Disease aetiology and progression shape the inter-patient multi-omics profile of clear cell renal carcinoma

Ricardo Cortez Cardoso Penha
International Agency for Research on Cancer

Alexandra Sexton-Oates
Rare Cancers Genomics Team (RCG), Genomic Epidemiology Branch (GEM), International Agency for Research on Cancer/World Health Organisation (IARC/WHO)

Sergey Senkin
International Agency for Research on Cancer (IARC/WHO) https://orcid.org/0000-0001-5848-005X

Hanla A. Park
International Agency for Research on Cancer https://orcid.org/0000-0001-8055-3729

Joshua Atkins
University of Oxford https://orcid.org/0000-0003-0821-1112

Ivana Holcatova
Charles University https://orcid.org/0000-0002-1366-0337

Anna Hornakova
1st Faculty of Medicine, Charles University https://orcid.org/0000-0002-1024-8056

Slavisa Savic
University Hospital https://orcid.org/0000-0002-3635-7437

Simona Ognjanovic
International Organization for Cancer Prevention and Research, Belgrade, Serbia

Beata Świątkowska
Nofer Institute of Occupational Medicine

Jolanta Lissowska
Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology https://orcid.org/0000-0003-2695-5799

David Zaridze
N.N. Blokhin National Medical Research Centre of Oncology

Anush Mukeria
Department of Epidemiology and Prevention, Russian N.N.Blokhin Cancer Research Centre, Moscow, Russian Federation

Vladimir Janout
Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic

Amelie Chabrier
International Agency for Research on Cancer
Vincent Cahais
IARC
Cyrille Cuenin
EpiGenomics and Mechanisms Branch (EGM); International Agency for Research on Cancer/World Health Organisation (IARC/WHO)
Ghislaine Scelo
IARC/WHO
Matthieu Foll
International Agency for Research on Cancer https://orcid.org/0000-0001-9006-8436
Zdenko Herceg
International Agency for Research on Cancer
Paul Brennan
International Agency for Research on Cancer (IARC/WHO)
Karl Smith-Byrne
Cancer Epidemiology Unit, University of Oxford
Nicolas Alcala
International Agency For Research On Cancer / World Health Organization https://orcid.org/0000-0002-5961-5064
James D. McKay
mckayj@iarc.who.int

International Agency for Research on Cancer, World Health Organization, Lyon, France

---

Article

Keywords:

Posted Date: February 13th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3891211/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Endogenous and exogenous processes are associated with distinctive molecular marks in somatic tissues, including human tumours. Here, we used integrative multi-omics analyses to infer sources of inter-patient somatic variation within clear cell renal cell carcinomas (ccRCC) and used them to explore how the disease aetiology and progression are reflected in the tumour DNA methylome, transcriptome, and somatic mutation profile. The main source of inter-patient variation within ccRCC tumours was associated with ageing, particularly cellular mitotic age estimated by DNA methylation (epiTOC2), clock-like DNA mutational signatures (SBS1/ID1), and telomere attrition, independent to chronological age. This component was associated with *PBRM1* and *SETD2* somatic cancer driver mutations, genome instability, tumor stage, grade, and ccRCC patient survival. Pan-cancer analysis supported the similar role of this molecular component in other cancer types. The ccRCC tumour microenvironment was another source of inter-patient variation, including a component associated with *BAP1* driver mutations, epigenetic regulation of epithelial-mesenchymal transition genes (i.e., *IL20RB, WT1*) and patient survival. An additional source of ccRCC inter-patient variation was linked to the epigenetic regulation of the xenobiotic metabolism gene *GSTP1*. This molecular component was associated with tobacco usage and tobacco-related genomic features, implying a relationship with tobacco-related carcinogenesis, but also present in tumours of never-smoking patients, potentially implicating it in other genotoxic effects. By considering how the tumour DNA methylome, transcriptome, and somatic mutation profile vary across patients, we provide novel insights into the endogenous and exogenous processes acting within ccRCC tumours and their relation to the disease aetiology and progression.

Introduction

Renal cell carcinoma (RCC) is the 16th most common cancer type worldwide, accounting for ~ 2% of all cancer patient deaths in 2020 (Sung et al., 2021). Clear cell RCC (ccRCC) is the most frequent histological subtype (~ 75%) (Hsieh et al., 2017). The incidence rates of ccRCC are higher in high-income countries, particularly in central and northern Europe, (Hsieh et al., 2017) with an increasing trend in incidence globally (Huang et al., 2022). Risk factors associated with ccRCC include age, sex, obesity, hypertension, and tobacco smoking; although they collectively explain less than 50% of the newly diagnosed cases (Hsieh et al., 2017).

A better understanding of the underlying molecular processes associated with ccRCC tumours could provide new insights into disease aetiology and how the tumour progresses. Endogenous and exogenous exposures leave distinct molecular marks through the course of a lifetime, detectable at DNA and RNA levels. Recurrent DNA mutation patterns, or DNA mutational signatures, have been linked to endogenous (e.g., cellular ageing, *APOBEC* activity) and exogenous exposures (e.g., tobacco smoke, aristolochic acid) (Alexandrov et al., 2015; Alexandrov et al., 2020, Scelo et al., 2014; Senkin et al., 2023). The methylome is similarly impacted by exogenous and endogenous processes (Herceg et al., 2018), with tobacco smoking, ageing and somatic driver mutations provoking aberrant DNA methylation patterns (Linehan et al., 1995; Gerlinger et al., 2012; Cancer Genome Atlas Research Network, 2013; Hakimi et al., 2013; Hannum et al.,

2013; Horvath et al., 2013; Guida et al., 2015; Joehanes et al., 2016; Şenbabaoğlu et al., 2016; Yang et al., 2016; Ricketts et al., 2018; Levine et al., 2018; Motzer et al., 2020; Halaburkova et al., 2020; Belsky et al., 2020; Chamberlain et al., 2022). Integrative multi-omics approaches, which detect sources of inter-individual variability, can improve representation of the molecular processes present in somatic tissues compared with a single-omic approach (Argelaguet et al., 2020; Cantini et al., 2021). In the current study, we have investigated if the integration of somatic DNA mutational signatures, cancer driver mutations, DNA methylome, and transcriptome profiles from well-characterized cohorts of ccRCC patients can provide novel insights into kidney cancer aetiology and disease progression.

Results and discussion

The current work used a two-stage study design. An initial discovery phase identified the sources of molecular variance (or latent factors) across tumours of ccRCC patients using the unsupervised Multi-Omics Factor Analysis (MOFA) approach to integrate DNA methylome, transcriptome, and somatic mutation profile data from whole-genome sequencing (WGS) data (see methods; Supplementary Fig. 1–2). We used LASSO regression (see methods, Supplementary Table 1) to select key features to establish signatures for ccRCC latent factors and attributed these signatures to independent cohorts of ccRCC patients in a validation phase. Associations between latent factors and molecular and epidemiological features were then tested within the discovery and validation phases. The characteristics of each cohort are described in Supplementary Table 2 and workflow of the study in Supplementary Fig. 1.

Description of molecular components in ccRCC tumours

Collectively, 31%, 41%, and 6% of the variance in DNA methylome, transcriptome, and somatic mutation profile data, respectively, were explained by the first six latent factors estimated by MOFA (Fig. 1). While pan-cancer analyses reported the global DNA hypomethylation of tumours in comparison with their histological normal material (Witte et al., 2014), the CpG sites associated with latent factors tended to be hypermethylated and annotated to functional regions of the genome (CpG islands, shores, and shelves within regulatory and coding regions) (Fig. 2A, Supplementary Tables 3–6, Source data file). There was no consistent evidence for associations between latent factors and global DNA methylation, measured by the mean methylation levels of *Alu* and *LINE1* transposable elements (Supplementary Table 7), implying that DNA methylation changes in specific functional regions of the genome contributed to most of the inter-patient variation in the DNA methylome. Important inter-patient variation in the gene expression levels (transcriptome) across ccRCC tumours were captured by the latent factors (Fig. 1). Pathway analysis showed that the gene expression levels with the highest loadings in each latent factor were enriched for cancer-related pathways, such as those involved on immune system, metabolism, cell cycle, cell plasticity and signalling, chromatin remodelling, and tissue development (Fig. 2B, Supplementary Table 8). Latent factors were also related to a range of DNA mutational signatures, including those implicated in endogenous (i.e., SBS1, SBS13) and exogenous (e.g., SBS4, DBS2) processes, as well as signatures where the aetiology remains unclear (i.e., SBS40a,b,c, ID5). Latent factors were also related to the presence of ccRCC somatic cancer driver mutations in genes that act as epigenetic regulators (*PBRM1, SETD2, BAP1*, and *KTM2C*) (Fig. 2C, Source data file).

In the following sections, we will report the key findings of latent factors 1 to 6 using MOFA (discovery) and their respective signatures (validation) to illustrate sources of ccRCC inter-patients' variability potentially attributed to endogenous and exogenous exposures (more details in MOFA subsection in methods).

The major ccRCC molecular component is linked to cellular mitotic age.

