Genome-wide association study identifies multiple risk loci for renal cell carcinoma

Ghislaine Scelo et al.

Previous genome-wide association studies (GWAS) have identified six risk loci for renal cell carcinoma (RCC). We conducted a meta-analysis of two new scans of 5,198 cases and 7,331 controls together with four existing scans, totalling 10,784 cases and 20,406 controls of European ancestry. Twenty-four loci were tested in an additional 3,182 cases and 6,301 controls. We confirm the six known RCC risk loci and identify seven new loci at 1p32.3 (rs4381241, $P = 3.1 \times 10^{-10}$), 3p22.1 (rs67311347, $P = 2.5 \times 10^{-9}$), 3q26.2 (rs10936602, $P = 8.8 \times 10^{-9}$), 8p21.3 (rs2241261, $P = 5.8 \times 10^{-9}$), 10q24.33-q25.1 (rs11813268, $P = 3.9 \times 10^{-8}$), 11q22.3 (rs74911261, $P = 2.1 \times 10^{-10}$) and 14q24.2 (rs4903064, $P = 2.2 \times 10^{-24}$). Expression quantitative trait analyses suggest plausible candidate genes at these regions that may contribute to RCC susceptibility.
Kidney cancer is the seventh most commonly diagnosed cancer in more developed regions of the world and incidence rates have been rising\(^1,2\). Renal cell carcinoma (RCC) comprises over 90% of kidney cancers and clear cell renal cell carcinoma (ccRCC) is the major histological subtype (~80% of RCC cases)\(^3\). Direct evidence for inherited predisposition to RCC is provided by a number of rare cancer syndromes with defined germline mutations in 11 genes (BAP1, FLCN, FH, MET, PTEN, SDHB, SDHC, SDHD, TSC1, TSC2 and VHL), that are associated with the development of different RCC subtypes\(^4,5\). While identification of these genes has led to important insights into the pathogenesis of RCC\(^5,6\), even collectively these diseases account for only a very small portion of the twofold increased risk of RCC seen in first-degree relatives of RCC patients\(^7,8\). Support for polygenic susceptibility to RCC has come from genome-wide association studies (GWAS) that have identified single-nucleotide polymorphisms (SNPs) at six loci influencing RCC risk in populations of European ancestry at chromosome bands 2p21, 2q22.3, 8q24.21, 11q13.3, 12p11.23 and 12q24.31 (refs 9–14). Here, we present findings from a meta-analysis of six GWAS scans of RCC; two new scans of 5,198 cases and 7,331 controls were combined with four previously published scans of 5,586 cases and 13,075 controls\(^9–11\), reaching a total of 10,784 cases and 20,406 controls, all of European ancestry. Twenty-four promising loci were further tested in an independent replication set of 3,182 cases and 6,301 controls drawn from three independent series (Fig. 1).

**Results**

**Discovery-phase findings.** For both the GWAS and replication sets, cases were restricted to invasive RCC (International Classification of Disease for Oncology second and third Edition topography code C64), including all histological subtypes, diagnosed in adults (that is, ≥18 years) (Supplementary Table 1). Comparable sample and SNP quality control exclusions were applied to the two new genotyped scans (Supplementary Online methods), which used the OmniExpress and Omni5M arrays, respectively. The discovery phase was conducted as a fixed-effect meta-analysis that included these two new scans together with four previously published scans (IARC-1, NCI-1, MDA and UK). The four previously reported scans were conducted using HumanHap 300 and 610 for IARC-1; 500 and 660w for NCI-1; 660w for MDA; and OmniExpress and HumanHap 1.2M for UK. Imputations were performed on all scans using 1,094 subjects from the 1,000 Genomes Project (phase 1 release 3) as the reference panel (Supplementary Online methods). Each discovery-stage data set was analysed individually assuming log-additive (trend) SNP effects, with the exception of the two IARC scans which were pooled and analysed together (Supplementary Online methods). We then performed a fixed-effects meta-analysis of 7,437,091 SNPs that were polymorphic in at least two data sets. Quantile–quantile plots of the combined results showed little evidence for inflation of the test statistics compared to the expected distribution (λ = 1.034; Supplementary Fig. 1). For visual representation, we provide a Manhattan plot summarizing the genome-wide SNP results in Supplementary Fig. 2.

