


















# Epigenomic charting and functional annotation of risk loci in renal cell carcinoma

Received: 22 October 2021

Accepted: 4 January 2023

Published online: 21 January 2023

 Check for updates

Amin H. Nassar <sup>1,2,3,12</sup>, Sarah Abou Alaiwi<sup>2,3,4,12</sup>, Sylvan C. Baca<sup>2,3,12</sup>, Elio Adib<sup>2,3</sup>, Rosario I. Corona <sup>5,6,7</sup>, Ji-Heui Seo<sup>3</sup>, Marcos A. S. Fonseca<sup>5,6</sup>, Sandor Spisak <sup>3,4,8</sup>, Talal El Zarif<sup>2,3</sup>, Viktoria Tisza<sup>3,4,8</sup>, David A. Braun <sup>1,3,8</sup>, Heng Du <sup>2</sup>, Monica He<sup>3</sup>, Abdallah Flaifel<sup>9</sup>, Michel Alchoueiry <sup>2</sup>, Thomas Denize<sup>9</sup>, Sayed G. Matar<sup>9</sup>, Andres Acosta<sup>9</sup>, Sachet Shukla <sup>3,10</sup>, Yue Hou<sup>3,10</sup>, John Steinharter<sup>3</sup>, Gabrielle Bouchard<sup>3</sup>, Jacob E. Berchuck <sup>2,3</sup>, Edward O'Connor<sup>3</sup>, Connor Bell <sup>3</sup>, Pier Vitale Nuzzo <sup>3</sup>, Gwo-Shu Mary Lee<sup>3</sup>, Sabina Signoretti<sup>9</sup>, Michelle S. Hirsch <sup>9</sup>, Mark Pomerantz<sup>3</sup>, Elizabeth Henske <sup>2,3</sup>, Alexander Gusev <sup>3,11</sup>, Kate Lawrenson <sup>5,6,7</sup>, Toni K. Choueiri <sup>2,3</sup> ✉, David J. Kwiatkowski <sup>2,3</sup> ✉ & Matthew L. Freedman <sup>3,4,8</sup> ✉

While the mutational and transcriptional landscapes of renal cell carcinoma (RCC) are well-known, the epigenome is poorly understood. We characterize the epigenome of clear cell (ccRCC), papillary (pRCC), and chromophobe RCC (chRCC) by using ChIP-seq, ATAC-Seq, RNA-seq, and SNP arrays. We integrate 153 individual data sets from 42 patients and nominate 50 histology-specific master transcription factors (MTF) to define RCC histologic subtypes, including *EPAS1* and *ETS-1* in ccRCC, *HNF1B* in pRCC, and *FOXI1* in chRCC. We confirm histology-specific MTFs via immunohistochemistry including a ccRCC-specific TF, BHLHE41. FOXI1 overexpression with knock-down of *EPAS1* in the 786-O ccRCC cell line induces transcriptional upregulation of chRCC-specific genes, *TFCP2L1*, *ATP6VOD2*, *KIT*, and *INSRR*, implicating FOXI1 as a MTF for chRCC. Integrating RCC GWAS risk SNPs with H3K27ac ChIP-seq and ATAC-seq data reveals that risk-variants are significantly enriched in allelically-imbalanced peaks. This epigenomic atlas in primary human samples provides a resource for future investigation.

In 2021, an estimated 76,080 adults in the United States will be diagnosed with kidney cancer, and an estimated 13,780 deaths from this disease will occur<sup>1</sup>. Renal cell carcinomas (RCCs), the most common family of kidney tumors and one of the top ten most common cancers in the US, are further stratified into several histologic subtypes. The most common subtype is clear cell RCC (ccRCC), accounting for ~75% of cases; papillary RCC (pRCC) and chromophobe RCC (chRCC) account for ~15% and 5% of cases, respectively<sup>2</sup>. These subtypes display divergent clinical behavior with regard to prognosis and response to therapeutic agents<sup>3–5</sup>. Large-scale molecular profiling efforts have

characterized the genomic and transcriptomic landscapes of ccRCC<sup>6</sup>, pRCC<sup>7</sup>, and chRCC<sup>8</sup>. These analyses revealed remarkable heterogeneity among these forms of RCC, with each subtype exhibiting distinctive somatic mutations, chromosomal copy number alterations, and gene expression profiles<sup>3</sup>. Notably, the histone modifications and sites of chromatin accessibility driving the transcriptional landscapes of RCC histotypes, and how this relates to kidney cancer heritability, have not been systematically explored.

Enhancers are *cis*-acting DNA regions that bind *trans*-acting proteins to determine cell-type-specific gene expression patterns and

responses to internal and external signals<sup>9,10</sup>. Chromatin immunoprecipitation and sequencing (ChIP-seq) of post-translational histone modifications (e.g., H3K27ac and H3K4me2) have identified millions of enhancers in mammalian genomes with the number of active enhancers in any given cell type estimated to be in the tens of thousands<sup>10,11</sup>. Profiling enhancers has emerged as a powerful tool to characterize critical transcription factors (TFs) driving cellular transcriptional states and to better understand germline risk variants. It has been established that cellular identity and function is determined primarily by a subset of TFs termed “master” TFs<sup>12–16</sup>. These TFs occupy active enhancers in the cell, and preferentially bind within exceptionally large enhancer domains termed “super-enhancers” (SEs) or stretch enhancers, that regulate genes required for establishing cell identity and function<sup>17,18</sup>. Moreover, master TFs participate in interconnected auto-regulatory circuitries or “cliques” that are self-reinforcing, show marked cell selectivity, and function to maintain cell state and/or cell survival<sup>17,19</sup>. In addition, analyses of population-based epigenomes have further revealed that expression quantitative trait loci (eQTLs) SNPs are often also associated with variation in nearby epigenomic features (such as active enhancers marked by H3K27ac) in coordinated regulatory modules<sup>20–25</sup>, motivating the use of epigenetic datasets for better functional characterization of these loci. Many of the prior epigenetic data in RCC were based on cell lines, which diverge from their original tumors and do not represent all histologic subtypes<sup>26</sup>.

Herein, we define the epigenetic architecture and circuitry of RCC across different histologic subtypes in 42 primary human RCC tumors using a combination of genome-wide H3K27ac, H3K4me2, and chromatin accessibility assays, and integrate data from targeted DNA sequencing and bulk RNA sequencing. We delineate distinctive enhancers operative in the different RCC histologies, nominate putative histology-specific master TFs, and prioritize RCC GWAS risk loci for functional validation.

## Results

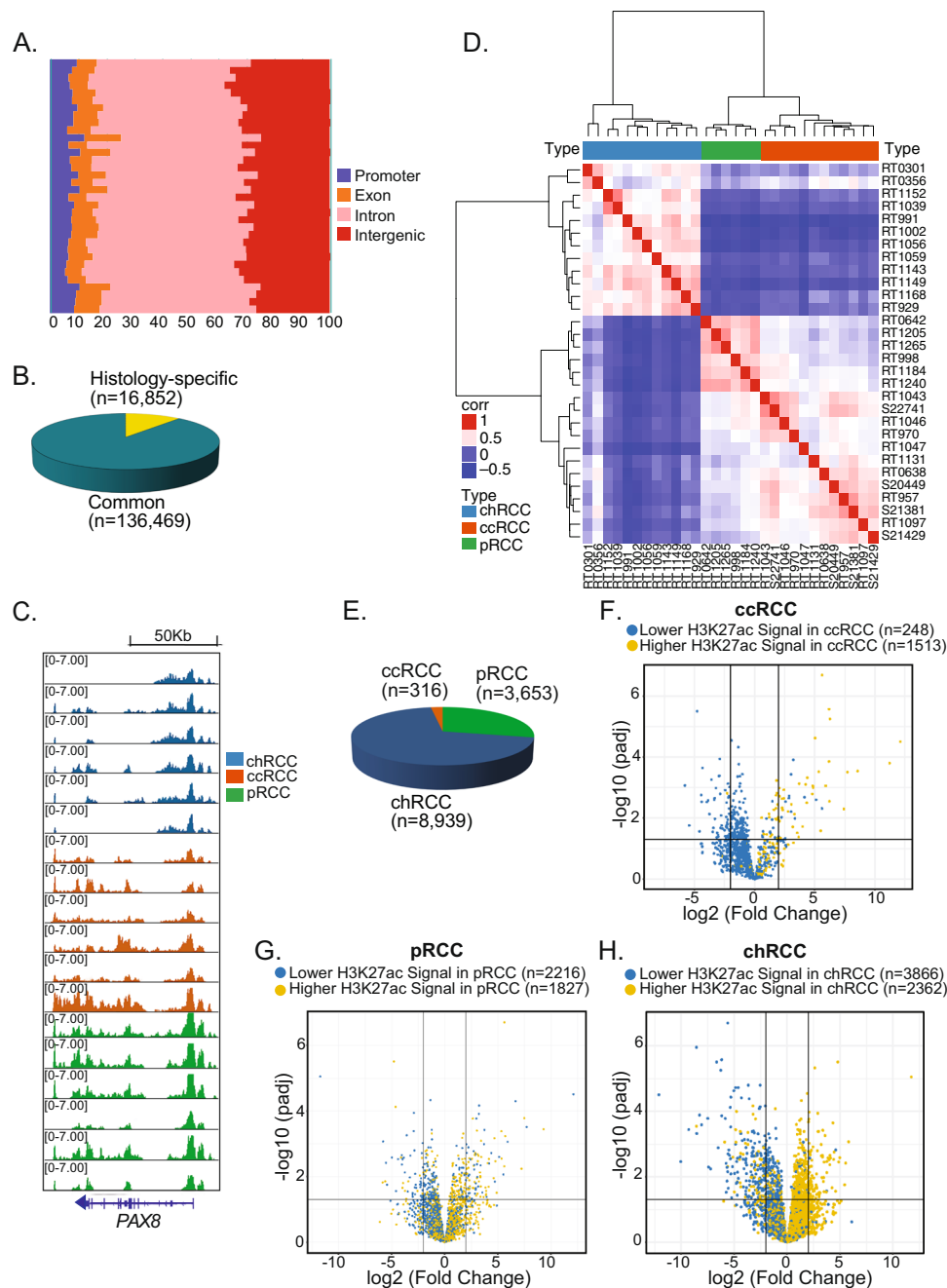
### Mapping the chromatin regulatory landscape of RCC

To examine the cistrome and to identify master TFs across the histologic subtypes of RCC, we performed histone chromatin immunoprecipitation followed by sequencing (ChIP-seq), the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), and RNA sequencing (RNA-seq) on 42 fresh frozen RCC tumor samples (24 ccRCC, 6 pRCC, 12 chRCC). Thirty-eight of 40 (95%) tissues were derived from radical nephrectomies. Of the 6 pRCC tumors, 4 were type I, 1 was type II, and 1 was unknown. Patients were grouped into two cohorts (1 and 2) as defined in Supplementary Data 1, Supplementary Fig. S1, and the Results section. We conducted H3K4me2 ChIP-seq to map both active and poised enhancers<sup>27</sup> and H3K27ac ChIP-seq to identify active promoters and enhancers<sup>28</sup>. ATAC-seq was performed to define the chromatin accessibility landscape, and RNA-seq was performed to capture the transcriptional programs of each RCC subtype. A total of 153 libraries were generated across the different data-types (Supplementary Data 1 and 2 and Supplementary Fig. S1). Using cohort 1 to assess the regulatory landscape across the different RCC histologies, a total of 153,321 promoter-distal (enhancer) H3K27ac ChIP-Seq peaks were identified across all the samples, most of which ( $n = 136,469$ ) were common to two or more histologies (Fig. 1A, B). For example, the *PAX8* locus is marked by H3K27ac in all samples. *PAX8* is a TF involved in early kidney embryogenesis and oncogenesis in RCC<sup>29,30</sup> and is a clinical diagnostic tool to help differentiate RCC from other malignancies<sup>31</sup> (Fig. 1C). Unsupervised hierarchical clustering (Fig. 1D) and principal component analysis (PCA) (Supplementary Fig. S2A) of H3K27ac peaks clearly segregated the three histologic types of RCC, and both analyses demonstrated that the H3K27ac landscape in chRCC was more distinct than either pRCC or ccRCC. Moreover, unsupervised hierarchical clustering of our cohort with an independent cohort of matched normal ( $n = 10$ ) from the KIRC TCGA cohort<sup>32</sup> showed a clean

separation of normal tissue from the different tumor histotypes (chRCC, ccRCC, pRCC, Supplementary Fig. S2B, S2C). In total, 16,852 peaks were significantly increased or decreased in one tumor histology compared to the other two (e.g., chromophobe versus clear cell and papillary) (“Methods”, false discovery rate (FDR) of 0.001 and least a fourfold difference in mean peak intensity between groups). In all, 12,908 sites were upregulated in one histology: 8939 were chRCC-specific, 3653 were pRCC-specific, and 316 were ccRCC-specific (Fig. 1E). These histology-specific H3K27ac peaks were differentially marked by H3K4me2 ChIP-Seq and were associated with open chromatin, strongly suggesting that they were histology-specific active enhancers (Supplementary Fig. S2D). Moreover, differential epigenetic sites correlated with the nearest gene expression difference ( $P$  value  $< 2 \times 10^{-16}$ , chi-square test) (Fig. 1F–H). GREAT<sup>33</sup> analysis of the 8939 chRCC-specific H3K27ac peaks revealed enrichment for genes involved in actin regulation, fatty acid oxidation, and ion transmembrane transporter activity (Fig. 2A), consistent with previously reported mRNA signature analysis showing an increased ion transmembrane transport signature in chRCC<sup>34</sup>. A similar analysis of the 3653 pRCC-specific sites showed enrichment for genes involved in renal system development (Fig. 2B). This is consistent with the notion that pRCC arises from embryonic nephrogenic rest precursor lesions, which persist during adult life<sup>35</sup>. Since there was a relatively small number of ccRCC-specific sites ( $n = 316$ ) in comparison to chRCC and pRCC, and the enhancer landscape of ccRCC resembled that of pRCC more than chRCC, we compared H3K27ac sites between ccRCC and pRCC only. The majority of H3K27ac sites were common to the two histologies ( $n = 113,786$ ), while 1265 sites were ccRCC-specific, and 2661 sites were pRCC-specific (Supplementary Figs. S2E and S3A). The 1265 ccRCC-enriched peaks were associated with genes involved in circulatory system development and angiogenesis (Fig. 2C), while the 2661 pRCC-enriched peak genes were again enriched for genes involved in kidney embryogenesis (Supplementary Fig. S3B). Similar analysis of the H3K4me2 peaks demonstrated clear separation of chRCC from the other two RCC subtypes with comparable histology-specific and common “poised” sites (Supplementary Fig. S3C–3F). H3K27ac and H3K4me2 ChIP-seq signals for all samples were strongly correlated (Pearson correlation,  $r = 0.73$ ) (Fig. 2D).