The major component of ccRCC inter-patients' variability (latent factor 1) was typified by DNA hypermethylation of CpG islands annotated to functional regions of the genome (Fig. 2A, Supplementary Table 4). These profiles were similar to those noted in cellular ageing processes (Marttila et al., 2015; Bell et al., 2019) and latent factor 1 was associated with chronological age (Table 1), prompting us to explore this ccRCC component in the context of biological ageing. Biological age is a complex process that reflects an individual's physiological state over time; varying aspects can be measured by epigenetic clocks (Rutledge et al., 2022). Latent factor 1 was correlated with a range of epigenetic clocks, showing the highest correlation with the age-adjusted mitotic-like clock epiTOC2 (Teschendorff, 2020) (Supplementary Fig. 3). EpiTOC2 is based on the CpG sites of Polycomb target genes that are unmethylated at birth but become progressively methylated as cells replicate, a process called cellular mitotic age (Yan et al., 2016, Teschendorff, 2020). Consistent with this, pathway analysis of gene expression levels correlated with latent factor 1 suggested an enrichment for Polycomb (*EZH2*) target genes (Supplementary Table 8). Linear regression models estimated that this epigenetic clock explained around 80% of the variance in latent factor 1 across ccRCC tumours (Fig. 3A). Latent factor 1 was also associated with other types of mitotic clocks, including clock-like DNA mutational signatures (SBS1/ID1) (Fig. 2C) and telomere attrition (Supplementary Table 9), reinforcing the hypothesis that these metrics act as proxies for the number of cell mitosis (Alexandrov et al, 2015; Yang et al., 2016, Alexandrov et al, 2020; Teschendorff, 2020). Similarly, phenotypic consequences consistent with higher mitotic counts, such as higher copy number alteration fraction of the genome and homologous recombination DNA repair deficiency, were associated with this ccRCC component (Supplementary Table 9). It was also associated with the presence of somatic cancer driver mutations in *PBRM1* and *SETD2*, chromatin remodeling genes involved in cell senescence (Lee et al., 2016) and proliferation (Dominguez et al., 2016; Cai et al., 2019) (Fig. 2A, Supplementary Table 9). Latent factor 1 was strongly associated with tumour stage and grade (Table 1, Fig. 3B). Patients in the top quintile of latent factor 1 were estimated to be 23 times more likely to be late-stage tumours (III and IV) and 9 times more likely to be high-grade tumours (grade 3–4), compared to those in the bottom quintile (Supplementary Table 9). Nevertheless, the levels of latent factor 1 considerably varied within ccRCC tumours within tumour stage and grades (Fig. 3B) and multivariate analysis suggested that the associations between latent factor 1 appeared not to be driven by tumour stage and grade alone (Supplementary Table 9). Consistent with the notion that cellular mitotic age is accelerated in tumour tissues, the patient's tumor material had marked higher values of latent factor 1 (tumours: mean of 0.43 ± 0.95; normal tissue: mean of -0.87 ± 0.20, $p = 1.6x10^{-45}$) than paired normal kidney tissue after adjustments for chronological age (Fig. 3C). Taken together, the main source of ccRCC inter-patients' variability in the tumour DNA methylome appears related to cellular

mitotic age, which may be influenced by proliferative effects of somatic mutations in *PBRM1* and *SETD2*. The top 5 epigenetically regulated genes with the highest loadings in latent factor 1 encode zinc-finger transcription factors with pleotropic role in cancer (Jen and Wang, 2016), such as *ZNF471* previously linked to ageing process (Marttila et al., 2015), and *SPON1*, essential to nephron formation in mice (Vidal et al., 2020) (Supplementary Table 3). Such genes may provide additional clues to explain how the deregulation of genes important to kidney homeostasis might contribute to the biological ageing process in ccRCC tumours.

The latent factor 1 signature was inferred across tumour samples from the TCGA pan-cancer cohorts (N = 8,040) (see methods, Supplementary Table 1). The signature for latent factor 1 explained on average 15% of variance in tumor DNA methylation in TCGA patient cohorts, ranging from 31% of variance explained in lymphoid neoplasm diffuse large B-cell lymphoma (TCGA-DLBC) to 5% in Uveal Melanoma (TCGA-UVM) (Supplementary Table 10). Independent to ccRCC tumours (TCGA-KIRC), the latent factor 1 signature was strongly associated with epiTOC2, dosage of the clock-like mutational signature SBS1, WGS-telomere length ratio, and copy number alterations in a joint model of TCGA tumours combined (Table 2). Intriguingly, these effects also appeared particularly prominent in other histological subtypes of kidney cancer (TCGA-KICH/chromophobe, TCGA-KIRP/papillary) as well as in other tumour types (e.g., adrenocortical carcinoma/TCGA-ACC, mesothelioma/TCGA-MESO, pancreatic/TCGA-PAAD) (Supplementary Table 10).

Molecular components related to the ccRCC tumour microenvironment.

To further explore how tumour microenvironment (TME) contributes to ccRCC inter-patients' variability across multi-omics layers, we imposed 65 gene expression TME signatures derived from previously published single-cell RNA sequencing (scRNA) data of ccRCC tumours (Li et al., 2022) into bulk ccRCC tumour transcriptomes (see methods). The scRNA-derived signatures were correlated within patient's tumours (Supplementary Fig. 4) and 27 representative signatures could be derived from the 65 TME signatures (see methods). Figure 4A describes the replicated associations between TME signatures and latent factors in the validation series, suggested from the discovery series (Supplementary Fig. 5A). Latent factors 2–6 were associated with TME signatures related to kidney epithelial, immune cell infiltrates, inflammation, epithelial-mesenchymal transition process (EMT) and cell proliferation processes (Fig. 4B), consistent with the TME making an important contribution to the inter-patient's variability in ccRCC (Li et al., 2022).

Latent factor 2 captured shared inter-patients' variability across omics layers (Fig. 1), associating with DNA methylation changes (annotated to CpG island and open sea, Fig. 2A), immune system-related pathways (Fig. 2B, Supplementary Table 8), and *BAP1* cancer driver mutations (Fig. 2C). Higher latent factor 2 levels were observed in male patient's tumours compared with female patient's tumours (Table 1). This ccRCC component was associated with the presence of TME signatures related to EMT process, proliferating cells (cycling endothelial cells and cell cycle kidney meta-programs) and myeloid cells (particularly fibronectin-positive tumour associated macrophages/FN1_TAM) (Fig. 4A,

Supplementary Fig. 5A), and late-stage, high-grade tumours (Table 1). Of the differentially methylated regions associated with latent factor 2, the expression of EMT related genes (*WT1, IL20RB, KRT19*) (Kalluri & Weinberg, 2010; Wang et al., 2020; Guo et al., 2022; Wu et al., 2023) appeared epigenetically upregulated (Supplementary Table 5). The CpG sites annotated to *IL20RB* (interleukin 20 receptor subunit beta) displayed the highest functional impact on its transcript levels (Supplementary Table 5), with higher levels of latent factor 2 associating with DNA hypomethylation (1st exon and 5-UTR) of *IL20RB* and respective upregulation of its RNA levels (Fig. 4B). The functional regulation of *IL20RB* expression by DNA methylation was pronounced in the presence of *BAP1* cancer driver mutations regardless of tumor stage in ccRCC (Supplementary Fig. 5B). Multivariate analyses suggested that *IL20RB* expression levels predicted TME features (i.e., cell cycle, EMT, and FN1_TAM) in addition to the effects of somatic cancer driver mutations in *BAP1* gene (Fig. 4C). Together, these results indicate that latent factor 2 is representing a complex relationship between the DNA hypomethylation in specific CpG sites and increased expression of EMT and cell proliferation related genes (i.e., *IL20RB*, *KRT19*, *WT1*), *BAP1* cancer driver mutations, and ccRCC tumour microenvironment components (i.e., FN1_TAM) that could contribute to disease progression and unfavorable prognosis of the ccRCC patients, particularly in men.

Latent factor 6, epigenetic regulation of GSTP1, and genotoxicity

An additional source of inter-patient's variability was an intriguing ccRCC component (latent factor 6) linked to DNA hypermethylation and gene expression changes (Fig. 1). This component was associated with tobacco smoking (Table 1), the environmental exposure robustly associated with ccRCC risk by observational studies (Hsieh et al., 2017). The DNA methylation patterns in CpG sites in proximity to CpG islands and gene bodies noted with latent factor 6 (Fig. 2A) were similar to those observed in tobacco smokers in blood (Guida et al., 2015; Joehanes et al., 2016; Plusquin et al., 2017; Svoboda et al., 2021); it was also associated with the dosage of tobacco-related DNA mutational (SBS4 and DBS2) (Fig. 2C) (Alexandrov et al., 2020) and methylation signatures (Fig. 5A) (Chamberlain et al., 2022). Of the differential methylated regions associated with latent factor 6, the CpG sites annotated to *GSTP1* (glutathione S-transferase pi) displayed the highest functional impact on its transcript levels (Supplementary Table 6). Latent factor 6 was related to hypermethylation and decreased expression of *GSTP1* (Fig. 5A), estimated to jointly explain around 36% of the variance in latent factor 6. *GSTP1* is a key enzyme involved in phase II detoxification of xenobiotics by glutathione conjugation and it has been implicated in the metabolism and clearance of a variety of genotoxic compounds (e.g., the carcinogens in tobacco smoke, cisplatin, mercury as well as endogenous free radicals) (Miller et al., 2003; Simic et al., 2009; Sawers et al., 2014; Shin et al., 2017). Epigenetic silencing of *GSTP1* is postulated to increase cellular sensitivity to genotoxic compounds (Su et al., 2007; Rønneberg et al., 2008, Cui et al., 2020). Consistent with this, when excluding the outlier effect of high mutation burden tumours from Romania, latent factor 6 was associated with total tumor mutation burden (Fig. 5A) among tobacco smokers, particularly in ever smokers (Supplementary Table 11). It is noteworthy that latent factor 6 was also present in never smokers and associated with DNA mutational signatures ID5 and SBS40b (Fig. 2C) commonly observed in ccRCC, suggesting that these DNA mutational signatures may be related to

