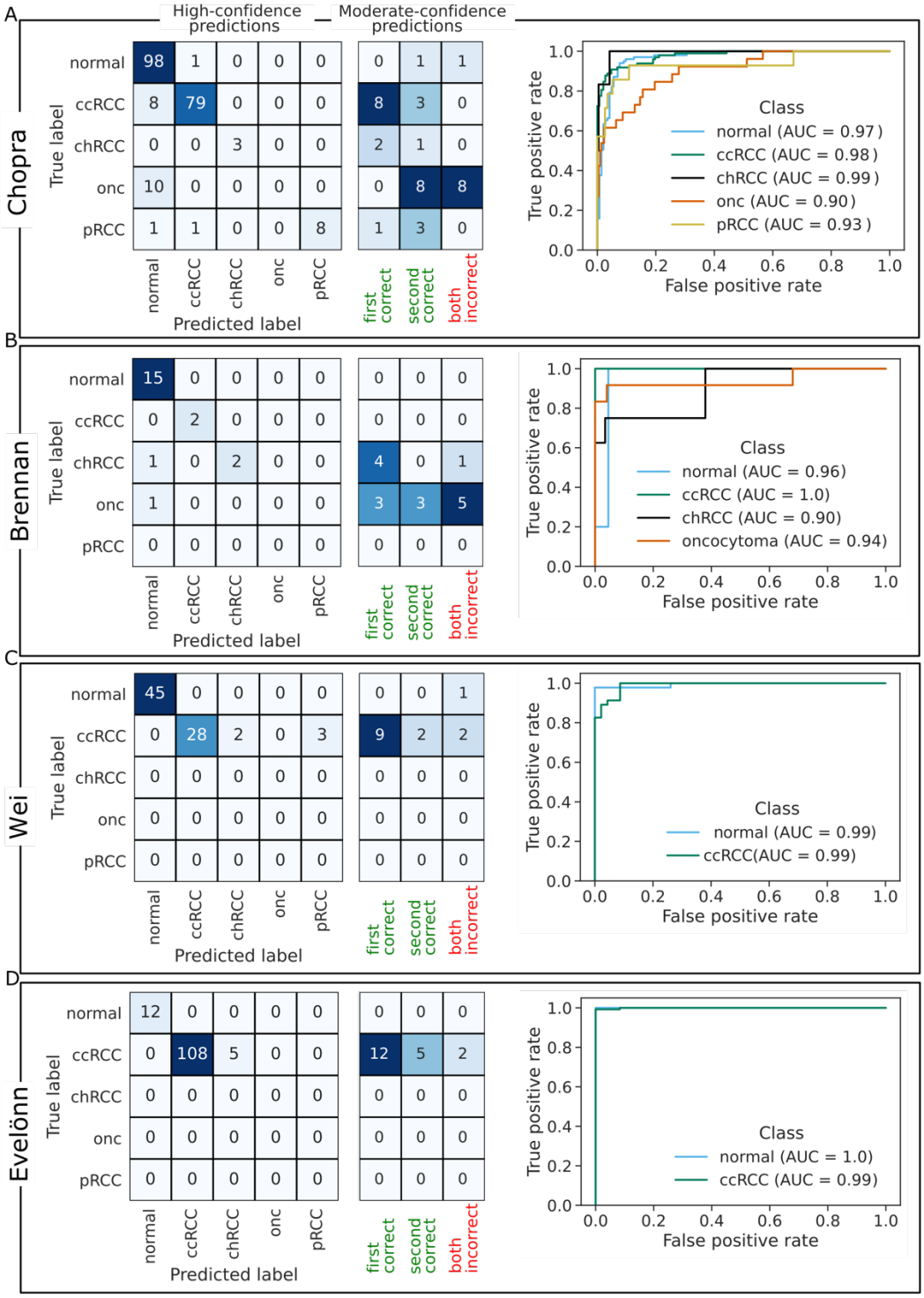


Figure 5.20: Confusion matrices and ROC curves, for the four external validation datasets

Results are shown for each dataset: Chopra validation (Panel A), Brennan (Panel B), Wei (Panel C) and Evelönn (Panel D). The confusion matrices show predicted and true labels for samples which have a high- and moderate-confidence prediction separately. For moderate-confidence predictions, there is a breakdown showing the number of samples correctly predicted in the first prediction, second prediction or incorrectly predicted (i.e. incorrect on first and second prediction). The ROC curves and area under the curve (AUC) do not take into account high- and moderate-confidence predictions.

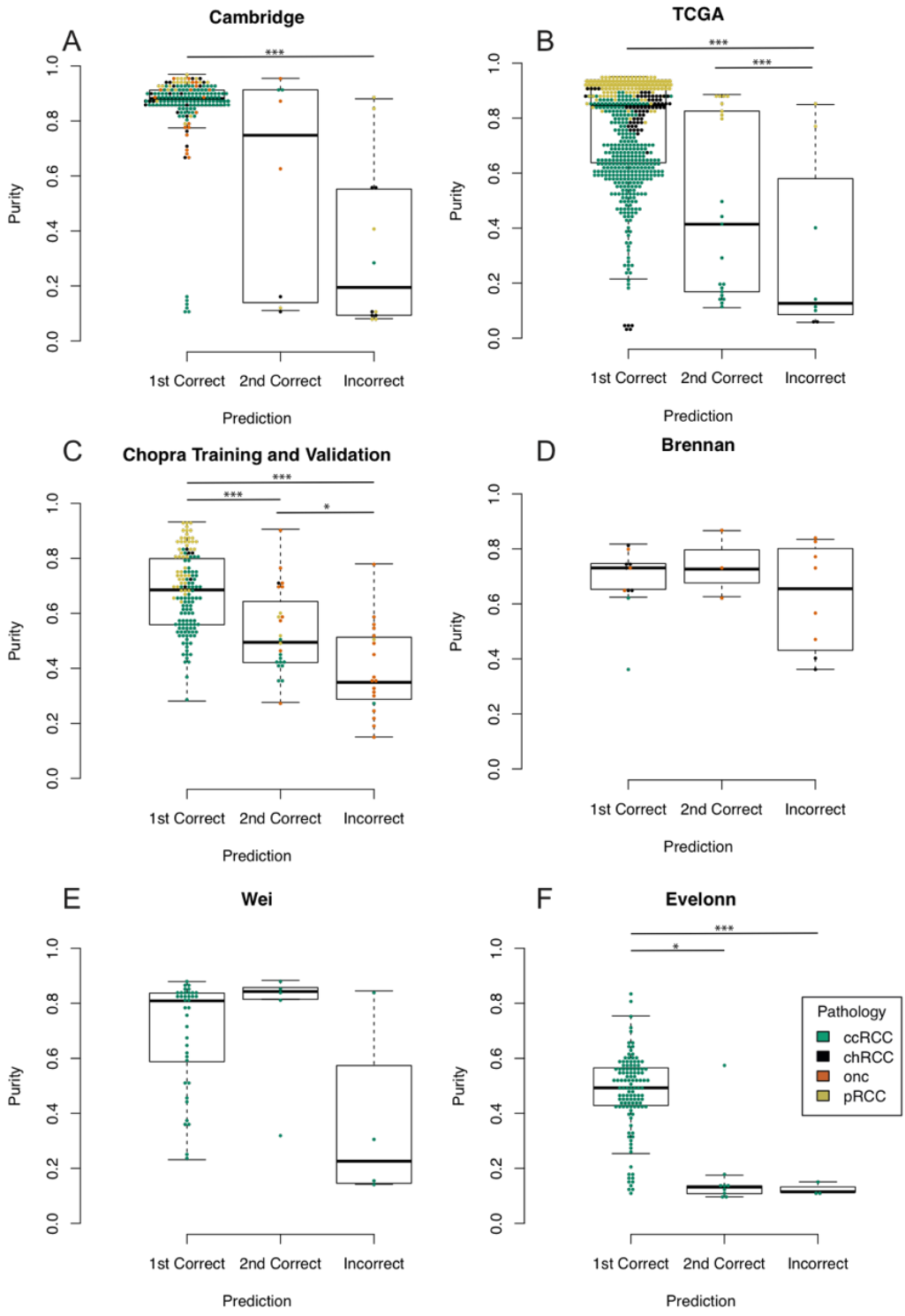


Figure 5.21: Purity in samples which are correctly predicted in the first prediction, second prediction or incorrectly predicted samples

Results are shown for each dataset: Cambridge (Panel A), TCGA (Panel B), Chopra Training and Validation (Panel C), Brennan (Panel D), Wei (Panel E) and Evelönn (Panel F). Incorrectly predicted samples were defined as samples that were incorrectly predicted on both first and second prediction. Adjusted p values are shown (* <0.05, **<0.009, *** <0.0009).

5.3.2.2.4 Exploring the impact of sample purity on MethylBoostER

Tumour purity was evaluated using methylation data using 'Infiniumpurify', which measures contamination with normal tissue. Normal kidney samples therefore do not have a purity estimate and are not included in this analysis. For each dataset, median purity was: 0.88 for Cambridge, 0.84 for TCGA, 0.80 for Chopra training, 0.58 for Chopra validation, 0.73 for Brennan, 0.81 for Wei and 0.48 for Evelönn (Figure 5.21A-F). Whilst Cambridge, TCGA, Chopra training and Wei consist of small tissue samples obtained during nephrectomy, the Chopra validation cohort consists entirely of ex-vivo core biopsy samples. The Brennan dataset contains small tissue samples obtained during nephrectomy and 4 fine needle aspirate (FNA) samples. Ex-vivo core biopsy and FNA attempt to reproduce clinical practice, where biopsy samples with limited tumour content may be obtained, and this explains the lower tumour purity in these two datasets. The low purity in the Evelönn dataset is unexpected and there is no obvious explanation other than potential poor-quality samples. Figure 5.22A depicts purity for samples that were predicted correctly in the first prediction, second prediction and incorrectly predicted samples, for all datasets combined. Median purity in samples that were correctly predicted on the first prediction was significantly higher than samples correct on second prediction, and those that were incorrectly predicted (0.82 vs 0.44 vs 0.29, adjusted p value <0.01 for all comparisons; Figure 5.22A). Figure 5.22B-C summarises purity and the prediction probability (for the first prediction), highlighting incorrectly predicted samples. Results are shown for all datasets combined, as well as individual datasets. There was a correlation between purity and probability of first prediction in Wei and Evelönn (Pearson correlation coefficient = 0.58 and 0.51 respectively, adjusted p value <0.01), but not in the other datasets (correlation <0.30 and/or adjusted p value >0.01). Figure 5.22B demonstrates that a subset of samples were incorrectly classified despite having a high-confidence prediction, and these samples tend to have lower purity.

Subsequently, I evaluated MethylBoostER accuracy at different purity levels, to identify a purity threshold which could be used to preclude sample inclusion in the model. Taking into account all datasets combined, Table 5.10 summarises the accuracy of first and second predictions, as well as the median probability of the first prediction, by tumour purity. In samples where purity was <0.2, there was a sharp drop off in accuracy, compared to samples in which purity was >0.2, suggesting that potentially a biopsy sample may have to be repeated if purity is below this level as MethylBoostER is highly likely to be inaccurate. Taking into account the entire cohort, 5% (61/1246) of samples had purity below this threshold, suggesting repeat biopsy would only be needed in a minority of cases. A limitation of the present analysis is that it was performed post-hoc, on both training/testing and validation datasets, therefore this remains to be externally validated. In

summary, selecting high-confidence predictions (i.e. probability of the first prediction >0.85) and removing low purity samples (purity <0.2) may maximise accuracy and ensure clinical utility.

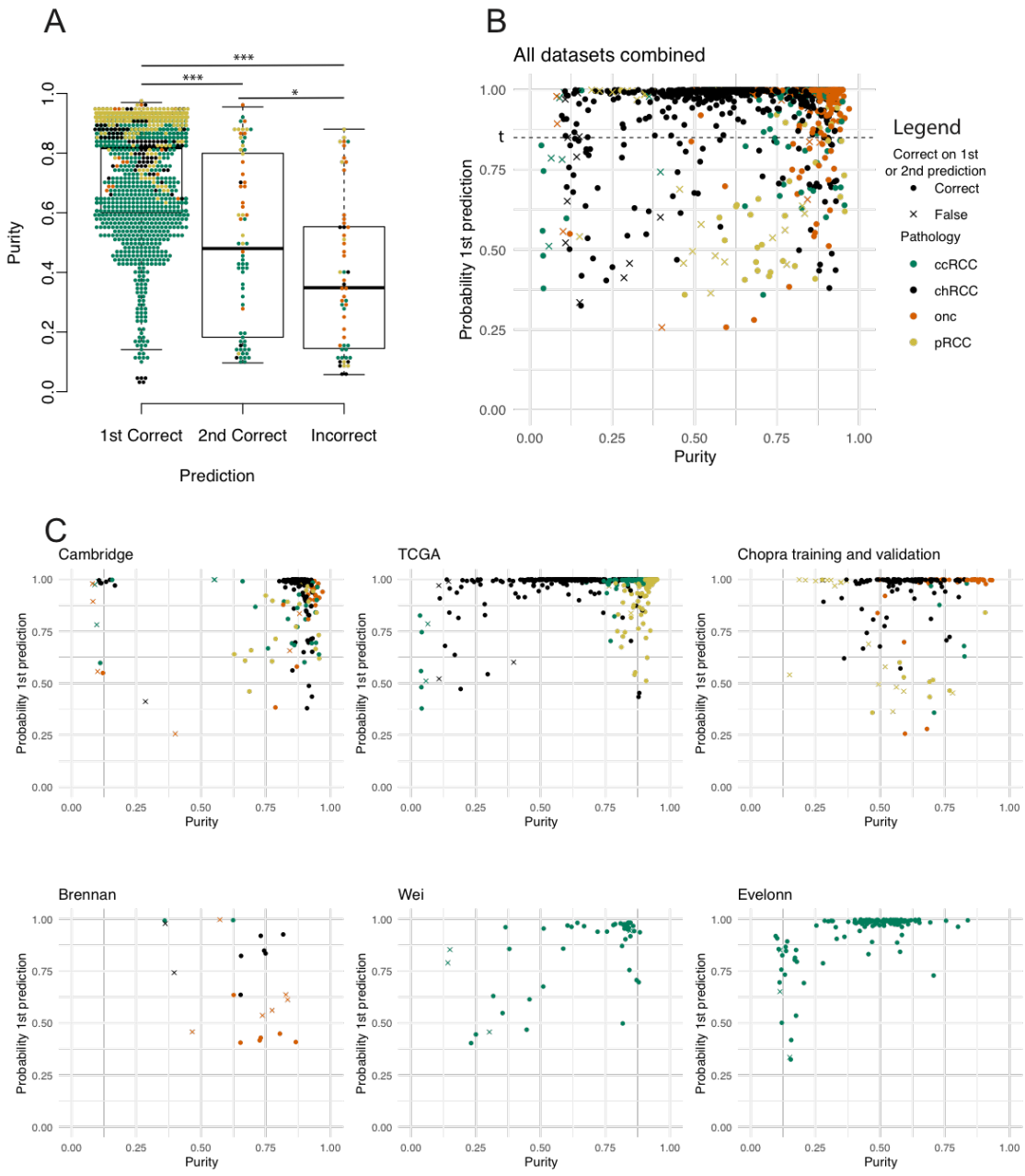


Figure 5.22: Sample purity and MethylBoostER output

Panel A- Purity in samples which are correctly predicted in the first prediction, second prediction or incorrectly predicted samples (i.e. incorrect on first and second prediction) combining all datasets. Adjusted p values are shown (* <0.05, **<0.009, *** <0.0009). Panel B and C- Purity and the prediction probability (for the first prediction), highlighting incorrectly predicted samples. Results are shown for all datasets combined (Panel B) and each individual dataset (Panel C). The threshold *t* = 0.85 indicating a high confidence prediction is shown. Samples which are incorrectly predicted (on both first and second prediction) are shown with a cross.

Table 5.10: MethylBoostER accuracy achieved for different purity thresholds

Normal kidney samples do not have a purity estimate and therefore are not included in this analysis, hence the total number of samples is 1246.

Purity	Median probability of 1st prediction	Accuracy of 1st prediction	Accuracy of 2nd prediction	Both predictions incorrect	Number of samples
0-0.1	0.78	0.42	0.08	0.5	12
0.1-0.2	0.85	0.37	0.39	0.24	49
0.2-0.3	0.99	0.71	0.08	0.21	24
0.3-0.4	0.98	0.6	0.1	0.3	30
0.4-0.5	0.99	0.82	0.14	0.04	95
0.5-0.6	1	0.91	0.04	0.05	159
0.6-0.7	1	0.97	0.03	0	159
0.7-0.8	0.99	0.92	0.05	0.04	111
0.8-0.9	0.99	0.94	0.04	0.02	294
0.9-1	1	0.99	0.01	0	313
All samples	1	0.9	0.06	0.04	1246

5.3.2.2.5 Exploring the impact of methylation heterogeneity on MethylBoostER

Methylation heterogeneity is relatively under-investigated; however it has important implications for diagnostic biomarker development and clinical applications. Multi-region samples from the same patient were evaluated to determine whether MethylBoostER predicts consistent results. Data were available for the Cambridge and Evelönn datasets, consisting of multiple samples obtained from different tumour regions and normal adjacent kidney from patients' nephrectomy specimens (Figure 5.23A). In the Cambridge dataset, multi-region samples were available (N=168) for 25 patients (18 ccRCC, 4 chRCC, 2 oncocytoma and 1 pRCC). In 92% (23/25) of patients, all multi-region samples were consistently predicted as being from the same pathological subtype; with 88% (22/25) achieving correct classifications for all samples (Figure 5.23B). Multi-region samples (N=17) were also available for 6 patients with ccRCC from Evelönn et al [127]. As shown in Figure 5.23C, 83% (5/6) of patients had a concordant prediction for all multi-region samples derived from the same patient. In three patients with ccRCC, some samples were classified as chRCC and in all cases, samples had a very low purity (i.e. below the 5th percentile), which may explain why these were misclassified. In TCGA, several ccRCC samples were re-classified as chRCC. For two patients from the Evelönn dataset (Figure 5.23C), all multi-region samples were predicted as chRCC rather than ccRCC, and it would be interesting to review the histopathology slides for these samples to determine whether they may indeed be re-classified on further review. Overall, these data suggest that multi-region samples achieve consistent predictions in 90% (28/31) of patients, meaning biopsies are likely to achieve

consistent diagnoses in clinical practice. Further exploration of methylation heterogeneity in ccRCC is performed in the next chapter (Chapter 6).

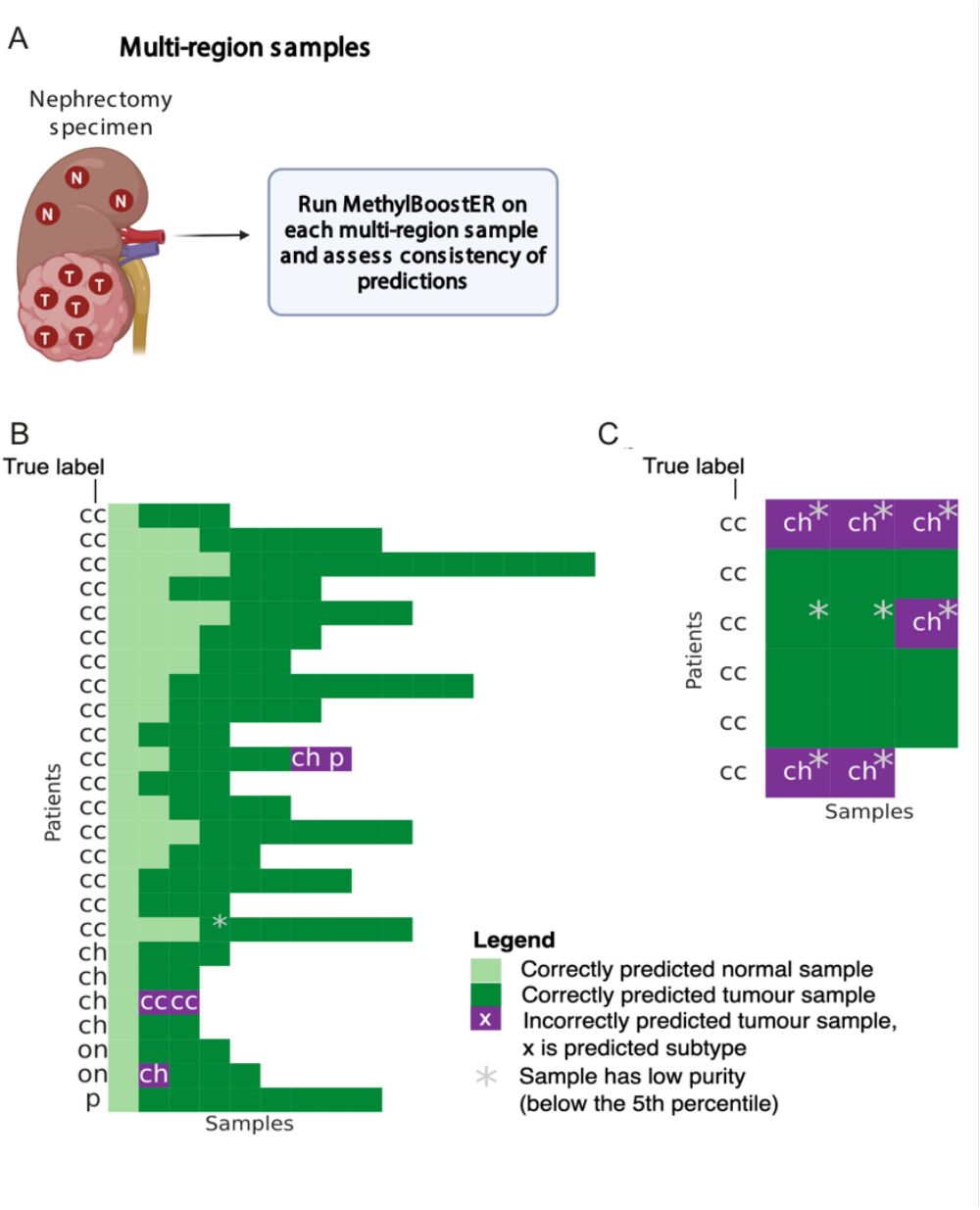


Figure 5.23: Classification results for multi-region samples

Panel A- Schematic showing analysis plan. Multi-region samples (i.e. multiple tumour and normal kidney samples) were obtained from each patient. Panel B and C- Diagram showing classification results for the Cambridge (Panel B) and Evelönn (Panel C) datasets. Each row is a patient, and the prediction for each multi-region sample is shown, along with whether the prediction was correct or incorrect (and the predicted class if incorrect). Samples with a low purity (below 5th centile) are highlighted with a star. Abbreviations: cc = ccRCC, ch=chRCC, on= oncocytoma, p=pRCC.

5.3.2.2.6 Potential clinical utility of MethylBoostER

Although MethylBoostER requires further validation, Figure 5.24 explores how it may be integrated in clinical practice in future (following further validation). Patients diagnosed with SRMs would undergo full clinical evaluation, including imaging and image-guided biopsy, which would then be analysed for DNA methylation, tumour purity assessment and MethylBoostER class prediction. Based on the results, if samples have low tumour purity, it may be advisable to repeat the biopsy as the model has low accuracy and may not be reliable in this setting. Additionally, if MethylBoostER predicts normal tissue, this is likely to suggest that despite image-guided biopsy, the renal mass may have been missed, and the biopsy should therefore be repeated. For a given sample, the MethylBoostER output would be integrated with clinical and imaging data to determine the most appropriate management option. High-confidence predictions indicating oncocytoma may suggest conservative management, whilst high-confidence predictions indicating malignant subtypes (ccRCC, pRCC and chRCC) would suggest active treatment. For moderate-confidence predictions, the two most likely predictions would be interpreted along with prediction probabilities, sample purity and clinical data, and management options discussed with the patients. The aim of this would be to improve the patient diagnostic pathway and enable patient-centred decision making regarding the most appropriate management options.

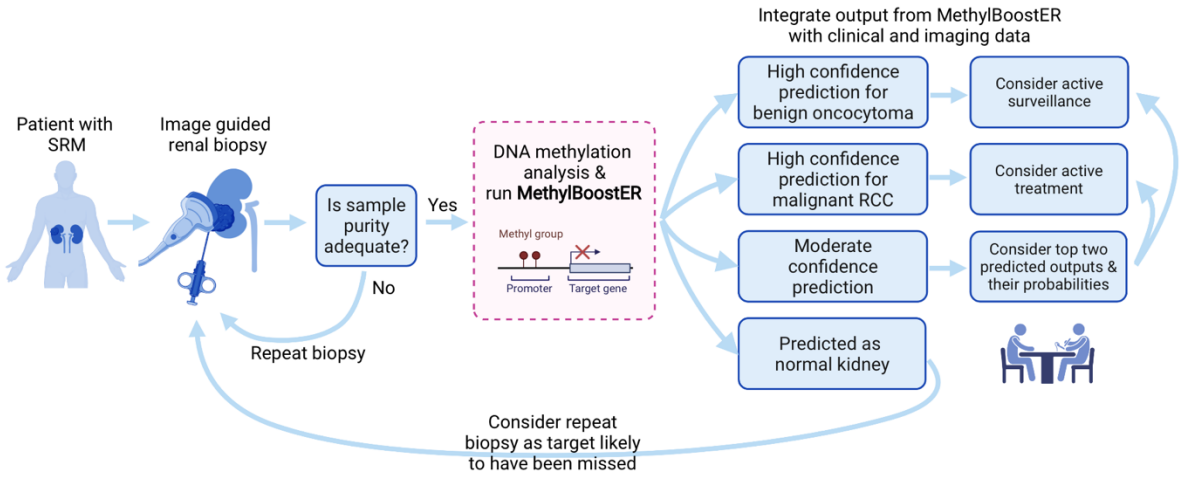


Figure 5.24: Proposed future integration of MethylBoostER into the existing clinical pathway for patients with SRMs

5.4 Discussion and future direction

5.4.1 Methylation & gene expression in pathological subtypes of renal tumours *vs* normal kidney

The first part of this chapter (section 5.3.1) focuses on characterising DNA methylation and gene expression in tissue derived from patients with common pathological subtypes of benign and malignant renal tumours. First, I compared normal kidney samples derived from patients with different pathological subtypes of renal tumours to evaluate the presence of a 'field-effect.' Field cancerization refers to the concept that normal tissue may acquire early changes that predispose to tumorigenesis [276, 277]. In normal kidney samples in my dataset, there was no evidence of clustering by adjacent tumour subtype in the unsupervised analysis. I then demonstrated there was no significant difference in methylation in normal kidney derived from patients with ccRCC and chRCC (two malignant tumours that have different cells of origin). Taken together, these results do not show evidence of a field effect (i.e. nothing to suggest that the normal kidney may display methylation changes which are precursors to a specific tumour subtype). Such a methylation field effect has been demonstrated in other organs, such as prostate tissue from patients with and without cancer [276, 277]. In the kidney, contradictory evidence exists in the literature. Arai et al showed that there are methylation differences in normal samples derived from patients with ccRCC versus chRCC [278, 279]. The authors hypothesize this is evidence that DNA methylation changes may occur in normal kidney which may predispose to cancer, and that these changes differ based on the tumour types' differing cells of origin. Malouf et al showed similar results to my own: DNA methylation was the same amongst normal kidney samples regardless of adjacent tumour subtype [117]. The literature suggests that on histopathological slide review, normal kidney samples derived from patients with different tumours show no obvious difference and there is also no difference compared to normal kidney from patients without tumours [278, 279]. Therefore, there is no evidence to suggest a field effect in normal kidney macro and microscopically, though this could still be the case on a molecular level. Whereas pre-malignant lesions have been identified in other tumour types (such as adenomatous polyps in colorectal cancer and CIN in cervical cancer), no such lesions have been identified for RCC. The identification of precursor lesions is a research priority as it could enable earlier cancer detection. Mitchell et al identified complex structural rearrangements (LOH at 3p and gains at 5q) which may occur in adolescence decades before ccRCC develops [71]. Since early events in the normal kidney affect structural rearrangements, it might be plausible that methylation changes may occur, though no evidence was noted in my dataset. My analysis

represents preliminary work based on a relatively small cohort of normal tissues, and future research (given more time and resources) could evaluate the concept of field effect more thoroughly. For instance, one could evaluate methylation disorder in a read (for example by calculating epipolymorphism) rather than methylation at individual CpGs. This idea is discussed more in detail in the discussion section of Chapter 6. It could also be useful to compare normal kidney tissue derived from patients with renal tumours and patients without tumours (for example from autopsies) to highlight changes that might be associated with tumorigenesis rather than pathological subtype specific changes [279].

Single cell RNA-seq studies have demonstrated that in benign renal cortex tissue, 50% of cells are noted to be proximal tubule (PT) cells, compared to only 10% in benign renal medullary tissue [228]. It is these PT cells that are the cell of origin for ccRCC and pRCC. In addition, the renal cortex has a high proportion of glomerular (vascular) cells and a low proportion of collecting duct cells relative to the medulla [227]. Lindgren et al evaluated RNA-seq data from 'normal kidney' samples from TCGA (including normal kidney from patients with ccRCC, pRCC and chRCC) and demonstrated that there were five subclusters which correspond to varying degrees of similarities with the normal cortex and medulla [227]. The authors go on to demonstrate that variation in gene expression amongst 'normal kidney' samples from TCGA may be explained by samples being collected from different anatomical locations of the normal kidney cortex and medulla. In my study, as in the vast majority of studies in the literature (including TCGA), there is no distinction between normal kidney derived from the cortex or medulla. It is therefore important for future research to document which part of the kidney normal samples have been taken from as this may affect results obtained from comparing tumour vs normal. Lindgren et al [227] analyse variation in gene expression in normal kidney; future research could evaluate whether DNA methylation changes recapitulate similar patterns.

I evaluated DNA methylation and gene expression data in four pathological subtypes of renal tumours (ccRCC, pRCC, chRCC and oncocytoma) and compared each of these to normal kidney, highlighting similarities and differences amongst subtypes. Compared to normal kidney, oncocytoma and chRCC were characterised by pronounced hypomethylation, whereas pRCC was characterised by hypermethylation and ccRCC demonstrated relatively more hypo than hypermethylated DMCs. Unsurprisingly, hypermethylated sites tended to be located in gene promoters whereas hypomethylated sites occurred at the gene body. In my analysis, subtypes derived from the same cell of origin demonstrated similar DMCs and DEGs compared to normal kidney (i.e. ccRCC and pRCC are derived from the PCT, whereas chRCC and oncocytoma are derived from the distal nephron), and

this was confirmed by the literature [74, 117]. Subsequently, I integrated data from methylation and gene expression to postulate which genes might be epigenetically regulated. One of the major findings was that in all four pathological subtypes, genes associated with kidney embryological development and cell differentiation (such as *WT1*, *ASS1*, *TGFBI*, *EPCAM*, *LGALS3*, *LHX1*, *HOXD10* and *HOXD7*) demonstrated a negative association between promoter methylation and gene expression, suggesting these are key pathways that may be involved in tumorigenesis and might be epigenetically regulated. DNA methylation plays a key role in kidney embryological development [280]. Studies evaluating renal embryological morphogenesis have identified a number of genes and key pathways including: *LHX1*, *WT1*, *GATA3*, *PAX* and *HOX* genes, genes involved in *Notch* and *Wnt* signalling, *PI3* and *MAP* kinase signalling, fibroblast growth factor receptors (*FGFR*) and genes involved in EMT and cell differentiation [281]. Indeed, these are the same genes and pathways which are disturbed in renal tumours in my analysis. Interestingly, many of the genes are also responsible for neuron development, such as *GREM1* and *NEURL* (which explains why often gene enrichment analysis in renal cancer reveals pathways associated with neuronal development).
My data also suggests a negative association between methylation and gene expression in genes which are involved in mediating EMT, cell adhesion and signalling pathways involving growth and differentiation (for example cadherins, claudins and integrins). This suggests that pathways which are required for normal cell differentiation during embryological development are deranged in renal tumorigenesis, and these changes may be epigenetically regulated. In addition, I show that methylation/expression patterns were noted to reflect the tumours' cell of origin. For example, markers of the distal nephron were under-expressed and demonstrated hypermethylation in tumours derived from the PCT, whilst they were over-expressed and hypomethylated in tumours originating from the distal nephron, and vice versa for PCT markers. Many of these genes encode for solute membrane transporters (such as aquaporins and members of the solute carrier family), which in the normal kidney have specific gene expression patterns along the tubule to reflect kidney physiological demands for molecule and water transport and are therefore cell-type-specific markers. Furthermore, my analysis suggests that in ccRCC vs normal tissue epigenetic changes which may have a functional relevance are noted in pathways involving immune cell signalling (especially T cell activation) and angiogenesis, in keeping with findings from the literature, which suggests these changes are characteristic of ccRCC [74]. These subtypes specific differences could be harnessed as potential diagnostic markers.

Further to my observation that genes involved in renal development, cell differentiation and proliferation may be epigenetically regulated, I discuss *TGFBI* more in detail, as it exemplifies

methylation and expression differences between pathological subtypes of renal tumours. I hypothesized that different gene functions may be cell-type-specific and context specific, which could explain the differences noted between pathological subtypes of renal tumours. For example, *TGFBI* (transforming growth factor beta induced protein) is involved in late branching and maturation of the ureteric bud, which eventually leads to the formation of the collecting duct [282]. In addition, *TGFBI* has been implicated in cell growth, proliferation and tumorigenesis; with upregulation noted in cancers of the aerodigestive tract, thyroid, brain, liver and gallbladder [283]. My results demonstrated hypomethylation and increased expression of *TGFBI* in ccRCC and pRCC, with hypermethylation and reduced expression in chRCC, and this was supported by the existing literature [283]. Increased expression in ccRCC and pRCC may represent *TGFBI*'s role in promoting proliferation and tumorigenesis, whereas reduced expression in chRCC may reflect de-differentiation of the collecting duct (representing *TGFBI*'s role in the development of the collecting duct cells, from which this subtype is derived). The strong negative correlation between methylation and gene expression noted in all subtypes, suggests this is in part mediated by methylation. A better understanding of the biology underlying the differences in tumour subtypes could lead to an enhanced understanding of the processes underlying tumorigenesis, as well as advances in identifying biologically meaningful diagnostic markers.

My analysis integrated data from methylation and gene expression to postulate which genes may be epigenetically regulated, however there are a number of limitations. Importantly, I was limited by the absence of matched DNA methylation and gene expression for all four subtypes. Future efforts should focus on generating such information, which would be invaluable given the absence of similar datasets in the public domain. It is acknowledged that an association between methylation and expression does not imply causation. In order to demonstrate causation, one could perform cell line experiments using de-methylating agents to observe the impact on gene expression. My analysis focused on methylation changes in the promoter region of genes, as these changes are more likely to be functionally relevant, however a more comprehensive analysis could evaluate changes in the gene body. Furthermore, I used methods which measure mRNA levels (such as RNA-seq), which may not necessarily reflect protein levels (as there may be post-translational modification, protein degradation etc) [133].

5.4.2 MethylBoostER machine learning model to predict pathological subtypes of benign and malignant renal tumours and normal tissue

Achieving an accurate diagnosis for patients with SRMs has been identified as a research priority [1], as this would reduce the number of individuals who are found to have benign disease post-operatively. Reducing over-treatment would benefit both patients with SRMs (avoiding unnecessary morbidity and risk of mortality) and the health service (freeing up resources to treat other diseases). In summary, data were combined on >1200 kidney tissue samples to develop MethylBoostER, a machine learning model which classifies samples into one of five classes (ccRCC, pRCC, chRCC, oncocytoma and normal kidney). The model was externally validated on >500 samples from four independent datasets, achieving a high accuracy (AUCs >0.90 for all sub-types). Although MethylBoostER is promising, a number of challenges must be discussed. Whilst the main aim of the analysis is to distinguish pathological subtypes of SRMs, due to insufficient sample sizes, data were pooled from both SRMs and larger tumours for the training/testing set and some of the validation cohorts. However, the model was validated on the Chopra dataset which consists exclusively of SRMs (245 samples) and demonstrated excellent accuracy and generalisability of results in smaller tumours and in low purity samples (e.g. Evelönn and fine needle aspirates in Brennan). The two classes which achieve the lowest AUCs in the external validation are chRCC and oncocytoma, which are classically the most difficult to distinguish due to their similar cell of origin and similar methylation features, as well as relatively low prevalence. Future work should include more samples from these two subtypes (whilst prioritising samples from patients with SRMs) and attempt to increase accuracy for chRCC and oncocytoma.

Ultimately, any diagnostic model is limited by the quality and number of samples which were used for training and testing. The TCGA re-classified 8 ccRCC samples as chRCC, although our analysis of gene expression and DNA methylation suggests that only 5 of these samples are chRCC, whilst the remaining 3 may be ccRCC. This highlights known difficulties in assigning sample class based on histology review and suggests that in theory samples from datasets other than TCGA might also be re-classified if they were to be reviewed by an expert uro-histopathologist. This has the potential to either increase or decrease the accuracy in MethylBoostER. Ideally, all samples would undergo review by an expert uro-histopathologist (and/or undergo molecular analysis) prior to being made publicly available (including my dataset), however due to limited resources this is often not possible. In future, MethylBoostER will also be run using the updated classification of TCGA samples to

evaluate whether this makes an impact on the model accuracy (this may be particularly interesting as it may improve prediction of chRCC, one of the worst performing classes).

In this thesis, I explore approaches to improve model accuracy, including the use of high- and moderate-confidence predictions, and an evaluation of tumour purity. A strength of the analysis is the inclusion of low purity samples (including ex-vivo core biopsies and FNA) to represent a real-world scenario. Although MethylBoostER can produce correct classifications even in samples with low tumour content (Table 5.10), MethylBoostER is likely to be inaccurate when purity is <0.20, suggesting that potentially a biopsy sample may have to be repeated if purity is below this threshold. Low tumour content in biopsy samples limits histopathological review in existing clinical practice, and has also been found to reduce accuracy of methylation models in the literature [121]. It is customary to take more than one biopsy sample and this may mitigate this challenge. Importantly, my work is the first classification model of renal subtypes to extensively explore methylation heterogeneity. MethylBoostER achieves consistent results when evaluating multi-region samples in 90% of patients, suggesting that methylation heterogeneity does not limit model applicability.

MethylBoostER can be interpreted in the context of existing models. In the literature, a number of different approaches have been taken to differentiate SRMs including analysing molecular markers based on gene expression profiles [240, 241, 284, 285], miRNA [286-288] or DNA methylation [119-121, 289]. Additionally, machine learning models have been developed to classify subtypes based on imaging features on CT [290, 291] or histopathological slides [292]. None of the models have been adopted in clinical practice, and they each have strengths and limitations. A large proportion of these studies have small sample sizes (<200 samples overall) [119, 284-286, 288-290, 293]. Numbers of chRCC and oncocytoma are often even more limited as these are the least prevalent, but most difficult to distinguish subtypes. Furthermore, studies often lack external validation, which is often a concern in biomarker research. A sufficiently large number of samples in the training set and external validation sets are crucial to reduce overfitting. MethylBoostER is trained on the largest DNA methylation cohort of renal tumour samples to date and is extensively validated (including >1700 samples overall), which is a strength of this work. Previously published molecular classifiers are often limited by focusing solely on distinguishing oncocytoma from chRCC, excluding the other, more common malignant subtypes [119, 121, 240, 284]. Although this approach simplifies classifiers (by reducing the number of predicted classes and ensuring even numbers of samples in each class), this reduces applicability in the real world as it becomes difficult to interpret the output when common diagnoses are missing. Our model was therefore designed to include the most common

pathological subtypes, although future models would ideally also include angiomyolipoma (the second most common benign tumour) and may consider inclusion of less common tumour types.

Existing models often rely on a limited number of markers (e.g. <100 markers) [119-121, 284], and this approach has both advantages and disadvantages. Small marker panels may be cheaper to develop, meaning they may be more readily accepted into clinical practice. However, a small panel such as the one created by Pires-Luis et al [119], which relies on promoter methylation of just 3 genes may be more prone to bias if there is measurement error in a single marker or missing data. MethylBoostER selects approximately 1300 methylation markers, making it more robust when applied to heterogeneous clinical samples. The large number of CpGs selected is an advantage if this were to be applied to liquid samples (rather than biopsy samples) in future. The use of liquid samples (such as plasma and/or urine) would be an ideal, non-invasive method to differentiate pathological subtypes of SRMs, replacing renal biopsy and therefore avoiding the risk of complications associated with the latter. Furthermore, liquid samples (blood or urine) could be taken in primary care or the outpatient clinic, avoiding the significant resource use associated with renal biopsy (biopsies are usually performed in the radiology day unit, and require 6 hours of bed rest following the procedure). In Chapter 7, I describe the Nimbus method, which was developed in our lab to evaluate DNA methylation markers in liquid samples. Since the levels of ctDNA are very low in patients with renal tumours (and are even lower in early stage disease), one of the potential strategies to improve circulating tumour DNA detection is to increase the number of targets analysed at high sequencing coverage [154], for example by targeting many thousands of methylation markers.

Future work should focus on improving the accuracy of MethylBoostER, as well as prospectively validating the model on in-vivo clinical biopsy samples. Advances in next generation sequencing and reducing costs means such tests could be used clinically in future. In practice, models which are based on data which are readily available (such as CT or histopathological slides) or relatively inexpensive, have a higher chance of being adopted. Importantly, MethylBoostER would not be used in isolation but combined with other clinical and imaging data. Ideally, a model would integrate multi-modal data including DNA methylation, imaging and clinical characteristics, and be built on information from multiple sites across the UK (or internationally) to increase sample sizes. Initiatives such as the Mark Foundation Institute for Integrated Cancer Medicine may facilitate this in future. Ultimately, an ideal model would predict patient outcome rather than simply predicting the pathological subtype of the renal tumour, and would take into account comorbidities and competing

risks of death. For example, although many chRCC are indolent, there are also aggressive chRCCs which would benefit from active treatment rather than surveillance. Future DNA methylation models might be able to provide information regarding tumour aggressiveness, for example by identifying samples with a CpG island methylator phenotype or other prognostic methylation markers. In conclusion, in this chapter, I have performed a comprehensive evaluation of DNA methylation in different pathological subtypes of renal tumours and have reflected on potential clinical utility in a diagnostic setting.

Chapter 6 DNA methylation heterogeneity in ccRCC tissue samples

6.1 Brief introduction

The study of intra tumoral heterogeneity (ITH) in ccRCC has diagnostic, predictive, and prognostic utility and has been identified as a research priority in kidney cancer [1]. ccRCC is characterised by significant genetic heterogeneity, for both mutations and somatic copy number aberrations (SCNAs). Aberrations that are present in all multi-region tumour samples from a patient are considered clonal, early events whereas those which demonstrate ITH are subclonal, late events in tumour evolution [294]. In a series of landmark studies led by the Swanton group, somatic mutational heterogeneity (i.e. presence of subclonal mutations) was noted in 100% of patients analysed and 73% of driver mutations were not identifiable in all multi-region samples from the same patient [82]. On average, 7 biopsies were needed to recapitulate ≥75% of driver mutations [87]. In addition, 75% of driver SCNA demonstrated heterogeneity across multi-region samples. For example, loss of chromosomes 4q, 8p, 14q and gains on chromosome 5q were clonal in some individuals with ccRCC but subclonal in other patients. Mutational and SCNA events were both highly concordant and correlated, suggesting both processes can often co-occur [82]. In particular, *BAP1* was the mutation which was associated with the highest prevalence of SCNAs within a clone, in keeping with this gene's known role in DNA repair and maintenance of DNA integrity [87, 295]. The degree of ITH and evolutionary pattern has been suggested to have prognostic potential in ccRCC [87, 296]. Patients with a high ITH index had reduced progression free survival compared to patients with a low ITH index (ITH index was defined as the number of subclonal drivers divided by the number of clonal driver events). Additionally, seven evolutionary subtypes were identified, each characterised by specific patterns of events, and these predicted the speed of metastatic progression [75]. Patient risk stratification is a clinical priority, therefore further characterisation of ITH, and the association with prognostic outcomes is warranted.

One in three patients with ccRCC undergoing treatment with curative intent will develop a recurrence, however accurate post-operative risk stratification scores to predict recurrence are lacking [5]. Prognostic risk scores are key to tailor surveillance protocols and may enable adjuvant therapy. Gene expression (transcriptomic) ITH has been widely observed, and this has hampered attempts to identify prognostic risk scores. Examining the S3 score, ITH was noted in 60% of patients [91]. Similarly, for the ccA/ccB score, ITH was noted in 80% of individuals [68]. Morphologic ITH,

encompassing tumour cytology, architecture and micro-environment, has also been noted [297]. For several decades it has been known that tumour grade, the presence of necrosis and rhomboid morphology vary across a tumour [298]. Higher architectural ITH (i.e. greater number of different patterns) was associated with higher grade, larger tumour size and stage [297]. More recently, there has been a growing interest in evaluating tumour immune cell and stromal infiltration and various classification systems have been developed. Clark et al identified heterogeneity between patients, with four major proteomic subtypes: CD8+ inflamed (i.e. high CD8+ infiltration), CD8- inflamed (i.e. high macrophages and fibroblasts), VEGF immune desert (i.e. high stromal and endothelial cells, angiogenesis and platelet degranulation) and metabolic immune desert (i.e. increased metabolic activation but reduced immune and stromal components). Alternatively, Şenbabaoğlu et al identified three distinct subgroups: T cell enriched, heterogeneously infiltrated, and non-infiltrated [299]. Single cell analysis has confirmed heterogeneity of immune cell infiltrates not only between patients, but also within some multi-region samples from the same patient [300]. T cells and tumour associated macrophages represent potential prognostic markers, as well being potential druggable targets [301]. Taken together, these studies suggest that there are varying degrees of ITH in different patients, ITH exists across different molecular levels, and different methods exist to classify ITH, with the potential for some of these to have prognostic value.

The high degree of genetic and morphological heterogeneity has several implications in both a research and clinical setting. Clinically, assessing only a small number of multi-region samples may miss the presence of negative prognostic markers and lead to under-estimating the aggressiveness of the tumour [87]. Studies evaluating prognostic biomarkers in ccRCC have failed to identify any clinically relevant markers that are externally validated. Heterogeneity between patients may contribute to this, as can heterogeneity within a patient (i.e. intra-tumoral heterogeneity) [68]. Indeed, sampling only one tumour region may hamper detection and validation of biomarkers as the marker of interest may be expressed in some but not all regions. For example, both *BAP1* mutations and chromosome 9p loss are predictors of poor prognosis in ccRCC, and have been shown to be subclonal [82]. Furthermore, analysing a limited number of multi-region samples may underestimate the prevalence of driver mutations and could limit our understanding of tumorigenesis. Conversely, this can also lead to an 'illusion of clonality', where a mutation may appear to be clonal if the analysis is limited to a small number of samples, whereas sampling a larger number of regions would reveal it to be subclonal [87]. A better understanding of tumour evolution may offer insights into renal carcinogenesis. Furthermore, clonal events may represent better diagnostic biomarkers (as they are more likely to be present in all regions sampled) and more appropriate drug targets (as

these are early events in cancer development). An analysis of subclonal aberrations may elucidate which changes may offer a selective advantage and lead to disease progression or drug resistance.

Due to the relatively low number of genetic mutations observed in ccRCC, there is a growing interest in DNA methylation as this represents an early event in tumorigenesis, with highly recurrent sets of changes [103]. However, only a very small number of studies have been performed assessing methylation ITH to date, and were limited by relatively small sample sizes and absence of sequence level methylation data [127, 135, 136]. These limited studies (described more in detail in the Background, Chapter 2) suggest a degree of relative methylation homogeneity, meaning that methylation markers could represent ideal diagnostic and prognostic targets. Therefore, in this chapter I aim to comprehensively characterise DNA methylation heterogeneity in tissue from ccRCC patients. A systematic approach, evaluating heterogeneity between patients, within a patient and within a sample, was adopted, as previously described (Figure 6.1) [207].

6.2 Chapter aims

1. Characterise DNA methylation heterogeneity by analysing multi-region tissue samples from patients with ccRCC, including evaluating:
    a. Heterogeneity between different patients
    b. Heterogeneity within a patient (i.e. multi-region samples from the same patient)
    c. Heterogeneity within a sample (i.e. epipolymorphism)
2. Compare phylo-epigenetic trees built using methylation data against phylogenetic trees built using SCNA data, for patients with ccRCC
3. Evaluate the association between methylation, epipolymorphism and gene expression
4. Use multi-region ccRCC samples to glean information regarding biomarker selection, including an evaluation of:
    a. Heterogeneity in significant differentially methylated cytosines (DMCs)
    b. Cell type decomposition
    c. Comparison of homogeneous versus heterogeneously methylated CpGs

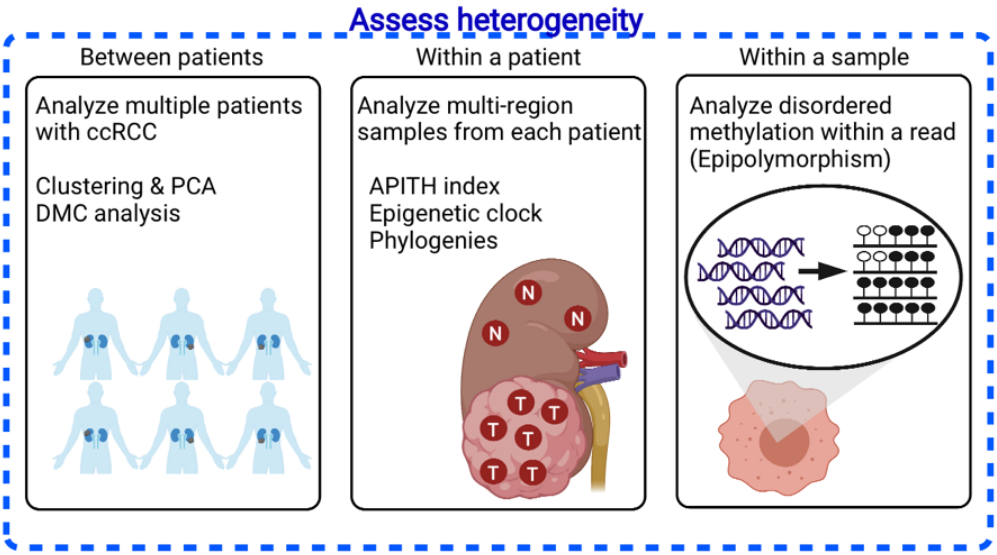


Figure 6.1: Schematic of analysis performed

Heterogeneity was evaluated between patients, within a patient and within a sample. Abbreviations: APITH (average pairwise intra-tumoral heterogeneity index), DMC (differentially methylated cytosines), PCA (principle component analysis).

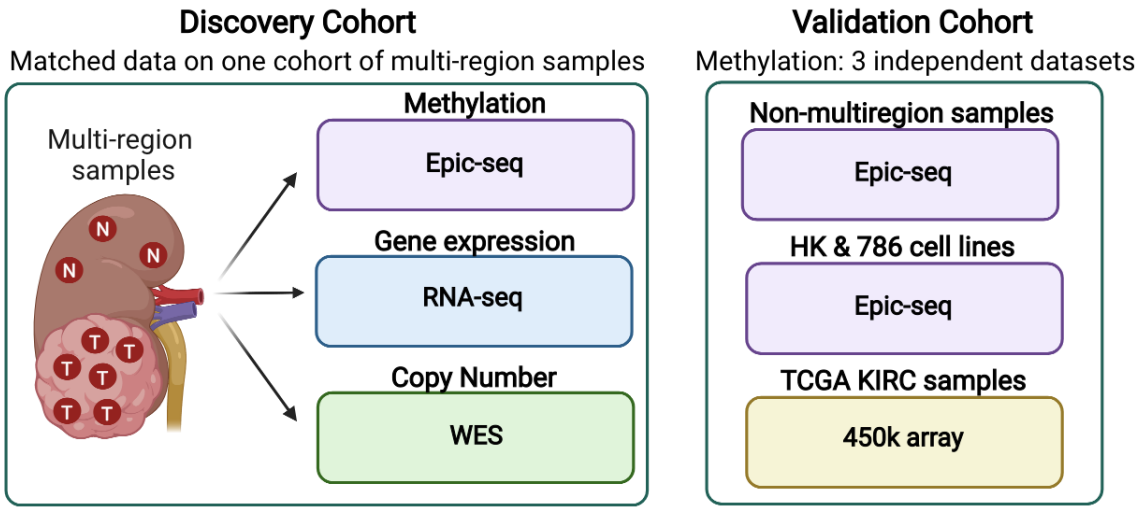


Figure 6.2: Summary of samples analysed

The discovery cohort consists of multi-region samples from patients with ccRCC, with matched Epic-seq, RNA-seq and WES (whole exome sequencing). In addition, methylation data were generated for a cohort of tumour and normal samples from patients with ccRCC (i.e. not multi-region samples) and cell lines. TCGA data were also obtained.

6.3 Results

The chapter can broadly be subdivided into two major sections. The first section focuses on evaluating methylation heterogeneity between different patients, within a patient (i.e. between tumour samples from the same patient) and within a sample (i.e. methylation pattern of individual reads) (Figure 6.1; sections 6.3.2 - 6.3.4). In the second part of the analysis, data from multi-region ccRCC samples was used to glean information regarding biomarker selection (section 6.3.5). This includes an analysis of heterogeneity in significant DMCs, cell type decomposition based on DNA methylation and a comparison of homogeneous versus heterogeneously methylated CpGs.

6.3.1.1 Sample overview

In brief, the exploratory methylation analysis was performed on a group of multi-region samples (fresh frozen tissue from nephrectomy specimens) obtained from patients with ccRCC. Subsequently, the results were validated on three methylation datasets: cell line samples, an independent cohort of ccRCC patients and the TCGA ccRCC (i.e. KIRC) samples. All data, except the publicly available TCGA dataset, were generated in house. A summary is provided in Figure 6.2 and the datasets are described more in detail below.

For the discovery cohort, I obtained matched methylation, gene expression and copy number data on multi-region samples derived from ccRCC patients (Figure 6.2). Multi-region samples were included in the analysis if at least 4 tissue specimens were available from the same patient (i.e. at least one normal and 3 tumour samples). Epic-seq methylation data were available for 136 multi-region samples from 18 patients (40 normal kidney and 96 ccRCC samples) (Table 6.1). For the analysis of methylation within-sample heterogeneity, one sample (5998.T4) was removed due to low coverage, therefore 135 samples were available in total. For a subset of these samples, matched whole exome sequencing (WES; 68 samples from 18 patients) and RNA-seq (47 samples from 7 patients) were generated (Table 6.1).

The results were validated using three independent methylation datasets (Figure 6.2). I generated Epic-seq methylation data for an independent cohort of 71 non-multiregion samples from ccRCC patients (30 normal kidney and 41 ccRCC samples). In this cohort, there were more tumour than normal samples as some patients did not have matched adjacent normal kidney tissue available. In addition, I generated Epic-seq for two cell lines: HK2 and 786-O cell lines, in vitro models for normal kidney proximal tubule epithelium and ccRCC respectively. Publicly available 450k array data were

obtained for the TCGA KIRC dataset (160 normal kidney and 325 ccRCC samples). Details regarding experimental methods and data analysis are found in the Methods (Chapter 4, section 4.7).

Table 6.1: Demographic and sample data for each ccRCC patient

Data are shown for 18 ccRCC patients. The number of normal (N) and tumour (T) multi-region samples are shown for each patient for methylation (Epic-seq), RNA-seq and WES. Missing samples are left blank (for example, patient 5532 had methylation data but no RNA-seq data). Clinical data are shown, including recurrence status (i.e. recurrence, no recurrence and metastases at diagnosis). One patient was lost to follow up (patient 7067). All patients had ≥4 years of follow-up.

	Characteristics at diagnosis						Follow up	Number of available samples					
Patient ID	Age (years)	T Stage	M stage	Size (cm)	Grade	Leibovich score	Recurrence	Epic-seq		RNA-seq		WES	
								N	T	N	T	N	T
5532	61	pT3a	M1	6	2	4	Mets at Dx	1	3			1	3
5644	61	pT2a	M0	7.5	3	5	No	1	7			1	4
5790	51	pT3a	M0	6.7	3	6	No	2	3			1	2
5799	53	pT1b	M0	6.5	2	2	No	3	7	2	1	1	4
5801	73	pT3a	M0	10.8	4	9	Recurrence	2	4				
5802	74	pT3a	M0	2.8	3	6	No	1	3			1	2
5813	42	pT3a	M0	8.7	2	4	Recurrence	3	7			1	7
5818	63	pT3a	M1	7.4	3	6	Mets at Dx	1	3			1	1
5826	76	pT1b	M0	6.1	3	4	No	1	3			1	2
5842	67	pT3a	M0	13.5	4	9	Recurrence	2	9			1	8
5848	65	pT1b	M0	5.4	3	3	Recurrence	3	3			1	3
5998	77	pT1b	M0	5.2	2	2	No	3	6	2	3	1	3
6262	65	pT3a	M0	10.2	4	8	Recurrence	4	6	4	3	1	5
6285	67	pT3a	M0	7.5	4	8	Recurrence	2	5			1	5
6300	62	pT3b	M0	13	4	9	Recurrence	4	11	3	9	1	8
7067	56	pT3a	M0	8	3	5	Unknown	3	5	3	6	1	2
7068	62	pT3a	M0	12	4	9	Recurrence	2	5	2	4	1	3
7281	65	pT3b	M0	8	4	8	Recurrence	2	6	2	3	1	3

Abbreviations: M = metastases, Mets at Dx= metastases at diagnosis, N= normal, RNA-seq= RNA sequencing, T= tumour, WES = whole exome sequencing

6.3.1.2 Sample purity

Prior to commencing the main analysis of ITH, I sought to explore sample purity as this may impact an evaluation of heterogeneity. Purity was evaluated in my samples using three orthogonal methods: DNA methylation analysis using 'InfiniumPurify', RNA-seq using 'ESTIMATE' and WES using 'ASCAT' (Figure 6.3A). The purity estimates derived from RNA-seq (including estimates of stromal and immune components) in my dataset were very similar to purity estimates in TCGA, suggesting my sample quality was within the norm of expected results (Figure 6.3B). Importantly, 'InfiniumPurify' determines tumour purity in the context of contamination with normal kidney (non-tumour) sample. This is in contrast with 'ESTIMATE', the RNA-seq based method, which determines tumour purity as a function of admixtures of immune and stromal cell components based on gene expression data. This difference explains why the correlation between RNA-seq and WES purity estimates was high (Pearson correlation coefficient = 0.81, p value = 1.6e-05), but the correlation with methylation based estimates was not significant (p value >0.05), since the methods measure contamination with different components (Figure 6.3C). For all subsequent analyses in this chapter, when referring to sample purity, estimates derived from WES (or RNA-seq if not available) are used. Given the high correlation between WES and RNA-seq based estimates, and the fact that WES and RNA-seq purity estimates were only available for a subset of samples, this method was deemed appropriate to increase the number of samples which had purity estimates.

6.3.2 Heterogeneity between patients

In order to evaluate heterogeneity between patients, I performed principal component analysis (PCA) using data from all available CpGs (N= ~1.1million CpGs) for the multi-region samples from 18 ccRCC patients. As expected and as shown in Chapter 5, samples cluster by pathological subtype, with normal kidney samples clustering very closely together (i.e. less variability in normal than tumour specimens) (Figure 6.4A). ccRCC samples are more heterogeneous, with some patient samples clustering more closely together (e.g. patient 5842) and some quite far away (e.g. 7068) (Figure 6.4B). The topmost variable CpGs were selected (after removal of all SNPs) for visualisation in a heatmap (i.e. CpGs with the highest variance in tumour samples). Selecting the top 10,000 CpGs with the highest variance in tumour demonstrates that samples cluster by patient rather than by subtype (i.e. normal and tumour do not cluster separately; Figure 6.4C). Selecting the top 50,000 CpGs with the highest variance in tumour demonstrate that broadly normal kidney samples cluster together, whilst tumours cluster by patient (Figure 6.4D). However, for patient 7068, all tumour samples do not cluster together, in keeping with the observation that this patient appears to have

the highest ITH on the PCA (Figure 6.4B). There is also evidence of heterogeneity between patients, with two major branches of the dendrogram. Overall, this suggests that in the majority of cases there is more variability between patients than within patients, and this is consistent despite using different cut-offs (top 10,000 and top 50,000 CpGs). This finding is in keeping with the limited available evidence on DNA methylation in multi-region ccRCC samples, which also points towards relative homogeneity within a patient [127, 135, 136].

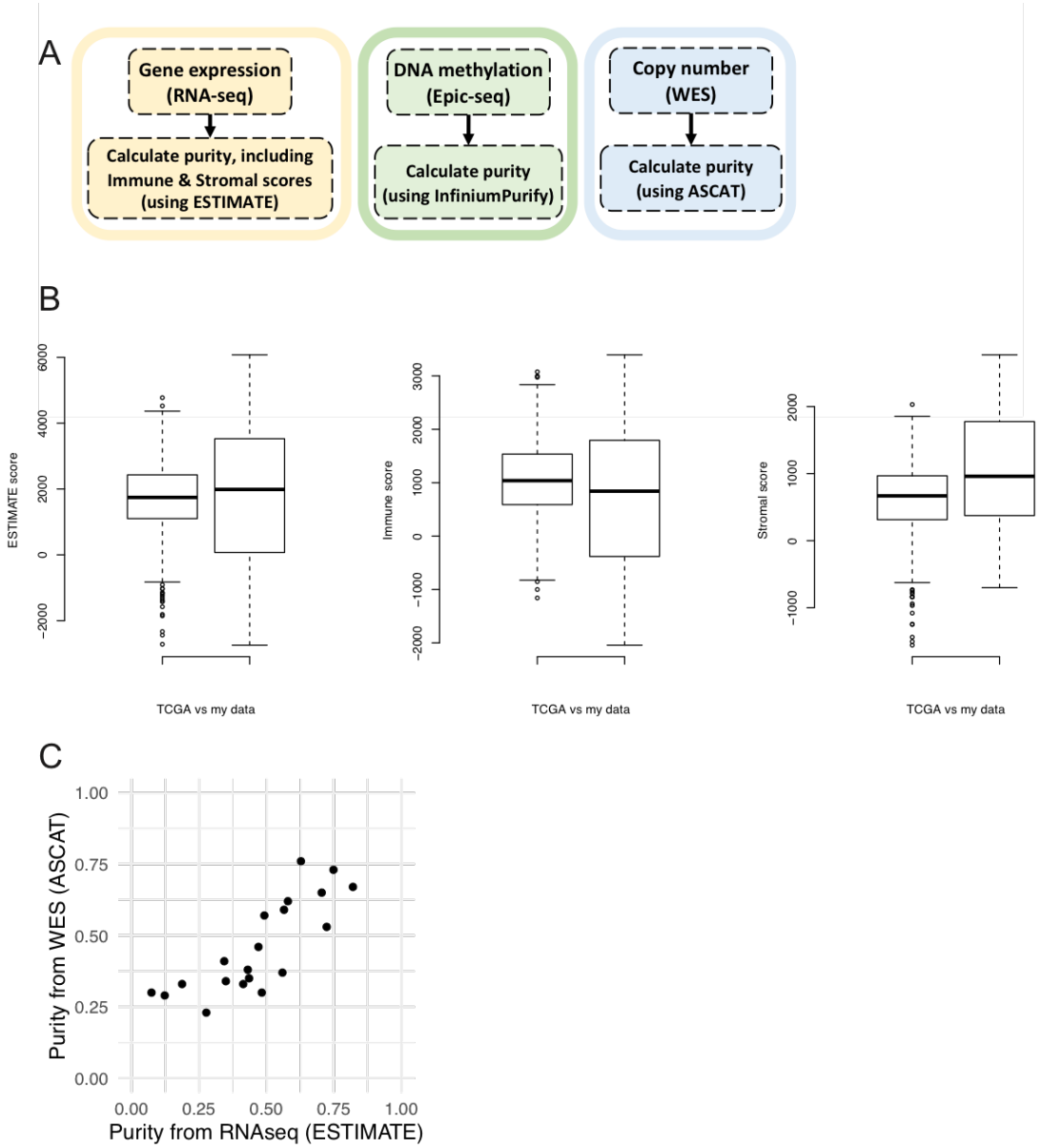


Figure 6.3: Purity assessment in multi-region samples

Panel A- Schematic summarising three orthogonal methods used to derive purity estimates: DNA methylation analysis using 'InfiniumPurify', RNA-seq using 'ESTIMATE' and whole exome sequencing (WES) using 'ASCAT'. Panel B- ESTIMATE scores, immune scores and stromal scores generated using RNA-seq data and the 'ESTIMATE' package for TCGA and my data. Panel C- Scatterplot of matched purity values for my samples obtained using RNA-seq and WES. Pearson correlation coefficient = 0.81.