De novo motif analysis of H3K27ac peaks enriched in each subtype identified four motifs that were highly enriched in chRCC, including one resembling a forkhead motif, and another resembling the motif associated with TFDP2 (Fig. 2E). FOXI1, a forkhead family TF, and TFDP2L1, closely related to TFDP2, have both been implicated in the development of intercalated cells of the kidney, the putative cell of origin of chRCC<sup>36</sup>. *FOXI1* and *TFDP2L1* gene loci were characterized by high H3K27ac signal in chRCC compared to the absent or markedly lower signal in ccRCC and pRCC (Fig. 2F, G). Motif enrichment analysis of putative ccRCC-specific enhancers identified a motif resembling the basic Leucine Zipper (bZIP) BATF motif, and another resembling the basic helix-loop-helix (bHLH) TF family member HIF2 $\alpha$  (also known as EPAS1) (Fig. 2H). *ETS1*, in the bZIP family with BATF, has been implicated in von-Hippel Lindau (VHL)-dependent ccRCC tumorigenesis<sup>37</sup>, and was highly marked by H3K27ac in ccRCC, less so in pRCC, and not in chRCC (Fig. 2I). *EPAS1* is well-known to be dysregulated in ccRCC due to the loss of the *VHL* protein function. It is the main driver of ccRCC and is upstream of multiple critical oncogenic pathways. Recent clinical trials with HIF2 inhibitors have shown clinical activity in advanced ccRCC patients<sup>38,39</sup>. The top-scoring pRCC-specific motif corresponded to HNF1 $\beta$  (Fig. 2J, K). *HNF1 $\beta$*  is a member of the homeodomain-containing superfamily of TFs, which is involved in nephrogenesis<sup>40</sup>, and was highly marked by H3K27ac specifically in pRCC (Fig. 2K).

RCC has recurrent mutations in genes involved in the chromatin remodeling and histone methylation pathways<sup>7,8,41–44</sup>. In our dataset, loss-of-function single-nucleotide variant and copy number alterations



**Fig. 1 | Landscape of H3K27ac signals across RCC histologies.** **A** Distribution of RCC H3K27ac peaks according to genomic region for 30 fresh frozen RCC tumor samples (12 chRCC, 6 pRCC, 12 ccRCC). **B** Numbers of histology-specific and common H3K27ac peaks. **C** H3K27ac profiles at *PAX8* in six representative samples from each RCC histology. **D** Hierarchical clustering of chRCC, ccRCC, and pRCC based on sample-to-sample pairwise correlation of the H3K27ac ChIP-seq peaks.

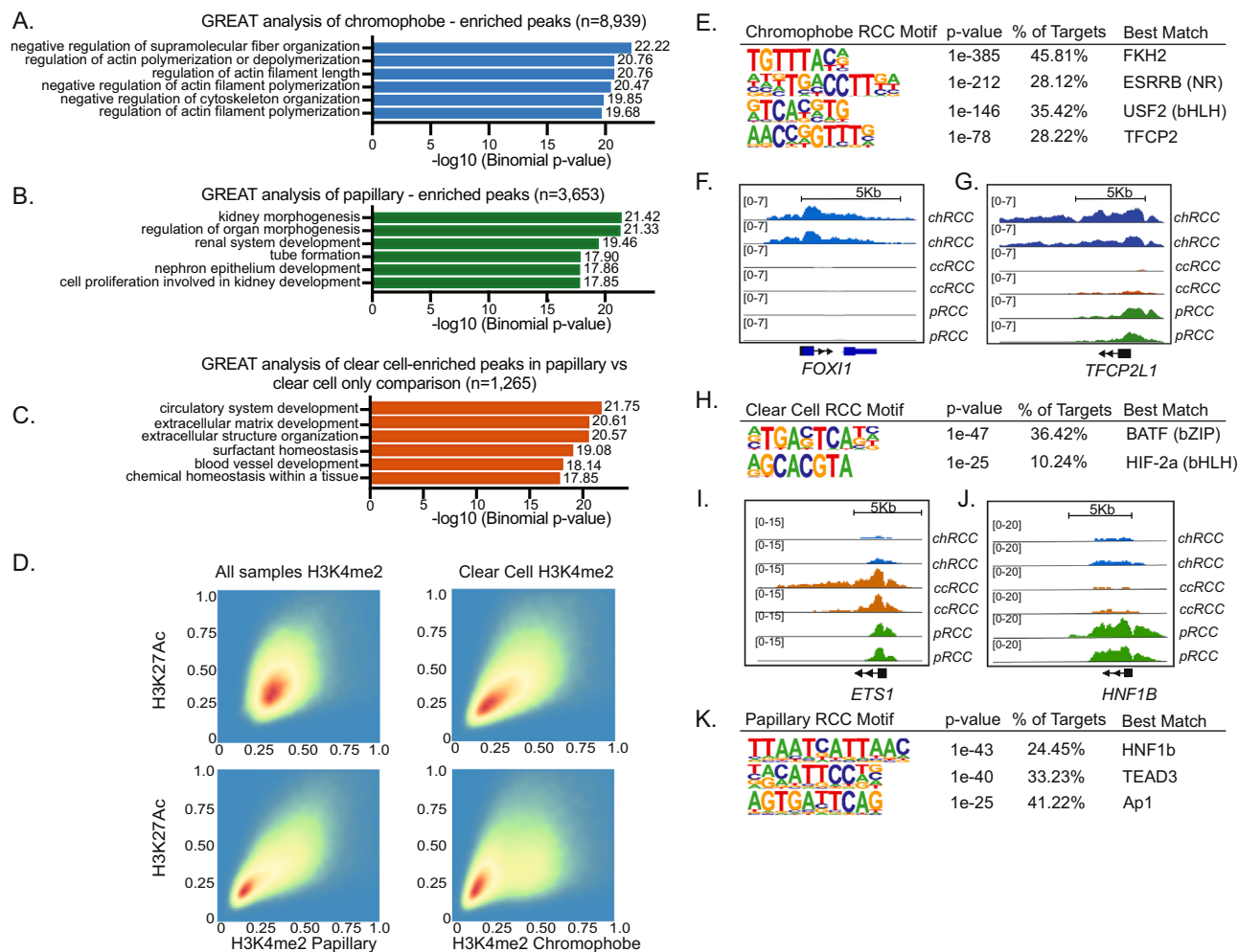
**E** Distribution of histology-specific H3K27ac peaks among RCC subtypes. **F–H** Volcano plots with the log change of gene expression (FPKM) in one histology compared to the other two histologies (**F** ccRCC vs. others, **G** pRCC vs. others, **H** chRCC vs. others). Two-sided *P* values were used and corrected for multiple comparison testing (FDR-adjusted *P* value < 0.05). RCC renal cell carcinoma, chRCC chromophobe RCC, ccRCC clear cell RCC, pRCC papillary RCC.

(see “Methods”) in such commonly mutated genes in RCC (*VHL*, *PBRM1*, *BAP1*, and *TP53*) did not correlate with global acetylation differences in both supervised and unsupervised analyses, albeit the small sizes were small. (Supplementary Fig. S4A–S4D and Supplementary Data 3 and 4).

### Specific master transcription factors define RCC subtypes

We next sought to systematically identify candidate histology-specific master TFs that define the three subtypes of RCC. Master TFs typically bind within SEs<sup>17,45</sup>, are often regulated by SEs, and regulate one another in a transcriptional core regulatory circuit (CRC)<sup>46</sup>. We

employed an integrative approach<sup>47,48</sup>, leveraging the RNA-Seq, ChIP-Seq, and ATAC-Seq data to identify candidate histology-specific master TFs (Fig. 3A). This approach aims to utilize orthogonal information to identify a consensus set of master TFs. We combined (1) expression data of differentially expressed TFs across RCC histologic subtypes; (2) TFs specific to RCC histologic subtypes relative to other cancer types (CaCTS)<sup>49</sup>; (3) differential SE-associated TFs among RCC histologic subtypes; and (4) TFs with histology-specific connectivity in regulatory cliques (see “Methods”). These four analyses identified more than 200 candidate TFs showing a histology-specific association in one or more analysis. We prioritized candidates for downstream validation by



**Fig. 2 | Epigenetic annotation of regulatory elements identifies enrichment of histology-specific pathways and TFs.** **A** GREAT analysis of chromophobe-enriched peaks ( $n = 8939$ ). Two-sided  $P$  values are shown. **B** GREAT analysis of papillary-enriched peaks ( $n = 3653$ ). Two-sided  $P$  values are shown. **C** GREAT analysis of clear cell-enriched peaks in clear cell vs. papillary only comparison ( $n = 1265$ ). **A–C** GREAT calculates statistical enrichments for association between genomic regions and annotations. Two-sided  $P$  values are shown. **D** Density map of correlation between H3K27ac versus H3K4me2 ChIP-seq peaks across subtypes. **E** Four most significantly enriched nucleotide motifs present in chRCC-specific sites by de novo motif analysis, limited by ATAC peaks. **F, G** H3K27ac profiles near

FOXI1 and TFCP2L1, respectively, in two representative samples of each histology (chRCC, ccRCC, pRCC). **H** Two most significantly enriched nucleotide motifs present in ccRCC specific sites by de novo motif analysis.  $P$  values are two-sided and unadjusted for multiple comparisons. **I, J** H3K27ac profiles near ETS1, and HNF1B, respectively, in two representative samples of RCC histology. **K** Three most significantly enriched nucleotide motifs present in pRCC-specific sites by de novo motif analysis.  $P$  values are two-sided and unadjusted for multiple comparisons. RCC renal cell carcinoma, chRCC chromophobe RCC, ccRCC clear cell RCC, pRCC papillary RCC.

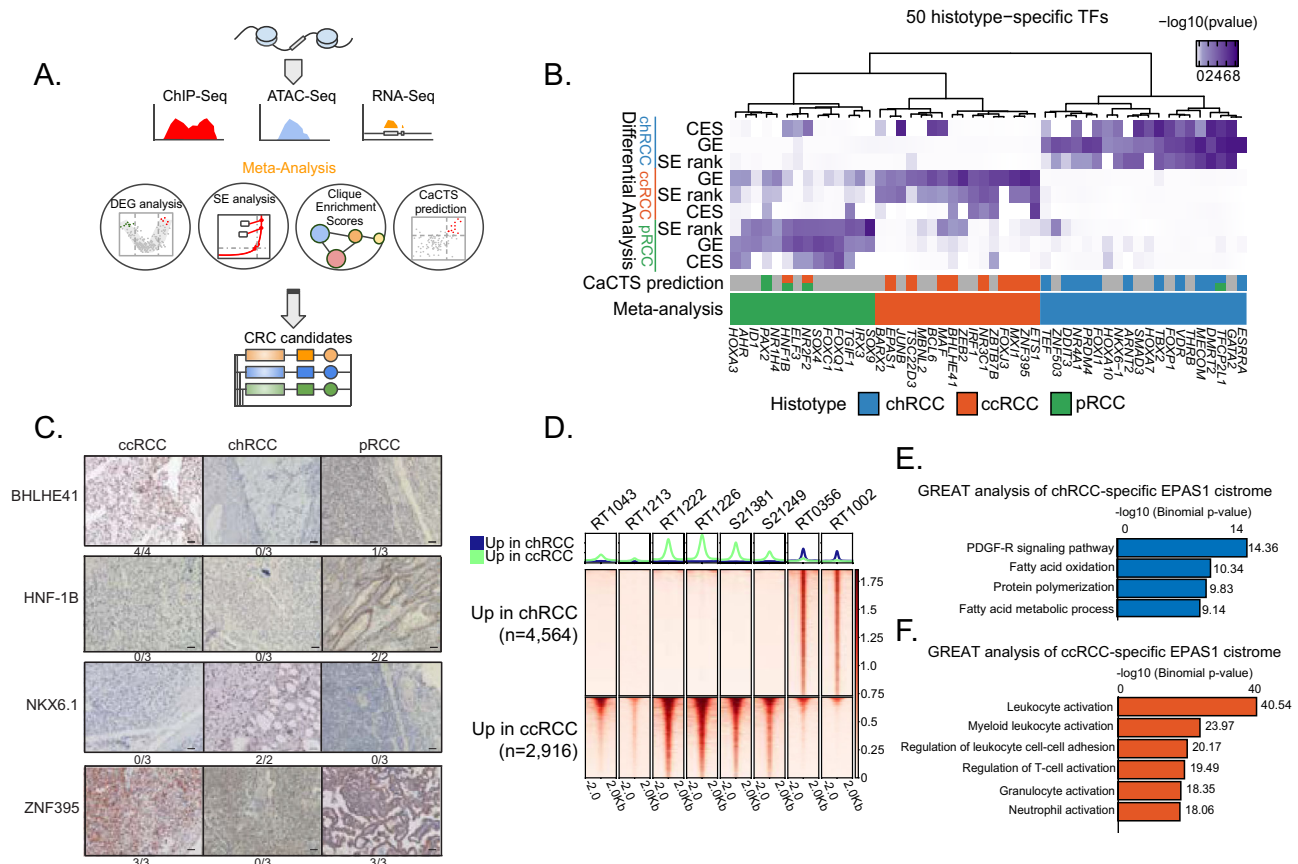
selecting those that were identified in more than one analysis (“Methods”, Supplementary Figs. S5A–S5) and S6 and Supplementary Data 5–11). This analysis highlighted 50 candidate histology-specific master TFs ( $N = 20$ , chRCC;  $N = 14$ , pRCC;  $N = 16$ , ccRCC) (Supplementary Data 12), including *FOXI1*, *TFCP2L1*, and *DMRT2* for chRCC; *EPAS1*, *ETS1*, *BARX2*, *ZNF395* for ccRCC, and *HNF1B* and *NR2F2* for pRCC (Fig. 3B). SE ranking, gene expression, and CES of the 50 histology-specific TFs selected using this meta-analysis approach clustered the samples tightly according to their respective histologies (Supplementary Fig. S6A).