In the meta-analysis, we observed associations that surpassed the level of genome-wide significance for all six of the previously reported GWAS loci at 2p21, 2q22.3, 8q24.21, 11q13.3, 12p11.23 and 12q24.31 (Supplementary Table 2). We did not find evidence to support a previously suggested locus marked by rs3845536 at 1q24.1 (ref. 15) (meta-analysis P = 0.0062).

For replication, we selected 24 SNPs marking 20 possible new-risk regions, based on a P value <5.0 × 10^{-7}. We also included two SNPs at the known 2p21 RCC risk locus that were potentially independent from the previously reported genome-wide significant SNPs in that region\(^9,16\). Four additional SNPs representing four promising loci (one of which was among the 20 previously mentioned regions) were also advanced from an analysis restricted to ccRCC (5,649 cases, 15,011 controls) based on the aforementioned P value criterion (Supplementary Data 1). For genotyping these markers using Taqman assays, highly correlated proxy variants were substituted for 14 SNPs for which a Taqman assay could not be optimized; two proxies per variant were selected for two SNPs in the region where the smallest P values were found. Thus, a total of 32 SNPs from 24 regions were genotyped and passed quality control metrics in three independent series totalling 3,182 cases and 6,301 controls (Fig. 1, Supplementary Table 3, Supplementary Data 1, Supplementary Online methods).

**Seven new loci associated with RCC risk.** In the combined analysis, SNPs at seven loci showed evidence for an association with RCC which was genome-wide significant: 1p32.3 (rs4381241, \(P = 3.1 \times 10^{-10}\)), 3p22.1 (rs67311347, \(P = 2.5 \times 10^{-8}\)), 3q26.2 (rs10936602, \(P = 8.8 \times 10^{-9}\)), 8p21.3 (rs2241261, \(P = 5.8 \times 10^{-9}\)), 10q24.33-q25.1 (rs11813268, \(P = 3.9 \times 10^{-8}\)), 11q22.3 (rs74911261, \(P = 2.1 \times 10^{-10}\)) and 14q24.2 (rs4903064, \(P = 2.2 \times 10^{-24}\)) (Table 1, Supplementary Data 1). None of SNP associations showed evidence for study heterogeneity. Regional LD plots for each locus are detailed in Supplementary Fig. 3. Restricting the analyses to ccRCC, no additional SNPs with genome-wide significant associations were identified (Supplementary Data 1).

We conducted further analyses of the genome-wide significant SNPs stratifying by sex and three established RCC risk factors: body mass index, smoking and hypertension (Supplementary Data 1).
Fig. 4). The most notable difference in risk was observed for the 14q24 variants that had a stronger effect in women than in men [for rs4903064, odds ratios: ORs (95% confidence interval: CI) of 1.36 (1.28–1.45) and 1.13 (1.08–1.19), respectively; heterogeneity $P = 7.4 \times 10^{-5}$]. Other observed differences across strata were of smaller magnitude (Supplementary Fig. 4). No notable findings were observed in additional SNP analyses of non-clear cell histologic subtypes (papillary, chromophobe; Supplementary Data 1) and case age at onset (<60 versus 60+) (Supplementary Data 2). For SNP rs76912165, which was not genome-wide significant overall, a trend for higher risk associated with stage 1 cases was observed (Supplementary Data 2).

We investigated whether rs6706003 and rs6755594 defined independent signals at the previously reported 2p21 locus. rs6706003 is minimally correlated with rs7579899 ($r^2 = 0.11$ in CEU)17 that was identified in the initial GWAS, and moderately correlated with rs12617313 ($r^2 = 0.61$), which was identified in a previous fine-mapping analysis16. By comparison, the correlation of rs6755594 with both of these sites is notably weaker ($r^2 = 0.04$ and 0.08, respectively). In conditional analyses of the GWAS data adjusting for rs7579899 and rs12617313, the rs6706003 signal was substantially reduced (OR 1.07, $P = 0.05$), while the rs6755594 signal was partially attenuated (OR 1.07, $P = 4.0 \times 10^{-4}$). On the basis of these findings, there is insufficient evidence to conclude that rs6755594 marks an independent locus in this region.