genotoxic compounds, as raised elsewhere (Senkin et al., 2023). ccRCC tumours with higher loadings of latent factor 6 tended to have lower levels of the gene expression signatures related to interferon gamma (*INFG*) response, implying a decrease in cellular response to *INFG* (Fig. 2B, Fig. 4A, Supplementary Table 8). *INFG* acts as a gatekeeper of ccRCC progression by restraining the clonal expansion of ccRCC cells (Perelli et al., 2023). The suppression of *INFG* response associated with latent factor 6 might also have contributed to formation of a permissive environment for the expansion of these tumour cells and disease progression. Latent factor 6 was also associated with the presence of cancer driver mutations (e.g., *KMT2C*) (Fig. 2C) and ccRCC tumours of female patients had higher loadings of this ccRCC component (Table 1), although how these relate to molecular processes captured by latent factor 6 remains to be elucidated. Interestingly, latent factor 6 was significantly higher in a subset of normal kidney tissues in comparison with the matched ccRCC tumours (mean of 0.95 ± 0.26 *vs.* mean of -0.47 ± 0.89, $p = 6.1x10^{-56}$) (Fig. 5B). While ccRCC tumours tended to have lower RNA levels of *GSTP1* than matched normal kidney tissues, a discrete distribution of DNA methylation levels of *GSTP1* in tumour samples were observed, with one subset of tumours displaying hypermethylation and another one hypomethylation of *GSTP1* CpG sites (Supplementary Fig. 6). Together, these findings suggest that there may be an impaired metabolism of potentially exogenous compounds, such as tobacco smoke exposure, in ccRCC tumours, leading cells to accumulate more somatic DNA mutations once exposed to such genotoxic compounds.

We then inferred latent factor 6 across tumour samples from the TCGA pan-cancer cohorts (see methods, Supplementary Table 12). The associations between latent factor 6, DNA methylation and RNA levels of *GSTP1*, tobacco methylation signature, and total mutation burden were also observed in papillary cell kidney cancer (Supplementary Table 12). However, this relationship varied substantially in other tumour types (Supplementary Table 12), which might reflect the differences in aetiology by tumour type.

The prognostic significance of ccRCC components

Finally, we investigated the prognostic value of the latent factors identified in ccRCC tumours in the discovery and validation sets. Latent factors were associated with ccRCC patients' survival (Fig. 6), with worse survival of ccRCC patients noted for higher values of latent factors 1 ($Validation_{model1}$: HR: 1.63, 95%CI = 1.35–1.98, $p = 4.1x10^{-7}$), latent factor 2 ($Validation_{model1}$: HR: 1.46, 95%CI = 1.16–1.84, p = 0.001), and latent factor 5 ($Validation_{model1}$: HR: 1.34, 95%CI = 1.11–1.63, p = 0.003) at baseline model (model1: sex and age at diagnosis). Despite the respective 14% and 17% attenuation in the effect of latent factors 2 ($Validation_{model2}$: HR: 1.33, 95%CI = 1.06–1.67, p = 0.014) and 5 ($Validation_{model2}$: HR: 1.23, 95%CI = 1.00-1.53, p = 0.055) on the overall survival of ccRCC patients after additional adjustments for tumour stage and grade, their association estimates remained consistent (Fig. 6). Furthermore, these latent factors were also associated with tumour stage and grade (Table 1). When considering other tumour types in the TCGA cohorts, we also noted a similar striking prognostic value for latent factor 1 in patients with papillary kidney cancer (TCGA-KIRP: HR: 2.63, 95%CI = 1.75–3.69, $p = 6.8x10^{-17}$) and other cancer types (e.g., TCGA-LGG, TCGA-ACC) (Supplementary Table 10). Our findings are in line with the

previously described role of *BAP1* somatic mutations, EMT, and tumour immune cells (factor 2), as well as cell proliferation (factor 5) on ccRCC patients' outcomes (Hakimi et al., 2013; Rickets et al., 2018; Motzer et al., 2020).

Discussion and conclusions

We have explored how the DNA methylome, transcriptome, and somatic mutation profiles of ccRCC tumors are shaped by disease aetiology and progression. The main biological sources of ccRCC inter-patient variation were detected using an integrative multi-omics analysis and then attributed into independent normal and tumour validation ccRCC datasets, as well as in other tumour types. We identified how ccRCC risk factors, such as cellular ageing processes, sex, and tobacco smoking behaviours impact the ccRCC tumour's genomic profiles, as well as gained novel insights into tumor microenvironment features and genotoxic agents related to disease aetiology and progression.

Kidney tissue has striking physiological heterogeneity determined by a unique vascularization structure with varying oxygen supply, energy demand, and different physiological functions. The proximal nephron segments are responsible for reabsorption of filtered fluid and solutes, whereas the distal nephrons concentrate urine and regulate salt excretion (Scholz et al., 2021). The function of the proximal nephron segment makes this part of the kidney particularly sensitive to metabolites derived from insults or risk factors. Given that ccRCC tumours arise from the proximal tubule (Young et al., 2018), this notion may explain the extent to which we observed genomic changes across ccRCC tumours to be associated with endogenous and exogenous process linked to risk factors and disease progression.

Some limitations of this study should be recognized. Firstly, our study is primarily focused on ccRCC tumours, which may have limited our ability to distinguish elements involved in disease aetiology (i.e., tobacco smoking) compared to those involved in the progression (tumour microenvironment features such as pro-inflammatory tumour-associated macrophages) of the tumour. We partially mitigated this by inferring the ccRCC molecular components in histologically normal kidney material and exploring molecular processes related to aetiological risk such as cellular ageing. Additional observational and genomic studies are needed to demonstrate factors involved in ccRCC aetiology and progression. Secondly, the integrative multi-omics approach used in this study (MOFA) is based on high-variance molecular variables across tumours capturing the principal sources of inter-patient's variation, and thus, minor sources of variation may not be resolved, which could explain the lack of ccRCC molecular components related to SBS22, the DNA mutational signature suspected to be provoked by the aristolochic acid exposure in ccRCC patients from Romania but only present in less than 15% of patients. The inference of signatures in the validation series, derived from whole-exome sequencing data in contrast to the WGS from the discovery set, is also not expected to be perfectly accurate, which may explain differences in effect sizes between the discovery and validation series. The high rates of missing data in the TCGA datasets also limited further validation of some findings, such as those related to body mass index, hypertension, and tobacco smoking status. WGS data were also absent in the validation cohorts limiting our ability to explore and replicate the relationship between latent factors and DNA

mutational signatures that are more difficult to attribute (e.g., SBS4, SBS40B) in the cohort we used for validation.

In summary, our study expands the current knowledge of the underlying biological processes in ccRCC tumours by unravelling molecular marks linked to endogenous and exogenous exposures detected at different omic layers. This includes ccRCC components linked to cellular mitotic age, tobacco smoke exposure, and tumour microenvironment, with potential prognostic value for patients.

Material and methods

Participants of the study

Discovery set. The participants included in the discovery set were part of the Mutographs project that was coordinated by the International Agency for Research on Cancer (IARC/WHO) with available WGS, transcriptome (microarray), and DNA methylation data (Supplementary Table 2). The participants included in the study met the following criteria (N = 151): age at diagnosis > = 18 yearsold (mean of 60.3 ± 10.7), reviewed diagnosis of primary ccRCC by pathologists following the guidelines from the International Cancer Genome Consortium, and no history of cancer treatment. The exclusion criteria were the non-availability of informed consent or suitable samples according to the protocol requirements. More details in Senkin et al., (2023). **Validation sets**. The IARC ccRCC validation cohort was composed of two IARC-led cohort studies, the K2 study and the NCI/IARC study in Central Europe and was used for the validation of transcriptome findings. Both hospital-based studies had transcriptome data available for ccRCC tumoural (N = 462) and normal adjacent kidney tissue samples (N = 256), with the same inclusion and exclusion criteria as the discovery set (Supplementary Table 2) and described elsewhere (Laskar et al., 2021). The TCGA cohort was used to validate the genomic findings of DNA methylation and related to somatic cancer driver mutations. IARC-ccRCC datasets with available transcriptome data based on microarray technology were used for validation of gene expression. The molecular and clinical information regarding the participants of TCGA cohorts is publicly available at https://portal.gdc.cancer.gov/ and DNA methylation data (normalized beta values), including primary tumours and matched normal adjacent tissues were obtained using TCGAbiolinks R package (version 2.22.3) (Colaprico et al., 2016).

DNA mutational signatures and cancer driver mutations

DNA mutational signatures and cancer driver mutations were obtained from WGS data from the Mutographs project, as described elsewhere (Senkin et al, 2023.). Briefly, WGS was conducted on Illumina HiSeqX platform (Ilumina, San Diego, CA, USA) with a target coverage of 40X and sequence reads were aligned to GRCh38 human reference genome. Somatic variant calling was performed using the standard Wellcome Sanger Institute's analysis pipeline (https://github.com/cancerit). The mutational matrices were generated by SigProfilerMatrixGenerator

(https://github.com/AlexandrovLab/SigProfilerMatrixGenerator), DNA mutational signatures were extracted using the default options of SigProfilerExtractor (https://github.com/AlexandrovLab/SigProfilerExtractor) (Islam et al., 2022), and activities of each DNA mutational signature were attributed along with the confidence intervals using the MSA tool (https://gitlab.com/s.senkin/MSA) (Senkin, 2021). For the identification of cancer driver mutations, dNdS approach restricting to a panel of known cancer genes (Martincorena et al., 2017) followed by a consensus annotation of candidate driver mutations using Cancer Gene Census (https://cancer.sanger.ac.uk/censu) and Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org) tools were used. Downsampling of WGS sequenced samples to whole-exome was performed using SigProfilerMatrixGenerator by applying the 'exome' option, which downsamples mutational matrices to the exome regions of the genome to explore the relationship between latent factors and DNA mutational signatures in TCGA cohorts.

