Online Methods

Genome-wide SNP Genotyping

New genome-wide SNP genotyping was conducted in three laboratories (**Supplementary Table 1**) using Illumina Infinium Beadchips available at the time of genotyping. All US samples were genotyped at the NCI Core Genotyping Facility (CGF, Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, Bethesda, USA) whereas the Centre National de Genotypage (CNG, Evry, France) genotyped all samples from Central Europe and HUNT2/Tromsø as well as cases from EPIC, UK and France. All Moscow samples were genotyped at the Kurchatov Scientific Center (KSC, Moscow, Russian Federation). Controls for the UK cases were drawn from data generated from the 1958 British Birth Cohort by the Wellcome Trust Sanger Institute as part of the Wellcome Trust Case Control Consortium (WTCCC)[10]. Controls from PLCO, ATBC and CPS-II were drawn from previously scanned subjects[28-30]. Controls for the EPIC cases were drawn from data generated from EPIC controls by CGF as part of the Pancreatic Cancer Cohort Consortium (PanScan)[31, 32].

Quality Control Assessment

Systematic quality control common to both centers was conducted separately for the European and US datasets before merging the two datasets, which included QC steps specific for the performance of different arrays at distinct times in the two main laboratories. For SNP assays, exclusions included those with less than 90% of completion rate, and SNPs with extreme deviation from fitness for Hardy-Weinberg proportion ($P < 1x10^{-7}$). Monomorphic assays observed in either cases or controls only, and SNPs with alleles ambiguously coded (AT and CG coding alleles), were excluded.

IARC/CNG Scan

After excluding 46 expected duplicate samples, the number of attempted DNA was 8,031. We excluded 4 pairs (8 samples) of expected duplicates that were not identical, 23 unexpected duplicate pairs (46 samples), and 112 samples with low (<95%) success rate. Samples were excluded if heterozygosity rates for autosomal chromosomes were >6 standard deviations from the mean. We further excluded one self-reported male and one female with abnormal x-chromosome heterozygosity rates (> 10% and < 20%, respectively). Utilizing a set of 12,000 un-linked SNPs (pair-wise $r^2 < 0.004$) common to all GWA arrays[33], 59 samples with less than 80% European ancestry were excluded based on STRUCTURE analysis[34]. Eleven samples were identified as first-degree relatives and excluded based on the identity by descent (IBD). A principal component analysis (PCA) using the EIGENSTRAT software excluded 83 additional samples detected as outliers (6 standard deviations to the mean)[35].

After these QC steps, of the 8,031 samples genotyped, 7,542 (2,461 cases and 5,081 controls) were retained. 577,547 SNPs were available for data pooling.

NCI Scan

2,109 samples (1,490 cases and 619 new controls) were genotyped on Illumina 610 or 660w chips at the Core Genotyping Facility. 2,874 previously scanned (on 550 or 610 chips) controls from PLCO, CPS-II and ATBC were included. Participants were excluded based on: 1) unanticipated inter-study duplicates (n=5); 2) completion rates lower than 92-94% as per the QC groups (n=38 samples); 3) abnormal heterozygosity values of less than 25% or greater than 35% (n=4; 2 overlap with low completion samples); 4) expected duplicates (n=50 pairs); 5) abnormal X chromosome heterozygosity (n = 10 and 6) phenotype exclusions (due to ineligibility or incomplete information)

(n = 57). Utilizing a set of 12,000 un-linked SNPs (pair-wise $r^2 < 0.004$) common to the GWAS chips used herein[33], 80 subjects with less than 85% European ancestry were excluded based on STRUCTURE analysis[34] and PCA[35]. For the known 50 duplicate pairs, concordance was 99.95%.

The final participant count for the association analysis was 1,311 cases and 3,424 controls. 585,576 SNPs were available for analysis in one or more studies.

Each participating study obtained informed consent from study participants and approval from its Institutional Review Board (IRB) for this study and obtained IRB certification permitting data sharing in accordance with the NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). The CGEMS data portal provides access to individual level data from the NCI scan ONLY to investigators from certified scientific institutions after approval of their submitted Data Access Request.

Merging Datasets

The post-QC datasets were merged normalizing strand differences when necessary. No incompatible encodings were detected, and the final dataset contained 586,069 SNPs (after excluding monomorphic and ambiguously coded AT and CG SNPs) for 3,772 cases and 8,505 controls.

