Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med 2016;374:135-45. DOI: 10.1056/NEJMoa1505917

Supplementary Appendix

Supplement to: The Cancer Genome Atlas Research Network*. Comprehensive Molecular Characterization of Papillary Renal Cell Carcinoma.

Contents

Supplementary Appendix

for

Comprehensive Molecular Characterization of Papillary Renal Cell Carcinoma

The Cancer Genome Atlas Research Network*

*Corresponding Author:
W. Marston Linehan, M.D.
Urologic Oncology Branch,
National Cancer Institute
Building 10 CRC Room 1-5940
Bethesda, MD 20892-1107 USA
Tel: 301-496-6353
Fax: 301-402-0922
Email: WML@nih.gov

Author Contributions

The Cancer Genome Atlas Research Network

The TCGA consortium contributed collectively to this study. Biospecimens were provided by the Tissue Source Sites and processed by the Biospecimen Core Resource. Data generation and analyses were performed by the Genome Sequencing Center, Genome Characterization Centers, and Genome Data Analysis centers. All data were released through the Data Coordinating Center. Project activities were coordinated by the National Cancer Institute and National Human Genome Research Institute Project Teams. Initial guidance in the project design was provided by the Disease Working Group.

Analysis and Disease Working Group:

Project leaders: W. Marston Linehan[1] and Paul T. Spellman[2]. Data Coordinator and Figure Coordinator: Chad J. Creighton[3]. Analysis Coordinator and Manuscript Coordinator: Christopher J. Ricketts[1]. Writing Team: W. Marston Linehan[1], Paul T. Spellman[2], Chad J. Creighton[3], Christopher J. Ricketts[1], Ramaprasad Srinivasan[1]. DNA Sequence analysis: David A Wheeler[3], Paul T. Spellman[2], Bradley A. Murray[4], Suzanne S. Fei[2], Caleb Davis[3], Laura Schmidt[1], Cathy D. Vocke[1], Myron Peto[2], Abu Amar M. Al Mamun[3], Christopher J. Ricketts[1]. Gene fusion analysis: Eve Shinbrot[3], Anurag Sethi[5], David A.Wheeler[3]. mRNA analysis: Samira Brooks[6], W. Kimryn Rathmell[6], Angela N. Brooks[4], Katherine A. Hoadley[6], Chad J. Creighton[3]. microRNA analysis: A. Gordon Robertson[7], Denise Brooks[7], Reanne Bowlby[7], Sara Sadeghi[7]. DNA methylation analysis: Hui Shen[8], Daniel J. Weisenberger[9], Moiz Bootwalla[9], Stephen B. Baylin[10], Peter W. Laird[8]. Copy number analysis: Bradley A. Murray[4], Andrew D. Cherniack[4], Gordon Saksena[4]. Protein analysis: Scott Haake[11], Jun Li[12], Han Liang[12], Yiling Lu[12], Gordon B. Mills[12], Rehan Akbani[12]. Pathway/Integrated Analysis: Mark D.M. Leiserson[13], Benjamin J. Raphael[13], Katherine A. Hoadley[6], Chad J. Creighton[3], Suzanne S. Fei[2], Pavana Anur[2], Donald Bottaro[1], Christopher J. Ricketts[1]. Pathology and Disease Group: Laurence Albiges[14], Nandita Barnabas[15], Toni K. Choueiri[14], Bogdan Czerniak[12], Andrew K. Godwin[16], A Ari Hakimi[17], Thai H. Ho[18], James Hsieh[17], Michael Ittmann[3], William Y. Kim[6], Bhavani Krishnan[6], W. Marston Linehan[1], Maria J. Merino[1], Kenna R. Mills Shaw[12], W. Kimryn Rathmell[6], Victor E. Reuter[17], Ed Reznik[17], Carl Simon Shelley[19], Brian Shuch[5], Sabina Signoretti[20], Ramaprasad Srinivasan[1], Pheroze Tamboli[12], George Thomas[2], Satish Tickoo[17]

International Genomics Consortium Biospecimen Core Resource:

Kenneth Burnett[21], Daniel Crain[21], Johanna Gardner[21], Kevin Lau[21], David Mallery[21], Scott Morris[21], Joseph D Paulauskis[21], Robert J Penny[21], Candace Shelton[21], W. Troy Shelton[21], Mark Sherman[21], Eric Thompson[21], Peggy Yena[21]

Nationwide Children's Hospital Biospecimen Core Resource:

Melissa T. Avedon[22], Jay Bowen[22], Julie M. Gastier-Foster[22], Mark Gerken[22], Kristen M. Leraas[22], Tara M. Lichtenberg[22], Nilsa C. Ramirez[22], Tracie Santos[22], Lisa Wise[22], Erik Zmuda[22]