**Newly identified loci and biological inferences.** To investigate plausible candidate variants and genes among the newly discovered loci for further study, we: (1) fine-mapped each locus, using 1000 Genome Phase 1, version 3 data (Supplementary Data 3); (2) screened non-coding annotation from ENCODE data using HaploReg v4.1 (ref. 18) and RegulomeDB v1.2 (ref. 19) to identify possible functional variants, primarily in cells of non-kidney origin but also in BC_kidney_01-11002 and BC_kidney_H12817N cell lines (Supplementary Data 3); and (3) performed expression quantitative trait locus (eQTL) analyses with genes located up to 3 Mb around the newly identified risk markers (or highly correlated proxies) using ccRCC and normal kidney tissue data from the Cancer Genome Atlas [Kidney Renal Clear Cell Carcinoma (KIRC) collection; 481 tumour and 71 normal tissue samples]20 and IARC (555 tumour and 234 normal tissue samples)21 (Supplementary Data 4).

The new highly significant locus marked by rs4903064 at 14q24 maps to the double PHD fingers 3 gene (DPF3), which encodes a histone acetylation and methyltransfer reader of the BAF and PBAF chromatin remodelling complexes. This locus contains a set of correlated SNPs ($r^2 > 0.8$ in 1000G EUR) that reside within the introns of DPF3 (Supplementary Data 3), of which only rs4903064 itself is annotated as likely to disrupt transcription factor binding (RegulomeDB score < 4)19. This variant is located within a region annotated as an enhancer in multiple tissues by the RoadMap project22 and is predicted to alter IRX2/IRX5 binding motifs. In an eQTL analysis, we observed a consistent pattern of increased DPF3 expression associated with the rs49030604 risk allele in both the KIRC and IARC data sets ($P = 5.5 \times 10^{-8}$ and $3.8 \times 10^{-9}$, respectively, Fig. 2, Supplementary Data 4). A consistent, but statistically weaker, expression pattern in the normal kidney tissue data sets of more limited sample size was also observed ($P = 0.15$ and 0.42, respectively). It is noteworthy that 14q24 is deleted in 22–45% of ccRCC20,23. While DPF3 mutation is rare in RCC20, somatic alterations of BAP1 and PBRM1, components of the BAF and PBAF complexes, respectively, are commonly seen in ccRCC24. In this regard, deregulation of this pathway is a common feature of RCC, and these data suggest that rs4903064 may play a role in RCC development through dysregulation of the DPF3 expression.

For the 1p32.3 locus marked by rs4381241, an intronic SNP within FAS-associated factor 1 (FAM1) that encodes a protein that can initiate or enhance FAS-mediated apoptosis, we identified several promising correlated variants with RegulomeDB scores, suggesting alteration of transcription factor binding (Supplementary Data 3) but did not observe a strong effect on expression (Supplementary Data 4). FAS-associated factor 1 facilitates the degradation of β-catenin, a transcriptional co-activator that stimulates expression of genes driving cell proliferation25. Constitutively activated β-catenin, induced by VHL inactivation, is an important pathway in ccRCC oncogenesis26. The rs4381241 risk allele is weakly correlated ($r^2 = 0.12$ in CEU) with the allele of another FAM1 variant (rs17106184) associated with reduced risk of type-2 diabetes and lower serum insulin post oral glucose challenge27,28.

The risk variant rs67311347 maps to a region of 3p22.1 that harbours several genes. Within the KIRC tumour tissue data, the risk-associated allele of the surrogate SNP rs9821249 ($r^2 = 0.97$ with rs67311347 in CEU) was weakly associated with higher expression of CTNNB1 ($P = 0.03$). This gene, located 706 kb away centromeric, is a strong candidate as it encodes the RCC proto-oncogene β-catenin, although this association was not seen within the IARC data set. In both normal tissue data sets, the risk-associated allele of rs67311347 was associated with a higher expression of ZNF620 ($P = 0.03$ and 0.02). This gene encodes the Zinc finger protein 620, but the function of this protein has not been well described.

The 8p21.3 risk variants rs2241261 and rs2889 (used as proxy for rs2241260, $P = 1.6 \times 10^{-9}$, $r^2 = 0.61$ with rs2241261 in CEU; Supplementary Data 1) are located 0.9 and 1.7 kb respectively from TNRFSF10B, a tumour suppressor gene encoding a mediator of apoptosis signalling29. In both the KIRC and IARC tumour tissue data ($P = 0.002$ and 0.03, respectively), the rs2241261 risk allele was associated with a decreased expression of GFA2, which encodes for cell-surface receptor for glial cell line-derived neurotrophic factor (GDNF) and neurturin (NTN), and mediates activation of the RET tyrosine kinase receptor (GliaI cell line-derived neurotrophic factor (Supplementary Data 4). A potential link with renal tissue function has not been described. Of the variants in strong LD with either rs2241261 or rs2889 ($r^2 > 0.8$ in 1000G EUR), only rs2889 is annotated as a strong regulatory candidate by RegulomeDB, predicted to be in a strong enhancer region and altering motifs for FOX family members of transcription factors (Supplementary Data 3).