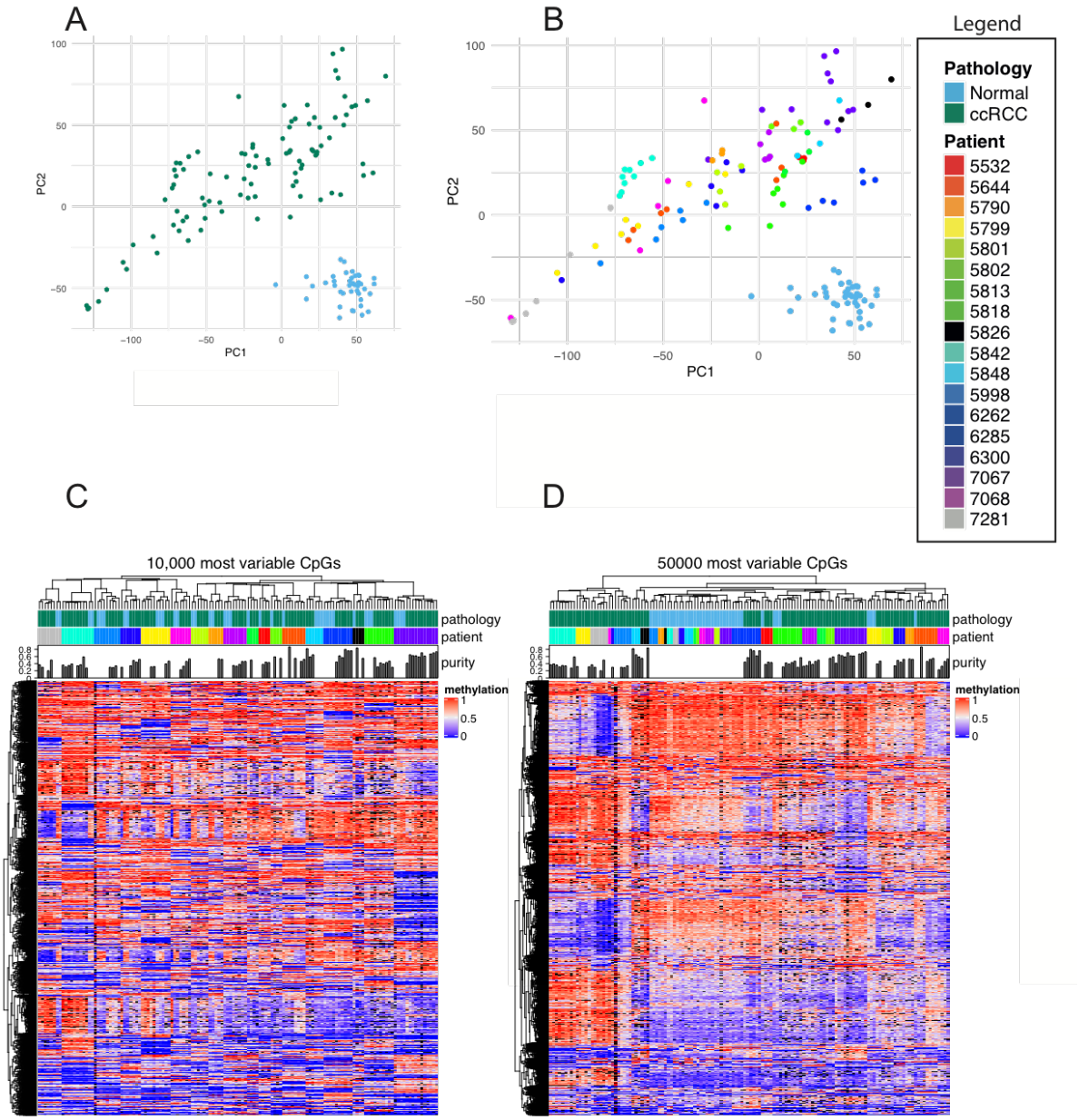


Figure 6.4: Methylation heterogeneity between patients

Panel A- Principal component analysis (PCA) demonstrating normal kidney and ccRCC samples. Panel B- PCA demonstrating normal kidney and ccRCC samples, highlighting multi-region tumour samples from the same patient. Panel C and D- Heatmaps visualising methylation values in the top 10,000 CpGs (Panel C) and top 50,000 CpGs (Panel D) with the highest variance in tumour samples. The top annotation demonstrates which multi-region tumour samples are derived from the same patient, along with tumour purity (estimated using WES or RNA-seq).

6.3.3 Heterogeneity within a patient

The following section explores heterogeneity within a patient (i.e. between samples derived from the same patient), by evaluating epigenetic age, the average pairwise ITH index, phylogenetic trees derived from methylation and SCNA, and prognostic risk scores.

6.3.3.1 DNA methylation accelerated ageing

I evaluated the DNA methylation age in multi-region samples derived from the same patient, to explore epigenetic ageing and heterogeneity. I calculated the predicted methylation age of each sample using Horvath's epigenetic clock [102]. In brief, the clock uses 353 CpGs and was developed using the 21k Illumina methylation array. In order to demonstrate reliability of methods, first I evaluated methylation age for TCGA ccRCC tumour (N=325) and normal kidney (N=160) tissue samples. As expected, a very good correlation was observed between chronological age and predicted DNA methylation age in normal kidney (correlation coefficient = 0.86; Figure 6.5A). Significantly higher methylation age was noted in tumour versus normal kidney TCGA samples (median methylation age 55 vs 48 years, p value <0.01) and this is in keeping with the existing literature (Figure 6.5B). For my Epic-seq data, in normal kidney samples the correlation between chronological age and predicted DNA methylation age was 0.42 (Figure 6.5A). The correlation increased to 0.57 if one outlier sample was removed (5998.N1a; chronological age 77 years, predicted age 44 years). This moderate correlation is likely due to the use of the Epic-seq platform (rather than 21k array) which leads to missing data for CpGs. In fact, 59 out of 353 CpGs were missing for all (100%) ccRCC samples, and 66 CpGs were missing in >50% of samples, and were derived by imputation (as described in section 4.7.2.1). Despite this, estimates of epigenetic age are still likely to be valuable in this cohort, and there is evidence of increased epigenetic age relative to chronological age in ccRCC samples compared to normal kidney samples. Indeed, epigenetic age was significantly higher in ccRCC compared to normal tissue (median methylation age 63 vs 59 years, p value <0.01; Figure 6.5B).

The predicted to chronological age ratio (PCAR) was calculated by dividing DNA methylation age by real age, where 1 represents a perfect match. In the majority of patients, the PCAR was relatively similar amongst tumour samples derived from the same patient (Figure 6.7A), except for patients 5532, 6300, 7067 and 7068 who had more variable estimates (maximum difference in PCAR amongst tumour samples was >30%). There were no obvious clinical parameters which might explain this observation. There was no association between PCAR and any clinical/prognostic parameters

including stage at diagnosis, tumour size and recurrence status (Wilcoxon test p value >0.05). Accelerated epigenetic ageing has been found to be associated with increased cancer risk, as well as prognosis in ccRCC patients [127], chRCC [302] and other cancers. There was no significant difference in PCAR in tumour samples from ccRCC patients from TCGA who were alive versus dead at follow up (Wilcoxon test p value >0.05). In summary, epigenetic age was consistently higher in tumour compared to normal kidney samples. I noted variation in DNA methylation age amongst multi-region tumour samples, which was more pronounced in some patients than others, however there were no obvious clinical correlates.

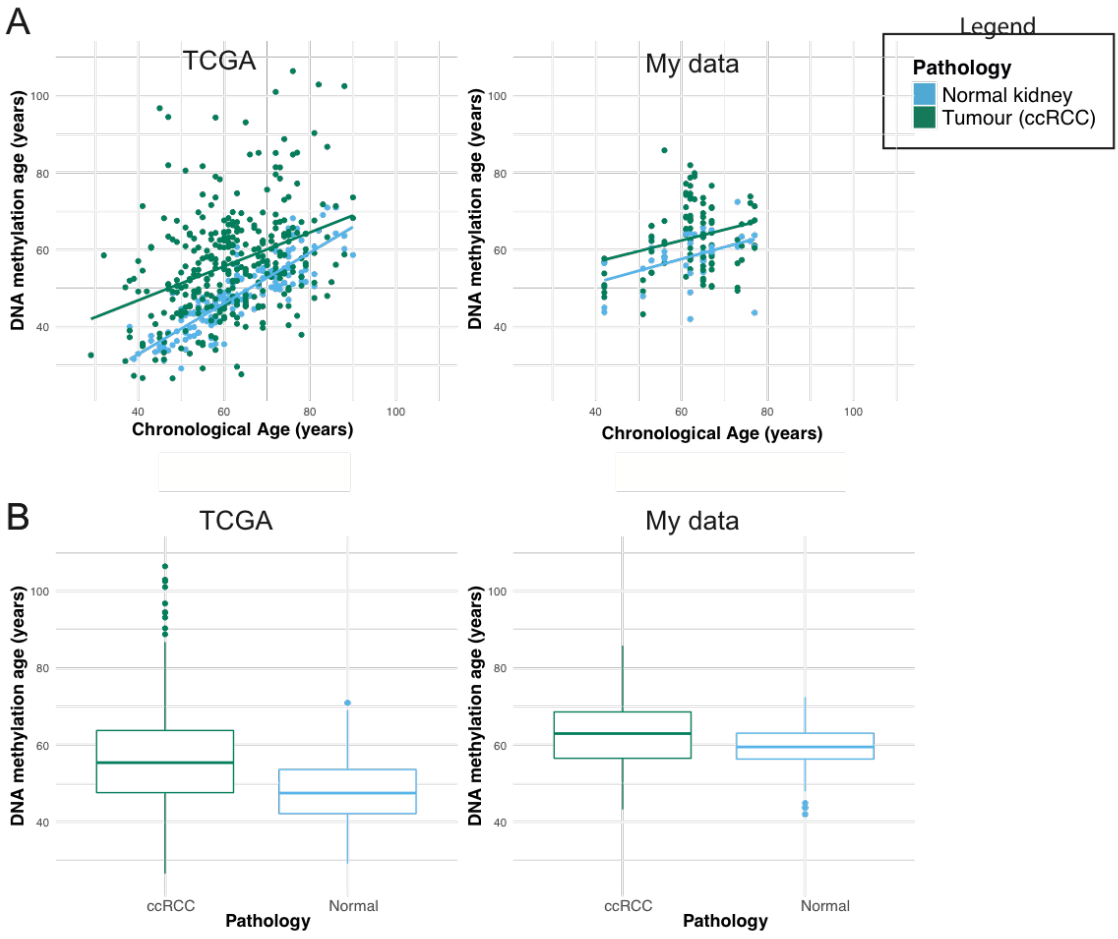


Figure 6.5: DNA methylation age versus chronological age in ccRCC and normal kidney samples

Panel A- DNA methylation age versus chronological age (in years) in TCGA data and my data. There is an obvious correlation between DNA methylation age and chronological age, which is more pronounced in normal tissue than ccRCC. Panel B- DNA methylation age is significantly higher in ccRCC compared to normal kidney samples in TCGA and my dataset (p value <0.01).

6.3.3.2 Methylation Average Pairwise ITH (APITH) index

Aiming to formally evaluate ITH in my samples, I identified the Average Pairwise ITH (APITH) index as a validated metric to quantify intra-tumoral heterogeneity independently of the number of tumour samples evaluated [209]. The APITH index may be derived using either methylation or copy number data [209]. I calculated the methylation APITH index as described by Hua et al (see section 4.7.2.2), for all CpGs as well as the top 5000 most variable CpGs [209]. In brief, first I calculated the pairwise Euclidean distance between tumour samples (for all CpGs and the top 5000 most variable CpGs), then obtained the average. In keeping with the results of the PCA (Figure 6.4B), patient 5842 had the lowest methylation APITH score (and therefore lowest ITH) and patient 7068 had the highest score (i.e. highest ITH). For both analyses (all CpGs and the top 5000 most variable CpGs), the methylation APITH index was not associated with clinical or prognostic factors. There was no significant difference in methylation APITH score by tumour stage (stage I-II vs III-IV), grade, Leibovich score (low vs intermediate/high) and recurrence status (no recurrence vs recurrence) (Wilcoxon test p value >0.05 for all comparisons; Figure 6.6A). There was also no correlation with tumour size (correlation coefficient=-0.07, p value >0.05). This suggests that in the present cohort, methylation ITH as measured by the APITH score, is not associated with clinical or prognostic factors. Interestingly, in the literature, methylation APITH scores were associated with overall survival and risk of distant metastases in lung cancer [209], but no association with prognostic factors was noted in papillary RCC [303].

Next, I sought to assess if the methylation APITH index may be confounded by differing tumour purity in multi-region samples from the same patient (i.e. high variance in tumour purity may lead to artificially high ITH and therefore misleading high methylation APITH index). Tumour purity was evaluated using WES or RNA-seq as described in section 6.3.1.2, and data were available for 72 samples from 16 patients. There was no obvious correlation between the methylation APITH index and the variance of the purity of tumour samples from the same patient (correlation= 0.03, p value >0.05, Figure 6.6B). Subsequently, I compared the APITH index derived from methylation data which were unadjusted versus adjusted for tumour purity, to assess whether this produced similar results. The analysis focused on patients where purity data were available for at least 3 tumour samples (N= 66 tumour samples from 13 patients). The percentage methylation at each CpG was adjusted for the purity of each sample, as previously described [199], and the APITH index was re-assessed. The results demonstrated that APITH index was relatively similar even after adjustment for tumour purity (Figure 6.6C); patient 7068 still had the highest degree of ITH and patient 5842 still had low levels of ITH. The APITH index is based on a large number of CpGs and therefore it appears to be

relatively robust to differences in sample purity. There was no significant difference between methylation APITH scores adjusted for tumour purity and any clinico-pathological factors (p value >0.05 for all comparisons; Figure 6.6C). In summary, this section suggests that the APITH index is a robust measure of ITH which is not affected by differing sample purity in multi-region samples from the same patient.

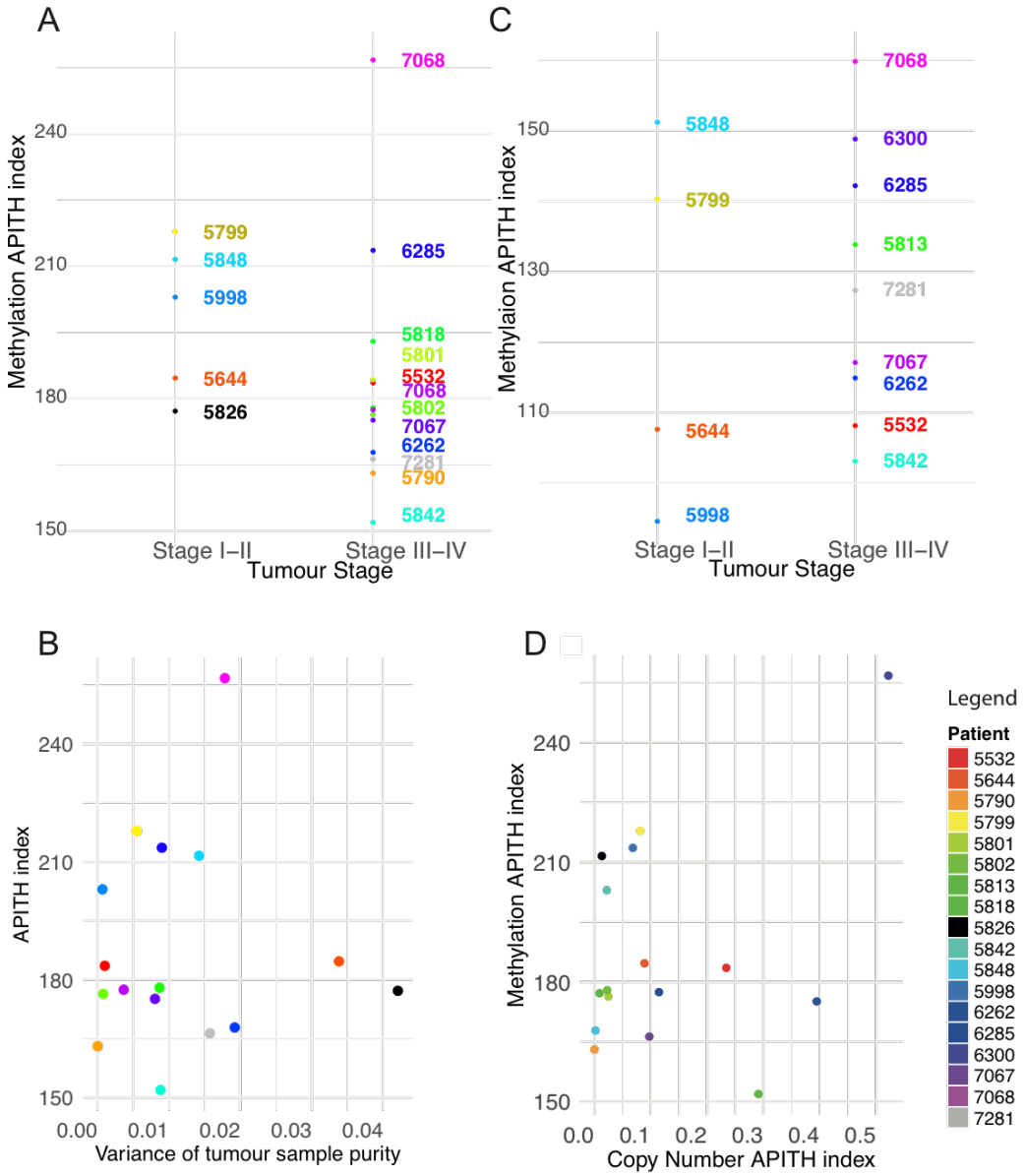


Figure 6.6: Methylation Average Pairwise ITH (APITH) index

Panel A- Methylation APITH index for each patient, by tumour stage. There was no significant difference in methylation APITH in patients with stage I-II versus III-IV disease (p value >0.05). Panel B- Scatterplot of Methylation APITH index and the variance of the sample purity for multi-region tumour samples derived from the same patient. Panel C- Methylation APITH index for each patient, derived from methylation data which are adjusted for tumour purity. Once again, there was no significant difference in methylation APITH in patients with stage I-II versus III-IV disease (p value >0.05). Panel D- Scatterplot of APITH derived using methylation versus copy number data.

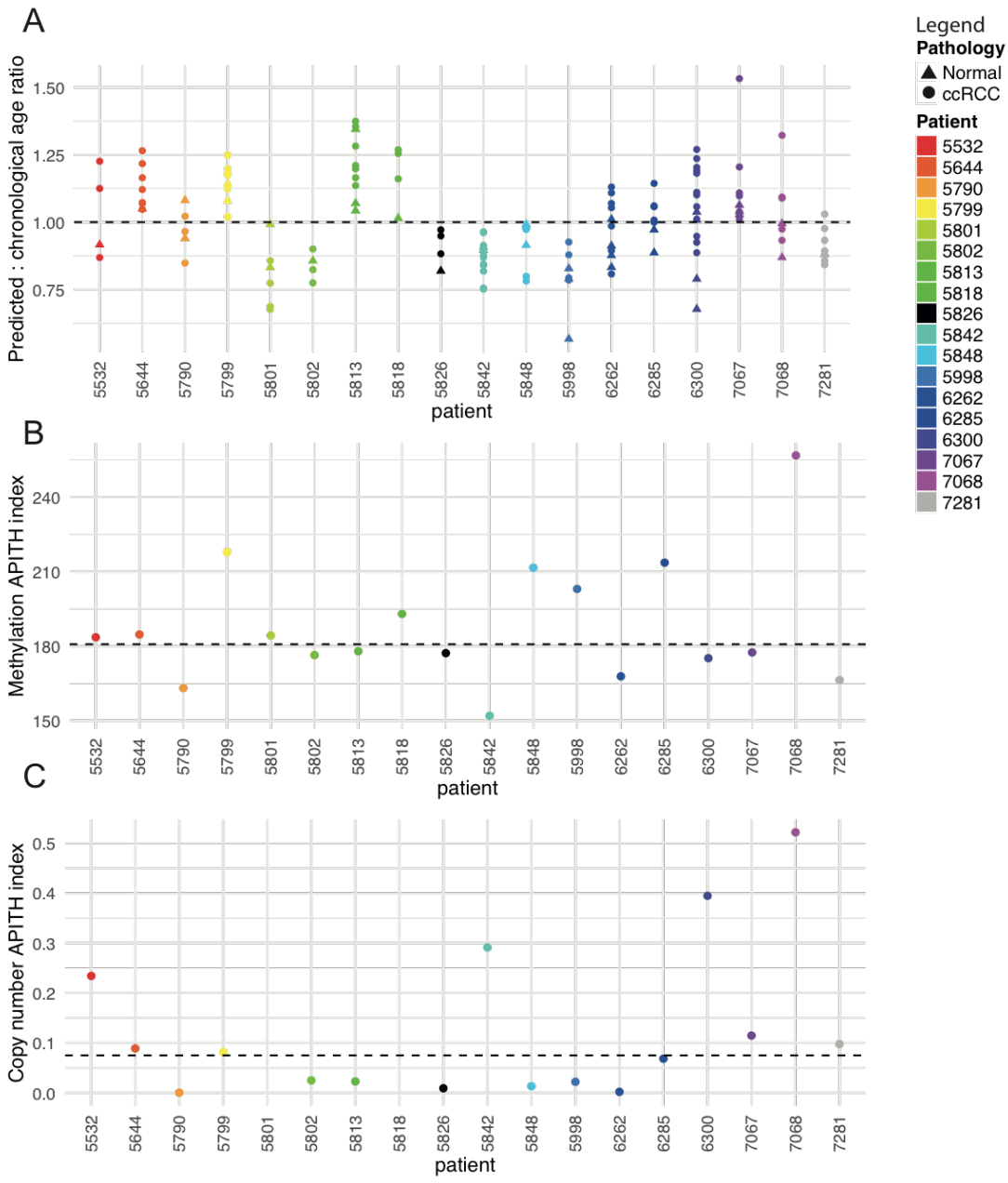


Figure 6.7: DNA methylation age and Average Pairwise ITH (APITH) index, by patient

Panel A- The predicted to chronological age ratio (PCAR) is shown for each sample, by patient. Tumour samples tend to have a higher PCAR than normal kidney. The dotted horizonal line represents a PCAR of 1, i.e. the DNA methylation and chronological age are the same. Panel B and C- APITH index derived from DNA methylation data (Panel B) and somatic copy number data (Panel C), by patient. The dotted horizonal line represents the median APITH.

6.3.3.3 Comparing APITH based on methylation and copy number data

In order to directly compare the degree of methylation heterogeneity against the degree of somatic copy number aberration (SCNA) heterogeneity, I compared the APITH index derived using methylation and copy number data respectively (Figure 6.7B and C). My collaborator, Dr Schwarz, generated the copy number APITH index by evaluating the percentage of the genome which is affected by private SCNA (using the 'ASCAT' package) in each sample and the average pairwise distance between samples, as previously described [209]. The SCNA APITH ranged between 0.0004 and 0.52, with a mean and median of 0.12 and 0.075 respectively (standard deviation 0.16). Interestingly, these values are relatively similar to SCNA APITH indices noted for lung adenocarcinoma multi-region samples in the literature: range 0-0.68, mean and median 0.18 and 0.16 respectively [209]. In my dataset, patient 7068 had the highest APITH derived by both SCNA and methylation, and was an obvious outlier. WES demonstrated subclonal whole genome doubling, therefore the entire genome has SCNA and the APITH was very high. There was a weak correlation between the methylation APITH and the copy number APITH in my dataset, although this did not achieve statistical significance (Pearson correlation coefficient = 0.33, p value = 0.216, Figure 6.6D). However, generally patients whose methylation APITH was above the median, also tended to have SCNA APITH above the median (Figure 6.7B and C). Further analysis of methylation versus SCNA heterogeneity is performed in section 6.3.3.4 (phylogenetic analysis). There was no significant difference in SCNA APITH score by tumour stage (stage I-II vs III-IV), grade, Leibovich score (low vs intermediate/high) and recurrence status (no recurrence vs recurrence) (Wilcoxon test p value >0.05 for all comparisons). There was also no association between APITH scores and predicted to chronological age ratio (Figure 6.7A).

Given no obvious associations were found between methylation and SCNA ITH and clinical nor prognostic parameters, I sought to evaluate whether the patients' recurrence status could be recapitulated using existing prognostic scores. Two prognostic scores were therefore evaluated: the Leibovich score, which is routinely used in clinical practice, and ClearCode34, the molecular score which has been most extensively validated (though not used clinically). The Leibovich score for each patient, along with tumour stage and recurrence status is summarised in Table 6.1 and Figure 6.8A. Tumour stage and Leibovich score were able to predict recurrence in some patients but not others, highlighting the need for better prognostic risk scores. For example, one patient with stage I disease developed a recurrence, whereas one stage II and two stage III patients remained recurrence free. Patients with a low Leibovich score (0-2) are generally considered low risk, whereas patients with intermediate (3-5) and high (≥6) scores are considered at higher risk of recurrence. In my cohort,

100% of patients (N=2/2) with a low score remained recurrence free, whilst 80% (N=8/10) of patients with a high score either had metastases at diagnosis or developed a recurrence post-operatively. 100% of patients with a Leibovich score ≥8 had a recurrence. In individuals with an intermediate score, 40% (2/5) did not develop a recurrence, highlighting that it is most difficult to accurately predict outcomes in intermediate risk patients.

ClearCode34 is a prognostic score which consists of a 34 gene classifier and has been externally validated in several studies [92, 93]. I therefore evaluated ClearCode34 in my multi-region ccRCC tissue samples using RNA-seq data (Figure 6.8B). Although samples from the same individual tend to cluster together in most cases, there are several multi-region samples which do not cluster by patient (for example samples from 5998, 7067 and 7068). This is in keeping with the known available literature which suggests that there is a degree of genetic ITH and that multi-region samples from the same patient can have different expression patterns for ClearCode34 genes [68]. The heatmap demonstrates that there are two main branches in the dendrogram, where samples from patients who did not have a recurrence tend to cluster on the right (i.e. patient 5799 and 5998). Samples from patient 7281 also cluster on the right. Review of clinical notes revealed that this individual received neo-adjuvant axitinib for 8 weeks pre-operatively as part of the NAXIVA trial. Studies suggest that treatment with neo-adjuvant tyrosine kinase inhibitors may lead to changes in gene expression, as well as significant increases in *VHL* promoter hypermethylation on post-operative nephrectomy tissue specimens [135, 304]. It may therefore be plausible that patient 7281 may cluster with patients who have not had a recurrence due to changes in gene expression following treatment. In summary, although the Leibovich score is able to predict recurrence in low and high-risk patients in my dataset, this was not possible for intermediate risk patients. Expression of ClearCode34 genes was not able to accurately distinguish between patients with and without a recurrence, though the analysis was limited by the small sample size. It is therefore not surprising that there was no obvious association between clinical parameters and methylation and SCNA ITH in my dataset.

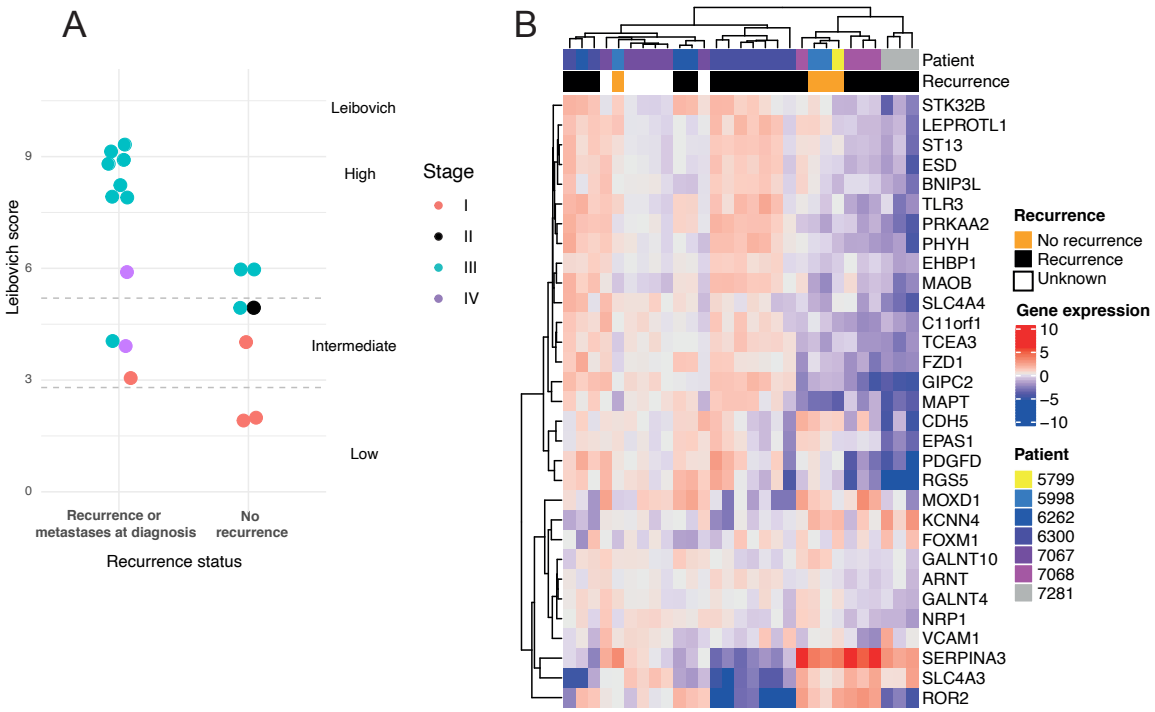


Figure 6.8: Prognostic scores in ccRCC

Panel A- Leibovich score for patients that did not have a recurrence versus patients who either had metastases at diagnosis or developed recurrence on follow up. Each dot represents a patient, where the colour of the dot shows tumour stage (stage I-IV). Individual Leibovich scores can be aggregated into three groups: low (0-2), intermediate (3-5) and high (≥6). Panel B- Gene expression is visualised for genes which contribute to the Clearcode34 prognostic risk score. Expression was derived from RNA-seq and is median centred and presented on a log2 scale. Multi-region samples from the same patient are shown, along with recurrence status (i.e. no recurrence, recurrence or unknown).

6.3.3.4 Comparing phylogenies based on methylation and copy number data

For each patient, phylogenies were inferred using methylation data and SCNA data respectively, and the phylo-epigenetic and phylogenetic trees were compared to assess similarities in evolutionary trajectories. Methylation phylogenies were inferred by me, whereas SCNA phylogenies were reconstructed by my collaborator Victoria Dumbrowe (Schwarz Group, Max Delbrück Center), as follows (details in section 4.7.2.3). In brief, for each patient, I selected the top 10% of CpGs with the highest variance in tumour samples and calculated the Euclidean distance matrix. Trees were subsequently inferred using the ordinary least squares minimum evolution algorithm [211], a method which has been extensively used in the literature for DNA methylation phylogenies [212, 213]. Phylogenetic trees were created by Miss Dombrowe using SCNA data from WES using 'MEDICC2' [215]. Following reference phasing of SCNA, 'MEDICC2' calculates the pairwise minimum-event distance between samples, and these data are used to construct phylogenetic trees using the neighbour joining algorithm [217].

I calculated the Robinson-Fould measure to compare similarities between trees (where 0 indicates identical trees and 1 indicates all tree splits are unique; Table 6.2). Of note, 6 patients were excluded from this analysis as data were available on only three tumour samples, and it was therefore not possible to adequately assesses similarities between methylation and SCNA trees. Patient 7068 is of particular interest as this individual was found to have both the highest SCNA and methylation derived APITH in section 6.3.3.3. However, unfortunately this was one of the patients which had to be excluded due to insufficient samples analysed by WES. Phylo-epigenetic and phylogenetic trees were compared for each patient in turn (N=8 patients) (Figure 6.9 and Figure 6.10). In two patients (i.e. 5532 and 7067) extremely similar methylation and SCNA trees were noted. Indeed, 5532 had identical tree topology (though different branch lengths), whereas for patient 7067 the trees were very similar, with the only difference being the order in which samples T5 and T6 emerge. Patients 5842, 6300 and 6262 had some consistent similarities between epigenetic and genetic trees, although some differences were also noted, and these are annotated on Figure 6.9 and Figure 6.10. For example, in patient 5842, sample T7 was the earliest point of divergence from normal in both methylation and SCNA trees. In patient 6262, samples T1 and T1z were evolutionarily more similar to each other than to sample T3, and this is consistent with the fact that T1 and T1z are anatomically very close to each other (only a few millimetres distance between the two samples). However, patients 5644, 5813 and 6285 had very different topologies. It was not possible to assess whether tree similarity was related to clinical factors due to the small sample size (all but one patient had stage III-IV disease, all patients had a recurrence except one individual who was lost to follow up). In

summary, a number of patients were identified in which methylation and SCNA trees were similar, suggesting potential co-evolution. Conversely, other patients (for example patient 5813) demonstrated distinctly different phylogenetic and phylo-epigenetic trees, indicating that potentially methylation and SCNA changes may have evolved separately.

Table 6.2: Robinson-Fould distance

Robinson-Fould distance comparing phylogenies derived using DNA methylation and copy number data.

Patient ID	Robinson-Fould distance
5644	0.77
5813	0.86
5842	0.69
6262	0.75
6285	0.87
6300	0.59
7067	0.49
5532	0 (identical trees)

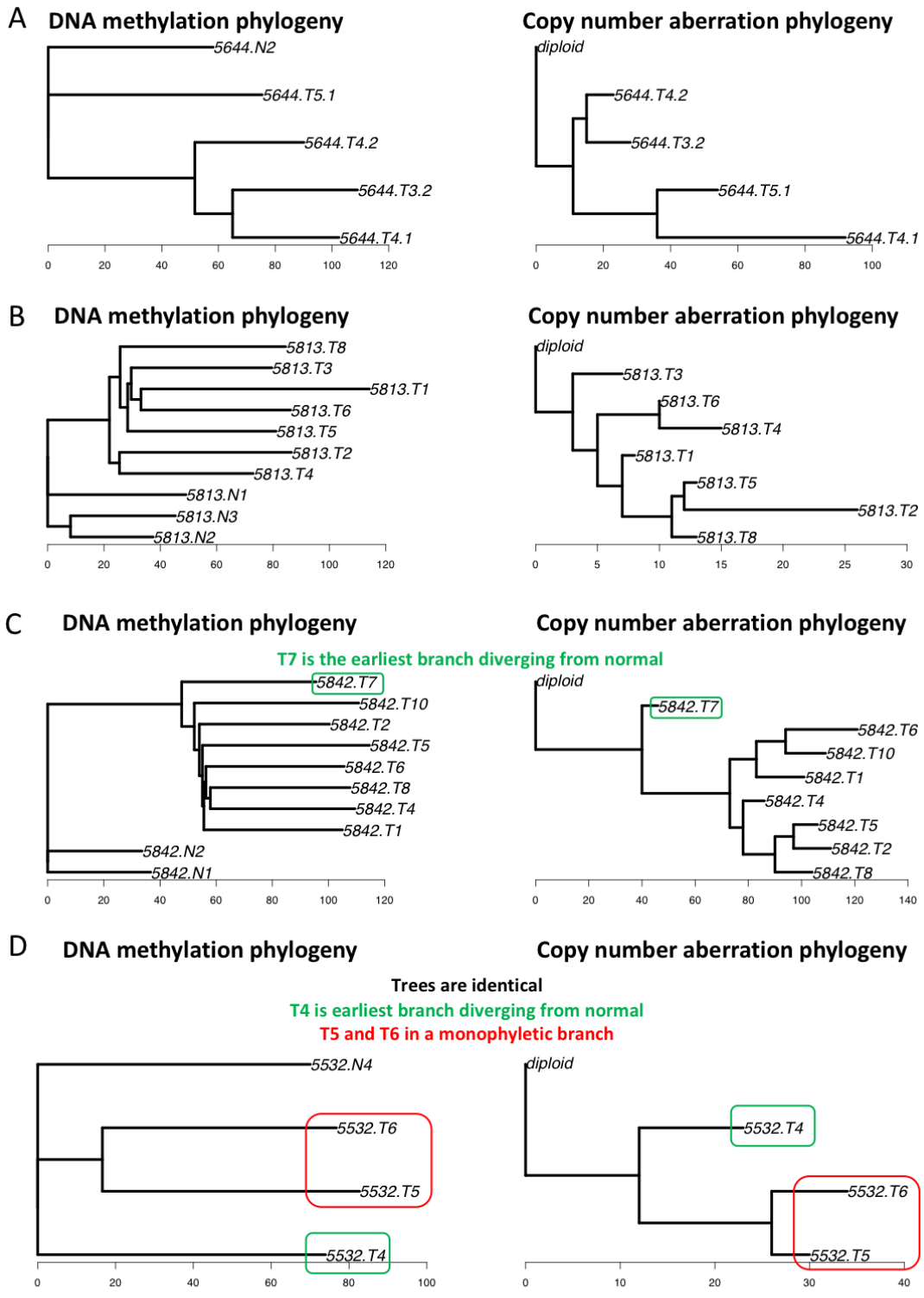


Figure 6.9: Phylogenies for patients 5644, 5813, 5842 and 5532

Phylogenies using DNA methylation and copy number data are compared for each patient, for patients 5644, 5813, 5842 and 5532. Monophyletic clades which are present in both phylo-epigenetic and phylogenetic trees are shown in red, with other similarities shown in green.

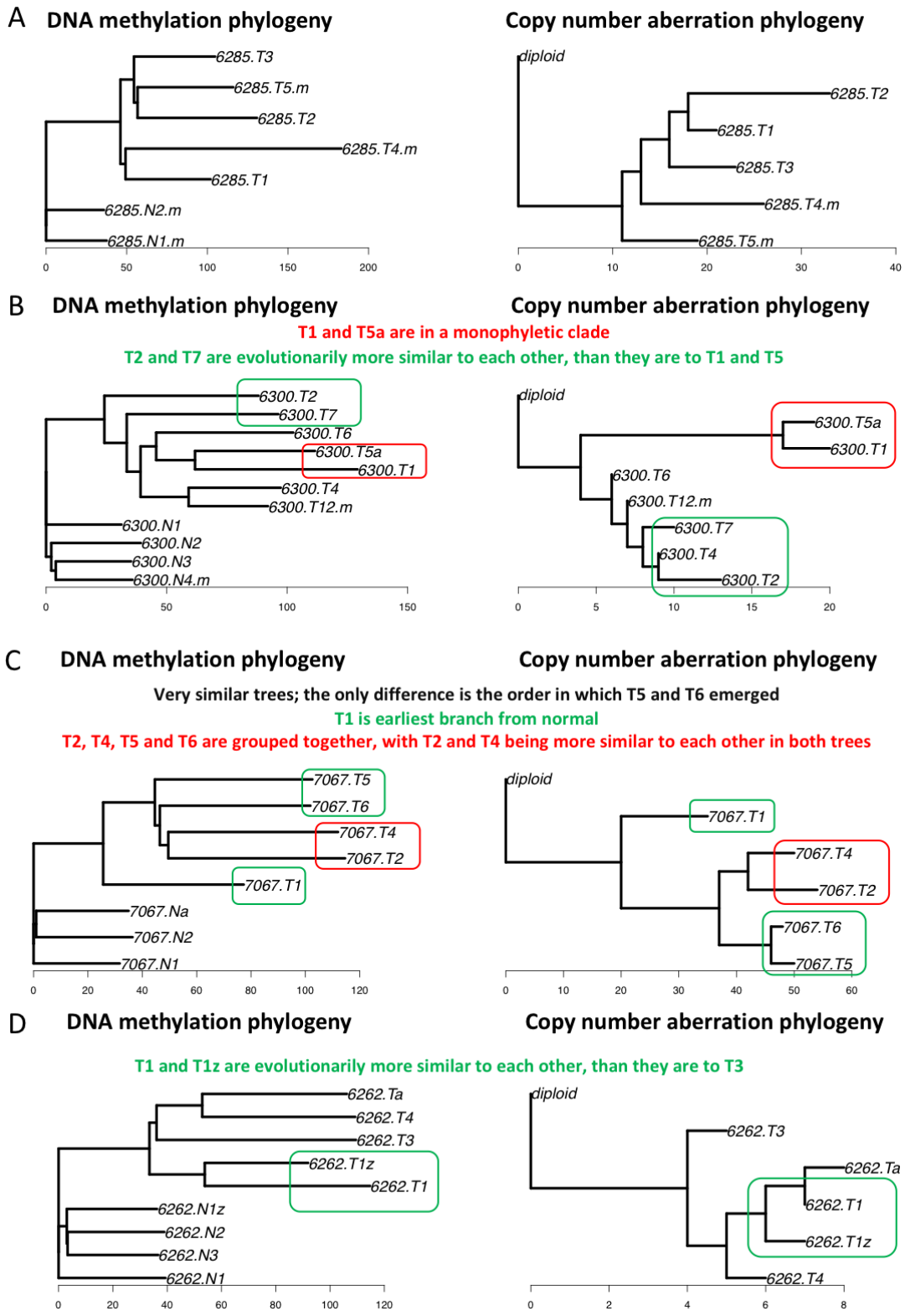


Figure 6.10: Phylogenies for patients 6285, 6300, 7067 and 6262

Phylogenies using DNA methylation and copy number data are compared for each patient, for patients 6285, 6300, 7067 and 6262. Monophyletic clades which are present in both phylo-epigenetic and phylogenetic trees are shown in red, with other similarities shown in green.

6.3.4 Heterogeneity within a sample

In order to assess methylation heterogeneity within a sample, I evaluated epipolymorphism, a concept which is described in greater depth in the Background Chapter; section 2.4.1. A single sequencing read enables us to evaluate the DNA methylation pattern derived from an individual cell (i.e. an epiallele). In brief, sequence level data were obtained for each sample and the methylation pattern was assessed at epigenetic loci (e-loci), which consist of 4 adjacent CpGs within a 150bp sequenced read window. Epipolymorphism and average methylation across each e-locus were calculated as described in section 4.7.3 (Figure 6.11). Epipolymorphism measures the diversity of epialleles within a sample. Values range between 0 (i.e., fully concordant methylation pattern in a read and all reads have the same pattern) and approaching 1 (i.e. highest degree of heterogeneity) [113] (Figure 6.11). Epipolymorphism in ccRCC tumour versus normal tissue was compared, as described by Landan et al [113] and Chen et al [219].

Disordered methylation within a read is thought to be a stochastic process, which may or may not have a functional relevance; with changes in the promoter region being more likely to be functional [110]. Both increased epipolymorphism and reduced epipolymorphism may be noted in normal relative to tumour tissue, and both these methylation changes may be associated with alterations in gene expression and tumorigenesis. It is hypothesized that disordered methylation within a read may be a precursor/intermediate state that can lead to regions of concordant differential methylation (i.e. DMRs) [110, 113]. For example, in kidney epithelial cells, an e-locus that has disordered methylation may be more likely to undergo a change in methylation across the read (to fully hyper or fully hypomethylated), compared to an e-locus with homogeneous/ordered methylation. This would be an example of higher epipolymorphism in normal relative to tumour tissue. Conversely, a kidney epithelial cell may have ordered methylation within a read, and stochastic gains in disordered methylation may be acquired leading to tumorigenesis (thus epipolymorphism would be higher in the tumour relative to normal).

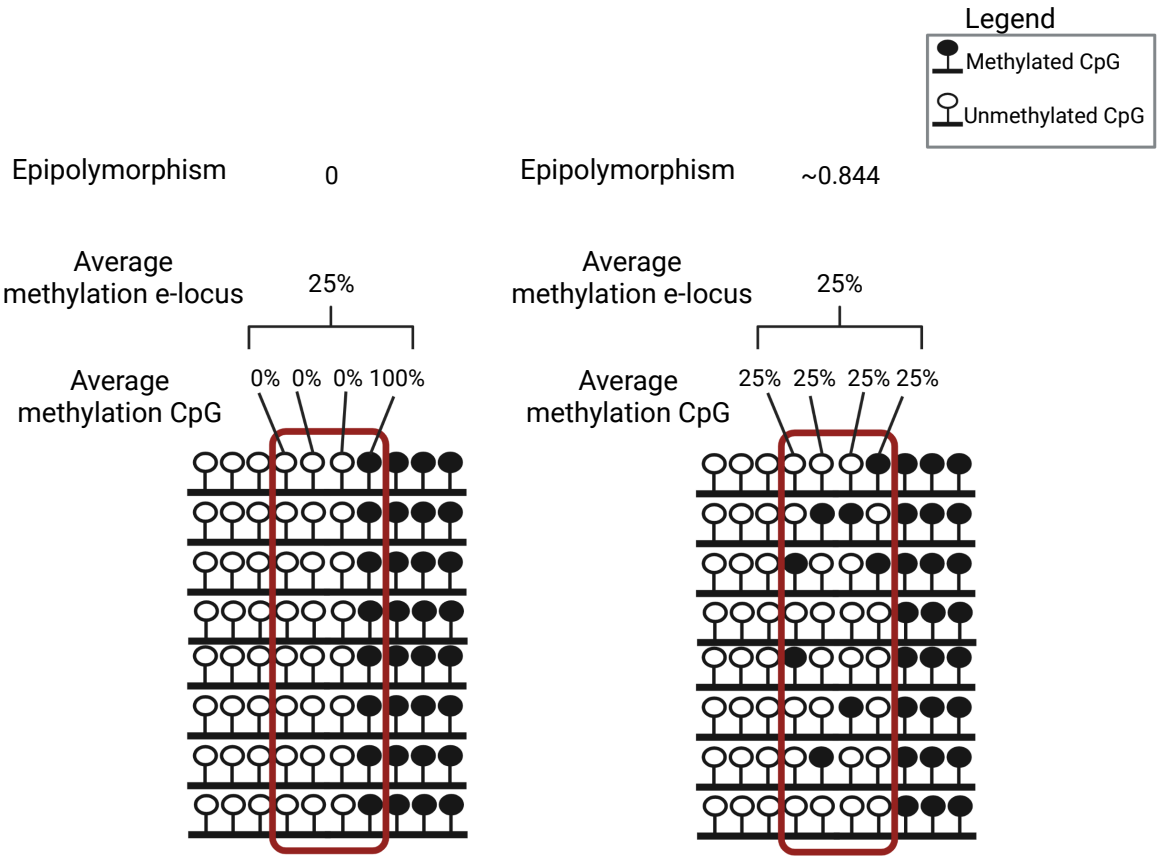


Figure 6.11: Schematic explanation of epipolymorphism & average methylation

Lollipops represent individual CpGs (black: methylated, white: unmethylated) and an e-locus is defined as four adjacent CpGs (red rectangle). The diagram demonstrates average methylation levels at each CpG, average methylation levels across an e-locus and epipolymorphism. Epipolymorphism measures how variable the methylation pattern is within and between reads. This explains how two samples may have the same average methylation across an e-locus, but different epipolymorphism values. This figure was adapted from [113].

6.3.4.1 Discovery in ccRCC versus normal tissue

Data were available for 138,412 e-loci, for 135 multi-region samples. For each e-locus (consisting of 4 adjacent CpGs), epipolymorphism and average methylation were derived for ccRCC tumour and normal samples. Methylation follows a bimodal distribution, with the majority of loci being fully methylated or unmethylated. In my data the relationship between epipolymorphism and the average methylation in an e-locus was U-shaped, in keeping with the existing literature [113] (Figure 6.12). Average methylation of 0 or 1 is associated with epipolymorphism of 0 (i.e. fully concordant methylation), whereas intermediate methylation across the e-locus is associated with varying levels of epipolymorphism (i.e. heterogeneous methylation). The density plot suggests that the majority of e-loci have fully concordant methylation (Figure 6.12).

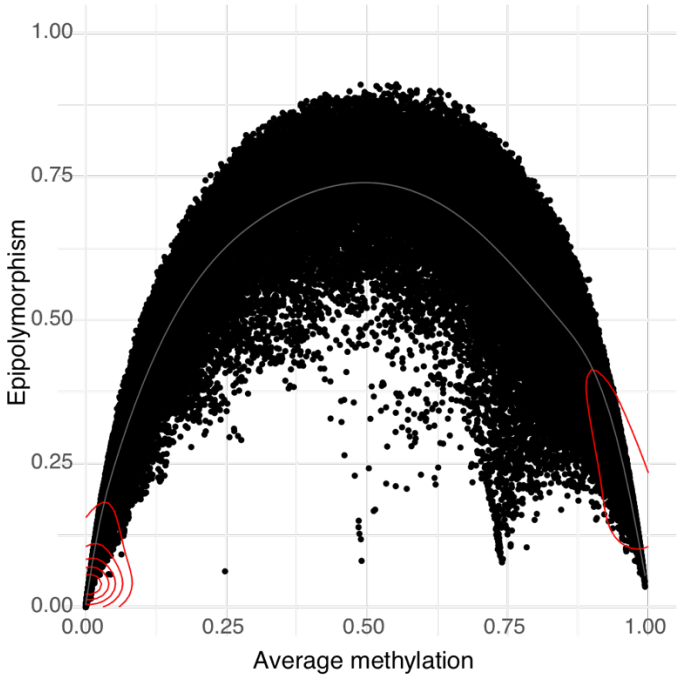


Figure 6.12: Epipolymorphism versus average methylation

Scatterplot of epipolymorphism versus average methylation at e-loci for the entire dataset (N= 138,412 e-loci). The red lines represent density plot contours, demonstrating that the majority of points display low epipolymorphism and methylation levels of 0 or 1. The grey line highlights the U-shaped distribution.

Next, I evaluated differential average methylation and differential epipolymorphism (i.e. the difference in each of these parameters in ccRCC vs normal tissue) and the relationship between these two parameters. E-loci were subdivided into those that have no significant methylation difference in ccRCC versus normal kidney and those with a significant difference. Examining e-loci

that demonstrate no significant average methylation difference in ccRCC tumour versus normal, it is evident that in the vast majority of cases (i.e. 95% of e-loci) this is accompanied by no significant difference in epipolymorphism (Figure 6.13A; Table 6.3). Therefore, these loci have concordant/ordered methylation within a read and have similar methylation patterns in tumour and normal. Conversely, 55% of e-loci with a significant methylation difference in ccRCC vs normal have significant differential epipolymorphism, compared to only 5% of e-loci with no methylation difference (p value < 2.2e-16) (Figure 6.13A, Table 6.3). This suggests that e-loci with a significant methylation difference in tumour versus normal, are more likely to be associated with disordered methylation than e-loci with no methylation difference. Subsequent analysis focused on exploring these sites of significant differential epipolymorphism between tumour and normal tissue.

I identified 28,300 e-loci that demonstrated significant differential epipolymorphism in ccRCC versus normal tissue, with 14,418 e-loci being significantly higher in ccRCC, and 13,882 e-loci having significantly higher epipolymorphism in normal tissue (adjusted p value <0.01, epipolymorphism difference > 0.1). To explore the relevance of the observed changes, these 28,300 significant e-loci were annotated to the nearest gene. E-loci with significantly higher epipolymorphism in ccRCC were located more commonly at gene promoters and in CpG islands, than e-loci with significantly higher epipolymorphism in normal kidney (34% vs 27% at gene promoters, and 56% vs 37% in CpG islands respectively) (Figure 6.13B and C). E-loci with significantly higher epipolymorphism in ccRCC vs normal tissue are not uniformly distributed across the genome, but rather occur more commonly at promoters and CpG islands, which are more likely to affect gene expression and suggests this increase in epipolymorphism might drive dysregulation of gene expression in cancer.

Subsequently, e-loci with significant differential epipolymorphism which are located in the promoter region were selected for enrichment analysis. There was a significant enrichment for genes involved in the following pathways: *Wnt* signalling (e.g. *MYC*, *WIF1*, *UBC*), cell junction organisation (including claudins, keratins, *CDH* genes, *SDK1*), solute carrier membrane (*SLC*) and voltage gated potassium channel (*KCN*) genes (all adjusted p values < 0.05). Disease ontology analysis revealed that there was a significant enrichment for genes which are known to be associated with renal cell carcinoma (47 out of 557 genes *vs* 380 out of 8007 genes in the background set, chi squared adjusted p value = 0.025) (Table 6.4). This suggests that although disordered methylation may be a stochastic process which occurs throughout the genome, ccRCC is associated with gains and losses of locally disordered methylation at the promoter region of known ccRCC genes. The following section externally validates this finding both in kidney cancer cell line data and in an independent cohort of ccRCC tissue samples.

Table 6.3: E-loci with a significant differential methylation and differential epipolymorphism in ccRCC vs normal tissue

E-loci (i.e. a locus containing four adjacent CpGs) were subdivided into those that have no significant methylation difference in ccRCC versus normal and those with a significant difference. Subsequently the proportion of e-loci with a significant epipolymorphism are shown. A significant methylation difference is defined as average methylation across a read (i.e. average methylation across the four CpGs), with a difference >15% in ccRCC versus normal and a q value of <0.01. A significant epipolymorphism difference is defined as epipolymorphism across a read (i.e. four adjacent CpGs) with a difference >0.1 in ccRCC versus normal and a q value of <0.01. It is evident that e-loci with a significant methylation difference in tumour versus normal, are more likely to be associated with a significant epipolymorphism difference than e-loci with no methylation difference.

	Overall number of e-loci	Proportion of e-loci with significantly higher epipolymorphism in Tumour versus Normal	Proportion of e-loci with significantly higher epipolymorphism in Normal versus Tumour
no significant methylation difference	91,520	0.3% (285/91,520)	4.4% (4040/91,520)
e-loci with significant hypermethylation in tumour (>15%, q value <0.01)	6761	26.1% (1763/6761)	31.7% (2146/6761)
e-loci with significant hypomethylation in tumour (>15%, q value <0.01)	3341	23.3% (777/3341)	27.4% (914/3341)

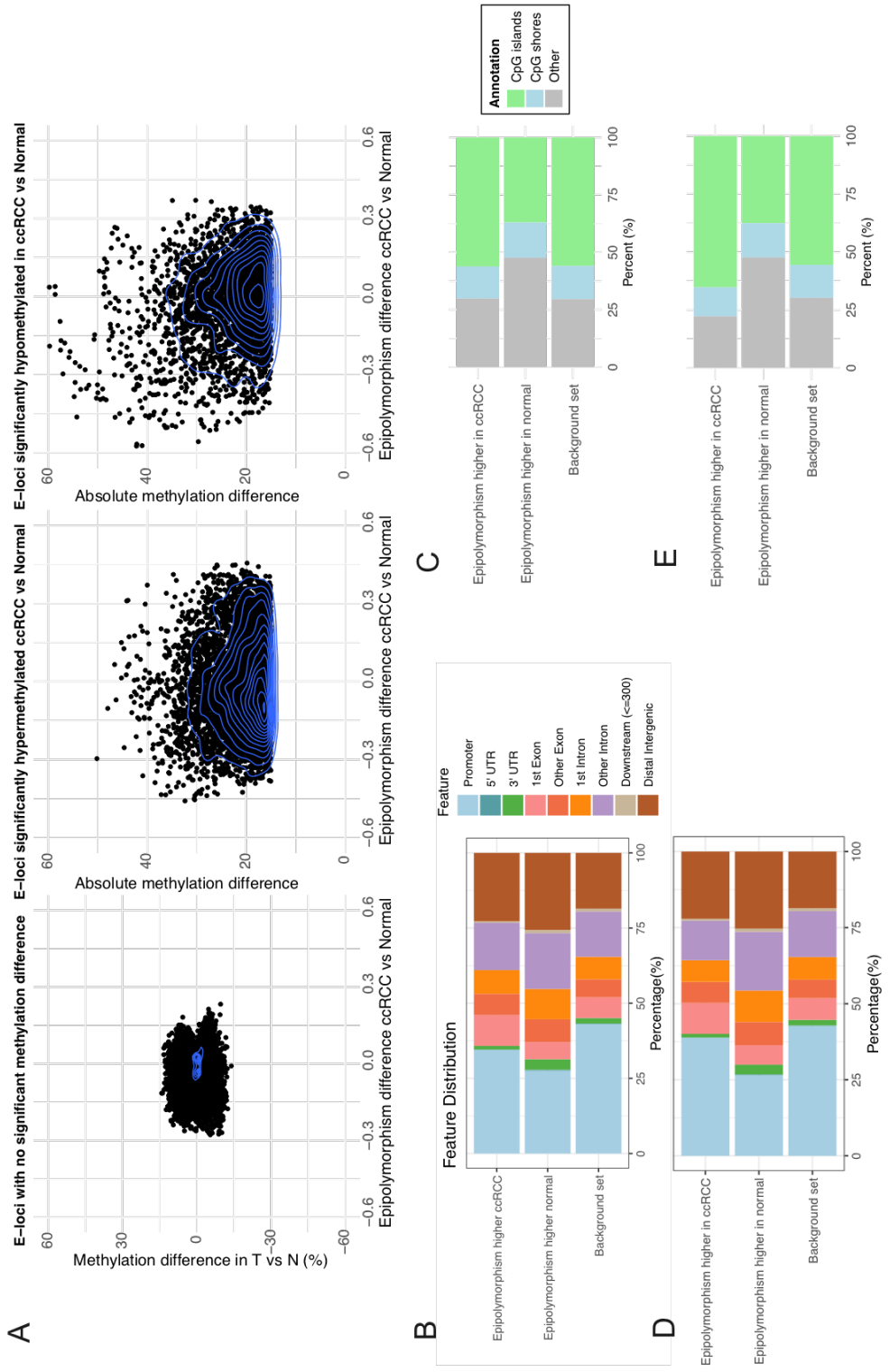


Figure 6.13: Differential epipolymorphism in ccRCC versus normal kidney tissue

Panel A- Scatterplot of differential epipolymorphism versus differential average methylation. E-loci are subdivided into those that have no significant average methylation difference in ccRCC versus normal kidney, those with significantly higher average methylation in ccRCC and those with significantly higher average methylation in normal kidney (panels left to right). Panel B-C- Annotation for e-loci with significant differential epipolymorphism in ccRCC versus normal kidney tissue in the discovery cohort (135 samples). Panel D-E- Annotation for e-loci with significant differential epipolymorphism in ccRCC versus normal kidney in the validation cohort (71 samples).

6.3.4.2 Validation in kidney cancer cell line data

Epipolymorphism at a given e-locus measures both disordered methylation within a read and between reads (i.e. the proportion of different epialleles). It is acknowledged that the proportion of different epialleles may increase due to gains/losses in methylation disorder across a read in different tumour subclones during tumorigenesis or mixtures of heterogeneous cell types. An evaluation of cell line data, which represent a single cell type, has been suggested as an approach to identify regions which display epipolymorphism beyond that which can be attributed to heterogeneous cell types [113]. Therefore, I compared epipolymorphism in the multi-region patient samples (ccRCC and adjacent normal tissue) and in the 786-O renal cancer cell line, as the latter represents a model system of ccRCC with 100% tumour purity (Figure 6.14). Overall, data were available for 109,198 e-loci for ccRCC, normal kidney and the 786-O cell line (Figure 6.14A). Evaluating e-loci with differential epipolymorphism in ccRCC vs normal kidney revealed that as expected the distribution of epipolymorphism values in the cancer cell line was more similar to ccRCC than normal tissue (Figure 6.14B and C). For example, I identified 10,713 e-loci where epipolymorphism was significantly higher in ccRCC vs normal kidney, and in 7132 of these e-loci, epipolymorphism in the 786-O cell line was even higher than tumour tissue (Figure 6.14C). This suggests that disordered methylation within a pure cell line may be associated with tumorigenesis, rather than being an artefact of cell contamination. Next, epipolymorphism was visualised at e-loci within the promoter region of eight selected genes known to be associated with RCC identified from section 6.3.4.1 (*MAD2L2*, *TRPC4*, *NDRG2*, *NOS2*, *TTYH2*, *RAB37*, *FRZB*, *TERT*) (Figure 6.14D). Epipolymorphism values in ccRCC tissue and the cancer cell line were higher than that for normal kidney (Figure 6.14D).

Subsequently, I compared epipolymorphism in the 786-O and the HK2 cell lines, to evaluate whether differential epipolymorphism may be noted in model systems of cancers vs controls, having removed the confounding effect of purity. The 786-O and HK2 cell lines represent ccRCC and normal renal proximal tubule epithelium respectively. I generated Epic-seq methylation data for 786-O (4 technical replicates) and HK2 cell lines (6 technical replicates). Data were available for 124,426 e-loci, of which 14,964 were found to have significantly higher epipolymorphism in 786-O, and 2198 significantly higher epipolymorphism in HK2 cell lines. My analysis then focused on the 47 genes which demonstrated differential epipolymorphism within their promoter in ccRCC versus normal tissue and are known to be associated with kidney cancer in disease ontology analysis (from section 6.3.4.1 above). Cell line data were missing for 3 of these genes. Out of the remaining 44 genes identified in ccRCC tissue, 28 genes also had differential epipolymorphism within their promoter

region in 786-O vs HK2 cell lines (epipolymorphism difference > 0.1, adjusted p value <0.05) (Table 6.4). Once again, the presence of differential epipolymorphism in the promoter region of these genes in the cell line, which is 100% pure, suggests methylation heterogeneity within a read may be associated with kidney cancer, rather than being an artefact of mixtures of heterogeneous groups of cells in kidney tissue. The literature suggests that methylation patterns in cell lines change secondary to immortalization and growth in culture [305]. HK2 cells therefore may not represent the methylation pattern in normal kidney and this could contribute to why some of the genes known to be associated with kidney cancer did not have differential epipolymorphism in HK2 vs 786-O cell lines.

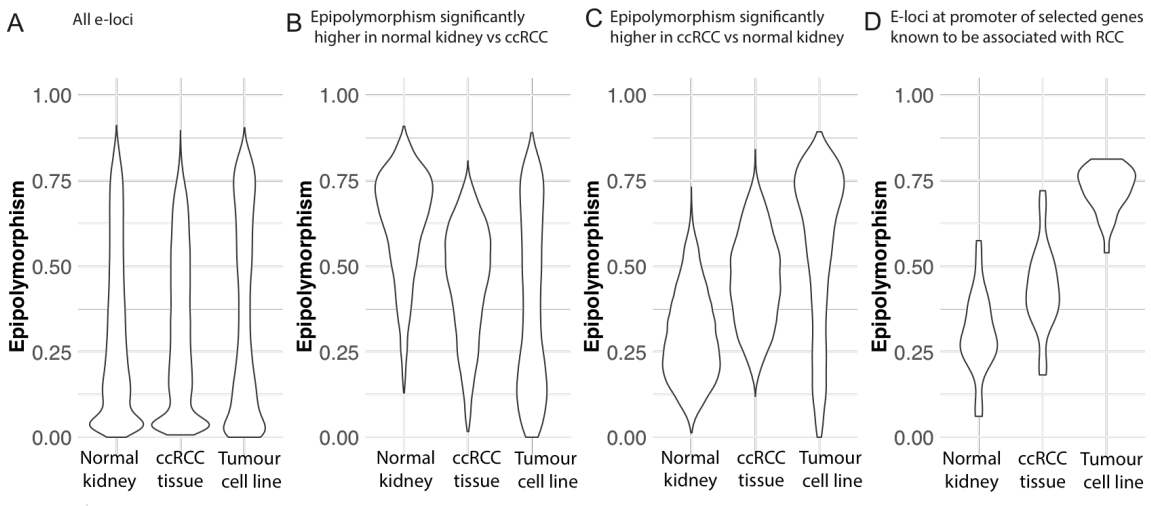


Figure 6.14: Epipolymorphism in normal kidney, ccRCC tissue and the 786-O ccRCC cell line

Epipolymorphism is shown in ccRCC tumour tissue, normal kidney tissue and the 786-O renal cancer cell line for: all e-loci (Panel A), e-loci with significantly higher epipolymorphism in normal kidney (Panel B) and e-loci with significantly higher epipolymorphism in ccRCC (Panel C). Panel D demonstrates epipolymorphism values for selected e-loci in the promoter region of 8 genes which are known to be associated with kidney cancer and were found to have significantly higher epipolymorphism in ccRCC vs normal kidney. These genes were selected to be visualised as an illustrative example.

6.3.4.3 External validation in an independent cohort of ccRCC tissue samples

I externally validated the results from the multi-region analysis in a separate cohort of 71 ccRCC and normal kidney samples (these were not multi-region samples). For these 71 samples, I generated Epic-seq methylation values and data were available for 148,096 e-loci. Of these, 21,094 were found to have significant differential epipolymorphism in ccRCC versus normal tissue, with 11,841 e-loci having significantly higher disordered methylation in ccRCC, and 9253 e-loci having significantly higher epipolymorphism in normal kidney. Once again, e-loci that had significantly higher epipolymorphism in ccRCC were located more commonly at gene promoters and in CpG islands, than e-loci with significantly higher epipolymorphism in normal kidney (39% vs 26% at gene promoters, and 65% vs 38% in CpG islands respectively) (Figure 6.13D and E). Out of the 47 RCC-associated genes which were found to have differential epipolymorphism in the first cohort (in section 6.3.4.1), 31 genes also demonstrated differential epipolymorphism in the external validation cohort (at the same e-loci) (Table 6.4). This provides further evidence that ccRCC is associated with gains and losses of locally disordered methylation at the promoter region of genes known to be associated with RCC.

Table 6.4: Genes known to be associated with ccRCC which demonstrate differential epipolymorphism

Genes known to be associated with ccRCC that were found to have significant differential epipolymorphism within their promoter region in ccRCC versus normal kidney tissue.