Transcriptome data

Processed transcriptome data of normal adjacent kidney and ccRCC tumour samples used for both discovery and IARC validation series were derived from previous studies (Laskar et al., 2021). Briefly, gene expression analysis was performed using Illumina HumanHT-12 v4 expression BeadChips (Ilumina, San Diego, CA, USA), restricting to samples with RNA integrity > 5. Raw probe intensities with signal-to-noise ratio > 9.5 were further processed via variance-stabilizing transformation and quantile normalization using lumi package in R (v2.5). The probe sequences were aligned to the hg19 human reference genome. For downstream analyses, only probes with detection rate (quality metric) > 5% in both paired normal and tumour samples were considered. Whenever multiple probes were mapped to a single gene, the probe with the highest detection rate was considered.

DNA methylation profiling

The DNA methylation analyses of new 121 ccRCC tumour samples were sequenced using Infinium Methylation EPIC (850K) Bead-Chip (Ilumina, San Diego, CA, USA) for the current study, as recently described elsewhere (Talukdar et al., 2021). Briefly, the DNA of samples underwent pre-processing steps as follows: bisulfite-conversion, whole-genome amplification, fragmentation, and hybridization with complementary probe sequences on Bead-Chip. The images of the arrays were captured by iScan system scanner (Ilumina, San Diego, CA, USA) and probe intensities were obtained by GenomeStudio Software (Ilumina, San Diego, CA, USA). The processing steps of probes were performed using the implemented functions in methylkey R package (https://github.com/IARCbioinfo/methylkey). DNA methylation status was estimated by the β value - signal from the methylated probe divided by the overall signal intensity. The methylation levels of CpG sites were described as a continuous β value range between 0 (no methylation) and 1 (full methylation). Sample-specific quality controls were performed interrogating DNA methylation predicted sex and sample clustering based on the overall signal intensity median of the methylated and unmethylated channels. One low-quality sample was excluded from further analyses. β values were normalized using functional normalization (FunNorm), and probes with missing rate > 20% or

flagged as ‘crossReactive’, ‘SNP’, and ‘XY’ were removed. SVA package was used to remove potential batch effects (v3.35.2). For regression purposes, β-values were converted to M-values. DNA methylation sites were annotated with the information provided by Illumina and the University of California Santa Cruz (UCSC) database (hg19).

Multi-Omics Factor Analysis (MOFA) and inference of latent factors

MOFA was performed to integrate the different omic layers (DNA methylome, transcriptome, and somatic mutational profile from WGS) of overlapping ccRCC tumour samples (R package MOFA2, v1.10.0). DNA methylation data were missing for 31 out of the 151 tumour samples (discovery set) and they were imputed by MOFA, as previously described (Argelaguet et al., 2018). As recommended by the software due to computational limitations, we selected the 5,000 most variable features across samples for DNA methylome (CpG sites) and transcriptome data (gene expression levels) as continuous variables (Supplementary Table 3). The somatic mutational profile derived from previous WGS (Senkin et al., 2013) was summarized as binary variables (presence or absence) of both ccRCC driver genes and DNA mutational signatures in different mutation contexts (SBS96, DBS72, ID83, CN48, and SV32), restricting to variables with more than five events in the discovery set. MOFA generated ten independent continuous latent factors that explained important sources of variance across omic layers of ccRCC tumours (Supplementary Fig. 2). Of note, we further analyzed latent factors 1 to 6 since no additional associations between latent factors 7 to 10 and epidemiological data were observed in the discovery set (Supplementary Table 13). Elbow statistic additionally supported the selection of the first six latent factors (Supplementary Fig. 2D).

To infer an approximation of the latent factors in the validation sets, we selected the most informative features within the overrepresented omic layer for each latent factor in the discovery set (Supplementary Table 1). Since more than 90% of the variance in DNA methylome and transcriptome layers were explained by latent factors, we used these omic layers to generate latent factor signatures. The features correlated with each latent factor (FDR < 0.05 for 30,000 tests) were selected as variables for the LASSO regression models. The LASSO tune parameters were chosen by resampling the discovery set using the tidymodels metapackage in R (v1.0.0; wrapper of glmnet) that by default sets to ten the minimum number of features when the most stringent penalty is applied (Table S1). The ten features selected by LASSO models were used to generate signatures that represented an approximation of each latent factor by adding up the scaled values of the normalized m-values (DNA methylation) or log2-transcripts per million (transcriptome) of each feature multiplied by the respective LASSO regression coefficients. These signatures were then used to infer the latent factors across TCGA cohorts, IARC normal, and tumour ccRCC samples (validation sets). Of note, the gene expression signatures for factors 2–5, calculated initially using transcriptome data (microarray), could also be applied to voom-transformed RNA-sequencing data (Law et al., 2014).

Calculating additional molecular variables

Global DNA methylation levels were estimated using the mean m-values of locus-specific repetitive elements (*LINE1* and *Alu*) across chromosomes using REMP package in R (v 1.24.0), since these repetitive elements were reported to be more accurate in estimating global DNA methylation levels than averaging the methylation levels of all CpG sites from 850K/450k Ilumina arrays (Lisanti et al., 2013; Zheng et al., 2017). DNA methylation-based clocks were calculated using methylclock (v1.6.0) and dnaMethyAge (v0.1.0) packages in R. The inference of immune cells from tumour microenvironment of bulk ccRCC tumour samples was performed using gene expression signatures based on the list of genes associated with ccRCC tumour cells identified by scRNA sequencing data (Li et al., 2022). We restricted our analyses to the cell signatures in which more than 75% of genes identified in the original study were also present in our study after quality control. For simplicity, we additionally pruned the gene signatures by original functional annotations, retaining independent signatures at $r < 0.80$ within ccRCC tumours. The cell signatures were calculated by adding up the scaled gene expression value of each gene belonging to a signature by sample. The DNA methylation signature for tobacco smoking status (epiTob) was calculated by adding up the methylation levels of the 5 CpG sites (cg05575921, cg26703534, cg23480021, cg08118908, cg00336149) that were previously associated with self-reported smoking status (Chamberlain et al., 2022). Molecular variables from TCGA cohorts were also included in the current study, such as the homologous recombination DNA repair deficiency score (Thorsson et al., 2018), the copy number alteration fraction of the genome (Knijnenburg et al., 2018), and telomere length ratio based on WGS data (Barthel et al., 2017).

Declarations

Ethics approval and consent to participate

Informed consent was obtained for all participants included in the discovery and validation sets. Ethical approvals were obtained from Local and Federal Research Ethics Committees, and from the IARC Ethics Committee. For the TCGA datasets, also used as validation set, the enrolling, collection, clinical and genomic data processing and distributions are subject to 45-CFR-46 (the “Common Rule”) governing protection of human research subjects. Under the revised TCGA consent policy, re-consent of still-living participants is no longer a program-imposed requirement. The Project Team described the best practices for informed consent for participating in TCGA in this memo (http://cancergenome.nih.gov/abouttcga/policies/informedconsent).

Consent for publication

Not applicable.

Availability of data and materials

WGS data and patient metadata are stored at EGA repository (study name: EGAS00001003542). DNA mutational signatures and cancer driver genes information are available in the supplementary tables of the original study (Senkin et al., 2023). Transcriptome data (microarray) of both normal and kidney

cancer samples are available on NCBI/GEO database (GSE167093). DNA methylation data (850k EPIC array) from 120 ccRCC tumour samples are available at xx repository (study name: ???). TCGA data were downloaded by TCGAbiolinks package in R from cBioPortal for cancer genomics' website (https://www.cbioportal.org/). Molecular metrics of genome instability in TCGA dataset are available in the respective supplementary tables of the previously published TCGA papers. R code used in the current work can be found on the project webpage here (https://github.com/ricardocortezcardoso/multi_omic_code).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded as part of the Mutographs team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (C98/A24032). Funding for gene expression data was provided by the US National Institutes of Health (NIH), National Cancer Institute (U01CA155309). Institut National du Cancer funded the DNA methylation data (GeniLuc2017-1-TABAC-03-CIRC-1-[TABAC17-022]).

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Authors' contributions

R.C.C.P. in conceptualization, formal analyses, methodology, wrote and review manuscript; A.S.O. in MOFA and global methylation analyses, and review manuscript; S.S. in DNA mutational signature analysis and review manuscript; H.L.P., M.F., Z.H., N.A. in interpretation of results and review manuscript; I.H., A.H., S.S., S.O., B.Ś., J.L., D.Z., A.M. in patient recruitment and sample coordination; A.C., V.C., C.C. in laboratory analysis; G.S. in resources for the 120 ccRCC tumour samples included in the DNA methylome in the discovery set; J.A., P.B., K.S. reviewed the manuscript; J.D.M. in supervision, conceptualization, project administration, funding acquisition, wrote and review manuscript.