To confirm whether similar patterns of TF-specificity are found in normal counterpart tissues, we turned to gene expression datasets from TCGA KICH, KIRC, and KIRP cohorts. GSEA analysis for histology-specific master TFs showed enrichment of clear cell and papillary RCC-specific transcription factors in tumors compared to normal (FDR corrected  $q$ -value  $< 0.01$ ). There was no significant enrichment of chromophobe RCC-specific transcription factors in tumors compared to normal. Of 50 master TFs, 34 (68%) had more significant expression in the tumor tissue compared to the normal counterpart (14/16 for

KIRC, 10/14 for KIRP, and 10/20 for KICH, Supplementary Fig. S6B–S6D).

Clinical correlative analyses from CheckMate cohorts (009/10/025) showed that among the 50 master TFs, high *BARX2* expression significantly correlated with improved overall survival in the entire cohort of patients with ccRCC (Supplementary Data 13 and Supplementary Fig. S6E) and in the subgroup treated with the anti-PD1, nivolumab (Supplementary Data 13 and Supplementary Fig. S6F).

To provide proof of concept validation at the protein level, we investigated the expression specificity and localization of master TFs by immunohistochemistry (IHC). We selected four representative master TFs that met the following criteria: (1) at least twofold change in gene expression of the histology-specific TF by bulk RNA sequencing (from TCGA) compared to the two other RCC histotypes, (2) high-quality antibodies for IHC, and (3) TF was not previously validated and implemented on a clinical level. For the 4 TFs (BHLHE41, HNF1 $\beta$ , NKX6.1, and ZNF395), we found significant changes in protein expression levels across histologic subtypes (Fig. 3C). More specifically, two ccRCC-specific master TFs (BHLHE41 and ZNF395), were



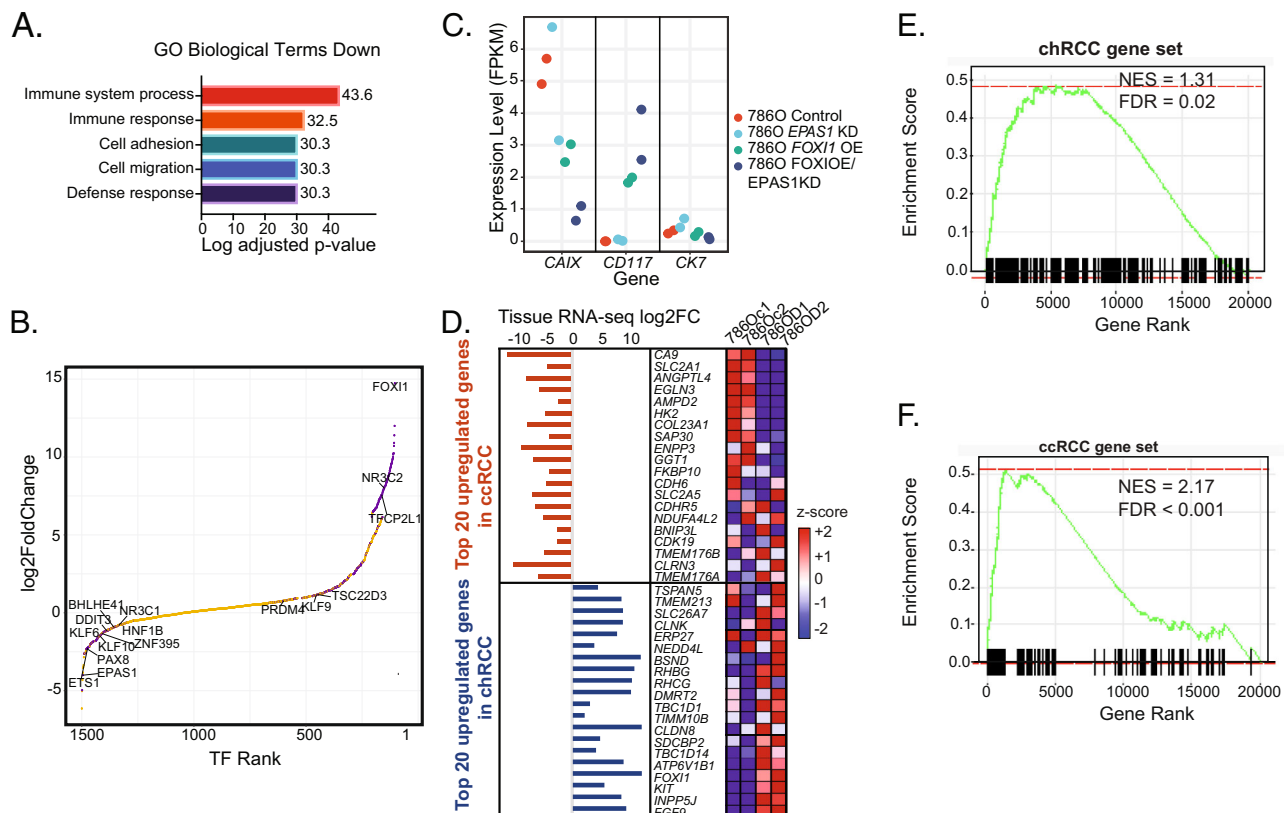
**Fig. 3 | Multi-dimensional integrative analysis identifies histology-specific master TFs.** **A** Overview of the approach used to nominate histology-specific master TFs participating in CRC. **B** Heatmap integrating the 50 histology-specific TFs identified by the meta-analysis approach (CES, differential expression, SE rank analysis, and CaCTS). **C** Representative immunohistochemical stainings of indicated antibodies in samples from ccRCC, chRCC, and pRCC tumors. Four histology-specific master TFs are shown. The number of positive tumors/number of tumors examined are indicated below each image (for ccRCC,  $n = 4$  for TF BHLHE41,  $n = 3$  for TFs HNF1B, NKX6.1, and ZNF395. For chRCC,  $n = 3$  for TFs BHLHE41, HNF1B, and ZNF395,  $n = 2$  for NKX6.1. For pRCC,  $n = 3$  for BHLHE41, NKX6.1, and ZNF395,  $n = 2$  for HNF1B). Scale bar is 50  $\mu\text{m}$ . **D** ChIP-seq binding profiles of EPAS1 across 6 ccRCC

and 2 chRCC human tissue samples. **E** GREAT analysis of chRCC-specific EPAS1 peaks relative to ccRCC. Two-sided  $P$  values are shown. **F** GREAT analysis of ccRCC-enriched EPAS1 peaks relative to chRCC. Two-sided  $P$  values are shown. **E**, **F** GREAT calculates statistical enrichments for the association between genomic regions and annotations. ChIP-seq chromatin immunoprecipitation sequencing, ATAC-seq assay for transposase-accessible chromatin sequencing, DEG differentially expressed genes, SE superenhancer, CaCTS cancer core transcription-factor specificity, CRC core regulatory circuitries, RCC renal cell carcinoma, chRCC chromophobe RCC, ccRCC clear cell RCC, pRCC papillary RCC. Source data are provided as a Source Data file.

highly expressed in ccRCC tumors. Both TFs had nuclear localization in four out of four and three out of three ccRCC samples, respectively. NKX6.1, a TF recently described as being expressed in chRCC, was detected nuclearly in two out of two chromophobe samples with no expression in either ccRCC or pRCC. We next confirmed histology-specific *in vivo* binding of the nominated master TF EPAS1 through examining the EPAS1 cisrome in ccRCC and chRCC. We characterized 2916 clear cell RCC-specific and 4564 chromophobe RCC-specific EPAS1-binding sites through performing EPAS1 ChIP on eight additional primary human tumors ( $n = 2$  chRCC;  $n = 6$  ccRCC, Fig. 3D). Subtype-specific EPAS1-binding sites were highly enriched for subtype-specific sites of H3K27ac. For instance, 2090 of the 4565 chromophobe-specific EPAS1-binding sites (46%) overlapped with chromophobe-specific H3K27ac sites ( $P < 2.2e-16$ ). This supports the conclusion that these differential EPAS1-binding sites are non-random and biologically relevant, because they coincide with regulatory elements that segregate closely with histology. ccRCC-specific EPAS1-binding sites were enriched for immune cell and white blood cell activation pathways and chRCC-specific EPAS1-binding sites were enriched for metabolic processes and fatty acid activation (Fig. 3E, F).

To further validate our approach of the nominated master TFs in driving the transcriptional identity of the different RCC histologies, we

manipulated TF expression in a ccRCC cell line. We hypothesized that overexpression of chRCC-specific TFs and suppression of ccRCC-specific TFs can shift the transcriptional landscape of the ccRCC cell line 786-O to become more chromophobe-like. *FOXII* scored as a chRCC-specific TF in 4/4 master TF analyses (Fig. 3B). Furthermore, *FOXII* is selectively expressed in intercalated cells (ICs), the putative cellular origin of chRCC<sup>36,50</sup> and is more highly expressed in chRCC than other cancer subtypes in the TCGA dataset (Supplementary Fig. S7A). We also manipulated the expression of *EPAS1* as a second target as it was highly specific for ccRCC in our integrative analysis (Fig. 3B and Supplementary Fig. S8B), and prior studies have characterized its role in the pathogenesis of ccRCC<sup>51</sup>. We overexpressed *FOXII* in the ccRCC cell line 786-O (*FOXII* OE); suppressed *EPAS1* (*EPAS1* KD) and simultaneously perturbed both genes in the same cell line (*FOXII* OE/*EPAS1* KD) (Supplementary Fig. S7C–7E and Supplementary Data 14). Expression changes are described in Supplementary Fig. S8A–S8F and Supplementary Data 15–18. Gene set enrichment analysis (GSEA) of downregulated genes with *FOXII* OE/*EPAS1* KD in 786-O showed enrichment of genes involved in the immune responses (Fig. 4A). Concomitant *FOXII* OE and *EPAS1* KD in 786-O cells resulted in upregulation of chRCC-specific master TFs such as *TFCP2L1* and *NR3C2* and down-regulation of ccRCC-specific master TFs such as *ETS-*



**Fig. 4 | Functional perturbation of two master TFs in a ccRCC cell line yields a more chRCC-like transcriptional profile.** **A** Downregulated GO biological terms in the cell line 786-O FOXI1 OE/EPAS1 KD. Two-sided adjusted *P* value corrected for multiple comparisons. **B** Rank order of differentially expressed TFs between 786-O CTRL and 786-O FOXI1 OE/EPAS1 KD by expression levels. Each TF dot represents one TF. Wald test from DESeq2 was used. Purple dots indicate adjusted *P* value <0.001 by DESeq Expression. **C** Gene expression levels of CAIX, CD117, and CK7 across different 786-O cell line conditions. **D** Heatmap showing the relative expression of the top 20 upregulated genes in ccRCC vs. chRCC and vice versa in