SNPs rs74911261 and rs1800057 are located 214 kb apart on 11q22.3 and are highly correlated ($r^2 = 0.83$ in CEU) non-synonymous variants, but for separate genes; rs74911261 (P144L) maps to KDEL, which encodes a protein localizing to the endoplasmic reticulum, while rs1800057 (P1054R) maps to the DNA repair gene ATM. The functional prediction tools SIFT30 and PolyPhen-2 (ref. 31) suggest that both amino acid substitutions are damaging. It is also plausible that they are correlated with regulatory variants that influence expression of nearby genes. In eQTL analyses, no consistent associations were detected. Only one of the five variants with strong LD to rs74911261 ($r^2 > 0.8$ in 1000G EUR) has a RegulomeDB score suggesting likely disruption of transcription factor binding (score < 4), rs141379009, and is located within a region annotated as an enhancer by the Roadmap project and predicted to alter a consensus Zfp105/ZNF35 binding motif (Supplementary Data 3). ATM mutations in RCC are uncommon20,23, and ataxia telangiectasia patients, though at markedly elevated cancer risk, have not been reported to frequently develop RCC32, questioning a direct role of ATM in RCC susceptibility.
Table 1 | Summary results for newly discovered loci associated with renal cell carcinoma.

<table>
<thead>
<tr>
<th>Locus</th>
<th>SNP*</th>
<th>Closest gene</th>
<th>Position (base pairs)</th>
<th>A/a</th>
<th>MAF†</th>
<th>Statistics</th>
<th>Discovery (10,784 cases; 20,406 controls)</th>
<th>Replication (3,182 cases; 6,301 controls)</th>
<th>Combined (13,966 cases; 26,707 controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p32.3</td>
<td>rs4381241</td>
<td>FAF1</td>
<td>50907438</td>
<td>T/C</td>
<td>0.44</td>
<td>OR (95% CI)</td>
<td>1.11 (1.07–1.15) 1.1 × 10^{-8}</td>
<td>1.11 (1.03–1.20) 8.7 × 10^{-3}</td>
<td>1.11 (1.07–1.15) 3.1 × 10^{-10}</td>
</tr>
<tr>
<td>3p22.1</td>
<td>rs6731347</td>
<td></td>
<td>40533243</td>
<td>G/A</td>
<td>0.31</td>
<td>OR (95% CI)</td>
<td>0.89 (0.86–0.93) 4.7 × 10^{-8}</td>
<td>0.94 (0.88–1.01) 8.8 × 10^{-2}</td>
<td>0.90 (0.87–0.94) 2.5 × 10^{-8}</td>
</tr>
<tr>
<td>3q26.2</td>
<td>rs10936602</td>
<td>LRRIQ4</td>
<td>169536637</td>
<td>T/C</td>
<td>0.27</td>
<td>OR (95% CI)</td>
<td>0.90 (0.86–0.94) 2.7 × 10^{-7}</td>
<td>0.91 (0.85–0.98) 9.7 × 10^{-3}</td>
<td>0.90 (0.87–0.93) 8.8 × 10^{-9}</td>
</tr>
<tr>
<td>8p21.3</td>
<td>rs2241261</td>
<td>RHOBT82/TFNRF510B</td>
<td>22876739</td>
<td>C/T</td>
<td>0.51</td>
<td>OR (95% CI)</td>
<td>1.10 (1.06–1.14)</td>
<td>1.10 (1.03–1.17)</td>
<td>1.10 (1.06–1.13)</td>
</tr>
<tr>
<td>10q24.33-</td>
<td>rs11813268</td>
<td>OBFC1</td>
<td>105682296</td>
<td>C/T</td>
<td>0.16</td>
<td>OR (95% CI)</td>
<td>1.13 (1.08–1.19)</td>
<td>1.10 (1.01–1.19)</td>
<td>1.12 (1.07–1.17)</td>
</tr>
<tr>
<td>q25.1</td>
<td>rs11813268</td>
<td>OBFC1</td>
<td>105682296</td>
<td>C/T</td>
<td>0.16</td>
<td>OR (95% CI)</td>
<td>1.13 (1.08–1.19)</td>
<td>1.10 (1.01–1.19)</td>
<td>1.12 (1.07–1.17)</td>
</tr>
<tr>
<td>11q22.3</td>
<td>rs7491126</td>
<td>KDELRC2</td>
<td>108357137</td>
<td>G/A</td>
<td>0.02</td>
<td>OR (95% CI)</td>
<td>1.42 (1.26–1.61)</td>
<td>1.38 (1.12–1.69)</td>
<td>1.41 (1.27–1.57)</td>
</tr>
<tr>
<td>rs1800057</td>
<td>ATM</td>
<td>108143456</td>
<td>C/G</td>
<td>0.02</td>
<td>1.40 (1.24–1.59)</td>
<td>1.30 (1.04–1.62)</td>
<td>1.38 (1.23–1.53)</td>
<td>9.0 × 10^{-9}</td>
<td></td>
</tr>
<tr>
<td>14q24.2</td>
<td>rs4903064</td>
<td>DPF3</td>
<td>73279420</td>
<td>T/C</td>
<td>0.23</td>
<td>OR (95% CI)</td>
<td>1.18 (1.13–1.23)</td>
<td>1.30 (1.21–1.39)</td>
<td>1.21 (1.16–1.25)</td>
</tr>
</tbody>
</table>