Acknowledgements

We would like to acknowledge Mutographs consortium, TCGA Research Network (https://www.cancer.gov/tcga), K2 and CCE studies, and the contribution of specimen donors and research groups involved in these resources.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A et al (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71(3):209–249. 10.3322/caac.21660
2. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M Renal cell car, Linehan WM, Lerman MI, Zbar B et al (1995) Identification of the von Hippel-Lindau (VHL) Gene: Its Role in Renal Cancer. JAMA. ; 273(7):564–570. 10.1001/jama.1995.03520310062031
3. Huang J, Leung DK, Chan EO, Lok V, Leung S, Wong I et al (2022) A Global Trend Analysis of Kidney Cancer Incidence and Mortality and Their Associations with Smoking, Alcohol Consumption, and Metabolic Syndrome. Eur Urol Focus 8(1):200–209. 10.1016/j.euf.2020.12.020
4. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S et al (2015) Clock-like mutational processes in human somatic cells. Nat Genet 47(12):1402–1407. 10.1038/ng.3441
5. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y et al (2020) The repertoire of mutational signatures in human cancer. Nature 578(7793):94–101. 10.1038/s41586-020-1943-3
6. Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzàlez-Porta M, Wozniak MB et al (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat Commun 5:5135. 10.1038/ncomms6135
7. Senkin S, Moody S, Díaz-Gay M, Abedi-Ardekani B, Cattiaux T, Ferreiro-Iglesias A et al (2023) Geographic variation of mutagenic exposures in kidney cancer genomes. Preprint at https://www.medrxiv.org/content/10.1101/2023.06.20.23291538v2
8. Herceg Z, Ghantous A, Wild CP, Sklias A, Casati L, Duthie SJ et al (2018) Roadmap for investigating epigenome deregulation and environmental origins of cancer. Int J Cancer 142(5):874–882. 10.1002/ijc.31014
9. Linehan WM, Lerman MI, Zbar B (1995) Identification of the von Hippel-Lindau (VHL) gene. Its role in renal cancer. JAMA 273(7):564–570 PMID:7837390
10. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D et al (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366(10):883–892. 10.1056/NEJMoa1113205
11. Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499(7456):43–49. 10.1038/nature12222
12. Hakimi AA, Ostrovnaya I, Reva B, Schultz N, Chen YB, Gonen M et al (2013) Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. Clin Cancer Res 19(12):3259–3267. 10.1158/1078-0432.CCR-12-3886
13. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S et al (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49(2):359–367. 10.1016/j.molcel.2012.10.016

14. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115. 10.1186/gb-2013-14-10-r115
15. Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Marttila Palli D et al (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet 24(8):2349–2359. 10.1093/hmg/ddu751
16. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR et al (2016) Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet 9(5):436–447. 10.1161/CIRCGENETICS.116.001506
17. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G et al (2016) Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol 17(1):231. 10.1186/s13059-016-1092-z
18. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD et al (2016) Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol 17(1):205. 10.1186/s13059-016-1064-3
19. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E et al (2018) The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep. ; 23(1):313–326.e5. 10.1016/j.celrep.2018.03.075. Nat Rev Dis Primers. 2017; 3:17009. DOI:10.1038/nrdp.2017.9
20. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S et al (2018) An epigenetic biomarker of aging for lifespan and healthspan. Aging 10(4):573–591. 10.18632/aging.101414
21. Motzer RJ, Banchereau R, Hamidi H, Powles T, McDermott D, Atkins MB et al (2020) Molecular Subsets in Renal Cancer Determine Outcome to Checkpoint and Angiogenesis Blockade. Cancer Cell 38(6):803–817e4. 10.1016/j.ccell.2020.10.011
22. Halaburkova A, Cahais V, Novoloaca A, Araujo MGDS, Khoueiry R, Ghantous A et al (2020) Pan-cancer multi-omics analysis and orthogonal experimental assessment of epigenetic driver genes. Genome Res 30(10):1517–1532. 10.1101/gr.268292.120
23. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X et al (2020) Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. Elife 9:e54870. 10.7554/eLife.54870
24. Chamberlain JD, Nusslé S, Chapatte L, Kinnaer C, Petrovic D, Pradervand S et al (2022) Blood DNA methylation signatures of lifestyle exposures: tobacco and alcohol consumption. Clin Epigenetics 14(1):155. 10.1186/s13148-022-01376-7
25. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC et al (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 14(6):e8124. 10.15252/msb.20178124
26. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E et al (2021) Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun 12(1):124. 10.1038/s41467-020-20430-7

27. Witte T, Plass C, Gerhauser C (2014) Pan-cancer patterns of DNA methylation. Genome Med 6(8):66. 10.1186/s13073-014-0066-6
28. Marttila S, Kananen L, Häyrynen S, Jylhävä J, Nevalainen T, Hervonen A et al (2015) Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics 16(1):179. 10.1186/s12864-015-1381-z
29. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT et al (2019) DNA methylation aging clocks: challenges and recommendations. Genome Biol 20(1):249. 10.1186/s13059-019-1824-y
30. Rutledge J, Oh H, Wyss-Coray T (2022) Measuring biological age using omics data. Nat Rev Genet 23(12):715–727. 10.1038/s41576-022-00511-7
31. Teschendorff AE (2020) A comparison of epigenetic mitotic-like clocks for cancer risk prediction. Genome Med 12(1):56. 10.1186/s13073-020-00752-3
32. Lee H, Dai F, Zhuang L, Xiao ZD, Kim J, Zhang Y et al (2016) BAF180 regulates cellular senescence and hematopoietic stem cell homeostasis through p21. Oncotarget 7(15):19134–19146. 10.18632/oncotarget.8102
33. Dominguez D, Tsai YH, Gomez N, Jha DK, Davis I, Wang Z (2016) A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. Cell Res 26(8):946–962. 10.1038/cr.2016.84
34. Cai W, Su L, Liao L, Liu ZZ, Langbein L, Dulaimi E et al (2019) PBRM1 acts as a p53 lysine-acetylation reader to suppress renal tumor growth. Nat Commun 10(1):5800. 10.1038/s41467-019-13608-1
35. Jen J, Wang YC (2016) Zinc finger proteins in cancer progression. J Biomed Sci 23(1):53. 10.1186/s12929-016-0269-9
36. Vidal VP, Jian-Motamedi F, Rekima S, Gregoire EP, Szenker-Ravi E, Leushacke M et al (2020) R-spondin signalling is essential for the maintenance and differentiation of mouse nephron progenitors. Elife 9:e53895. 10.7554/eLife.53895
37. Li R, Ferdinand JR, Loudon KW, Bowyer GS, Laidlaw S, Muyas F et al (2022) Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. Cancer Cell 40(12):1583–1599e10. 10.1016/j.ccell.2022.11.001
38. Kalluri R, Weinberg RA (2009) The basics of epithelial-mesenchymal transition. J Clin Invest 119(6):1420–1428. 10.1172/JCI39104
39. Wang W, He J, Lu H, Kong Q, Lin S (2020) KRT8 and KRT19, associated with EMT, are hypomethylated and overexpressed in lung adenocarcinoma and link to unfavorable prognosis. Biosci Rep 40(7):BSR20193468. 10.1042/BSR20193468
40. Guo H, Jiang S, Sun H, Shi B, Li Y, Zhou N et al (2022) Identification of IL20RB as a Novel Prognostic and Therapeutic Biomarker in Clear Cell Renal Cell Carcinoma. Dis Markers. ; 2022:9443407. 10.1155/2022/9443407
41. Wu LL, Yuan SF, Lin QY, Chen GM, Zhang W, Zheng WE et al (2023) Construction and validation of risk model of EMT-related prognostic genes for kidney renal clear cell carcinoma. J Gene Med

25(11):e3549. 10.1002/jgm.3549

42. Plusquin M, Guida F, Polidoro S, Vermeulen R, Raaschou-Nielsen O, Campanella G et al (2017) DNA methylation and exposure to ambient air pollution in two prospective cohorts. Environ Int 108:127–136. 10.1016/j.envint.2017.08.006
43. Svoboda LK, Neier K, Wang K, Cavalcante RG, Rygiel CA, Tsai Z et al (2021) Tissue and sex-specific programming of DNA methylation by perinatal lead exposure: implications for environmental epigenetics studies. Epigenetics 16(10):1102–1122. 10.1080/15592294.2020.1841872
44. Miller DP, De Vivo I, Neuberg D, Wain JC, Lynch TJ, Su L et al (2003) Association between self-reported environmental tobacco smoke exposure and lung cancer: modification by GSTP1 polymorphism. Int J Cancer 104(6):758–763. 10.1002/ijc.10989
45. Simic T, Savic-Radojevic A, Pljesa-Ercegovac M, Matic M, Mimic-Oka J (2009) Glutathione S-transferases in kidney and urinary bladder tumors. Nat Rev Urol 6(5):281–289. 10.1038/nrurol.2009.49
46. Sawers L, Ferguson MJ, Ihrig BR, Young HC, Chakravarty P, Wolf CR et al (2014) Glutathione S-transferase P1 (GSTP1) directly influences platinum drug chemosensitivity in ovarian tumour cell lines. Br J Cancer 111(6):1150–1158. 10.1038/bjc.2014.386
47. Shin YJ, Kim KA, Kim ES, Kim JH, Kim HS, Ha M et al (2018) Identification of aldo-keto reductase (AKR7A1) and glutathione S-transferase pi (GSTP1) as novel renal damage biomarkers following exposure to mercury. Hum Exp Toxicol 37(10):1025–1036. 10.1177/0960327117751234
48. Su PF, Lee TC, Lin PJ, Lee PH, Jeng YM, Chen CH et al (2007) Differential DNA methylation associated with hepatitis B virus infection in hepatocellular carcinoma. Int J Cancer 121(6):1257–1264. 10.1002/ijc.22849
49. Rønneberg JA, Tost J, Solvang HK, Alnaes GI, Johansen FE, Brendeford EM et al (2008) GSTP1 promoter haplotypes affect DNA methylation levels and promoter activity in breast carcinomas. Cancer Res 68(14):5562–5571. 10.1158/0008-5472.CAN-07-5828
50. Cui J, Li G, Yin J, Li L, Tan Y, Wei H et al (2020) GSTP1 and cancer: Expression, methylation, polymorphisms and signaling (Review). Int J Oncol 56(4):867–878. 10.3892/ijo.2020.4979
51. Perelli L, Carbone F, Zhang L, Huang JK, Le C, Khan H et al (2023) Interferon signaling promotes tolerance to chromosomal instability during metastatic evolution in renal cancer. Nat Cancer 4(7):984–1000. 10.1038/s43018-023-00584-1
52. Scholz H, Boivin FJ, Schmidt-Ott KM, Bachmann S, Eckardt KU, Scholl UI et al (2021) Kidney physiology and susceptibility to acute kidney injury: implications for renoprotection. Nat Rev Nephrol 17(5):335–349. 10.1038/s41581-021-00394-7
53. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR et al (2018) Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 361(6402):594–599. 10.1126/science.aat1699
54. Laskar RS, Li P, Ecsedi S, Abedi-Ardekani B, Durand G, Robinot N et al (2021) Sexual dimorphism in cancer: insights from transcriptional signatures in kidney tissue and renal cell carcinoma. Hum Mol