Statistical analysis

Associations between the 586,069 SNPs and the risk of kidney cancer were estimated using unconditional logistic regression by the odds ratio (OR) and 95% confidence interval (CI) using the multivariate unconditional logistic regression assuming a co-dominant/trend genetic model (the effect of the variant by log-additive model with 1 degree of freedom). PCA analysis revealed two significant ($p<0.05$) eigenvectors when included in the NULL model (logistic regression with

dummy variables for sex, country and study for the US). The main effect model was adjusted by sex, country, two eigenvectors showing significant effect ($p<0.05$) in the NULL model, and study for the US studies. For the replication studies, both an unadjusted and adjusted analysis were conducted; adjustment included sex, country (study), smoking status (current, former and never), body mass index and diagnosis of hypertension.

The estimated inflation factors of the test statistic, were 1.011 for IARC/CNG scan, 1.016 for the NCI scan, and 1.018 for the pooled scan. All p-values and confidence intervals were corrected for the appropriate observed inflation factor (genomic control).[36]

Replication and TaqMan genotyping

In order to select a set of top-ranked SNPs for further follow-up, we initially combined the European and US datasets through a meta-analysis. Genomic control was applied to the IARC/CNG and NCI scans separately[36], and the results were subsequently combined using a fixed effects meta-analysis model, and per-allele trend effect estimates and p-values were computed using inverse variance weighting (**first column of Table 1**). The individual level genome-wide data were subsequently pooled and association results of the six SNPs selected for replication was combined with results from the replication studies by meta-analysis (**third column of Table 1**). A separate analysis of the six SNPs selected for replication is shown in **Supplementary Table 3** using alternative genetic models, namely, the dominant and recessive models. The association results of the six SNPs selected for replication is also shown separately for each study participating in the GWAS in **Supplementary Table 4**.

TaqMan genotyping assays (ABI, Foster City, CA) for replication were optimized for 5 of 6 SNPs in the three notable regions to validate the Illumina results. rs11894252 could not be

manufactured but instead, rs1867785 (r^2 = 1.0 in CEU HapMap Phase II) was optimized[12]. TaqMan assays for replication were genotyped in three centers, MD Anderson Cancer Center (Houston, TX), Nijmegen, The Netherlands and IARC. Concordance of known duplicates was greater than 99%. In an analysis of 1,126 samples from three studies scanned at NCI, the comparison of the Illumina calls with the TaqMan assays showed a concordance of 98.7-100%; no shifts from wild type to homozygotes were observed. The Illumina Infinium genotype probe cluster plots for the 4 SNPs achieving genome-wide significance, rs11894252, rs7579899, rs7105934 and rs4765623, are shown in **Supplementary Figure 3.**

Imputation

In order to further interrogate the loci associated with RCC, we imputed additional SNPs within 1 Mb on either side of the implicated SNPs using the MACH software and data from the 1000 Genomes project as scaffold[13]. Unconditional logistic regression as implemented in the ProbABEL[37] software was used to analyze the posterior SNP dosages from MACH, adjusting for sex, country, two eigenvectors showing significant effect ($p<0.05$) in the NULL model, and study for the US studies. Association results for all SNPs with R^2 (squared correlation between imputed and true genotypes) above 0.3 and minor allele frequency above 0.05 in the regions of 2p21 (*EPAS1*), 11q13.3, and 12q24.31 (*SCARB1*), are shown in Supplementary Table 2. Also shown in Supplementary Table 2 are the association results for each imputed SNP after adjusting for one of the implicated SNPs in each region.

Data Analysis

Data analysis and management was performed with GLU (Genotyping Library and Utilities version 1.0), PLINK, SAS® version 9.2 (Raleigh, NC, USA), Eigenstrat, MACH, and ProbABEL.

URLs:

CGEMS portal: http://cgems.cancer.gov/

CGF: http://cgf.nci.nih.gov/

GLU: http://code.google.com/p/glu-genetics/

EIGENSTRAT: http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm

STRUCTURE: http://pritch.bsd.uchicago.edu/structure.html

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/

SAS: http://www.sas.com/

MACH: http://www.sph.umich.edu/csg/abecasis/mach/index.html

ProbABEL: http://mga.bionet.nsc.ru/~yurii/ABEL/