	Identified in the original cohort of ccRCC vs normal tissue (135 multiregion samples)	Genes with differential epipolymorphism in HK2 vs 786-O cell line	Genes with differential epipolymorphism in external validation cohort of ccRCC vs normal tissue (71 samples)
Genes with significant differential epipolymorphism in their promoter region	47 genes: *MAD2L2, AGT, CD44, KRT18, TRPC4, NDRG2, CD276, NOS2, TTYH2, RAB37, FRZB, TERT, TBXT, MUC3A, PRKCZ, TP53BP2, TGFBR3, HIF1AN, FGFR2, ALOX5, ADM, IL18, ST3GAL4, WT1, PXN, KRT7, WIF1, MOK, PRKCH, IGF1R, ABCC1, CXCL16, HPN, CGB3, BSG, BCL2L11, JAG1, EGF, DUSP1, SOD2, HLA-G, HLA-A, TAP2, PODXL, MAD1L1, IGF2BP3, CCN3, MYC*	28 genes: *MAD2L2, PRKCZ, AGT, ALOX5, IL18, WT1, PXN, KRT7, KRT18, WIF1, TRPC4, NDRG2, CXCL16, TTYH2, RAB37, HPN, CGB3, BCL2L11, FRZB, JAG1, EGF TERT, DUSP1, SOD2, HLA-A, MUC3A, MAD1L1, CCN3*	31 genes: *MAD2L2, WT1, CD44, TRPC4, NDRG2, CD276, NOS2, TTYH2, RAB37, TERT, PRKCZ, AGT, FGFR2, ADM, IL18, WT1, KRT7, WIF1, PRKCH, IGF1R, ABCC1, BSG, BCL2L11, JAG1, EGF, SOD2, HLA-G, HLA-A, TAP2, MAD1L1, MYC*

6.3.4.4 Association between disordered methylation and gene expression

A key question is whether differential epipolymorphism has a functional relevance on gene expression. In order to explore the association between gene expression, methylation and epipolymorphism, Epic-seq and matched RNA-seq data were obtained for a subset of multi-region samples (N=47). The analysis was limited to e-loci which have significant differential epipolymorphism in ccRCC vs normal kidney and are located within the promoter region of genes, as these changes are more likely to be functional (N= 7536 e-loci). I was interested in exploring the overall effect of epipolymorphism on gene expression. Therefore, I developed a linear model to predict gene expression based on epipolymorphism. Out of the 7536 e-loci assessed, I identified 1870 e-loci (in the promoter region of 475 unique genes) where epipolymorphism was a significant predictor of gene expression in univariate analysis (BH adjusted p value <0.05). This included genes known to be associated with kidney cancer identified in section 6.3.4.1 (*ALOX5*, *WT1*, *CD44*, *KRT7*, *KRT18*, *CD276*, *CXCL16*, *RAB37*, *BCL2L11*, *JAG1*, *EGF*, *HLA-A*, *IGF2BP3*, *SLC16A3*, *DPP6*). In the majority of cases, the correlation between epipolymorphism and gene expression was negative. Indeed, a negative correlation was seen in 87% (N= 976/1119) of e-loci with higher epipolymorphism in ccRCC, and 60% (N=449/751) of e-loci with higher epipolymorphism in normal kidney. This provides evidence of an association between promoter epipolymorphism and gene expression. In particular, increased disordered methylation within the promoter in ccRCC tends to be associated with transcriptional repression compared to normal kidney.

One possible hypothesis is that the association between epipolymorphism and gene expression is driven by average methylation. To ascertain the effect of epipolymorphism beyond methylation, I subsequently evaluated a linear model predicting gene expression based on methylation alone or methylation and epipolymorphism, as previously performed by Landau et al [110]. This enabled me to assess whether the inclusion of epipolymorphism can increase the model's predictive ability, which would imply that epipolymorphism is an independent predictor of gene expression. My analysis identified 216 e-loci (at the promoter region of 103 unique genes) where the addition of epipolymorphism resulted in significant improvements in the model compared to a model based on methylation alone (i.e. significant increases in adjusted R^2, likelihood ratio test BH adjusted p value <0.05; Figure 6.15A and B). These 216 e-loci were examined more closely. They consist of 43 e-loci (in the promoter of 21 genes) with significantly higher epipolymorphism in ccRCC, and 173 e-loci (in the promoter of 82 genes) with significantly higher epipolymorphism in normal tissue (Table 6.5 and Table 6.6). Many of the genes identified in Table 6.5 and Table 6.6 have previously been implicated in cancer disease biology and ccRCC in particular. E-loci with significantly higher epipolymorphism in

tumour include genes coding for cytokeratins (*KRT18*), membrane transporters (*SLC22A31*), zinc finger transcription factors (*ZNF728*) and genes involved in cell proliferation (e.g. *RASL11B*; [306]). E-loci with significantly higher epipolymorphism in normal kidney include genes coding for membrane transporters (*SLC16A3*), glomerular markers (*NFASC*), claudins which play a role in cell adhesion (*CLDN4*, *CLDN23*), genes involved in cellular response to hypoxia (*PSMF1*, *UBC*, *UBE2D2*), Notch signalling (*ARRB2*, *JAG1*, *UBC*), apoptosis (*SERPINB9*), angiogenesis and migration (*IGFBP7* [307, 308]), kidney epithelial morphogenesis (*SMTNL2* [309, 310]), kidney cancer tumorigenesis (*LPCAT1* [311]) and zinc finger transcription factors (*ZNF888*). Similarly, Landau et al showed a significant association between epipolymorphism and gene expression in genes coding for Zinc finger proteins [110]. Furthermore, my analysis identified genes coding for histone proteins (*H2BC13*, *H2AC11*, *H2AC17*, *H2AC16*, *H2BC11*, *H2BC7*, *H4C12*) in which epipolymorphism was significantly higher in normal tissue compared to ccRCC, and epipolymorphism may contribute to gene expression (Table 6.6). This finding warrants further future investigation. Interestingly, a number of the identified genes are involved in immunoregulatory interactions between lymphoid and non-lymphoid cells. ccRCC tissue contains more immune cells than normal kidney tissue (see later section 6.3.5.1.1). If the effect on epipolymorphism was simply an artefact secondary to low tumour purity (contamination with immune cells), one would expect epipolymorphism to be higher in ccRCC for these genes. However, my data demonstrates significantly increased promoter epipolymorphism in normal tissue compared to ccRCC for these genes (as well as an association with gene expression), suggesting these changes may be related to tumorigenesis rather than purity.

Table 6.5: Linear models to predict gene expression, for e-loci with significantly higher epipolymorphism in ccRCC

Linear models to predict gene expression based on methylation and epipolymorphism versus methylation alone, for e-loci with significantly higher epipolymorphism in ccRCC compared to normal tissue. The likelihood ratio test was used to compare adjusted R^2 values for the two models. Where multiple e-loci were found to be significant, the number of significant e-loci are reported and the e-locus with the lowest adjusted p value is shown in the table. Genes are ranked based on descending adjusted R^2 for the linear model using methylation and epipolymorphism.

Gene	Correlation coefficient: Epipolymorphism vs gene expression	Linear model adjusted R^2, Gene expression predicted by methylation	Linear model adjusted R^2, Gene expression predicted by methylation and epipolymorphism	Likelihood ratio test adjusted p value	Number of significant e-loci in the promoter region of the gene
RASL11B	-0.78	0.42	0.62	0.005	2
ZNF728	-0.78	0.45	0.59	0.029	2
DPP6	-0.66	0.14	0.58	0.0003	4
KRBA1	0.76	0.31	0.56	0.004	4
HS3ST3B1	-0.73	0.3	0.53	0.006	5
KRT18	0.73	0.37	0.51	0.044	1
ICA1	-0.68	0.21	0.47	0.007	1
LINC02693	-0.66	0.29	0.46	0.03	1
DLGAP1	-0.66	0.17	0.44	0.007	3
LRAT	0.60	0.16	0.44	0.007	4
ESPNP	-0.59	0.16	0.42	0.009	2
MAN1C1	-0.67	0.26	0.42	0.049	1
CD4	0.66	-0.02	0.41	0.004	1
MIR663AHG	-0.46	-0.02	0.40	0.007	1
PLAC9	-0.40	-0.01	0.33	0.015	1
LINC01287	0.14	0.07	0.30	0.044	1
PRR5L	-0.39	-0.01	0.28	0.009	3
ANKRD18CP	-0.31	-0.02	0.27	0.012	1
SLC22A31	-0.37	-0.02	0.27	0.01	2
CD8B2	-0.24	-0.01	0.24	0.03	1
PRDM6	-0.40	-0.02	0.23	0.031	2

Table 6.6: Linear models to predict gene expression, for e-loci with significantly higher epipolymorphism in normal kidney

Linear models to predict gene expression based on methylation and epipolymorphism versus methylation alone, for e-loci with significantly higher epipolymorphism in normal kidney compared to ccRCC. The likelihood ratio test was used to compare adjusted R^2 values for the two models. Where multiple e-loci were found to be significant, the number of significant e-loci are reported and the e-locus with the lowest adjusted p value is shown in the table. Genes are ranked based on descending adjusted R^2 for the linear model using methylation and epipolymorphism. A total of 82 genes were identified (adjusted p value <0.05), however only the top-most genes are shown here.

Gene	Correlation coefficient: Epipolymorphism vs gene expression	Linear model adjusted R^2, Gene expression predicted by methylation	Linear model adjusted R^2, Gene expression predicted by methylation and epipolymorphism	Likelihood ratio test adjusted p value	Number of significant e-loci in the promoter region of the gene
SLC16A3	-0.92	0.77	0.84	0.001	3
H2BC13	-0.69	0.21	0.63	7.67E-06	4
MFHAS1	0.70	0.24	0.63	8.20E-06	8
H2AC11	-0.72	0.11	0.54	1.14E-05	5
PDZD2	0.54	0.19	0.51	0.001	1
B3GNTL1	-0.69	0.25	0.47	0.005	1
CLDN4	0.39	0.25	0.46	0.005	3
H2AC16	-0.20	-0.01	0.45	2.69E-05	8
H2BC11	-0.67	0.26	0.44	0.009	2
LPCAT1	-0.60	0.01	0.44	6.08E-05	2
UBE2D2	-0.67	0.08	0.44	1.86E-04	10
NINJ2	-0.68	0.06	0.43	0.001	5
IGF2BP3	-0.63	0.13	0.42	0.002	1
NFASC	0.56	0.02	0.41	0.0003	1
RRM2	-0.54	0.09	0.38	0.002	3
METRNL	0.20	0.05	0.35	0.003	3
TUBBP5	0.51	-0.01	0.35	0.001	2
SMTNL2	0.59	-0.02	0.33	0.002	12
SNX29P2	0.54	0	0.33	0.004	1
ATAD5	-0.51	0.1	0.32	0.009	2
MSC	-0.55	0.02	0.32	0.009	2
OLFML2A	0.24	0.06	0.32	0.005	2
UBC	-0.54	0.04	0.32	0.004	4
TBC1D14	0.56	-0.01	0.31	0.002	2
ZNF888	-0.56	0.02	0.3	0.004	1
LMF1-AS1	0.53	0.01	0.26	0.009	1

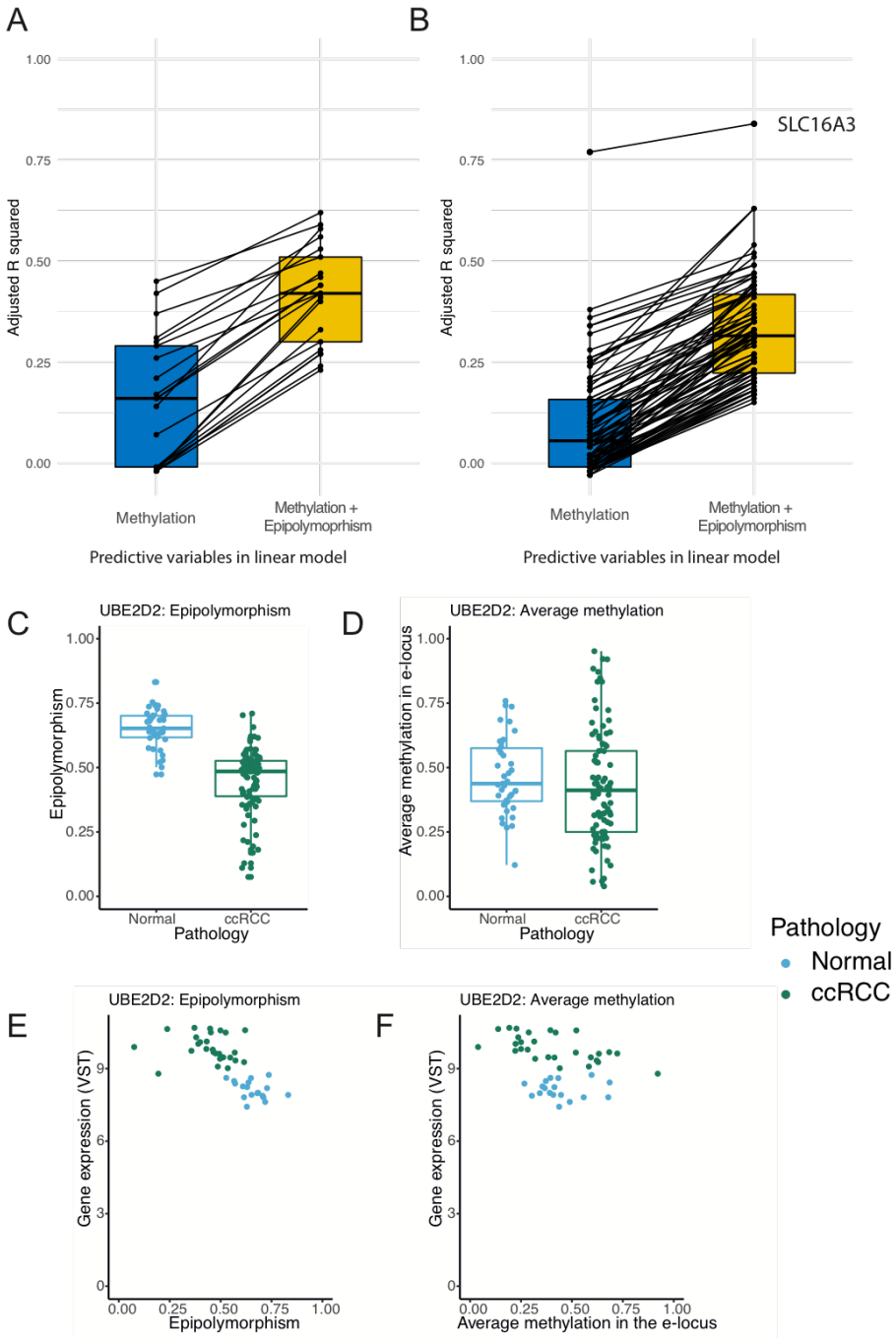


Figure 6.15: Epipolymorphism, methylation and gene expression

Panel A and B – Adjusted R^2 for a linear model to predict gene expression based on methylation alone versus methylation and epipolymorphism. Results are shown for genes which had a statistically significant improvement in the R^2 from Table 6.5 and Table 6.6. Results are shown separately for e-loci with significantly higher epipolymorphism in ccRCC (Panel A) and e-loci with significantly higher epipolymorphism in normal kidney (Panel B). *SLC16A3* is the gene with the highest adjusted R^2 and is an outlier, therefore it was explored more in detail in Figure 6.16. Panel C and D- Epipolymorphism (Panel C) and average methylation (Panel D) in ccRCC and normal kidney tissue at the *UBE2D2* gene promoter. Results are shown for all 135 tissue samples for which DNA methylation data were available. Panel E and F- *UBE2D2* gene expression versus epipolymorphism (Panel E) and average methylation (Panel F) in ccRCC and normal kidney tissue. Results are shown for 47 tissue samples for which matched DNA methylation and RNA-seq data were available. VST refers to the variance stabilising transformation applied to gene expression

The following section will discuss three of these genes (*UBE2D2*, *SLC16A3* and *DPP6*) in more detail as illustrative examples. I identified 10 e-loci within the promoter region of the *UBE2D2* gene in which epipolymorphism predicted gene expression independently of methylation (significant improvement in the adjusted R^2 following inclusion of epipolymorphism in the linear model compared with methylation alone, Figure 6.15). There was significantly lower *UBE2D2* expression in normal tissue compared to ccRCC, and this was accompanied by significantly higher promoter epipolymorphism, with no evidence of a difference in average methylation across the read (Figure 6.15C-F). There was a strong negative correlation between epipolymorphism and gene expression (Pearson correlation coefficient = -0.67, adjusted p value = 1.80e-05), with no evidence of an association between methylation and gene expression (adjusted p value >0.05; Figure 6.15). For the best performing e-locus, a linear model predicting gene expression based on epipolymorphism and methylation produced a significant improvement compared to a model using methylation alone (improvement in adjusted R^2 from 0.08 to 0.44), which is not surprising given the absence of a methylation difference. *UBE2D2* codes for Ubiquitin Conjugating Enzyme E2 D2, which is a member of the E2 ubiquitin-conjugating enzyme family, which facilitates proteasome degradation of various proteins, including tp53. *UBE2D2* facilitates the cell's response to hypoxia and is regulated by the HIF transcription factor [312]. My data demonstrates increased *UBE2D2* expression in ccRCC. The literature suggests increased *UBE2D2* causes reduced *tp53*, which in turn is pro-tumorigenic (*tp53* mutations are the most common alteration in cancer [313]). In summary, my data is the first to suggest evidence of significantly disordered methylation in normal kidney compared to ccRCC. The disordered methylation stretches for relatively large regions within the *UBE2D2* gene promoter (i.e. 10 e-loci) and may contributed to altered gene expression.

For e-loci with higher epipolymorphism in normal kidney, *SLC16A3* was the top-ranking gene with the highest R^2 value for the linear model containing methylation and epipolymorphism to predict gene expression (adjusted R^2= 0.84; Figure 6.15B and Table 6.6). This suggests that a high proportion of the variability observed in gene expression is explained by the model based on methylation and epipolymorphism. This gene encodes for the monocarboxylate transporter MCT4, which transports lactate (along with H+) generated during glycolysis across the plasma membrane and plays a key role in mediating the Warburg effect in ccRCC [314]. MCT4 overexpression is associated with increased cell proliferation, reduced apoptosis and tumorigenesis [314]. In >85% of ccRCC patients, there is MCT4 gene over-expression and associated hypomethylation at specific sites in the gene promoter of tumour tissue compared to adjacent normal kidney [315]. Fisel et al [315] demonstrated that gene expression is regulated by methylation by performing RCC cell line experiments comparing promoter/reporter fusion plasmids containing either methylated or mock-methylated *SLC16A3*

promoter fragments. Significantly reduced promoter activity was noted in the mock-methylated compared to methylated cells. This is clinically relevant because high *SLC16A3* expression has been found to be a negative predictor of survival in several independent ccRCC cohorts [314-316], and there is potential for therapeutic targeting with MCT4 inhibitor drugs. In addition, lactate transport, which is mediated by MCT4, can be non-invasively imaged using Hyperpolarized ^{13}C-pyruvate magnetic resonance imaging [317] and its role in a diagnostic and prognostic setting is under investigation. In agreement with the available literature, my data confirmed that in ccRCC there is *SCL16A3* promoter hypomethylation and this is associated with increased gene expression compared to normal kidney (Figure 6.16). In addition, epipolymorphism is significantly different in 20 e-loci in the promoter region of the gene in ccRCC versus normal, with disordered methylation in normal tissue and ordered methylation (i.e. concordant hypomethylation) in ccRCC and this is associated with changes in gene expression (Figure 6.16). This could suggest that disordered methylation in the normal tissue is an early event which predisposes to hypomethylation in ccRCC tissue and this has a functional relevance on gene expression.

For e-loci with higher epipolymorphism in ccRCC, *DPP6* was selected as an illustrative example (Table 6.5). *DPP6* (Dipeptidyl Peptidase Like 6) is a peptidase membrane glycoprotein which mediates *KCND2* voltage-gated potassium channels and regulates cell differentiation, proliferation, apoptosis and tumorigenesis [318]. Sheikh et al demonstrated that *DPP6* gene expression is regulated by DNA methylation and that this drives neuronal cell differentiation and regulates a cell-type specific phenotype [318]. Similarly, hypermethylation and reduced gene expression have been noted in other malignancies (including pancreatic cancer, melanoma and acute myeloid leukaemia) [319-322]. In my analysis of renal tissue methylation, there was an increase in disordered methylation at the *DPP6* promoter, as well as average hypermethylation in ccRCC vs normal kidney. A model predicting gene expression based on epipolymorphism and average methylation significantly increased the adjusted R^2 compared to a model based on methylation alone (increase in adjusted R^2 from 0.14 to 0.58, adjusted p value = 0.0003). A negative association was noted between promoter epipolymorphism and gene expression (correlation = -0.66), which may be consistent with epigenetic silencing of *DPP6* in ccRCC. Previous studies have reported reduced *DPP6* gene expression as an independent prognostic marker in ccRCC tissue [321] and separate studies reported promoter hypermethylation in ccRCC, associated with increasing tumour grade, stage and a predictor of metastatic recurrence [323]. My work is the first to directly link *DPP6* methylation and gene expression in ccRCC, and the first in any biological system to characterise epipolymorphism at this gene locus. In summary, this analysis suggests that epipolymorphism within a gene promoter may be an independent regulator of gene expression in addition to overall methylation.

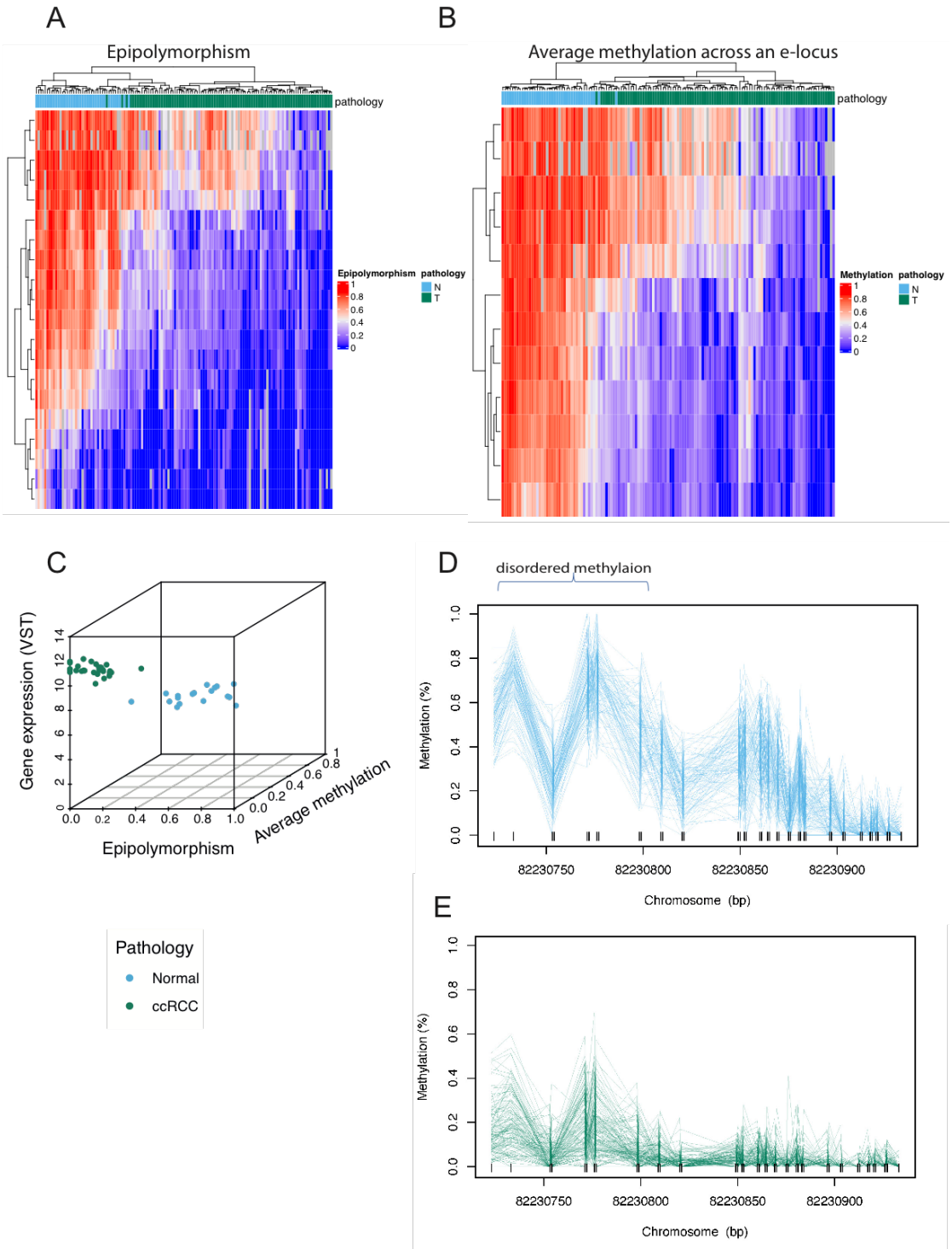


Figure 6.16: Epipolymorphism, methylation and gene expression for e-loci within the promoter region of the SLC16A3 gene

Panel A and B- Heatmap demonstrating epipolymorphism (Panel A) and average methylation (Panel B) in each e-locus in the promoter region of *SLC16A3*, in ccRCC and normal kidney tissue. Panel C- Scatterplot of gene expression versus epipolymorphism and average methylation in a 3D scale. Epipolymorphism and average methylation are higher for normal kidney than ccRCC, and there is associated reduced expression in normal kidney. Panel D and E- Methylation levels along the promoter region of the *SLC16A3* gene. In ccRCC, there is global hypomethylation (Panel E), whereas in normal kidney there is evidence of disordered methylation (hypermethylation and hypomethylation of adjacent CpGs; Panel D).

6.3.5 Multi-region samples to inform biomarker selection

The following section focuses on integrating data from multi-region ccRCC samples to glean information regarding biomarker selection and potential clinical implications. Two complementary analyses are performed, shown below.

1) Analysis of significant DMCs that differentiate ccRCC vs normal kidney, exploring how cell type composition or contamination may affect selection of DMCs as potential biomarkers (Section 6.3.5.1).
2) Evaluation of CpGs which are homogeneously versus heterogeneously methylated within a patient, and recurrent across the cohort. This analysis will enable an assessment of methylation changes that may be early or late events and therefore have different biomarker applications (Section 6.3.5.2).

6.3.5.1 Exploring how cell type composition may affect selection of DMCs

I postulated that differing cell type composition may impact the DMCs that are called as being significantly different in ccRCC vs normal kidney, and that exploring this will help select more informative tumour markers. First, I used 'methylKit' to call significant DMCs between ccRCC and normal in the 136 multi-region samples (with patient as a covariate, see Methods 4.4.3). In this case, all SNP types were removed to ensure that when multi-region samples from the same patient cluster together, this is due to similar patterns of DNA methylation rather than SNPs. After removal of SNPs, 107,947 DMCs were identified and the top 10% of DMCs with the lowest variance in tumour samples were compared to the top 10% of DMCs with the highest variance in tumour samples (Figure 6.17A and B).

DMCs with low variance in multi-region tumour samples have both low between and within sample heterogeneity, and methylation patterns do not seem to be related to tumour purity (Figure 6.17A). These DMCs clearly separate tumour versus normal and multi-region samples from the same patient cluster together. These could therefore represent ideal biomarker targets for diagnostic clinical applications. Gene set enrichment analysis of DMCs in the gene promoter region identified pathways associated with cell-cell communication and cell adhesion (e.g. *SDK1*), claudins (*CLDN8*, *CLDN10*, *CLDN14*) and signal regulatory protein family interactions (e.g. protein kinases such as *PTK2* and *SRC*), which are key tumorigenic pathways. This implies these DMCs are likely to represent ccRCC specific changes.

Evaluating DMCs with the highest tumour variance demonstrates two major branches of the dendrogram (Figure 6.17B). One branch includes normal samples and ccRCC samples with high tumour purity and the other branch consists of ccRCC samples with low purity. Interestingly, low purity samples cluster away from normal kidney. Indeed, purity is significantly higher in samples that cluster more closely with normal tissue than those that cluster away from normal (mean purity 59.4% vs 39.9%, p value = 3.9e-8, Figure 6.17C). These DMCs with high variance in ccRCC were annotated to proximal genes, and gene set enrichment was performed for those in the promoter region. GSEA highlighted pathways associated with neutrophil degranulation (51/464 vs 480/10,654; adjusted p value= 2.28e-06). Taken together, these two results suggest that a proportion of DMCs that are called as significantly different between ccRCC and normal tissue, may not actually represent markers that are associated with ccRCC tumours, but rather may be markers associated with immune cell infiltration secondary to low tumour purity. Any attempt to use these as diagnostic biomarkers in clinical practice would therefore be hampered by heterogeneity associated with sample purity. The following section focuses on characterising cell type composition in my tissue samples and exploring the clinical utility of this analysis.

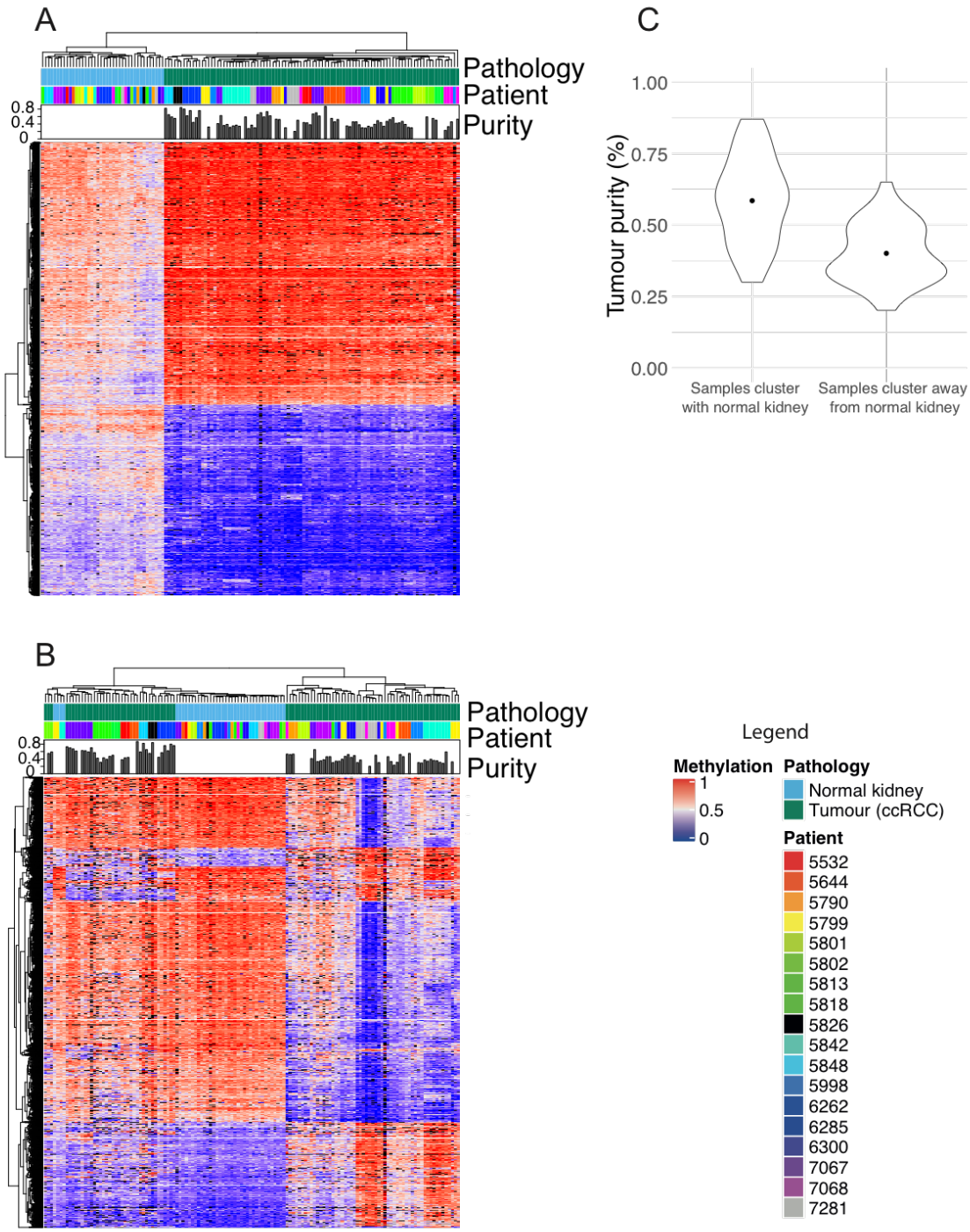


Figure 6.17: DMCs with low and high variance in tumours

Panel A-B: Heatmap of methylation levels for ccRCC and normal kidney at selected differentially methylated cytosines (DMCs). DMCs were identified and the top 10% with the lowest variance in tumour samples (Panel A) were compared to the top 10% of DMCs with the highest variance in tumour samples (Panel B). The top annotation bar shows tumour purity for each sample (estimated using WES or RNA-seq). Panel C- Tumour purity for ccRCC samples which clustered either with normal samples, or away from normal samples in the dendrogram accompanying the heatmap in Panel B. Tumour purity for each sample was calculated using WES or RNA-seq. Purity was significantly higher in samples that cluster more closely with normal tissue than those that cluster away from normal (mean purity 59.4% vs 39.9%, p value = 3.9e-8).

6.3.5.1.1 Characterising cell type composition in multi-region samples

It is recognised that bulk tissue data consists of signal from multiple cell types, and several methods have been developed to perform deconvolution of cell type components. This is particularly interesting because ccRCC is characterised by a high immune cell and stromal infiltrate compared to other malignancies [203]. Studies suggest the presence of immune cell heterogeneity both between and within patients, and this appears to have prognostic implications [299]. I therefore sought to explore purity and cell type composition in my multi-region samples using three orthogonal methods: DNA methylation analysis, RNA-seq and WES (Figure 6.18).

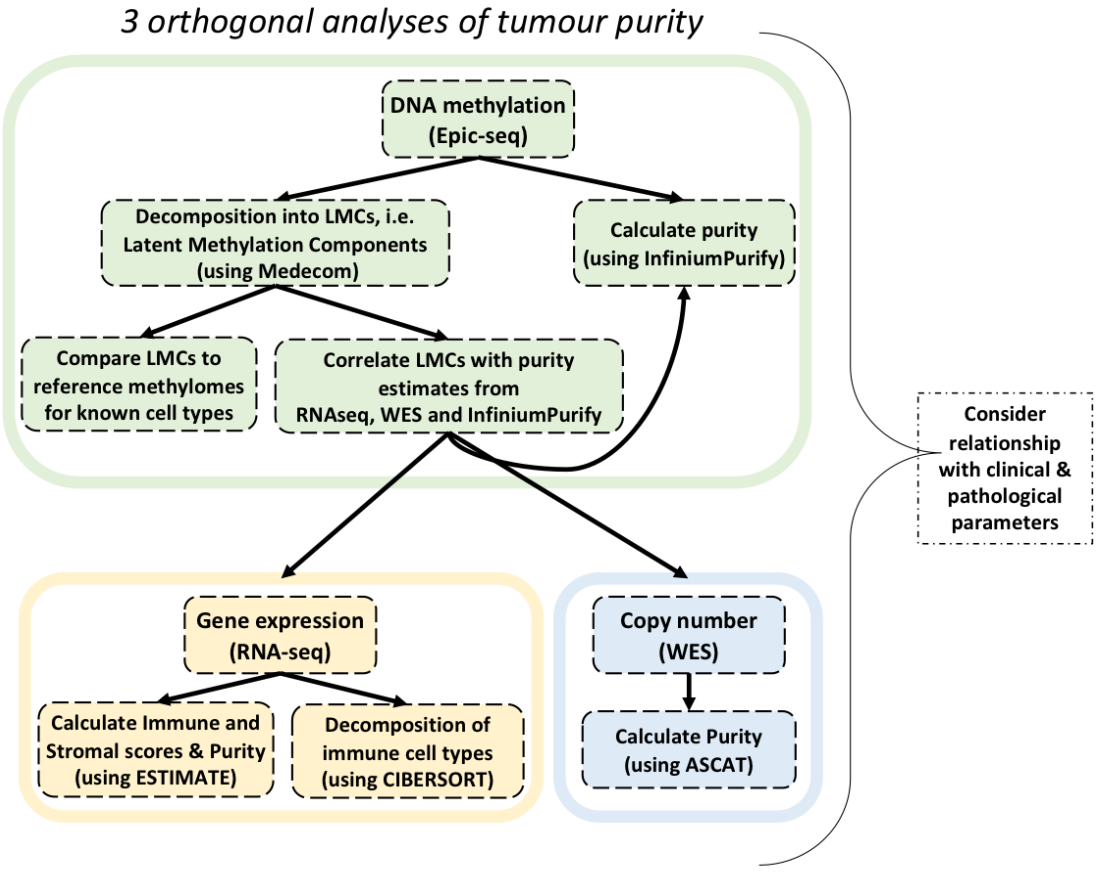


Figure 6.18: Schematic demonstrating purity and cell type decomposition analysis

Schematic demonstrating methods used to explore purity and cell type decomposition in my multi-region samples using three orthogonal methods: DNA methylation analysis, RNA-seq and WES. In brief, I used the 'MeDeCom' package to deconvolute DNA methylation data into latent methylation components (LMCs). I also used 'CIBERSORTx' to deconvolute RNA-seq data into immune cell types. Subsequently, I evaluated whether the proportion of LMCs in each tissue sample was correlated with purity estimates calculated from WES, RNA-seq and methylation data.

In order to determine cell type composition in my samples, I performed reference-free cell type deconvolution based on methylation data. A 'reference-free' method does not rely on reference methylomes from the literature, and this was deemed the best option due to the absence of reliable, good-quality kidney tissue specific reference methylomes. I selected the top 10% of DMCs with the highest variance in tumour samples identified in section 6.3.5.1 and performed reference-free deconvolution using the 'MeDeCom' package [200, 201].

'MeDeCom' uses regularized non-negative matrix factorization to decompose the DNA methylation matrix into two matrices: cell-type-specific latent methylation components (LMCs) and the proportion of LMCs in each sample [201] (Figure 6.19A). LMCs represent the reference methylomes of unknown cell populations. The following model parameters were selected, as these minimized the cross-validation error: K=7 (i.e., the number of LMCs) and λ = 0.01 (i.e., regularization parameter). Thus, the samples were deconvoluted into 7 LMCs, which represent different cell types. The subsequent analysis focuses on characterizing these LMCs and determining whether these have clinical utility.

First, the proportion of each LMC present in each of the multi-region samples was visualized in a heatmap (Figure 6.19B). Normal tissue samples cluster together, and are characterised by a high proportion of LMC5, suggesting this LMC represents normal kidney cells. Tumour samples consist of two major branches. Generally, multi-region tumour samples from the same patient tend to cluster together. Subsequently, I performed an orthogonal analysis to determine the nature of the LMCs:

1) The proportion of LMCs in each sample was correlated with purity estimates calculated from WES, RNA-seq and methylation data
2) The methylomes of the LMCs were compared to publicly available reference methylomes for known cell types

Table 6.7 summarizes the results of this analysis. In summary, LMC5 appears to represent normal kidney epithelium, LMC4 is likely to be a marker of ccRCC, whereas LMC1 and LMC3 are likely to represent immune cells (LMC2 and LMC6 remain unclear). The proportion of LMC5 and LMC4 in each tissue sample was compared with purity estimates derived from WES, RNA-seq and methylation data (Figure 6.19C). LMC5 was positively correlated with RNA-seq purity estimates in normal tissue, but not ccRCC tumours, suggesting LMC5 represents normal kidney. In tumours, LMC5 was negatively correlated with purity estimates from 'InfiniumPurify', a method which determines tumour purity as a function of contamination with normal tissue. In other words, tumour

samples which had a low purity (i.e. high contamination with normal kidney), had a higher content of LMC5. This is consistent with my previous observation from Figure 6.19B, that LMC5 represents normal kidney epithelial cells. LMC4 may represent kidney epithelium in ccRCC tumours. Indeed, the content of LMC4 in each sample was correlated with RNA-seq purity estimate for tumours, but not for normal kidney. In addition, LMC4 was also positively correlated with purity estimates derived by WES (Figure 6.19C).

Figure 6.19: Decomposition of bulk DNA methylation data into latent methylation components

See figure on following page

Panel A- Schematic explaining how bulk DNA methylation data are decomposed into cell-type-specific latent methylation components (LMCs) and the proportion of LMCs in each sample. Panel B- Heatmap demonstrating the 7 LMCs, and the proportion of each of these LMCs in the tissue samples. For each sample, the top annotation bar shows the pathology (ccRCC vs normal kidney), sample purity and patient ID (from which the sample was derived). Panel C - Heatmap demonstrating correlation values (cor) for the proportion of each LMC and purity values, for tissue samples. Purity values are derived using three independent methods: WES, RNA-seq ('ESIMATE') and DNA methylation ('InfiniumPurify'). 'ESTIMATE' is the only method which provides purity estimates for normal samples, therefore these are shown separately in the heatmap. Panel D- Heatmap demonstrating correlation values (cor) between each LMC and reference methylomes for various cell types. For each cell type, the accompanying code denotes the reference from which it was derived (from publicly available reference methylomes). In Panel C and D, a significant positive and negative correlation are shown in red and blue respectively. White denotes no significant correlation (p value > 0.05). Panel D- Cluster dendrogram obtained using methylomes for the seven LMCs and reference methylomes. LMC1 and LMC3 cluster with immune cells, on the left branch of the dendrogram.

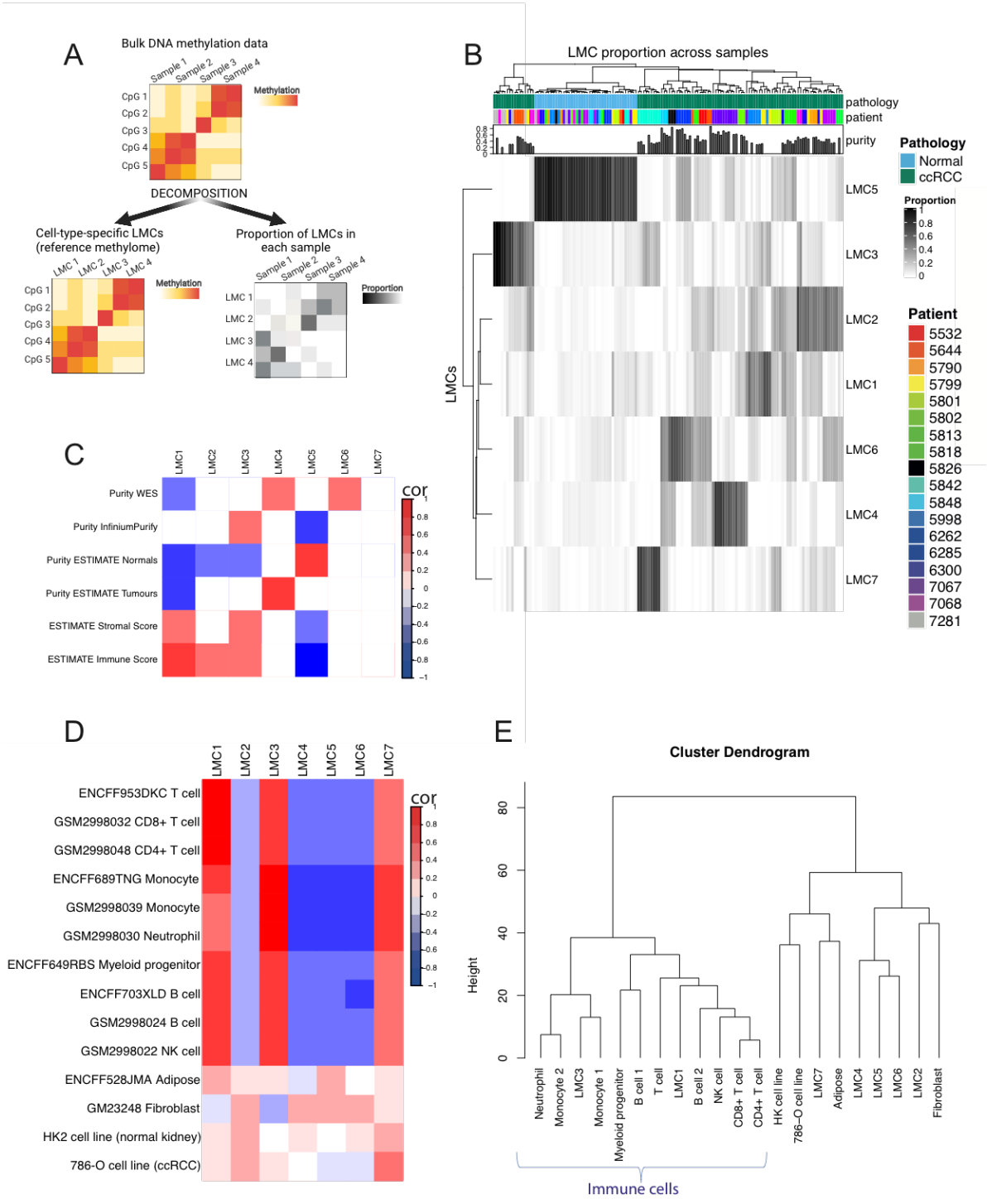


Figure 6.19: Decomposition of bulk DNA methylation data into latent methylation components

Table 6.7: Summary of results of latent methylation components analysis

Latent methylation components (LMCs) are shown. All correlations (cor) shown in the table had an adjusted p value < 0.001.

	Hypothesis	Correlation with purity estimates (RNA-seq, Epic-seq and WES)	Correlation with reference methylomes
LMC1	Represents immune cell, likely T cells	LMC1 was correlated with the immune score from RNA-seq (cor=0.60)	LMC1 was correlated with T cells (cor=0.83) and other immune cells to a lesser degree
LMC2	Cell type unclear		
LMC3	Represents immune cell, likely tumour associated macrophages	LMC3 was correlated with the immune score from RNA-seq (cor=0.50)	LMC3 was correlated with monocytes (cor = 0.96) and neutrophils (cor=0.95)
LMC4	Represents kidney epithelium in ccRCC tumours	LMC4 was correlated with RNA-seq purity estimates for tumour samples (cor = 0.62) and there was no correlation with purity estimates for normal kidney samples. LMC4 was also correlated with purity estimates from WES (cor=0.49).	
LMC5	Represents normal (non-cancerous) kidney epithelium	LMC5 was correlated with RNA-seq purity estimate in normal tissue (cor=0.72) and there was no correlation with purity estimates for ccRCC tumour samples. LMC5 was also negatively correlated with Infiniumpurify (cor=-0.71) which measures contamination with normal kidney.	
LMC6	Cell type unclear		
LMC7	Cell type unclear		

As already mentioned, to my knowledge there are no existing good-quality kidney tissue specific reference methylomes. Therefore, I obtained reference methylomes from multiple publicly available sources (see Methods section 4.6). LMC1 was highly positively correlated with the reference methylome for T cells derived from two separate sources (correlation = 0.83), suggesting it may indeed represent this cell type (Figure 6.19D and E). This hypothesis was corroborated by the fact that the proportion of LMC1 in each tissue sample correlated with the immune score calculated using RNA-seq (Table 6.7). LMC3 was also positively correlated with the immune score and had a strong correlation with the reference methylome for monocytes and neutrophils, meaning it may indeed represent tumour associated macrophages. Although both LMC1 and LMC3 are likely to represent immune cells, it is difficult to determine which type of cell based on this analysis alone. It is unclear exactly which cell type LMC7 may be, however it was noted to be weakly correlated with the 786-O methylomes (a ccRCC cell line) (Figure 6.19D). Interestingly, none of the LMCs had similar methylation levels to HK2 cells, a cell line representing normal kidney epithelium. This could be due to epigenetic changes that occur in immortalized cell lines [136]. I also sought to assess evidence of heterogeneity within a patient. As already noted in Figure 6.19B, tumour samples from the same patient tend to cluster together, suggesting higher heterogeneity between than within patients.

Subsequently, I focused my analysis on LMC1 and LMC3 as these two components are likely to represent immune cell infiltration and the literature suggests that immune composition may have prognostic potential in ccRCC. The proportion of LMC1 in each sample was evaluated to assess the relationship with clinical parameters. The proportion of LMC1 was significantly higher in ccRCC compared to normal kidney samples (median 9% versus 3%, adjusted p value= 0.004, Figure 6.20B). Interestingly, LMC1 was significantly higher in tumour samples with favourable clinical prognostic parameters (Figure 6.20A). Indeed, LMC1 was significantly higher in low grade disease (25% in grade 2 versus 5% in grades 3-4, adjusted p value=5.5e-09), lower stage (16% stage I-II versus 7% versus III-IV, adjusted p value= 0.028), lower Leibovich score (43% low risk versus 6% intermediate-high risk, adjusted p value=1.1e-06) and better prognosis (18% no recurrence vs 6% recurrence or metastases at diagnosis, adjusted p value= 0.0004). The fact that LMC1 was consistently higher for all four clinical parameters suggests this may be relevant and less likely to have been noted by chance. LMC3 was not related to any clinical parameter. In summary, evaluating DNA methylation data demonstrated that LMC1 content (which is hypothesized to represent immune cell infiltration) was higher in tumour samples relative to normal, and within tumour samples, higher levels were associated with a better prognosis.

An orthogonal analysis was performed using RNA-seq data to calculate an immune score for each sample, and this was noted to follow a similar pattern as was noted for LMC1 (Figure 6.20C and D). Indeed, whilst the immune score was significantly higher in ccRCC compared to normal kidney (median score 1725 vs -636, adjusted p value = 4.0e-10), a higher immune score was also associated with more favourable prognostic factors (lower stage and grade, adjusted p value <0.05 for all comparisons). It was not possible to assess the impact of Leibovich score and recurrence status separately as the same 4 samples which were stage I-II disease were also the ones who had a low risk Leibovich score and did not have a recurrence. The immune score was also negatively correlated with tumour size (correlation= -0.61, p value = 0.0005). However, these results are limited by the small number of samples in the favourable prognostic group and ideally larger sample sizes would be available. The immune score was relatively similar in most multi-region tumour samples from the same patient, except for patient 6300 which had higher variability (Figure 6.20D). Next, I used 'CIBERSORTx' [205] to deconvolute bulk RNA-seq data from my samples to try to understand the underlying immune cell types which may contribute to LMC1 and the immune score pattern noted (Figure 6.20F and G). 'CIBERSORTx' estimates the proportion of different immune cells that are present in a sample (hence the total sum is 100% of immune cells, see Figure 6.20F). There was a significantly higher proportion of M2 macrophages and CD8+ T cells in ccRCC compared to normal tissue (median 8% vs 2%, adjusted p value= 0.004 and median 44% vs 32%, adjusted p value = 0.001 respectively; Figure 6.20G). Unfortunately, the small sample size and large number of immune cell types precluded an evaluation of cell types by clinical parameters.

In summary, this section has two main findings. First, I demonstrated that a number of DMCs which are called as significantly different in ccRCC vs normal kidney may actually be markers of immune cells rather than tumour. This needs to be taken into consideration when applying the DMCs as biomarkers for different applications (e.g. diagnostic vs prognostic setting, plasma ctDNA vs tissue markers). Secondly, I showed that bulk DNA methylation profiles from my tissue samples can be deconvoluted into 7 cell types, where LMC4 and LMC5 are likely to represent normal kidney and tumour respectively, whereas LCM1 and LMC3 are likely to represent immune cells. The proportion of LMC1 in tumour samples is consistently associated with favourable prognostic clinical parameters (lower grade, stage, tumour size and recurrence), which warrants further investigation and validation in larger datasets.

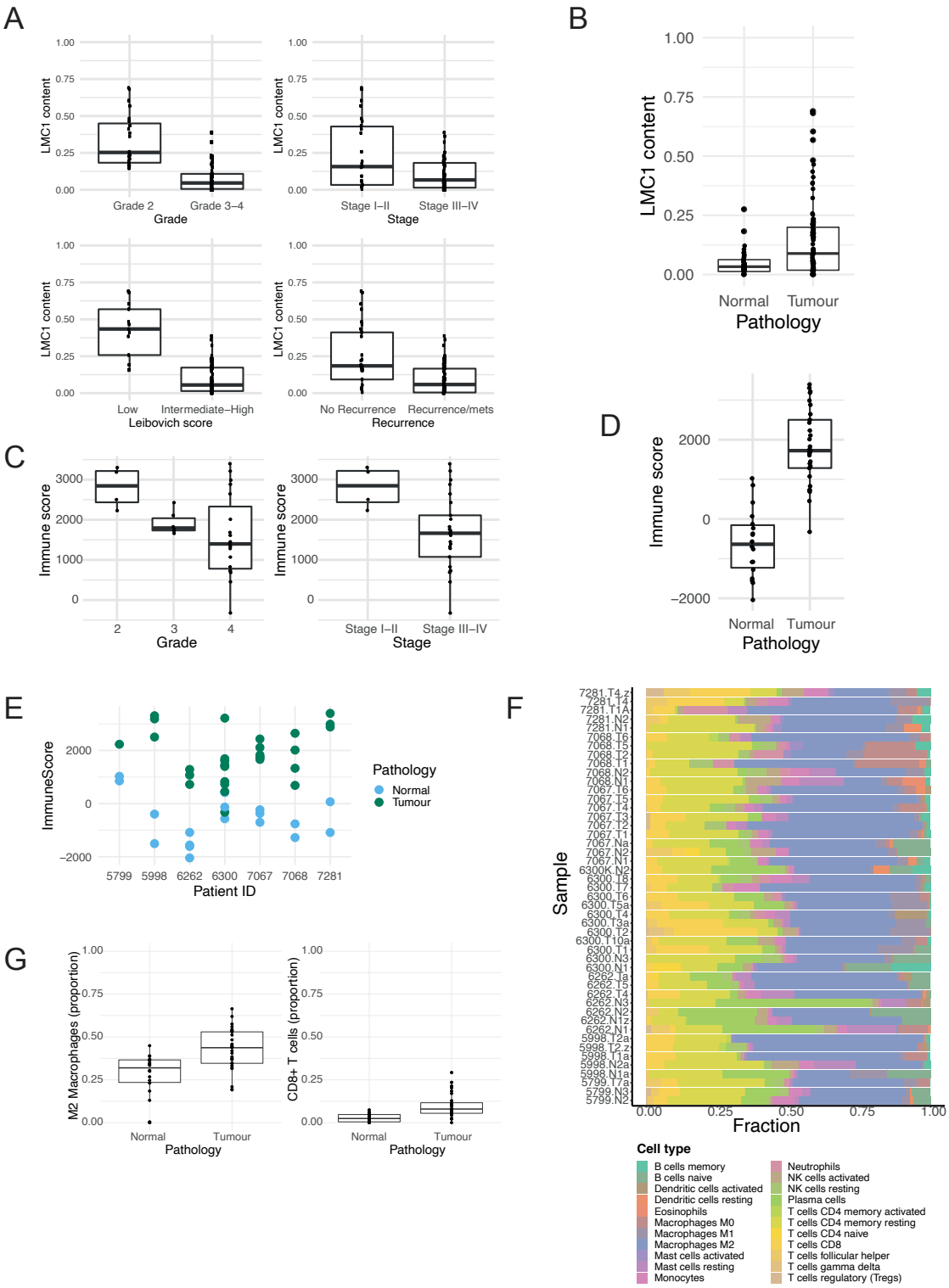


Figure 6.20: Immune cell components in tumour and normal samples, by clinical parameters

Panel A and B- Boxplots depicting the LMC1 (latent methylation component 1) content obtained from methylation deconvolution analysis. LMC1 component in tumour samples is shown by grade, stage, Leibovich score and recurrence status (Panel A). LMC1 content is also shown in ccRCC vs normal kidney (Panel B). Panel C, D and E- Boxplots depicting the immune score obtained from RNA-seq analysis (using 'ESTIMATE') for tumour samples, by grade and stage (Panel C). Results for tumour vs normal samples are shown overall (Panel D) and by patient (Panel E). Panel F and G- Results of immune cell decomposition from RNA-seq analysis (using 'CIBERTSORTx'), for each patient (F), and in tumour versus normal samples (G).

6.3.5.2 Exploring Homogeneously vs Heterogeneously methylated markers

6.3.5.2.1 Rationale

I aimed to explore methylation heterogeneity to gain an insight into the timing of methylation changes and thus improve our ability to select clinically useful diagnostic and prognostic markers. I therefore evaluated homogeneously and heterogeneously methylated CpGs, as these may be useful diagnostic and prognostic markers respectively (Figure 6.21A). CpGs that are differentially methylated in ccRCC versus normal tissue and demonstrate homogeneous methylation patterns amongst tumour samples (i.e. low ITH) are postulated to be clonal, early events in tumorigenesis [324]. This idea is similar to the concept that gene mutations that are present in all multi-region tumour samples from a patient are clonal, early events (such as *VHL*, *PBRM1*, *SETD2* etc) whereas mutations which demonstrate high ITH are subclonal, late events in tumour evolution [294]. Selecting homogeneously methylated DMCs would be key in a diagnostic setting. For example, when selecting methylation markers which can be applied to renal tumour biopsies or liquid biopsies to differentiate pathological subtypes of tumours (as described in Chapter 5). In a tumour biopsy setting, the ideal markers would have homogeneous methylation patterns within a patient and between patients in order to provide consistent diagnoses without being limited by ITH. In the case of liquid biopsies (i.e. plasma or urine cell free DNA analysis), it is postulated that early/truncal events are more likely to be represented in the plasma/urine than subclonal changes. Conversely, investigating CpGs that demonstrate heterogeneous methylation amongst tumour samples (i.e. high ITH) could elucidate methylation changes associated with aggressive disease which confer a selective advantage. These markers are therefore more likely to be useful for prognostic applications (Figure 6.21A).

6.3.5.2.2 Definition

CpGs were defined as homogeneous and heterogeneous as described by Hao et al in a seminal analysis of ITH in oesophageal cancer [220] (Figure 6.21B; Methods section 4.7.4). In summary, first CpGs were identified that were differentially methylated in tumour versus normal kidney for each patient (i.e. $\geq$25% methylation difference on average between tumour and normal samples within a patient). Subsequently, these CpGs were subdivided into homogeneous markers (i.e. methylation levels similar in tumour samples from the same patient) and heterogeneous markers (methylation levels different in tumour samples from the same patient; Figure 6.21B). The choice of thresholds is discussed in the Methods section 4.7.4. Hao et al found that differing the choice of thresholds

produced broadly similar results [220] (just more or less stringent numbers of CpGs called), and this was also observed in my data. Subsequently, methylation markers that were present in at least 6 out of 18 patients (i.e. ≥33% of the whole cohort) were selected as being recurrent, and therefore more likely to be clinically significant.

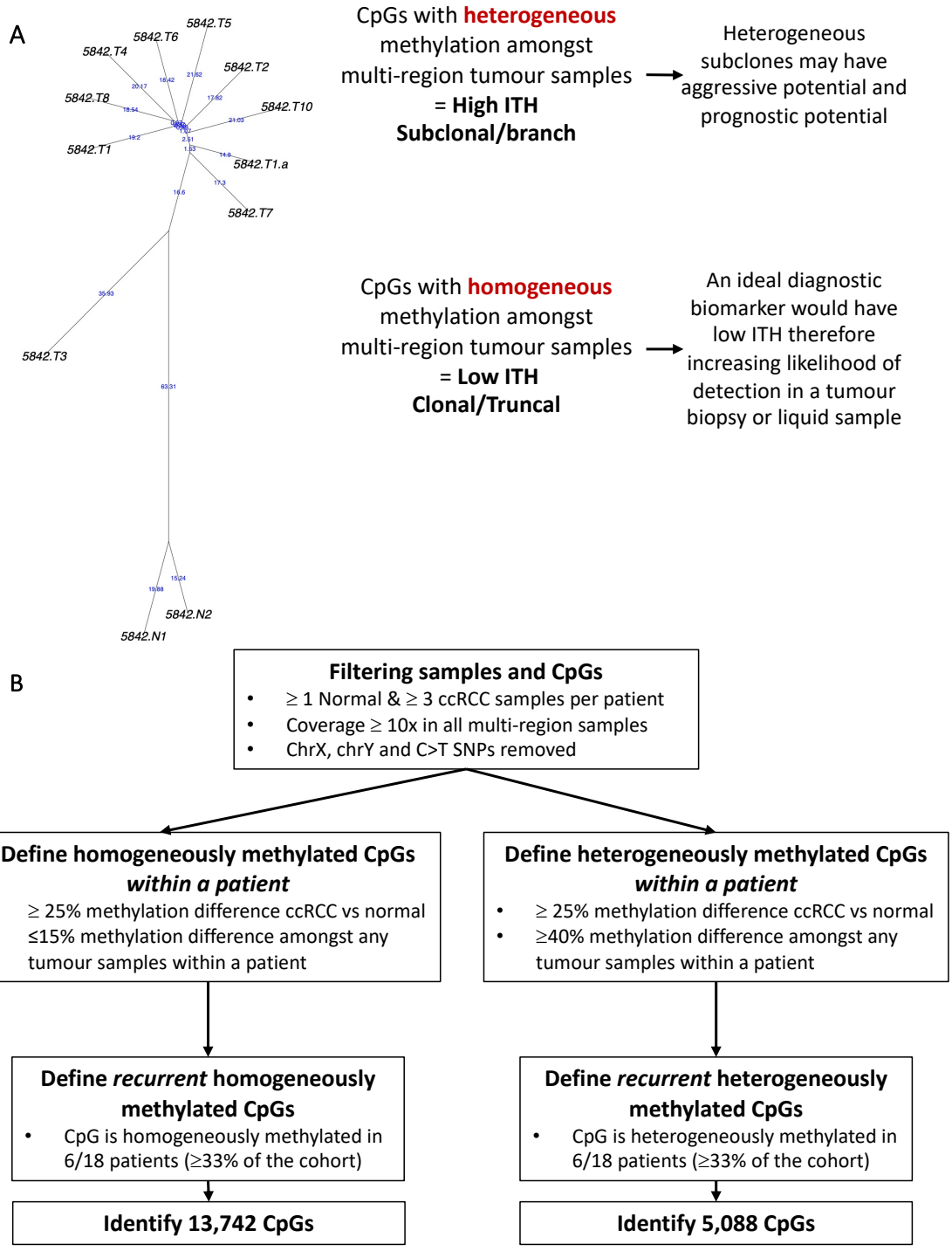


Figure 6.21: Rationale and definition of homogeneously and heterogeneously methylated CpGs

Rationale (Panel A) and definition (Panel B) of heterogeneously and homogeneously methylated CpGs [220].

6.3.5.2.3 Homogeneous vs Heterogenous methylation markers

Using the definition illustrated above, homogeneous (N=13,742) and heterogeneous methylation markers (N=5088) were identified. Figure 6.22A summarises the number of homogeneous and heterogeneous methylation markers by patient. A strong correlation was noted between the number of heterogeneous CpGs and the methylation APITH index for each patient (correlation coefficient = 0.65, p value= 0.003), confirming these are reliable measures of methylation heterogeneity (Figure 6.22B). The number of homogeneous and heterogeneous methylation markers was not related to any clinical characteristics such as tumour size, grade, stage or recurrence status (p value >0.05 for all comparisons).

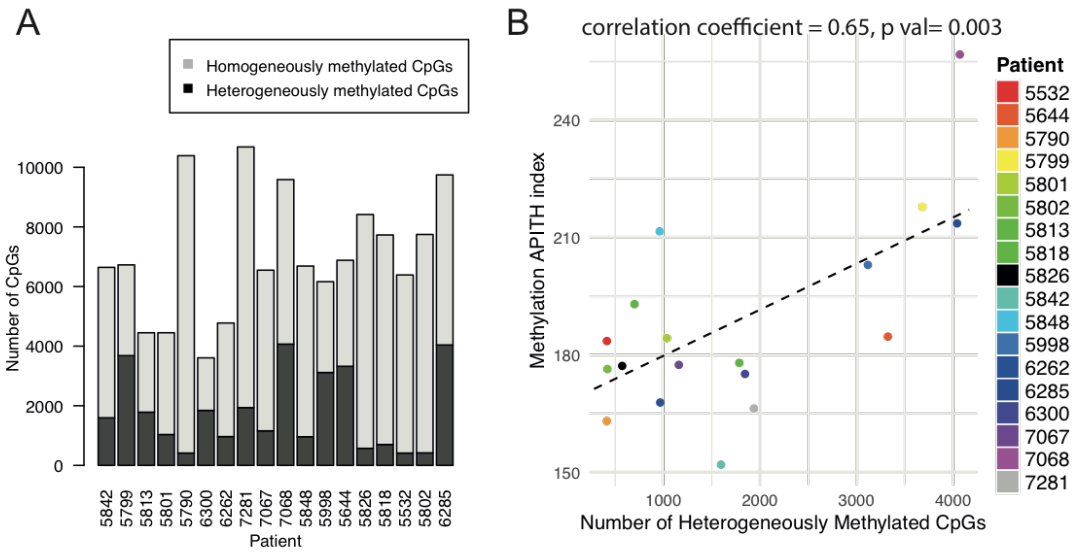


Figure 6.22: Homogeneously and heterogeneously methylated CpGs, by patient

Panel A- Barplot summarising the number of homogeneously and heterogeneously methylated CpGs, for each patient. Panel C- Scatterplot showing the number of heterogeneously methylated CpGs and methylation Average Pairwise intra-tumoral heterogeneity (APITH) index by patient. There was a strong positive correlation between the two (Pearson correlation coefficient = 0.65).