Genet 30(5):343–355. 10.1093/hmg/ddab031

55. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D et al (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 44(8):e71. 10.1093/nar/gkv1507
56. Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN et al (2022) Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genom 2(11):None. 10.1016/j.xgen.2022.100179
57. Senkin S (2021) MSA: reproducible mutational signature attribution with confidence based on simulations. BMC Bioinformatics 22(1):540. 10.1186/s12859-021-04450-8
58. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P et al (2017) Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171(5):1029–1041e21. 10.1016/j.cell.2017.09.042
59. Talukdar FR, Soares Lima SC, Khoueiry R, Laskar RS, Cuenin C, Sorroche BP et al (2021) Genome-Wide DNA Methylation Profiling of Esophageal Squamous Cell Carcinoma from Global High-Incidence Regions Identifies Crucial Genes and Potential Cancer Markers. Cancer Res 81(10):2612–2624. 10.1158/0008-5472.CAN-20-3445
60. Law CW, Chen Y, Shi W, Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15(2):R29. 10.1186/gb-2014-15-2-r29
61. Lisanti S, Omar WA, Tomaszewski B, De Prins S, Jacobs G, Koppen G et al (2013) Comparison of methods for quantification of global DNA methylation in human cells and tissues. PLoS ONE 8(11):e79044. 10.1371/journal.pone.0079044
62. Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W et al (2017) Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res 45(15):8697–8711. 10.1093/nar/gkx587
63. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH et al (2018) The Immune Landscape of Cancer. Immunity 48(4):812–830. 10.1016/j.immuni.2018.03.023
64. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD et al (2018) Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep 23(1):239–254. 10.1016/j.celrep.2018.03.076
65. Barthel FP, Wei W, Tang M, Martinez-Ledesma E, Hu X, Amin SB et al (2017) Systematic analysis of telomere length and somatic alterations in 31 cancer types. Nat Genet 49(3):349–357. 10.1038/ng.3781

Tables

Tables 1 and 2 are available in the Supplementary Files section.

Supplementary Tables

Supplementary Tables are not available with this version

Figures

Figure 1

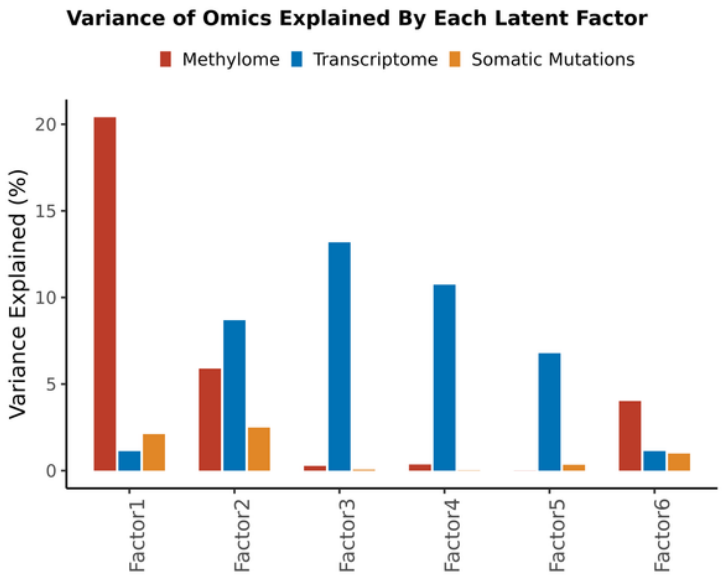

Figure 1

Overview of latent factors 1 to 6. Percentage of variance in each omic layer across tumour samples from the discovery set (all molecular variables included in the Multi-Omics Factor Analysis) explained by latent factors. Methylome (DNA methylation) in red, transcriptome (microarray) in blue, and somatic mutations (cancer driver mutations and DNA mutational signatures derived from whole-genome sequencing data) in yellow.

Figure 2

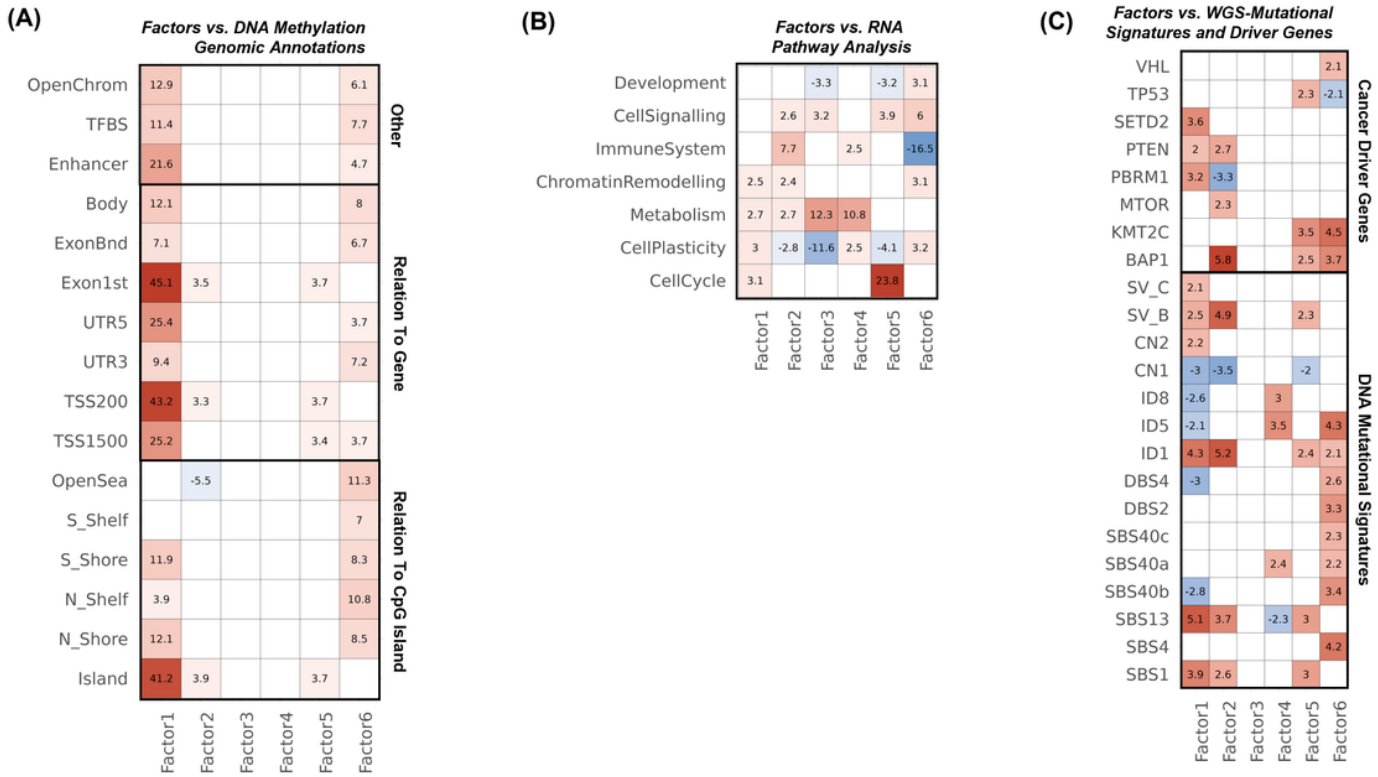

Figure 2

Associations between latent factors and molecular features of ccRCC tumours in the discovery set. Heatmaps showing the Z-scores (beta divided by standard error) of linear regression analyses between latent factors (outcome) and molecular features related to the three omic layers included in the Multi-Omics Factor Analysis (MOFA) for DNA methylome and somatic profile. (A) For the DNA methylome layer (N=120), the average beta methylation levels of the 5,000 MOFA CpG sites by genomic annotation related to CpG island (island, shores, shelves, and open sea), gene (proximal/TSS200 and distal/TSS1500 promoters, UTRs, exons, and body), and other regulatory regions (open chromatin, transcription factor binding site/TFBS, and enhancer) were used as predictors. (B) For the transcriptome layer, pathway analysis was performed for the top 500 gene expression levels correlated with latent factors. The top 10 biological pathways by latent factor (FDR<0.05) were then manually annotated into functional groups (Development, CellSignalling, ImmuneSystem, ChromatinRemodelling, Metabolism, CellPlasticity, and CellCycle), and normalized enrichment scores were represented as sum up by functional weighted. (C) The somatic profile based on whole genome sequence (WGS, N=151) was represented by ccRCC driver mutations (binary; presence or absence) and DNA mutational signatures

(continuous). Regression models included age at diagnosis, sex, and country of origin as covariates. Values represented as shades of red (Z-scores>0) and blue (Z-score<0). The associations that passed multiple-testing correction (FDR<0.05) within each group of variables were represented. Tobacco-related (SBS4, DBS2), clock-like (SBS1, ID1), *APOBEC* (SBS13), copy number (CN) and structural variation (SV) DNA mutational signatures.