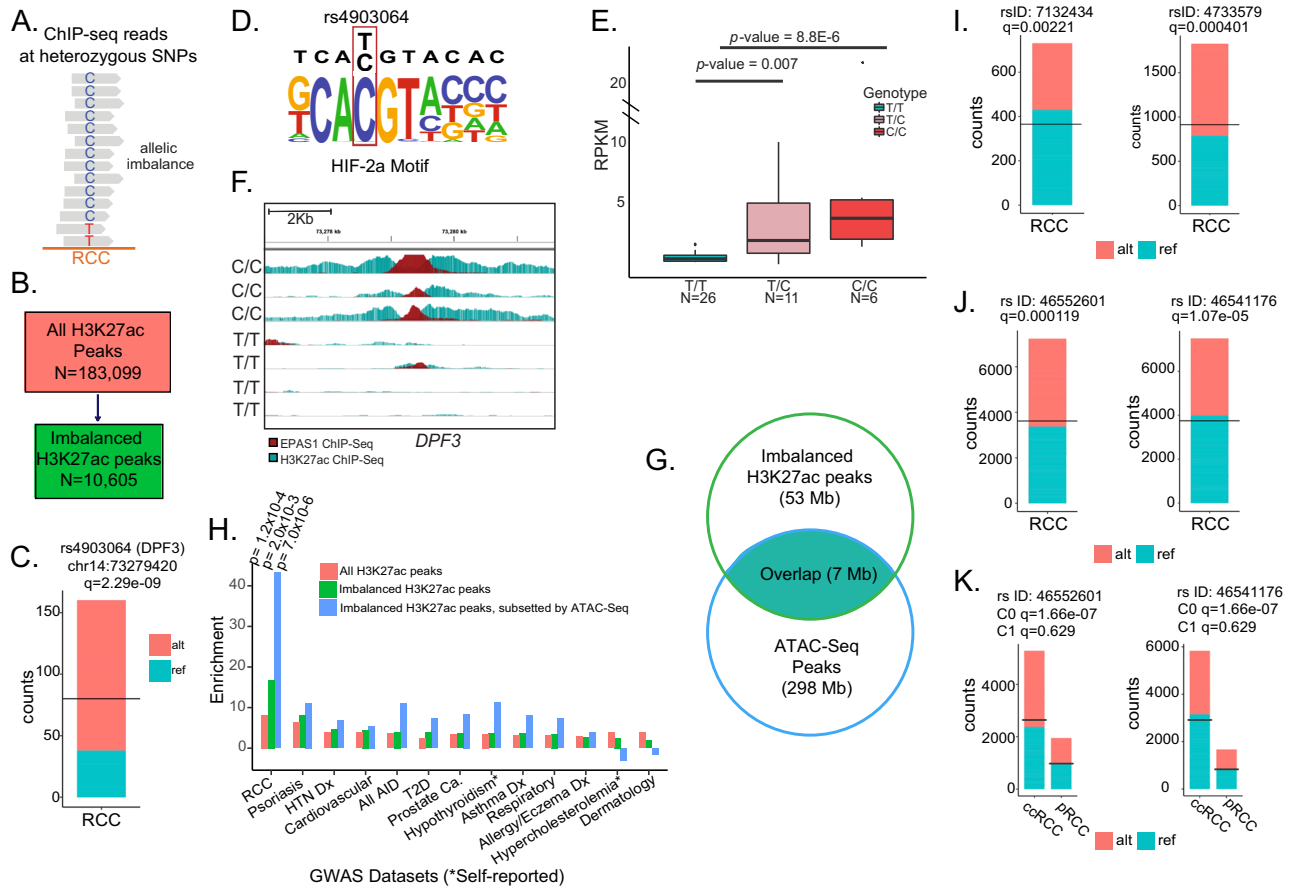
the different cell line conditions. **E** GSEA analysis showing the top 100 upregulated genes from chRCC vs. ccRCC dataset in 786-O FOXI1 OE/EPAS1 KD vs. 786-O CTRL comparison. **F** GSEA analysis of the top 100 upregulated genes in ccRCC vs. chRCC in 786-O CTRL vs. 786-O FOXI1 OE/EPAS1 KD. CTRL control, KD knockdown, OE overexpression, TF transcription factor, GSEA gene set enrichment analysis, 786Oc1 CTRL1, 786Oc2 CTRL2, 786OD1 786-O FOXI1 OE/EPAS1 KD1, 786OD2 786-O FOXI1 OE/EPAS1 KD2, NES Normalized Enrichment Score, FDR false discovery rate, FC fold change, RCC renal cell carcinoma, chRCC chromophobe RCC, ccRCC clear cell RCC, pRCC papillary RCC.

1 and ZNF395 compared to control cells. (Fig. 4B). Further supervised analysis showed that CAIX, a downstream target of EPAS1 was significantly overexpressed in the WT 786-O cell line compared to EPAS1 KO/FOXI1 OE cell line. In contrast, CD117, a well-known immunohistochemical marker of chRCC<sup>52</sup>, was overexpressed in the EPAS1 KO/FOXI1 OE cell line compared to WT 786-O cell line. Although CK7 expression is a supportive marker clinically for chRCC<sup>53</sup>, its expression was uniform across WT 786-O and manipulated cell line states (Fig. 4C). Of note, there were only 109 differentially expressed genes between the FOXI1 OE/EPAS1 KD cell and FOXI1 OE only cell line ( $P < 0.05$ ,  $q < 0.01$ , Supplementary Fig. S8D and Supplementary Data 17).

We then compared differentially expressed genes between our experimental conditions in 786-O and our RCC tumor expression data (Fig. 4D and Supplementary Fig. S8G). Defining chRCC and ccRCC gene sets as the top 100 upregulated genes for each histology relative to the other, we showed that the chRCC gene set is enriched among upregulated genes in 786-O FOXI1 OE/EPAS1 KD cell line vs. control (Fig. 4E), and the ccRCC gene set is enriched among upregulated genes in 786-O control vs. FOXI1 OE/EPAS1 KD cell line (Fig. 4F). The FOXI1 OE/EPAS1 KD 786-O cell line demonstrated significantly higher expression of chRCC-specific candidate master TFs, TFCP2L1, GATA2, DDIT3, NKX6-1, and lower expression of ccRCC-specific candidate master TFs, ZNF395 and TSC22D3. Differential TFs in the 786-O FOXI1 OE/EPAS1 KD vs. 786O CTRL comparison overlapped more significantly with differential TFs from the chRCC vs. ccRCC human

samples comparison than from the ccRCC vs. pRCC human sample comparison (Supplementary Data 19). In summary, these data show that overexpression of a single chRCC master transcription-factor candidate, FOXI1, in the ccRCC cell line 786-O, with or without knockdown of EPAS1, led to marked expression changes driving the cell line to be more like a chRCC cell line without any modification to the set of mutations present in 786-O cells.

**Allelic imbalance annotates germline RCC risk variants.** Allelic imbalance, the differential allelic representation of heterozygous single-nucleotide polymorphisms (SNPs) in ChIP-seq reads, provides an in vivo comparison of *cis*-regulatory activity between two haplotypes (Fig. 5A). As such, allelic imbalance can highlight the functional relevance of candidate causal variants<sup>22,32,54–57</sup> at loci that have been associated with RCC risk through GWAS. A key advantage of profiling epigenomes from many individuals is the ability to capture heterozygous sites and to measure the effect of TF binding on regulatory elements through analysis of allelic imbalance. To this end, we assessed chromatin allelic imbalance in ChIP-seq reads from these RCC samples in order to nominate causal risk SNPs from a genome-wide association study of RCC. We applied stratAS<sup>32</sup> to H3K27ac ChIP-seq data from 20 ccRCC and 6 pRCC (“Methods”). We identified 10,605 chromatin-imbalanced H3K27ac peaks—which we defined as peaks with one or more imbalanced SNPs after correction for multiple hypothesis testing—across 183,099 peaks tested in the combined ccRCC and pRCC sample sets (Fig. 5B and Supplementary Data 20).



**Fig. 5 | Allelically imbalanced H3K27ac peaks in ccRCC and pRCC.** **A** Schematic of allelic imbalance at heterozygous SNPs. **B** Schematic showing subset of allelically imbalanced H3K27ac peaks from total H3K27ac peaks in RCC (pRCC and ccRCC) samples. **C** AI at rs4903064 (chr14:73279420; DPF3) in RCC. **D** rs4903064 C-allele sequence context creates a HIF-2α binding site. **E** RPKM values for the EPAS1 ChIP signal in the peaks around rs4903064 in 43 samples. Genotype is shown on the X axis. Box boundaries correspond to 1st and 3rd quartiles; whiskers extend to a maximum of 1.5× the interquartile range. Two-sided unadjusted p-values are shown. Statistical analysis was performed using Kruskal–Wallis test. **F** Overlaid EPAS1 and H3K27ac ChIP-Seq tracks for seven samples within -10 Kb of genomic coordinates of rs4903064. Individual sample genotypes are shown. **G** Venn diagram showing overlap of imbalanced H3K27ac peaks with ATAC-seq peaks in RCC. **H** Enrichment of risk SNPs from GWAS for RCC and other diseases in

H3K27ac peaks with differential allelic imbalance, compared to all H3K27ac peaks. Empiric one-sided P value is indicated (unadjusted for multiple comparison). Statistical analysis was performed using Kruskal–Wallis test. GWAS risk SNPs rs7132434 and rs4733579 demonstrating allelic imbalance in RCC. **I** AI at Chr12 SNP rs7132423 and Chr8 SNP rs4733579 shown. **J** AI at Chr2 SNPs within EPAS1, rs46552601 and rs46541176 shown. **K** Chr2 SNPs within EPAS1, with imbalance plots split by histology. Adjusted Q values for imbalance are indicated. ChIP-seq chromatin immunoprecipitation sequencing, Ca cancer, HTN hypertension, Dx diagnosis, T2D type 2 diabetes mellitus, AID autoimmune disease, C0 reference allele, C1 alternate allele, AI allelic imbalance, RCC renal cell carcinoma, chrRCC chromophobe RCC, ccRCC clear cell RCC, pRCC papillary RCC, SNP single-nucleotide polymorphism.

We hypothesized that chromatin AI peaks correspond to regulatory elements that are bound by master TFs. At these elements, the presence of *trans*-acting factors (i.e., master TFs) may result in regulatory element activity that are observable as H3K27ac AI peaks. One example is rs4903064 (chr14:73279420; 14q24.2), an eQTL for *DPF3*<sup>58</sup> (Fig. 5C). rs4903064, which is allelically imbalanced in our H3K27ac dataset, has been shown to be a GWAS risk variant in all three histologic subtypes (ccRCC, chrRCC, and pRCC), and the altered C-allele of this SNP has been suggested to create a HIF-binding motif (Fig. 5D)<sup>59</sup>. To confirm this, we used EPAS1 transcription-factor ChIP-Seq data in an independent cohort of 43 samples derived from 23 patients with ccRCC to study the effect of the GWAS risk variant rs4903064 (see “Methods”) on EPAS1 binding. Analysis of 11 primary, 21 metastatic ccRCC samples, and 11 normal renal tissue samples showed that tumors with homozygous C/C alleles were significantly enriched for EPAS1 peaks (Fig. 5E, F) in both tumors and normal tissue and that rs4903064 is a *EPAS1* cQTL. This confirms prior literature that the C-allele creates a HIF-binding site<sup>59</sup>.

We applied our chromatin AI analysis to annotate risk SNPs identified by a GWAS for RCC<sup>58</sup>. Chromatin AI peaks were highly enriched (16.7-fold) for RCC GWAS risk variants as assessed by LD score regression analysis ( $P=1.9 \times 10^{-4}$ )<sup>60</sup>. Subsetting imbalanced H3K27ac peaks to regions of accessible chromatin where TFs are likely to bind (as assessed by ATAC-Seq from RCC tissues<sup>61</sup>) resulted in substantial additional enrichment (43.2-fold;  $P=7.0 \times 10^{-6}$ , Fig. 5G, H). This enrichment represents more than fivefold that of the total set of H3K27ac peaks, which are themselves enriched 8.1-fold ( $P=1.2 \times 10^{-4}$ ). By contrast, multiple other GWAS phenotypes from the UK Biobank<sup>62</sup> showed substantially less enrichment compared to RCC GWAS SNPs, indicating the specificity of this enrichment for RCC (Fig. 5H).

Using this method, we were able to fine-map a total of 30 risk SNPs (Supplementary Data 21), with some examples highlighted. Rs7132434, located on chr 12, has been characterized as a functional variant that alters AP-1 binding leading to upregulation of *BHLHE41*, thus promoting tumor growth through induction of *IL-17*<sup>63</sup> (Fig. 5I). Another example is rs4733579 which is located on chr 8 and is in LD with

rs35252396 (Fig. 5I). The latter has been shown to contribute to RCC susceptibility through regulating MYC and PVT1 expression<sup>64</sup>. Of note, rs35252396 is an indel and therefore cannot be assessed by stratAS. Our analysis also highlighted rs46552601 and rs46541176 (Fig. 5J), both located on chr2p21 within the *EPAS1* gene, where at least 59 SNPs have been fine mapped. These two SNPs are allelically imbalanced by H3K27Ac peaks only in ccRCC but not in pRCC, consistent with the specific role of *EPAS1* (*HIF2-α*) in ccRCC pathogenesis (Fig. 5K). Further work is recommended to functionally validate the mechanism of these two SNPs in ccRCC pathogenesis. Finally, rs4765623 located in the *SCARBI* locus has been recently validated in RCC progression<sup>32</sup>.