First the analysis focused on homogeneously methylated CpGs, which were visualised in a heatmap, confirming low inter and intra tumour heterogeneity at these sites (Figure 6.23A). Normal kidney and tumour samples are clearly clustered in two separate branches of the dendrogram. There is evident homogeneous methylation amongst tumour samples derived from the same patient, and between patients. The heatmap confirmed that the methods used identified homogeneous methylation markers as expected. Next, these homogeneous markers were validated in two independent cohorts, 71 ccRCC vs normal kidney samples analysed by Epic-seq and TCGA ccRCC vs

normal samples (Figure 6.23B and C respectively). In both cohorts it is evident that the CpGs clearly distinguish ccRCC from normal tissue and there is relative methylation homogeneity amongst ccRCC samples from different patients. To explore these CpGs further, I performed gene set enrichment and gene ontology analysis. Homogeneously methylated CpGs were enriched at the promoter region of genes that are known to be associated with ccRCC; including signalling by receptor tyrosine kinase and *VEGF*, cancer cell motility, cell cycle progression (integrin pathway) and interferon signalling (Figure 6.23D). An over-representation test for biological processes highlighted genes involved in 'regulation of cell migration and adhesion' and 'nephron tubule epithelial cell differentiation', including *GATA3*, *STAT1*, *LIF*, *WWTR1*, *MEF2C*, *MTSS1* (highlighted in Chapter 5). These methylation changes are likely to be truncal/early events in RCC tumorigenesis. Interestingly, my analysis identified homogeneous methylation at genes involved in interferon signalling pathways (Figure 6.23D). Interferon signalling has previously been found to be a feature of ccRCC and to be associated with *VHL*, *BAP1* and *SETD2* mutations, which are early events in tumorigenesis [133]. Studies have shown that *SETD2* mediates the methylation of the *STAT1* gene, which in turn mediates interferon gamma signalling [325]. *STAT1* can also be regulated by hypoxia and *VHL* inactivation in ccRCC, and is regarded as a tumour suppressor [326]. Interferon gamma signalling is involved in antigen presentation and the cancer immune response, including regulating *PDL1* expression in malignant cells [133]. In vitro experiments show that treatment of ccRCC cell lines with a demethylating agent (5-aza-2'-deoxycitidine) leads to increased interferon gamma gene expression and interferon induced apoptosis [103]. This is particularly noteworthy because it raises the possibility of ccRCC patient treatment with demethylating agents to augment the response to immunotherapies [103]. Next, I identified genes that had >10 homogeneously methylated CpGs in their promoter region, as a set of high confidence genes that are likely to be truncal (i.e. likely to have occurred early in cancer evolution). These genes were: *SLC16A3* (22 CpGs in the promoter region), *EHBP1L1* (20 CpGs), *SEPTIN9* (17 CpGs), *SLC16A1* (12 CpGs), *RIN1* (11 CpGs), *NDUFA4L2* (11 CpGs) and *APOBEC3D* (11 CpGs). Importantly, *SLC16A3* was also found to have homogeneous methylation within a read in ccRCC samples (i.e. low epipolymorphism values), as well as promoter hypomethylation and increased gene expression (see section 6.3.4.4).

GSEA demonstrated that heterogeneously methylated CpGs (i.e. high ITH) were enriched for the Rho GTPase cycle and regulatory signals, as well as neutrophil degranulation (Figure 6.22D). Studies have shown that increased neutrophil infiltration is associated with worse survival in ccRCC [327]. Neutrophil associated genes could either be a marker of aggressive disease (i.e. leading to increased inflammation) or an artefact of low tumour purity (i.e. if there is a high immune cell content in the

tumour sample analysed). The latter is a recognised limitation of methylation analysis of bulk tissue, which I attempted to overcome in my deconvolution analysis in section 6.3.5.1 and by analysing only high purity samples (see section 6.3.5.2.4 below).

6.3.5.2.4 Homogeneous vs Heterogenous methylation markers in samples with higher purity

To avoid low purity from confounding the analysis of informative CpGs, I repeated the analysis in section 6.3.5.2.3, excluding low purity samples. Samples were therefore selected if purity (as measured by WES or RNA-seq) was above the median (0.44). Patients were included if 3 or more tumour samples were available (to enable an analysis of heterogeneous and homogeneously methylated markers). Overall multi-region samples with higher purity were available for 5 patients (patients 6300, 6262, 5848, 5644, 6285). An analysis of homogeneous and heterogeneously methylated CpGs was subsequently performed.

Limiting the analysis to 5 patients with high purity samples identified 9955 CpGs that are homogeneously methylated in at least 3 out of 5 patients. GSEA was performed and this once again highlighted pathways associated with Interferon gamma signalling and focal adhesion (34/531 vs 201/8093, adjusted p= 7.47e-05). This included growth factors (*VEGFA* and epidermal growth factor *EGF*), several proto-oncogenes (*SRC*) and tyrosine kinase genes (*RAP1A*, *PTK2*, *FYN*, *FLT4*), genes encoding for adhesion molecules such as integrins and laminins, and regulators of apoptosis (*BCL2*). These are the same genes which were identified as homogenously methylated CpGs in section 6.3.5.2.3. These would therefore represent good diagnostic biomarker choices.

CpGs that are heterogeneously methylated in at least 3 out of 5 patients were identified (N= 7021), annotated to the nearest gene and those in the promoter region were selected for GSEA. This highlighted the Wnt signaling pathway (23/501 vs 166/8093, adjusted p value 2.07e-02), including *SERPINF1*, *DKK2*, *SFRP1*, *CCN4*, *SMAD3*, *CCND1*, *WNT7B*. Additionally, number of genes known to be implicated in ccRCC were also identified including *ZSCAN18* and *SHMT2*. A number of these genes (*DKK2*, *SFRP1*, *CCND1, SERPINF1*, *SMAD3*, *SHMT2*) have been shown to be associated with prognosis and implicated in tumour progression in ccRCC, suggesting that these subclonal changes may be markers of aggressive disease [328-332]. Interestingly, a systematic review of prognostic methylation markers identified conflicting results for *SFRP1* [67]. Whilst two studies found promoter methylation was associated with poor survival [123], one study demonstrated this was associated

with more favourable survival [333]. My work has suggested heterogeneous methylation patterns at *SFRP1*, and this could provide some explanation for the conflicting results and difficulties validating prognostic markers in ccRCC.

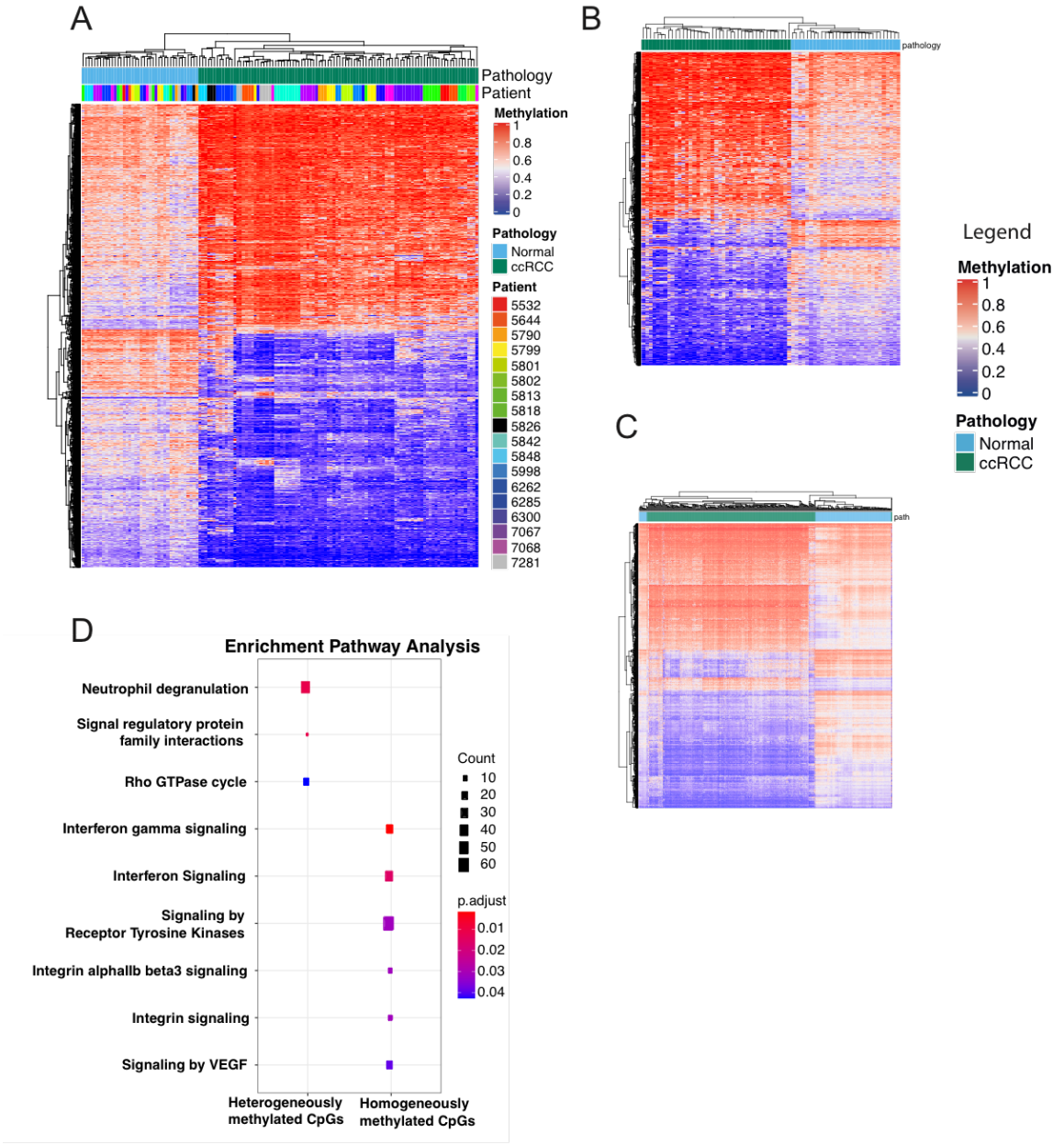


Figure 6.23: Homogeneously and heterogeneously methylated CpGs

Panel A- Heatmap illustrating homogeneously methylated CpGs in the discovery cohort (N=135 multi-region samples from 18 patients). This confirms low inter and intra tumour heterogeneity at these CpGs. Panel B and C- Heatmap illustrating homogeneously methylated CpGs in the two validation cohorts, which consists of an independent cohort of 71 ccRCC and normal samples (Panel B) and TCGA samples (Panel C). Panel D- Gene set enrichment analysis of homogeneously and heterogeneously methylated CpGs.

6.4 Discussion and future direction

In summary, the main findings of this chapter are enumerated below. The following discussion reflects on what my analysis adds to our current understanding of ccRCC disease biology and potential clinical implications of these findings.

1) Epigenetic inter tumoral heterogeneity dominated over intra tumoral heterogeneity in most cases (i.e. there was more heterogeneity between than within patients) (section 6.3.2).
2) Methylation and SCNA ITH were compared by evaluating the APITH index and by comparing phylo-epigenetic with phylogenetic trees. There were no associations between ITH and clinical nor prognostic parameters, though this may be due to the limited sample size (section 6.3.3).
3) Although disordered methylation is believed to be a stochastic process, there was evidence of significant differential epipolymorphism between ccRCC and normal kidney at the promoter region of genes known to be implicated in kidney cancer. This finding was confirmed in an independent cohort of ccRCC patients (section 6.3.4).
4) I identified a select number of genes in which gene expression was associated with promoter epipolymorphism, after adjusting for average methylation levels. This suggests that epipolymorphism may contribute to modulating gene expression independently of average methylation levels at the promoter (section 6.3.4.4).
5) Evaluating DMCs which were called as significantly different in ccRCC vs normal kidney suggested that some DMCs represent contamination with immune cells rather than tumour intrinsic changes. This is a useful finding that needs to be considered when performing future biomarker studies (section 6.3.5.1).
6) Bulk DNA methylation profiles from my tissue samples were deconvoluted into 7 cell types, where LMC4 and LMC5 are likely to represent normal kidney and tumour respectively, whereas LCM1 and LMC3 are likely to represent immune cells. The proportion of LMC1 in tumour samples was consistently associated with favourable prognostic clinical parameters (lower grade, stage, tumour size and prognosis) (section 6.3.5.1.1). This novel finding warrants further investigation.
7) I identified methylation markers that are homogeneously and heterogeneously methylated within a patient and recurrent across the cohort of patients; and I reflect on their utility as diagnostic and prognostic markers (section 6.3.5.2). In particular, homogeneously methylated markers would be useful in a diagnostic setting.

6.4.1 Heterogeneity between patients, within patients and within a sample

The first section of the chapter explored heterogeneity between patients, within a patient and within a sample. In keeping with the findings from Chapter 5, I observed greater heterogeneity between tumour samples than between normal samples. Overall, there appeared to be more heterogeneity between different patients than within samples from the same patient. This was consistent with three available studies in the literature, which suggested relative methylation homogeneity [127, 135, 136]. Studies evaluating colorectal and lung cancer also found that heterogeneity between patients was dominant over heterogeneity within a patient [209, 334]. This observation suggests that many methylation changes are early events in tumorigenesis, which may provide added benefit as diagnostic markers, over highly heterogeneous alterations, such as mutations. It was noted that some patients displayed more ITH than others (e.g. patient 7068 had the highest ITH), however no association was found with clinical factors. This may be due to the relatively small sample size (N=18) or due to limited or incomplete clinical follow-up. Renal cancer registries suggest that approximately 80% of patients who develop a recurrence will do so within 5 years [335]. My cohort was relatively mature (all patients had $\geq$4 years of follow-up), although some patients may develop a late recurrence beyond this period. Furthermore, only the presence of recurrence, rather than the timing of recurrence was known. Future studies could evaluate time to recurrence and focus on larger sample sizes. Alternatively, it may be that the relationship between methylation ITH and prognosis is not linear, as is the case for mutation ITH. Indeed, Turajlic et al noted that whilst ccRCC patients with low mutational ITH have attenuated progression, patients with both low mutational ITH and high chromosomal instability index have rapid progression [87].

My analysis compared phylogenetic trees derived from copy number data and phylo-epigenetic trees derived from methylation data to glean information regarding tumour evolution. This is the first such analysis to be performed in ccRCC. I demonstrated that whilst phylo-epigenetic and phylogenetic trees were very similar in some patients (e.g. patients 5532 and 7067), they were very different in other individuals (e.g. patients 5644, 5813, 6285). Unfortunately, data were only available for 8 patients and once again no clinical or prognostic correlates were identified. Studies in prostate cancer and glioma suggest that similarities between methylation and SCNA phylogenies are to be expected [324]. Future work in ccRCC samples could involve assessing mutation calls from WES and creating phylogenetic trees based on these data. This would allow a comparison against SCNA and methylation, to compare evolutionary patterns and evaluate co-occurrence of genetic and epigenetic changes (e.g. convergence on similar driver gene pathways). Another challenge in comparing genetic and epigenetic phylogenies is the relative lack of phylo-epigenetic inference

methods, which I aim to address in the future. I am currently planning further methodological research into bioinformatic approaches to model phylo-epigenetic evolution. For this thesis, I created phylo-epigenetic trees using the Euclidean distance and ordinary least squares minimum evolution algorithm with the 'Ape' package, an approach widely used in the literature [212, 213]. A potential limitation of the method is that it treats individual cytosines as independent events which may be methylated or unmethylated, disregarding co-methylation in favour of a simplified model [107]. To address this, I am collaborating with Victoria Dombrowe and bioinformaticians at Dr Schwarz's group, to develop a novel method to create phylo-epigenetic trees. The method, which will be named 'MYTHICC', involves summarising methylation data into three discrete states (unmethylated, partially methylation and methylated) and segmenting the genome based on these methylation blocks. Subsequently, blocks are compared between samples to evaluate transitions between the three methylation states and a minimum event distance is used to infer evolutionary trees. Future work will involve continuing the method development and comparing trees created using the three techniques (i.e. Methylation Euclidean distance, MYTHICC and MEDICC2 using SCNA) to evaluate whether segmentation provides more accurate tree topology compared to existing methods. Importantly, my analysis demonstrated that both tumour and normal tissue are characterised by disordered methylation (i.e. evidence of differential epipolymorphism) and this information must be taken into account whilst defining methylation blocks. A potential limitation in my analysis is that multi-region samples were spatially but not temporally separated. Such spatially and/or temporally separated samples could include: tumour thrombi obtained during IVC thrombectomy, biopsies from local recurrences, biopsy or metastasectomy specimens from distant metastases (e.g. adrenalectomy or lung metastasectomy samples from Papworth hospital). In future, assessing methylation in these samples could improve our understanding of tumour evolution through time, help identify aggressive subclones and potential actionable targets. In summary, I performed a comprehensive analysis of heterogeneity between patients and within a patient, though unfortunately no clinical correlates were identified.

My analysis represents the first evaluation of epipolymorphism, a measure of methylation heterogeneity within a sample, in ccRCC. A strength of my work is the use of sequence level data using Epic-seq, which allows biological discovery not amenable to array-based methods (such as 450k array). My main finding was that differential epipolymorphism was noted between ccRCC and normal tissue, in the promoter region of genes which are known to be associated with kidney cancer. This observation was identified in a primary cohort of samples (i.e. 135 multi-region samples) and externally validated in a separate cohort of 71 non multi-region samples. Although

gains in epipolymorphism are believed to be a stochastic process, these results suggest that disordered methylation may accumulate in functionally relevant loci which are known to contribute to tumorigenesis in ccRCC. A strength of this work is the external validation in an independent cohort of patients and the exploration in cell line data (to remove the confounding effect of tumour purity). However, DNA methylation changes may occur in cell lines secondary to immortalization and growth in-vitro [136, 305], meaning they may not represent the ideal model system. Furthermore, I evaluated methylation and RNA-seq data in matched tissue samples and demonstrated that differential epipolymorphism in the gene promoter was a predictor of gene expression. Some genes were characterised by increased disordered methylation in ccRCC (e.g. *DPP6*), whereas others were characterised by more ordered methylation in ccRCC (e.g. *SLC16A3* and *UBE2D2*) compared to normal tissue, and these changes were associated with gene expression. The association between promoter epipolymorphism and gene expression was negative in most cases, in keeping with transcriptional silencing. A limitation is that methylation and epipolymorphism are by definition correlated and thus their effects on gene expression are difficult to disentangle. Further, my analysis did not determine the causal relationship between gene expression and epipolymorphism. Despite these caveats, these novel results suggest that epipolymorphism within a gene promoter may be an independent regulator of gene expression in addition to overall methylation in ccRCC.

Future work should focus on understanding the mechanism underlying differential epipolymorphism in ccRCC versus normal kidney and whether these changes predispose to cancer progression. Hu et al studied methylome evolution by evaluating temporally distinct samples along the lung carcinogenesis pathway including lung precancers, preinvasive, and early invasive lung adenocarcinomas [336]. The study authors demonstrated that there were higher levels of epipolymorphism as cancer progressed from earlier to later stage disease and suggested that DNA methylation had undergone evolutionary drift. It may be that there is an evolutionary drift from normal tissue to ccRCC tumour tissue. Further methylation analysis of temporal samples in ccRCC would help elucidate this. In addition, a comparison of epipolymorphism in normal kidney from patients with ccRCC against normal kidney from patients without tumours would enable us to determine if the disordered methylation seen in normal kidney is secondary to premalignant processes or a feature of normal renal epithelium. For example, our collaborators (Dr Thomas Mitchell and colleagues) have access to kidneys donated for transplantation and subsequently deemed unsuitable and therefore used for research.

6.4.2 Clinical utility of heterogeneity analysis

The second section of this chapter focused on extracting useful learning points regarding biomarker selection, which may be gleaned by evaluating methylation heterogeneity. I identified CpGs which have homogeneous and heterogeneous methylation in multi-region samples from each individual patient and are highly recurrent across patients in the cohort. Homogeneous CpGs were noted in key pathways which are known to be involved in RCC (such as VEGF and tyrosine kinase signalling, cell motility and cell cycle progression), suggesting these are clonal, early steps in tumorigenesis. These could represent ideal diagnostic biomarkers for tissue biopsies or liquid samples. In Chapter 5, we develop a machine learning model that uses methylation data to predict pathological subtypes of renal tumours. In the vast majority of cases (90%), sampling multiple regions from the same tumour yielded consistent classification results. This is in keeping with the notion that there are early recurrent sets of methylation changes.

In addition, I identified CpGs which have heterogeneous methylation. Heterogeneous methylation patterns may be secondary to subclonal tumour changes or due to admixtures of cell types due to varying sample purity. Though it may not be possible to disentangle these two processes completely, they highlight informative considerations which must be taken into account in future biomarker studies and clinical applications. I attempted to address this challenge by assessing methylation data after adjustment for purity using 'Infiniumpurify', however this method introduced noise into the methylation matrix and did not provide added benefit. I also repeated the analysis excluding low purity samples and this enabled the identification of heterogeneously methylated CpGs which are likely to be subclonal events as they coincide with putative prognostic markers in ccRCC (such as *SFRP1*, *DKK2* and *CCND1*) [67]. These methylation changes may be late events in tumorigenesis and could represent markers of tumour aggressiveness. Furthermore, heterogeneous methylation within these gene promoters could explain difficulties validating these prognostic markers noted in the literature [67].

I also evaluated DMCs that distinguish ccRCC and normal tissue and demonstrated that a proportion of these markers (i.e. those with a high variance in tumour) may represent immune cell markers rather than tumour intrinsic changes. Any attempt to use these as diagnostic biomarkers in clinical practice would therefore be hampered by heterogeneity associated with sample purity. Furthermore, these would not be useful markers for plasma cell free DNA analysis due to the high contamination from gDNA from lysed immune cells. This demonstrates the added benefit of the present analysis along-side existing methods of DMC calling. Simply removing samples with a low

tumour purity below a certain threshold could in theory overcome this challenge, however my results suggest that low purity samples may still provide clinically useful data. Another approach could be to select DMCs with a low variance amongst tumours, or to filter DMCs against the methylation levels in immune cells. The latter approach was undertaken in Chapter 7 whilst selecting biomarkers for cell free DNA analysis. Ultimately, future studies evaluating DMCs in ccRCC versus normal kidney should adjust for variance in tumour purity.

Furthermore, I performed reference-free deconvolution of bulk methylome data into seven latent methylation components (LMCs). One of these, LMC1, which is likely to represent immune cells, was noted to be associated with clinical parameters. LMC1 levels were higher in tumours which were lower grade, stage and did not develop a recurrence. However, the present analysis was unable to determine the exact cell type that LMC1 represents, though I hypothesized LMC1 represents a T cell population. Further work in larger independent cohorts is necessary to explore this interesting finding. ccRCC has been found to have a high immune cell infiltrate relative to other cancers, and in particular a high T cell infiltrate [299, 337]. Generally, presence of a high immune T cell infiltrate is a good prognostic marker in other solid cancers (i.e. increased CD8+ T cells are associated with better survival) and is considered a sign of anti-tumour adaptive immune response [338]. In ccRCC, studies initially showed the contrary: a high number of CD8+ T cells was associated with poor overall survival [339]. Recently, studies have focused on characterising T cell subtypes, demonstrating that CD8+ T cells may indeed be activated, inhibited or exhausted and this determines the prognostic impact [339]. Different T cell subpopulations are associated with different prognoses, due to their different roles (for example Tregs are immunosuppressive) [340]. Senbabaoglu et al [299] found that Th17 cells and $CD8^+$ T/Treg ratio were associated with improved survival, whereas Th2 and Tregs were associated with worse survival. Similarly, macrophage populations noted in renal cancers may be subdivided into pro-inflammatory M1 cells and anti-inflammatory M2 cells [228], highlighting the importance of characterising immune cell subpopulations. Although the immune composition in ccRCC is not a focus of my PhD project, this is a very interesting area for further research.

My findings on methylation heterogeneity have important implications for future biomarker studies. Heterogeneity between patients may hamper biomarker identification. For example, a systematic review of prognostic methylation markers in ccRCC demonstrated that *GREM1* was associated with patient survival in some studies, but not others [67, 341]. Van Vlodrop et al showed that the prevalence of hypermethylation in ccRCC patients ranged between 20 and 50% depending on the CpG island evaluated [342], and this could explain the conflicting results noted. The presence of

disordered methylation within an e-locus may contribute to difficulties in externally validating biomarkers. For example, one CpG may demonstrate hypermethylation relative to normal, whereas the adjacent CpG may demonstrate hypomethylation. Distinct methods (e.g. 450k array, RRBS and Epic-seq) may cover different CpGs within the promoter region of the same gene, therefore studies utilising different platforms may fail to show consistent results. These findings highlight the importance of evaluating (and reporting) CpG level data, rather than simply gene level data. Studies have shown that transcriptional silencing may be regulated by hypermethylation at a single CpG or a handful of sites (rather than requiring hypermethylation of an entire region) [343]. This could explain why epipolymorphism predicts gene expression, independently from average methylation, in my data. Disordered methylation per se could be a valuable biomarker of ccRCC in selected genes. The high level of methylation heterogeneity (epipolymorphism) within a locus will have profound implications for the design and interpretation of diagnostic assays (depending on the target CpGs analysed).

In summary, to the best of my knowledge, this analysis represents the most systematic characterization of ccRCC methylation heterogeneity. Although a handful of studies have evaluated methylation heterogeneity in ccRCC [127, 135, 136], my work is novel in that it compares methylation and SCNA heterogeneity, adopts a comprehensive approach and has the largest sample sizes to date. My analysis is the first to evaluate epipolymorphism using sequence level methylation data (Epic-seq) in kidney cancer, a method which is considered the most informative approach in the absence of single cell methylome sequencing [344], and provides added value over array-based methods such as 450k array. In conclusion, this chapter comprehensively characterises methylation heterogeneity in ccRCC tissue between patients, within a patient and within a sample. Importantly, I focus on how this analysis can help us select better biomarkers and how future research studies can avoid common pitfalls.

Chapter 7 DNA methylation in liquid samples

7.1 Brief introduction

Cell free DNA (cfDNA) has recently attracted substantial interest from the translational research community due to its potential role as a cancer biomarker [149]. My thesis has identified two main research questions of interest, namely improved diagnosis of small renal masses (SRMs) and improved risk-stratification in patients with non-metastatic RCC, both of which could be greatly enhanced by tumour derived cfDNA (ctDNA) detection. For example, ctDNA could be used as a non-invasive liquid biopsy (replacing renal biopsy) to aid diagnosis in patients with SRMs, therefore reducing overdiagnosis and overtreatment. There are, however, several challenges associated with ctDNA detection and for RCC in particular.

Levels of ctDNA vary between cancer types and early reports suggested that patients with ccRCC may have lower ctDNA levels compared to other malignancies [153, 155-157, 345]. For example, Zhang et al performed deep sequencing of cfDNA (targeting single nucleotide variants, insertions/deletions, copy number aberrations and chromosomal rearrangements) and detected ctDNA in ~56% of patients with renal cancer (63% in stage IV RCC). Evaluating >10,000 cfDNA samples, demonstrated that detection rates in RCC were lower than 22 other cancer types [157]. It was postulated that low detection rates may be partly explained by the use of a pan-cancer assay, however, RCC-specific mutational panels have also yielded disappointing results. Pal et al evaluated the Guardant360 assay (Guardant Health) in plasma cfDNA derived from 220 patients with metastatic RCC (mRCC) and reported detection rates of 79% [346]. The Guardant360 test detects single nucleotide variants (SNV), insertions/deletions (indels) and copy number amplifications in 73 genes known to be affected in RCC. The study by Pal et al identified a median of 1 genomic alteration per patient (IQR: 0-3), demonstrating how the low frequency of genomic changes may hamper detection even in advanced disease. Increasing the number of mutations targeted or targeting patient-specific mutations has been postulated as an alternative strategy to enhance sensitivity. Smith et al sequenced nephrectomy samples to identify patient-specific mutations and generated custom panels for ctDNA detection using the 'Integration of Variant Reads-Tailored Panel Sequencing' (INVAR-TAPAS) pipeline [158]. Despite adopting a sensitive, personalised approach to mutational analysis, detection rates were disappointing once again. Including patients of all ccRCC stages, ctDNA was only detected in ~50% of pre-operative samples. In a separate study by the same group, using the same INVAR-TAPAS pipeline yielded much lower ctDNA detection rates in patients with RCC (42%) compared to patients with melanoma (96%), glioma (75%), breast (100%) and non-

small cell lung cancer (63%) [347]. The largest mutational analysis of ctDNA in renal cancer to date (N=920 samples), demonstrated detection rates of 72% in mRCC, despite this being the group of patients with the most advanced disease burden and therefore highest likelihood of detection [348]. It is unclear whether ctDNA levels are truly lower in RCC compared to other malignancies (due to some yet undiscovered biological process) or whether detection is limited by extensive intra-tumoural genomic heterogeneity and the low frequency of recurrent mutations [349]. This has driven the search for alternative strategies.

One of the potential approaches to improve sensitivity is to increase the number of targets analysed at high sequencing coverage [154], for example by targeting methylation markers which are orders of magnitude more abundant than mutations and are frequently early (clonal/truncal) events. In other cancer types, this has proved a successful strategy. Chen et al developed PanSeer, evaluating methylation at 10,613 CpG sites across 477 genomic regions in plasma derived cfDNA to detect five cancer types [350]. The test achieved excellent accuracy in patients with known cancer. In addition, PanSeer identified ctDNA in 95% (95% CI: 89–98%) of asymptomatic individuals who were later diagnosed with cancer up to four years prior to the development of symptoms, demonstrating the potential for use as an early diagnostic test [350]. Similarly, a high profile biotech company (GRAIL) evaluated >100,000 informative methylation regions (including >1 million CpGs) to detect 12 cancer types in cancer patients and healthy controls [351]. Sensitivity for the 12 cancers (stages I-IV) was 77.9% (95% CI: 75.0% to 80.7%), increasing with more advanced disease stage [151]. However, once again, it was noted that sensitivity in RCC, using this pan-cancer assay, was much lower than other cancer subtypes [151,351]. I hypothesized that targeting RCC-specific methylation markers may provide superior results compared to pan-cancer markers. Thus far, perhaps the most promising results were reported by Nuzzo et al [159]. The authors evaluated ctDNA methylation in patients with RCC using 'cell-free methylated DNA immunoprecipitation and high-throughput sequencing' (cfMeDIP–seq). This method enabled detection of ctDNA in patients with RCC across the spectrum of disease severity, achieving an overall Receiver Operating Characteristic area under the curve (ROC AUC) of 0.99. In a head-to-head comparison, cfMeDIP-seq was significantly more sensitive than ctDNA mutational variant analysis in 34 patients with mRCC [352]. Whereas mutational analysis detected variants in 21% of mRCC patients, cfMeDIP-seq achieved 100% sensitivity with 88% specificity. This suggests methylation analysis of ctDNA in renal cancer may be a promising strategy warranting further investigation.

In summary, these initial reports suggest that methylation markers may enable superior detection of ctDNA in RCC compared to mutational analysis, however, low levels of ctDNA may pose a significant challenge. Proximal sampling, i.e. collecting cfDNA from fluids closer to the tumour, has been proposed as a method to increase detection of ctDNA in many cancer types, however this has not yet been trialled in RCC. I postulated that in patients undergoing renal biopsy, blood and fluid may be collected non-invasively from the biopsy site using a needle and syringe, and that this (referred to as 'post-biopsy fluid') represents a proximal sample. I hypothesized that this fluid may be enriched for ctDNA secondary to tumour puncture during biopsy (akin to fine needle aspiration) and that bleeding directly from the tumour may also contain higher levels of ctDNA than peripheral venous blood (a distal sample). I sought to explore this topic further, as this strategy could represent a useful adjunct to aid diagnosis in patients with SRMs undergoing renal biopsy where plasma ctDNA levels might be prohibitively low.

Therefore, in this chapter, I aimed to develop a methylation panel to detect ctDNA in ccRCC patients. In addition, I compared ctDNA detection rates using methylation versus mutational analysis, and evaluated proximal versus distal sampling strategies. This chapter represents an initial study to investigate the feasibility of this approach in a relatively small number of patients with a spectrum of disease severity. If successful, it is envisioned that this work may inform the design of a larger study in future.

7.2 Chapter aims

Overall, this chapter aimed to adapt Nimbus (Non-destructive Integration of Methylation to Boost Underlying Signals), a novel method for targeted methylation analysis established in our research group, to detect cfDNA derived from patients with and without ccRCC. The following aims were addressed:

1. Design a panel for targeted ctDNA methylation analysis by identifying DNA methylation markers that differentiate ccRCC from normal kidney in nephrectomy tissue samples
2. Assess ccRCC Nimbus panel performance using experimental quality control metrics
3. Determine ctDNA detection rates for ccRCC patients compared to cancer-free controls using Nimbus targeted methylation analysis of plasma cfDNA
4. Compare ctDNA detection rates in ccRCC patients using methylation (Nimbus) compared to mutational analysis (INVAR-TAPAS) and explore associations with clinical parameters
5. Explore proximal sampling as a method to increase ctDNA detection in RCC

7.3 Results

This chapter is divided into two main sections. The first part describes the Nimbus panel design and an initial evaluation of quality control metrics (sections 7.3.1-7.3.3). The second section covers Nimbus analysis in plasma cfDNA and proximal samples from patients with and without renal tumours (sections 7.3.4-7.3.6; Figure 7.1).

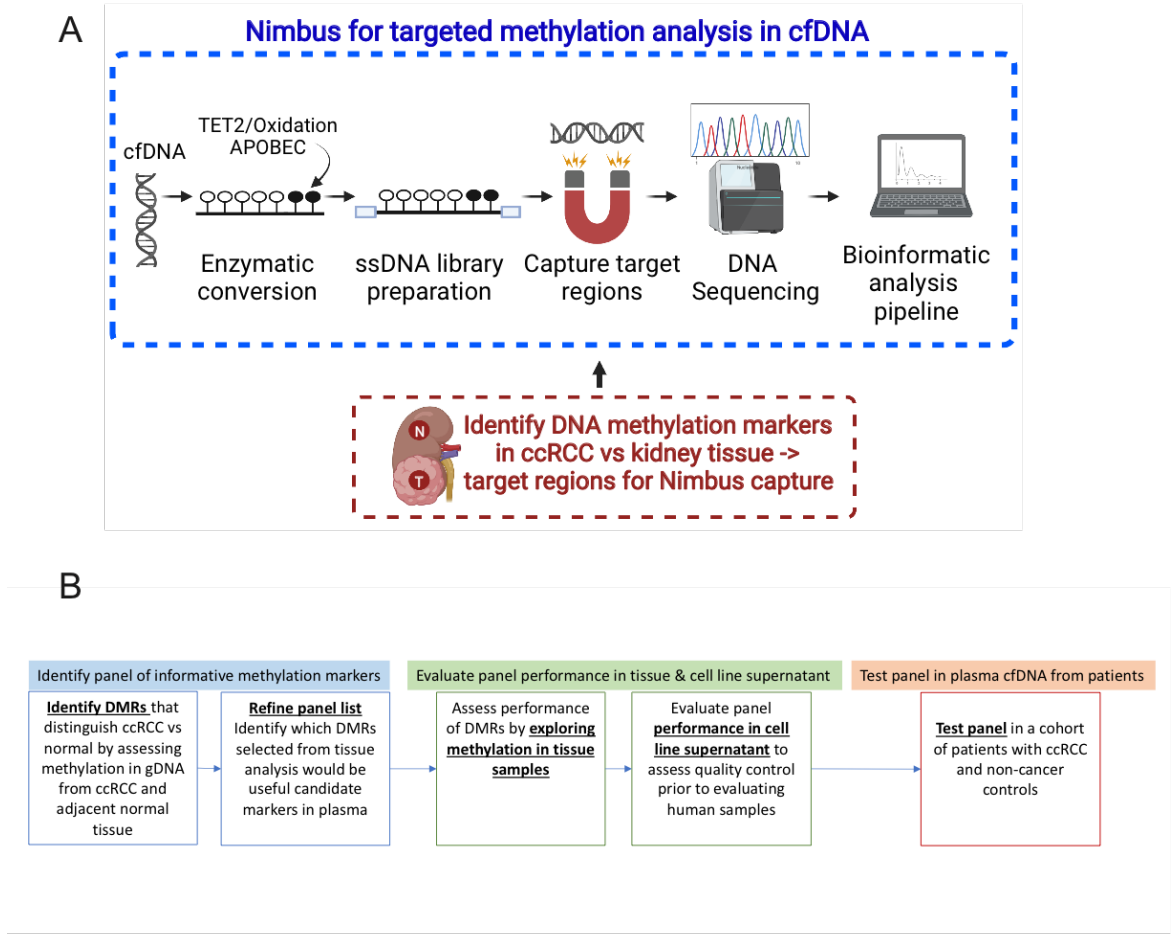


Figure 7.1: Overview of workflow for ctDNA methylation analysis methods

Panel A- Nimbus couples enzymatic conversion, single stranded (ssDNA) library preparation and targeted capture sequencing along with a bespoke bioinformatics pipeline to derive tumour-specific signals. Tissue analysis is performed to determine differentially methylated regions (DMRs) that can distinguish ccRCC from normal kidney and these markers are included in the Nimbus panel (i.e. regions of interest to be captured, see Panel B). Panel B- Methods used to identify informative methylation markers and evaluate performance of the panel. Genome-wide DNA methylation analysis was performed in tissue in order to identify a panel of DMRs that can distinguish ccRCC from normal kidney. These markers were applied to DNA isolated from tissue and cell line supernatant (to assess quality control metrics) prior to proceeding to human cfDNA analysis.

7.3.1 Nimbus: targeted DNA methylation analysis in cfDNA

Non-destructive Integration of Methylation to Boost Underlying Signals (Nimbus) is a novel, highly sensitive method for cfDNA analysis. Nimbus couples optimized cfDNA methylation library preparation with a sophisticated bioinformatics pipeline to increase ctDNA detection rates. Nimbus refers to both wet lab and dry lab experimental methods (Figure 7.1A). This workflow was developed by Drs Park and Lach (postdoctoral research associates in the Massie laboratory) and I played an active role in method development and optimization. In summary, enzymatic conversion is performed using the NEB Next Enzymatic Methyl-seq Kit (New England BioLabs). TET2 oxidises methylated cytosines, subsequently APOBEC deaminates unmethylated cytosines into uracils whilst sparing methylated cytosines due to the previous TET2 action. Enzymatic conversion is non-destructive, therefore reducing fragmentation and loss of cfDNA compared to biochemically harsh bisulphite conversion (Park et al, unpublished). Conversion is followed by optimized single stranded library preparation and targeted hybridization capture of thousands of methylation markers (Figure 7.1A). The method has previously been successfully applied to plasma ctDNA from prostate cancer patients, and I sought to adapt Nimbus to enable detection of ctDNA in ccRCC patient samples.

In order to adapt Nimbus to my desired application, I first sought to create a custom methylation marker panel specific to ccRCC (Figure 7.1A-B). I performed genome-wide DNA methylation analysis in tissue to identify a panel of differentially methylated regions (DMRs) that can distinguish ccRCC from normal kidney, and are therefore useful methylation markers for ctDNA analysis. Subsequently, I refined these tissue-derived DMRs to select those which were most informative in plasma cfDNA (as described in the Methods section 4.8.1). Capture probes were designed to target these DMRs. I applied this marker panel to DNA isolated from tissue and cell line supernatant (to assess quality control metrics) prior to proceeding to an analysis of human cfDNA.

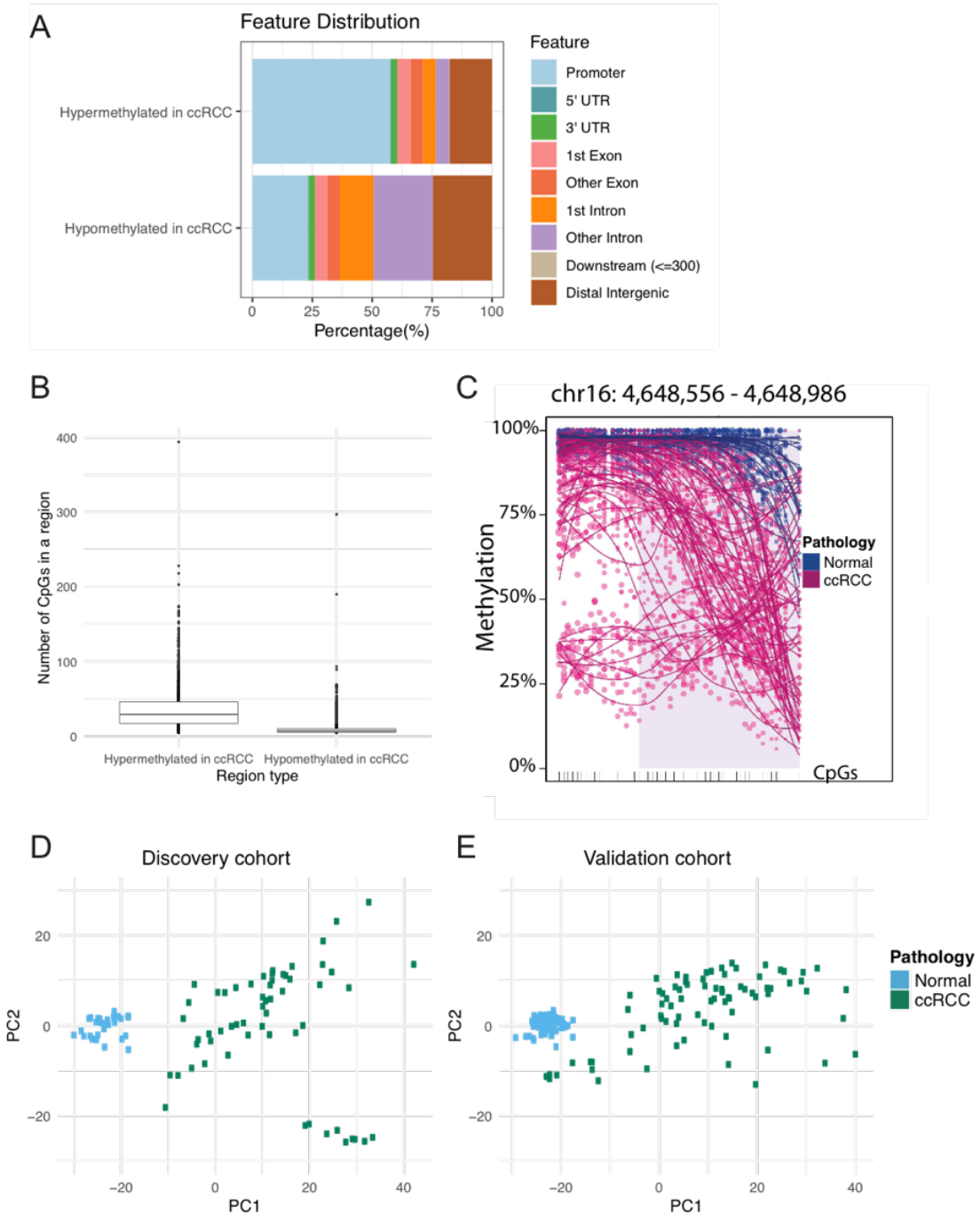


Figure 7.2: Tissue analysis in ccRCC vs normal kidney to determine methylation panel for Nimbus cfDNA analysis

Panel A- Annotation of differentially methylated regions (DMRs) derived from tissue analysis comparing ccRCC vs normal kidney. Hypermethylated DMRs are more commonly located in gene promoters, whereas hypomethylated DMRs are located in distal intergenic regions and introns. Panel B- Boxplots demonstrating the number of CpGs within a DMR, for hypermethylated and hypomethylated regions. Panel C- Locus plot of a representative hypomethylated DMR. Methylation levels (y axis) are shown for each CpG (x axis) within the region. There is clear hypomethylation in ccRCC (pink) relative to normal kidney (blue). Panel D and E- Principal component analysis (PCA) plot using data from the 5801 DMRs identified in tissue, in the discovery cohort (N=71 ccRCC and normal kidney samples), as well as an independent validation cohort (N=159 ccRCC and normal kidney samples). The discovery cohort refers to the original set of samples that the 5801 DMRs were identified in.

7.3.2 Selection of informative methylation marker panel in tissue to be used in Nimbus

I aimed to identify DNA methylation markers in tissue that differentiate ccRCC from normal kidney, which could be applied to detect ccRCC ctDNA in plasma. To allow testing of the Nimbus ctDNA methylation assay within the timelines for my PhD project, the Nimbus ccRCC targeted panel was designed using an initial dataset from my tissue analysis (N=75 samples from 21 patients). While this cohort was sufficient to identify a set of recurrent DNA methylation markers, future iterations could be extended to include the larger set of markers identified in Chapter 5 and Chapter 6.

Figure 7.1B summarises the workflow used to identify and evaluate the informative methylation marker panel for ccRCC. I generated sequence-level DNA methylation data using Epic-seq in fresh frozen tissue samples from patients undergoing curative or cytoreductive nephrectomy at Addenbrooke's Hospital (see Methods section 4.8.1). Multi-region samples (N=75) were obtained, consisting of renal tumour and adjacent normal tissue. This included 53 ccRCC tissue samples from 15 patients, and 22 normal kidney samples from 21 patients (15 from patients with ccRCC, 2 from patients with oncocytomas and 4 from patients with chRCC). Using the 'dmrseq' package in R, I identified 16,441 genomic regions that were differentially methylated in ccRCC versus normal kidney, at a q value of < 0.01 [188].

In order to identify which DMRs selected from tissue analysis would be useful candidate markers in blood plasma, data were obtained from 32 healthy (cancer-free) controls from a previously published study [221]. DMRs were selected if there was a >60% methylation difference between ccRCC tissue and healthy control cfDNA samples, to maximise chances of detecting a methylation difference. This analysis identified 5801 informative DMRs (2183 hypermethylated and 3618 hypomethylated) in ccRCC compared to normal kidney and healthy control cfDNA. The proportion of hyper- and hypomethylated DMRs is in keeping with results of Chapter 5, which demonstrated that indeed when comparing ccRCC versus normal tissue, there are approximately 1.6 times more hypomethylated than hypermethylated regions. As expected, the hypermethylated DMRs were more likely to be located in gene promoters, whereas hypomethylated DMRs were more commonly located in distal intergenic or intronic regions (Figure 7.2A). In hypermethylated regions there were a median of 29 CpGs (IQR: 17-46), whereas for hypomethylated regions there were a median of 7 CpGs (IQR: 5-10; Figure 7.2B). This is also in keeping with known data, since hypermethylation tends to occur at CpG islands, regions with a high CpG density.

I explored methylation levels at these 5801 DMRs in tissue samples to confirm that these were appropriate markers to take forward into subsequent analysis. Figure 7.2C shows a representative DMR included in the panel, demonstrating clear hypomethylation in ccRCC tissue compared to normal kidney, as expected. Methylation levels in normal kidney tissue (shown in blue) tend to have low variance, whilst there is a much greater spread of methylation levels in ccRCC (shown in pink). Next, I evaluated Epic-seq data in an independent cohort of ccRCC and normal kidney tissue samples to ensure results were reproducible and generalisable (N=159). I created a PCA plot using data from these 5801 DMRs in the discovery cohort (N=71), as well as the validation cohort (N=159). Figure 7.2D and E suggest that once again there was more variance in tumour than normal kidney, possibly reflecting tumour cellularity or differences in clonality of these markers between samples, in keeping with results from Chapter 5 and Chapter 6. In both the discovery and validation cohorts, the selected panel of CpGs differentiated tumour versus normal tissue (Figure 7.2D-E), suggesting these may be appropriate markers to take forward into cfDNA analysis.

7.3.3 Nimbus quality control metrics

As already alluded to, Nimbus is a novel approach developed in our research group and I played an active role in method development. Herein, I compare Nimbus to a commercially available method (Epic-seq) and evaluate quality control (QC) metrics in cell lines, to ensure adequate performance prior to proceeding to human cfDNA samples.

7.3.3.1 Comparison with established commercially available methods

In order to compare Nimbus against commercially available methods, I analysed a gDNA sample derived from human ccRCC tissue using both the established Epic-seq (Illumina) and our in-house Nimbus protocol. A tissue sample was selected for this comparison because the minimum input for Epic-seq is 500ng, precluding analysis of low input cfDNA samples. I focused my analysis on CpGs which achieved ≥10x minimum coverage using both methods (N= 129,797 CpGs). Correlation between Epic-seq and Nimbus was 0.94, and the Bland Altman plot demonstrates that the vast majority of CpGs have the same percentage methylation using the two methods (i.e. for the majority of CpGs the difference between values obtained using the two methods is zero; Figure 7.3A).

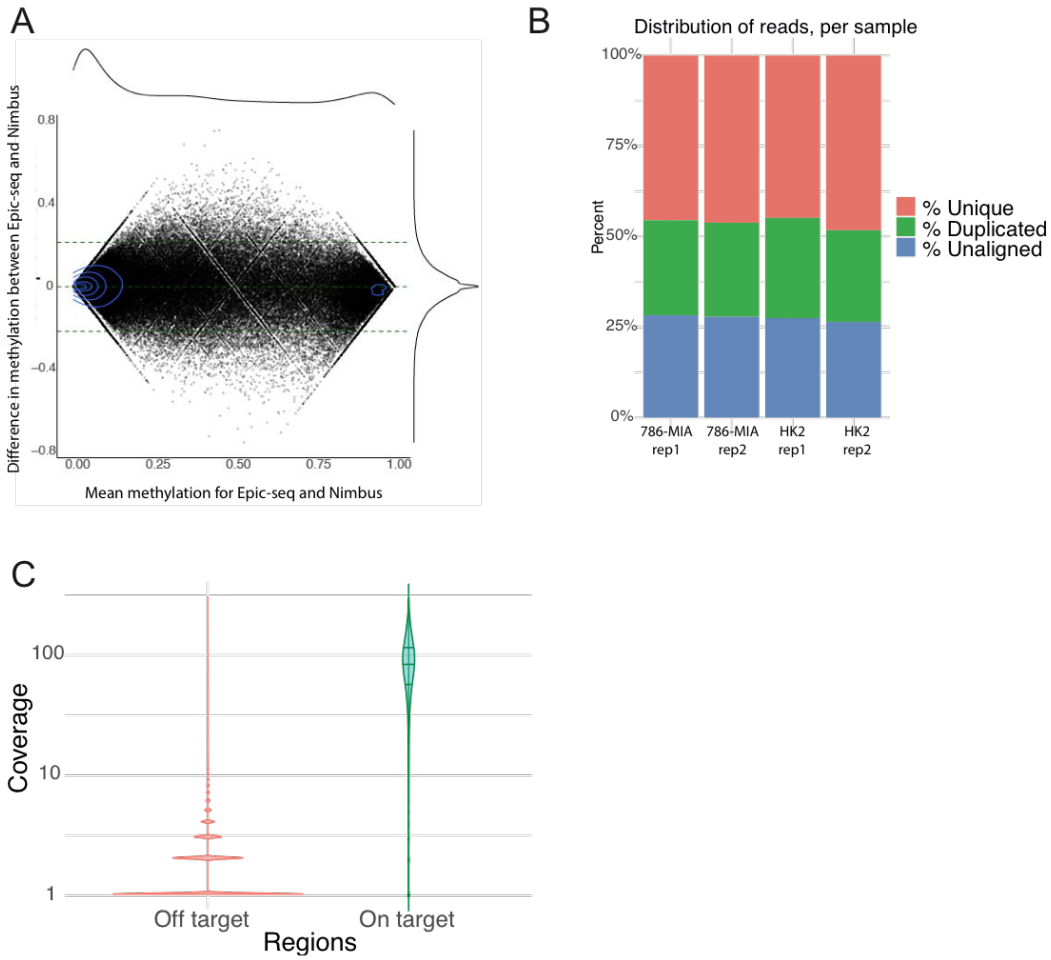


Figure 7.3: Quality control metrics for cell line supernatant and gDNA sequenced using Nimbus

Panel A- Bland Altman plot for one gDNA ccRCC tissue sample sequenced using Epic-seq and Nimbus. Density plots are shown above the x- and y- axes. Panel B- Distribution of sequencing reads for four cell line samples analysed using Nimbus, showing the proportion of unaligned, duplicate, and unique reads. 786-M1A cell line replicates (rep1 and rep2) represent a model system of ccRCC, whereas HK2 cell line replicates (rep1 and rep2) represent a model system of renal proximal tubule cells. Panel C- Sequencing coverage for CpGs which are on-target (ccRCC DMR panel) and off-target respectively.

7.3.3.2 Quality control in cell lines

I applied Nimbus to DNA extracted from cell line supernatant to ensure adequate performance and baseline QC in low-input, fragmented cfDNA prior to evaluating human plasma samples. I therefore applied the Nimbus ccRCC panel to DNA derived from HK2 cells (2 technical replicates) and 786-M1A cells (2 technical replicates), which represent model systems for normal kidney epithelial cells and ccRCC respectively. Cell line supernatant (rather than gDNA) was selected as a surrogate for highly fragmented cfDNA, both of which exhibit a nucleosomal fragmentation pattern. For these samples, median alignment rate was 61.6% (IQR: 61.0-62.6%) and median duplication rate was 36.6% (IQR: 35.9-37.3%; Figure 7.3B). Samples achieved a median of ~13 million unique reads (median: 1.33e+07; IQR: 1.19e+07 to 1.48e+07). Figure 7.3B shows the percentage of reads which aligned to the human reference genome, and of those, the proportion which represent uniquely mapped versus duplicate reads. Consistent results were achieved for all samples analysed, demonstrating technical reproducibility of the Nimbus workflow. Median CHH and CHG methylation were 0.3% (IQR: 0.3-0.3) and 0.4% (IQR: 0.4%-0.4%), indicating there is no evidence of under-conversion and the tight distribution of values demonstrates consistency of assay performance.

Next, the efficiency of the capture panel was evaluated using standard metrics. In this cell line experiment, 99% of the regions of interest were captured at a coverage ≥1x (median 99.4%, IQR: 99.4-99.5%). The on-target rate is defined as the percentage of aligned and unique bases which are captured by the panel (i.e. bases which are located within the target panel ±150bp, the latter being the length of a sequencing read). The median on-target rate was 26.5% (IQR: 26.1-27.0%) and the median coverage on-target was 88x (IQR: 78-99x). The on-target bases are highly enriched compared to off-target bases (median fold enrichment 126, IQR: 123-128). Figure 7.3C demonstrates coverage in on-target versus off-target CpGs (rather than all bases) and suggests high on-target rates; coverage is orders of magnitude higher for captured regions of interest, compared to non-target regions. Limiting the analysis to the target regions of interest and coverage ≥10x, methylation results were reproducible, with technical replicates achieving a Pearson correlation coefficient of 0.97 and 0.98 for HK2 and 786-M1A cell lines respectively. Overall, this analysis demonstrated sufficiently satisfactory QC metrics to proceed to an analysis of human cfDNA samples.

7.3.3.3 Quality control in plasma samples

Next, I performed Nimbus on 67 cfDNA samples derived from human plasma and evaluated QC metrics to ensure the method was achieving expected standards using low input cfDNA. These samples were derived from 30 ccRCC patients and 37 controls. Further details regarding cohort characteristics are described in section 7.3.4. Median alignment to the reference genome was 70.4% (IQR: 67.2-72.3%; Figure 7.4A). The median duplication rate was 51.3% (IQR: 46.4-58.8%; Figure 7.4A), which was considerably higher than that for cell lines (discussed further in section 7.4.1). Following alignment and de-duplication, samples achieved a median of ~9 million unique reads (median: 9.05e+06; IQR: 7.46e+06 to 1.06e+07) (Figure 7.4B). In the pilot study (N=67), there was only one sample which achieved a rate >1% for CHH and CHG methylation (i.e. non CpG methylation), which indicates under-conversion (sample 6022). Median CHH and CHG methylation were 0.2% (IQR: 0.2-0.2%, max: 1.2%) and 0.2% (IQR: 0.2-0.2%, max: 1.1%) respectively (Figure 7.4C), demonstrating high efficiency conversion of unmethylated cytosines across this cohort of cfDNA samples.

Subsequently, QC metrics relating to the capture panel were assessed. Generally, 99% of the regions of interest were captured at a coverage ≥ 1x (median 99.3%, IQR: 99.2-99.4%), which is similar to that achieved in cell line experiments. The median on-target rate was 11.1% (IQR: 9.0-13.5%), however 4 samples achieved an on-target rate <5% (samples 7387, 5826, 5848 and 5998; Figure 7.4D). The median coverage on-target was 25x (IQR: 19-35), however 4 samples achieved <10x median coverage on-target (samples 5826, 5846, 5848, 5998; Figure 7.4E). Median fold-enrichment on-target compared to off-target was 53-fold (IQR: 42-64; Figure 7.4F). In summary, although relatively low on-target rate and low coverage were noted (particularly in some samples), the high on-target fold enrichment is reassuring. Taken together, these results suggest generally consistent performance of the ccRCC Nimbus panel. Generally, QC metrics were sufficiently satisfactory to proceed to analysis of cfDNA detection, however, there is scope for further optimization of library preparation and sequencing strategies to improve the ccRCC Nimbus workflow in future studies (as discussed in section 7.4.1).

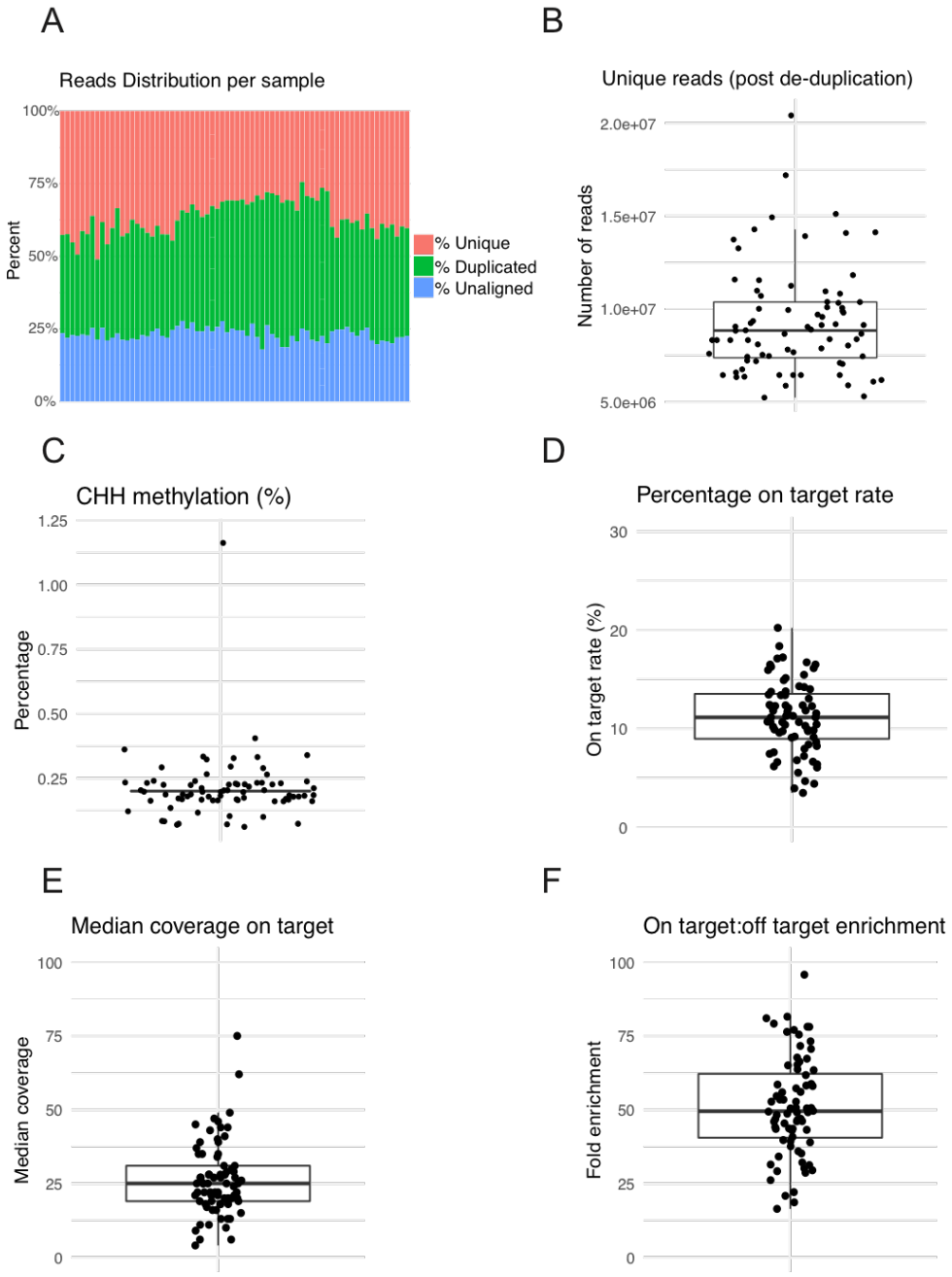


Figure 7.4: Quality control for human plasma cfDNA sequenced using Nimbus

Panel A- Distribution of sequenced reads for plasma cfDNA samples analysed using Nimbus, showing the proportion of unaligned, duplicate, and unique reads. Panel B- Boxplot showing the total number of unique reads obtained following alignment and de-duplication. Panel C- Boxplot of non-CpG (CHH) methylation signal detected. Panel D- Boxplot showing the on-target rate for the ccRCC Nimbus panel. Panel E- Median coverage for on-target bases. Panel F- Fold enrichment in on-target compared to off-target bases.

7.3.4 Targeted methylation analysis in plasma cfDNA from patients with ccRCC and controls

I performed Nimbus on a discovery cohort consisting of plasma cfDNA derived from 67 patients with ccRCC (N=30) and cancer-free controls (N=37). Patients with ccRCC were selected across the disease spectrum to enable an evaluation of ctDNA detection rates at different stages (6 patients with stage I-II, 6 patients with stage III and 18 with stage IV). Patient demographics were known for a subset of the cohort (Table 7.1). Cancer-free controls (N=37) consisted of individuals who did not have (known) evidence of renal cancer. These are referred to as 'controls' for the duration of the Chapter. Controls were selected pragmatically based on sample availability and came from a cohort of individuals with benign disease on prostate biopsy and no evidence of cancer on clinical follow up. Unfortunately, clinical and demographic data were not known for all patients at the time of writing (although this data has been requested).

The following section describes the approach used to calculate Nimbus assay scores and how this was applied to my discovery cohort (plasma cfDNA derived from 30 ccRCC patients and 37 controls). Nimbus includes an automated bioinformatics pipeline, which produces a Nimbus score for each sample. As described above (section 7.3.2), I identified DNA methylation markers included in the Nimbus panel by evaluating tissue (i.e. DMRs that differentiate ccRCC vs normal kidney) and filtered these against healthy plasma cfDNA methylomes. DMRs (N=5801) in the panel can thus be subdivided into those that are hypo- and hypermethylated in ccRCC vs normal kidney, and these will be referred to as such for the remainder of the Chapter. It is acknowledged that a number of these DMRs may not be adequate markers in plasma cfDNA, which is enriched in blood cells (whereas normal kidney was the comparator in tissue). Thus, the first step in the Nimbus pipeline is to determine which of the DMRs included in the panel are informative in plasma by assessing methylation levels in cfDNA from controls, for each read. Twenty cfDNA samples were randomly selected from controls to determine background methylation levels in cancer-free individuals (referred to as the 'background set' hereafter). DMRs were considered informative if background methylation levels in the background set were different to the expected methylation levels from ccRCC tissue. Using this method, 1347 hypomethylated and 1529 hypermethylated informative DMRs were identified (uninformative DMRs were excluded).

