Supplementary File 1: Material and Methods

TCGA project: Clinical and RNA-seq data related to 460 ccRCC samples were collected from TCGA. The list of those cases appears in Table S5.

RNA sequencing: In brief, first strand cDNA was synthesized from 100 ng of DNase1-treated total RNA using a mix of DNA/RNA chimeric primers that hybridize to both the 5' portion of the poly(A) sequence and randomly across the transcript. Second strand synthesis produced double-stranded cDNA, which was amplified using single primer isothermal strand-displacement amplification. The resultant cDNA was fragmented to 200 bp (mean fragment size) with the S220 Focused-Ultrasonicator (Covaris) and used to make barcoded sequencing libraries on the SPRI-TE Nucleic Acid Extractor (Beckman-Coulter). Libraries were quantitated by qPCR (KAPA Systems), multiplexed and sequenced, 4 samples per lane, on the HiSeq2000 using 75 bp paired-end sequencing. This method allows for the sequencing of mRNA and non-polyadenylated RNA including histone mRNAs, precursors for Cajal body specific small RNAs, and lncRNAs; the resulting data were analyzed with the current Illumina pipeline to generate raw fastq files.

Identification of differentially expressed genes from RNA-seq: We first counted the overlaps between the mapped reads and genomic features, such as genes/exons using htseq-count script distributed with the HTSeq package. It has been shown that a small number of highly expressed genes can consume a significant amount of the total sequence. As this can change between lanes and experimental condition along with library size, we performed between-sample normalization when testing for differential expression. The choice of normalization is not independent of the test used to determine if any genes are significantly differentially expressed between conditions. We used the scaling factor normalization method as it preserves the count nature of the data and has been shown to be an effective means of improving the detection of differential expression (1). To perform the normalization and the test for differential expression with a negative binomial model between conditions, we chose to use the Bioconductor package DESeq (version 1.11.0), although there are other options available (2). Specifically, for each gene, a generalized linear model (GLM) was fit to compare the expressions of four types, MtRCC, ccRCC, papillary RCC and normal kidney tissue, and to adjust the effects from two batches, MD Anderson and TCGA. The Benjamini-Hochberg method was used to control the FDR (3). For genes with FDR < 0.01, pairwise comparisons were performed between MtRCC vs. normal kidney tissue, MtRCC vs. ccRCC and MtRCC vs. papillary RCC. The Holm method was applied to calculate the adjusted *P*-values of pairwise comparisons (4).

Exome sequencing: Briefly, 3 micrograms of DNA from each sample were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. Whole-exome sequencing was performed, and paired-end sequencing (2 x 76 bp) was carried out using the Illumina HiSeq 2000; the resulting data were analyzed with the Illumina pipeline to generate raw fastq files. The coverage of our samples varied between 46x–100x.

Somatic mutation detection from whole-exome sequencing: After raw paired-end reads from whole-exome resequencing were aligned/mapped to the human genome reference (hg19) and PCR duplicate reads were removed by Mosaik aligner, we then analyzed the resulting alignments using the Bayesian model-based software GigaBayes/FreeBayes that enables the efficient analysis of billions of aligned short-read sequences (5). The program evaluates each aligned base and base quality value at each position to indicate putative single-nucleotide variations (SNVs) and short insertions/deletions (indels), and their corresponding SNV probability value (PSNV). Base quality values are converted to base probabilities corresponding to each of the four possible nucleotides. Using a Bayesian formulation, a PSNV (or indel probability value, as appropriate) is calculated as the likelihood that multiple different alleles are present between the reference genome sequence and the reads aligned at that position. If the probability value exceeds a prespecified threshold, the SNV or indel candidate is reported in the output. In this study, we used a PSNV cutoff value (0.9) to define a high-confidence SNV or short indel candidate. We also filtered out all known SNVs/indels in the UCSC dbSNP 135 and 1000 human genome project SNP database, and kept any mutations, which are in the Catalogue of Somatic Mutations in Cancer (COSMIC) database curated by Wellcome Trust Sanger Institute. We then determined the somatic status of each SNV (or indel) by comparing the genotypes and likelihood between matched normal and tumor samples. The somatic status of a specific SNV/indel was reported if the matched normal had wild allele-based homozygous genotype and the tumor had heterozygous or mutant allele-based homozygous genotype with a certain cutoff of genotype likelihood/*P*-value (0.99). Finally, each somatic mutation or indel was annotated with its functional effect by SIFT to determine whether a mutation candidate was synonymous or non-synonymous (benign or deleterious).

References

1. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol;11:R25.

2. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol;11:R106.

3. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57:289-300.

4. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1979:65-70.

5. Marth GT, Korf I, Yandell MD, et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet 1999;23:452-6.