*SNP with lowest P value within locus. For 1q22.3, results shown for two non-synonymous SNPs in KDELRC2 (rs74991126, Pro144Leu) and ATM (rs1800057, Pro1054Arg; r² = 0.83 in CEU).

†Minor allele frequency among all controls (n = 26,707). Odds ratios (OR) are shown for the minor allele, assuming a log-additive (trend) SNP effect.

For the remaining two new RCC risk loci, in silico analyses and eQTL did not indicate altered regulation of a plausible candidate gene. For each of these loci, we identified SNPs that correlate with low RegulomeDB scores for intriguing nearby candidate genes (Supplementary Data 3). The marker SNP rs10936602 maps to 8p21.3, a region amplified in 15% of ccRCC tumours in KIRC28; several notable nearby genes could represent possible candidate genes, including MECOM, a transcriptional regulator frequently amplified in RCC29, and TERC, encoding a component of telomerase, in which mutations cause autosomal dominant dyskeratosis congenita and aplastic anaemia30. This risk variant is moderately correlated with variants previously associated with telomere length and risk of several malignancies, including multiple myeloma, chronic lymphocytic leukaemia, bladder cancer, glioma and colorectal cancer (rs10936602, rs12696304, rs1920116; r² = 0.66, 0.58 and 0.80, respectively)34–39. The 10q24 risk variant rs11813268 is located 4 kb upstream of OBFC1, a gene identified in GWAS and laboratory investigation as a regulator of human telomere length40. This risk variant is highly correlated with SNPs associated with leucocyte telomere length (rs4387287, rs9419958 and rs9420907; r² = 0.99, 0.82 and 0.82, respectively)40, and to a lesser degree with melanoma (rs2995264, r² = 0.52)41, suggesting the underlying basis for RCC risk may be mediated through a common pathway.

Polygenic risk score analysis and explained heritability. Additional analyses were conducted by generating a polygenic risk score (PRS) from 13 SNPs mapping to the six previously reported and seven newly identified susceptibility loci (Supplementary Table 5). Accepting the caveat of the winner’s curse phenomenon, whereby the strength of SNP associations may have been overestimated, subjects in the highest decile of the PRS had a threefold increased risk of RCC relative to the lowest decile (OR 3.24, 95% CI 2.86–3.67; P = 1.2 × 10^{-76}). Stratifying by histological subtypes, the PRS was most strongly associated with clear cell RCC (per unit increase: OR 3.24, 95% CI 2.91–3.62; P = 3.4 × 10^{-106}), with a weaker association for chromophobe RCC (OR 2.34, 95% CI 1.58–3.46; P = 3.4 × 10^{-5}) and papillary RCC (OR 1.83, 95% CI 1.44–2.32; P = 5.3 × 10^{-7}). The PRS did not significantly differ between cases aged <60 versus ≥60 at diagnosis or across cancer stage (Supplementary Table 5). Using Genome-Wide Complex Trait Analysis (GCTA), we estimate that the heritability and familial relative risk of RCC attributable to all common variation were 14.2% (SE = 0.023) and 1.52 (SE = 0.10), respectively. After excluding established and newly identified loci, the estimates were 12.8% (SE = 0.023) and 1.46 (SE = 0.10), respectively. On the basis of these estimates, ~90% of the heritability and familial risk remains to be elucidated.