TFs can read the genetic code and bind *cis*-regulatory elements to activate or repress gene expression. We hypothesized that loci with significant chromatin AI are likely harbingers of TF binding and thus are associated with the regulation of gene expression in an allele-specific manner. We leveraged chromatin AI from our ccRCC H3K27Ac dataset and allele-specific expression (ASE) using RNA-seq data from the KIRC TCGA cohort. SNP loci with at least 50 reads in both the ccRCC H3K27Ac dataset and RNA-seq KIRC TCGA samples were retained. Retained SNPs were classified as chromatin allelically imbalanced or allelically balanced in ccRCC. Chromatin allelically imbalanced SNPs harbored H3K27ac peaks with a significant skew towards the alternate allele compared to the wild-type allele ( $P < 0.05$ ). For the 22,162 chromatin allelically imbalanced SNPs, we matched a background set of chromatin allelically balanced SNPs lying within the H3K27ac consensus peak set. Allele-specific expression of genes whose transcription start site lies within 50 Kb of the chromatin allelically imbalanced and balanced SNPs was analyzed in the TCGA KIRC dataset ( $N = 412$ ) with a significance cutoff of  $P$ -adjusted  $< 0.01$ . Of the unique genes analyzed, chromatin allelically imbalanced SNPs were significantly more likely to lie within 50 kb of a gene with ASE (1170/2646, 44%) compared to the background set of chromatin allelically balanced SNPs (65/940, 6.9%,  $P = 8.4 \times 10^{-25}$ , Supplementary Data 22–25).

## Discussion

In this study, we report a compendium of ATAC-Seq, RNA-Seq, and histone modification data across the three most common RCC subtypes. Our goal was to describe enhancer programs in primary human samples from patients with RCC and to identify histology-specific epigenetic mechanisms that may underlie differential clinical features and prognosis. Critically, this resource includes chRCC, an understudied cancer type with scarce *in vitro* and *in vivo* models and molecular profiling data, and no proven therapeutic strategies for metastatic disease.

In line with histological, transcriptional and mutational data, ccRCC, pRCC, and chRCC exhibit distinct epigenetic landscapes. chRCC clearly separated from ccRCC and pRCC, both by H3K27ac and H3K4me2 landscapes. Motif-based analysis revealed the enrichment of distinct TF binding sites in H3K27ac peaks in ccRCC vs. pRCC, suggesting that distinct TFs were active and driving histology-specific regulatory pathways and downstream transcriptional programs. Using *EPAS1* TF ChIP-Seq as an example, we demonstrate that histology-specific binding of *EPAS1* mediates independent pathways in different RCC histologies.

Prior work in medulloblastoma and ependymoma demonstrated how molecularly-defined cancer subgroups exhibit specific core regulatory circuitries<sup>65</sup>. Herein, we nominated 50 histology-specific TFs in RCC based on a convergence of evidence, including high-level expression, specificity of expression across RCC subtypes, super-enhancer association, and connectivity. Our experimental results demonstrate that upregulating a chRCC-specific TF, FOXI1, can partially reprogram a ccRCC cell line, 786-O, towards a chRCC transcriptional phenotype; consistent with results from other tumor and normal cell types showing that master TFs govern the transcriptional identity of the cell or tissue.

Our work furthermore sheds light on multiple subtype-specific master TFs, the role of which is yet to be exactly determined. *HNF1β*, a candidate pRCC master TF, has been shown to be expressed in the embryonic kidney and in pRCC, and it is upregulated in pRCC compared to other kidney cancer histologies<sup>7,66,67</sup>. *HNF1β* gene is located on chromosome 17 and is frequently impacted by copy number gains in pRCC<sup>66</sup> further providing grounds for its role in pRCC pathogenesis. *ETS-1*, a candidate ccRCC master TF known to have a role in cancer pathogenesis as well as angiogenesis and hematopoietic stem cell differentiation, is another example<sup>68–71</sup>. As a proto-oncogenic factor, *ETS-1* is capable of activating genes associated with angiogenesis, metastasis and invasive behavior in multiple tumor types<sup>72–74</sup>. In addition, *ETS-1* expression correlates with microvessel density in some non-glial tumors and is an independent negative prognostic marker in different tumor entities such as breast, ovarian, pancreatic, and colorectal cancers<sup>72,75,76</sup>. Nonetheless, the role of *ETS-1* in RCC has mainly focused on its interaction with *EPAS1* (*HIF2-α*)<sup>77</sup>. Here, we show that *ETS-1* may play a role as a master TF specific to ccRCC and not chRCC or pRCC. Future studies should focus on understanding the global epigenomic effect of *ETS1* in driving ccRCC pathogenesis.

Using clinical trial data (CheckMate 009/010/025), we suggest that high *BARX2* expression may be associated with longer survival among patients treated with nivolumab but not everolimus. Prospective data in the metastatic ccRCC space are needed to confirm the role of *BARX2* expression as a predictive biomarker in patients treated with immune checkpoint inhibitors.

In routine practice, pathologists use a myriad of targets to differentiate across RCC histotypes. Our approach has the potential to narrow master TFs to clinically meaningful ones and augments the current armamentarium with additional TFs, including BHLHE41 (ccRCC) and NKX6.1 (chRCC) that can be stained to better characterize tumors.

BHLHE41 was previously shown in triple-negative breast cancer to counteract expression of HIF-target genes by promoting HIF proteasomal degradation in a process independent of VHL or hypoxia<sup>78</sup>. Elevated protein expression of BHLHE41 in ccRCC has not been described previously. Using a systematic approach, we found histology-specific protein expression of BHLHE41 in ccRCC compared to pRCC and chRCC. In ccRCC where HIF activation is foundational, further investigation of the relation between BHLHE41 and HIF activation is warranted.

GWAS have identified hundreds of putative cancer-risk loci<sup>79</sup>. These studies have led to the conclusion that cancer is driven by thousands of variants with individual small effects, and that cancer-risk variants predominantly lie in non-coding regions of the genome<sup>80,81</sup>. More recently, it has been realized that GWAS heritability is enriched at variants that lie in tissue-relevant epigenomic features<sup>20,60,82–84</sup>. Emerging evidence suggests that many GWAS variants function by altering an existing or creating a TF binding site. This in turn regulates the local chromatin regulatory landscape by altering the enhancer landscape and thus modulating the expression of target genes<sup>85</sup>. We used our population-scale epigenomic data to further empower the discovery of *cis*-regulatory elements that may mediate between the risk variant and target gene and thus constrain putative targets for experimental validation. We showed that RCC GWAS risk SNPs are enriched in regulatory elements, consistent with prior studies<sup>20,60,82–84</sup>. Profiling these regulatory elements across multiple (genetically diverse) individuals allowed us to identify regions of chromatin allelic imbalance, which are further enriched for GWAS SNPs, and subsetting these regions to ATAC-seq peaks where TFs are likely to bind, showed even more enrichment for GWAS SNPs. In RCC, there are 13 recognized risk loci, but the causal variants have been identified for few, and underlying mechanisms of risk remain elusive for most regions<sup>88</sup>. Our approach can help pinpoint causal SNPs and provide candidate mechanisms by identifying small numbers of SNPs that may alter binding sites for



specific TFs, an example of which is the rs4903064 locus. Recent work around the rs4903064 locus<sup>59</sup> showed that it had an allele-specific effect on DPF3 expression in ACHN and HEK293T cell lines as assessed by massively parallel reporter assay, confirmatory luciferase assays, and eQTL analyses. The rs4903064-C RCC risk allele was shown to create a HIF-binding site and enhance gene expression. Increased expression of DPF3 conferred a growth advantage to cells by at least two pathways: inhibition of apoptosis via CEMIP and activation of STAT3 via IL23R. The authors also showed that DPF3-overexpressing cells showed higher T-cell mediated cytotoxicity compared to controls. In a separate effort, Protze et al.<sup>86</sup> used 23 tumor tissue specimens and two primary ccRCC cell lines and showed that the risk SNP is located within an active enhancer region, in turn creating a EPAS1-binding motif. They also showed that HIF-mediated DPF3 regulation depends on the presence of the C-risk allele. Moreover, DPF3 deletion in proximal tubular cells decreased cell growth, suggesting a role for DPF3 in cell proliferation.

Our work has several limitations. First, experimental validation was limited and performed on two master TFs in one ccRCC cell line 786-O. Future work focused on validation of the role of other candidate master TFs in lineage plasticity and across multiple RCC cell lines is warranted. Second, it is important to highlight that type 1 and type 2 papillary RCC are distinctive entities, and that although our work can globally highlight differences across RCC histotypes, the small sample size of the pRCC cohort limited more granular assessment by pRCC types.

In conclusion, this work combines and integrates ChIP-Seq, ATAC-seq, and RNA-seq data into a unified analysis to capture the deep complexity of the epigenomic landscape of RCC. The majority of epigenetic analyses have been conducted in immortalized or malignant cancer cell lines<sup>10,48</sup>, which inevitably have diverged significantly in many respects from primary tumor tissues, as seen in prior studies from medulloblastoma and diffuse large B-cell lymphoma<sup>26,65</sup>. In addition, studies have demonstrated that the vast majority of available RCC cell lines are derived from ccRCC, with very few pRCC lines, and relative absence of any chRCC lines<sup>87,88</sup>. To date, most of our knowledge in the epigenomic space of RCC primary tissue is derived from older DNA methylation analysis<sup>6–8</sup>, limited to one specific subtype of RCC<sup>37</sup>, or confined to only one epigenetic dimension<sup>61</sup>. By integrating several epigenetic technologies, we discovered master TFs that drive histology-specific transcriptional programs in RCC and showed that regulatory elements with allelic imbalance are specifically enriched for RCC GWAS risk SNPs relative to GWAS risk SNPs for other disease features. Overall, our work highlights the histology-specific nature of these regulatory elements and the opportunity of cataloging the epigenetic landscapes in expanded patient cohorts and across various tumor types.

## Methods

### Patients and samples

All tumor samples at DFCI were obtained at the time of resection at Brigham and Women's Hospital, and were collected under a DFCI/Harvard Cancer Center IRB-approved protocol (01-045) with the written informed consent of all patients. No monetary compensation was offered for patient participation. The samples were then stored fresh and frozen in the Gelb Center biobank at the Dana-Farber Cancer Institute, under a protocol approved by the MGB Institutional Review Board. Patients in the CheckMate 009/010/025 clinical trials consented to an institutional review board (IRB) approved protocol to participate in the respective clinical trials and to have their samples collected for tumor and germline sequencing.

The two patient cohorts from DFCI are defined in Supplementary Data 1. Cohort 1 was used to define the regulatory landscape across the different RCC histologies. On the other hand, cohort 2, was used primarily for the Allelic Imbalance analysis to increase the power of our

statistical analysis. In cohort 1, 28 of 30 (93%) tumor samples were obtained at radical nephrectomy, and 2 of 30 (7%) were from metastatic sites (Supplementary Data 1). Cohort 2 was used for the allelic imbalance, which consisted of 14 patients from cohort 1 in addition to 12 additional independent patients with ccRCC. Overall, 42 total samples were profiled (Supplementary Fig. S1). All specimens were reviewed by pathologists with expertise in genitourinary malignancies (TD, AF, SM, AA, MH, SS) to confirm the diagnosis, histological subtype, tumor grade, and stage (Supplementary Data 1). An additional cohort of 43 samples derived from 23 patients with ccRCC was characterized to study the effect of EPAS1 binding at the GWAS risk variant rs4903064 using EPAS1 ChIP-Seq (Supplementary Data 1).

### Chromatin immunoprecipitation for histone and TF marks

We performed chromatin immunoprecipitation (ChIP) for histone marks (H3K27ac and H3K4me2) in primary human tumors<sup>89</sup>. Briefly, 20–30 mg of frozen tissue was pulverized using the CryoPREP dry impactor system (Covaris). The tissue was then fixed using 1% formaldehyde (Thermo Fisher) in PBS for 10 min at room temperature and then quenched with 125 mM glycine. The tissue was then lysed in ice-cold lysis buffer (50 mM Tris, 10 mM EDTA, 1% SDS with protease inhibitor), and the chromatin was sheared to 300–800 base pair using the Covaris E220 sonicator (105-watt peak incident power, 5% duty cycle, 200 cycles/burst). Five volumes of dilution buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl pH 8.1) were then added, and a portion was taken for DNA preparation. 5–20 µg DNA equivalent was then incubated with antibodies (H3K27ac, Diagenode, C15410196, 1 µg/IP dilution; H3K4me2, Diagenode, C15410035, 1 µg/IP dilution), along with protein A and protein G beads (Life Technologies) with constant mild shaking at 4 °C overnight. The beads were then washed three times each with Low-Salt Wash Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 7.5, 150 mM NaCl), High-Salt Wash Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 7.5, 500 mM NaCl), and LiCl Wash Buffer (10 mM Tris pH 7.5, 250 mM LiCl, 1% NP-40, 1% Na-Doc, 1 mM EDTA) and rinsed with TE buffer (pH 8.0) once. For EPAS1 (HIF2-α) ChIP (Abcam, ab199, 1:100 dilution), the same amount of tissue was additionally fixed with 0.2 mM DSG (Thermo Fisher, Catalogue number: 20593) for 25 min followed by 1% fixation with FA for 20 min. The rest of the protocol per above.

Sequencing libraries were generated from purified input and IP sample DNA using the ThruPLEX-FD Prep Kit (Rubicon Genomics). Libraries were sequenced using 150-base paired-end reads on an Illumina platform (Novogene). Data quality is shown in Supplementary Data 26.

