Scelo, G; Purdue, MP; Brown, KM; Johansson, M; Wang, Z; Eckel-Passow, JE; Ye, Y; Hofmann, JN; Choi, J; Foll, M; +113 more... Gaborieau, V; Machiela, MJ; Colli, LM; Li, P; Sampson, JN; Abedi-Ardekani, B; Besse, C; Blanche, H; Boland, A; Burdette, L; Chabrier, A; Durand, G; Le Calvez-Kelm, F; Prokhortchouk, E; Robinot, N; Skryabin, KG; Wozniak, MB; Yeager, M; Basta-Jovanovic, G; Dzamic, Z; Foretova, L; Holcatova, I; Janout, V; Mates, D; Mukeriya, A; Rascu, S; Zaridze, D; Bencko, V; Cybulski, C; Fabianova, E; Jinga, V; Lissowska, J; Lubinski, J; Navratilova, M; Rudnai, P; Szeszenia-Dabrowska, N; Benhamou, S; Cancel-Tassin, G; Cussenot, O; Baglietto, L; Boeing, H; Khaw, KT; Weiderpass, E; Ljungberg, B; Sitaram, RT; Bruinsma, F; Jordan, SJ; Severi, G; Winship, I; Hveem, K; Vatten, LJ; Fletcher, T; Koppova, K; Larsson, SC; Wolk, A; Banks, RE; Selby, PJ; Easton, DF; Pharoah, P; Andreotti, G; Freeman, LEB; Koutros, S; Albanes, D; Männistö, S; Weinstein, S; Clark, PE; Edwards, TL; Lipworth, L; Gapstur, SM; Stevens, VL; Carol, H; Freedman, ML; Pomerantz, MM; Cho, E; Kraft, P; Preston, MA; Wilson, KM; Michael Gaziano, J; Sesso, HD; Black, A; Freedman, ND; Huang, WY; Anema, JG; Kahnoski, RJ; Lane, BR; Noyes, SL; Petillo, D; Teh, BT; Peters, U; White, E; Anderson, GL; Johnson, L; Luo, J; Buring, J; Lee, IM; Chow, WH; Moore, LE; Wood, C; Eisen, T; Henrion, M; Larkin, J; Barman, P; Leibovich, BC; Choueiri, TK; Mark Lathrop, G; Rothman, N; Deleuze, JF; McKay, JD; Parker, AS; Wu, X; Houlston, RS; Brennan, P; Chanock, SJ; (2017) Genome-wide association study identifies multiple risk loci for renal cell carcinoma. Nature communications, 8. p. 15724. ISSN 2041-1723 DOI: https://doi.org/10.1038/ncomms15724

Downloaded from: http://researchonline.lshtm.ac.uk/3962251/

DOI: https://doi.org/10.1038/ncomms15724

Usage Guidelines:

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
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, \( P = 2.5 \times 10^{-8} \)), 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, \( P = 2.2 \times 10^{-24} \)). 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. 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). 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. While identification of these genes has led to important insights into the pathogenesis of RCC, 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. Support for polygenic susceptibility to RCC has been reported scans of 5,586 cases and 13,075 controls, reaching a genome-wide significant SNPs in that region. 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. 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.

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, $P = 2.5 \times 10^{-9}$), 3q26 (rs10936602, $P = 8.8 \times 10^{-9}$), 8p21.3 (rs2241261, $P = 5.8 \times 10^{-9}$), 10q24.33-q25.1 (rs11813268, $P = 3.9 \times 10^{-9}$), 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 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 Table 2).
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 analysis \(^{16}\). 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, \(P = 4.0 \times 10^{-5}\)). 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 sets of more than 1000 Genome Project tissue samples \(^{21}\) (Supplementary Data 4). A consistent, but statistically weaker, expression pattern in the normal kidney tissue data sets of more than 1000 Genome Project tissue samples \(^{21}\) (Supplementary Data 4) was observed in additional SNP analyses of non-clear cell 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 \(\beta\)-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 signalling \(^{25}\). 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 GFRα2, 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 (Glial 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 KDEL2, which encodes a protein localizing to the endoplasmic reticulum, while rs1800057 (P1054R) maps to the DNA repair gene ATM. The functional prediction tools SIFT \(^{30}\) 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 uncommon, and ataxia telangiectasia patients, though at markedly elevated cancer risk, have not been reported to frequently develop RCC, questioning a direct role of ATM in RCC susceptibility.
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 3q26.2, a region amplified in 15% of ccRCC tumours in KIRC20; several notable nearby genes could represent possible candidate genes, including MECom, a transcriptional regulator frequently amplified in RCC20, and TERC, encoding a component of telomerase, in which mutations cause autosomal dominant dyskeratosis congenita and aplastic anaemia23. 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 (rs10936599, rs12696304, rs1920116; \( r^2 = 0.66, 0.58 \) and 0.80, respectively)24–26. 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; \( r^2 = 0.99, 0.82 \) and 0.82, respectively)40, and to a lesser degree with melanoma (rs2995264, \( r^2 = 0.52 \))41, suggesting the underlying basis for RCC risk may be mediated through a common pathway.