TCGA Project Team:

John A. Demchok[23], Ina Felau[23], Carolyn M. Hutter[24], Margi Sheth[23], Heidi J. Sofia[24], Roy Tarnuzzer[23], Zhining Wang[23], Liming Yang[23], Jean C. Zenklusen[23], Jiashan (Julia) Zhang[23]

Data Coordinating Center:

Brenda Ayala[25], Julien Baboud[25], Sudha Chudamani[26], Jia Liu[26], Laxmi Lolla[26], Rashi Naresh[25], Todd Pihl[25], Qiang Sun[25], Yunhu Wan[25], Ye Wu[26]

Genome Characterization/Sequencing Centers:

Adrian Ally[7], Miruna Balasundaram[7], Saianand Balu[6], Rameen Beroukhim[4], Tom Bodenheimer[6], Christian Buhay[3], Yaron S.N. Butterfield[7], Rebecca Carlsen[7], Scott L. Carter[4], Hsu Chao[3], Eric Chuah[7], Amanda Clarke[7], Kyle R. Covington[3], Mahmoud Dahdouli[3], Ninad Dewal[3], Noreen Dhalla[7], HarshaVardhan Doddapaneni[3], Jennifer A. Drummond[3], Stacey B. Gabriel[4], Richard A. Gibbs[3], Ranabir Guin[7], Walker Hale[3], Alicia Hawes[3], D. Neil Hayes[6], Robert A. Holt[7], Alan P. Hoyle[6], Stuart R. Jefferys[6], Steven J.M. Jones[7], Corbin D. Jones[6], Divya Kalra[3], Christie Kovar[3], Lora Lewis[3], Jie Li[3], Yussanne Ma[7], Marco A. Marra[7], Michael Mayo[7], Shaowu Meng[6], Matthew Meyerson[4], Piotr A. Mieczkowski[6], Richard A. Moore[7], Donna Morton[3], Lisle E. Mose[6], Andrew J. Mungall[7], Donna Muzny[3], Joel S. Parker[6], Charles M. Perou[6], Jeffrey Roach[6], Jacqueline E. Schein[7], Steven E. Schumacher[4], Yan Shi[6], Janae V. Simons[6], Payal Sipahimalani[7], Tara Skelly[6], Matthew G. Soloway[6], Carrie Sougnez[4], Angela Tam[7], Donghui Tan[6], Nina Thiessen[7], Umadevi Veluvolu[6], Min Wang[3], Matthew D. Wilkerson[6], Tina Wong[7], Junyuan Wu[6], Liu Xi[3], Jane Zhou[3]

Genome Data Analysis Centers:

Jason Bedford[5], Fengju Chen[3], Yao Fu[5], Mark Gerstein[5], David Haussler[27], Katayoon Kasaian[7], Phillip Lai[9], Shiyun Ling[12], Amie Radenbaugh[27], David Van Den Berg[9], John N. Weinstein[12], Jingchun Zhu[27]

Tissue Source Sites:

Monique Albert[28], Iakovina Alexopoulou[29], Jeremiah J Andersen[19], J. Todd Auman[6], John Bartlett[28], Sheldon Bastacky[30], Julie Bergsten[31], Michael L. Blute[32], Lori Boice[6], Roni J. Bollag[33], Jeff Boyd[34], Erik Castle[18], Ying-Bei Chen[17], John C. Cheville[35], Erin Curley[21],

Benjamin Davies[30], April DeVolk[31], Rajiv Dhir[30], Laura Dike[36], John Eckman[31], Jay Engel[37], Jodi Harr[31], Ronald Hrebinko[30], Mei Huang[6], Lori Huelsenbeck-Dill[38], Mary Iacocca[38], Bruce Jacobs[30], Michael Lobis[38], Jodi K. Maranchie[30], Scott McMeekin[39], Jerome Myers[31], Joel Nelson[30], Jeremy Parfitt[40], Anil Parwani[30], Nicholas Petrelli[38], Brenda Rabeno[38], Somak Roy[30], Andrew L. Salner[41], Joel Slaton[39], Melissa Stanton[18], R. Houston Thompson[35], Leigh Thorne[6], Kelinda Tucker[31], Paul M. Weinberger[33], Cythnia Winemiller[31], Leigh Anne Zach[31], Rosemary Zuna[39]

Author Affiliations:

1. Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892
2. Oregon Health & Science University, Portland, OR 97239
3. Baylor College of Medicine, Houston, TX 77030
4. The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, MA 02142
5. Yale University, New Haven, CT 06520
6. University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
7. Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 4S6
8. Van Andel Research Institute, Grand Rapids, MI 49503
9. University of Southern California, Los Angeles, CA 90033
10. Johns Hopkins University, Baltimore, MD 21287
11. H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612
12. Univ. of Texas MD Anderson Cancer Center, Houston, TX, 77030
13. Brown University, Providence, RI, 02912, USA
14. Dana-Farber Cancer Institute, Boston, MA 02215
15. Asterand Bioscience, Detroit, MI 48202
16. University of Kansas Medical Center, Kansas City, KS 66160
17. Memorial Sloan Kettering Cancer Center, New York, NY 10065
18. Mayo Clinic, Scottsdale, Arizona 85255
19. Gundersen Medical Foundation, La Crosse, WI 54601
20. Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115
21. The International Genomics Consortium, Phoenix, AZ 85004
22. The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205
23. National Cancer Institute, National Institutes of Health, Bethesda, MD 20892
24. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892
25. SRA International, Inc., 4300 Fair Lakes Court, Fairfax, VA 22033
26. Leidos Biomedical Research, Inc. Frederick National Laboratory for Cancer Research, Rockville MD 20850
27. University of California Santa Cruz Genomics Institute, Santa Cruz, CA 95064
28. Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada

29. St. Joseph's Healthcare Hamilton, Hamilton, Ontario L8N 3Z5, Canada
30. University of Pittsburgh Medical Center Presbyterian University Hospital, Pittsburgh, PA 15213
31. Penrose-St. Francis Health Services, Colorado Springs, CO 80907
32. Massachusetts General Hospital, Boston, MA 02114
33. Georgia Regents University, Augusta GA 30912
34. Fox Chase Cancer Center, Philadelphia, PA 19111
35. Mayo Clinic, Rochester, MN 55905
36. TechOne Site 501, Detroit, MI 48202
37. Kingston General Hospital, Kingston, Ontario K7L 5H6, Canada
38. Helen F Graham Cancer Center at Christiana Care Health Systems, Newark, DE 19713
39. University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104
40. London Health Sciences Centre, London, Ontario N6A 5A5, Canada
41. Gray Cancer Center Hartford Hospital, Hartford, CT 06106

Experimental Procedures

Biospecimens

Sample inclusion criteria

Surgical resection biospecimens were collected from patients diagnosed with renal papillary carcinoma, and had not received prior treatment for their disease (chemotherapy or radiotherapy). Institutional review boards at each tissue source site reviewed protocols and consent documentation and approved submission of cases to TCGA. Cases were staged according to the American Joint Committee on Cancer (AJCC). Each frozen primary tumor specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the tissue source site). Specimens were shipped overnight using a cryoport that maintained an average temperature of less than -180°C.

Pathology quality control was performed on each tumor and normal tissue (if available) specimen from either a frozen section slide prepared by the BCR or from a frozen section slide prepared by the TSS. Hematoxylin and eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumor specimen was histologically consistent with the allowable renal papillary carcinomas and the adjacent tissue specimen contained no tumor cells. Tumor-adjacent tissue from Type II papillary carcinomas was acceptable as the sole germline control. Tumor-adjacent tissues from papillary carcinomas other than Type II were characterized if accompanied by DNA from a patient-matched blood specimen. The percent tumor nuclei, percent necrosis, and other pathology annotations were also assessed. Tumor samples with ≥60% tumor nuclei and ≤20% or less necrosis were submitted for nucleic acid extraction.

The TSSs contributing biospecimens included in this manuscript include: ABS, Asterand, Inc., Baylor, Catholic Health Initiative - Penrose St. Francis Health Services, Catholic Health Initiative - St. Joseph's Medical Center Cancer Institute, Christiana Care Health Services, Inc., Cleveland Clinic, Fox Chase Cancer Center, Gundersen Lutheran, Hartford Hospital, International Genomics Consortium, ILSbio, LLC., Mayo Clinic, MD Anderson, Memorial Sloan Kettering Cancer Center, National Cancer Institute Urologic Oncology Branch, Ontario Institute for Cancer Research, Roswell Park Cancer Institute, University of North Carolina, University of Oklahoma Health Sciences Center, and the University of Pittsburgh.

Approximately 83% of renal papillary carcinoma cases (consisting of a primary tumor and a germline control) submitted to the BCR and processed passed quality control metrics. Tumor tissue from 209 cases was submitted for reverse phase protein array analysis. The data freeze included 161 cases from KIRP batches 51, 71, 162, 194, 209, 246, 266, 281, 299, and 325.