Figure 3

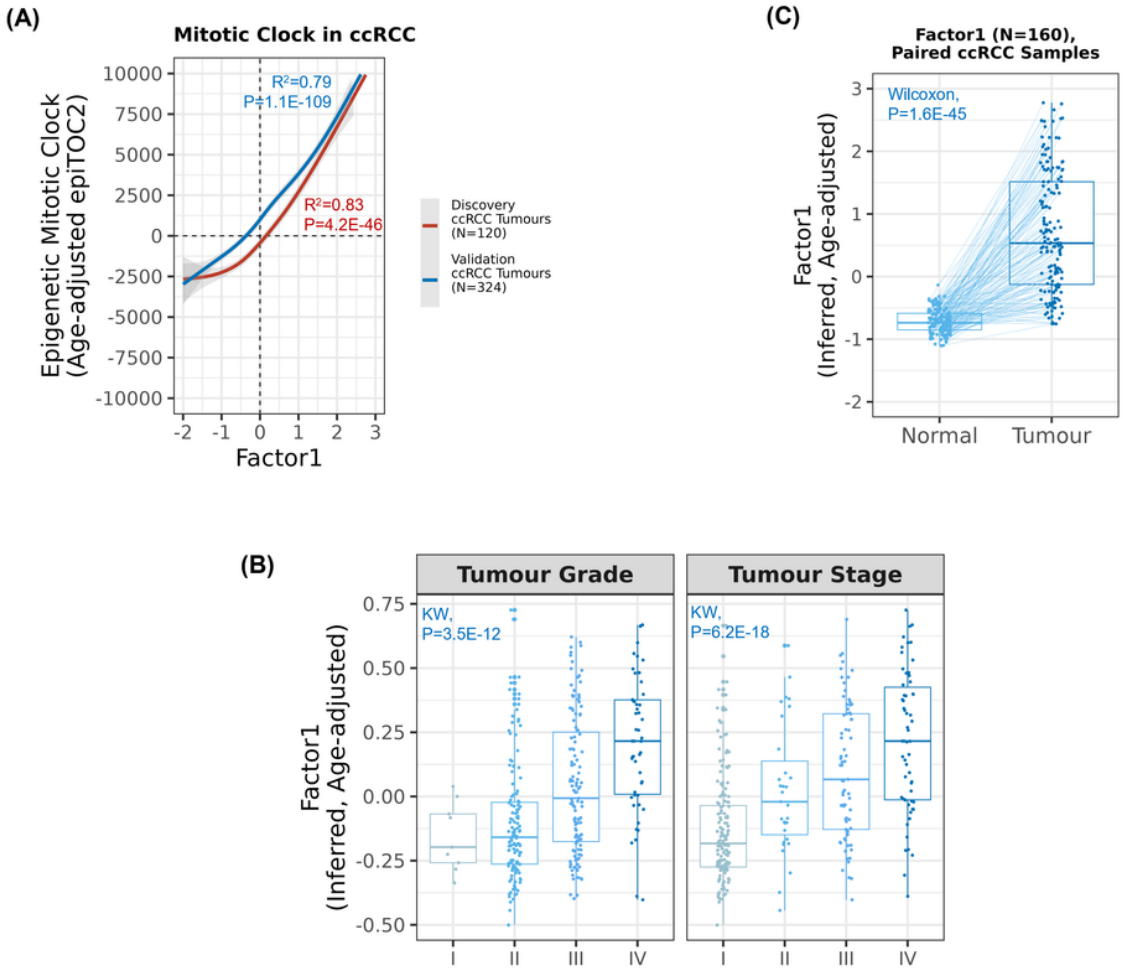

Figure 3

Relationship between latent factor 1, the mitotic-like epigenetic clock epiTOC2, and prognosis. Analyses were performed using the residuals of the cellular mitotic age epigenetic clock epiTOC2 after adjusting by chronological age. (A) Univariate linear regression between latent factor 1 and the age-adjusted epiTOC2 (Discovery: N=120, Validation/TCGA-KIRC: N=324) in ccRCC tumours. (B) Age-adjusted residuals of latent factor 1 across different ccRCC tumour stages (N=322) and grades (N=320) from TCGA validation set (N=323). Statistical comparison between multiple means was performed using Kruskal-Wallis's test. (C) Comparison of paired normal adjacent kidney tissues (light blue) and ccRCC tumours (dark blue) for age-adjusted epiTOC2 (TCGA-KIRC: N=160 pairs). Lines connect matched samples. P-values from Wilcoxon

signed-rank test that calculate differences of means between matched samples. P-values < 0.05 were considered statically significant. Observed latent factor 1 used for the regression models in the discovery set while the latent factor 1 signature was used for the analyses in the validation set.

Figure 4

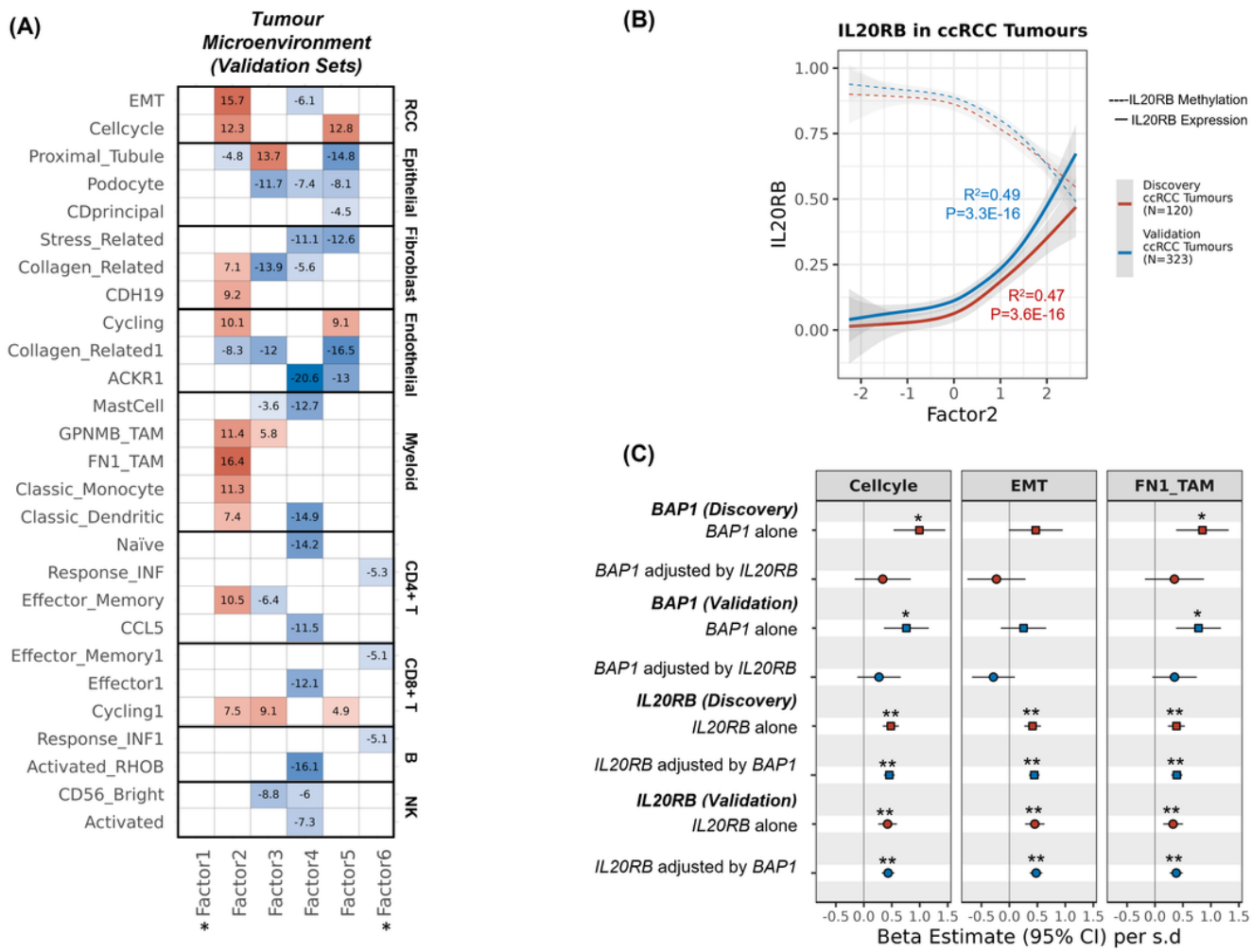

Figure 4

Association between inferred tumour microenvironment cells and factors in ccRCC. (A) Heatmap plot is showing the statistically significant associations between latent factors 1 to 6 and the 27 representative tumour microenvironment signatures in ccRCC tumours. The association estimates were derived from the analyses in the validation sets (IARC ccRCC series: N=462 for latent factors 2-5; TCGA-KIRC: N=323 for latent factors 1 and 6) after adjustments by covariates (sex, age at diagnosis, and country of origin whenever possible), restricting to the associations that passed multiple-testing correction (false discover rate < 0.05, 162 tests) in both discovery and validation ccRCC sets. The associations were represented as Z-scores (beta divided by standard error; Z-scores>0 in shades of red; Z-score<0 in shades of blue). The ccRCC tumour microenvironment signatures (CD4+ T, B, NK, endothelial, myeloid, CD8+ T, epithelial and fibroblast cells) and kidney cancer meta programs/RCC (epithelial-to-mesenchymal transition/EMT and

cell cycle) were derived from single-cell RNA sequencing data published by Li et al., (2022). (B) Univariate regression lines representing the relationship between rescaled values (0 to 1) of the average DNA methylation (dashed line) of the CpG sites (cg06392589, cg12293186) and the RNA levels (solid line) of *IL20RB* gene, and latent factor 2 in the discovery and validation ccRCC tumour datasets. R^2 values mean the variance in latent factor 2 explained by DNA methylation and expression of *IL20RB* across samples. (C) Forest plot representing the multivariable regression analyses between key tumour microenvironment signatures related to latent factor 2 (outcome: cell cycle, epithelial-mesenchymal transition/EMT, and fibronectin 1 positive tumour-associated macrophages/FN1_TAM) and the presence of *BAP1* cancer driver mutations and/or *IL20RB* expression levels in both discovery (red; IARC ccRCC serie; N=120) and validation (blue; TCGA-KIRC; N=269) ccRCC tumour sets. Beta estimates were represented as an increase in the effect of the selected features (*BAP1* alone or adjusted by *IL20RB*, and vice-versa) per 1 unit of standard deviation increase in ccRCC tumour microenvironment signatures. * p<0.01; ** p<0.001.