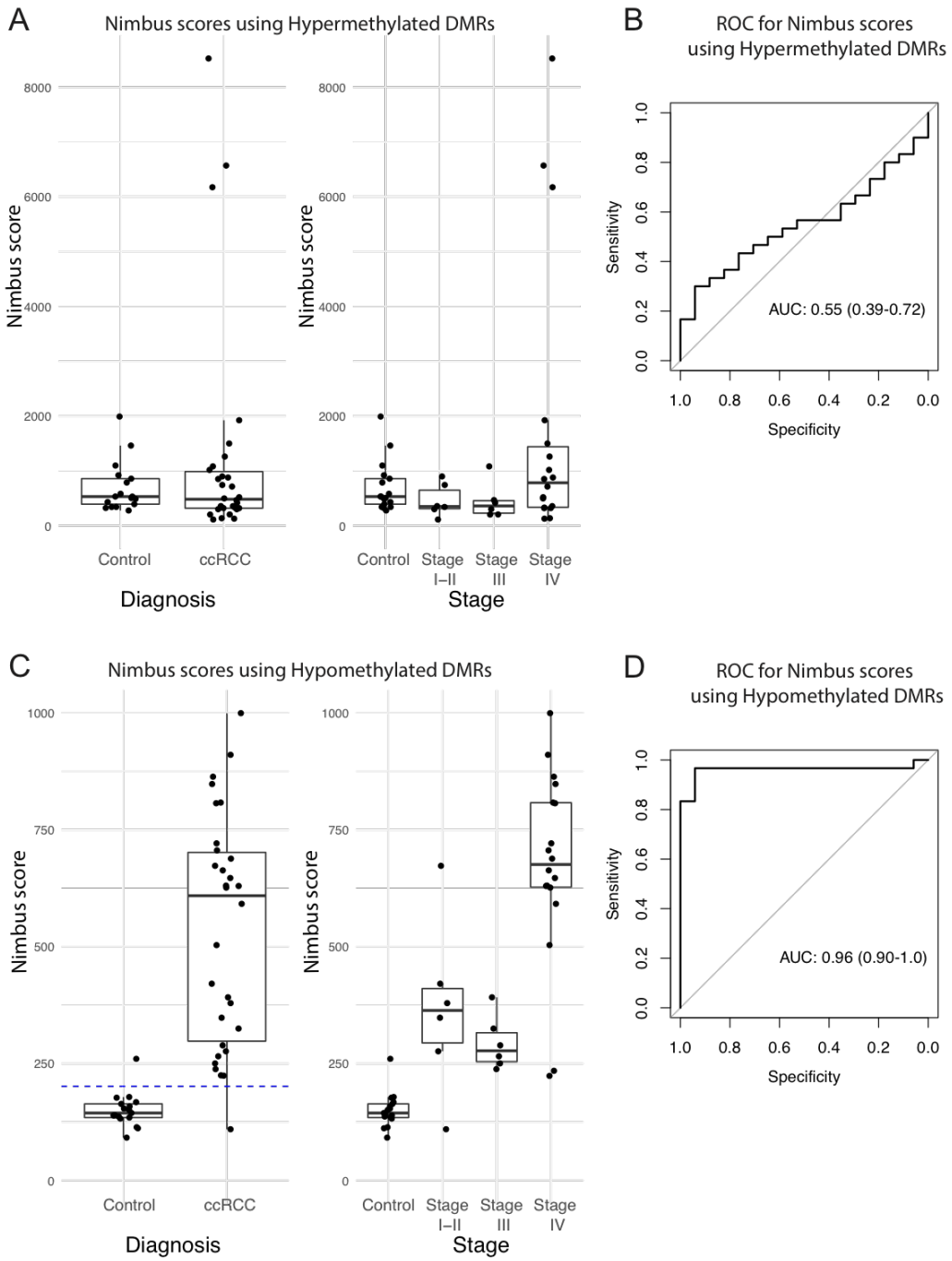


Figure 7.5: Nimbus scores for plasma cfDNA samples obtained from ccRCC patients and controls

Panel A- Nimbus scores, derived using hypermethylated differentially methylated regions (DMRs), for plasma cfDNA samples. The boxplots on the left demonstrate scores in ccRCC patient plasma samples (N=30) versus control (N=17). The panel on the right subdivides ccRCC samples by stage (stage I-II, stage III and stage IV). Panel B- Receiver operating characteristic (ROC) curve for Nimbus scores derived using hypermethylated DMRs, to distinguish ccRCC patients from controls. The area under the curve (AUC) is shown. Panel C- Nimbus scores, derived using hypomethylated DMRs, for plasma cfDNA samples. The boxplots on the left demonstrate scores in ccRCC patient plasma samples (N=30) versus control (N=17). The panel on the right subdivides ccRCC samples by stage (stage I-II, stage III and stage IV). The dotted blue line indicates the Nimbus score threshold which maximizes sensitivity and specificity in the training cohort (score=201), achieving 0.941 specificity and 0.967 sensitivity. Panel D- ROC curve (and AUC) for Nimbus scores derived using hypomethylated DMRs, to distinguish ccRCC patients from controls.

Subsequently, I evaluated methylation levels using Nimbus scores in cfDNA samples from 30 ccRCC and the remaining 17 controls (Nimbus scores generated by Dr Lach, Massie Group). For each cfDNA sample, each read is assessed in turn, and the bioinformatic pipeline calculates the score (similar to the likelihood) that the read is tumour derived by comparing the distribution of methylation levels in the sample compared to the background set (i.e. 20 separate cancer-free controls). Individual scores per read were summarised into an overall score per patient, by adding the log of the individual read level scores. This produces an overall Nimbus score that can be compared for individual samples. Importantly, reads which contain non-CpG methylation (CHH/CHG) >1% are discarded as these demonstrate evidence of DNA fragment under-conversion (e.g. resulting from non-denatured template molecules in the conversion reaction). This approach enables us to discard individual reads rather than an entire sample based on non CpG-methylation, thus maximizing available data. In addition, Nimbus scores are normalised based on the number of high-quality reads per million, to enable accurate comparisons between samples which are not confounded by variations in total sequencing read counts. High-quality reads were defined as reads which contain at least one CpG, did not demonstrate non-CpG methylation >1% and have a high probability of being correctly aligned to the reference genome (mapping quality score ≥40).

I assessed Nimbus scores for DMRs which are hypermethylated and hypomethylated in ccRCC separately, to determine which provides superior discriminatory power to distinguishing ccRCC from controls. Evaluating hypermethylated DMRs, there was no significant difference in Nimbus scores in cfDNA derived from ccRCC versus controls; Benjamini Hochberg (BH) adjusted p value >0.05 (Figure 7.5A). There was a trend towards higher Nimbus scores in patients with metastatic compared to non-metastatic disease (median 783, IQR: 337-1440 vs median 350, IQR: 278-537, BH adjusted p value = 0.078), however this was mostly driven by three patients with stage IV RCC who had very high scores and were obvious outliers (Figure 7.5A). Figure 7.5B demonstrates the receiver operating characteristic (ROC) curve for hypermethylated DMRs, highlighting sensitivity and specificity at different Nimbus score thresholds. As expected based on Figure 7.5A, the area under the curve (AUC) was very low.

Next, I evaluated hypomethylated DMRs and demonstrated a clear, significantly higher score in ccRCC compared to controls (median 609, IQR: 298-701 vs median 144, IQR: 134-164; BH adjusted p value =7.9e-09; Figure 7.5C). Scores were significantly higher in stage IV disease than stage I-III (median 676, IQR: 627-808 vs median 307, IQR: 262-382, BH adjusted p value = 0.0009, Figure 7.5C). Therefore, hypomethylated DMRs were selected to be taken forward for all subsequent analyses.

Figure 7.5D shows the ROC curve for hypomethylated DMRs, resulting in an AUC of 0.96 (95% CI: 0.90-1.0). The Nimbus score threshold which maximizes sensitivity and specificity was 201, achieving 0.94 specificity and 0.97 sensitivity (i.e. misclassifies 1/17 control and 1/30 ccRCC samples, as shown in Figure 7.5C). One ccRCC sample misclassified as control was derived from a patient with stage I disease (patient 5848). This sample also had lower QC metrics than the rest of the cohort (as described in section 7.3.3.3). 6.4 million unique reads were obtained, the on-target rate was 3.9% and median coverage on-target was 6x, indicating a low number of unique molecules which may limit ctDNA detection. The inability to detect ctDNA using Nimbus may be due to early tumour stage (presumed lower total ctDNA levels) and reduced assay sensitivity due to low numbers of unique ctDNA molecules and associated sampling error. Overall, these results suggest that the Nimbus score (based on DMRs which are hypomethylated in ccRCC) can differentiate cfDNA derived from patients with ccRCC versus controls, with a high sensitivity and specificity. Validation in an independent, larger cohort is warranted.

7.3.5 Comparing methylation and mutational ctDNA analysis

One of my initial hypotheses was that targeted methylation analysis of ctDNA may achieve superior detection rates than mutational analysis, therefore I sought to compare both methods in a subset of cfDNA plasma samples (N=14). For 14 of the ccRCC patients included in my Nimbus analysis, I was able to obtain data from Dr Chris Smith (Rosenfeld Group) derived from the INVAR-TAPAS method for patient-specific mutational analysis in cfDNA [158]. Table 7.1 summarises patient and sample characteristics. In brief, Dr Smith first sequenced multi-region nephrectomy samples using whole exome sequencing (WES) to identify patient-specific mutations, subsequently a custom panel was created targeting bespoke SNVs as well as 109 genes which are commonly mutated in ccRCC. The panel was applied to blood plasma cfDNA samples collected prior to nephrectomy. INVAR-TAPAS includes a bioinformatic pipeline which evaluates plasma ctDNA mutations whilst modelling background noise in control samples to allow custom error suppression at patient-specific mutations. cfDNA fragment size is also used to weight variant reads and enrich for tumour derived signal [347]. INVAR-TAPAS estimates a global mutant allele fraction (MAF) for each sample, where a MAF of 0% indicates ctDNA was not detectable. First, I evaluated INVAR-TAPAS results for the 14 ccRCC samples, subsequently I compared these against my Nimbus methylation data.

Table 7.1: Characteristics of ccRCC cfDNA samples analysed using Nimbus & INVAR-TAPAS

Clinical and sample data for 14 ccRCC cfDNA samples analysed using both methylation (Nimbus) and mutational (INVAR-TAPAS) approaches. The number of patient-specific mutations targeted by INVAR-TAPAS is shown for each patient. Some of the information included in the table was derived from Smith et al [158].

Patient ID	Age/ Sex	BMI	Stage	Tumour Size	Vascular invasion	Necrosis	Ki67 sum	Pseudocapsule	Leibovich score	Mutations targeted by INVAR
5401	77F	34	pT3a	11 cm	Renal Vein	Yes	7	Focal	5	156
5532	62M	43	pT3aM1	6 cm	No	No	25	Focal	4	211
5626	76M	32	pT3aM1	4.5 cm	No	No	42	Focal	4	73
5644	62M	43	pT2a	7.5 cm	No	No	24	Focal	5	279
5799	54M	34	pT1b	6.5 cm	No	No	41	Focal	2	426
5801	74M	28	pT3a	10.8cm	Renal Vein	Yes	56	Yes	8	321
5802	75F	22	pT3a	2.8 cm	Renal Vein	Focal	42	No	6	223
5813	42M	28	pT3a	8.7 cm	IVC	No	29	No	4	371
5818	64F	24	pT3aM1	7.4 cm	Renal Vein	Yes	18	Focal	6	388
5826	77F	33	pT1b	6.1 cm	No	No	37	Focal	4	327
5827	60F	32	pT1a	2.5 cm	No	No	36	No	0	257
5846	62F	30	pT3a	23 cm	Renal Vein	No	20	Focal	6	80
5848	66F	41	pT1b	5.4 cm	No	No	31	No	3	148
5998	65M	24	pT1b	5.2 cm	No	No	25	Yes	2	321

Using INVAR-TAPAS, ctDNA was detected in 50% (7/14) of plasma samples from patients with ccRCC. I postulated that ctDNA detection may be related to the number of patient-specific mutations that were identified in tissue and subsequently targeted in cfDNA analysis (Table 7.1). However, the number of mutations targeted was not significantly different in patients in which ctDNA was detected compared to those where ctDNA was not detected (median 279, IQR: 190-346 vs median 257, IQR: 180-324; p value >0.05). Thus, in this small cohort, there was no evidence to suggest that the low ctDNA detection rates were due to insufficient numbers of mutations being targeted (although this could still be a contributing factor). I noted a significant correlation between tumour size and MAF estimated by INVAR-TAPAS (Pearson correlation coefficient = 0.61, BH adjusted p value= 0.019, Figure 7.6A). Furthermore, MAF was significantly higher in patients with vascular invasion compared to those without (median MAF 0.00011, IQR: 0.00007-0.00015 vs median MAF 0, IQR: 0-0; BH adjusted p value 0.0024; Figure 7.6B). Vascular involvement was defined as renal vein or inferior vena cava (IVC) invasion, and indeed the patient with IVC invasion had the second highest MAF. These findings are consistent with the original study by Dr Smith [158], despite the smaller

sample size in my cohort. Interestingly, ctDNA was not detectable by INVAR-TAPAS in 2 out of 3 patients with mRCC and it was these two patients (patients 5532 and 5626) which were noted to have smaller tumours (4.5cm and 6cm) and absence of renal vein involvement. In summary, ctDNA MAF was associated with increased tumour size and vascular invasion, suggesting tumour proliferation and shedding of ctDNA into the bloodstream may determine the amount of ctDNA present which may be detected by targeted mutational analysis.

Next, I compared plasma ctDNA detection rates using my ccRCC Nimbus methylation assay to the INVAR-TAPAS mutational analysis. For Nimbus, I evaluated hypomethylated DMRs as these were superior to hypermethylated DMRs and used the previously defined threshold (Nimbus score > 201; see section 7.3.4). ctDNA detection rates were significantly higher using Nimbus (13/14 patients) compared to INVAR-TAPAS (7/14 patients) (Detection rate 93% vs 50%; p value = 0.036). I aimed to evaluate whether there was a quantitative relationship between Nimbus scores and INVAR-TAPAS output. There was no significant difference in Nimbus score in patients in which ctDNA was detected by INVAR-TAPAS compared to those where ctDNA was not detected (p value > 0.05; Figure 7.6C). There was no significant correlation between ctDNA MAF estimated by INVAR-TAPAS and Nimbus scores (p value > 0.05). However, some consistencies were noted amongst the two methods. The one false-negative using Nimbus (patient 5848) did not have ctDNA detected by INVAR-TAPAS. Furthermore, two mRCC plasma samples (patients 5532 and 5626) had relatively lower Nimbus scores (score <300) compared to other mRCC samples, and INVAR-TAPAS was unable to detect ctDNA in these individuals. Furthermore, I sought to explore the relationship between the Nimbus score and clinical parameters. There was no significant association (BH adjusted p value > 0.05) between Nimbus scores and the following patient and tumour characteristics: age, sex, BMI, tumour size, vascular invasion, tumour proliferation (ki67) and presence of a pseudocapsule (Figure 7.6D-E). However, results may be biased by the small total sample size (N=14) and limited numbers in subgroup analyses. In summary, ctDNA detection rates using targeted methylation analysis via Nimbus were significantly higher than detection rates using INVAR-TAPAS, suggesting that Nimbus might have greater clinical utility. Further validation is warranted in larger cohorts of samples with linked clinical and outcome data.

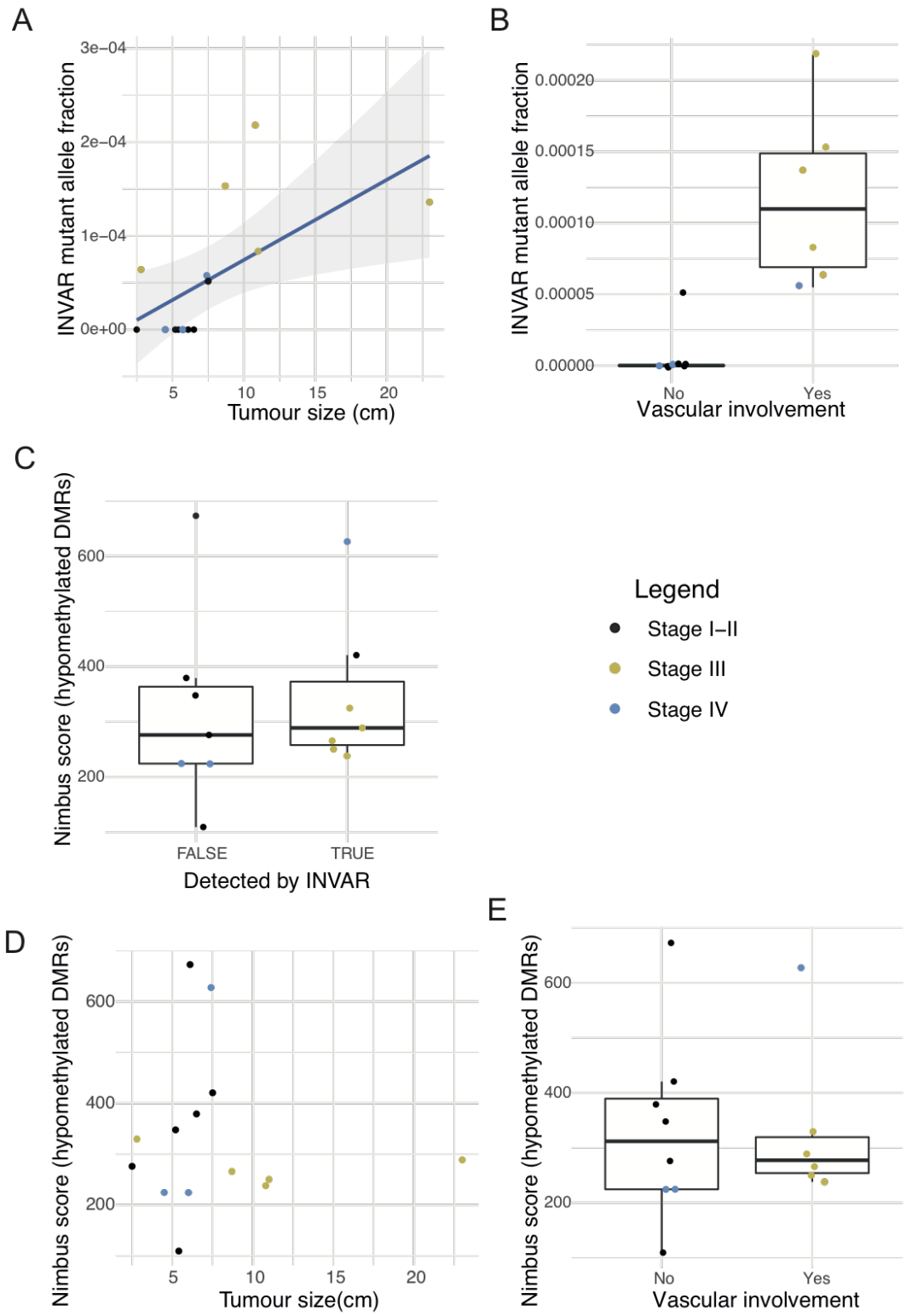


Figure 7.6: INVAR-TAPAS mutational analysis in plasma cfDNA samples obtained from ccRCC patients and controls

Panel A- Mutant allele fraction obtained using INVAR-TAPAS versus tumour size (Pearson correlation coefficient= 0.61, adj p val= 0.019). Panel B- Mutant allele fraction obtained using INVAR-TAPAS, for patients with and without vascular involvement (adj p val = 0.0024). Panel C- Nimbus scores for ccRCC patients in which ctDNA was detected by INVAR-TAPAS compared to those where ctDNA was not detected (p val > 0.05). Panel D- Nimbus scores vs tumour size. There was no significant correlation. Panel E- Nimbus scores for patients with and without vascular involvement (p val >0.05).

7.3.6 Methylation analysis of cfDNA derived from proximal vs distal samples

I hypothesized that proximal sampling might yield increased ctDNA detection rates compared to distal samples, which may be a useful clinical strategy for patients with SRMs who are expected to have low levels of plasma ctDNA. I therefore performed a small study to establish the feasibility of collecting proximal and distal samples in patients with renal tumours, with the ultimate aim of assessing whether proximal samples can lead to increased detection rates. For each patient, I collected matched renal biopsy, post-biopsy fluid (proximal sample) and plasma (distal sample). Post-biopsy fluid was defined as serosanguinous fluid collected from the local site following renal biopsy (a needle and syringe were used to aspirate liquid). The feasibility of proximal sample collection is described below, followed by the results of the Nimbus ccRCC panel analysis in proximal and distal samples.

Fifteen patients undergoing renal biopsy as standard of care were approached, and 100% consented to participate in the study (DIAMOND biobanking study; REC ID 03/018). Biopsy indication was variable and included: diagnosis for indeterminate SRM or SRM considering surveillance, diagnosis of mRCC, or biopsy requirement for inclusion into another clinical trial at Addenbrooke's Hospital (EASE and NAXIVA trials). Of the 15, two patients were excluded as the biopsy was cancelled due to clinical reasons (e.g. lesion too small to biopsy) and two further patients were excluded following biopsy results (one biopsy revealed metastatic melanoma, one biopsy was non-diagnostic). Post-biopsy fluid and plasma were therefore available for 11 patients, and tissue biopsy was available from 8 of these. The pathology and tumour stage of the cohort was mixed, with 3 patients having oncocytoma and 8 patients having ccRCC (including stage I, III and IV disease). The volume of post-biopsy fluid collected ranged between <1ml and 10ml (median: 2ml, IQR: 1-3ml; Table 7.2). The higher the volume of post-biopsy fluid, the higher the cfDNA content which can be extracted and processed using Nimbus. It was therefore unsurprising to find a strong positive correlation between the volume of post-biopsy fluid collected and the number of high-quality sequencing reads obtained using Nimbus (Pearson correlation= 0.76, p value= 0.006; Figure 7.7A). There were no obvious statistically significant determinants of post-biopsy fluid volume; there was no association with pathological diagnosis (ccRCC vs oncocytoma), biopsy technique (co-axial needle vs biopsy gun), size or stage. In summary, this suggests that patients are willing to participate in proximal sampling studies and that collecting post-biopsy fluid is feasible, although very low volumes were obtained in some patients.

Table 7.2: Clinical and sample data for 11 patients undergoing renal biopsy

Peripheral blood, biopsy tissue and post-biopsy fluid were collected for each patient, however biopsies were missing for three individuals. Pathological stage is included, where known. The asterisk (*) indicates that the patient did not have surgery, therefore stage is presumed (radiological stage is reported rather than pathological stage).

Patient ID	Age/ Sex	Co-axial needle	Volume of post biopsy fluid	Pathology	Tumour Size	Stage	Biopsy available
5046	78M	Yes	2ml	Oncocytoma	4.7 cm	Benign	Yes
7425	69F	Yes	5ml	ccRCC	5.5 cm	ypT3a	Yes
7445	78M	No	~1ml + saline	ccRCC	4.1 cm	Stage I *	Yes
7447	71M	Yes	10ml	ccRCC	4.8 cm	ypT3a	Yes
7459	65M	Yes	2ml	ccRCC	6.0 cm	Stage IV	Yes
7735	79F	No	2ml	Oncocytoma	6.2 cm	Benign	Yes
7740	67M	No	2ml	ccRCC	4.4 cm	pT1b	No
7752	74M	No	~1ml + saline	Oncocytoma	2.5 cm	Benign	Yes
7755	59F	No	4ml	ccRCC	3.5 cm	pT3a	No
7756	79M	Yes	~1ml + saline	ccRCC	3.9 cm	Stage I *	No
7757	57M	No	~1ml + saline	ccRCC	4.1 cm	pT3a	Yes

Nimbus scores were generated for cfDNA derived from matched samples (plasma and post-biopsy fluid) and I compared these to assess signal in distal vs proximal samples. Evaluating paired data, Nimbus scores were significantly higher in post-biopsy fluid compared to plasma (BH adjusted p value = 0.015; Figure 7.7B). Indeed, post-biopsy fluid Nimbus scores were higher than plasma Nimbus scores in 82% (9/11) of patients. Subsequently, I evaluated gDNA from renal biopsy tissue as a proof-of-principle, as these samples would be expected to have tumour derived DNA content orders of magnitude higher than cell free DNA from plasma. Nimbus scores in biopsy tissue were significantly higher than plasma Nimbus scores (BH adjusted p value = 0.015; Figure 7.7B). In 87.5% (7/8) of patients, Nimbus scores derived using biopsy tissue were higher than those derived from post-biopsy aspirate, however this did not reach statistical significance (BH adjusted p value = 0.18; Figure 7.7B). The higher scores noted in biopsy tissue compared to cfDNA implies Nimbus is measuring reads which originate from kidney tissue and are therefore more likely to be tumour derived. The utility of proximal sampling is best exemplified for patient 7459, who had widespread metastases. Nimbus scores in plasma were relatively low, however this individual had both the highest Nimbus score in post-biopsy fluid and renal biopsy, and is indeed an obvious outlier in Figure 7.7B.

Furthermore, I sought to evaluate if any clinical parameters were associated with Nimbus scores, namely tumour stage and pathological subtype. In patients with ccRCC (N=8), there was no significant difference in Nimbus scores in patients with stage I vs stage III-IV disease, for cfDNA derived from plasma or post-biopsy fluid (p value >0.05, Figure 7.7C-D). This may be due to the limited sample size, as in the discovery cohort (N=30, section 7.3.4), Nimbus scores were significantly higher in patients with metastatic vs non-metastatic ccRCC. The present analysis aimed to evaluate cfDNA in patients with ccRCC, but by chance, 3 patients in the study were diagnosed with oncocytoma following renal biopsy. There was no significant difference in Nimbus scores in patients with ccRCC compared to oncocytoma in cfDNA derived from plasma or post-biopsy fluid (p value >0.05, Figure 7.7E-F). The small samples size (only 3 patients with oncocytoma) and absence of oncocytoma-specific markers in the methylation panel limit the interpretation of this analysis, however I look forward to exploring this in future. Taken together, these data suggest tumour derived cfDNA may be enriched in post-biopsy fluid compared to plasma and that further research into this topic is warranted.

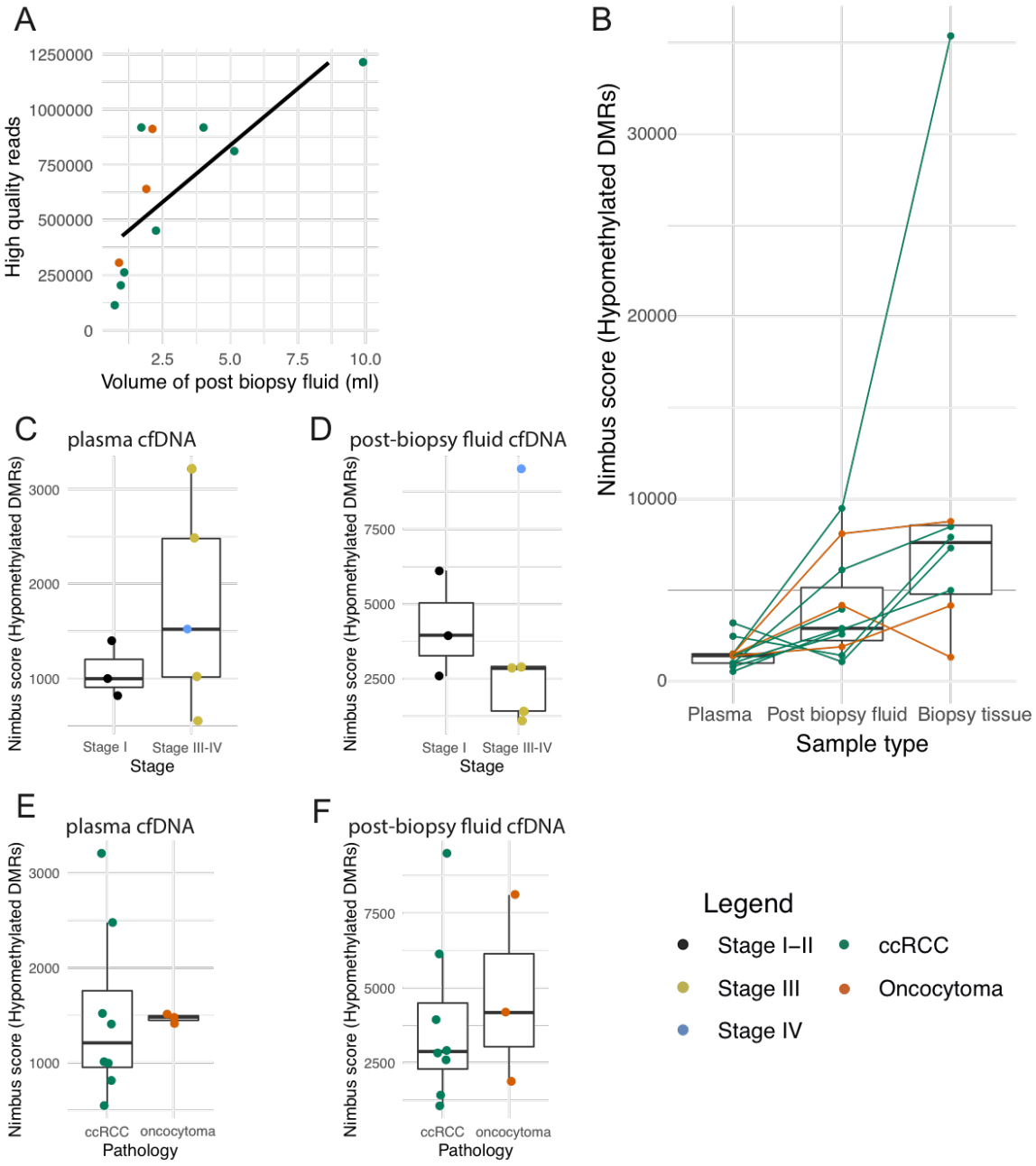


Figure 7.7: Nimbus scores in cfDNA derived from proximal versus distal samples, in patients with renal tumours

Panel A- Scatterplot demonstrating the volume of post-biopsy fluid collected and the number of high-quality sequencing reads obtained using Nimbus. A strong positive correlation was noted (Pearson correlation= 0.76, p value= 0.006). Panel B- Nimbus scores for cfDNA samples derived from plasma, post-biopsy fluid and renal biopsy tissue. Nimbus scores were derived using hypomethylated differentially methylated regions (DMRs). Panel C and D- Nimbus scores, derived using hypomethylated DMRs, for cfDNA samples derived from plasma (Panel C) and post-biopsy fluid (Panel D) of ccRCC patients, presented by tumour stage. Panel E and F- Nimbus scores, derived using hypomethylated DMRs, for cfDNA samples derived from plasma (Panel E) and post-biopsy fluid (Panel F) of patients with ccRCC versus oncocytoma.

7.4 Discussion and future direction

In summary, in this chapter I aimed to evaluate ctDNA detection in patients with renal tumours vs cancer-free controls, via targeted DNA methylation analysis using the Nimbus method. Nimbus is a novel experimental method (consisting of enzymatic conversion, library preparation and targeted capture) and a bespoke bioinformatics pipeline for ctDNA methylation analysis which was recently developed in our research group (Massie laboratory). First, I evaluated methylation in gDNA from ccRCC and adjacent normal tissue in order to identify methylation markers that distinguish ccRCC tumour from normal kidney. I subsequently investigated these markers in liquid samples using Nimbus, including cell line supernatant and human plasma cfDNA. Cell line work was performed to assess quality control metrics and confirmed acceptability prior to proceeding to patient samples. In plasma cfDNA, I demonstrated that Nimbus achieves high accuracy in distinguishing ccRCC from controls (AUC=0.96) and that detection rates are superior compared to mutational ctDNA analysis using INVAR-TAPAS. Lastly, I demonstrate that Nimbus signal may be enriched in post-biopsy fluid compared to plasma, suggesting the potential utility of proximal sampling.

7.4.1 The Nimbus method and quality control metrics

I evaluated Nimbus QC metrics, reflecting on current methodological strengths and potential challenges, and highlighting areas for future method optimization. I compared QC metrics in DNA derived from cell line supernatant and human plasma. Following sequencing, alignment rates to the reference genome were ~70%. This is in keeping with published methods evaluating DNA methylation analysis in ctDNA which achieved rates of 60-80% [353]. Non-CpG methylation rates, a measure of under-conversion in the first enzymatic step, were also very consistent (<1% in all but one sample). Comparing cell line and human cfDNA sample results, the main finding was that the latter demonstrated lower numbers of unique reads (which was more pronounced in certain samples) and was associated with a high duplication rate, low on-target rate and low coverage. The high duplication rate is likely to reflect the low starting material in cfDNA (low number of unique molecules), which necessitates use of a high number of PCR cycles. Indeed, the starting material for cell line experiments was 50ng, whereas only 10ng were used for cfDNA due to sample availability, and this represents a real-world challenge. The low starting input is exacerbated by steps within the protocol which lead to DNA degradation or loss. Nimbus uses enzymatic conversion rather than bisulphite conversion to minimize DNA degradation associated with the latter (estimated to be ~90%) [354]. The enzymatic method therefore retains the cfDNA's native fragmentation pattern and reduces DNA loss. However, the Nimbus protocol requires multiple bead-based washes during

enzymatic conversion and library preparation (prior to PCR amplification) which leads to loss of complexity and fewer unique molecules. Future optimization experiments will aim to minimise DNA loss during sequential wash steps. In addition, alternative library preparation methods, which include fewer wash steps, can be trialled (for example NEB EM-seq) [355].

Important considerations for capture-based methods include minimising off-target rates and maximising efficiency of recovering target loci. The low on-target rates in my data (median ~11%) may suggest relatively inefficient capture, and the low coverage may impact the sensitivity to detect markers of interest. The relatively low on-target rate may be consistent with deep redundant sequencing (i.e. saturating coverage of target regions, resulting in higher on-target duplication rates and inflation of off-target regions in de-duplicated data). In contrast, the high on-target fold-enrichment shows that the ccRCC capture panel has successfully been enriched, although the low on-target unique coverage reflects the limited complexity (genome equivalents) in the final sequencing data. Down sampling analysis and alternative capture-based methods are currently being explored to determine whether on-target rates can be improved in my future work, aiming to increase enrichment of desired targets. These optimization steps could improve the efficiency of our current method, which would be key to allow future integration into clinical practice. Although beyond the scope of the present analysis, future work could also explore pre-analytical steps (such as type of blood collection tube used, centrifugation and extraction protocols) to maximise ctDNA yields and enable simultaneous collection of multiple analytes (such as ctDNA, miRNAs, proteins etc) [356].

7.4.2 Defining a methylation panel for targeted analysis in Nimbus

I hypothesize that refining my target methylation panel could lead to further methodological improvements. In section 4.8.1, I identified significant DMRs in ccRCC versus normal kidney tissue and included these in a methylation panel which was evaluated in cfDNA using Nimbus. A number of these DMRs did not prove to be informative in cfDNA analysis. Intuitively, it might be postulated that DMRs which are hypermethylated in ccRCC might be better markers than hypomethylated DMRs as the former contain more CpGs per read (as shown in Figure 7.2). However, my results demonstrate that regions which are hypomethylated in ccRCC distinguish tumour from control cfDNA whilst hypermethylated regions did not, an observation which was consistent in other cancer types tested in our group. Preliminary work evaluating ctDNA using Nimbus in patients with glioblastoma, prostate and oesophageal cancer compared to controls demonstrated the superiority of hypo- over hypermethylated regions (work undertaken by Dr Park and Dr Lach, postdoctoral fellows within my

research group). Ziller et al [134] found that hypomethylated DMRs are more cell-type specific than hypermethylated DMRs, which may explain these findings. Furthermore, it may be that hypomethylated DMRs demonstrate less biological or technical variability than hypermethylated DMRs, and this is something which would be interesting to assess in future work. In Chapter 6, I evaluated methylation heterogeneity in ccRCC between patients, within a patient and within a sample. This analysis could be adapted to select more appropriate methylation markers for inclusion in the Nimbus panel. Ideally, methylation markers would display low epipolymorphism (i.e. reduced disordered methylation within and between reads) and reduced inter-patient heterogeneity, to increase ctDNA detection. Although my original intention was to identify methylation markers in Chapter 5 and Chapter 6, and subsequently evaluate them in liquid samples, due to time constraints, I had to perform the two analyses in parallel. Moving forward, my aim is to refine the current list of DMRs by removing non-informative markers, including homogeneous clonal hypomethylated markers and excluding hypermethylated DMRs. Creating a focused, smaller capture panel could therefore enable greater depth of sequencing at reduced costs. The cost of the current Nimbus method is ~£350 per sample, which could be reduced to ~£125 per sample in future by refining the panel size, using more affordable library preparation kits and pooling more samples for sequencing.

7.4.3 Nimbus for targeted methylation analysis in cfDNA

ctDNA detection has multiple potential applications, and regarding the two clinical questions addressed in this thesis, this would be: (a) liquid biopsies to differentiate SRMs to improve the diagnostic pathway and (b) detection of post-operative minimal residual disease to improve risk-stratification in ccRCC patients treated with curative intent. However, controversy exists regarding detection rates of ctDNA in patients with renal cancer, with some studies reporting levels below 50% [158, 345, 357] and one study suggesting 97% [159]. I postulated that targeting thousands of RCC-specific methylation markers would improve detection rates compared to existing approaches relying on mutational analysis. I therefore aimed to evaluate detection rates of ctDNA in ccRCC using my current panel of methylation markers and our newly developed Nimbus methodology. Following a successful proof of principle study, Nimbus could be investigated in future large cohort studies to address the clinical scenarios listed above.

My principal finding is that the Nimbus score (based on DMRs which are hypomethylated in ccRCC) can differentiate cfDNA derived from patients with ccRCC versus controls. The AUC in the discovery cohort was very high (AUC=0.96). Future work should therefore focus on externally validating these results in larger, independent datasets. My discovery cohort consisted of plasma from 67 patients

with ccRCC and cancer-free controls, in order to determine detection rates and evaluate feasibility of the current methodology. The sample size of approximately 50 to 100 was chosen as this is an achievable number in view of current sample availability. Previously published work had similar sample sizes [158, 159]. Controls were selected pragmatically from patients with benign prostate biopsies and no known evidence of RCC. It is acknowledged that this is not an ideal cohort as it excludes women. One advantage is these controls represent a relatively similar age (and therefore likely similar comorbidity profile) to renal cancer patients and therefore provide a more appropriate comparison than commercially available cfDNA from young, healthy individuals. Moss et al found that cfDNA content is twice as high in older adults (67-97 year-olds) compared to young individuals (19-30 year-olds), suggesting the importance of age-appropriate controls to limit bias [101]. Future work on RCC ctDNA methylation analysis should include larger, comorbidity matched and gender balanced control cohorts in which the absence of renal cancer has been confirmed (for example by recruiting patients with normal abdominal imaging following investigation for haematuria).

I demonstrated that Nimbus achieves superior detection rates than patient-specific mutational analysis using INVAR-TAPAS in my dataset. On a practical level, Nimbus has the advantage that it does not rely on patient-specific panels (as does INVAR-TAPAS), therefore enabling prospective use in patients who have not had tissue sampling or tumour sequencing. Furthermore, not having to create custom panels reduces costs and increases speed, therefore increasing ease of use in clinical practice. The reason for low ctDNA detection rates using mutational analysis remains incompletely understood. INVAR-TAPAS has previously shown ctDNA detection rates of 1 part per million [347], although this requires high input cfDNA and/or larger numbers of patient-specific mutations than can be routinely achieved for ccRCC using WES. In my cohort, ctDNA was only detected in 50% of patients with ccRCC using INVAR-TAPAS. Evaluating a larger number of mutations/genomic alterations improves the detection limit, however surprisingly, in my cohort there was no association between the number of patient-specific mutations targeted by INVAR-TAPAS and detection rates. The MAF estimated using INVAR-TAPAS was associated with increased tumour size and presence of vascular invasion. However, 2 patients with mRCC (patients 5532 and 5626) did not have detectable ctDNA (and it was these patients who had smaller tumours and no venous involvement). It is likely that larger tumours and tumours with direct vascular access are able to shed more ctDNA. It may be that in patients with mRCC, circulating tumour derived cfDNA may represent the metastasis rather than the primary kidney tissue sampled to derive patient-specific mutations for inclusion in the INVAR-TAPAS panel. Turajlic et al [75] evaluated primary tumours and synchronous metastases in TRACERx and demonstrated that although driver genes are often shared between the two, some

aberrations are present in the metastases and absent in the primary. These two mRCC patient (patients 5532 and 5626) had relatively lower Nimbus scores (score <300) compared to other mRCC samples, which could suggest that the MAF may have been below the limit of detection of INVAR-TAPAS. Future studies evaluating the biology underlying ctDNA shedding by renal tumours is key. I attempted to explore the association between clinico-pathological parameters and Nimbus scores in my dataset (for example tumour proliferation using ki67 and presence/absence of a pseudocapsule) to characterise determinants of ctDNA shedding. However, my study had a limited sample size, therefore only very strong associations could have been identified. In summary, Nimbus demonstrated superior detection rates compared to mutational analysis and this warrants further external validation as a potentially clinically useful strategy.

Comparing Nimbus results to existing cfDNA studies in renal cancer from the literature suggests that Nimbus is highly competitive. In my discovery cohort, I applied Nimbus to plasma cfDNA from 30 ccRCC patients and 17 controls and achieved a ROC AUC of 0.96. Selecting a threshold of >201 for hypomethylated targets, ctDNA was detected in 97% (29/30) of ccRCC patients, including stages I-IV. The patient which was not detected had stage I disease. As already mentioned in the Background section 2.5, studies performing mutational ctDNA analysis achieved a lower performance than Nimbus. For example, research evaluating mutation panels in genes known to be affected in RCC achieved detection rates of 72-79% in patients with metastatic ccRCC (rates were much lower in non-metastatic disease) [346, 348]. Patient specific mutational panels (such as INVAR-TAPAS) also achieved low detection rates (~42-50%) as did pan-cancer methylation panels (such as GRAIL; sensitivity <50%) [358]. In summary, to our knowledge, only one study achieved detection rates similar to Nimbus. Nuzzo et al [159] utilised cfMeDIP-seq to analyse methylation in cfDNA from patients with ccRCC and controls. In brief, the method uses antibody-based immunoprecipitation to pull down (and therefore enrich for) methylated cfDNA fragments, which are then sequenced. The authors report a ROC AUC of 0.99, with ctDNA being detected in 97% (67/69) of ccRCC patients, which is very similar to Nimbus' performance. Due to time constraints, it was not possible to benchmark Nimbus against cfMeDIP-seq, though this could be performed in future. Interestingly, whilst I selected my methylation target panel for Nimbus using kidney tissue, Nuzzo et al identify DMRs directly in plasma (by methylation calling in ccRCC vs control plasma). Since tumour derived signal in plasma is very low (estimated <0.01% in most cfDNA samples) it is likely that Nuzzo et al's strategy may identify methylation changes within immune cell populations which are different in ccRCC patients versus controls, rather than tumour-cell intrinsic changes. Inflammation is a hallmark of cancer [233]; ccRCC is considered an inflammatory cancer and several immune-associated

prognostic parameters have been identified (including neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio and C-reactive protein) [359]. This may explain the high AUCs noted by Nuzzo and suggests that although cfMeDIP-seq may not necessarily detect tumour-derived signal, this may still be a potentially useful strategy to detect ctDNA in ccRCC versus controls. Since I identified my target methylation panel by comparing differential methylation in ccRCC tissue versus normal kidney, it is more likely that Nimbus detects tumour derived signal. Some of the signal detected by Nimbus might be originating from kidney tissue secondary to generic kidney damage (for example due to localised pressure secondary to tumour growth) rather than being tumour specific. This could potentially explain why Nimbus scores were not correlated with mutant allele fraction estimated by INVAR-TAPAS nor clinical parameters (such as tumour size and vascular involvement) and why Nimbus scores were high in patients with oncocytoma (discussed further below). Alternatively, the lack of correlation between INVAR-TAPAS mutant allele fraction and Nimbus scores could also be due to low sample sizes (MAF was only available for 7 patients). Further exploration of my methylation capture panel could help to disentangle this, leading to further insights and refinements for ctDNA detection in RCC.

My analysis is the first to compare ctDNA detection in proximal vs distal samples in patients with renal tumours, and this represents a key strength of my work. My results suggest that proximal sampling using post-biopsy fluid may enable an enrichment of Nimbus signal compared to peripheral venous samples, which could be useful in patients with SRMs and low plasma cfDNA. At present renal biopsy may be non-diagnostic for a number of reasons (insufficient sample, necrosis, target lesion missed or difficulty differentiating pathological subtypes on histopathology). It is envisioned that targeted methylation analysis of cfDNA derived from post-biopsy fluid could serve as a diagnostic adjunct in patients undergoing renal biopsy. One challenge is that low volumes of post-biopsy fluid (<1ml) were obtained in some patients, and this could limit future clinical applications.

Nimbus was also able to detect a signal in cfDNA from patients with oncocytoma. However, the current methylation panel was seemingly unable to distinguish ccRCC from oncocytoma (although sample numbers were very low). This is not surprising as the DMRs included in the panel were not designed to detect oncocytoma or to discriminate these benign lesions from ccRCC. As mentioned above, Nimbus may also include methylation changes that distinguish kidney tissue from plasma cfDNA (which is mostly immune cells) in addition to tumour-specific changes in methylation. Therefore, in future, the Nimbus methylation panel will be refined to include CpGs which distinguish pathological subtypes of renal tumours, for example those identified in Chapter 5, by

MethylBoostER. Smith et al [158] were able to detect ctDNA in patients with oncocytoma and chRCC using INVAR-TAPAS, suggesting this is feasible. Evidently, increased understanding of ctDNA biology (including cfDNA secretion and degradation) will translate into improved detection rates by informing optimal sampling strategies.

In conclusion, this chapter consists of a proof-of-concept suggesting Nimbus can detect ctDNA in ccRCC with excellent accuracy and that the signal in proximal samples is likely to be higher than distal samples. This paves the way for further external validation studies and testing in patients with SRMs. Longitudinal sampling pre- and post-nephrectomy could be performed to assess whether Nimbus scores are reduced following tumour excision, which would suggest Nimbus measures tumour derived signal. In addition, it would be interesting to evaluate Nimbus scores in cfDNA derived from urine, as an alternative distal sample. Smith et al and Nuzzo et al suggested this might be a useful method to detect ctDNA in patients with renal cancers [158, 159], though in both studies plasma cfDNA yielded superior detection rates. The advantage of using urine over plasma is that it is truly non-invasive, can be collected in larger volumes and is preferred by patients [360]. cfDNA could also be coupled with protein markers (such as aquaporin 1 and perilipin 2 in urine or KIM-1 in blood) [6, 145, 361] to increase sensitivity and specificity. In conclusion, I present a method which enables sensitive ctDNA detection in patients with renal cancers using thousands of methylation markers, achieving superior detection rates compared to most currently available platforms. I elucidate further optimisation steps and following external validation, it is envisioned that this method could have tangible clinical applications.

Chapter 8 Summary, future direction and conclusion

In this thesis, I aimed to characterise DNA methylation in tissue and liquid samples from patients with renal tumours. In Chapter 5, I evaluate DNA methylation in tissue derived from patients with common benign and malignant renal tumours, highlight putative epigenetically regulated genes and explore the potential for diagnostic applications. In Chapter 6, I comprehensively map the landscape of DNA methylation heterogeneity in ccRCC tissue and lastly, I assess targeted methylation analysis for ctDNA detection in liquid samples in Chapter 7. Herein, I summarise and integrate the findings of my three results chapters and discuss potential clinical implications.

In patients with SRMs, the current diagnostic pathway (consisting of imaging and renal biopsy) is unable to confidently distinguish pathological subtypes of renal tumours and differentiate benign from malignant disease. As a result, 20% of patients are found to have benign disease post-operatively, meaning patients are undergoing unnecessary kidney surgery with associated risks of morbidity and mortality. To address this unmet clinical need, we developed MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours), a machine learning model which uses DNA methylation data to classify tissue samples into common pathological subtypes of malignant and benign renal tumours (ccRCC, pRCC, chRCC and oncocytoma) and normal kidney. MethylBoostER was externally validated on four independent datasets, achieving a high accuracy (AUCs >0.90 for all subtypes). One of the main challenges for model development and external validation is the lack of publicly available methylation data, in particular for less common subtypes (e.g. chRCC, oncocytoma and rarer subtypes) and for early-stage tumours. Further external validation is warranted in patients with SRMs, using prospectively collected biopsy samples (rather than nephrectomy tissue which has a much higher tumour content). Combining samples and well-annotated clinical data from multiple UK or international sites may enable such large-scale validation studies in future. As discussed in section 5.4.2, an ideal model would integrate multi-modal data including DNA methylation, imaging and clinical characteristics and would predict patient outcome rather than simply predicting the pathological subtype of the renal tumour, whilst taking into account competing risks of death. Once a model has been refined and finalized, it must be tested prospectively in a clinical setting and benchmarked against the existing standard of care. Prior to the model being adopted in clinical practice, it is crucial to demonstrate clinical utility (i.e. improved diagnostic accuracy and better patient outcomes), cost-effectiveness and feasibility of the test within an NHS setting.

I explored the association between MethylBoostER accuracy and sample purity as it is acknowledged that biopsy samples often have low tumour content, and this represents a real-world challenge limiting clinical application. There was a sharp drop off in accuracy in samples with purity <0.2, suggesting that potentially biopsy may have to be repeated if purity is below this threshold. In a clinical scenario, it is envisioned that patients with SRMs would undergo renal biopsy, and MethylBoostER could represent a useful diagnostic adjunct in addition to histopathological review of biopsy tissue. However, both histopathological slide review and the MethylBoostER model may be limited in biopsy samples with low tumour content, and it is envisioned that liquid biopsies could be used to increase diagnostic confidence and avoid repeat biopsy in this setting. In Chapter 7, I demonstrated the feasibility of detecting cfDNA using targeted methylation analysis (using the Nimbus method) in plasma and post-biopsy fluid from patients with renal tumours. In the discovery cohort, targeted methylation analysis of plasma cfDNA was able to distinguish ccRCC from controls with an AUC of 0.96, suggesting this could potentially be used clinically in the future. However, the existing literature suggests that levels of ctDNA tend to be lower in patients with early-stage renal tumours [151, 158], limiting the potential use of plasma ctDNA as a liquid biopsy in patients with SRMs. These previous reports motivated the study of post-biopsy fluid as a route to improve detection rates through proximal sampling. My results suggested that post-biopsy fluid, which is collected at the time of renal biopsy, may be enriched for tumour derived DNA compared to peripheral venous samples. In future, I plan to create an updated targeted methylation panel for Nimbus, including methylation markers that distinguish pathological subtypes of renal tumours (for example CpGs from MethylBoostER). The Nimbus experimental method can be applied to DNA derived from renal biopsy, post-biopsy fluid, plasma and even potentially urine (though the latter was not tested in this thesis). In an ideal scenario, plasma cfDNA would be used as a non-invasive liquid biopsy to determine a diagnosis in patients with SRMs (first line option; Figure 8.1). Individuals with undetectable plasma cfDNA would undergo renal biopsy and collection of post-biopsy fluid, which would be processed using targeted methylation analysis to identify the most likely diagnosis (though the small volumes of post-biopsy fluid collected in some patients may limit applicability). This could have the potential to reduce the number of patients undergoing unnecessary renal surgery for benign tumours. Multi-sample analysis and a targeted panel containing thousands of methylation markers could increase likelihood of detection. However, maintaining testing costs reasonably low is necessary to ensure feasibility in clinical practice, therefore future studies should also evaluate the minimum panel size and lower cost methods to detect these molecular signatures.

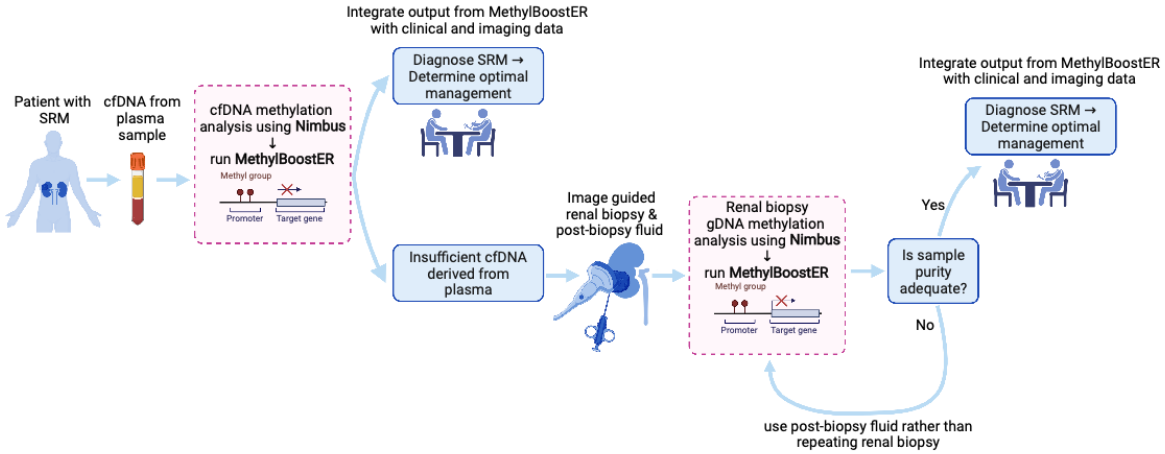


Figure 8.1: Proposed future integration of Nimbus and MethylBoostER into clinical practice

Ideally, plasma cfDNA would be used as a non-invasive liquid biopsy to determine a diagnosis in patients with SRMs. Targeted methylation analysis would be performed using Nimbus, and the methylation data would be used to run the MethylBoostER machine learning model. Individuals with undetectable plasma cfDNA would undergo renal biopsy and collection of post-biopsy fluid. The latter sample type would be particularly useful in patients where there is low plasma cfDNA and the renal biopsy either contains low tumour content or the target lesion was missed. MethylBoostER output would be integrated with clinical and pathological data (including patients' comorbidity profile) to enable patient-centred management.

Although genetic heterogeneity has been extensively studied, relatively little is known regarding methylation heterogeneity in ccRCC. An improved understanding of the latter could offer insights into tumour biology and inform biomarker selection. I therefore sought to characterise DNA methylation heterogeneity in ccRCC tissue in Chapter 6. Homogeneously methylated CpGs across multi-region samples are postulated to be early events in tumorigenesis and they represent ideal targets for diagnostic applications, for example in diagnosing SRMs. Markers with low intra-tumoral heterogeneity are key to ensure that diagnosis is consistent and independent of the region biopsied. In Chapter 5, I demonstrate that MethylBoostER achieves consistent classification results in multi-region samples from the same patients in 90% of individuals. Refining target selection by prioritising homogeneously methylated CpGs could improve this statistic, enabling clinical utility. Furthermore, markers with low inter-patient heterogeneity are more likely to be detected in liquid samples when using the same targets for multiple individuals (which is preferable to a costly patient-specific approach). Conversely, heterogeneously methylated CpGs could represent late, subclonal events in tumour evolution (although they could also represent heterogeneous cell types which constitute bulk methylation profiles from tissue). It may not be possible to disentangle these two processes completely, however, an awareness of this can lead to more informative biomarker selection. For example, in Chapter 6, I demonstrated that a proportion of differentially methylated cytosines that distinguish ccRCC and normal tissue (i.e. those with a high variance in tumour) may represent

immune cell markers rather than tumour intrinsic changes. These would therefore be uninformative markers in plasma cfDNA (as the majority of the non-tumour signal comes from immune cells), and this should be taken into account when designing future versions of the Nimbus capture panel. In summary, integrating data from Chapter 5 to Chapter 7 is likely to improve the marker selection for inclusion in future iterations of the RCC Nimbus target panel.

Another clinical priority explored in my thesis is the need for improved prognostic stratification for ccRCC patients. The literature suggests that the degree of genetic intra-tumoral heterogeneity (ITH) and pattern of tumour evolution (e.g. *VHL* monodriver vs branched evolution) has prognostic potential in ccRCC [87]. This prompted me to comprehensively evaluate methylation heterogeneity in Chapter 6. However, no association with clinical parameters was identified, potentially due to the small sample size. I also compared phylo-epigenetic trees against phylogenetic trees derived from copy number data. My analysis would be rendered more complete by deriving tumour phylogenies based on mutational data and comparing these with phylo-epigenetic trees, and I am planning to undertake this work with my collaborators at the Max Delbrück Center in Berlin. Although a number of prognostic methylation markers have been proposed, very few of these have been successfully externally validated and some demonstrate conflicting results across different studies [67]. In Chapter 6, I demonstrated that heterogeneously methylated CpGs coincide with putative prognostic methylation markers in ccRCC (such as *SFRP1*, *DKK2* and *CCND1*), which could explain difficulties in external validation noted in the literature, especially when different platforms are used.

A key novel result from Chapter 6, was the identification of locally disordered methylation (i.e. differential epipolymorphism) at the promoter region of genes known to be associated with ccRCC. Clinically this is relevant because disordered methylation within a read may contribute to difficulties validating prognostic markers due to different CpGs being evaluated by disparate platforms (for example Epic-seq vs 450k array). Importantly, my analysis revealed that epipolymorphism within a gene promoter may be an independent regulator of gene expression in addition to gains or losses in average methylation, which is a novel insight into transcriptional regulation in ccRCC. Assessing disordered methylation within a read may in itself be a useful cancer biomarker and should be explored further in future studies.

Given that identifying individual prognostic methylation markers has proven to be elusive, a different approach towards risk stratification in ccRCC needs to be adopted. For example, monitoring of minimal residual disease post-operatively by serially evaluating plasma cfDNA in

patients who have undergone curative nephrectomy could represent a useful strategy. The half-life of cfDNA is between 15 minutes and 2.5 hours [149], suggesting dynamic fluctuations could accurately reflect changes in tumour burden. Having demonstrated that cfDNA can be detected in plasma from ccRCC patients using Nimbus with a high accuracy (Chapter 7), I hope to evaluate serial samples to determine if there is any association with clinical relapse. The ultimate aim would be to use targeted methylation analysis in ctDNA to identify high-risk patients who would benefit from adjuvant therapy or increased follow-up, and to identity relapse early, enabling improved survival.

In conclusion, I have comprehensively characterised DNA methylation in patients with ccRCC and other common pathological subtypes of renal tumours, evaluating both tissue and liquid samples. My findings will help researchers select more informative biomarkers for both diagnostic and prognostic applications, hopefully reducing the number of markers which fail to be validated. Although biotechnology companies, such as GRAIL and GuardantHealth, have invested in large scale studies of ctDNA, these often utilise pan-cancer markers to screen healthy individuals. Focusing on RCC specific markers and tailoring the search to a specific clinical question (for example differentiating SRMs or risk-stratifying ccRCC patients) rather than a large heterogeneous population, will ensure that my work has a high chance of translating to clinically relevant results in future.

Chapter 9 References

1. Rossi SH, Blick C, Handforth C, Brown JE, Stewart GD, Renal Cancer Gap Analysis Collaborative. Essential Research Priorities in Renal Cancer: A Modified Delphi Consensus Statement. Eur Urol Focus. 2020;6(5):991-8.
2. Rossi SH, Fielding A, Blick C, Handforth C, Brown JE, Stewart GD. Setting Research Priorities in Partnership with Patients to Provide Patient-centred Urological Cancer Care. Eur Urol. 2019;75(6):891-3.
3. Rossi SH, Stewart GD. Renal Cancer. In: Pang K, Osman NI, Catto J, Chapple C, editors. Basic Urological Sciences. 1 ed: CRC Press; 2021.
4. Rossi SH, Stewart GD. Epidemiology and screening for renal cancer. In: Anderson CJ, Patel HRH, editors. Renal Cancer: Current and Future Innovations. 1 ed: Springer Nature; 2022.
5. Usher-Smith JA, Li L, Roberts L, Harrison H, Rossi SH, Sharp SJ, et al. Risk models for recurrence and survival after kidney cancer: a systematic review. BJU Int. 2021. *In press.*
6. Flitcroft JG, Verheyen J, Vemulkar T, Welbourne EN, Rossi SH, Welsh SJ, et al. Early detection of kidney cancer using urinary proteins: a truly non-invasive strategy. BJU Int. 2022;129(3):290-303.
7. Rossi SH*, Newsham I* et al. Accurate detection of benign and malignant renal tumour subtypes with MethylBoostER: an epigenetic marker driven learning framework. Science Advances. 2022. *Accepted.*
8. Sciacovelli M, Dugourd A, Jimenez LV, Yang M… Rossi SH et al. Nitrogen partitioning between branched-chain amino acids and urea cycle enzymes sustains renal cancer progression. bioRxiv. 2021:2021.09.17.460635.
9. RECOVERY Trial Group, Horby P et al. Dexamethasone in Hospitalized Patients with Covid-19. The New England journal of medicine. 2021;384(8):693-704.
10. GlobalSurg Collaborative. SARS-CoV-2 infection and venous thromboembolism after surgery: an international prospective cohort study. Anaesthesia. 2022;77(1):28-39.
11. GlobalSurg Collaborative. Effects of pre-operative isolation on postoperative pulmonary complications after elective surgery: an international prospective cohort study. Anaesthesia. 2021;76(11):1454-64.
12. Covidsurg Collaborative. SARS-CoV-2 vaccination modelling for safe surgery to save lives: data from an international prospective cohort study. Br J Surg. 2021;108(9):1056-63.
13. GlobalSurg Collaborative. Timing of surgery following SARS-CoV-2 infection: an international prospective cohort study. Anaesthesia. 2021;76(6):748-58.
14. Bergamaschi L et al. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. Immunity. 2021;54(6):1257-75 e8.
15. Znaor A, Lortet-Tieulent J, Laversanne M, Jemal A, et al. International variations and trends in renal cell carcinoma incidence and mortality. Eur Urol. 2015;67(3):519-30.
16. Cancer Research UK kidney cancer statistics 2016. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/kidney-cancer#heading-Zero.
17. Smittenaar CR, Petersen KA, Stewart K, Moitt N. Cancer incidence and mortality projections in the UK until 2035. Br J Cancer. 2016;115(9):1147-55.
18. Cancer Research UK Kidney Cancer statistics. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/kidney-cancer.
19. Rossi SH, Klatte T, Usher-Smith J, Stewart GD. Epidemiology and screening for renal cancer. World J Urol. 2018;36(9):1341-53.
20. Capitanio U, Bensalah K, Bex A, Boorjian SA, et al. Epidemiology of Renal Cell Carcinoma. Eur Urol. 2019;75(1):74-84.

21. Pelc NJ. Recent and future directions in CT imaging. Ann Biomed Eng. 2014;42(2):260-8.
22. Welch HG, Skinner JS, Schroeck FR, Zhou W, et al. Regional Variation of Computed Tomographic Imaging in the United States and the Risk of Nephrectomy. JAMA Intern Med. 2017.
23. Wernli KJ, Rutter CM, Dachman AH, Zafar HM. Suspected extracolonic neoplasms detected on CT colonography: literature review and possible outcomes. Acad Radiol. 2013;20(6):667-74.
24. Ricketts CJ, De Cubas AA, Fan H, Smith CC, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep. 2018;23(12):3698.
25. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018;361(6402):594-9.
26. Wein AJ, Kavoussi LR, Partin AW, Peters CA. Campbell-Walsh urology. 11 ed: Elsevier; 2016.
27. Ljungberg B, Albiges L, Abu-Ghanem Y, Bensalah K, et al. European Association of Urology Guidelines on Renal Cell Carcinoma: The 2019 Update. Eur Urol. 2019.
28. Ginzburg S, Tomaszewski JJ, Kutikov A. Focal ablation therapy for renal cancer in the era of active surveillance and minimally invasive partial nephrectomy. Nat Rev Urol. 2017;14(11):669-82.
29. Rossi SH, Prezzi D, Kelly-Morland C, Goh V. Imaging for the diagnosis and response assessment of renal tumours. World J Urol. 2018;36(12):1927-42.
30. Millet I, Doyon FC, Hoa D, Thuret R, et al. Characterization of small solid renal lesions: can benign and malignant tumors be differentiated with CT? AJR Am J Roentgenol. 2011;197(4):887-96.
31. Akdogan B, Gudeloglu A, Inci K, Gunay LM, et al. Prevalence and predictors of benign lesions in renal masses smaller than 7 cm presumed to be renal cell carcinoma. Clin Genitourin Cancer. 2012;10(2):121-5.
32. Marconi L, Dabestani S, Lam TB, Hofmann F, et al. Systematic Review and Meta-analysis of Diagnostic Accuracy of Percutaneous Renal Tumour Biopsy. Eur Urol. 2016;69(4):660-73.
33. Poggio ED, McClelland RL, Blank KN, Hansen S, et al. Systematic Review and Meta-Analysis of Native Kidney Biopsy Complications. Clin J Am Soc Nephrol. 2020;15(11):1595-602.
34. Macklin PS, Sullivan ME, Tapping CR, Cranston DW, et al. Tumour Seeding in the Tract of Percutaneous Renal Tumour Biopsy: A Report on Seven Cases from a UK Tertiary Referral Centre. Eur Urol. 2019;75(5):861-7.
35. Marconi L, Dabestani S, Lam TB, Hofmann F, et al. Systematic Review and Meta-analysis of Diagnostic Accuracy of Percutaneous Renal Tumour Biopsy. Eur Urol. 2016;69(4):660-73.
36. Patel HD, Druskin SC, Rowe SP, Pierorazio PM, et al. Surgical histopathology for suspected oncocytoma on renal mass biopsy: a systematic review & meta-analysis. BJU Int. 2017;119(5):661-6.
37. Nguyen KA, Nolte AC, Alimi O, Hsiang W, et al. Determinants of Active Surveillance in Patients With Small Renal Masses. Urology. 2019;123:167-73.
38. Kim JH, Li S, Khandwala Y, Chung KJ, et al. Association of Prevalence of Benign Pathologic Findings After Partial Nephrectomy With Preoperative Imaging Patterns in the United States From 2007 to 2014. JAMA Surg. 2019;154(3):225-31.
39. Johnson DC, Vukina J, Smith AB, Meyer AM, et al. Preoperatively misclassified, surgically removed benign renal masses: a systematic review of surgical series and United States population level burden estimate. J Urol. 2015;193(1):30-5.
40. Sohlberg EM, Metzner TJ, Leppert JT. The Harms of Overdiagnosis and Overtreatment in Patients with Small Renal Masses: A Mini-review. Eur Urol Focus. 2019;5(6):943-5.
41. Ingels A, Duc S, Bensalah K, Bigot P, et al. Postoperative outcomes of elderly patients undergoing partial nephrectomy. Sci Rep. 2021;11(1):17201.
42. Pierorazio PM, Johnson MH, Ball MW, Gorin MA, et al. Five-year analysis of a multi-institutional prospective clinical trial of delayed intervention and surveillance for small renal masses: the DISSRM registry. Eur Urol. 2015;68(3):408-15.
43. McIntosh AG, Ristau BT, Ruth K, Jennings R, et al. Active Surveillance for Localized Renal Masses: Tumor Growth, Delayed Intervention Rates, and >5-yr Clinical Outcomes. Eur Urol. 2018;74(2):157-64.