Discussion. Our meta-analysis of six GWAS scans identified seven new RCC susceptibility loci. Our findings provide further evidence for polygenic susceptibility to RCC. Future investigation of the genes targeted by the risk SNPs is likely to yield increased insight into the development of RCC. We estimate that the risk loci so far identified for RCC account for only about 10% of the familial risk of RCC. Although the power of our study to detect the major common loci (MAF > 0.2) conferring risk ≥ 1.2 was high (~80%), we had low power to detect alleles with smaller effects and/or MAF < 0.1. By implication, variants with such profiles probably represent a much larger class of susceptibility loci for RCC and hence a large number of variants remain to be discovered. In parallel, whole-exome and whole-genome sequencing of genetically enriched cases selected according to early age of onset or family history would provide new
opportunities to discover rare variants associated with RCC. As more RCC susceptibility alleles are discovered, deciphering the biological basis of risk variants should provide new insights into the biology of RCC that may lead to new approaches to prevention, early detection and therapeutic intervention.

Methods

Informed consent and study approval. Each participating study obtained informed consent from the study participants and approval from its Institutional Review Board (for the IARC scans and replication: IARC Ethics Committee; for the MDA scans and replication: Institutional Review Board of the University of Texas MD Anderson Cancer Center; for the UK scan: Royal Marsden NHS Trust ethics committee; for the NCI scans: NCI Special Studies Institutional Review Board, The Vanderbilt Institutional Review Board, The Emory University Institutional Review Board, Dana-Farber/Harvard Cancer Center institutional review board, Institutional Review Board of the Harvard T.H. Chan School of Public Health, Institutional Review Board of Brigham and Women’s Hospital, Van Andel Research Institute Institutional Review Board, Spectrum Health Institutional Review Board and Fred Hutchinson Cancer Research Center Institutional Review Board; for the Mayo replication: Mayo Clinic institutional review board.

Genome-wide SNP genotyping. Genome-wide SNP genotyping for two new scans was coordinated by the National Cancer Institute (NCI-2; NCI, Bethesda, Maryland, USA) and the International Agency for Research on Cancer (IARC-2; IARC, Lyon, France). The NCI-2 samples, obtained from 13 studies conducted in the US and Finland (Supplementary Table 1), were genotyped at the NCI Cancer Genomics Research Laboratory (CGR, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA) using the Illumina OmniExpress array. The NCI-2 scan included controls previously genotyped by Illumina OmniExpress, or Omni 2.5Marray from some of the participating studies (ATBC, CPSII, HPFS, NHS, PLCO and WHI; Supplementary Table 1). IARC-2 samples, obtained from six studies conducted in Europe and Australia (Supplementary Table 1), were genotyped at the Centre National de Genotypage, Commissariat à l’énergie atomique et aux énergies alternatives (CNG, CEA, Evry and Paris) using the Illumina Omni 5 M arrays. Additional controls (N = 447) from one study (IARC K2) were also included, which had been genotyped on the OmniExpress array at Johns Hopkins Center for Inherited Disease Research.

Quality control assessment. The quality control exclusions for the four previously published scans have been reported9–11. For the two new scans, quality control was conducted separately at each institution using comparable exclusions. For the new IARC-2 scan, a total of 5,424 samples were genotyped on the Centre National de Genotypage, Commissariat à l’énergie atomique et aux énergies alternatives (CNG, CEA, Evry and Paris) using the Illumina Omni 5 M arrays. Additional controls (N = 447) from one study (IARC K2) were also included, which had been genotyped on the OmniExpress array at Johns Hopkins Center for Inherited Disease Research.