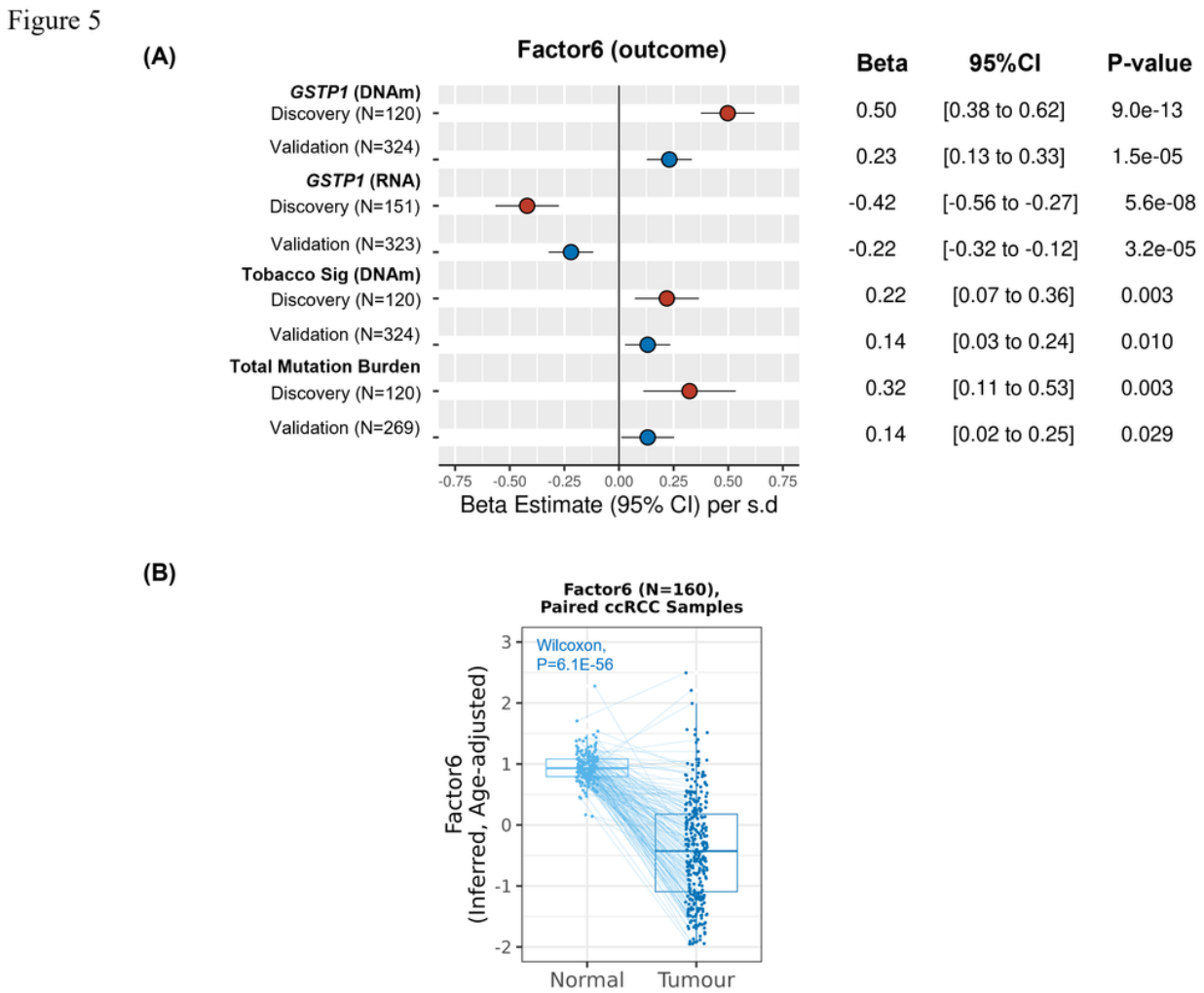

Figure 5

Associations between latent factor 6 and molecular features related to exogenous exposures in ccRCC tumours. Forest plot showing the results of multivariable regression analyses of latent factor 6 (outcome) and *GSTP1*methylation (Average m-values of CpG sites annotated to *GSTP1*, DNAm, continuous) and gene expression levels (RNA, continuous), and DNA methylation signature of tobacco smoking trained to predict self-reported tobacco smoking status (5 CpG sites, continuous, epiTob), and total mutation burden (whole-genome for discovery and whole-exome for validation) in both discovery (N=120 for DNA methylation and 151 for gene expression data) and validation (TCGA-KIRC; N=324) sets. Covariates used in the regression models were sex, age at diagnosis, and country of origin (whenever possible). For analyses of total mutation burden, ccRCC cases from Romania (N=31) were excluded from the discovery set. Beta estimates were represented as an increase in the effect of the selected features per 1 unit of standard deviation increase in latent factor 6. Blue dots when discovery and red dots when validation ccRCC tumour cohorts. P values < 0.05 were considered statically significant. Observed latent factors used for the regression models in the discovery set while signatures for the same latent factors were used for the analyses in the validation set.

Figure 6

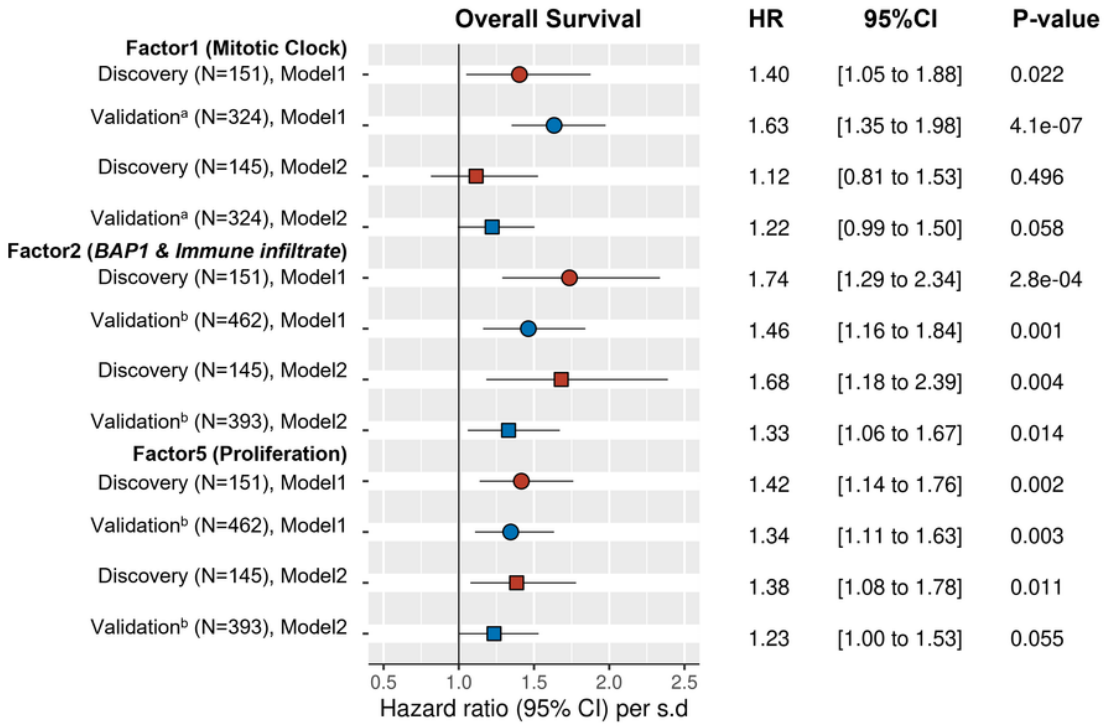

Figure 6

Molecular components associated with prognosis of ccRCC patients. Cox proportional-hazards models for assessing overall survival of ccRCC patients by latent factors 1 (mitotic-like epigenetic clock epiTOC2), 2 (related to *BAP1* cancer driver mutations and pro-inflammatory immune cells), and 5 (cell cycle) adjusting for age at diagnosis, sex (model 1; circle shape), and additionally by tumour stage and grade (I+II *vs.* III+IV; model 2; square shape) in the discovery (red) and validation (blue) datasets. Hazard ratios (HR) represented as an increase in relative mortality risk per 1 unit of standard deviation increase in factors. Two different validation sets were used according to the factors, TCGA-KIRC (a; N=324) for latent factor 1 and IARC series sets (b; N=462) for latent factors 2 and 5. Observed latent factors used for the regression models in the discovery set while signatures for the same latent factors were used for the analyses in the validation sets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Table2.xlsx
- Supp.FiguresLegends.docx
- SupplementaryFigures.pdf