### 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 T/C</td>
<td>0.44</td>
<td>1.11 (1.07–1.15)</td>
<td>1.11 (1.03–1.20)</td>
<td>1.11 (1.07–1.15)</td>
<td>3.1 x 10^{-8}</td>
<td>0%</td>
</tr>
<tr>
<td>3p22.1</td>
<td>rs67311347</td>
<td>G/A</td>
<td>0.31</td>
<td>0.89 (0.86–0.93)</td>
<td>0.94 (0.88–1.01)</td>
<td>0.90 (0.87–0.94)</td>
<td>2.5 x 10^{-8}</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>3q26.2</td>
<td>rs10936602</td>
<td>LRR1Q4</td>
<td>169536637 T/C</td>
<td>0.27</td>
<td>0.90 (0.86–0.94)</td>
<td>0.91 (0.85–0.98)</td>
<td>0.90 (0.87–0.93)</td>
<td>8.8 x 10^{-9}</td>
<td>11%</td>
</tr>
<tr>
<td>8p21.3</td>
<td>rs2241261</td>
<td>RHOBTB2/ TNR5F1OB</td>
<td>22876739 C/T</td>
<td>0.51</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>
<td>5.8 x 10^{-9}</td>
<td>21%</td>
</tr>
<tr>
<td>10q24.33-q25.1</td>
<td>rs11813268</td>
<td>OBFC1</td>
<td>105682296 C/T</td>
<td>0.16</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>
<td>3.9 x 10^{-8}</td>
<td>0%</td>
</tr>
<tr>
<td>11q22.3</td>
<td>rs7491126</td>
<td>KDELC2</td>
<td>108357137 G/A</td>
<td>0.02</td>
<td>0.91 (0.86–0.94)</td>
<td>0.91 (0.85–0.98)</td>
<td>0.90 (0.87–0.93)</td>
<td>8.8 x 10^{-9}</td>
<td>11%</td>
</tr>
<tr>
<td>14q24.2</td>
<td>rs4903064</td>
<td>DPF3</td>
<td>73279420 T/C</td>
<td>0.28</td>
<td>1.18 (1.13–1.23)</td>
<td>1.20 (1.13–1.29)</td>
<td>1.21 (1.16–1.25)</td>
<td>2.2 x 10^{-9}</td>
<td>36%</td>
</tr>
</tbody>
</table>

*SNP with lowest P value within locus. For 1q22.3, results shown for two non-synonymous SNPs in KDELC2 (rs7491126, Pro144Leu) and ATM (rs1800057, Pro1054Arg; \( r^2 = 0.83 \) 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.

**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 \( \geq 1.2 \) 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 evidence.
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.

**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 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.

**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 2.5M array 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 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 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 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 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 pairs. 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 excessively low mean heterozygosity for ChrX SNPs. For the cleaned data including genotypes for 2,808 individuals, we next pooled a dataset of 6,421 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 (admixtute proportion for CEU <80%), both members of a pair of unexpected within-subject 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 data 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-2), IARC (IARC-1), the University of Texas MD Anderson Cancer Center (MDA)\(^5\), and the Institute of Cancer Research, UK (UK)\(^3\), 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 data set 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\(^4\). Imputed SNPs with sufficient accuracy as assessed by \(r^2 \geq 0.3\) 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 last-dense chip across the different scans) and removing them before running the imputation algorithm. MAFs calculated from the genotyping data correlated with \(r^2 > 0.99\) with MAFs calculated from the imputed dosage data. Finally, top SNPs were technically validated through Taqman genotyping in the IARC and NCI-2 scan (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 2.2.1 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 IARC-1 and IARC-2. Eigenvectors were considered significant if \(P < 0.05\) from the Trace–Widom statistics. In the IARC series, all 19 eigenvectors were significantly associated with the country of recruitment. We additionally conducted analyses restricted to cRCC. The SNP association results from each data set were combined by meta-analysis using a fixed-effects model. Heterogeneity in genetic effects across data sets was assessed using the DerSimonian–Laird random-effects model. In addition to quality control steps taken for the original GWAS, we removed SNPs with a missing rate \(>5\%\) of genotypes and adjusted for sex, substudy and the top 20 eigenvectors.