### ATAC-seq

Briefly, 20 mg of frozen tissue were resuspended and dounced in 1000 µl of HB. Nuclei were filtered using a 70-µm Flowmi strainer, isolated using iodixanol density-gradient centrifugation method, and washed with RSB buffer<sup>61</sup>. In all, 50,000 nuclei were resuspended in 50 µl of transposition mix (2.5 µl transposase (100 nM final), 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, and 5 µl water) by pipetting up and down six times. Transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1000 r.p.m. Reactions were cleaned with Qiagen columns. Libraries were amplified and prepared using the Omni-ATAC protocol<sup>90,91</sup> and sequenced on an Illumina platform (Novogene) using 150-base paired-end reads. Data quality is shown in Supplementary Data 26.

### ChIP-seq data analysis

ChIP-sequencing reads were demultiplexed using Illumina bcl2fastq v2.18 and aligned to the human genome build hg19 using the Burrows-Wheeler Aligner (BWA) version 0.7.15<sup>92</sup>. Non-uniquely mapping and redundant reads were discarded. MACS v2.1.1.20140616<sup>93</sup> was used for

ChIP-seq peak calling with a  $q$ -value (FDR) threshold of 0.01. ChIP-seq data quality was evaluated by a variety of measures, including total peak number, FrIP (fraction of reads in peak) score, number of high-confidence peaks (enriched >tenfold over background), and percent of peak overlap with DHS peaks derived from the ENCODE project. ChIP-seq peaks were assessed for overlap with gene features and CpG islands using *annotatr*<sup>94</sup>. *IGV*<sup>95</sup> was used to visualize normalized ChIP-seq read counts at specific genomic loci. ChIP-seq heatmaps were generated with *deepTools*<sup>96</sup> and show normalized read counts at the peak center  $\pm 2$  kb unless otherwise noted. Overlap of ChIP-seq peaks was assessed using *BEDTools*. Peaks were considered overlapping if they shared one or more nucleotides.

### Annotation of histology-specific enriched ChIP-seq peaks

Sample-sample clustering, principal component analysis, and identification of lineage-enriched peaks were performed using *Mapmaker* (<https://bitbucket.org/cfce/mapmaker>), a ChIP-seq analysis pipeline implemented with *Snakemake*<sup>97</sup>. ChIP-seq data from ccRCC, chRCC, and pRCC were compared to identify H3K27ac and H3K4me2 peaks with significant enrichment in either histology. A union set of peaks for each histone modification was created using *BEDTools* v2.26.0. narrowPeak calls from *MACS* were used for H3K27ac and H3K4me2. The number of unique aligned reads overlapping each peak in each sample was calculated from BAM files using *BEDTools*. Read counts for each peak were normalized to the total number of mapped reads for each sample. Quantile normalization was applied to this matrix of normalized read counts. Using *DESeq2*<sup>98</sup>, lineage-enriched peaks were identified at the indicated FDR-adjusted  $P$  value ( $P_{\text{adj}}$ ) and log<sub>2</sub> fold-change cutoffs (H3K27ac and H3K4me2,  $P$ -adjusted <0.001, log<sub>2</sub> fold change >3). Unsupervised hierarchical clustering was performed based on the Spearman correlation between samples. Principal component analysis was performed using the *prcomp* R function. Enriched de novo motifs in differential peaks were detected using *HOMER* version 4.7. The top non-redundant motifs were ranked by  $P_{\text{adj}}$ . The *GREAT* tool<sup>33</sup> was used to assess for enrichment of Gene Ontology (GO) and MSigDB perturbation annotations among genes near differential ChIP-seq peaks, assigning each peak to the nearest gene within 500 kb.

### RNA methods

RNA was extracted using the Qiagen RNeasy Mini Kit (Cat No./ID: 74104) as recommended, from frozen tumor samples adjacent to those samples used for ChIP. RNA-seq libraries were constructed from 1  $\mu$ g total RNA using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit according to the manufacturer's protocol. Barcoded libraries were pooled and sequenced on the Illumina HiSeq 2500 generating 50-bp paired-end reads. FASTQ files were processed using the *VIPER* workflow<sup>99</sup>. Read alignment to human genome build hg19 was performed with *STAR*<sup>100</sup>. *Cufflinks* was used to assemble transcript-level expression data from filtered alignments<sup>101</sup>.

### DNA extraction and mutation analysis

DNA was either (1) extracted from OCT-frozen primary tissue using the Qiagen QIAamp DNA Mini Kit (Cat No./ID: 51304) ( $n=9$ ); or (2) extracted from tumor regions consisting of at least 20% tumor cells from unstained slides using the QIAamp DNA FFPE Tissue Kit (Qiagen) according to the manufacturer's instructions ( $n=19$ ). DNA quantification was performed by Nanodrop and Pico-Green assays.

Mutational analysis was performed using our institutional CLIA-certified targeted panel sequencing, *Oncopanel*<sup>102</sup> without normal DNA analysis. Called variants were excluded if observed at a frequency  $\geq 0.1\%$  in the Exome Aggregation Consortium (ExAC) database<sup>103</sup>, as they were considered likely germline variants. Loss-of-function variants were defined as nonsense mutations, frameshift insertions or deletions, or splice site alterations affecting consensus nucleotides (Supplementary Data 3). Missense mutations were only

included if deemed pathogenic by (1) both *SIFT*<sup>104</sup> and *Polyphen-2*<sup>105</sup> and (2) reported as oncogenic or likely oncogenic in *cBioPortal* (TCGA Pan-cancer Atlas studies)<sup>106,107</sup>. CNVs were called by *Oncopanel*, and here we retained only homozygous deletions (Supplementary Data 4).

### Histology-specific expression of TFs

Normalized gene expression was computed for all 1631 known TFs<sup>108,109</sup> in all 28 RCC samples with RNA-Seq data (11 chRCC, 11 ccRCC, and 6 pRCC). To identify TFs whose expression was significantly higher in one histology than another, we compared the normalized TF expression using the Wilcoxon rank-sum two-sample one-sided test. TFs with an average normalized expression less than 10 FPKM in the histotype with the highest expression were excluded. Multiple hypothesis correction was done using Benjamini-Hochberg<sup>102</sup>. Unsupervised clustering of the samples using the 105 differentially expressed TFs, that had log<sub>2</sub> fold change  $\geq 1$ , FDR < 0.1, and minimum of 10 transcripts per million was performed to separate the RCC samples according to histology (Supplementary Fig. S5A and Supplementary Data 5 and 6).

### Cancer core transcription-factor specificity algorithm

Cancer Core Transcription factor Specificity (CaCTS) method is developed and described in detail in ref. <sup>110</sup>, which compares the expression level of a TF in each RCC histotype with all other cancer types. The CaCTS algorithm has been previously developed as an R package and has been deposited in GitHub (<https://github.com/lawrenson-lab/CaCTS>) and Zenodo (<https://doi.org/10.5281/zenodo.5234007>).

Briefly, we used pan-cancer RNA-sequencing data from The Cancer Genome Atlas to identify RCC histology-specific candidate master TFs. Considering a set of transcription factors detailed previously<sup>108,109</sup>, we got a list of 1671 unique TFs. We selected only the ones expressed in the TCGA pan-cancer dataset<sup>49</sup>. CaCTS score defines cell-specific TFs for each cancer type based on the RNA expression information. The specificity of expression of each TF, or "CaCTS score", was calculated by comparing its expression level in the query tumor type to that in the remaining TCGA tumor types. A high CaCTS score is therefore assigned to factors with high-level expression in the query tumor types as compared to background dataset. The output of the CaCTS algorithm is a list of all TFs ranked by CaCTS scores in each of the three RCC subtypes. The main procedure consists in a Jensen-Shannon Divergence (JSD) score. We first adjusted the normalized expression values, to handle negative values, by shifting the values to 0 and the maximum expression value. Then, to define a representative sample to each tumor type, we applied the mean of the expression values considering each tumor type individually. Mean values of 0 were replaced with  $1 \times 10^{-17}$ . We followed the same principles described by D'Alessio et al.<sup>13</sup> aiming to quantify the divergence between the transcription-factor expression across different cell types. We created two same-sized vectors to represent the observed pattern and the "ideal" pattern. For the observed pattern, the vector was formed by values from the expression mean profiles of the query cell type and the background dataset. Each element in this vector were divided by the sum of all elements. For the "ideal" pattern, the vector was formed by a value of 1 at the position equivalent to that of the query cell type and zeroes at all other positions. For example, considering ccRCC as query and the other 33 cancer types as background, the position 1, that corresponds to ccRCC position, was formed by a value 1 and the other positions were formed by value 0. From these two vectors, the JSD function was performed with the R package *jsd* version 0.1 and a cell-type-specificity score was obtained for each TF. The candidate list for each cancer type was defined by considering the top 5% of high JSD and top 5% of ranked expression mean. This analysis identified 36 TFs that were highly and uniquely expressed in one of the three RCC histotypes in comparison

to TCGA cancers (Fig. 3B, Supplementary Fig. S5B–S5D, and Supplementary Data 7).

### Superenhancer rank analysis

Super-enhancers (SEs) and typical enhancers (TEs) were identified using ROSE2 (<https://pyipi.org/project/rose2/>)<sup>17</sup> (Supplementary Fig. S9 and Supplementary Data 8). Both SEs and TEs were ranked based on integrated ChIP-seq signal<sup>17</sup>, where the SE with rank 1 has the highest ChIP-seq signal. SEs were then associated to the nearest gene, giving an SE rank score for each gene<sup>17</sup>. For each transcription-factor-sample pair, we identified the minimum TF SE rank for all SEs associated with that TF, or assigned the maximum SE rank found in all samples (1477) if the TF was not associated to any SE in that sample. TFs with an average superenhancer rank greater than 1000 were not considered further. We then compared the SE ranks for each TF and histology against the other histologies using the Wilcoxon rank-sum two-sample one-sided test. Multiple hypothesis correction was done using the Benjamini–Hochberg Procedure, and  $FDR \leq 0.10$  was considered significant. In all, 4073 SE region-associated genes were detected in at least one RCC sample, including 377 TFs. Unsupervised hierarchical clustering of the SE-associated TFs partitioned the samples according to their histology (Supplementary Fig. S5E) as did similar analysis of the 119 TFs with differential SE rank by histology ( $FDR < 0.1$ ) (Supplementary Fig. S5F and Supplementary Data 9).

### ATAC-seq normalization

Publicly available ATAC-seq data from 16 ccRCCs and 34 pRCCs<sup>61</sup>, and newly generated ATAC-seq data for seven chRCC samples were integrated and normalized. chRCC ATAC-seq peaks sets were normalized<sup>111</sup> by computing a “score per million” (spm), which is the individual peak score ( $-\log_{10}(P \text{ value})$ ) by the total sum of all peak scores in the sample multiplied by a million, and peaks with  $\text{spm} \leq 5$  were filtered out. Second, to generate the chRCC-specific ATAC-seq peak set, we conserved regions covered by at least two samples, breaking up peaks if necessary. Lastly, we removed regions that overlap repeat regions considering the UCSC Table browser (<http://genome.ucsc.edu/cgi-bin/hgTables>) “Repeats” group and “Repeatmasker” track. Only regions that contain the full length as  $N$  were removed.

### Clique enrichment score analysis

SE-associated TFs are typically wired in auto-regulatory networks, called “cliques”, that cooperatively regulate their own expression and expression of an extended network of genes to create the transcriptional fingerprint of the tissue<sup>65</sup>. To identify highly interconnected TFs, we integrated H3K27ac and ATAC-seq data and used Coltrons to identify cliques and determine the clique enrichment score (CES). CES for each TF was calculated using clique assignments performed using Coltrons<sup>112</sup>. Coltrons uses a motif-based approach employing both SE data and normalized ATAC-seq data to restrict the motif search to regions of open chromatin, to build networks of transcription factors. It identifies cliques of various sizes reflecting putative interaction between TFs in driving TF expression. The CES equals the fraction of cliques containing each expressed TF. Using the CES, we performed clustering analysis with the following parameters: distance = Canberra, agglomeration method = ward.D2. To select RCC histology-specific TFs using the CES, we restricted to a minimum average CES of 0.05, i.e., a TF has to be present, on average, in at least 5% of all cliques in a given RCC histological subtype. 167 TFs had a positive CES in one or more tumors, and unsupervised clustering of the samples based on TF CES separated samples by histology (Supplementary Fig. S5G). In total, 87 of the 167 TFs (52%) showed differential connectivity between the three histologies, meaning that on average, they participated in a significantly larger number of cliques in one histology compared to the other (one-sided Mann–Whitney  $U$  test of one histotype vs. the rest,

$FDR < 0.1$ , minimum CES = 0.05) (Supplementary Fig. S5H and Supplementary Data 10–11).

### Master TF analysis

These four analyses (Differential expression, CaCTS, SE rank, and CES) identified more than 200 candidate TFs showing a histology-specific association in one or more analyses. However, each analysis has potential limitations. Differential gene expression is not discriminative enough and is expected to have too many false positives. SE rank analysis relies heavily on enhancer-gene assignment, which is biased against long-distance interactions. CES analysis is a motif-based approach and depends on the prior knowledge of TF motifs, which are incompletely characterized. Finally, CaCTS is designed to compare pan-cancer TF expression, so it can miss TFs that are also important in other cancer types, which is why we used a meta-analysis approach where the TF has to be significant in one of the three analyses (CaCTS, SE rank, and CES) as well as significantly differentially expressed between the three histologies (Supplementary Data 12).