44. Gerlinger M, Rowan AJ, Horswell S, Math M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. NEJM. 2012;366(10):883-92.
45. Leao RR, Richard PO, Jewett MA. The role of biopsy for small renal masses. Int J Surg. 2016;36(Pt C):513-7.
46. Leao RR, Richard PO, Jewett MA. Indications for biopsy and the current status of focal therapy for renal tumours. Transl Androl Urol. 2015;4(3):283-93.
47. Pierorazio PM, Hyams ES, Mullins JK, Allaf ME. Active surveillance for small renal masses. Rev Urol. 2012;14(1-2):13-9.
48. Nayak JG, Patel P, Bjazevic J, Liu Z, et al. Clinical outcomes following laparoscopic management of pT3 renal masses: A large, multi-institutional cohort. Can Urol Assoc J. 2015;9(11-12):397-402.
49. Veccia A, Falagario U, Martini A, Marchioni M, et al. Upstaging to pT3a in Patients Undergoing Partial or Radical Nephrectomy for cT1 Renal Tumors: A Systematic Review and Meta-analysis of Outcomes and Predictive Factors. Eur Urol Focus. 2021;7(3):574-81.
50. Jewett MA, Mattar K, Basiuk J, Morash CG, et al. Active surveillance of small renal masses: progression patterns of early stage kidney cancer. Eur Urol. 2011;60(1):39-44.
51. Volpe A, Panzarella T, Rendon RA, Haider MA, et al. The natural history of incidentally detected small renal masses. Cancer. 2004;100(4):738-45.
52. Uzosike AC, Patel HD, Alam R, Schwen ZR, et al. Growth Kinetics of Small Renal Masses on Active Surveillance: Variability and Results from the DISSRM Registry. J Urol. 2018;199(3):641-8.
53. Jang A, Patel HD, Riffon M, Gorin MA, et al. Multiple growth periods predict unfavourable pathology in patients with small renal masses. BJU Int. 2018;121(5):732-6.
54. Gordetsky J, Eich ML, Garapati M, Del Carmen Rodriguez Pena M, et al. Active Surveillance of Small Renal Masses. Urology. 2019;123:157-66.
55. Dabestani S, Thorstenson A, Lindblad P, Harmenberg U, et al. Renal cell carcinoma recurrences and metastases in primary non-metastatic patients: a population-based study. World J Urol. 2016;34(8):1081-6.
56. Klatte T, Rossi SH, Stewart GD. Prognostic factors and prognostic models for renal cell carcinoma: a literature review. World J Urol. 2018;36(12):1943-52.
57. Bai Y, Li S, Jia Z, Ding Y, et al. Adjuvant therapy for locally advanced renal cell carcinoma: A meta-analysis and systematic review. Urol Oncol. 2018;36(2):79 e1- e10.
58. Choueiri TK, Tomczak P, Park SH, Venugopal B, et al. Adjuvant Pembrolizumab after Nephrectomy in Renal-Cell Carcinoma. NEJM. 2021;385(8):683-94.
59. Leibovich BC, Blute ML, Cheville JC, Lohse CM, et al. Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: a stratification tool for prospective clinical trials. Cancer. 2003;97(7):1663-71.
60. Zisman A, Pantuck AJ, Dorey F, Said JW, et al. Improved prognostication of renal cell carcinoma using an integrated staging system. J Clin Oncol. 2001;19(6):1649-57.
61. Frank I, Blute ML, Cheville JC, Lohse CM, et al. 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. 2002;168(6):2395-400.
62. Karakiewicz PI, Briganti A, Chun FK, Trinh QD, et al. Multi-institutional validation of a new renal cancer-specific survival nomogram. J Clin Oncol. 2007;25(11):1316-22.
63. Sorbellini M, Kattan MW, Snyder ME, Reuter V, et al. A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. J Urol. 2005;173(1):48-51.
64. Buti S, Puligandla M, Bersanelli M, DiPaola RS, et al. Validation of a new prognostic model to easily predict outcome in renal cell carcinoma: the GRANT score applied to the ASSURE trial population. Annals of oncology. 2018;29(7):1604.
65. Graham J, Dudani S, Heng DYC. Prognostication in Kidney Cancer: Recent Advances and Future Directions. J Clin Oncol. 2018:JCO2018790147.

66. Petitprez F, Ayadi M, de Reynies A, Fridman WH, et al. Review of Prognostic Expression Markers for Clear Cell Renal Cell Carcinoma. Front Oncol. 2021;11:643065.
67. Joosten SC, Deckers IA, Aarts MJ, Hoeben A, et al. Prognostic DNA methylation markers for renal cell carcinoma: a systematic review. Epigenomics. 2017;9(9):1243-57.
68. Gulati S, Martinez P, Joshi T, Birkbak NJ, et al. Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. Eur Urol. 2014;66(5):936-48.
69. Buttner F, Winter S, Rausch S, Hennenlotter J, et al. Clinical utility of the S3-score for molecular prediction of outcome in non-metastatic and metastatic clear cell renal cell carcinoma. BMC Med. 2018;16(1):108.
70. The Cancer Genome Atlas Research. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43-9.
71. Mitchell TJ, Turajlic S, Rowan A, Nicol D, et al. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell. 2018;173(3):611-23 e17.
72. The Cancer Genome Atlas Research, Linehan WM, Spellman PT, Ricketts CJ, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. NEJM. 2016;374(2):135-45.
73. Casuscelli J, Weinhold N, Gundem G, Wang L, et al. Genomic landscape and evolution of metastatic chromophobe renal cell carcinoma. JCI Insight. 2017;2(12).
74. Chen F, Zhang Y, Senbabaoglu Y, Ciriello G, et al. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. Cell Rep. 2016;14(10):2476-89.
75. Turajlic S, Xu H, Litchfield K, Rowan A, et al. Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell. 2018;173(3):581-94 e12.
76. Mitchell TJ, Rossi SH, Klatte T, Stewart GD. Genomics and clinical correlates of renal cell carcinoma. World J Urol. 2018;36(12):1899-911.
77. Turajlic S, Larkin J, Swanton C. SnapShot: Renal Cell Carcinoma. Cell. 2015;163(6):1556- e1.
78. Schmidt L, Junker K, Weirich G, Glenn G, et al. Two North American families with hereditary papillary renal carcinoma and identical novel mutations in the MET proto-oncogene. Cancer Res. 1998;58(8):1719-22.
79. Warren AY, Harrison D. WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. World J Urol. 2018. 2018;36(12):1913-26.
80. Maher ER. Hereditary renal cell carcinoma syndromes: diagnosis, surveillance and management. World J Urol. 2018;36(12):1891-8.
81. Verine J, Pluvinage A, Bousquet G, Lehmann-Che J, et al. Hereditary renal cancer syndromes: an update of a systematic review. Eur Urol. 2010;58(5):701-10.
82. Gerlinger M, Horswell S, Larkin J, Rowan AJ, et al. Genomic architecture & evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature genetics. 2014;46(3):225-33.
83. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.
84. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
85. Venkatesan S, Swanton C. Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. Am Soc Clin Oncol Educ Book. 2016;35:e141-9.
86. Sottoriva A, Kang H, Ma Z, Graham TA, et al. A Big Bang model of human colorectal tumor growth. Nature genetics. 2015;47(3):209-16.
87. Turajlic S, Xu H, Litchfield K, Rowan A, et al. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. Cell. 2018;173(3):595-610 e11.
88. Huang Y, Wang J, Jia P, Li X, et al. Clonal architectures predict clinical outcome in clear cell renal cell carcinoma. Nat Commun. 2019;10(1):1245.
89. Rini B, Goddard A, Knezevic D, Maddala T, et al. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. Lancet Oncol. 2015;16(6):676-85.

90. Morgan TM, Mehra R, Tiemeny P, Wolf JS, et al. A Multigene Signature Based on Cell Cycle Proliferation Improves Prediction of Mortality Within 5 Yr of Radical Nephrectomy for Renal Cell Carcinoma. Eur Urol. 2018;73(5):763-9.
91. Buttner F, Winter S, Rausch S, Reustle A, et al. Survival Prediction of Clear Cell Renal Cell Carcinoma Based on Gene Expression Similarity to the Proximal Tubule of the Nephron. Eur Urol. 2015;68(6):1016-20.
92. Brooks SA, Brannon AR, Parker JS, Fisher JC, et al. ClearCode34: A prognostic risk predictor for localized clear cell renal cell carcinoma. Eur Urol. 2014;66(1):77-84.
93. Ghatalia P, Rathmell WK. Systematic Review: ClearCode 34 - A Validated Prognostic Signature in Clear Cell Renal Cell Carcinoma (ccRCC). Kidney Cancer. 2018;2(1):23-9.
94. Brannon AR, Reddy A, Seiler M, Arreola A, et al. Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. Genes Cancer. 2010;1(2):152-63.
95. Brannon AR, Haake SM, Hacker KE, Pruthi RS, et al. Meta-analysis of clear cell renal cell carcinoma gene expression defines a variant subgroup and identifies gender influences on tumor biology. Eur Urol. 2012;61(2):258-68.
96. de Velasco G, Culhane AC, Fay AP, Hakimi AA, et al. Molecular Subtypes Improve Prognostic Value of International Metastatic Renal Cell Carcinoma Database Consortium Prognostic Model. Oncologist. 2017;22(3):286-92.
97. Bird A. DNA methylation patterns and epigenetic memory. Genes & development. 2002;16(1):6-21.
98. Field AE, Robertson NA, Wang T, Havas A, et al. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Mol Cell. 2018;71(6):882-95.
99. Xie H, Wang M, de Andrade A, Bonaldo Mde F, et al. Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic Acids Res. 2011;39(10):4099-108.
100. Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet. 2017;18(9):517-34.
101. Moss J, Magenheim J, Neiman D, Zemmour H, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9(1):5068.
102. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013;14(10):R115.
103. Lasseigne BN, Brooks JD. The Role of DNA Methylation in Renal Cell Carcinoma. Mol Diagn Ther. 2018;22(4):431-42.
104. Wagner JR, Busche S, Ge B, Kwan T, et al. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome biology. 2014;15(2):R37.
105. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nature reviews Molecular cell biology. 2019;20(10):590-607.
106. Tubio JMC, Li Y, Ju YS, Martincorena I, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science. 2014;345(6196):1251343.
107. Guo S, Diep D, Plongthongkum N, Fung HL, et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nature genetics. 2017;49(4):635-42.
108. Shi J, Xu J, Chen YE, Li JS, et al. The concurrence of DNA methylation and demethylation is associated with transcription regulation. Nat Commun. 2021;12(1):5285.
109. Feinberg AP. Epigenetic stochasticity, nuclear structure and cancer: the implications for medicine. J Intern Med. 2014;276(1):5-11.

110. Landau DA, Clement K, Ziller MJ, Boyle P, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell. 2014;26(6):813-25.
111. Hernando-Herraez I, Evano B, Stubbs T, Commere PH, et al. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. Nat Commun. 2019;10(1):4361.
112. Pan H, Jiang Y, Boi M, Tabbo F, et al. Epigenomic evolution in diffuse large B-cell lymphomas. Nat Commun. 2015;6:6921.
113. Landan G, Cohen NM, Mukamel Z, Bar A, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nature genetics. 2012;44(11):1207-14.
114. Joosten SC, Smits KM, Aarts MJ, Melotte V, et al. Epigenetics in renal cell cancer: mechanisms and clinical applications. Nat Rev Urol. 2018;15(7):430-51.
115. Lommen K, Vaes N, Aarts MJ, van Roermund JG, et al. Diagnostic DNA Methylation Biomarkers for Renal Cell Carcinoma: A Systematic Review. Eur Urol Oncol. 2021;4(2):215-26.
116. Kubiliute R, Jarmalaite S. Epigenetic Biomarkers of Renal Cell Carcinoma for Liquid Biopsy Tests. Int J Mol Sci. 2021;22(16).
117. Malouf GG, Su X, Zhang J, Creighton CJ, et al. DNA Methylation Signature Reveals Cell Ontogeny of Renal Cell Carcinomas. Clin Cancer Res. 2016;22(24):6236-46.
118. Davis CF, Ricketts CJ, Wang M, Yang L, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014;26(3):319-30.
119. Pires-Luis AS, Costa-Pinheiro P, Ferreira MJ, Antunes L, et al. Identification of clear cell renal cell carcinoma and oncocytoma using a three-gene promoter methylation panel. Journal of translational medicine. 2017;15(1):149.
120. Chopra S, Liu J, Alemozaffar M, Nichols PW, et al. Improving needle biopsy accuracy in small renal mass using tumor-specific DNA methylation markers. Oncotarget. 2017;8(3):5439-48.
121. Brennan K, Metzner TJ, Kao CS, Massie CE, et al. Development of a DNA Methylation-Based Diagnostic Signature to Distinguish Benign Oncocytoma From Renal Cell Carcinoma. JCO Precis Oncol. 2020;4.
122. Morris MR, Latif F. The epigenetic landscape of renal cancer. Nat Rev Nephrol. 2017;13(1):47-60.
123. Ricketts CJ, Hill VK, Linehan WM. Tumor-specific hypermethylation of epigenetic biomarkers, including SFRP1, predicts for poorer survival in patients from the TCGA Kidney Renal Clear Cell Carcinoma (KIRC) project. PLoS One. 2014;9(1):e85621.
124. Angulo JC, Manini C, Lopez JI, Pueyo A, et al. The Role of Epigenetics in the Progression of Clear Cell Renal Cell Carcinoma and the Basis for Future Epigenetic Treatments. Cancers (Basel). 2021;13(9).
125. Peters I, Merseburger AS, Tezval H, Lafos M, et al. The Prognostic Value of DNA Methylation Markers in Renal Cell Cancer: A Systematic Review. Kidney Cancer. 2020;4:3-13.
126. Wei JH, Haddad A, Wu KJ, Zhao HW, et al. A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. Nat Commun. 2015;6:8699.
127. Evelonn EA, Landfors M, Haider Z, Kohn L, et al. DNA methylation associates with survival in non-metastatic clear cell renal cell carcinoma. BMC cancer. 2019;19(1):65.
128. Juo YY, Johnston FM, Zhang DY, Juo HH, et al. Prognostic value of CpG island methylator phenotype among colorectal cancer patients: a systematic review and meta-analysis. Annals of oncology. 2014;25(12):2314-27.
129. Issa JP. CpG island methylator phenotype in cancer. Nature reviews Cancer. 2004;4(12):988.
130. Wang Q, Wang G, Liu C, He X. Prognostic value of CpG island methylator phenotype among hepatocellular carcinoma patients: A systematic review and meta-analysis. Int J Surg. 2018;54(Pt A):92-9.

131. Hughes LA, Melotte V, de Schrijver J, de Maat M, et al. The CpG island methylator phenotype: what's in a name? Cancer Res. 2013;73(19):5858-68.
132. Arai E, Chiku S, Mori T, Gotoh M, et al. Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. Carcinogenesis. 2012;33(8):1487-93.
133. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2020;180(1):207.
134. Ziller MJ, Gu H, Muller F, Donaghey J, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477-81.
135. Stewart GD, Powles T, Van Neste C, Meynert A, et al. Dynamic epigenetic changes to VHL occur with sunitinib in metastatic clear cell renal cancer. Oncotarget. 2016;7(18):25241-50.
136. Winter S, Fisel P, Buttner F, Rausch S, et al. Methylomes of renal cell lines and tumors or metastases differ significantly with impact on pharmacogenes. Sci Rep. 2016;6:29930.
137. Booth MJ, Ost TW, Beraldi D, Bell NM, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013;8(10):1841-51.
138. Smeets E, Lynch AG, Prekovic S, Van den Broeck T, et al. The role of TET-mediated DNA hydroxymethylation in prostate cancer. Mol Cell Endocrinol. 2018;462(Pt A):41-55.
139. Bachman M, Uribe-Lewis S, Yang X, Williams M, et al. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nat Chem. 2014;6(12):1049-55.
140. Chen K, Zhang J, Guo Z, Ma Q, et al. Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. Cell Res. 2016;26(1):103-18.
141. Liu Y, Hu Z, Cheng J, Siejka-Zielinska P, et al. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. Nat Commun. 2021;12(1):618.
142. Liu Y, Rosikiewicz W, Pan Z, Jillette N, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome biology. 2021;22(1):295.
143. Feng S, Zhong Z, Wang M, Jacobsen SE. Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics Chromatin. 2020;13(1):42.
144. Li S, Garrett-Bakelman F, Perl AE, Luger SM, et al. Dynamic evolution of clonal epialleles revealed by methclone. Genome biology. 2014;15(9):472.
145. Morrissey JJ, Mellnick VM, Luo J, Siegel MJ, et al. Evaluation of Urine Aquaporin-1 and Perilipin-2 Concentrations as Biomarkers to Screen for Renal Cell Carcinoma: A Prospective Cohort Study. JAMA Oncol. 2015;1(2):204-12.
146. Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. Cancer Res. 2012;72(23):6097-101.
147. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
148. Available from: https://www.cancerresearchuk.org/sites/default/files/diagnostic.pdf.
149. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nature reviews Cancer. 2017;17(4):223-38.
150. Duffy MJ, Diamandis EP, Crown J. Circulating tumor DNA (ctDNA) as a pan-cancer screening test: is it finally on the horizon? Clin Chem Lab Med. 2021;59(8):1353-61.
151. Liu MC, Oxnard GR, Klein EA, Swanton C, et al. Sensitive and specific multi-cancer detection & localization using methylation signatures in cell-free DNA. Annals of oncology. 2020;31(6):745-59.
152. Cheng THT, Jiang P, Tam JCW, Sun X, et al. Genome-wide bisulfite sequencing reveals the origin & time-dependent fragmentation of urinary cfDNA. Clinical biochemistry. 2017;50(9):496-501.
153. Cristiano S, Leal A, Phallen J, Fiksel J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385-9.

154. Shen SY, Singhania R, Fehringer G, Chakravarthy A, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563(7732):579-83.
155. Zill OA, Banks KC, Fairclough SR, Mortimer SA, et al. The Landscape of Actionable Genomic Alterations in Cell-Free Circulating Tumor DNA from 21,807 Advanced Cancer Patients. Clin Cancer Res. 2018;24(15):3528-38.
156. Bettegowda C, Sausen M, Leary RJ, Kinde I, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224):224ra24.
157. Zhang Y, Yao Y, Xu Y, Li L, et al. Pan-cancer circulating tumor DNA detection in over 10,000 Chinese patients. Nat Commun. 2021;12(1):11.
158. Smith CG, Moser T, Mouliere F, Field-Rayner J, et al. Comprehensive characterization of cell-free tumor DNA in plasma and urine of patients with renal tumors. Genome Med. 2020;12(1):23.
159. Nuzzo PV, Berchuck JE, Korthauer K, Spisak S, et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. Nat Med. 2020;26(7):1041-3.
160. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10(466).
161. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372(6538).
162. Cristiano S, Leal A, Phallen J, Fiksel J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385-9.
163. Heitzer E, Perakis S, Geigl JB, Speicher MR. The potential of liquid biopsies for the early detection of cancer. NPJ Precis Oncol. 2017;1(1):36.
164. Buscail E, Chiche L, Laurent C, Vendrely V, et al. Tumor-proximal liquid biopsy to improve diagnostic and prognostic performances of circulating tumor cells. Mol Oncol. 2019;13(9):1811-26.
165. Chemi F, Rothwell DG, McGranahan N, Gulati S, et al. Pulmonary venous circulating tumor cell dissemination before tumor resection and disease relapse. Nat Med. 2019;25(10):1534-9.
166. Palmela Leitao T, Miranda M, Polido J, Morais J, et al. Circulating tumor cell detection methods in renal cell carcinoma: A systematic review. Crit Rev Oncol Hematol. 2021;161:103331.
167. Wang Y, Li L, Douville C, Cohen JD, et al. Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Sci Transl Med. 2018;10(433).
168. Cohen JD, Javed AA, Thoburn C, Wong F, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(38):10202-7.
169. Li G, Bilal I, Gentil-Perret A, Feng G, et al. CA9 as a molecular marker for differential diagnosis of cystic renal tumors. Urol Oncol. 2012;30(4):463-8.
170. Risberg B, Tsui DWY, Biggs H, Ruiz-Valdepenas Martin de Almagro A, et al. Effects of Collection and Processing Procedures on Plasma Circulating Cell-Free DNA from Cancer Patients. J Mol Diagn. 2018;20(6):883-92.
171. Streck Cell-Free DNA blood collection tubes 2021. Available from: https://www.streck.com/products/stabilization/cell-free-dna-bct/.
172. Ryan MJ, Johnson G, Kirk J, Fuerstenberg SM, et al. HK-2: an immortalized proximal tubule epithelial cell line from normal adult human kidney. Kidney Int. 1994;45(1):48-57.
173. Williams RD, Elliott AY, Stein N, Fraley EE. In vitro cultivation of human renal cell cancer. II. Characterization of cell lines. In Vitro. 1978;14(9):779-86.
174. Vanharanta S, Shu W, Brenet F, Hakimi AA, et al. Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer. Nat Med. 2013;19(1):50-6.
175. Rodrigues P, Patel SA, Harewood L, Olan I, et al. NF-kappaB-Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis. Cancer Discov. 2018;8(7):850-65.
176. Jones J, Otu H, Spentzos D, Kolia S, et al. Gene signatures of progression and metastasis in renal cell cancer. Clin Cancer Res. 2005;11(16):5730-9.

177. Colaprico A, Silva TC, Olsen C, Garofano L, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.
178. Diez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. Epigenetics Chromatin. 2015;8:22.
179. Cerami E, Gao J, Dogrusoz U, Gross BE, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401-4.
180. Davis CA, Hitz BC, Sloan CA, Chan ET, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46(D1):D794-D801.
181. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25(14):1841-2.
182. Salas LA, Koestler DC, Butler RA, Hansen HM, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome biology. 2018;19(1):64.
183. Picard Tools. Broad Institute, GitHub repository. Available from: http://broadinstitute.github.io/picard/.
184. Patro R, Duggal G, Love MI, Irizarry RA, et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.
185. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.
186. National centre for biotechnology information single nucleotide polymorphisms. Homo sapiens HG38 National centre for biotechnology information; 2021. Available from: https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/00-common_all.vcf.gz.
187. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome biology. 2012;13(10):R87.
188. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics. 2019;20(3):367-83.
189. Naeem H, Wong NC, Chatterton Z, Hong MK, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics. 2014;15:51.
190. Price ME, Cotton AM, Lam LL, Farre P, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013;6(1):4.
191. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382-3.
192. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics. 2019;35(17):3151-3.
193. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
194. Klaus B, Reisenauer S. An end to end workflow for differential gene expression using Affymetrix microarrays. F1000Res. 2016;5:1384.
195. Ritchie ME, Phipson B, Wu D, Hu Y, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
196. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4:1521.
197. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.
198. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.

199. Qin Y, Feng H, Chen M, Wu H, et al. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. Genes Dis. 2018;5(1):43-5.
200. Scherer M, Nazarov PV, Toth R, Sahay S, et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz. Nat Protoc. 2020;15(10):3240-63.
201. Lutsik P, Slawski M, Gasparoni G, Vedeneev N, et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome biology. 2017;18(1):55.
202. Hastie T, Tibshirani R, Balasubramanian N, Chu G. impute: impute: Imputation for microarray data. R package version 1.66.0. 2021
203. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
204. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(39):16910-5.
205. Newman AM, Steen CB, Liu CL, Gentles AJ, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773-82.
206. Sturm G, Finotello F, List M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. Methods Mol Biol. 2020;2120:223-32.
207. Sheffield NC, Pierron G, Klughammer J, Datlinger P, et al. DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. Nat Med. 2017;23(3):386-95.
208. Restrepo P, Bubie A, Craig AJ, Labgaa I, et al. Intra-tumoral epigenetic heterogeneity and aberrant molecular clocks in hepatocellular carcinoma. Preprint on medrxiv. 2021. Available from: https://www.medrxiv.org/content/10.1101/2021.03.22.21253654v1.
209. Hua X, Zhao W, Pesatori AC, Consonni D, et al. Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma. Nat Commun. 2020;11(1):2459.
210. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526-8.
211. Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol. 2002;9(5):687-705.
212. Mazor T, Pankov A, Johnson BE, Hong C, et al. DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. Cancer Cell. 2015;28(3):307-17.
213. Liu SJ, Magill ST, Vasudevan HN, Hilz S, et al. Multiplatform Molecular Profiling Reveals Epigenomic Intratumor Heterogeneity in Ependymoma. Cell Rep. 2020;30(5):1300-9 e5.
214. Schwarz RF, Trinh A, Sipos B, Brenton JD, et al. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014;10(4):e1003535.
215. Petkovic M, Watkins TBK, Colliver EC, Laskina S, et al. Whole-genome doubling-aware copy number phylogenies for cancer evolution with MEDICC2 2021. Available from: https://www.biorxiv.org/content/10.1101/2021.02.28.433227v1.
216. Watkins TBK, Lim EL, Petkovic M, Elizalde S, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020;587(7832):126-32.
217. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25.
218. Smith MR. Information theoretic generalized Robinson-Foulds metrics for comparing phylogenetic trees. Bioinformatics. 2020;36(20):5007-13.
219. Chen X, Ashoor H, Musich R, Wang J, et al. epihet for intra-tumoral epigenetic heterogeneity analysis and visualization. Sci Rep. 2021;11(1):376.
220. Hao JJ, Lin DC, Dinh HQ, Mayakonda A, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nature genetics. 2016;48(12):1500-7.

221. Sun K, Jiang P, Chan KC, Wong J, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(40):E5503-12.
222. Lehnert B. Package BlandAltmanLeh 2015. Available from: https://cran.r-project.org/web/packages/BlandAltmanLeh/BlandAltmanLeh.pdf.
223. Robin X, Turck N, Hainard A, Tiberti N, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
224. Srivastava S, Koay EJ, Borowsky AD, De Marzo AM, et al. Cancer overdiagnosis: a biological challenge and clinical dilemma. Nature reviews Cancer. 2019;19(6):349-58.
225. British Medical Journal 'Too much Medicine'. Available from: https://www.bmj.com/too-much-medicine.
226. Chen W, Zhuang J, Wang PP, Jiang J, et al. DNA methylation-based classification and identification of renal cell carcinoma prognosis-subgroups. Cancer Cell Int. 2019;19:185.
227. Lindgren D, Eriksson P, Krawczyk K, Nilsson H, et al. Cell-Type-Specific Gene Programs of the Normal Human Nephron Define Kidney Cancer Subtypes. Cell Rep. 2017;20(6):1476-89.
228. Zhang Y, Narayanan SP, Mannan R, Raskind G, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proceedings of the National Academy of Sciences of the United States of America. 2021;118(24).
229. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene. 2002;21(35):5400-13.
230. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews Cancer. 2011;11(10):726-34.
231. Sun G, Zhang X, Liang J, Pan X, et al. Integrated Molecular Characterization of Fumarate Hydratase-deficient Renal Cell Carcinoma. Clin Cancer Res. 2021;27(6):1734-43.
232. Rathmell WK, Rathmell JC, Linehan WM. Metabolic Pathways in Kidney Cancer: Current Therapies and Future Directions. J Clin Oncol. 2018:JCO2018792309.
233. Hanahan D and Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
234. Wu P, Xiang T, Wang J, Lv R, et al. Identification of immunization-related new prognostic biomarkers for papillary renal cell carcinoma by integrated bioinformatics analysis. BMC Med Genomics. 2021;14(1):241.
235. Joshi S, Tolkunov D, Aviv H, Hakimi AA, et al. The Genomic Landscape of Renal Oncocytoma Identifies a Metabolic Barrier to Tumorigenesis. Cell Rep. 2015;13(9):1895-908.
236. Wozniak MB, Le Calvez-Kelm F, Abedi-Ardekani B, Byrnes G, et al. Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. PLoS One. 2013;8(3):e57886.
237. Devetzi M, Goulielmaki M, Khoury N, Spandidos DA, et al. Geneticallymodified stem cells in treatment of human diseases: Tissue kallikrein (KLK1)based targeted therapy (Review). Int J Mol Med. 2018;41(3):1177-86.
238. Sotiropoulou G, Pampalakis G, Diamandis EP. Functional roles of human kallikrein-related peptidases. The Journal of biological chemistry. 2009;284(48):32989-94.
239. Tailor PD, Kodeboyina SK, Bai S, Patel N, et al. Diagnostic and prognostic biomarker potential of kallikrein family genes in different cancer types. Oncotarget. 2018;9(25):17876-88.
240. Skala SL, Wang X, Zhang Y, Mannan R, et al. Next-generation RNA Sequencing-based Biomarker Characterization of Chromophobe Renal Cell Carcinoma and Related Oncocytic Neoplasms. Eur Urol. 2020;78(1):63-74.
241. Wang Q, Gan H, Chen C, Sun Y, et al. Identification and validation of a 44-gene expression signature for the classification of renal cell carcinomas. J Exp Clin Cancer Res. 2017;36(1):176.
242. Habuka M, Fagerberg L, Hallstrom BM, Kampf C, et al. The kidney transcriptome and proteome defined by transcriptomics and antibody-based profiling. PLoS One. 2014;9(12):e116125.

243. Yusenko MV, Zubakov D, Kovacs G. Gene expression profiling of chromophobe renal cell carcinomas and renal oncocytomas by Affymetrix GeneChip using pooled and individual tumours. Int J Biol Sci. 2009;5(6):517-27.
244. Tang SW, Yang TC, Lin WC, Chang WH, et al. Nicotinamide N-methyltransferase induces cellular invasion through activating matrix metalloproteinase-2 expression in clear cell renal cell carcinoma cells. Carcinogenesis. 2011;32(2):138-45.
245. Nam HY, Chandrashekar DS, Kundu A, Shelar S, et al. Integrative Epigenetic and Gene Expression Analysis of Renal Tumor Progression to Metastasis. Molecular cancer research. 2019;17(1):84-96.
246. Nagy A, Banyai D, Semjen D, Beothe T, et al. Sciellin is a marker for papillary renal cell tumours. Virchows Arch. 2015.
247. Lechpammer M, Resnick MB, Sabo E, Yakirevich E, et al. The diagnostic and prognostic utility of claudin expression in renal cell neoplasms. Mod Pathol. 2008;21(11):1320-9.
248. Yu AS. Claudins and the kidney. J Am Soc Nephrol. 2015;26(1):11-9.
249. Fritzsche FR, Oelrich B, Johannsen M, Kristiansen I, et al. Claudin-1 protein expression is a prognostic marker of patient survival in renal cell carcinomas. Clin Cancer Res. 2008;14(21):7035-42.
250. Shenoy N, Vallumsetla N, Zou Y, Galeas JN, et al. Role of DNA methylation in renal cell carcinoma. J Hematol Oncol. 2015;8:88.
251. Zhang T, Niu X, Liao L, Cho EA, et al. The contributions of HIF-target genes to tumor growth in RCC. PLoS One. 2013;8(11):e80544.
252. Cho M, Uemura H, Kim SC, Kawada Y, et al. Hypomethylation of the MN/CA9 promoter and upregulated MN/CA9 expression in human renal cell carcinoma. Br J Cancer. 2001;85(4):563-7.
253. Ashida S, Nishimori I, Tanimura M, Onishi S, et al. Effects of von Hippel-Lindau gene mutation and methylation status on expression of transmembrane carbonic anhydrases in renal cell carcinoma. Journal of cancer research and clinical oncology. 2002;128(10):561-8.
254. Pastorekova S, Gillies RJ. The role of carbonic anhydrase IX in cancer development: links to hypoxia, acidosis, and beyond. Cancer Metastasis Rev. 2019;38(1-2):65-77.
255. Jakubickova L, Biesova Z, Pastorekova S, Kettmann R, et al. Methylation of the CA9 promoter can modulate expression of the tumor-associated carbonic anhydrase IX in dense carcinoma cell lines. Int J Oncol. 2005;26(4):1121-7.
256. Nakamura J, Kitajima Y, Kai K, Hashiguchi K, et al. Expression of hypoxic marker CA IX is regulated by site-specific DNA methylation and is associated with the histology of gastric cancer. Am J Pathol. 2011;178(2):515-24.
257. Cho M, Grabmaier K, Kitahori Y, Hiasa Y, et al. Activation of the MN/CA9 gene is associated with hypomethylation in human renal cell carcinoma cell lines. Mol Carcinog. 2000;27(3):184-9.
258. Di-Poi N, Zakany J, Duboule D. Distinct roles and regulations for HoxD genes in metanephric kidney development. PLoS Genet. 2007;3(12):e232.
259. Fu Y, Li F, Zhao DY, Zhang JS, et al. Interaction between Tbx1 and Hoxd10 and connection with TGFbeta-BMP signal pathway during kidney development. Gene. 2014;536(1):197-202.
260. Shenoy US, Adiga D, Kabekkodu SP, Hunter KD, et al. Molecular implications of HOX genes targeting multiple signaling pathways in cancer. Cell Biol Toxicol. 2021.
261. Hakami F, Darda L, Stafford P, Woll P, et al. The roles of HOXD10 in the development and progression of head and neck squamous cell carcinoma (HNSCC). Br J Cancer. 2014;111(4):807-16.
262. Wang L, Chen S, Xue M, Zhong J, et al. Homeobox D10 gene, a candidate tumor suppressor, is downregulated through promoter hypermethylation and associated with gastric carcinogenesis. Mol Med. 2012;18:389-400.
263. Rivera MN, Haber DA. Wilms' tumour: connecting tumorigenesis and organ development in the kidney. Nature reviews Cancer. 2005;5(9):699-712.
264. Yang L, Han Y, Suarez Saiz F, Minden MD. A tumor suppressor and oncogene: the WT1 story. Leukemia. 2007;21(5):868-76.

265. Liu Z, Wan Y, Yang M, Qi X, et al. Identification of methylation-driven genes related to the prognosis of papillary renal cell carcinoma: a study based on The Cancer Genome Atlas. Cancer Cell Int. 2020;20:235.
266. Liu G, Guo Z, Zhang Q, Liu Z, et al. AHNAK2 Promotes Migration, Invasion, and Epithelial-Mesenchymal Transition in Lung Adenocarcinoma Cells via the TGF-beta/Smad3 Pathway. Onco Targets Ther. 2020;13:12893-903.
267. Wang M, Li X, Zhang J, Yang Q, et al. AHNAK2 is a Novel Prognostic Marker and Oncogenic Protein for Clear Cell Renal Cell Carcinoma. Theranostics. 2017;7(5):1100-13.
268. Belge H, Gailly P, Schwaller B, Loffing J, et al. Renal expression of parvalbumin is critical for NaCl handling and response to diuretics. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(37):14849-54.
269. Merseburger AS, Hennenlotter J, Simon P, Ohneseit PA, et al. Cathepsin D expression in renal cell cancer-clinical implications. Eur Urol. 2005;48(3):519-26.
270. Ketterer S, Mitschke J, Ketscher A, Schlimpert M, et al. Cathepsin D deficiency in mammary epithelium transiently stalls breast cancer by interference with mTORC1 signaling. Nat Commun. 2020;11(1):5133.
271. Trimpert C, Wesche D, de Groot T, Pimentel Rodriguez MM, et al. NDFIP allows NEDD4/NEDD4L-induced AQP2 ubiquitination and degradation. PLoS One. 2017;12(9):e0183774.
272. Dong H, Zhu L, Sun J, Zhang Y, et al. Pan-cancer Analysis of NEDD4L and Its Tumor Suppressor Effects in Clear Cell Renal Cell Carcinoma. J Cancer. 2021;12(20):6242-53.
273. Desmedt V, Desmedt S, Delanghe JR, Speeckaert R, et al. Galectin-3 in Renal Pathology: More Than Just an Innocent Bystander. Am J Nephrol. 2016;43(5):305-17.
274. Dancer JY, Truong LD, Zhai Q, Shen SS. Expression of Galectin-3 in renal neoplasms: a diagnostic, possible prognostic marker. Arch Pathol Lab Med. 2010;134(1):90-4.
275. Merseburger AS, Kramer MW, Hennenlotter J, Serth J, et al. Loss of galectin-3 expression correlates with clear cell renal carcinoma progression and reduced survival. World J Urol. 2008;26(6):637-42.
276. Moller M, Strand SH, Mundbjerg K, Liang G, et al. Heterogeneous patterns of DNA methylation-based field effects in histologically normal prostate tissue from cancer patients. Sci Rep. 2017;7:40636.
277. Brikun I, Nusskern D, Gillen D, Lynn A, et al. A panel of DNA methylation markers reveals extensive methylation in histologically benign prostate biopsy cores from cancer patients. Biomark Res. 2014;2(1):25.
278. Arai E, Wakai-Ushijima S, Fujimoto H, Hosoda F, et al. Genome-wide DNA methylation profiles in renal tumors of various histological subtypes and non-tumorous renal tissues. Pathobiology. 2011;78(1):1-9.
279. Arai E, Kanai Y, Ushijima S, Fujimoto H, et al. Regional DNA hypermethylation and DNA methyltransferase (DNMT) 1 protein overexpression in both renal tumors and corresponding nontumorous renal tissues. Int J Cancer. 2006;119(2):288-96.
280. Wanner N, Vornweg J, Combes A, Wilson S, et al. DNA Methyltransferase 1 Controls Nephron Progenitor Cell Renewal and Differentiation. J Am Soc Nephrol. 2019;30(1):63-78.
281. Patel SR, Dressler GR. The genetics and epigenetics of kidney development. Semin Nephrol. 2013;33(4):314-26.
282. Shah MM, Sampogna RV, Sakurai H, Bush KT, et al. Branching morphogenesis and kidney disease. Development. 2004;131(7):1449-62.
283. Chen Y, Zhao H, Feng Y, Ye Q, et al. Pan-Cancer Analysis of the Associations of TGFBI Expression With Prognosis and Immune Characteristics. Front Mol Biosci. 2021;8:745649.
284. McGillivray PD, Ueno D, Pooli A, Mendhiratta N, et al. Distinguishing Benign Renal Tumors with an Oncocytic Gene Expression (ONEX) Classifier. Eur Urol. 2021;79(1):107-11.
285. Sanford T, Chung PH, Reinish A, Valera V, et al. Molecular sub-classification of renal epithelial tumors using meta-analysis of gene expression microarrays. PLoS One. 2011;6(7):e21260.

286. Youssef YM, White NM, Grigull J, Krizova A, et al. Accurate molecular classification of kidney cancer subtypes using microRNA signature. Eur Urol. 2011;59(5):721-30.
287. Silva-Santos RM, Costa-Pinheiro P, Luis A, Antunes L, et al. MicroRNA profile: a promising ancillary tool for accurate renal cell tumour diagnosis. Br J Cancer. 2013;109(10):2646-53.
288. Di Meo A, Saleeb R, Wala SJ, Khella HW, et al. A miRNA-based classification of renal cell carcinoma subtypes by PCR and in situ hybridization. Oncotarget. 2018;9(2):2092-104.
289. Ibragimova I, Slifker MJ, Maradeo ME, Banumathy G, et al. Genome-wide promoter methylome of small renal masses. PLoS One. 2013;8(10):e77309.
290. Wang P, Pei X, Yin XP, Ren JL, et al. Radiomics models based on enhanced computed tomography to distinguish clear cell from non-clear cell renal cell carcinomas. Sci Rep. 2021;11(1):13729.
291. Kocak B, Kus EA, Yardimci AH, Bektas CT, et al. Machine Learning in Radiomic Renal Mass Characterization: Fundamentals, Applications, Challenges, and Future Directions. AJR Am J Roentgenol. 2020;215(4):920-8.
292. Zhu M, Ren B, Richards R, Suriawinata M, et al. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Sci Rep. 2021;11(1):7080.
293. Lasseigne BN, Burwell TC, Patil MA, Absher DM, et al. DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma. BMC Med. 2014;12:235.
294. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28.
295. Masclef L, Ahmed O, Estavoyer B, Larrivee B, et al. Roles and mechanisms of BAP1 deubiquitinase in tumor suppression. Cell Death Differ. 2021;28(2):606-25.
296. Huang Z, Hua Y, Tian Y, Qin C, et al. High expression of fructose-bisphosphate aldolase A induces progression of renal cell carcinoma. Oncol Rep. 2018;39(6):2996-3006.
297. Cai Q, Christie A, Rajaram S, Zhou Q, et al. Ontological analyses reveal clinically-significant clear cell renal cell carcinoma subtypes with convergent evolutionary trajectories into an aggressive type. EBioMedicine. 2020;51:102526.
298. Lopez JI, Guarch R, Larrinaga G, Corominas-Cishek A, et al. Cell heterogeneity in clear cell renal cell carcinoma. APMIS. 2013;121(12):1187-91.
299. Senbabaoglu Y, Gejman RS, Winer AG, Liu M, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome biology. 2016;17(1):231.
300. Krishna C, DiNatale RG, Kuo F, Srivastava RM, et al. Single-cell sequencing links multiregional immune landscapes and tissue-resident T cells in ccRCC to tumor topology and therapy efficacy. Cancer Cell. 2021;39(5):662-77 e6.
301. Chevrier S, Levine JH, Zanotelli VRT, Silina K, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell. 2017;169(4):736-49 e18.
302. Lin Q, Wagner W. Epigenetic Aging Signatures Are Coherently Modified in Cancer. PLoS Genet. 2015;11(6):e1005334.
303. Zhu B, Poeta ML, Costantini M, Zhang T, et al. The genomic and epigenomic evolutionary history of papillary renal cell carcinomas. Nat Commun. 2020;11(1):3096.
304. Hwang HS, Go H, Park JM, Yoon SY, et al. Epithelial-mesenchymal transition as a mechanism of resistance to tyrosine kinase inhibitors in clear cell renal cell carcinoma. Lab Invest. 2019;99(5):659-70.
305. Smiraglia DJ, Rush LJ, Fruhwald MC, Dai Z, et al. Excessive CpG island hypermethylation in cancer cell lines vs primary human malignancies. Human molecular genetics. 2001;10(13):1413-9.
306. He H, Dai J, Zhuo R, Zhao J, et al. Study on the mechanism behind lncRNA MEG3 affecting clear cell renal cell carcinoma by regulating miR-7/RASL11B signaling. J Cell Physiol. 2018;233(12):9503-15.

307. Rupp C, Scherzer M, Rudisch A, Unger C, et al. IGFBP7, a novel tumor stroma marker, with growth-promoting effects in colon cancer through a paracrine tumor-stroma interaction. Oncogene. 2015;34(7):815-25.
308. Suzuki H, Igarashi S, Nojima M, Maruyama R, et al. IGFBP7 is a p53-responsive gene specifically silenced in colorectal cancer with CpG island methylator phenotype. Carcinogenesis. 2010;31(3):342-9.
309. Gordon EA, Whisenant TC, Zeller M, Kaake RM, et al. Combining docking site and phosphosite predictions to find new substrates: identification of smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate. Cell Signal. 2013;25(12):2518-29.
310. Galvez-Santisteban M, Rodriguez-Fraticelli AE, Bryant DM, Vergarajauregui S, et al. Synaptotagmin-like proteins control the formation of a single apical membrane domain in epithelial cells. Nature cell biology. 2012;14(8):838-49.
311. Du Y, Wang Q, Zhang X, Wang X, et al. Lysophosphatidylcholine acyltransferase 1 upregulation and concomitant phospholipid alterations in clear cell renal cell carcinoma. J Exp Clin Cancer Res. 2017;36(1):66.
312. Siebenthall KT, Miller CP, Vierstra JD, Mathieu J, et al. Integrated epigenomic profiling reveals endogenous retrovirus reactivation in renal cell carcinoma. EBioMedicine. 2019;41:427-42.
313. Mantovani F, Collavin L, Del Sal G. Mutant p53 as a guardian of the cancer cell. Cell Death Differ. 2019;26(2):199-212.
314. Gerlinger M, Santos CR, Spencer-Dene B, Martinez P, et al. Genome-wide RNA interference analysis of renal carcinoma survival regulators identifies MCT4 as a Warburg effect metabolic target. J Pathol. 2012;227(2):146-56.
315. Fisel P, Kruck S, Winter S, Bedke J, et al. DNA methylation of the SLC16A3 promoter regulates expression of the human lactate transporter MCT4 in renal cancer with consequences for clinical outcome. Clin Cancer Res. 2013;19(18):5170-81.
316. Fisel P, Stuhler V, Bedke J, Winter S, et al. MCT4 surpasses the prognostic relevance of the ancillary protein CD147 in clear cell renal cell carcinoma. Oncotarget. 2015;6(31):30615-27.
317. Keshari KR, Sriram R, Koelsch BL, Van Criekinge M, et al. Hyperpolarized 13C-pyruvate magnetic resonance reveals rapid lactate export in metastatic renal cell carcinomas. Cancer Res. 2013;73(2):529-38.
318. Sheikh MA, Malik YS, Yu H, Lai M, et al. Epigenetic regulation of Dpp6 expression by Dnmt3b and its novel role in the inhibition of RA induced neuronal differentiation of P19 cells. PLoS One. 2013;8(2):e55826.
319. Zhao X, Cao D, Ren Z, Liu Z, et al. Dipeptidyl peptidase like 6 promoter methylation is a potential prognostic biomarker for pancreatic ductal adenocarcinoma. Biosci Rep. 2020;40(7).
320. Saied MH, Marzec J, Khalid S, Smith P, et al. Genome wide analysis of acute myeloid leukemia reveal leukemia specific methylome and subtype specific hypomethylation of repeats. PLoS One. 2012;7(3):e33213.
321. Zhang Z, Lin E, Zhuang H, Xie L, et al. Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. Cancer Cell Int. 2020;20:27.
322. McGuinness C, Wesley UV. Dipeptidyl peptidase IV (DPPIV), a candidate tumor suppressor gene in melanomas is silenced by promoter methylation. Front Biosci. 2008;13:2435-43.
323. Kang HW, Park H, Seo SP, Byun YJ, et al. Methylation Signature for Prediction of Progression Free Survival in Surgically Treated Clear Cell Renal Cell Carcinoma. J Korean Med Sci. 2019;34(19):e144.
324. Mazor T, Pankov A, Song JS, Costello JF. Intratumoral Heterogeneity of the Epigenome. Cancer Cell. 2016;29(4):440-51.
325. Chen K, Liu J, Liu S, Xia M, et al. Methyltransferase SETD2-Mediated Methylation of STAT1 Is Critical for Interferon Antiviral Activity. Cell. 2017;170(3):492-506 e14.
326. Ivanov SV, Salnikow K, Ivanova AV, Bai L, et al. Hypoxic repression of STAT1 and its downstream genes by a pVHL/HIF-1 target DEC1/STRA13. Oncogene. 2007;26(6):802-12.

327. Tessier-Cloutier B, Twa DD, Marzban M, Kalina J, et al. The presence of tumour-infiltrating neutrophils is an independent adverse prognostic feature in clear cell renal cell carcinoma. J Pathol Clin Res. 2021;7(4):385-96.
328. Wang QS, Li F, Liao ZQ, Li K, et al. Low level of Cyclin-D1 correlates with worse prognosis of clear cell renal cell carcinoma patients. Cancer Med. 2019;8(9):4100-9.
329. Hirata H, Hinoda Y, Nakajima K, Kawamoto K, et al. Wnt antagonist gene DKK2 is epigenetically silenced and inhibits renal cancer progression through apoptotic and cell cycle pathways. Clin Cancer Res. 2009;15(18):5678-87.
330. Du GW, Yan X, Chen Z, Zhang RJ, et al. Identification of transforming growth factor beta induced (TGFBI) as an immune-related prognostic factor in clear cell renal cell carcinoma (ccRCC). Aging (Albany NY). 2020;12(9):8484-505.
331. Park JH, Lee C, Suh JH, Chae JY, et al. Nuclear expression of Smad proteins and its prognostic significance in clear cell renal cell carcinoma. Human pathology. 2013;44(10):2047-54.
332. Wang H, Chong T, Li BY, Chen XS, et al. Evaluating the clinical significance of SHMT2 and its co-expressed gene in human kidney cancer. Biol Res. 2020;53(1):46.
333. Atschekzei F, Hennenlotter J, Janisch S, Grosshennig A, et al. SFRP1 CpG island methylation locus is associated with renal cell cancer susceptibility and disease recurrence. Epigenetics. 2012;7(5):447-57.
334. Uchi R, Takahashi Y, Niida A, Shimamura T, et al. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. PLoS Genet. 2016;12(2):e1005778.
335. Brookman-May SD, May M, Shariat SF, Novara G, et al. Time to recurrence is a significant predictor of cancer-specific survival after recurrence in patients with recurrent renal cell carcinoma--results from a comprehensive multi-centre database (CORONA/SATURN-Project). BJU Int. 2013;112(7):909-16.
336. Hu X, Estecio MR, Chen R, Reuben A, et al. Evolution of DNA methylome from precancerous lesions to invasive lung adenocarcinomas. Nat Commun. 2021;12(1):687.
337. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6:8971.
338. Borcherding N, Vishwakarma A, Voigt AP, Bellizzi A, et al. Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. Commun Biol. 2021;4(1):122.
339. Simonaggio A, Epaillard N, Pobel C, Moreira M, et al. Tumor Microenvironment Features as Predictive Biomarkers of Response to Immune Checkpoint Inhibitors (ICI) in Metastatic Clear Cell Renal Cell Carcinoma (mccRCC). Cancers (Basel). 2021;13(2).
340. Geissler K, Fornara P, Lautenschlager C, Holzhausen HJ, et al. Immune signature of tumor infiltrating immune cells in renal cancer. Oncoimmunology. 2015;4(1):e985082.
341. Morris MR, Ricketts C, Gentle D, Abdulrahman M, et al. Identification of candidate tumour suppressor genes frequently methylated in renal cell carcinoma. Oncogene. 2010;29(14):2104-17.
342. van Vlodrop IJ, Baldewijns MM, Smits KM, Schouten LJ, et al. Prognostic significance of Gremlin1 (GREM1) promoter CpG island hypermethylation in clear cell renal cell carcinoma. Am J Pathol. 2010;176(2):575-84.
343. van Vlodrop IJ, Niessen HE, Derks S, Baldewijns MM, et al. Analysis of promoter CpG island hypermethylation in cancer: location, location, location! Clin Cancer Res. 2011;17(13):4225-31.
344. Klughammer J, Kiesel B, Roetzer T, Fortelny N, et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat Med. 2018;24(10):1611-24.
345. Taylor WC. Comment on 'Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA' by M. C. Liu et al. Annals of oncology. 2020;31(9):1266-7.
346. Pal SK, Sonpavde G, Agarwal N, Vogelzang NJ, et al. Evolution of Circulating Tumor DNA Profile from First-line to Subsequent Therapy in Metastatic Renal Cell Carcinoma. Eur Urol. 2017;72(4):557-64.

347. Wan JCM, Heider K, Gale D, Murphy S, et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Sci Transl Med. 2020;12(548).
348. Zengin ZB, Weipert C, Salgia NJ, Dizman N, et al. Complementary Role of Circulating Tumor DNA Assessment and Tissue Genomic Profiling in Metastatic Renal Cell Carcinoma. Clin Cancer Res. 2021;27(17):4807-13.
349. Green EA, Li R, Albiges L, Choueiri TK, et al. Clinical Utility of Cell-free and Circulating Tumor DNA in Kidney and Bladder Cancer: A Critical Review of Current Literature. Eur Urol Oncol. 2021;4(6):893-903.
350. Chen X, Gole J, Gore A, He Q, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 2020;11(1):3475.
351. Klein EA, Richards D, Cohn A, Tummala M, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Annals of oncology. 2021;32(9):1167-77.
352. Lasseter K, Nassar AH, Hamieh L, Berchuck JE, et al. Plasma cell-free DNA variant analysis compared with methylated DNA analysis in renal cell carcinoma. Genet Med. 2020;22(8):1366-73.
353. Liang N, Li B, Jia Z, Wang C, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. Nat Biomed Eng. 2021;5(6):586-99.
354. Hong SR, Shin KJ. Bisulfite-Converted DNA Quantity Evaluation: A Multiplex Quantitative Real-Time PCR System for Evaluation of Bisulfite Conversion. Front Genet. 2021;12:618955.
355. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res. 2021.
356. Murray MJ, Watson HL, Ward D, Bailey S, et al. "Future-Proofing" Blood Processing for Measurement of Circulating miRNAs in Samples from Biobanks and Prospective Clinical Trials. Cancer Epidemiol Biomarkers Prev. 2018;27(2):208-18.
357. Yamamoto Y, Uemura M, Fujita M, Maejima K, et al. Clinical significance of the mutational landscape and fragmentation of circulating tumor DNA in renal cell carcinoma. Cancer Sci. 2019;110(2):617-28.
358. Klein B, Clinnick L, Chesler J, Stranieri A, et al. Supporting Regional Aged Care Nursing Staff to Manage Residents' Behavioural and Psychological Symptoms of Dementia, in Real Time, Using the Nurses' Behavioural Assistant (NBA): A Pilot Site 'End-User Attitudes' Trial. Stud Health Technol Inform. 2018;246:24-8.
359. Semeniuk-Wojtas A, Lubas A, Stec R, Syrylo T, et al. Neutrophil-to-lymphocyte Ratio, Platelet-to-lymphocyte Ratio, and C-reactive Protein as New and Simple Prognostic Factors in Patients With Metastatic Renal Cell Cancer Treated With Tyrosine Kinase Inhibitors: A Systemic Review and Meta-analysis. Clin Genitourin Cancer. 2018;16(3):e685-e93.
360. Harvey-Kelly LLW, Harrison H, Rossi SH, Griffin SJ, et al. Public attitudes towards screening for kidney cancer: an online survey. BMC Urol. 2020;20(1):170.
361. Cohen JD, Li L, Wang Y, Thoburn C, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359(6378):926-30.

Appendix 1

Accurate detection of benign and malignant renal tumour sub-types with MethylBoostER: an epigenetic marker driven learning framework

Sabrina H. Rossi[1†], Izzy Newsham[2†], Sara Pita[1], Kevin Brennan[3], Gahee Park[1], Christopher G. Smith[4,5], Radoslaw P. Lach[1], Thomas Mitchell[6,7], Junfan Huang[2], Anne Babbage[1], Anne Y. Warren[8], John T. Leppert[9], Grant D. Stewart[6], Olivier Gevaert[3], Charles E. Massie[1*], Shamith A. Samarajiwa[2*]

[1]Department of Oncology, University of Cambridge; Hutchison-MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK.

[2]MRC Cancer Unit, University of Cambridge; Hutchison-MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK.

[3]Stanford Center for Biomedical Informatics Research, Department of Medicine and Department of Biomedical Data Science, Stanford University; Stanford, CA, USA.

[4]Cancer Research UK Cambridge Institute, University of Cambridge; Cambridge, UK.

[5]Cancer Research UK Major Centre, Cambridge, UK.

[6]Department of Surgery, University of Cambridge; Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, UK.

[7]Wellcome Trust Sanger Institute; Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

[8]Department of Histopathology, University of Cambridge; Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, UK.

[9]Department of Urology, Stanford University School of Medicine, Stanford University; Stanford, CA, USA.

[†]Equal contribution

[*]To whom correspondence should be addressed;E-mail: ss861@cam.ac.uk, cem45@cam.ac.uk

One Sentence Summary:

MethylBoostER is a machine learning model which uses DNA methylation data from tissue samples to accurately predict pathological sub-types of renal tumours, with the aim of improving diagnostic accuracy and therefore tailoring patient treatment.

Abstract:

The incidence of renal cell carcinoma (RCC), and small renal masses (SRMs) in particular, is rising due to widespread incidental detection. Imaging and renal biopsy are unable to accurately differentiate malignant from benign tumours, consequently 20% of patients with SRMs undergo unnecessary surgery for benign disease. We therefore develop MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours), a machine learning model leveraging DNA methylation data from >1500 tissue samples, to classify pathological sub-types of renal tumours (clear cell (ccRCC), papillary (pRCC), chromophobe (chRCC), benign oncocytoma) and normal kidney. DNA methylation data were integrated from our dataset and publicly available sources (N=1228) and used to train and test XGBoost models using four-fold cross validation. The prediction accuracy in the testing set was 0.960; with high class-wise Area Under the Receiver Operating Characteristic Curves (ROC AUC's): 0.994 (ccRCC), 0.992 (pRCC), 0.988 (chRCC), 1.00 (oncocytoma), 1.00 (normal). External validation was performed on >500 samples from four independent datasets, achieving AUC's >0.89 for all predicted classes. MethylBoostER distinguishes between high and moderate-confidence predictions (based on predicted probabilities >0.85), resulting in an output which is interpretable and clinically useful. Furthermore, MethylBoostER provides consistent classification of multi-region samples from the same patient in 90% of individuals, demonstrating methylation heterogeneity does not limit model applicability. We also explore the impact of tumour purity on predictions to represent real-world clinical practice. In summary, MethylBoostER accurately differentiates pathological sub-types of renal tumours and may provide a more confident pre-surgical diagnosis to guide treatment decision-making.

Main Text:

INTRODUCTION

Renal cell carcinoma (RCC) is amongst the top 10 most common cancers worldwide. Incidence rates have increased by 47% in the last ten years, making it one of the fastest accelerating cancers, with rates projected to rise even further in the future (*1*). This is attributed to the increasing prevalence of known risk factors and increased incidental detection, secondary to widespread use of imaging for unrelated symptoms (*2*). Small renal masses (SRMs), defined as <4cm in size, are therefore increasingly detected and encompass a mixture of potential diagnoses that comprise dozens of histological and molecular tumour sub-types (*3, 4*). The major sub-types include clear cell (ccRCC), papillary (pRCC), chromophobe (chRCC) RCC or benign oncocytomas. Differentiating malignant from benign solid SRM represents a challenge both on imaging and renal biopsy. Renal biopsy may be non-diagnostic due to small tumour size, challenging anatomy, low biopsy tumour content or tumour heterogeneity (*5–7*). As a result of no biopsy being undertaken or the inability to conclude a histological diagnosis based on biopsy samples, approximately 20% of SRMs removed at surgery are found to be benign post-operatively (*8*). Consequently, a significant number of patients undergo unnecessary surgery for a benign condition, with associated post-operative risks of morbidity and mortality, and long-term impact on renal function. The rates of post-operative complications following minimally invasive surgery are: blood transfusion (5%), re-operation (2-5%), respiratory complications (1-7%) and death (4%) (*9*). Conversely, a meta-analysis demonstrated one in four renal biopsies reported as oncocytoma are found to be malignant RCC following surgical excision (*5*), risking false negatives, late diagnosis and deleterious outcomes. Improved diagnosis and differentiation of SRM has been identified as a key research focus in RCC by an international research priority setting initiative (*10*). Each of these pathological sub-types has characteristic genetic and molecular

features, such that it is argued that RCC is not a single disease (*11*). ccRCC and pRCC are derived from the proximal convoluted tubule in the renal cortex, whereas chRCC and oncocytoma are derived from the distal nephron (*12, 13*). DNA methylation changes are abundant, genome-wide early events in renal tumourigenesis, and are specific to the cell of origin, and so represent an ideal diagnostic target in this setting (*14–17*). Analysis of DNA methylation markers therefore has the potential to support and improve the current diagnostic pathway for renal cancer, providing a much more confident pre-surgical diagnosis to guide treatment decision-making.

In this study, we propose MethylBoostER (Methylation and XGBoost for Evaluation of Renal tumours). MethylBoostER is a machine learning model based on XGBoost that differentiates pathological sub-types of renal tumours, using DNA methylation markers identified in large tissue datasets. We externally validate MethylBoostER on four independent datasets and also evaluate multi-region samples from the same patient to assess the role of intra-tumoural heterogeneity on MethylBoostER's performance. Additionally, we externally validate our results on ex-vivo core biopsy tissue samples (which simulate in-vivo patient biopsies) and fine needle aspirates (FNA), low purity samples representing real world applications and evaluate the role of tumour purity on MethylBoostER's performance to assess clinical utility. MethylBoostER is trained on the largest DNA methylation cohort of renal tumour samples to date and is extensively validated, demonstrating accurate predictions across international cohorts and demonstrates potential future clinical utility.

RESULTS

DNA methylation sequencing in a cohort of patients with renal tumours

DNA methylation data were obtained for 1228 samples from three data sources, namely Cambridge samples (N=319), The Cancer Genome Atlas (TCGA) (N=872) and the Chopra training set (*18*) (N=37) (Figure 2a and b). The former dataset contained information on 3 million CpGs

measured using EPIC-seq, whilst the latter two were obtained via the 450K Illumina array. The merged datasets were obtained by overlapping CpGs within +/- 50bp of Illumina probes and overall contained data on 158,670 probes. This dataset (N=1228) was split into training and testing sets (via four-fold cross validation, see Methods), with subsequent validation being performed on four additional independent datasets (Chopra validation (*18*), Wei (*19*), Brennan (*20*) and Evelönn (*21*)), and will be referred to as such hereafter.

MethylBoostER accurately distinguishes pathological sub-types of renal tumours

We trained multiclass XGBoost machine learning models to classify four pathological sub-types of renal tumours (ccRCC, chRCC, oncocytoma, pRCC) and normal tissue. Four-fold nested cross validation was used (a method to randomly split the dataset into training (75%) and testing (25%) sets over four iterations) so that the testing set performance could be evaluated over the whole training/testing set. Consequently, this results in four trained models. We display summary results for the four models (individual results displayed in Figures S1, S2, S4, S5 and S6).

The performance over the training/testing set was high, with an average accuracy of 0.960. Table S1 lists additional performance metrics, including metrics for each renal tumour sub-type (forthwith called 'class'), which shows that the models could distinguish the normal class perfectly, and the chRCC class was the most difficult to classify correctly. Figure 2c shows the confusion matrix, which depicts how many samples, of a given true class, were predicted as each class. Most samples (1179/1228) are on the diagonal, indicating they were correctly predicted. Out of the 1228 samples evaluated, common misclassifications were in predicting ccRCC as pRCC (13 samples) and predicting pRCC as ccRCC or chRCC (9 and 8 samples, respectively). Importantly, although 3 oncocytoma samples were predicted as malignant RCC,

no malignant samples were classed as benign oncocytoma or normal tissue. Figure 2e (and Figure S1) shows the Receiver Operating Characteristic (ROC) curves (and Precision-Recall curves) for each class, again indicating that the models achieved a high testing set performance.

A large motivating factor for using XGBoost models is their interpretability – we can examine which input features were learnt by the models when classifying the five classes. We selected only these learned features, and used them to visualise the training/testing set, as shown in Figure 2d (and Figure S2). This shows that although there are dataset-specific clusters within each class, the samples now form class-wise clusters. This suggests that the features that the models learn are not dataset specific but are features that distinguish each class in all three datasets.

High- and moderate-confidence predictions make model outputs clinically more informative

Given a sample and a trained XGBoost model, we can obtain the model's predicted probabilities for that sample belonging to each class (renal tumour sub-type). This reveals the confidence of the model's prediction. Figure 3a shows the predicted probabilities of the predicted class over the testing sets, which indicates that while most predictions are certain, some are less certain (13.2% of samples were predicted with a probability of < 0.95), with probabilities as low as 0.256.

We utilised this information to separate the models' predictions into two categories: high-confidence and moderate-confidence predictions. High-confidence predictions are where the model's confidence is greater than a threshold, and we output a single answer – the predicted class. Moderate-confidence predictions are where the model's confidence is lower than the threshold, and we output two answers – the predicted class (the first most likely prediction) and the second most likely prediction. We refer to these two outputs as *first prediction* and *second prediction*, respectively. The threshold was set to 0.85 (see Figure 3b), which maximises the

accuracy of both high- and moderate-confidence predictions and the fraction of high-confidence predictions over the testing set. In clinical practice in the case of moderate predictions, the clinician could take into account the two predicted classes and integrate this information with clinical, histopathology and imaging features to conclude the most likely diagnosis (see Figure 3c). The average accuracy of the high-confidence predictions over the testing sets was 0.982. The average accuracy of the moderate-confidence predictions over the testing sets was 0.668 (considering only the first prediction), but was 0.871 when the prediction was treated as correct if the *first or second* prediction was correct. See Figure S3 for the testing set confusion matrix split by high- and moderate-confidence predictions.