**Replication genotyping and analysis.** After filtering out previous GWAS-identified SNPs, we selected for replication 32 SNPs with association \(P < 5 \times 10^{-7}\). A separate set of 3,182 cases and 6,301 controls of European ancestry were genotyped at three institutions (IARC: 1,674 cases and 4,222 controls; Mayo Clinic: 909 cases and 1,479 controls; MDA: 599 cases and 600 controls) for replication. Genotyping at IARC and MDA was conducted by Taqman assay (Applied Biosystems, CA, USA), while the Mayo Clinic samples were genotyped using a combination of MassARRAY (Agena Bioscience, Inc., CA, USA) and Taqman assays. The associations with each SNP (per minor allele/trend) were computed individually for each institution (IARC: adjusted for sex and study; Mayo Clinic: age and sex; MDA: age and sex) and combined with the discovery-stage results through fixed-effects meta-analysis.

**Polygenic risk score and analyses of additional RCC phenotypes.**PRS was calculated for 13 SNPs, one from each of the six previously identified loci and seven newly identified RCC risk loci (rs1105934, rs4765623, rs1781341, rs1894252, rs12105918, rs6470588, rs4381241, rs67313147, rs19036602, rs2241261, rs18113268, rs74911261 and rs4903064), as follows:

\[
PR_{IS} = \sum_{i=1}^{13} w_{i} x_{i}
\]

where PRS is the risk score for individual \(i\), \(x_{i}\) is the number of risk alleles for the \(j\)th variant and \(w_{i}\) is the weight [\(\ln(OR)\)] of the \(j\)th variant. Associations with the PRS and individual SNPs selected for replication were computed for the following RCC phenotypes: papillary and chromophobe RCC histologies (through case-control analyses); age at onset (<60 versus 60+ years at diagnosis; case-only analyses) and stage (2, 3 and 4 versus 1; case-only analyses). The stage-stratified analyses were restricted to the IARC data sets, for which these data were available.

**Technical validation of imputed SNPs.** To technically validate our imputation findings, we genotyped the 32 SNPs carried over for replication by Taqman assay in a subset of samples from the NCI-2 and IARC-1/2 scans (\(n = 566\) and 6,402 respectively). The concordances between imputed and directly assayed genotypes are detailed in Supplementary Table 4.

**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 January 2016). We quantified expression as normalized read counts 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 v3.1.3. IARC: For a subset of cases from the IARC K2 and the CE studies (Supplementary Table 1), we conducted expression analysis of renal normal and tumour tissue samples using conducted using Illumina HumanHT-12 v4 expression BeadChips (Illumina, Inc., San Diego) for samples with RNA integrity (RIN) \(>5\). Raw expression intensities of samples with signal-to-noise ratio \(>9.5\) were processed with variance-stabilizing transformation and quantile normalization with lumi package\(^5\) as reported by Wozniak et al.\(^21\). The 50 mer sequences of probes were mapped to human reference genome hg19 downloaded from UCSC Genome Browser database (http://genome.ucsc.edu/, accessed on 15 November 2014) using BWA\(^6\) to demarcate positional relationships between corresponding probes/genes and SNPs. In total, 234 normal and 555 tumour tissue samples from confirmed clear cell RCC cases were used to test for allele-specific increases in gene expression for genes within a 6 Mb window under linear trend assumption. Analyses were performed using R v3.1.3.

**Data availability.** The scan IARC-2 obtained Institutional Review Board (IRB) 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 (KDRIISK)”; url: http://www.ncbi.nlm.nih.gov/projects/gapl/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: EGAS0000102336). The NCI-2 scan will be posted on dbGaP. TCGA data were accessed at the following url: http://gdc-portal.nci.nih.gov/projects/TCGA-KIRC.

**References**

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 (U01CA153509) 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 (U01CA153509). 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/

**How to cite this article:** Sceolo, G. et al. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* 8, 15724 (2017).

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

© The Author(s) 2017

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 that is 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/