### Immunohistochemistry

For immunohistochemistry (IHC), 5-mm FFPE tissue sections were deparaffinized, rehydrated, and subjected to heat-induced antigen retrieval (0.01 mol/L sodium citrate tribasic dihydrate, pH 6.0 or EDTA retrieval buffer, pH 9.0) followed by treatment with 3%  $\text{H}_2\text{O}_2$  for 15 min at room temperature to block endogenous peroxidase activity. Sections were incubated overnight at 4 °C with primary antibodies against mouse HNF1 $\beta$  (Santa Cruz, sc-130407, 1:50 dilution), mouse BHLHE41 (ThermoFisher, TA806146, 1:200 dilution), rabbit NKX6.1 (Cell signaling, #54551, 1:50 dilution), rabbit ZNF395 (Lsbio, LS-B5647-100, 1:500 dilution). Sections were then stained with appropriate secondary antibodies per manufacturer recommendations (ImmPRESS HRP anti-rabbit and anti-mouse IgG Polymer detection kits: MP-7451; MP-7402, Vector laboratories). Slides were counterstained with hematoxylin, dehydrated, and mounted. Slides were scanned at  $\times 20$  magnification and positively stained tumor cells for HNF1 $\beta$ , BHLHE41, NKX6.1, or ZNF395 were determined.

### Cell lines

786-O cell line was obtained from a certified commercial vendor (ATCC, CRL-1932) and grown in DMEM (Invitrogen) supplemented with 10% fetal bovine serum (Sigma), 100 U/mL penicillin, and 100  $\mu\text{g}/\text{mL}$  streptomycin (Invitrogen). No further authentication was performed. Cell lines tested negative for mycoplasma contamination.

### FOXII overexpression in 786-O cell line (FOXII OE)

The full-length *FOXII* open reading frame (ORF) without stop codon was amplified by gene-specific cloning primers (Supplementary Data 27) using cDNA from an RCC cell line which expressed FOXII endogenously (Obtained from L. H.’s lab) and cloned into mammalian expression lentiviral vector (pFUGW) derivative (Addgene #52962) AgeI and BamHI sites. Correct *FOXII* ORF sequence was verified by Sanger sequencing, expression level was measured by qRT-PCR and correct protein size was verified by immunoblot technique. Stable FOXII overexpressing cell lines were created by lentiviral transduction using the FOXII containing transfer vector and pMD2.G (Addgene #12259) and psPAX (Addgene#12259) lentiviral packaging vectors. Transduced cells were selected and maintained under blasticidin containing (10  $\mu\text{g}/\text{mL}$ ) selective culture media.

### EPAS1 shRNA knockdown in 786-O cell line (EPAS1 KD)

Short hairpin RNAs (shRNA) were designed against the FOXII 3' UTR region using the Broad Gene Perturbation Web Portal (GPP): <https://portals.broadinstitute.org/gpp/public/> and non-human targeting control (NTC) gRNA was included and cloned (detailed protocol available at <http://www.broadinstitute.org/rnai/public/resources/protocols>).

Briefly, for each shRNA (Supplementary Data 27) complementary single-stranded oligonucleotides were synthesized (Invitrogen) with cloning sites, phosphorylated, annealed, and shRNA cassettes were ligated into pLKO.1 (Addgene #8453) shRNA expression vector containing puromycin selection marker. After bacterial transformation (Stbl3, Invitrogen) individual clones were picked and regenerated, and correct shRNA sequences were verified by Sanger sequencing. Suppression effects were tested by qRT-PCR and immunoblot techniques.

Lentiviral particles were generated for each shRNA experiment by transforming 786-O cells with shRNA transfer vectors and lentiviral packaging mix (pMD2.G psPAX). Lentiviral particle containing media were collected and filtered using 0.45- $\mu$ m pore size syringe filters (Corning) after 48 h post transfection and used for treatment of previously plated 786-O cell line. Media was changed after 24 h and replaced by puromycin (2  $\mu$ g/ml) containing selective culture medium. Puromycin-resistant cells were maintained and collected after 3 days of selection and total RNA was isolated using RNeasy Mini kit (Qiagen).

### Creation of double manipulated cell lines (FOX11 OE/EPAS1 KD)

786-O/FOX11 OE cells were cultured and maintained in DMEM media as described above. Cells from the two different conditions were seeded in parallel six-well plates and 24 h later, cells were infected with lentivirus containing shRNA targeting EPAS1 and negative control shRNA. Cells were selected by puromycin for 72 h following infection, equal cell numbers were harvested for RNA extraction and immunoblotting.

### Gene expression analyses

For targeted gene expression, qRT-PCR 500 ng total RNA (Qiagen) was reverse transcribed (High Capacity Reverse transcription kit, Life-Technologies) and cDNA was diluted (20 $\times$ ). SYBR Green assay was performed on Light Cycler 480 instrument (2 $\times$  Probe Master Mix, Roche). All primer sequences are listed in Supplementary Data 27. Relative gene expression was calculated based on the ddCT method (Pfaffl 2001). Each sample was measured by two technical replicates. *ACTB* gene was used as housekeeping genes to normalize the samples.

For gene expression profile analysis, RNA-Seq was performed from the different conditions using poly A capture method from total RNA by Novogene sequencing service.

### Chromatin and gene expression allelic imbalance analysis

DNA samples from three pRCC and ten ccRCC samples were genotyped on a commercially available genotyping array (Infinium Global Screening Array-24, version 1.0; Illumina) at the Broad Institute Genomic Services, Cambridge, Massachusetts. We used a total of 642,824 SNP markers. H3K27ac ChIP-seq reads from the same 10 ccRCC and 3 pRCCs were analyzed for the imbalance of heterozygous SNP alleles using stratAS (<https://github.com/gusevlab/stratAS>; Gusev et al., submitted; <https://www.biorxiv.org/content/10.1101/631150v1.full.pdf>, deposited in Zenodo (<https://doi.org/10.5281/zenodo.7373647>), hereby denoted chromatin allelic imbalance. Several upstream steps were performed to boost power and accuracy of chromatin allelic imbalance detection.

First, genotypes were imputed from SNP array data with STITCH<sup>113</sup>, using the 1000 Genomes Phase 3 reference panel. Subsequent analyses were limited to SNPs with allele fraction >5% in the Haplotype Reference Consortium version 1.1<sup>102</sup>. Genotypes were then phased with Eagle2 using the Sanger Imputation Service (<https://imputation.sanger.ac.uk/>). Heterozygous SNPs were filtered for mapping bias using the WASP pipeline<sup>114</sup> and allele-specific read counts were tabulated using ASEReadCounter from the Genome Analysis Toolkit v3.8<sup>115</sup>.

Briefly, stratAS identifies chromatin allelic imbalance—and differential allelic imbalance between two-sample groups—by modeling the

reads from heterozygous SNPs with a beta-binomial distribution (<https://doi.org/10.1101/631150v1.full.pdf>). stratAS takes advantage of haplotype phasing to sum read counts from heterozygous SNPs from each phased haplotype for each individual within a given ChIP-seq peak. stratAS models the reads from individual *i* overlapping heterozygous germline SNP *j* as:  $R_{alt,i} | R_{ref,i} \sim \text{BetaBin}(\pi_j, \rho_j)$ , where  $\pi$  is the mean allelic ratio and  $\rho$  is a locally-defined, per-individual sequence read correlation parameter reflecting overdispersion.

We tested differential chromatin imbalance between ccRCC and pRCC by the likelihood ratio test between the models  $\pi_{clearcell,j} = \pi_{pRCC,j}$  and  $\pi_{clearcell,j} \neq \pi_{pRCC,j}$ , maximizing the likelihood of each model by a standard 1-dimensional optimization.

Copy number profiles were estimated from read representation in ChIP-seq inputs using qdnaseq v1.18. Copy number profiles were used in the modeling of the overdispersion parameter  $\rho$ . This is motivated by larger degrees of overdispersion in regions of cancer-associated copy number alterations.  $\rho$  is estimated for each individual from all heterozygous read-carrying SNPs across ten declines of estimated copy number levels stratAS params.R script, with the following options: `-min_snps 50, -min_cov 5, -group_snp TRUE, -group 10`.

We tested variants within a consensus set of H3K27ac peaks. This peak set was derived by dividing the genome into 50 bp windows and including any window with peaks in two or more samples. We restricted tests to peaks with at least one read supporting each haplotype (`-min_cov 1`). The following additional parameters were set for the stratAS script: `-max_rho 0.2, -window -1, -min_cov 1, and -fill_cnv TRUE`. *P* values for allelic imbalance, and differential imbalance between ccRCC and pRCC, were corrected for multiple hypothesis testing as follows. First, all SNPs within a given peak were adjusted by Bonferroni correction. The corrected *P* value of the most significant SNP was taken as the *P* value for imbalance of the peak. We then obtained FDR-adjusted *q*-values for all peaks using the *q*-value R package (v4.0.1). Peaks with *q* < 0.05 were considered significantly imbalanced. Differentially imbalanced peaks were grouped into categories for motif enrichment analysis. Peaks with ccRCC-specific chromatin imbalance were defined as those with (1) a significant imbalance in ccRCC, (2), significant differential imbalance between ccRCC and pRCC, and (3), no significant imbalance in the stratAS analysis of pRCC alone. Conversely, a set of peaks with pRCC-specific imbalance was defined by (1) chromatin imbalance in pRCC, (2) differential chromatin imbalance between the two groups, and (3), no significant chromatin imbalance in ccRCC. These two groups were analyzed for motif enrichment using homer v4.7. The background for enrichment was specified as all peaks imbalanced in pRCC and in ccRCC in the separate stratAS analyses of these groups. The top three “known results” motif categories were reported.

The union set of chromatin-imbalanced peaks (i.e., in pRCC, ccRCC, or in the combined or differential comparisons) were assessed for enrichment of significant SNPs from genome-wide association studies (GWAS) of RCC<sup>58</sup>, prostate cancer (6), psoriasis<sup>116</sup>, cardiovascular disease, diagnosis of hypertension<sup>117</sup>, self-reported hypercholesterolemia, type 2 diabetes mellitus<sup>118</sup>, self-reported hypothyroidism<sup>119</sup>, all autoimmune diseases, dermatologic disorders, diagnosis of asthma, respiratory disease, and diagnosis of allergy or eczema. For each GWAS, the number of genome-wide significant SNPs overlapping chromatin-imbalanced H3K27ac peaks was counted and divided by the number of base pairs covered by these peaks. The same process was performed with a background set of peaks randomly sampled over 10,000 iterations from all RCC H3K27ac peaks. The relative overlap of chromatin-imbalanced peaks was divided by the relative overlap of background H3K27ac peaks to estimate relative enrichment. Enrichment was also calculated relative to random peaks, matched to the chromatin-imbalanced peaks for number, peak size, and chromosome. Empiric one-sided *P* values were calculated from the iterations.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Sequencing reads are aligned to the human genome build hg19. The DFCI datasets generated in this study, including ChIP-seq, RNA-seq, and ATAC-seq data (R0929TAT, R0957TAT, R0998TAT, R1097TAT, R1149TAT, R1152TAT, R991TAT) have been deposited in the Gene Expression Omnibus (GEO) database under accession code [GSE188486](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE188486). The TCGA Kidney Renal Clear Cell Carcinoma (KIRC) publicly available data used in this study are available in a public repository from the Broad Institute Firehose Pipeline (<http://gdac.broadinstitute.org>). All clinical and correlative sequencing data from publicly available. CheckMate 009/010/025 publicly available data used in this study are available in the European Genome-Phenome Archive database under accession codes [EGAS00001004291](https://www.ebi.ac.uk/ena/record/EGAS00001004291) and [EGAS00001004292](https://www.ebi.ac.uk/ena/record/EGAS00001004292)<sup>120</sup>. ATAC-Seq data publicly available data used in this study are available in the Genomic Data Commons database (<https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>)<sup>61</sup>. The remaining data are available within the Article or Supplementary Information. Any other queries about the data used in this study should be directed to the corresponding authors of this study. Source data are provided with this paper.

## Code availability

Algorithms used for data analysis are all publicly available from the indicated references in the Methods section.