External validation on four independent datasets verify the generalisability of MethylBoostER to new data

We externally validated MethylBoostER on four independent datasets, namely Chopra validation (N=245), Brennan (N=37), Wei (N=92) and Evelönn (N=144) (*18–21*). They each contain different numbers of samples and a different class distribution, as shown in Figure 4a. The normal and ccRCC classes are the most frequent, which reflects clinical prevalence. We evaluated our models on these datasets (without the high- and moderate-confidence predictions method, so just taking the first prediction for all samples) and the average accuracy for Chopra validation, Brennan, Wei and Evelönn was 0.824, 0.703, 0.875, and 0.894 respectively (see Table S2 for additional metrics, and Figure S4a for confusion matrices).

The models had lower performance for chRCC and oncocytoma - oncocytoma samples were frequently predicted as chRCC or normal. In Chopra validation, the oncocytomas predicted incorrectly on both the first and second prediction were predicted as normal (11 samples), chRCC (4 samples) and pRCC (2 samples). In Brennan, the incorrect oncocytomas (on both first and second predictions) were predicted as normal (1 sample), and chRCC (5 samples). The rel-

atively lower performance in chRCC and oncocytoma may be due to the low proportion of these classes in the training set, reflecting a real-world challenge, as these two are the least common pathological sub-types. In addition, previous reports suggest that methylation in these two classes are most similar, in keeping with their common cellular origin (*13*). We found that incorporating high- and moderate-confidence predictions lead to improved results. Figure 4b and Table S3 show the accuracy and Matthews Correlation Coefficient (MCC) scores for the high- and moderate-confidence predictions. As seen in Figure 4b, the moderate-confidence first predictions have an accuracy of < 0.65 for all four datasets, and the second predictions have an accuracy of < 0.45 for all four datasets. When we combine them, we see that the first or second predictions reach an accuracy of > 0.7. As expected, the accuracy of the high-confidence predictions is higher, and is > 0.9 for all datasets.

Many moderate-confidence predictions are correct on the second prediction, as shown in Figures 4c,d,e,f (and Figure S4b for the other models). For example, 16 samples in the Chopra validation dataset had an incorrect first prediction but correct second prediction (see Figure 4c). These samples would have simply been incorrectly predicted without the use of this high- and moderate-confidence prediction method. We also show that all four datasets achieved high class-wise ROC AUC's, as shown in Figures 4c,d,e,f (see Figure S5 for the other models' ROC curves and all Precision-Recall curves).

MethylBoostER provides consistent classification of multi-region samples from the same patient

ccRCC is characterised by a high degree of genetic intratumoral heterogeneity (ITH), with over 60% of somatic mutations not detectable across all multi-region samples (*22*). On average, seven multi-region samples are needed to detect over 75% of genetic variants (*23*). Very little is known about methylation variation in ccRCC, although a handful of reports suggests relative

homogeneity (*24–27*), therefore highlighting the potential relevance of diagnostic methylation markers. In the Cambridge dataset, multi-region samples were available (N=168) for 25 patients (18 ccRCC, 4 chRCC, 2 oncocytoma and 1 pRCC). In 92% (23/25) of patients, all multi-region samples were predicted consistently as being from the same pathological sub-type; with 88% (22/25) achieving correct classification for all samples (see Figure 5a). Multi-region samples (N=17) were also available for 6 patients with ccRCC from the Evelönn external validation dataset (*24*). As shown in Figure 5b, 83% (5/6) of patients had a concordant prediction for all multi-region samples derived from the same patient, although the model struggles to differentiate chRCC from ccRCC. We noted that incorrectly predicted samples had a very low tumour purity. When the Evelönn dataset is visualised using dimensionality reduction (see Figure S6), almost all the incorrectly predicted samples (including all the incorrectly predicted multi-region samples) can be seen to cluster away from the rest of the ccRCC samples, indicating these samples have a different methylation pattern (potentially related to their lower tumour purity, as shown in Figure S6b).

Impact of Tumour purity on MethylBoostER

Tumour purity was calculated using methylation beta values and InfiniumPurify, as previously described (*20, 28*). This method determines tumour purity in the context of contamination with normal (non-tumour) kidney sample. As expected, median sample purity was lower for the Chopra and Brennan datasets compared to TCGA and Cambridge, as the former include ex-vivo core biopsy samples (i.e., simulated biopsies using nephrectomy specimens) and fine needle aspirates respectively. Out of all the datasets, the Evelönn data appear to have the lowest tumour purity (median purity: 0.88 Cambridge, 0.84 TCGA, 0.80 Chopra training, 0.58 Chopra validation, 0.73 Brennan, 0.81 Wei and 0.48 Evelönn). Figure 6a summarizes purity for samples that were predicted correctly in the first prediction, second prediction and incorrectly predicted

samples (results for individual datasets are shown in Figure S7). Median purity in samples that were incorrectly predicted was significantly lower than samples correctly predicted on the first or second prediction (0.27 vs 0.82 vs 0.44, p value $<$0.01 for all comparisons; Figure 6a). We noted that three of the samples that were misclassified by our model were also incorrectly classified by Brennan et al. (*20*) and that the study authors postulated this was due to low sample purity. Figure 6b and c summarise the purity and prediction probability (for the first prediction), highlighting incorrectly predicted samples. We found a correlation between purity and the probability of first prediction in Wei and Evelönn (Pearson's's correlation coefficient 0.58 and 0.51, adjusted p value < 0.01), but not the other datasets (correlation < 0.30 and/or adjusted p value > 0.01). There are a number of samples which are incorrectly predicted despite having a high-confidence prediction, and these tend to have lower purity. Taking into account all datasets combined, Table S4 summarises the model accuracy, as well as the median probability of the first prediction, by tumour purity. This further demonstrates that samples with higher tumour purity are more accurately predicted, for example samples with purity ≥ 0.9 obtain an accuracy of 0.99. There is a sharp drop-off in accuracy in samples with a purity < 0.20 compared to samples in which purity was > 0.2, suggesting that potentially a biopsy sample may have to be repeated if purity is below this level. This analysis was performed post-hoc on both training/testing and validation datasets, and there were a limited number of low purity samples (5% of the cohort; Table S4). Therefore, these results (and a purity threshold that warrants repeat biopsy) remain to be externally validated.

Features selected by MethylBoostER are associated with carcinogenic processes

The features that the MethylBoostER models utilised during classification were obtained. Each of the four XGBoost models selected 1490, 1697, 1476 and 1331 features, respectively. An

analysis of the genomic location of the features revealed that the selected features do not follow the same genomic location distribution as the background, as shown in Figure 7a. In the promoter regions close to the TSS ($\leq$1kb), there is a lower percentage of features compared to the background set (68% compared to 78%). However, in the promoter regions further away from the TSS (1-2kb and 2-3kb away), there is a higher percentage of features compared to the background set (8.7% compared to 6.3%, and 4.3% compared to 2.3%). There is also a higher percentage of features in introns (5.1% compared to 3.1% for 1st introns, and 8.7% compared to 5.5% for other introns).

The features from MethylBoostER were mapped to proximal genes and analysed at the gene ontology and pathway level. The enriched Gene Ontology (GO) terms are visualised as a network in Figure 7b. The GO analysis detected a neuronal and muscle developmental signature, carcinogenic signalling (Wnt, MAPK, and Ras signalling pathways) and both positive and negative regulation of transcription, amongst others. The significant enrichment (see Table S5 for the adjusted p-values of all gene list comparisons) of transcription factors (TF Checkpoint (*29*)) and epigenetic regulators (Epifactor db (*30*)) indicate the perturbation of transcriptional regulators and chromatin remodelers in renal cancer. Whilst an RCC related gene set (RCC Harmonizome/Diseases db (*31*)) and a ccRCC related putative driver gene set (*32*) were enriched, a similar set of pRCC genes was not enriched (pRCC Harmonizome/Diseases db (*31*)). The feature set also indicated a strong metastatic signature (Human Cancer Metastasis Database (*33*)), with enriched gene ontology terms related to Motility, Axon guidance, and Cell adhesion. Enrichment of epithelial mesenchymal transition (EMT)-related genes (dbEMT2 (*34*)) provide further evidence for this metastatic signature. In addition, the gene list was enriched for tumour suppressors, oncogenes, and fusion genes (COSMIC Cancer Gene Census (*35*)).

We also checked whether the results of the gene-wise GO analysis were consistent with the results from the Genomic Regions Enrichment of Annotations Tool (GREAT) (*36*), which

is designed for localised regions that are not necessarily within genes. This also enabled the features that were not mapped to genes to be included. Figure 7c shows the significant GO terms resulting from this analysis, which demonstrates that they are similar to the results of the gene-wise GO analysis. The consistent themes were development, cell adhesion, cell motility, cell signalling and neurogenesis. A few functions were unique to this GREAT analysis, such as immune response and extracellular matrix organization, and a few were unique to the gene-wise analysis, such as regulation of transcription and response to cortisol stimulus.

Pathway analysis revealed enriched pathways covering similar functions as found during the GO analysis. The most enriched pathways were: Wnt signalling pathway (adj.p-value=$2.79\text{x}10^{-15}$), Cadherin signalling pathway (adj.p-value=$2.92\text{x}10^{-5}$) and Netrin-1 signalling (adj.p-value=$4.62\text{x}10^{-03}$). Other enriched pathways included axon guidance, cell junction organization and signalling by Receptor Tyrosine Kinases.

Next, we focused our analysis on the consistently salient features – features that were selected by all four XGBoost models. There were 38 such features, which mapped to 45 genes. Out of these 45 genes common to all 4 models, 9 have not been previously associated with kidney function, renal disease or cancer (ANKMY1, EHMT1, H2AW, JDP2, H2BU1, KIAA1143, KIF15, LINC01960, SF3A2). As detailed in Table S6, an analysis of the literature shows that the others have been associated with ccRCC (*37–39*), chRCC and renal cell carcinoma (*40–42*). These genes have also been linked to kidney development (*43*), gene expression in kidney, adhesion, motility, invasion, and metastasis (*44–47*), and poor survival in renal cancer and other cancer types (*48–50*). In addition, they contribute to increased survival, act as tumour suppressors, promote cancer stemness (*44*), contribute to carcinogenesis, proliferation, and tumour growth and have been identified in other cancer types. These associations demonstrate that the methylation features identified by MethylBoostER are biologically relevant. The methylation distribution of all 38 features and their genes can be seen in Figure S8. We confirmed

these findings by carrying out an entity-relationship natural language processing (NLP) analysis of approximately 30 million PubMed abstracts, using the feature mapped gene set and a dictionary of relationship terms based on enriched biological processes identified above. Modelling this information from 28579 interactions (each supported by multiple publications) as an entity-relationship network, and subsequent network topological analysis where relationship strength was visualised as degree (number of genes directly linked to a relationship) and relationship importance as betweenness centrality highlighted the association of these genes with carcinogenesis, development, and tumour microenvironment interactions. This corroborates the functional bioinformatics analyses and confirms that these genes have been previously associated with carcinogenic processes, tumour micro-environment, metabolism, metastasis, immune and inflammatory responses and are known to influence patient survival and prognosis in the biomedical literature (Figure S10 and Supplementary File-1 containing NLP derived gene relationship pairs and their frequency).

DISCUSSION

The rising incidence of SRMs drives the need to improve the diagnostic pathway, to avoid over-treatment of patients with benign disease and optimize health resource use. The challenge consists of differentiating benign oncocytoma from malignant sub-types, the most common being ccRCC, pRCC and chRCC. Previously published molecular classifiers are often limited by focusing solely on distinguishing oncocytoma from chRCC, excluding the other, more common sub-types and therefore reducing applicability in the real world (*20, 51–54*). In addition, a number of these studies had small sample sizes ($<$200 samples) (*52, 54, 55*). Existing models often rely on a limited number of markers (e.g. $<$100 markers), making them less robust when applied to heterogeneous clinical samples and limiting potential future applicability (*20, 52–54, 56*) .

Machine learning models based on DNA methylation have demonstrated excellent accu-

racy in other cancer types, including lung, brain, breast malignancies and sarcomas (*57–59*). The ability of machine learning models trained on DNA methylation data to classify cancer of unknown primary with excellent accuracy, leading to benefits in overall survival, serves as a testament to DNA methylation as a unique marker of cell identity (*60*). We have therefore developed an XGBoost machine learning model (MethylBoostER), leveraging methylation data from over 1200 patient samples. We demonstrate that the model accurately predicts pathological sub-types of renal tumours in the training and testing set, with ROC AUCs over 0.95 for all sub-types evaluated. The extensive external validation in four independent methylation datasets, totalling over 500 patient samples, is a measure of the robustness of our method. We demonstrate the high accuracy of the model with AUCs over 0.89 for all sub-types, in all independent datasets. Unique features of our model are its interpretability, ease of use in a clinical setting and an approach that reduces the clinical importance of ITH. We envision that in future, following further validation, MethylBoostER could be integrated in clinical practice to improve the patient diagnostic pathway (Figure 8). Individuals with SRMs would undergo imaging-guided tumour biopsy, which would be processed for DNA methylation and the MethylBoostER model would be used to predict pathological sub-types, serving as an adjunct to treatment decision-making. For high-confidence predictions the output is limited to the most likely diagnosis, whereas for moderate-confidence predictions the model will supply two class predictions and the clinician would be encouraged to integrate imaging and histological data. A strength of our work is the output of moderate and high-confidence predictions, which overall increases model accuracy and empowers clinicians to make patient-centred decisions which take into account both clinical and methylation data. A recognized clinical challenge is that biopsy accuracy may be hampered by genetic ITH or reduced tumour content, especially in smaller or more difficult to access tumours. We therefore address these potential drawbacks. We demonstrate that unlike the known extensive genetic ITH (*22*), analysing methylation patterns in multi-region samples

from the same patient lead to consistent diagnoses in the vast majority of cases (92% in the Cambridge and 83% in the Evelönn datasets). In addition, we assessed low purity samples (including ex-vivo biopsies and fine needle aspirates) to represent a real world scenario. We show that samples that are incorrectly predicted by our model have significantly lower purity than correctly predicted samples. The association between purity and model accuracy was also previously noted by Brennan et al. (*20*). It is customary to take more than one sample at the time of biopsy, which may help overcome this challenge, alternatively low purity samples may require repeat biopsy.

In the testing set, the most common misclassifications involved ccRCC, pRCC and chRCC, whilst in the external validation sets, chRCC and oncocytoma were the classes with the lowest AUCs. The tumours' shared cells of origin (proximal vs distal nephron) may explain some of these results. Problems in accurately differentiating between ccRCC and chRCC on histopathology are another well-known challenge. In TCGA, 15 cases were initially classified as ccRCC on histopathological slide review, however, these were later re-reviewed by specialist uro-histopathologists and reclassified as chRCC (*11, 61*). We noted that 8 of these samples were included in our testing and training dataset, and MethylBoostER classifies 5 as chRCC and 3 as ccRCC. The former suggests that our model can correctly classify chRCC samples better than a standard pathologist, and more akin to a specialist uro-histopathologist. We explored the methylation and gene expression profiles of TCGA samples (Figures S9a and b) and demonstrate that the 3 samples which our model classifies as ccRCC (TCGA participant IDs 4821, 4688 and 4696), cluster more closely with ccRCC than chRCC based on both methylation and gene expression data. We hypothesize that these samples may indeed be ccRCC (i.e., the first classification was correct rather than the re-classification). This highlights existing challenges in diagnosing sub-types using existing histopathology methods and emphasizes the need to produce accurate predictive models. Finally, another challenge is that predictive models are trained

on datasets in which the true diagnosis is based on histopathology, and if this is incorrect, it may bias the model. To ensure that the predictions of MethylBoostER were not heavily biased by these 8 mislabelled samples, we re-trained the models with the re-classified labels (8 ccRCCs were re-classified as chRCC, and a number of non-RCC were samples removed). The results over the testing sets were similar, as shown in Figures S9c and d, demonstrating that the small number of incorrect labels do not largely affect the results.

Carcinogenesis is invariably accompanied by perturbations in gene and epigenome regulation. Aberrant DNA methylation changes on promoters, enhancers, and gene bodies contribute to alteration of gene expression and consequently affect signalling, regulatory and metabolic pathways. Renal carcinoma has been associated with MET (*62*), Hippo (*63*), Wnt (*64*), MAPK (*65*), NRF-ARE, PI3K/AKT/mTOR (*66*), metabolic, angiogeneic and immune checkpoint associated pathways (*67*). Gene ontology and pathway enrichment analysis of the MethylBoostER features mapped to genes show a strong association with genes, processes and pathways involved with carcinogenesis. Particularly, oncogenes and tumour suppressors (COSMIC cancer gene census), cancer-associated pathways (KEGG), putative ccRCC driver genes, and RCC associated genes were enriched. In addition, biological processes associated with metastasis (motility, cell adhesion and axonal guidance) and EMT were also strongly enriched. Comparison with curated gene sets in metastasis (HCMDB) and EMT (dbEMT2) databases further confirmed this. Other signatures such as neuronal and muscle development, Wnt, MAP Kinase and Ras signalling pathways, nucleoside metabolism, T-cell activity are also enriched. Unsurprisingly, enrichment of both positive and negative regulation of transcription coincides with known consequences of DNA methylation. Furthermore, enrichment of TFs and epigenetic regulators further supported this strong transcriptional regulation signature. Interestingly, genes associated with response to cortisol stimulus are enriched in our feature set. Serum cortisol levels have been found to be significantly higher in RCC and are associated with higher tu-

mour grade (*68*). The enrichment in our feature set suggest a possible epigenetic regulatory mechanism for this process. Finally, we show that the CpG features selected by the models are enriched for cancer-related genes, and the features selected by all four models are enriched for well-known renal cell carcinoma genes. We use literature analysis to demonstrate that the majority of classification features are associated with genes involved in kidney pathology, cancer or carcinogenic processes, demonstrating the clear discernment in model selected methylation patterns. Associations between methylation and gene expression are outside the scope of the present analysis, which aimed to demonstrate that the model selects salient features which are biologically relevant and easily interpretable.

In summary, we develop MethylBoostER, a machine learning model that predicts pathological sub-types of benign and malignant renal tumours. The relatively modest number of patients diagnosed with certain tumour sub-types (especially chRCC and oncocytoma) and limited access to patient samples can hamper the application of machine learning based approaches. We test the model on renal tumours of any stage, to increase the sample size, and demonstrate high accuracy in the external validation set consisting solely of SRM (e.g., Chopra validation dataset) and low purity samples (e.g. Evelönn and fine needle aspirates in Brennan). Future studies should aim to obtain larger sample sizes, and focus on multi-modal integration with imaging, patient clinical characteristics and epigenetic data.

MATERIALS AND METHODS

Cambridge samples

Patients with benign and malignant renal tumours undergoing curative or cytoreductive nephrectomy at Addenbrooke's Hospital were recruited to the DIAMOND study. Ethics approval and patient consent were obtained (Research Ethics Committee REC ID 03/018). Fresh frozen tissue was stored at -80°C. Where available, multi-region tumour samples were taken along with adjacent normal kidney tissue. DNA was extracted from a small section of frozen tissue, using the commercially available AllPrep DNA/RNA Mini Kit (QIAGEN) according to the manufacturer's protocol.

DNA samples (10ng/µl, 500ng total) were sheared using the S220 Focused-ultrasonicator (Covaris) to generate dsDNA fragments. The D1000 ScreenTape System (Agilent) was used to ensure >60% of DNA fragments were between 100 and 300bp long, with a mean fragment size of 180-200bp. Tissue methylation analysis was performed using the TruSeq Methyl Capture EPIC Library Preparation Kit (Illumina), using the manufacturer's protocol (a.k.a. Epic-seq). This consists of a capture-based method targeting approximately 3 million CpGs. Four samples are multiplexed in each capture reaction using sample indexing adaptors. The protocol involves hybridization of biotin-tagged probes to gDNA followed by capture using streptavidin beads (two hybridization-capture steps) followed by bisulfite conversion at 54°C for two hours. Following Polymerase Chain Reaction (PCR) amplification, uracils are copied as thymines, with resulting libraries consisting of two families of dsDNA molecules (originating from Watson and Crick strands), with a high thymine to cytosine ratio. Twelve samples were pooled for sequencing on the HiSeq4000 Illumina Sequencing platform (single end 150bp read) using two lanes per library pool. Technical replicates were performed for cell line data to assess assay reproducibility (correlation = 0.97). Sequenced data were trimmed (TrimGalore v0.4.4) and

aligned to the bisulfite converted human reference genome (GRCh38/hg38). Methylation calling was performed using the Bismark suite of tools (v0.22.1). Trimming and alignment reports were compiled using MultiQC (v1.7). Data were included in downstream analysis if a depth of greater than 10x coverage was achieved, to reduce the risk of false positive calling. In addition, samples were removed if non CpG methylation was greater than 1%.

Publicly available data

The Cancer Genome Atlas (TCGA) data for ccRCC (KIRC), pRCC (KIRP) and chRCC (KICH), were obtained via the TCGAbiolinks package in R (*69*). Only CpG probes with less than 5% missing data were kept. The Chopra datasets, one for training and one for testing, were downloaded from the Open Science Framework, repository OSF.IO/Y8BH2 (*18*) and the AML samples were excluded. The Wei dataset (*19*) and Evelönn dataset (*21*) were obtained from the Gene Expression Omnibus, with GEO accession numbers GSE61441 and GSE113501, respectively. The Brennan dataset was obtained directly from the study authors (*20*). In all cases, methylation data were evaluated using the Illumina Infinium Human DNA Methylation 450 platform (a.k.a. 450k array).

Pre-processing

Training/testing dataset

The training/testing dataset comprised of the TCGA data, the Chopra training data, and the Cambridge samples. The following pre-processing steps were performed on this training/testing set. CpG probes found in two blacklists (*70, 71*) for the 450k array were removed, as well as probes at the site of C/T and G/A SNPs, probes that map to multiple regions, and probes at repeat regions. In addition, CpG probes located on the sex chromosomes were omitted to remove gender bias. The 450k array (Chopra training samples and TCGA samples) and EPIC-

seq (Cambridge data) cover different CpG sites. In order to combine data from the two methods, Epic-seq beta values within 50bp of the 450K probes were averaged, as adjacent CpGs tend to be co-methylated (*72*). Subsequently, probes that were missing for the whole of one dataset (TCGA, Chopra training data or Cambridge) were removed, to avoid dataset specific bias. This resulted in a dataset with 158,670 CpG probes.

Lastly, the beta values were converted to M-values, as M-values have been shown to be more homoscedastic (*73*). They are computed using the following equation:

$$M = \log_2 \left(\frac{\beta}{1 - \beta} \right)$$

Where M-values were calculated to be infinity (due to β = 1), they were set to the maximum finite value of the training/testing data, and where M-values were calculated to be minus infinity (due to β = 0), they were set to the minimum finite value of the training/test data.

External datasets

The external datasets used were the Chopra validation data (which is independent of the Chopra training data), the Brennan dataset, the Wei dataset and Evelönn dataset. The same 158,670 probes in the pre-processed training/testing dataset were selected for all four external datasets.

Data visualisation

Dimensionality reduction was carried out using umap-learn Python package (version 0.3.10) (*74*). NaN values were converted to the value 0 before the transformation.

MethylBoostER: The XGBoost classification model

A multiclass extreme gradient boosting classifier model was constructed using the xgboost Python package (version 0.90) (*75*). Four-fold nested cross validation was implemented on the training/testing set, with integrated hyperparameter optimisation.

Four-fold nested cross validation

The training/testing set was split up into four stratified outer folds, so that each fold keeps 25% of the dataset for testing. This was split so that multiple samples from the same patient were all in the same fold, to avoid data leakage between folds. These folds were iterated through, treating each one as the testing set in each iteration. Within these iterations, the remaining dataset was treated as a training/validation set.

Inner rounds of four-fold cross validation were performed on this training/validation set in order to find the optimal hyperparameters. For each inner fold iteration, a model was trained on the training set (75% of the training/validation set) and validated on the validation set (25% of the training/validation set). These inner rounds were repeated for various sets of hyperparameters (the hyperparameters tested were dictated by the Tree of Parzen Estimators algorithm in the hyperopt package (*76*)).

The hyperparameters that resulted in the best average MCC score on the validation sets (averaged across all four inner folds) were selected, and a model was trained with these parameters on the whole training/validation set. This resulted in a trained model with optimal hyperparameters for each outer fold, and the performance of these models on their fold's testing set was reported. This method of cross validation ensured that the models had the optimal hyperparameters, they were evaluated on completely unseen test data (specifically data that was not used to select the hyperparameters, to avoid inflated scores), and the performance on the whole training/testing dataset was reported.

Hyperparameters and training details

Two parameters were manually set, namely subsample = 0.5 and colsample bytree = 0.5 in order to help reduce overfitting and run time (*75*). Many of the remaining parameters were found using hyperparameter optimisation implemented in the hyperopt Python package (version

0.2.5) (*76*). The maximum number of evaluations was set to 10 and the parallelism was set to 2, to reduce the run time. The search spaces for each parameter were:

- number of trees: 50-500
- learning rate: 0.05-0.5 (sampled from a log uniform distribution)
- max tree depth: 2 or 3 (intentionally set very low to help reduce overfitting)
- L1 regularisation term: 0-1
- L2 regularisation term: 0-1

During model training, Gaussian noise with mean 0 and variance 0.2 was added into the training data to help reduce overfitting. Early stopping was also applied to reduce overfitting, with the number of early stopping rounds set to 5 and multiclass logloss as the evaluation metric used. Samples were weighted to avoid both patient bias and to mitigate the class bias. Balanced class weights were calculated using the 'compute class weight function' in the sklearn package (*77*), which ensures samples in less common classes are weighted higher. The same function was used to generate balanced patient weights, so that samples from patients with many samples were weighted lower. The two weights were multiplied together, in order to get a weight for each sample to be used during training.

Evaluation metrics

Accuracy, precision, recall, F_1 and the Matthews Correlation Coefficient (MCC) were used to evaluate the models' performance. F_1 is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The MCC was also reported as it is a performance measure that is not affected by large class imbalances, unlike accuracy. Receiver Operating Characteristic (ROC) curves and Precision

Recall (PR) curves were also plotted, with each class plotted separately (i.e., using the one-vs-all strategy). ROC curves show the false positive rate and the true positive rate over all possible values of the classification threshold. PR curves show the precision and recall over all possible values of the classification threshold. The Area Under the Curve (AUC) was also reported for both of these curves, with 1 as the best possible AUC score.

High- and moderate-confidence predictions

The model's output probability was used as a confidence measure, and the predictions were split into two categories: high-confidence, where the output probability was larger than a specified threshold, t, and moderate-confidence, where the output probability was less than t. For moderate-confidence predictions, we outputted the model's top two predictions (i.e., the two classes with the two highest output probabilities).

The parameter t was chosen based on the testing set. Three metrics were plotted for values of t between 0.5 and 1 (in increments of 0.025). These metrics were: the fraction of high-confidence predictions, the accuracy of the high-confidence predictions, and the accuracy of the moderate-confidence predictions where a prediction was treated as correct if the correct class was in the model's top two predictions. In order to remove noise in these accuracy scores, simple linear models were fitted (i.e., approximated accuracy = a*t + b) and the approximated accuracy scores were used to estimate t. For each value of t, these three metrics were averaged and the value of t where the average was the largest was taken. This was t = 0.85 and was validated on the external validation datasets.

Tumour purity analysis

Tumour purity was obtained for each sample using methylation beta values using the InfiniumPurify package in R (*28*) as previously described (*20*). Purity in correctly predicted and

incorrectly predicted samples was compared using the Wilcoxon signed rank sum test with correction for multiple testing.

Feature analysis

MethylBoostER feature importance and annotation

Feature importance values were obtained from the trained XGBoost models, which used the gain measure as the feature importance. All features with an importance greater than zero were selected. For each selected feature, the CpG was mapped to a gene if it was within a gene or within 1500bp upstream of a gene.

Genomic location of features

The R package ChIPseeker (*78*) (version 1.30.2) was used for genomic location annotation and visualisation. The R package EnsDb.Hsapiens.v86 (*79*) (version 2.99.0) was used for the annotation source. The background set used was the total set of input features.

Over-Representation analysis (ORA)

Gene Ontology enrichment analysis was carried out using the Biological Networks Gene Ontology (BINGO v3.0.3) package (*80*) and network visualization was carried out using Cytoscape (v3.9.0) (*81*). Pathway enrichment analysis was carried out using the Panther analysis tool (release 16) (*82*) with Panther and Reactome pathway collections. Statistical over-representation was carried out using Fisher's Exact test with false discovery rate (FDR) correction. Over-representation analysis was also performed on eight gene lists: dbEMT2 (*34*), COSMIC Cancer Gene Census (*35*), a list of putative driver genes for ccRCC (*32*), pRCC and RCC genes from the Diseases database (*31*), TF Checkpoint (*29*) for transcription factors, EpiFactors for epigentics regulators (*30*) and Human Cancer Metastatic Database (HCMDB) for metastatic genes (*33*). Enrichment of gene sets was determined using Fisher's Exact test with FDR correction, using

the Python package SciPy (*83*). The background set used in ORA were the total set of input features mapped to genes, and Ensembl gene IDs were used where specified (if not, gene symbols were used).

For the localised region GO enrichment analysis, the R package rGREAT (*84*) (version 1.26.0) was used, and the background set was the total set of input features. The significant (adjusted p-value $\leq$0.05) Biological Process terms that best summarised the results (with redundant terms manually excluded) were visualised.

Literature mining entity relationships

The Chilibot text mining web application (*85*) was manually searched with the 45 genes common to all four models together with the terms "kidney", "renal", "RCC" and "cancer". 30 million PubMed abstracts were also searched for entity-relationships using the Pangaea natural language processing Python package (*86*) with a renal carcinoma associated relation term list and a collection of HGNC gene symbols (4217 genes) derived from MethylBoostER's features mapped to genes (5012 genes). Entity-Relationship terms with frequencies $\geq$ 3 were used to build a gene relationship network using Cytoscape (v3.9.0).

Supplementary Materials

Fig. S1. ROC and Precision-Recall curves for all four models on the test set.

Fig. S2. UMAP of samples in the training/test set, for all four models.

Fig. S3. Testing set confusion matrix split by high- and moderate-confidence predictions.

Fig. S4. Performance of all four XGBoost models on all four external datasets.

Fig. S5. Performance of all four XGBoost models on all four external datasets

Fig. S6. UMAP representation of the Evel¨onn external dataset.

Fig. S7. Sample purity for samples which are predicted correctly on the first prediction.

Fig. S8. Methylation distribution of the 38 features that were selected by all four XGBoost models.

Fig. S9. Evaluation of Ricketts et al. sample reclassification.

Fig. S10. Topology analysis of methylation features in NLP derived Entity-Relationships network.

Table S1. Average performance metrics on the testing set.

Table S2. Performance metrics for all external validation datasets, averaged across all four models.

Table S3. Accuracy and MCC for all external validation datasets, split by high- and moderate confidence predictions.

Table S4. Accuracy of both first and second predictions and sample numbers for different levels of purity.

Table S5. Over Representation Analysis of gene sets.

Table S6. The literature evidence and associations for the 45 genes common to all models.

Data File S1. NLP derived gene-relationship pairs.

References and Notes

1. U. Capitanio, K. Bensalah, A. Bex, S. A. Boorjian, F. Bray, J. Coleman, J. L. Gore, M. Sun, C. Wood, P. Russo, Epidemiology of renal cell carcinoma. *Eur Urol* **75**, 74-84 (2019).

2. H. G. Welch, J. S. Skinner, F. R. Schroeck, W. Zhou, W. C. Black, Regional variation of computed tomographic imaging in the united states and the risk of nephrectomy. *JAMA Intern Med* **178**, 221-227 (2018).

3. B. Shuch, A. Amin, A. J. Armstrong, J. N. Eble, V. Ficarra, A. Lopez-Beltran, G. Martignoni, B. I. Rini, A. Kutikov, Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. *European urology* **67**, 85–97 (2015).

4. H. Moch, P. Humphrey, T. Ulbright, V. Reuter, *WHO classification of tumours of the urinary system and male genital organs (International Agency for Research on Cancer/World Health Organization Classification of Tumours)*, vol. 8 (WHO/IARC Press, Lyon, France, 2016), fourth edn.

5. H. D. Patel, S. C. Druskin, S. P. Rowe, P. M. Pierorazio, M. A. Gorin, M. E. Allaf, Surgical histopathology for suspected oncocytoma on renal mass biopsy: a systematic review and meta-analysis. *BJU Int* **119**, 661-666 (2017).

6. L. Marconi, S. Dabestani, T. B. Lam, F. Hofmann, F. Stewart, J. Norrie, A. Bex, K. Bensalah, S. E. Canfield, M. Hora, M. A. Kuczyk, A. S. Merseburger, P. F. A. Mulders, T. Powles, M. Staehler, B. Ljungberg, A. Volpe, Systematic review and meta-analysis of diagnostic accuracy of percutaneous renal tumour biopsy. *Eur Urol* **69**, 660-673 (2016).

7. J. J. Tomaszewski, R. G. Uzzo, M. C. Smaldone, Heterogeneity and renal mass biopsy: a review of its role and reliability. *Cancer Biol Med* **11**, 162-72 (2014).

8. D. C. Johnson, J. Vukina, A. B. Smith, A. M. Meyer, S. B. Wheeler, T. M. Kuo, H. J. Tan, M. E. Woods, M. C. Raynor, E. M. Wallen, R. S. Pruthi, M. E. Nielsen, Preoperatively misclassified, surgically removed benign renal masses: a systematic review of surgical series and united states population level burden estimate. *J Urol* **193**, 30-5 (2015).

9. E. M. Sohlberg, T. J. Metzner, J. T. Leppert, The harms of overdiagnosis and overtreatment in patients with small renal masses: A mini-review. *Eur Urol Focus* **5**, 943-945 (2019).

10. S. H. Rossi, C. Blick, C. Handforth, J. E. Brown, G. D. Stewart, C. Renal Cancer Gap Analysis, Essential research priorities in renal cancer: A modified delphi consensus statement. *Eur Urol Focus* (2019).

11. C. J. Ricketts, A. A. De Cubas, H. Fan, C. C. Smith, M. Lang, E. Reznik, R. Bowlby, E. A. Gibb, R. Akbani, R. Beroukhim, D. P. Bottaro, T. K. Choueiri, R. A. Gibbs, A. K. Godwin, S. Haake, A. A. Hakimi, E. P. Henske, J. J. Hsieh, T. H. Ho, R. S. Kanchi, B. Krishnan, D. J. Kwiatkowski, W. Lui, M. J. Merino, G. B. Mills, J. Myers, M. L. Nickerson, V. E. Reuter, L. S. Schmidt, C. S. Shelley, H. Shen, B. Shuch, S. Signoretti, R. Srinivasan, P. Tamboli, G. Thomas, B. G. Vincent, C. D. Vocke, D. A. Wheeler, L. Yang, W. Y. Kim, A. G. Robertson, N. Cancer Genome Atlas Research, P. T. Spellman, W. K. Rathmell, W. M. Linehan, The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep* **23**, 3698 (2018).

12. M. D. Young, T. J. Mitchell, F. A. Vieira Braga, M. G. B. Tran, B. J. Stewart, J. R. Ferdinand, G. Collord, R. A. Botting, D. M. Popescu, K. W. Loudon, R. Vento-Tormo, E. Stephenson, A. Cagan, S. J. Farndon, M. Del Castillo Velasco-Herrera, C. Guzzo, N. Richoz, L. Mamanova, T. Aho, J. N. Armitage, A. C. P. Riddick, I. Mushtaq, S. Farrell, D. Rampling, J. Nicholson, A. Filby, J. Burge, S. Lisgo, P. H. Maxwell, S. Lindsay, A. Y. Warren, G. D. Stewart, N. Sebire, N. Coleman, M. Haniffa, S. A. Teichmann, M. Clatworthy, S. Behjati, Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594-599 (2018).

13. G. G. Malouf, X. Su, J. Zhang, C. J. Creighton, T. H. Ho, Y. Lu, N. J. Raynal, J. A. Karam, P. Tamboli, F. Allanick, R. Mouawad, J. P. Spano, D. Khayat, C. G. Wood, J. Jelinek, N. M. Tannir, Dna methylation signature reveals cell ontogeny of renal cell carcinomas. *Clin Cancer Res* **22**, 6236-6246 (2016).

14. B. N. Lasseigne, J. D. Brooks, The role of dna methylation in renal cell carcinoma. *Mol Diagn Ther* **22**, 431-442 (2018).

15. K. Lommen, N. Vaes, M. J. Aarts, J. G. van Roermund, L. J. Schouten, E. Oosterwijk, V. Melotte, V. C. Tjan-Heijnen, M. van Engeland, K. M. Smits, Diagnostic dna methylation biomarkers for renal cell carcinoma: A systematic review. *Eur Urol Oncol* **4**, 215-226 (2021).

16. W. Chen, J. Zhuang, P. P. Wang, J. Jiang, C. Lin, P. Zeng, Y. Liang, X. Zhang, Y. Dai, H. Diao, Dna methylation-based classification and identification of renal cell carcinoma prognosis-subgroups. *Cancer Cell Int* **19**, 185 (2019).

17. J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K. Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, Y. Dor, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease. *Nat Commun* **9**, 5068 (2018).

18. S. Chopra, J. Liu, M. Alemozaffar, P. W. Nichols, M. Aron, D. J. Weisenberger, C. K. Collings, S. Syan, B. Hu, M. Desai, M. Aron, V. Duddalwar, I. Gill, G. Liang, K. D. Siegmund, Improving needle biopsy accuracy in small renal mass using tumor-specific DNA methylation markers. *Oncotarget* **8**, 5439–5448 (2017).

19. J. H. Wei, A. Haddad, K. J. Wu, H. W. Zhao, P. Kapur, Z. L. Zhang, L. Y. Zhao, Z. H. Chen, Y. Y. Zhou, J. C. Zhou, B. Wang, Y. H. Yu, M. Y. Cai, D. Xie, B. Liao, C. X. Li, P. X. Li, Z. R. Wang, F. J. Zhou, L. Shi, Q. Z. Liu, Z. L. Gao, D. L. He, W. Chen, J. T. Hsieh, Q. Z. Li, V. Margulis, J. H. Luo, A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. *Nature Communications* **6**, 1–11 (2015).

20. K. Brennan, T. J. Metzner, C.-S. Kao, C. E. Massie, G. D. Stewart, R. W. Haile, J. D. Brooks, M. P. Hitchins, J. T. Leppert, O. Gevaert, Development of a DNA Methylation–Based Diagnostic Signature to Distinguish Benign Oncocytoma From Renal Cell Carcinoma. *JCO Precision Oncology* pp. 1141–1151 (2020).

21. E. A. Evelönn, M. Landfors, Z. Haider, L. Köhn, B. Ljungberg, G. Roos, S. Degerman, DNA methylation associates with survival in non-metastatic clear cell renal cell carcinoma. *BMC Cancer* **19**, 1–16 (2019).

22. M. Gerlinger, A. J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, C. Swanton, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892 (2012).

23. S. Turajlic, *et al.*, Deterministic evolutionary trajectories influence primary tumor growth: Tracerx renal. *Cell* **173**, 595-610 e11 (2018).

24. E. A. Evelonn, M. Landfors, Z. Haider, L. Kohn, B. Ljungberg, G. Roos, S. Degerman, Dna methylation associates with survival in non-metastatic clear cell renal cell carcinoma. *BMC Cancer* **19**, 65 (2019).

25. K. Sharpe, G. D. Stewart, A. Mackay, C. Van Neste, C. Rofe, D. Berney, I. Kayani, A. Bex, E. Wan, F. C. O'Mahony, M. O'Donnell, S. Chowdhury, R. Doshi, C. Ho-Yen, M. Gerlinger, D. Baker, N. Smith, B. Davies, A. Sahdev, E. Boleti, T. De Meyer, W. Van Criekinge, L. Beltran, Y. J. Lu, D. J. Harrison, A. R. Reynolds, T. Powles, The effect of vegf-targeted

therapy on biomarker expression in sequential tissue from patients with metastatic clear cell renal cancer. *Clin Cancer Res* **19**, 6924-34 (2013).

26. S. Winter, P. Fisel, F. Buttner, S. Rausch, D. D'Amico, J. Hennenlotter, S. Kruck, A. T. Nies, A. Stenzl, K. Junker, M. Scharpf, U. Hofmann, H. van der Kuip, F. Fend, G. Ott, A. Agaimy, A. Hartmann, J. Bedke, M. Schwab, E. Schaeffeler, Methylomes of renal cell lines and tumors or metastases differ significantly with impact on pharmacogenes. *Sci Rep* **6**, 29930 (2016).

27. G. D. Stewart, T. Powles, C. Van Neste, A. Meynert, F. O'Mahony, A. Laird, D. Deforce, F. Van Nieuwerburgh, G. Trooskens, W. Van Criekinge, T. De Meyer, D. J. Harrison, Dynamic epigenetic changes to vhl occur with sunitinib in metastatic clear cell renal cancer. *Oncotarget* **7**, 25241-50 (2016).

28. Y. Qin, H. Feng, M. Chen, H. Wu, X. Zheng, Infiniumpurify: An r package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis* **5**, 43-45 (2018).

29. S. Tripathi, K. R. Christie, R. Balakrishnan, R. Huntley, D. P. Hill, L. Thommesen, J. A. Blake, M. Kuiper, A. Lægreid, Gene ontology annotation of sequence-specific dna binding transcription factors: setting the stage for a large-scale curation effort. *Database* **2013** (2013).

30. Y. A. Medvedeva, A. Lennartsson, R. Ehsani, I. V. Kulakovskiy, I. E. Vorontsov, P. Panahandeh, G. Khimulya, T. Kasukawa, F. Drabløs, F. Consortium, *et al.*, Epifactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015** (2015).

31. S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, L. J. Jensen, Diseases: Text mining and data integration of disease–gene associations. *Methods* **74**, 83–89 (2015).

32. S. Turajlic, H. Xu, K. Litchfield, A. Rowan, T. Chambers, J. I. Lopez, D. Nicol, T. O'Brien, J. Larkin, S. Horswell, M. Stares, L. Au, M. Jamal-Hanjani, B. Challacombe, A. Chandra, S. Hazell, C. Eichler-Jonsson, A. Soultati, S. Chowdhury, S. Rudman, J. Lynch, A. Fernando, G. Stamp, E. Nye, F. Jabbar, L. Spain, S. Lall, R. Guarch, M. Falzon, I. Proctor, L. Pickering, M. Gore, T. B. Watkins, S. Ward, A. Stewart, R. DiNatale, M. F. Becerra, E. Reznik, J. J. Hsieh, T. A. Richmond, G. F. Mayhew, S. M. Hill, C. D. McNally, C. Jones, H. Rosenbaum, S. Stanislaw, D. L. Burgess, N. R. Alexander, C. Swanton, Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173**, 581–594.e12 (2018).

33. G. Zheng, Y. Ma, Y. Zou, A. Yin, W. Li, D. Dong, Hcmdb: the human cancer metastasis database. *Nucleic acids research* **46**, D950–D955 (2018).

34. M. Zhao, Y. Liu, C. Zheng, H. Qu, dbEMT 2.0: An updated database for epithelial-mesenchymal transition genes with experimentally verified information and precalculated regulation information for cancer metastasis. *Journal of Genetics and Genomics* **46**, 595–597 (2019).

35. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941-D947 (2018).

36. C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495–501 (2010).

37. V. Aggelis, R. A. Craven, J. Peng, P. Harnden, D. A. Cairns, E. R. Maher, R. Tonge, P. J. Selby, R. E. Banks, Proteomic identification of differentially expressed plasma membrane proteins in renal cell carcinoma by stable isotope labelling of a von hippel-lindau transfectant cell line model. *Proteomics* **9**, 2118–2130 (2009).

38. L. Wang, Z. Peng, K. Wang, Y. Qi, Y. Yang, Y. Zhang, X. An, S. Luo, J. Zheng, NDUFA4L2 is associated with clear cell renal cell carcinoma malignancy and is regulated by ELK1. *PeerJ* **2017**, 1–15 (2017).

39. X. Bao, J. Duan, Y. Yan, X. Ma, Y. Zhang, H. Wang, D. Ni, S. Wu, C. Peng, Y. Fan, *et al.*, Upregulation of long noncoding rna pvt1 predicts unfavorable prognosis in patients with clear cell renal cell carcinoma. *Cancer Biomarkers* **21**, 55–63 (2018).

40. Q. T. Pham, N. Oue, Y. Yamamoto, Y. Shigematsu, Y. Sekino, N. Sakamoto, K. Sentani, N. Uraoka, M. Tiwari, W. Yasui, The expression of bts-2 enhances cell growth and invasiveness in renal cell carcinoma. *Anticancer research* **37**, 2853–2860 (2017).

41. P. Fisel, S. Kruck, S. Winter, J. Bedke, J. Hennenlotter, A. T. Nies, M. Scharpf, F. Fend, A. Stenzl, M. Schwab, E. Schaeffeler, DNA methylation of the SLC16A3 promoter regulates expression of the human lactate transporter MCT4 in renal Cancer with consequences for clinical outcome. *Clinical Cancer Research* **19**, 5170–5181 (2013).

42. P. Fisel, V. Stühler, J. Bedke, S. Winter, S. Rausch, J. Hennenlotter, A. T. Nies, A. Stenzl, M. Scharpf, F. Fend, S. Kruck, M. Schwab, E. Schaeffeler, MCT4 surpasses the prognostic relevance of the ancillary protein CD147 in clear cell renal cell carcinoma. *Oncotarget* **6**, 30615–30627 (2015).

43. W. Cizelsky, A. Tata, M. Kühl, S. J. Kühl, The wnt/jnk signaling target gene alcam is required for embryonic kidney development. *Development* **141**, 2064–2074 (2014).

44. F. Ferragut, V. S. Vachetta, M. F. Troncoso, G. A. Rabinovich, M. T. Elola, Alcam/cd166: A pleiotropic mediator of cell adhesion, stemness and cancer progression. *Cytokine & Growth Factor Reviews* (2021).

45. S. Zhao, M. S. Geybels, A. Leonardson, R. Rubicz, S. Kolb, Q. Yan, B. Klotzle, M. Bibikova, A. Hurtado-Coll, D. Troyer, *et al.*, Epigenome-wide tumor dna methylation profiling identifies novel prognostic biomarkers of metastatic-lethal progression in men diagnosed with clinically localized prostate cancer. *Clinical Cancer Research* **23**, 311–319 (2017).

46. W. Naushad, W. D. Mahauad-Fernandez, C. M. Okeoma, Structural determinant of bst-2-mediated regulation of breast cancer cell motility: a role for cytoplasmic tail tyrosine residues. *Oncotarget* **8**, 110221 (2017).

47. T. Yang, H. Zhou, P. Liu, L. Yan, W. Yao, K. Chen, J. Zeng, H. Li, J. Hu, H. Xu, *et al.*, lncrna pvt1 and its splicing variant function as competing endogenous rna to regulate clear cell renal cell carcinoma progression. *Oncotarget* **8**, 85353 (2017).

48. H. Wang, T. Chong, B. Y. Li, X. S. Chen, W. B. Zhen, Evaluating the clinical significance of SHMT2 and its co-expressed gene in human kidney cancer. *Biological Research* **53**, 1–10 (2020).

49. M. Lechpammer, M. B. Resnick, E. Sabo, E. Yakirevich, W. O. Greaves, K. T. Sciandra, R. Tavares, L. C. Noble, R. A. DeLellis, L. J. Wang, The diagnostic and prognostic utility of claudin expression in renal cell neoplasms. *Modern Pathology* **21**, 1320–1329 (2008).

50. B. Zou, D. Wang, K. Xu, J.-l. Liu, D.-y. Yuan, Z. Meng, B. Zhang, Prognostic value of long non-coding rna plasmacytoma variant translocation1 in human solid tumors: A meta-analysis. *Medicine* **98** (2019).

51. S. L. Skala, X. Wang, Y. Zhang, R. Mannan, L. Wang, S. P. Narayanan, P. Vats, F. Su, J. Chen, X. Cao, J. Siddiqui, P. Argani, M. P. Cieslik, T. J. Giordano, A. M. Chinnaiyan, S. M. Dhanasekaran, R. Mehra, Next-generation rna sequencing-based biomarker characterization of chromophobe renal cell carcinoma and related oncocytic neoplasms. *Eur Urol* **78**, 63-74 (2020).

52. P. D. McGillivray, D. Ueno, A. Pooli, N. Mendhiratta, J. S. Syed, K. A. Nguyen, P. G. Schulam, P. A. Humphrey, A. J. Adeniran, P. C. Boutros, B. Shuch, Distinguishing benign renal tumors with an oncocytic gene expression (onex) classifier. *Eur Urol* **79**, 107-111 (2021).

53. S. Durinck, E. W. Stawiski, A. Pavia-Jimenez, Z. Modrusan, P. Kapur, B. S. Jaiswal, N. Zhang, V. Toffessi-Tcheuyap, T. T. Nguyen, K. B. Pahuja, Y. J. Chen, S. Saleem, S. Chaudhuri, S. Heldens, M. Jackson, S. Pena-Llopis, J. Guillory, K. Toy, C. Ha, C. J. Harris, E. Holloman, H. M. Hill, J. Stinson, C. S. Rivers, V. Janakiraman, W. Wang, L. N. Kinch, N. V. Grishin, P. M. Haverty, B. Chow, J. S. Gehring, J. Reeder, G. Pau, T. D. Wu, V. Margulis, Y. Lotan, A. Sagalowsky, I. Pedrosa, F. J. de Sauvage, J. Brugarolas, S. Seshagiri, Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. *Nat Genet* **47**, 13-21 (2015).

54. A. S. Pires-Luis, P. Costa-Pinheiro, M. J. Ferreira, L. Antunes, F. Lobo, J. Oliveira, R. Henrique, C. Jeronimo, Identification of clear cell renal cell carcinoma and oncocytoma using a three-gene promoter methylation panel. *J Transl Med* **15**, 149 (2017).

55. I. Ibragimova, M. J. Slifker, M. E. Maradeo, G. Banumathy, E. Dulaimi, R. G. Uzzo, P. Cairns, Genome-wide promoter methylome of small renal masses. *PLoS One* **8**, e77309 (2013).

56. A. A. Slater, M. Alokail, D. Gentle, M. Yao, G. Kovacs, E. R. Maher, F. Latif, Dna methylation profiling distinguishes histological subtypes of renal cell carcinoma. *Epigenetics* **8**, 252-67 (2013).

57. C. Koelsche, *et al.*, Sarcoma classification by dna methylation profiling. *Nat Commun* **12**, 498 (2021).

58. D. Capper, *et al.*, Dna methylation-based classification of central nervous system tumours. *Nature* **555**, 469-474 (2018).

59. P. Jurmeister, M. Bockmayr, P. Seegerer, T. Bockmayr, D. Treue, G. Montavon, C. Vollbrecht, A. Arnold, D. Teichmann, K. Bressem, U. Schuller, M. von Laffert, K. R. Muller, D. Capper, F. Klauschen, Machine learning analysis of dna methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci Transl Med* **11** (2019).

60. S. Moran, A. Martinez-Cardus, S. Sayols, E. Musulen, C. Balana, A. Estival-Gonzalez, C. Moutinho, H. Heyn, A. Diaz-Lagares, M. C. de Moura, G. M. Stella, P. M. Comoglio, M. Ruiz-Miro, X. Matias-Guiu, R. Pazo-Cid, A. Anton, R. Lopez-Lopez, G. Soler, F. Longo, I. Guerra, S. Fernandez, Y. Assenov, C. Plass, R. Morales, J. Carles, D. Bowtell, L. Mileshkin, D. Sia, R. Tothill, J. Tabernero, J. M. Llovet, M. Esteller, Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* **17**, 1386-1395 (2016).

61. F. Chen, Y. Zhang, Y. Senbabaoglu, G. Ciriello, L. Yang, E. Reznik, B. Shuch, G. Micevic, G. De Velasco, E. Shinbrot, M. S. Noble, Y. Lu, K. R. Covington, L. Xi, J. A. Drummond, D. Muzny, H. Kang, J. Lee, P. Tamboli, V. Reuter, C. S. Shelley, B. A. Kaipparettu, D. P. Bottaro, A. K. Godwin, R. A. Gibbs, G. Getz, R. Kucherlapati, P. J. Park, C. Sander, E. P.

Henske, J. H. Zhou, D. J. Kwiatkowski, T. H. Ho, T. K. Choueiri, J. J. Hsieh, R. Akbani, G. B. Mills, A. A. Hakimi, D. A. Wheeler, C. J. Creighton, Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep* **14**, 2476-89 (2016).

62. T. K. Choueiri, U. Vaishampayan, J. E. Rosenberg, T. F. Logan, A. L. Harzstark, R. M. Bukowski, B. I. Rini, S. Srinivas, M. N. Stein, L. M. Adams, *et al.*, Phase ii and biomarker study of the dual met/vegfr2 inhibitor foretinib in patients with papillary renal cell carcinoma. *Journal of Clinical Oncology* **31**, 181 (2013).

63. U. Schütte, S. Bisht, L. C. Heukamp, M. Kebschull, A. Florin, J. Haarmann, P. Hoffmann, G. Bendas, R. Buettner, P. Brossart, *et al.*, Hippo signaling mediates proliferation, invasiveness, and metastatic potential of clear cell renal cell carcinoma. *Translational oncology* **7**, 309–321 (2014).

64. Ż. Piotrowska, M. Niezgoda, G. Młynarczyk, M. Acewicz, I. Kasacka, Comparative assessment of the wnt/β-catenin pathway, cacybp/sip, and the immunoproteasome subunit lmp7 in various histological types of renal cell carcinoma. *Frontiers in Oncology* **10**, 2530 (2020).

65. D. Huang, Y. Ding, W.-M. Luo, S. Bender, C.-N. Qian, E. Kort, Z.-F. Zhang, K. Vanden-Beldt, N. S. Duesbery, J. H. Resau, *et al.*, Inhibition of mapk kinase signaling pathways suppressed renal cell carcinoma growth and angiogenesis in vivo. *Cancer research* **68**, 81–88 (2008).

66. D. Miricescu, D. G. Balan, A. Tulin, O. Stiru, I. A. Vacaroiu, D. A. Mihai, C. C. Popa, R. I. Papacocea, M. Enyedi, N. A. Sorin, *et al.*, Pi3k/akt/mtor signalling pathway involvement in renal cell carcinoma pathogenesis. *Experimental and Therapeutic Medicine* **21**, 1–7 (2021).

67. L. C. Harshman, C. G. Drake, T. K. Choueiri, Pd-1 blockade in renal cell carcinoma: to equilibrium and beyond. *Cancer immunology research* **2**, 1132–1141 (2014).

68. T. Rasmuson, B. Ljungberg, K. Grankvist, J. Jacobsen, T. Olsson, Increased serum cortisol levels are associated with high tumour grade in patients with renal cell carcinoma. *Acta Oncologica* **40**, 83–87 (2001).

69. A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, H. Noushmehr, Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research* (2015).

70. H. Naeem, N. C. Wong, Z. Chatterton, M. K. H. Hong, J. S. Pedersen, N. M. Corcoran, C. M. Hovens, G. Macintyre, Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* **15**, 51 (2014).

71. M. E. Price, A. M. Cotton, L. L. Lam, P. Farré, E. Emberly, C. J. Brown, W. P. Robinson, M. S. Kobor, Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics and Chromatin* **6**, 1–15 (2013).

72. S. Guo, D. Diep, N. Plongthongkum, H. L. Fung, K. Zhang, K. Zhang, Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna. *Nat Genet* **49**, 635-642 (2017).

73. P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, S. M. Lin, Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* **11**, 587 (2010).

74. L. McInnes, J. Healy, N. Saul, L. Grossberger, Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* **3**, 861 (2018).

75. T. Chen, C. Guestrin, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

76. J. Bergstra, D. Yamins, D. Cox, Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *International Conference on Machine Learning* **28** (2013).

77. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

78. G. Yu, L.-G. Wang, Q.-Y. He, Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383 (2015).

79. J. Rainer, *EnsDb.Hsapiens.v86: Ensembl based annotation package* (2017). R package version 2.99.0.

80. S. Maere, K. Heymans, M. Kuiper, Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).

81. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).

82. H. Mi, D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou, T. Mushayamaha, P. D. Thomas, Panther version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive api. *Nucleic acids research* **49**, D394–D403 (2021).

83. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

84. Z. Gu, *rGREAT: Client for GREAT Analysis* (2021). Https://github.com/jokergoo/rGREAT, http://great.stanford.edu/public/html/.

85. H. Chen, B. M. Sharp, Content-rich biological network constructed by mining pubmed abstracts. *BMC bioinformatics* **5**, 1–13 (2004).

86. L. Pirvan, S. A. Samarajiwa, Pangaea: A modular and extensible collection of tools for mining context dependent gene relationships from the biomedical literature. *bioRxiv* (2020).

Acknowledgments:

We are grateful to Mr. Vincent Gnanapragasam, chief investigator of the DIAMOND study (REC 03/018), for access to renal tumour clinical samples. Infrastructure for the DIAMOND study was provided by the Cancer Research UK Cambridge Cancer Centre (Major Centre Award C9685/A25117) and NIHR Biomedical Research Centre. We are also grateful to Wing-Kit Leung for support with wet lab methods.

Funding:

S.H.R. is funded by a Cancer Research UK Clinical Doctoral Fellowship. I.N. is supported by a UK MRC doctoral training award. S.A.S. is supported by the UK Medical Research Council funding (MC UU 12022/10) and Isaac Newton Trust/Wellcome Trust ISSF/University of Cambridge grant (S.A.S. and J.H.). C.E.M. is funded by a CRUK Fellowship award (A26718). G.D.S. is supported by The Mark Foundation for Cancer Research, the Cancer Research UK Cambridge Centre [C9685/A25177] and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). A.Y.W is supported by The Mark Foundation for Cancer Research, the Cancer Research UK Cambridge Centre [C9685/A25177] and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The Human Research Tissue Bank at Addenbrooke's Hospital is supported by the NIHR Cambridge Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions:

The study was conceived and designed by S.H.R., I.N., C.E.M. and S.A.S., methodology development, data analysis and interpretation was carried out by I.N., S.H.R., C.E.M. and S.A.S. and J.H contributed to natural language processing and analysis. S.P., R.L. and K.B. also assisted with data analysis. S.A.S., G.D.S., J.T.L. and O.G. provided supervision and clinical input was provided by T.J.M., G.D.S. and J.T.L. A.B. and S.H.R. carried out wet lab experiments, and G.P. provided laboratory supervision. The manuscript was written by S.H.R., I.N., C.E.M. and S.A.S. Provision of study materials and patient data was carried out by S.H.R., C.E.M., G.D.S., J.T.L., C.S., T.J.M., O.G., K.B. Collection and Assembly of data was by S.H.R, and I.N. Pathology review and advice on data interpretation was provided by A.Y.W. All authors read and approved the manuscript.

Competing interests:

G.D.S. has received educational grants from Pfizer, AstraZeneca and Intuitive Surgical; consultancy fees from Pfizer, Merck, EUSA Pharma and CMR Surgical; Travel expenses from Pfizer and Speaker fees from Pfizer. All other authors have no competing interest.

Data and code availability

The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Public datasets are available from: The Cancer Genome Atlas (TCGA), obtained via the TCGAbiolinks package (*69*), Wei et al. (*19*) (GEO accession number: GSE61441), Evelönn et al. (*21*) (GEO accession number: GSE113501) and Chopra et al. (*18*) (Open Science Framework, repository OSF.IO/Y8BH2). The Brennan dataset was obtained directly from the study authors (*20*). The code for this project will be made publicly available on GitHub on peer-reviewed publication.

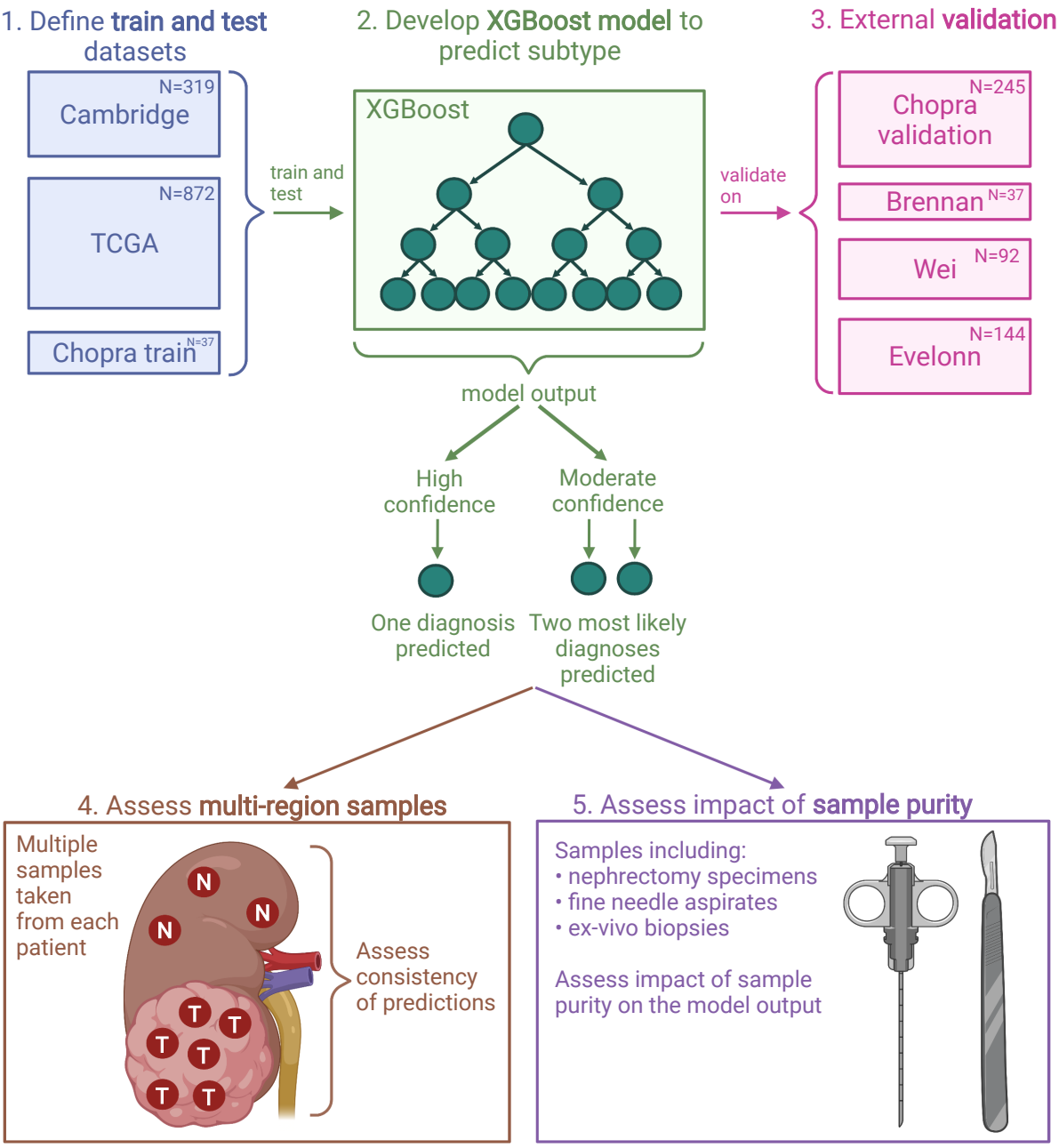


Figure 1: Graphical abstract

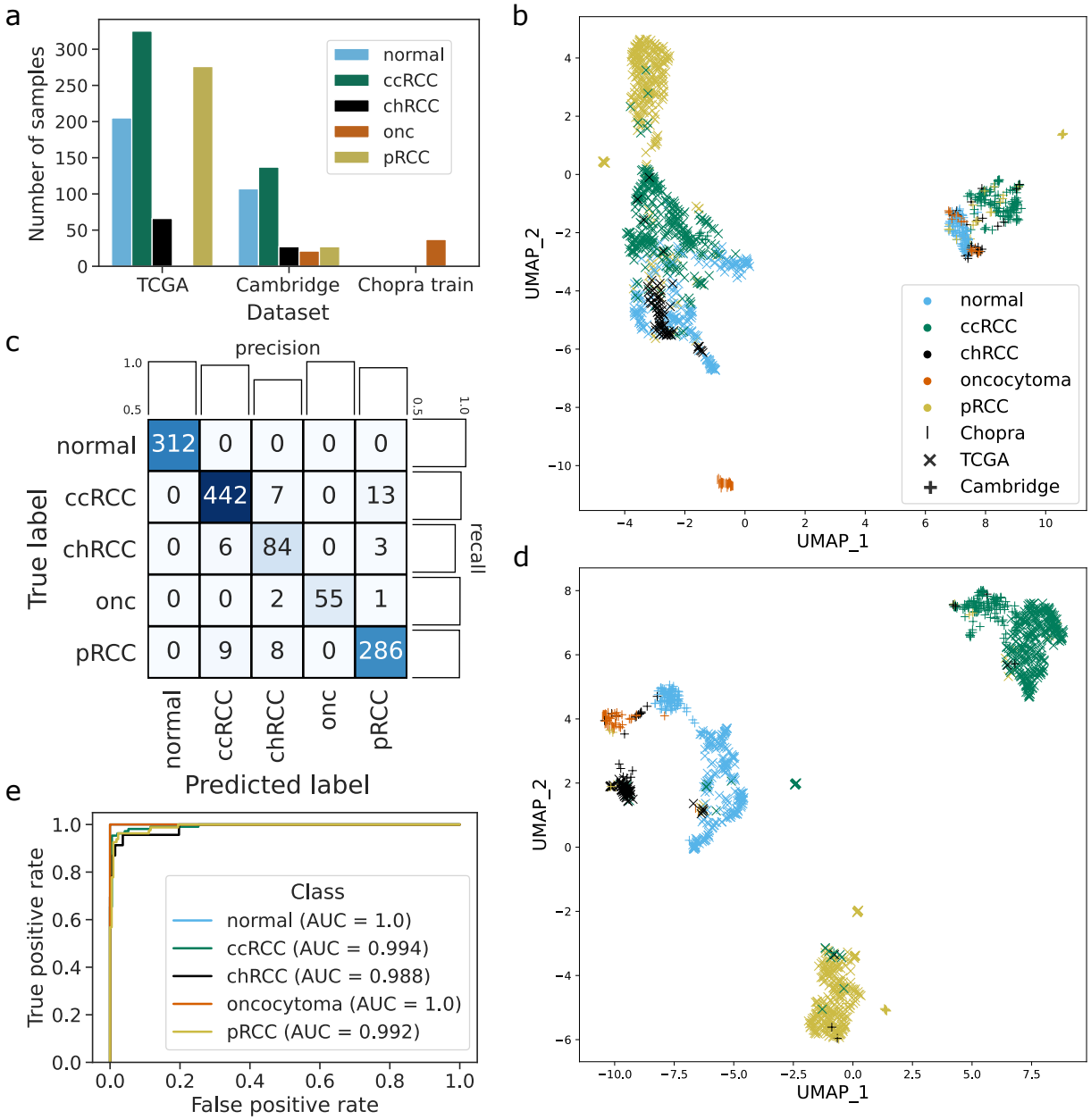


Figure 2: Data characteristics and testing set performance. **a** Number of samples in each class for the training/testing set. **b** UMAP representation of the training/test dataset, using all input features. **c** Confusion matrix displaying the testing set performance, with precision and recall bars. **d** UMAP representation of the training/test dataset, using the input features learnt by the XGBoost model. **e** ROC curves over the testing set, split by class.