For the new IARC-2 scan, a total of 5,424 samples were genotyped on the Centre National de Genotypage, Commissariat à l’énergie atomique et aux énergies alternatives (CNG, CEA, Evry and Paris) using the Illumina Omni 5 M arrays. Additional controls (N = 447) from one study (IARC K2) were also included, which had been genotyped on the OmniExpress array at Johns Hopkins Center for Inherited Disease Research.
exclusion, the concordance rate was >99.9% for 66 pairs of blind duplicate samples. After removing duplicates, a dataset including 2,820 unique samples was advanced to further assess quality control at the subject level. In addition, we excluded 10 sex-discordant individuals and two individuals with exceptionally low mean heterozygosity for ChrX SNPs. For the cleaned dataset including genotypes for 2,808 individuals, we next pooled data from 4,221 previously scanned controls (HumanOmni2.5M or HumanOmniExpress array) from the ATBC, CPSII, HPFS, NHS, PLCO and WHI studies (Supplementary Table 1). After merging the newly scanned data with the previously scanned controls, we obtained genotypes for 7,029 individuals. Subsequently, we excluded data for 204 non-CEU individuals (admixture proportion for CEU < 80%), both members of a pair of unexpected within-family duplicate samples, one from each of eight unexpected cross-study duplicate pairs, and one from each of eight related pairs (two parent–child pairs and six sibling pairs). The final analytic dataset comprised 6,808 individuals (2,417 cases, 4,391 controls) for 678,580 loci.

Statistical analysis. The statistical analysis included summary data from four previously published scans conducted at the NCI (NCI-1)9, IARC (IARC-1)9, the University of Texas MD Anderson Cancer Center (MDA)30 and the Institute of Cancer Research, UK (UK)31, as well as the two new scans from NCI (NCI-2) and IARC (IARC-2). The IARC-1 and IARC-2 data were pooled, resulting in five separate discovery-stage data sets. Imputation was performed separately for each scan dataset using SNPs of minor allele frequency ≥0.01 (≥0.05 for the IARC data set), with 1,000 Genomes Project data (phase 1 release 3) used as a reference set. IMPUTE2 version 2.2.2 was used for imputation of the NCI-1, NCI-2, MDA and UK data sets, while Minimac version 3 was used for the IARC data set. Imputed SNPs with sufficient accuracy as assessed by r² ≥0.3 for both IMPUTE2 and Minimac were retained for the analysis. We further assessed the quality of imputation by randomly selecting 10% of genotyped SNPs on chromosome 1 within the IARC-1 series (which used the least-dense chip across the different scans) and removing them before running the imputation algorithm. MAFs calculated from the genotyping data correlated with r² >0.99 with MAFs calculated from the imputed dosage data. Finally, top SNPs were technically validated through Taqman genotyping in the IARC and NCI-2 scans (Supplementary Table 4). After imputation, genotypes for 7,437,091 SNPs were available for analysis. Association testing with RCC was conducted separately for each data set assuming log-additive (trend) SNP effects using SNPTEST version 2.2 at NCI and R version 3.2.3 at IARC. The model covariates varied by data set; for the previous scans, we used the same covariates as in the initially published analyses. The covariates were as follows: sex and study for NCI-1 (no statistically significant eigenvectors present in null model); sex and four significant eigenvectors for NCI-2; age, sex and two significant eigenvectors for MDA; no covariates for the UK; and sex, study, and 19 significant eigenvectors for IARC-1 and IARC-2. Eigenvectors were considered significant if P<0.05 from the Tracy–Widom statistics. In the IARC series, all 19 eigenvectors were significantly associated with the country of recruitment. We additionally conducted analyses restricted to cases with a missing rate <10% or a HWE P value >0.05, and removed outlier samples with expression values exceeding 1.5 times the interquartile range. Linear trend tests were used to test for allele-specific increases in gene expression for genes within a 6 Mb window. Analyses were performed using R v3.1.1. IARC: For a subset of cases from the IARC K2 and the CE studies (Supplementary Table 1), we conducted a disease stage-stratified analysis with a linearmixed-effects model with an interaction term between stage and SNP. The stage-stratified analyses were restricted to the IARC data sets, for which these data were available.

Gene expression data and eQTL analysis. KIRC: Genotyping and RNAseq data for the KIRC TCGA samples (481 tumours and 71 normal renal tissues) were downloaded from The Cancer Genome Atlas database (http://cancergenome.nih.gov/), accessed on 15 November 2016. We quantified expression as normalized read counts and removed outlier samples with expression values exceeding 1.5 times the interquartile range. For each of the 24 normal and 555 tumour samples from the KIRC TCGA data set, we used the same covariates as in the initially published analyses. The stage-stratified analyses were restricted to cases with a missing rate <10% or a HWE P value >0.05, and removed outlier samples with expression values exceeding 1.5 times the interquartile range. Linear trend tests were used to test for allele-specific increases in gene expression for genes within a 6 Mb window. Analyses were performed using R v3.1.1. IARC: For a subset of cases from the IARC K2 and the CE studies (Supplementary Table 1), we conducted a disease stage-stratified analysis with a linearmixed-effects model with an interaction term between stage and SNP. The stage-stratified analyses were restricted to the IARC data sets, for which these data were available.