## References

- Society, A. C. *Cancer Facts & Figures 2020* (American Cancer Society, Atlanta, 2020).
- Hsieh, J. J. et al. Renal cell carcinoma. *Nat. Rev. Dis. Prim.* **3**, 17009 (2017).
- Ricketts, C. J. et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* **23**, 3698 (2018).
- Cohen, H. T. & McGovern, F. J. Renal-cell carcinoma. *N. Engl. J. Med.* **353**, 2477–2490 (2005).
- Jonasch, E., Gao, J. & Rathmell, W. K. Renal cell carcinoma. *BMJ* **349**, g4797 (2014).
- Cancer Genome Atlas Research Network Analysis Working Group. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
- Cancer Genome Atlas Research Network Analysis Working Group. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
- Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
- Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
- Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
- Buganim, Y., Faddah, D. A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14**, 427–439 (2013).
- D'Alessio, A. C. et al. A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep.* **5**, 763–775 (2015).
- Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587–594 (2009).
- Morris, S. A. & Daley, G. Q. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Res.* **23**, 33–48 (2013).
- Sancho-Martinez, I., Baek, S. H. & Izpisua Belmonte, J. C. Lineage conversion methodologies meet the reprogramming toolbox. *Nat. Cell Biol.* **14**, 892–899 (2012).
- Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Parker, S. C. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
- Neph, S. et al. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 e1324 (2016).
- Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
- Grubert, F. et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
- Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).
- Waszak, S. M. et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
- Chapuy, B. et al. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**, 777–790 (2013).
- Russ, B. E. et al. Regulation of H3K4me3 at transcriptional enhancers characterizes acquisition of virus-specific CD8(+) T cell-lineage-specific function. *Cell Rep.* **21**, 3624–3636 (2017).
- Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why. *Mol. Cell* **49**, 825–837 (2013).
- Traykova-Brauch, M. et al. An efficient and versatile system for acute and chronic modulation of renal tubular function in transgenic mice. *Nat. Med.* **14**, 979–984 (2008).
- Bleu, M. et al. PAX8 activates metabolic genes via enhancer elements in renal cell carcinoma. *Nat. Commun.* **10**, 3739 (2019).
- Tong, G. X. et al. Expression of PAX8 in normal and neoplastic renal tissues: an immunohistochemical study. *Mod. Pathol.* **22**, 1218–1227 (2009).
- Gusev, A. et al. Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in renal cell carcinoma. Preprint at <https://www.biorxiv.org/content/10.1101/631150v1.full.pdf> (2019).
- McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- Ricketts, C. J. et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* **23**, 313–326 e315 (2018).
- Kovacs, G. Molecular cytogenetics of renal cell tumors. *Adv. Cancer Res.* **62**, 89–124 (1993).
- Werth, M. et al. Transcription factor TFCP2L1 patterns cells in the mouse kidney collecting ducts. *eLife* **6**, e24265 (2017).
- Yao, X. et al. VHL deficiency drives enhancer activation of oncogenes in clear cell renal cell carcinoma. *Cancer Discov.* **7**, 1284–1305 (2017).

38. Choueiri, T. K. et al. Inhibition of hypoxia-inducible factor-2alpha in renal cell carcinoma with belzutifan: a phase 1 trial and biomarker analysis. *Nat. Med.* **27**, 802–805 (2021).
39. Choueiri, T. K. & Kaelin, W. G. Jr. Targeting the HIF2-VEGF axis in renal cell carcinoma. *Nat. Med.* **26**, 1519–1530 (2020).
40. Heliot, C. et al. HNF1B controls proximal-intermediate nephron segment identity in vertebrates by regulating Notch signalling components and *Irx1/2*. *Development* **140**, 873–885 (2013).
41. Chen, F. et al. Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep.* **14**, 2476–2489 (2016).
42. Sato, Y. et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
43. Dalgliesh, G. L. et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360–363 (2010).
44. Varela, I. et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542 (2011).
45. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
46. Boyer, L. A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
47. Ament, S. A. et al. New meta-analysis tools reveal common transcriptional regulatory basis for multiple determinants of behavior. *Proc. Natl Acad. Sci. USA* **109**, E1801–E1810 (2012).
48. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
49. Reddy, J. et al. Predicting master transcription factors from pan-cancer expression data. <https://doi.org/10.1101/839142> (2019).
50. Lindgren, D. et al. Cell-type-specific gene programs of the normal human nephron define kidney cancer subtypes. *Cell Rep.* **20**, 1476–1489 (2017).
51. Shen, C. & Kaelin, W. G. Jr. The VHL/HIF axis in clear cell renal carcinoma. *Semin. Cancer Biol.* **23**, 18–25 (2013).
52. Liu, L. et al. Immunohistochemical analysis of chromophobe renal cell carcinoma, renal oncocytoma, and clear cell carcinoma: an optimal and practical panel for differential diagnosis. *Arch. Pathol. Lab. Med.* **131**, 1290–1297 (2007).
53. Mathers, M. E., Pollock, A. M., Marsh, C. & O'Donnell, M. Cytokeratin 7: a useful adjunct in the diagnosis of chromophobe renal cell carcinoma. *Histopathology* **40**, 563–567 (2002).
54. Baca, S. C. et al. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *Nat. Genet.* **54**, 1364–1375 (2022).
55. Benaglio, P. et al. Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nat. Genet.* **51**, 1506–1517 (2019).
56. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
57. Wang, A. T. et al. Allele-specific QTL fine mapping with PLASMA. *Am. J. Hum. Genet.* **106**, 170–187 (2020).
58. Scelo, G. et al. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* **8**, 15724 (2017).
59. Colli, L. M. et al. Altered regulation of DPF3, a member of the SWI/SNF complexes, underlies the 14q24 renal cancer susceptibility locus. *Am. J. Hum. Genet.* **108**, 1590–1610 (2021).
60. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
61. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
62. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
63. Bigot, P. et al. Functional characterization of the 12p12.1 renal cancer-susceptibility locus implicates BHLHE41. *Nat. Commun.* **7**, 12098 (2016).
64. Grampp, S. et al. Genetic variation at the 8q24.21 renal cancer susceptibility locus affects HIF binding to a MYC enhancer. *Nat. Commun.* **7**, 13183 (2016).
65. Lin, C. Y. et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* **530**, 57–62 (2016).
66. Szponar, A., Yussenko, M. V., Kuiper, R., van Kessel, A. G. & Kovacs, G. Genomic profiling of papillary renal cell tumours identifies small regions of DNA alterations: a possible role of HNF1B in tumour development. *Histopathology* **58**, 934–943 (2011).
67. Tang, Z. et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
68. Anderson, M. K., Hernandez-Hoyos, G., Diamond, R. A. & Rothenberg, E. V. Precise developmental regulation of Ets family transcription factors during specification and commitment to the T cell lineage. *Development* **126**, 3131–3148 (1999).
69. Iwasaka, C., Tanaka, K., Abe, M. & Sato, Y. Ets-1 regulates angiogenesis by inducing the expression of urokinase-type plasminogen activator and matrix metalloproteinase-1 and the migration of vascular endothelial cells. *J. Cell Physiol.* **169**, 522–531 (1996).
70. Leprince, D. et al. A putative second cell-derived oncogene of the avian leukaemia retrovirus E26. *Nature* **306**, 395–397 (1983).
71. Nunn, M. F., Seeburg, P. H., Moscovici, C. & Duesberg, P. H. Tripartite structure of the avian erythroblastosis virus E26 transforming gene. *Nature* **306**, 391–395 (1983).
72. Dittmer, J. The biology of the Ets1 proto-oncogene. *Mol. Cancer* **2**, 29 (2003).
73. Hahne, J. C. et al. The transcription factor ETS-1: its role in tumour development and strategies for its inhibition. *Mini Rev. Med. Chem.* **8**, 1095–1105 (2008).
74. Kim, E. et al. Tumor suppressor p53 inhibits transcriptional activation of invasion gene thromboxane synthase mediated by the proto-oncogenic factor ets-1. *Oncogene* **22**, 7716–7727 (2003).
75. Davidson, B. et al. Ets-1 messenger RNA expression is a novel marker of poor survival in ovarian carcinoma. *Clin. Cancer Res.* **7**, 551–557 (2001).
76. Span, P. N. et al. Expression of the transcription factor Ets-1 is an independent prognostic marker for relapse-free survival in breast cancer. *Oncogene* **21**, 8506–8509 (2002).
77. Elvert, G. et al. Cooperative interaction of hypoxia-inducible factor-2alpha (HIF-2alpha) and Ets-1 in the transcriptional activation of vascular endothelial growth factor receptor-2 (Flk-1). *J. Biol. Chem.* **278**, 7520–7530 (2003).
78. Montagner, M. et al. SHARPI suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. *Nature* **487**, 380–384 (2012).
79. Freedman, M. L. et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* **43**, 513–518 (2011).
80. Ricano-Ponce, I. & Wijmenga, C. Mapping of immune-mediated disease genes. *Annu Rev. Genomics Hum. Genet.* **14**, 325–353 (2013).
81. Hindorf, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
82. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
83. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

84. Reshef, Y. A. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).
85. Link, V. M. et al. Analysis of genetically diverse macrophages reveals local and domain-wide mechanisms that control transcription factor binding and function. *Cell* **173**, 1796–1809 e1717 (2018).
86. Protze, J. et al. The renal cancer risk allele at 14q24.2 activates a novel hypoxia-inducible transcription factor-binding enhancer of DPF3 expression. *J. Biol. Chem.* **298**, 101699 (2022).
87. Brodaczewska, K. K., Szczylik, C., Fiedorowicz, M., Porta, C. & Czarnecka, A. M. Choosing the right cell line for renal cell cancer research. *Mol. Cancer* **15**, 83 (2016).
88. Sinha, R. et al. Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nat. Commun.* **8**, 15165 (2017).
89. Pomerantz, M. M. et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).
90. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21–29 (2015).
91. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
92. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
93. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
94. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
95. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
96. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
97. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
98. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
99. Cornwell, M. et al. VIPER: visualization pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinforma.* **19**, 135 (2018).
100. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
101. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
102. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
103. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
104. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
105. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
106. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
107. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).
108. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
109. Saint-Andre, V. et al. Models of human core transcriptional regulatory circuitries. *Genome Res.* **26**, 385–396 (2016).
110. Reddy, J. et al. Predicting master transcription factors from pan-cancer expression data. *Sci. Adv.* **7**, eabf6123 (2021).
111. Siebenthal, K. T. et al. Integrated epigenomic profiling reveals endogenous retrovirus reactivation in renal cell carcinoma. *EBio-Medicine* **41**, 427–442 (2019).
112. Ott, C. J. et al. Enhancer architecture and essential core regulatory circuitry of chronic lymphocytic leukemia. *Cancer Cell* **34**, 982–995 e987 (2018).
113. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).
114. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
115. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
116. Tsoi, L. C. et al. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat. Commun.* **8**, 15382 (2017).
117. Giri, A. et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
118. Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).
119. Wellcome Trust Case Control, C. et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
120. Braun, D. A. et al. Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat. Med.* **26**, 909–918 (2020).

## Acknowledgements

We are grateful for all our patients treated at Dana-Farber Cancer Institute. This work is supported by the United States Department of Defense (Idea Award, W81XWH-19-1-0554—M.L. Freedman, M.M. Pomerantz, A. Nassar, S. Abou Alaiwi). M.L. Freedman is supported by the Claudia Adams Barr Program for Innovative Cancer Research, the H.L. Snyder Medical Research Foundation, the Cutler Family Fund for Prevention and Early Detection, the Donahue Family Fund, the Department of Defense Awards W81XWH-21-1-0234 (M.L.F.), W81XWH-21-1-0339 (M.L.F.), W81XWH-19-1-0554 (M.L.F.), NIH Awards R01CA251555 (M.L.F.), R01CA227237 (M.L.F.), and a PCF Challenge Award. T.K. Choueiri is supported in part by the Dana-Farber/Harvard Cancer Center Kidney SPOR (2P50CA101942-16) and Program 5P30CA006516-56 (T.K.C.), the Kohlberg Chair at Harvard Medical School and the Trust Family, Michael Brigham, Pan Mass Challenge, Hinda and Arthur Marcus Fund and Loker Pinard Funds for Kidney Cancer Research at DFCI. M.M. Pomerantz is also supported by John and Ann Hall; Rebecca and Nathan Milikowsky Family Foundation.

## Author contributions

A.H.N., S.A.A., T.K.C., D.J.K., and M.L.F. conceived the study; A.H.N., S.A.A., S.C.B., E.A., R.I.C., and M.A.S.F. analyzed the data under the joint supervision of D.J.K., M.L.F., and K.L.; A.H.N. and S.A.A. wrote the manuscript under the joint supervision of T.K.C., D.J.K., and M.L.F.; A.H.N. and S.A.A. performed ChIP-seq, ATAC-seq, RNA-seq, and SNP array experiments with assistance from T.E.Z. and J.S.; S.Spisak., V.T. performed shRNA experiments under the supervision of M.L.F. S.G. assisted with the genotype imputation; A.F., A.A., T.D., S.G.M., A.A.,

S. Signoretti, and M.S.H. reviewed the pathology slides; S.C.B., E.A., and A.G. analyzed the allelic imbalance data. D.A.B., H.D., M.H., S.S., Y.H., J.H., G.B., J.E.B., E.O., C.B., P.V.N., G.M.L., M.P., E.H., and M.A. assisted with procurement of clinical samples. All authors assisted in writing and revising the manuscript. T.K.C., D.J.K., and M.L.F. are co-senior authors for this work.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-35833-5>.

**Correspondence** and requests for materials should be addressed to Toni K. Choueiri, David J. Kwiatkowski or Matthew L. Freedman.

**Peer review information** *Nature Communications* thanks Chunliang Li, Xifeng Wu and Hao Zeng for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

<sup>1</sup>Department of Hematology/Oncology, Yale New Haven Hospital, New Haven, CT 06510, USA. <sup>2</sup>Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. <sup>3</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. <sup>4</sup>Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA. <sup>5</sup>Women's Cancer Research Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>6</sup>Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>7</sup>Center for Bioinformatics and Functional Genomics, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>8</sup>The Eli and Edythe L. Broad Institute, Cambridge, MA 02142, USA. <sup>9</sup>Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA. <sup>10</sup>Translational Immunogenomics Lab, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>11</sup>McGraw/Patterson Center for Population Sciences, Dana-Farber Cancer Institute, Boston, MA 02115, USA. <sup>12</sup>These authors contributed equally: Amin H. Nassar, Sarah Abou Alaiwi, Sylvan C. Baca. ✉e-mail: [toni\\_choueiri@dfci.harvard.edu](mailto:toni_choueiri@dfci.harvard.edu); [dk@rics.bwh.harvard.edu](mailto:dk@rics.bwh.harvard.edu); [mfreedman@partners.org](mailto:mfreedman@partners.org)