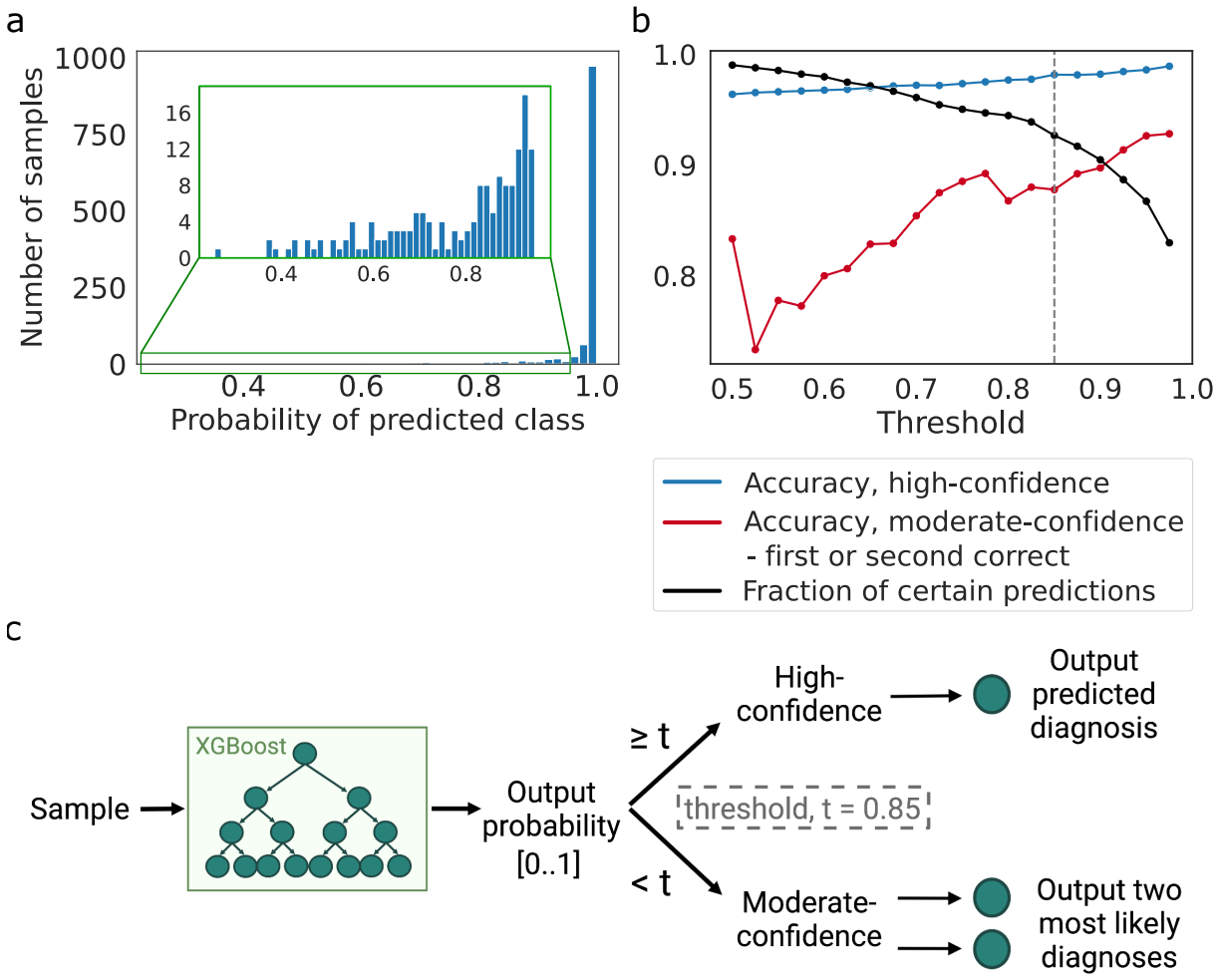


Figure 3: High- and moderate-confidence predictions. **a** Histogram of the model's probabilities of the predicted class for the testing sets. **b** Line plot showing how the testing set accuracy scores and fraction of high-confidence predictions vary as the threshold changes. The vertical dotted line indicates the chosen threshold, 0.85. **c** Graphical overview of the prediction process with high- and moderate-confidence predictions.

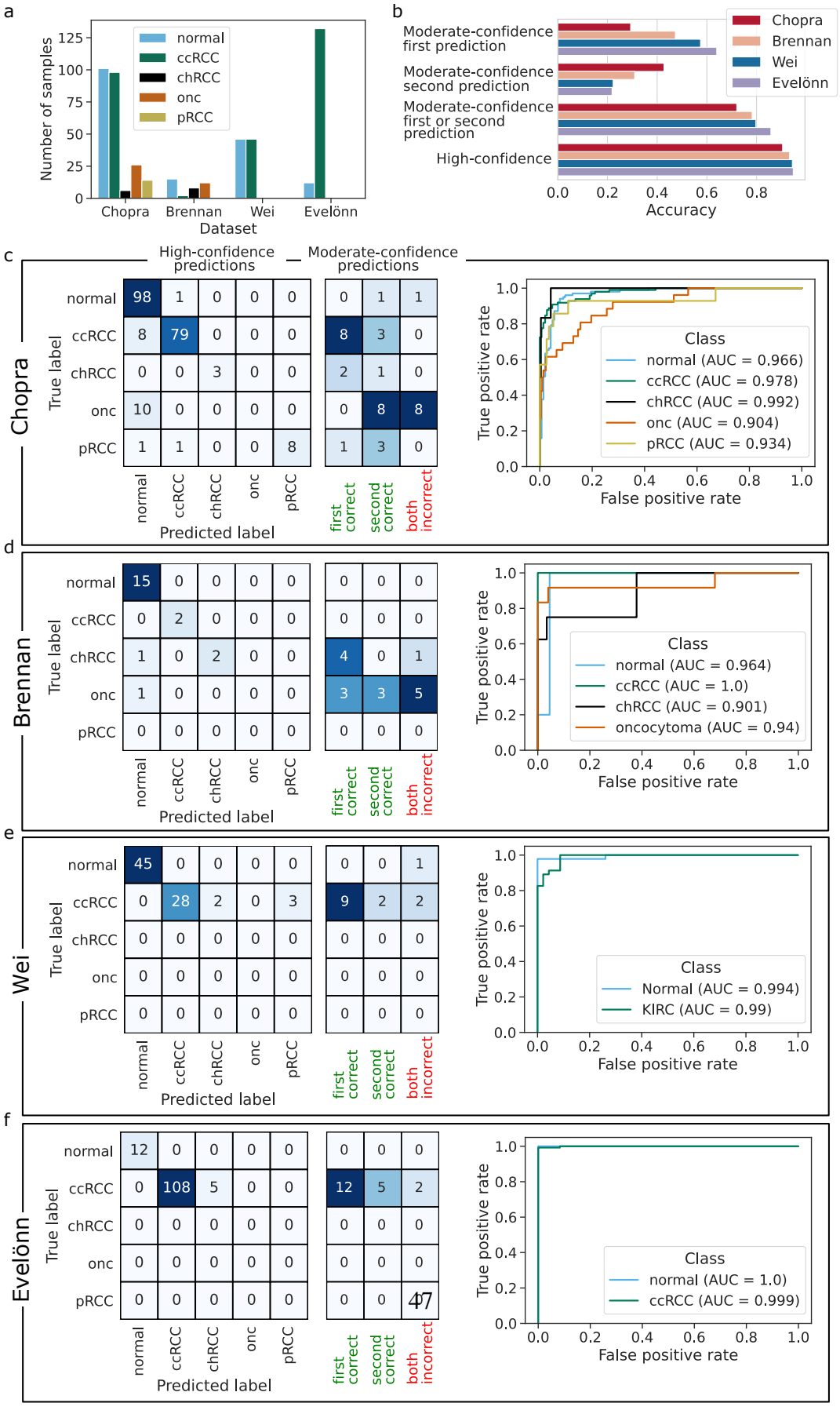

a
Number of samples
125
100
75
50
25
0
normal
ccRCC
chRCC
onc
pRCC
Chopra
Brennan
Wei
Evelönn
Dataset
b
Moderate-confidence first prediction
Moderate-confidence second prediction
Moderate-confidence first or second prediction
High-confidence
Chopra
Brennan
Wei
Evelönn
0.0
0.2
0.4
0.6
0.8
Accuracy
c
Chopra
High-confidence predictions
Moderate-confidence predictions
True label
normal 98 1 0 0 0
ccRCC 8 79 0 0 0
chRCC 0 0 3 0 0
onc 10 0 0 0 0
pRCC 1 1 0 0 8
0 1 1
8 3 0
2 1 0
0 8 8
1 3 0
normal
ccRCC
chRCC
onc
pRCC
first correct
second correct
both incorrect
Predicted label
True positive rate
False positive rate
Class
normal (AUC = 0.966)
ccRCC (AUC = 0.978)
chRCC (AUC = 0.992)
onc (AUC = 0.904)
pRCC (AUC = 0.934)
d
Brennan
True label
normal 15 0 0 0 0
ccRCC 0 2 0 0 0
chRCC 1 0 2 0 0
onc 1 0 0 0 0
pRCC 0 0 0 0 0
0 0 0
0 0 0
4 0 1
3 3 5
0 0 0
Predicted label
True positive rate
False positive rate
Class
normal (AUC = 0.964)
ccRCC (AUC = 1.0)
chRCC (AUC = 0.901)
oncocytoma (AUC = 0.94)
e
Wei
True label
normal 45 0 0 0 0
ccRCC 0 28 2 0 3
chRCC 0 0 0 0 0
onc 0 0 0 0 0
pRCC 0 0 0 0 0
0 0 1
9 2 2
0 0 0
0 0 0
0 0 0
Predicted label
True positive rate
False positive rate
Class
Normal (AUC = 0.994)
KIRC (AUC = 0.99)
f
Evelönn
True label
normal 12 0 0 0 0
ccRCC 0 108 5 0 0
chRCC 0 0 0 0 0
onc 0 0 0 0 0
pRCC 0 0 0 0 0
0 0 0
12 5 2
0 0 0
0 0 0
0 0 47
Predicted label
True positive rate
False positive rate
Class
normal (AUC = 1.0)
ccRCC (AUC = 0.999)

Figure 4: External validation on four independent datasets **a** Number of samples in each class for each dataset. **b** Accuracy for high- and moderate-confidence predictions for each external dataset. 'First or second prediction' indicates that a prediction is treated as correct if its first or second prediction was correct. **c, d, e** and **f** Confusion matrices for both high- and moderate-confidence predictions and ROC curves, split by class, for each external dataset. For the moderate-confidence confusion matrices, the x-axis is split into first correct (i.e., the first prediction was correct), second correct (i.e., the second prediction was correct), and both incorrect (i.e., both first and second predictions were incorrect).

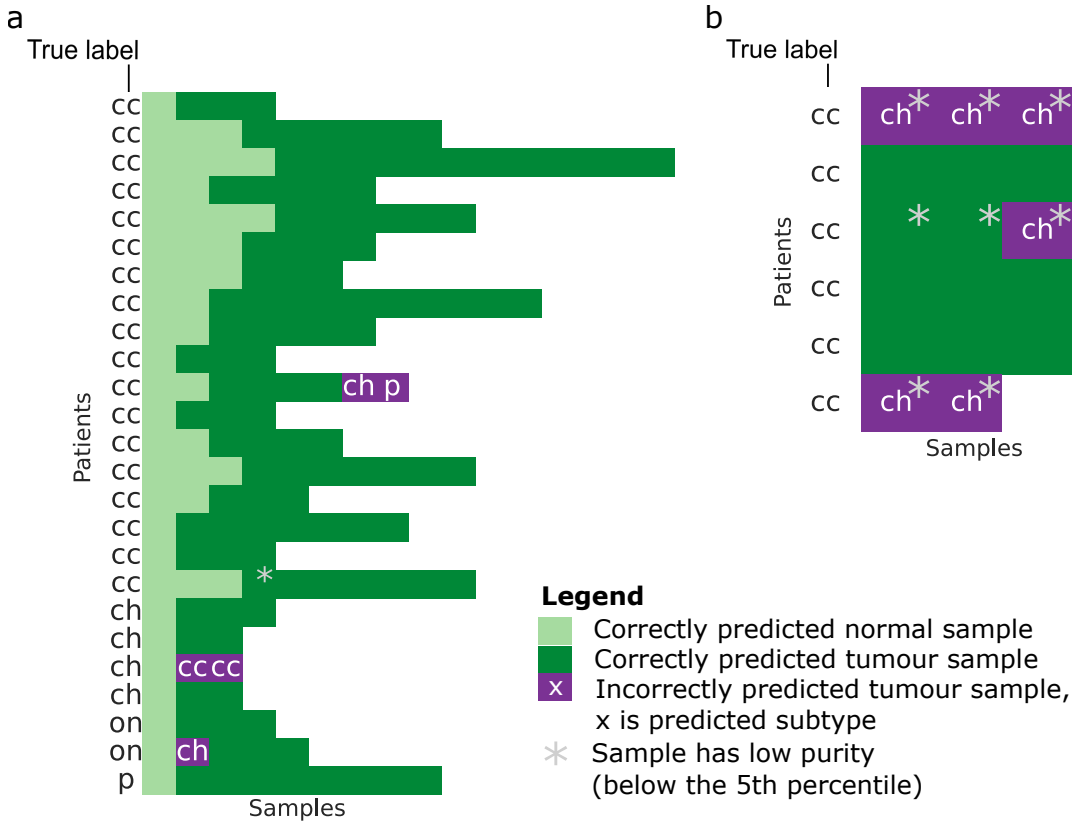


Figure 5: Classification of multi-region samples. **a** Diagram visualising the model's predictions of multi-region samples for each patient in the Cambridge dataset. **b** Diagram visualising the model's predictions of multi-region samples for each patient in the Evelönn dataset.

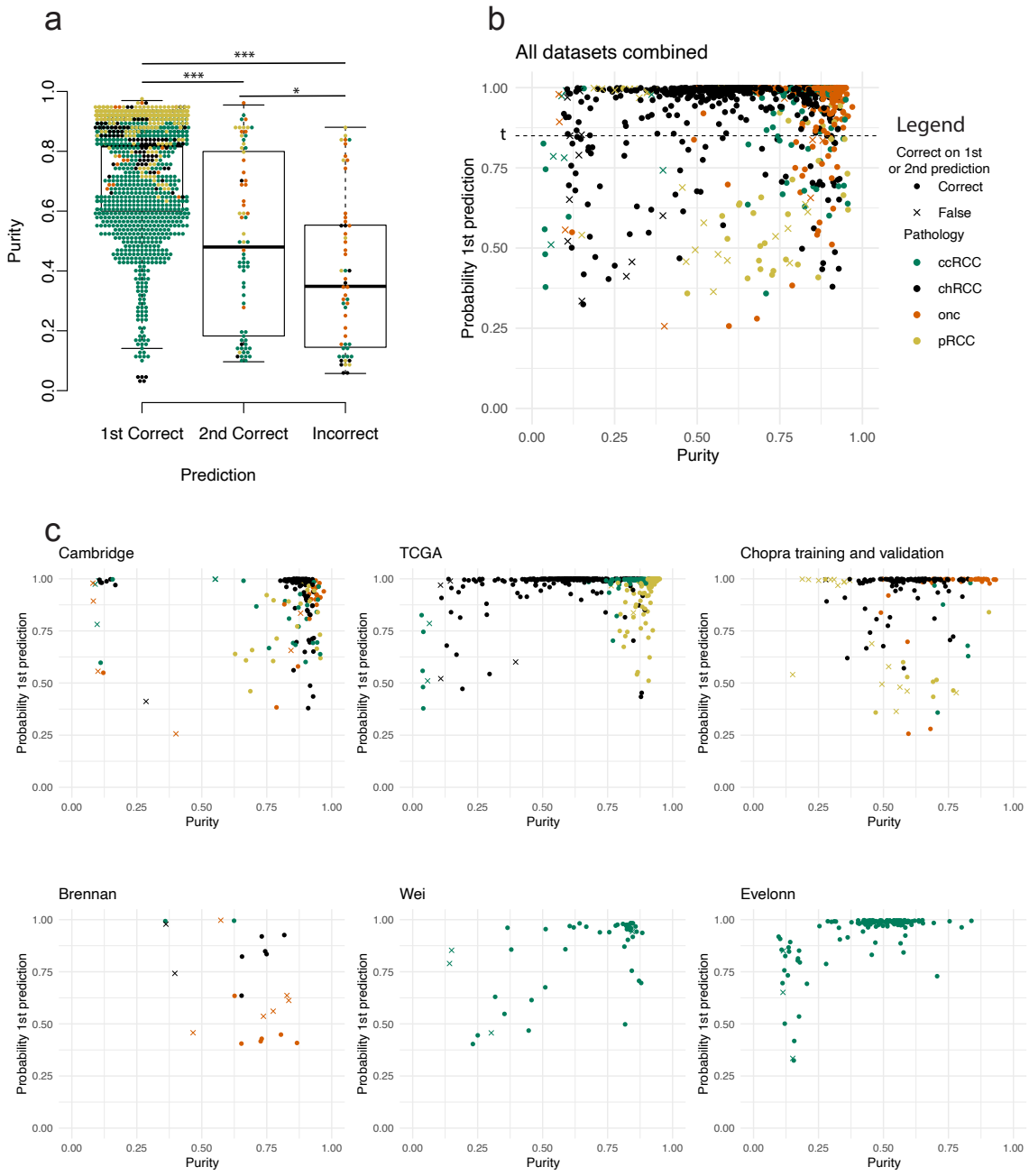


Figure 6: Sample purity and MethylBoostER output **a** Sample purity for samples which are predicted correctly on the first prediction (1st correct), second prediction (2nd correct) and incorrectly predicted samples (incorrect) on both predictions. Data is shown for all datasets combined, with pathological sub-types shown in different colors. Adjusted p values are shown (* <0.05, *** <0.0009) **b and c** Sample purity and the probability of the first prediction is demonstrated for all datasets combined (b) and each dataset individually (c). The threshold t = 0.85 indicating a high-confidence prediction is shown. Samples which are incorrectly predicted (in both first and second prediction) are indicated with a cross.

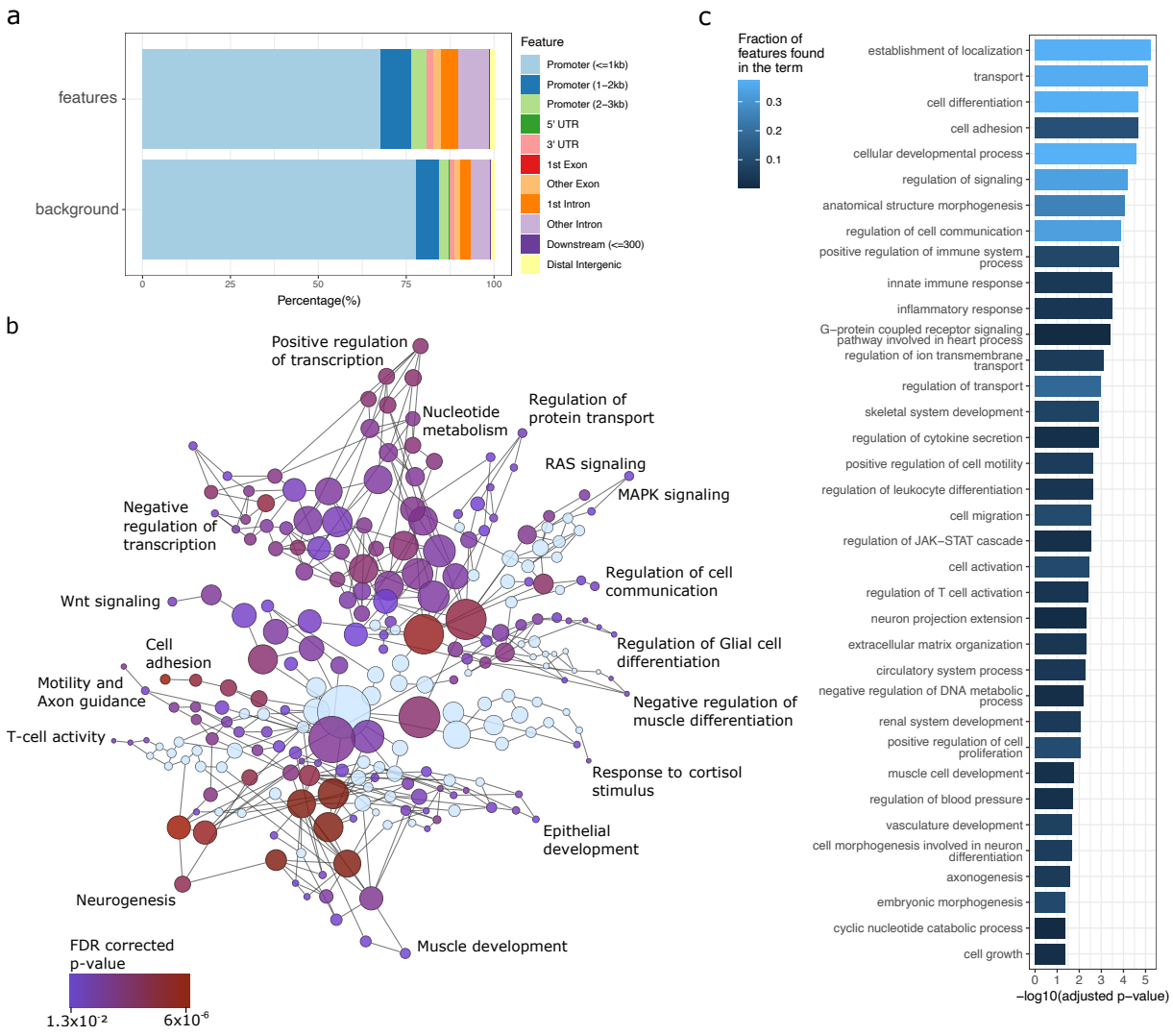


Figure 7: The genomic location and functional annotation of the features selected by Methly-BoostER. **a** The distribution of genomic locations (relative to genes) for the selected features, compared to the background (the total set of input features). **b** Enriched Gene Ontology terms from the Biological Process category represented as a network, where each branch represents a different functional category. Results obtained from the gene-wise GO analysis. **c** Enriched Gene Ontology terms from the Biological Process category represented as a barplot. Results obtained from the localised region GO analysis.

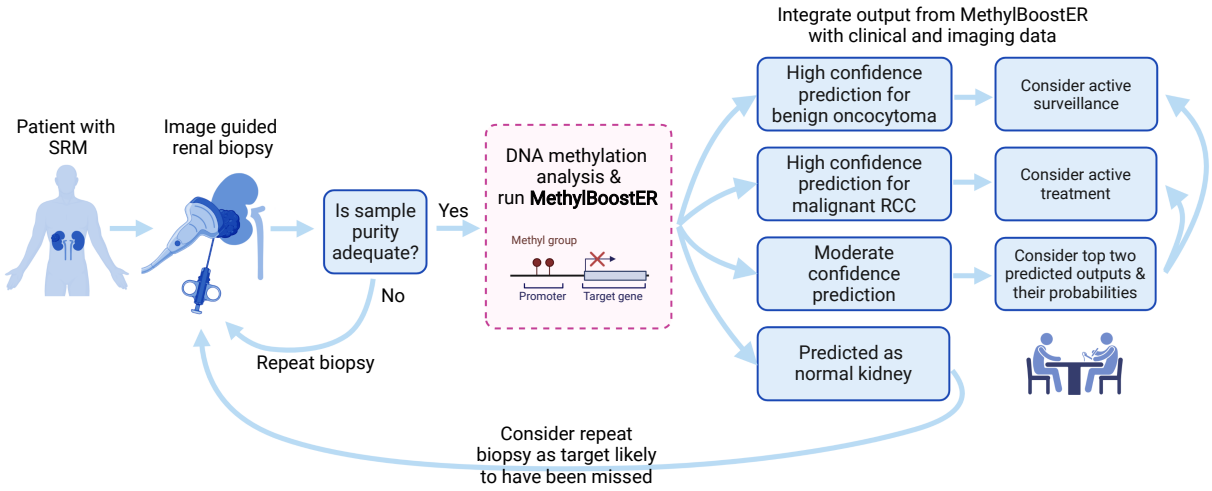


Figure 8: Proposed future integration of MethylBoostER model into the existing diagnostic pathway for patients with small renal masses (SRM). Patients would have an image-guided renal biopsy and biopsy samples would undergo DNA methylation analysis. MethylBoostER results would be interpreted in the context of integration with clinical and imaging data. For high-confidence predictions, MethylBoostER would predict one class, where benign oncocytoma and malignant RCC would likely be managed with active surveillance and active treatment respectively. In moderate-confidence predictions, the two classes with the highest probabilities would be predicted. Samples with low purity or cases in which MethylBoostER predicts normal kidney (suggesting the target lesion was missed) would prompt repeat biopsy.

Supplementary Figures

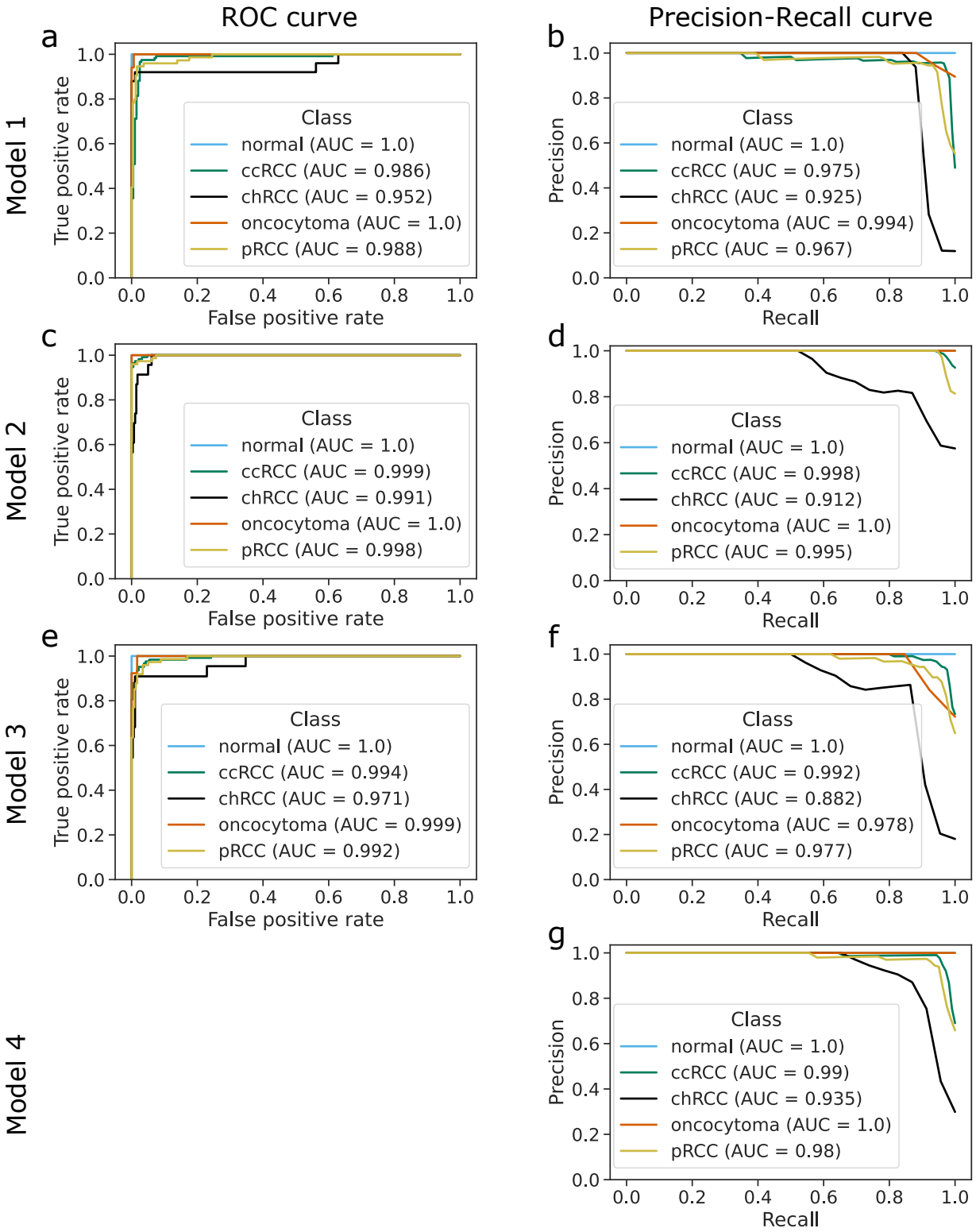


Figure S1: ROC and Precision-Recall curves for all four models on the testset. The ROC curve for model 4 can be found in the main Figure 2e.

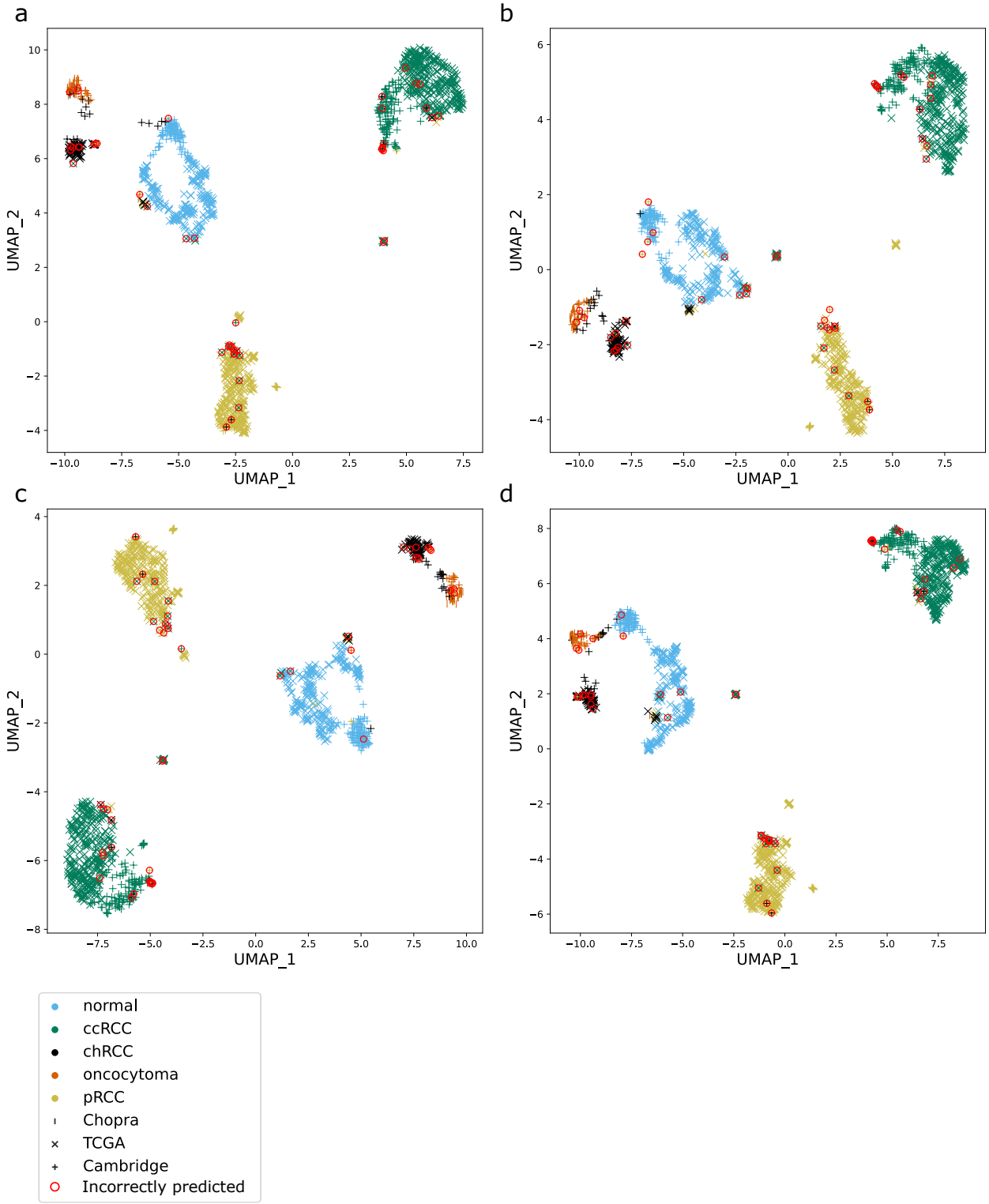


Figure S2: UMAP of samples in the training/test set, for all four models (subfigures a, b, c and d represent model 1, 2, 3 and 4, respectively). Only features selected by each XGBoost model were input into the UMAP algorithm.

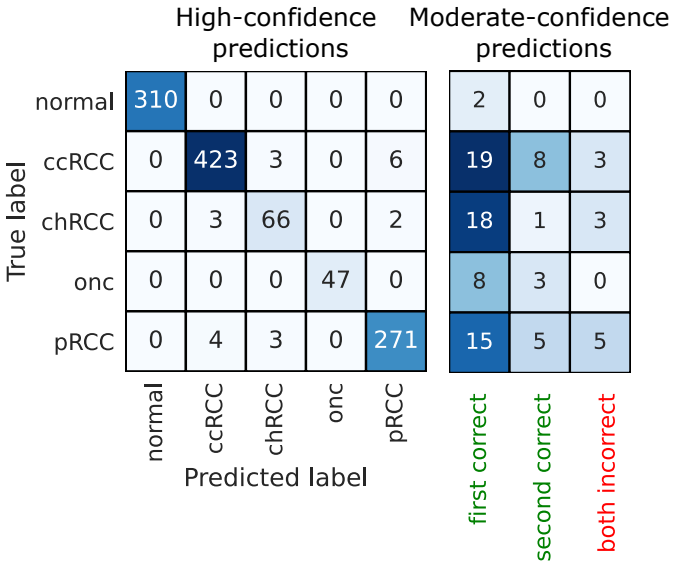


Figure S3: Testing set confusion matrix split by high- and moderate-confidence predictions. For moderate-confidence predictions, we show whether the first prediction was correct, the second prediction was correct, or whether both predictions were incorrect.

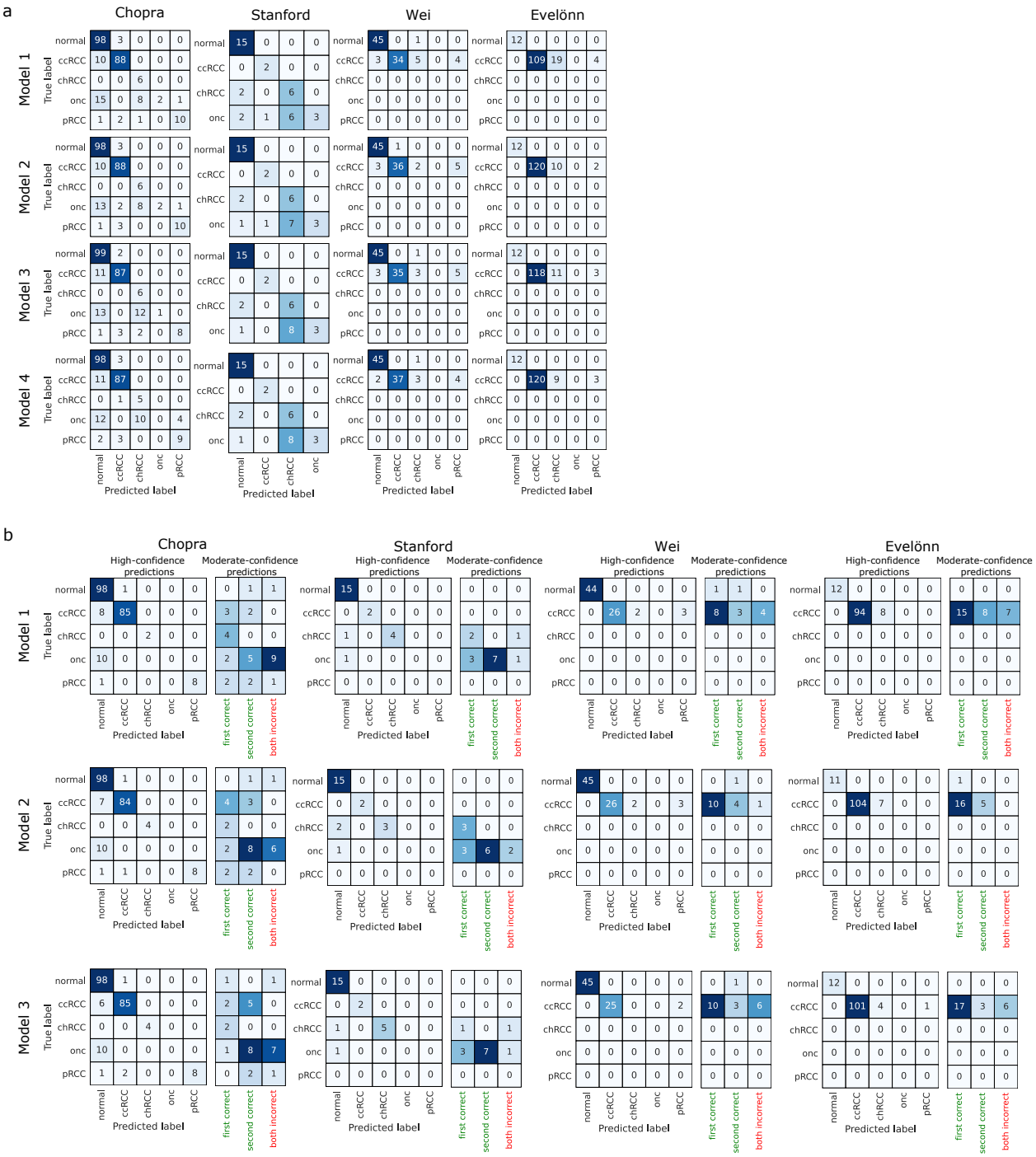


Figure S4: Performance of all four XGBoost models on all four external datasets. **a** Confusion matrices for all models on all datasets. **b** Confusion matrices split by high- and moderate-confidence predictions. For Model 4, see main Figure 4.

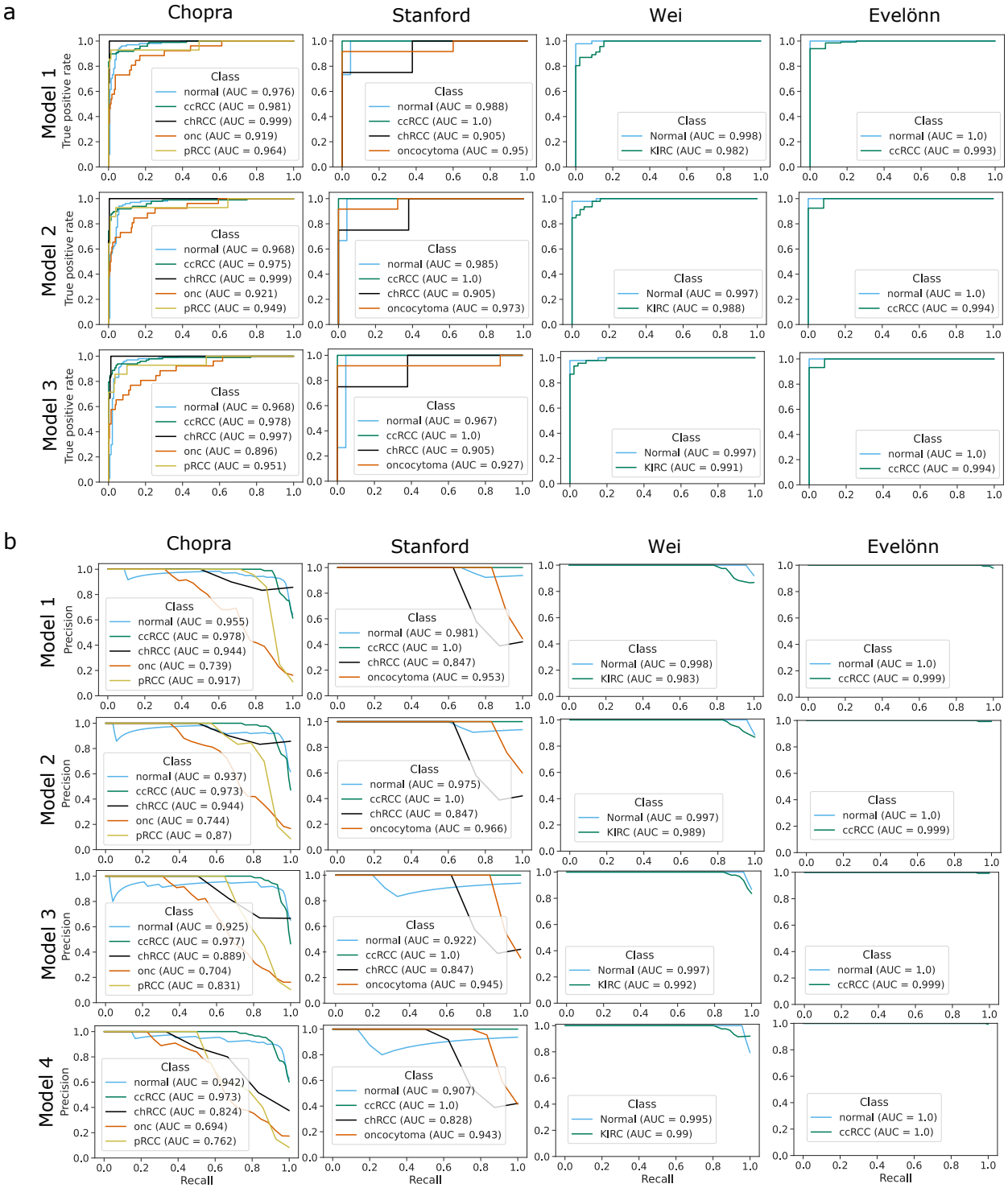


Figure S5: Performance of all four XGBoost models on all four external datasets, shown through ROC curves (**a**) and Precision-Recall curves (**b**). For Model 4's ROC curves, see main Figure 4.

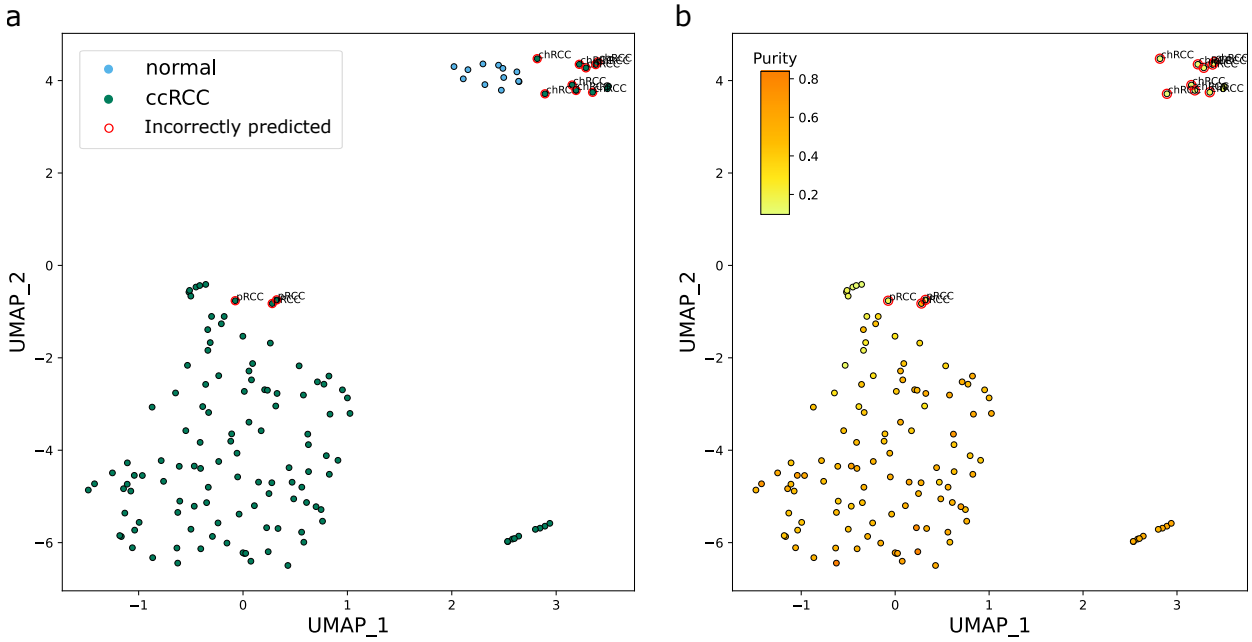


Figure S6: UMAP representation of the Evelönn external dataset, using the input features learnt by XGBoost model 4. Incorrectly predicted samples are highlighted and labelled with the model's prediction. **a** Colour represents diagnosis. **b** Colour represents tumour purity, and only ccRCC samples are shown.

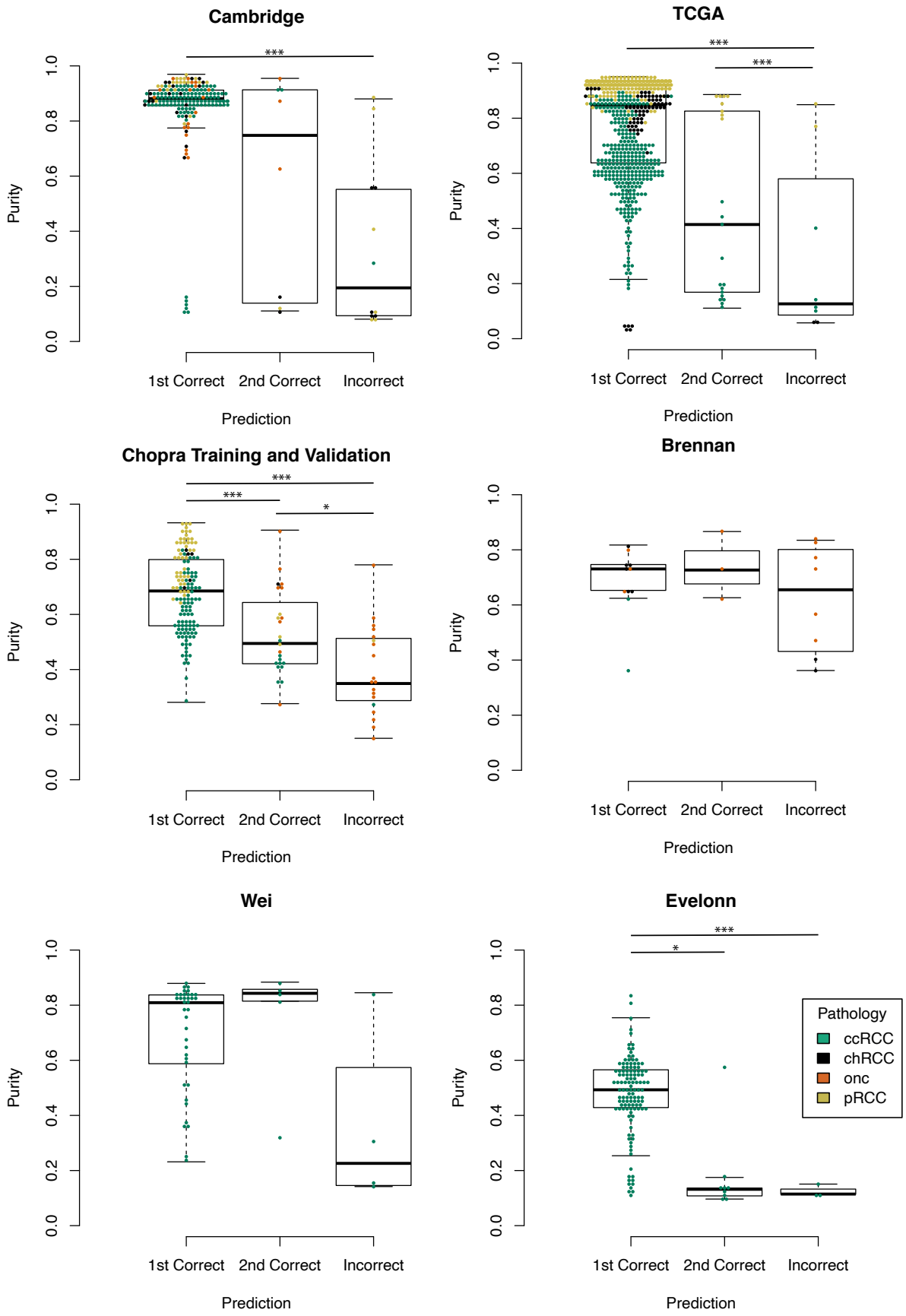
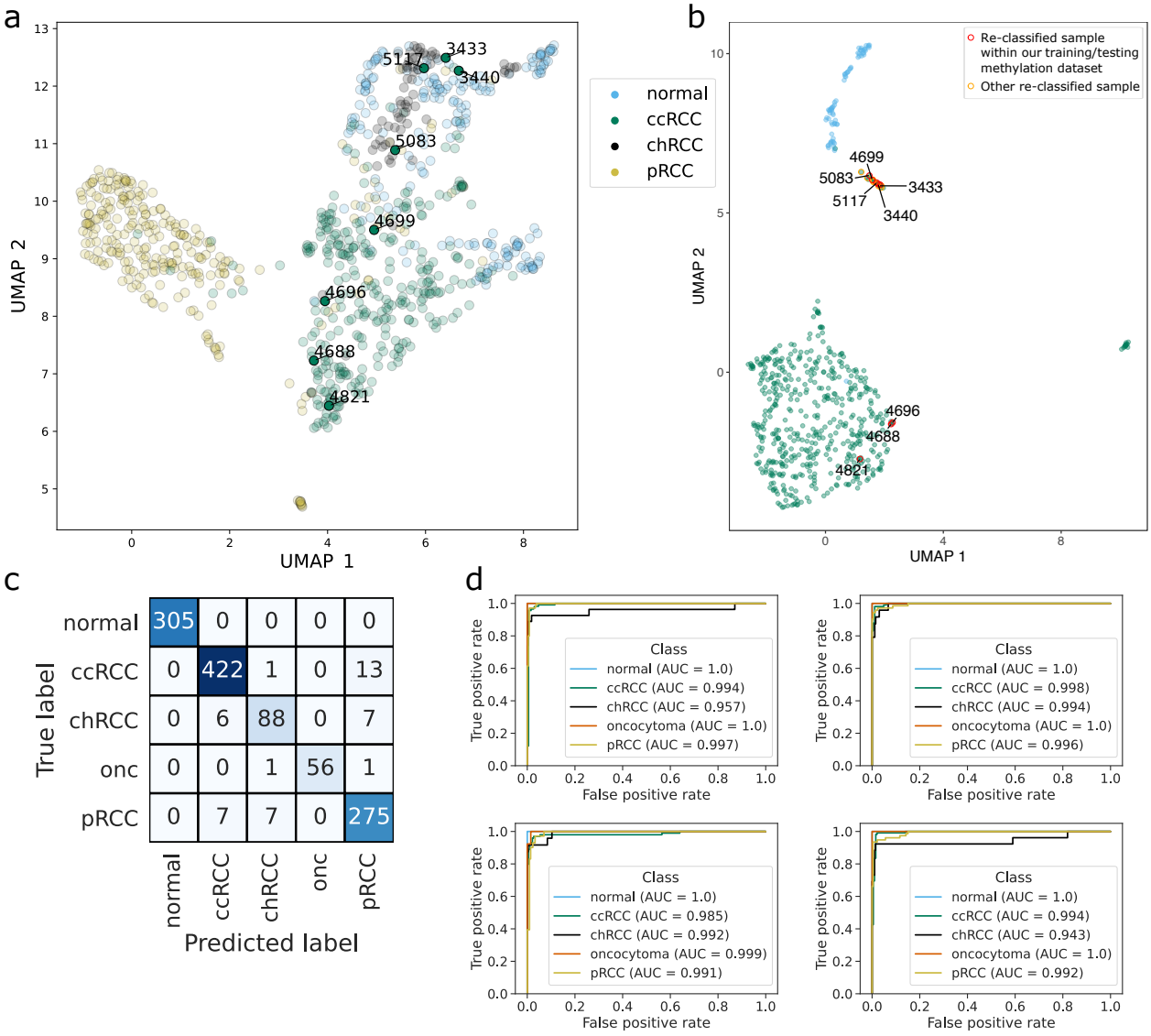


Figure S7: Sample purity for samples which are predicted correctly on the first prediction (1st correct), second prediction (2nd correct) and incorrectly predicted samples (incorrect) on both predictions. Data is shown for each individual dataset, with pathological subtypes shown in different colors. Adjusted p values are shown (* <0.05, ** <0.009, *** <0.0009)

Figure S8: Methylation distribution of the 38 features that were selected by all four XGBoost models. Each swarm plot shows the M-value distribution for each diagnosis. Where a feature maps to a gene (or genes), the gene diagram is positioned directly below, with the feature location marked in red. The Transcriptional Start Site (TSS) for each gene is also indicated and the x-coordinates are chromosome positions.



True label \ Predicted label	normal	ccRCC	chRCC	onc	pRCC
normal	305	0	0	0	0
ccRCC	0	422	1	0	13
chRCC	0	6	88	0	7
onc	0	0	1	56	1
pRCC	0	7	7	0	275

Figure S9: **a** UMAP of all TCGA samples in the training/testing methylation dataset. The eight samples in our dataset that were re-classified by Ricketts et al. as chRCC are indicated and labelled with their TCGA participant ID. **b** UMAP of the KIRC TCGA RNA-Seq dataset. The samples that were re-classified by Ricketts et al. as chRCC are indicated, and the eight samples that are also in the methylation dataset are labelled with their TCGA participant ID. 'Other re-classified sample' refers to re-classified samples that are not included in the methylation dataset. **c** and **d** Confusion matrix and ROC curves of the testing set results after training MethylBoostER with the re-classified labels (eight ccRCC samples re-classified as chRCC, and a number of non-RCC samples removed).

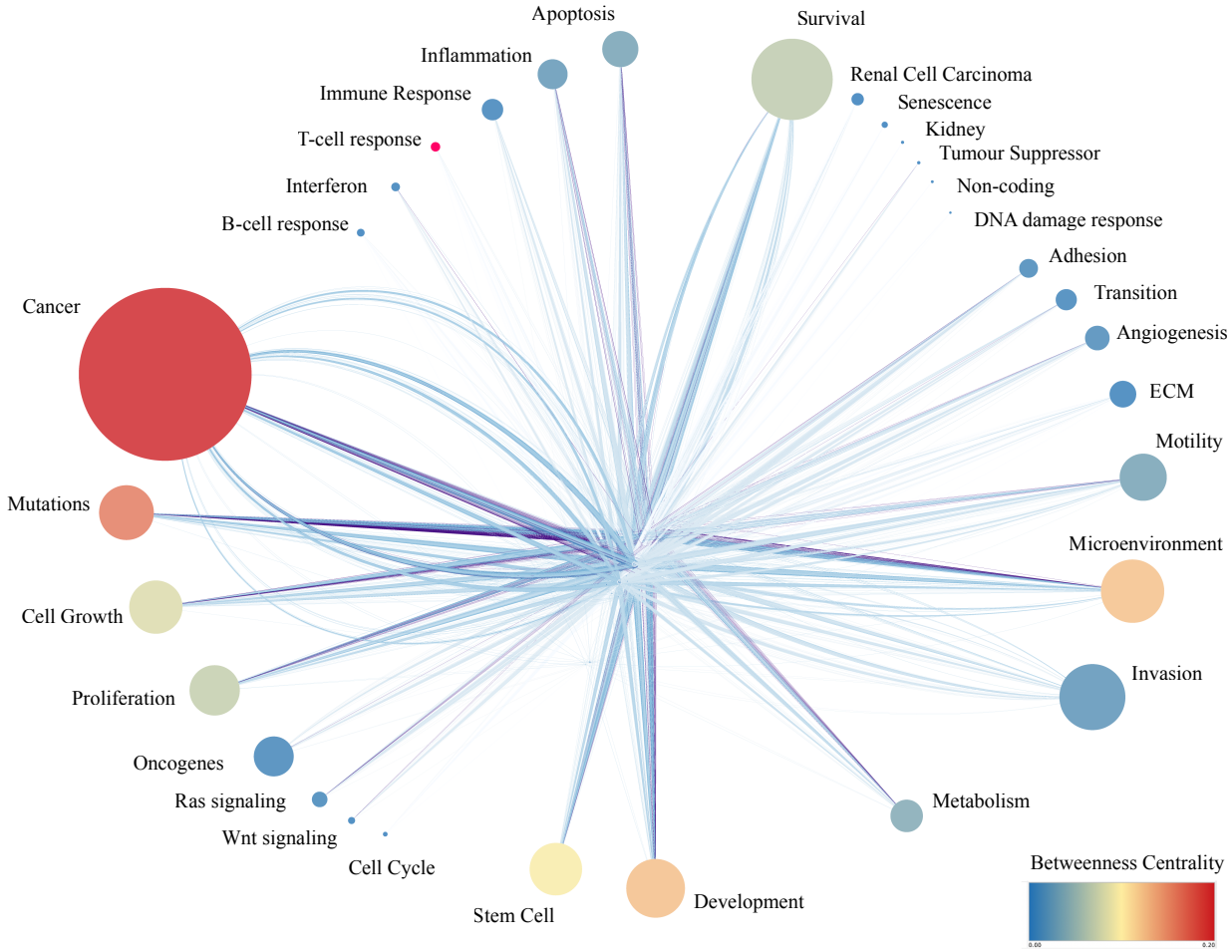


Figure S10: **a** Topology analysis of methylation feature Entity-Relationships. Visualization of 28579 entity-relationships between MethylBoostER derived methylation feature mapped genes and renal carcinoma related relationship terms text mined from 30 million PubMed abstracts. Node size represents local connectivity(degree) and node colour represents global importance (betweenness centrality). Edge colour represents gene publication frequency, with darker edges representing high values.

Supplementary Tables

Table S1: Average performance metrics on the testing set

	normal	ccRCC	chRCC	onc	pRCC
F1	1.0	0.962	0.865	0.974	0.944
Precision	1.0	0.967	0.832	1.000	0.944
Recall	1.0	0.957	0.903	0.949	0.944
Accuracy			0.960		
MCC			0.945		

Table S2: Performance metrics for all external validation datasets, averaged across all four models.

		normal	ccRCC	chRCC	oncocytoma	pRCC
Chopra	precision	0.797	0.934	0.365	0.750	0.878
	recall	0.973	0.893	0.958	0.048	0.661
	F1	0.876	0.913	0.527	0.090	0.748
Stanford	precision	0.822	0.833	0.455	1.00	
	recall	1.000	1.000	0.750	0.25	
	F1	0.902	0.900	0.566	0.40	
Wei	precision	0.942	0.993			
	recall	0.978	0.772			
	F1	0.960	0.868			
Evelönn	precision	1.0	1.000			
	recall	1.0	0.884			
	F1	1.0	0.938			

Table S3: Accuracy and MCC for all external validation datasets, split by high- and moderate-confidence predictions.

	High-confidence				Moderate-confidence, first or second prediction			
	Chopra	Brennan	Wei	Evelönn	Chopra	Brennan	Wei	Evelönn
Accuracy	0.905	0.933	0.944	0.948	0.720	0.782	0.797	0.858
MCC	0.842	0.783	0.895	0.794	0.668	0.632	0.439	0.250

Table S4: Accuracy of both first and second predictions and sample numbers for different levels of purity, averaged over all datasets (both the training/testing set and all external validation sets). Normal kidney samples do not have a purity estimated and therefore are not included in this analysis, hence the total number of samples is 1246.

Purity	Median probability of 1st prediction	Accuracy of 1st prediction	Accuracy of 2nd prediction	Both predictions incorrect	Number of samples
0.0-0.1	0.78	0.42	0.08	0.5	12
0.1-0.2	0.85	0.37	0.39	0.24	49
0.2-0.3	0.99	0.71	0.08	0.21	24
0.3-0.4	0.98	0.6	0.1	0.3	30
0.4-0.5	0.99	0.82	0.14	0.04	95
0.5-0.6	1	0.91	0.04	0.05	159
0.6-0.7	1	0.97	0.03	0	159
0.7-0.8	0.99	0.92	0.05	0.04	111
0.8-0.9	0.99	0.94	0.04	0.02	294
0.9-1.0	1	0.99	0.01	0	313
All samples	1	0.9	0.06	0.04	1246

Table S5: MethylBoostER's features were mapped to genes and compared to a number of gene lists using Fisher's tests. Here are the gene list details and adjusted p-values showing the significance of the overlap.

Gene list description	Name	Size of gene list	Size of overlap	Adjusted p-value
Human Transcription Factors	TF Checkpoint (*29*)	3479	701	5.97×10^{-15}
Epigenetic regulators	Epifactor db (*30*)	720	147	1.27×10^{-2}
RCC related gene set	RCC Harmonizome/Diseases db (*31*)	829	193	9.17×10^{-7}
ccRCC putative driver genes	ccRCC driver gene panel (*32*)	230	59	1.07×10^{-4}
pRCC related gene set	pRCC Harmonizome/Diseases db (*31*)	81	19	1.16×10^{-1}
Genes associated with tumor metastasis	HCMDB (*33*)	1939	478	2.07×10^{-16}
Epithelial-mesenchymal transition genes	dbEMT2 (*34*)	1184	294	2.58×10^{-16}
Genes that play a causal role in cancer	COSMIC Cancer Gene Census genes (*35*)	723	190	5.85×10^{-13}

Table S6: The literature evidence and associations for the 45 genes common to all models

Association	**Genes**	**References**
ccRCC	ALCAM, PVT1, VMP1, NDUFA4L2, CLDN1	(*37–39*)
chRCC	SHMIT2	
Renal cell carcinoma	TESC, BST2, THRA, CLDN1, SLC16A3, KHDRBS2, MIR21	(*40–42*)
Kidney development	ALCAM, FOXN1, SRGAP3	(*43*)
Gene expression in kidney	SSC4D, KLF12	
Adhesion, motility, invasion and metastasis	ALCAM, ATP11A, NCOR2, PVT1, TESC, BST2, VMP1, HDLBP, CLDN1, SLC16A3, DLGAP1	(*44–47*)
Poor survival in renal cancer and other cancer types	PVT1, SHMT2, VMP1, NDUFA4L2, SC3A2, CLDN1, KLK11	(*48–50*)
Increased survival	THRA	
Act as tumour suppressors	SCIRT, NR1D1	
Promote cancer stemness	ALCAM, NCOR2	(*44*)
Carcinogenesis, proliferation and tumour growth	PVT1, SHMT2, TESC, BST2, SLC16A3, MIR1204, DLGAP1	
Have been identified in other cancer types	CHRM2, ZP2	
Not previously associated with kidney function, renal disease or cancer	ANKMY1, EHMT1, H2AW, JDP2, H2BU1, KIAA1143, KIF15, LINC01960, SF3A2	