Data availability. The scan IARC-2 obtained Institutional Review Board certification permitting data sharing in accordance with the US NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted GWAS. Data are accessible on dbGaP (study name: ‘Pooled Genome-Wide Analysis of Kidney Cancer Risk (KIDRISK)’; http://www.ncbi.nlm.nih.gov/projects/gcp/cgi-bin/study.cgi?study_id=phs001271.v1.p1). Similarly, the NCI-1 scan is accessible on dbGaP (phs000351.v1.p1). Data from IARC-1 and MDA scans are available from Paul Brennan and Xifeng Wu, respectively, upon reasonable request. The UK scan data will be made available on the European Genome-phenome Archive database (accession number: EGAS00001002336). The NCI-2 scan will be posted on dbGaP. TCGA data were accessed at the following url: https://gdac-portal.nci.nih.gov/projects/TCGA-KIRC.

References
32. Suarez, F.
34. Codd, V.

**Acknowledgements**

The authors thank all of the participants who took part in this research and the funders and support staff who made this study possible. Funding for the genome-wide genotyping was provided by the US National Institutes of Health (NIH), National Cancer Institute (U01CA155309) for those studies coordinated by IARC and by the intramural research program of the National Cancer Institute, US NIH, for those studies coordinated by the NCI. Funding for the IARC gene expression and eQTL study was provided by the US National Institutes of Health (NIH), National Cancer Institute (U01CA155309). Additional acknowledgments can be found in Supplementary Note.

**Author contributions**


**Additional information**

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/


**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
School of Population and Global Health, The University of Melbourne, Carlton, Victoria 3053, Australia. Human Genetics Foundation (HuGeF), 10126 Turin, Italy. Department of Medicine, The University of Melbourne, Melbourne, Victoria 3010, Australia. Genetic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger 7600, Norway. Department of Public Health and General Practice, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim 7491, Norway. London School of Hygiene and Tropical Medicine, University of London, London WC1H 9SH, UK. Institute of Environmental Medicine, Karolinska Institutet, 171 77 Stockholm, Sweden. Leeds Institute of Cancer and Pathology, University of Leeds, Cancer Research Building, St James’s University Hospital, Leeds LS9 7TF, UK. Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK. Department of Health, National Institute for Health and Welfare, 00271 Helsinki, Finland. Vanderbilt-Ingram Cancer Center, Department of Urology, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. Vanderbilt-Ingram Cancer Center, Division of Epidemiology, Department of Medicine, Institute for Medicine and Public Health, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee 37209, USA. Vanderbilt-Ingram Cancer Center, Division of Epidemiology, Department of Medicine, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, Tennessee 37203, USA. American Cancer Society, Atlanta, Georgia 30303, USA. Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. Warren Alpert Medical School of Brown University, Providence, Rhode Island 02903, USA. Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. Brigham and Women’s Hospital and VA Boston, Boston, Massachusetts 02115, USA. Division of Urology, Spectrum Health, Grand Rapids, Michigan 49503, USA. College of Human Medicine, Michigan State University, Grand Rapids, Michigan 49503, USA. Van Andel Research Institute, Center for Cancer Genomics and Quantitative Biology, Grand Rapids, Michigan 49503, USA. Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. Department of Epidemiology and Biostatistics, School of Public Health Indiana University Bloomington, Bloomington, Indiana 47405, USA. Department of Urology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. Department of Oncology, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK. Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. Medical Oncology, Royal Marsden NHS Foundation Trust, London SW3 6JJ, UK. Department of Urology, Mayo Medical School and Mayo Clinic, Rochester, Minnesota 55902, USA. McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada H3A 0G1. Department of Health Sciences Research, Mayo Clinic, Jacksonville, Florida 32224, USA. Division of Genetics and Epidemiology, The Institute of Cancer Research, London SW7 3RP, UK. These authors contributed equally to this work. These authors jointly supervised this work.