Sample Processing

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

RNA samples were quantified by measuring Abs_{260} with a UV spectrophotometer and DNA quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify that tumor DNA and germline DNA representing a case were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen (Hilden, Germany) for REPLI-g whole genome amplification using a 100 µg reaction scale. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only analytes with RIN ≥7.0 were included in this study. Only cases yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study.

Samples with residual tumor tissue were considered for proteomics analysis. When available, a 10 to 20 mg piece of snap-frozen tumor adjacent to the piece used for molecular sequencing and characterization was submitted to MD Anderson for reverse phase protein array analysis.

Histopathological Review

The tumor samples collected for the TCGA KIRP project were selected based on an histological diagnosis of papillary renal cell carcinoma (PRCC). While this is a perfectly acceptable selection criterion, it does not take into account the greater degree of heterogeneity present within this selection criterion than for instance the selection criteria of clear cell histology for the TCGA KIRC project. In general, papillary kidney tumors can be separated histologically into Type 1 PRCC, that has a distinct specific histology, or Type 2 PRCC, that can represent multiple types of histologies that are not Type 1 PRCC. Additionally, some papillary kidney tumors may demonstrate no distinct histology other that general papillae-like histology and are given no specific type but described as unclassified papillary RCC. These differing histological types may represent quite different tumors and, although some of the submitted tumors had been assigned a papillary type, this was not sufficient to effectively produce accurate sub-groups. Due to the potential importance of the heterogeneity, it was decided that the tumor samples should be reviewed by a panel of six pathologists with experience and expertise in assessing papillary kidney cancer to produce a greater depth of analysis for these samples and produce three basic sub-groups consisting of either Type 1 PRCC, Type 2 PRCC or Unclassified PRCC. These sub-groups could be used for specific analyses, such as mutation analysis, as the present of significantly mutated genes specific to one subtype may be obscured when all the tumors are assessed as a single group. Furthermore, the accurate diagnosis of papillary kidney cancer can be challenging and all samples were re-assessed using available data to confirm their status as papillary kidney tumors and allow the removal of any questionable samples.

The six person pathology review panel assessed diagnostic slides from 131 of the potential 167 papillary kidney tumors and provided one of the following classifications for each slide:

- Type 1 papillary RCC
- Type 2 papillary RCC, including:
  - HLRCC
  - TFE3 RCC
  - Papillary RCC with oncocytic features
  - Papillary RCC with clear cell
- Unclassified papillary RCC
- Not papillary RCC, including:
  - Metanephric Adenoma/Tumor
  - Mucinous tubulo and spindle

For each of these 136 tumors a sub-grouping of either Type 1 PRCC, Type 2 PRCC or Unclassified PRCC was assigned based on the consensus of these six reviews. For the remaining 36 tumors that were not assessed by the six person pathology review panel the original TCGA pathology reports were reassessed to provide a confirmed sub-group where possible. Of the potential papillary tumors, 161 were confirmed to be PRCC consisting of 75 Type 1 PRCCs, 60 Type 2 PRCCs and 26 Unclassified PRCC. The remaining 6 tumors were either urothelial cancer (TCGA-A4-7287, TCGA-B3-4104, TCGA-HE-7130) or simply not PRCC (TCGA-A4-7828, TCGA-AL-3467, TCGA-HE-A5NK).

Copy Number Analysis Methods

SNP Array-Based Copy Number Analysis:

DNA from each tumor or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described.[1] Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy number at each probe locus.[2] For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor.[3,4] This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation.[5] As part of this process of copy number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from the TCGA germline samples from the ovarian cancer analysis and from samples of this cohort. Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile.[6] Significance of copy number alterations were assessed from the segmented data using GISTIC2.0 (Version 2.0.22).[5] Briefly, GISTIC2.0 deconstructs somatic copy-number alterations into broad and focal events and applies a probabilistic framework to identify location and significance levels of somatic copy-number alterations. For the purpose of this analysis, we defined an arm-level event as any event spanning more than 50% of a chromosome arm. For copy number based clustering, tumors were clustered based on log2 copy number at regions revealed by GISTIC analysis. Clustering was done in R based on Euclidean distance using Ward's method. Allelic and integer copy number, tumor purity, and tumor ploidy were calculated using the ABSOLUTE algorithm.[7]

Exome Mutation Analysis Methods

Library Preparation: Illumina HiSeq

DNA samples were constructed into Illumina paired-end pre-capture libraries according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications as described in the *BCM-HGSC Illumina Barcoded Paired-End Capture Library Preparation* protocol. Libraries were prepared using Beckman robotic workstations (Biomek NXp and FXp models). The complete protocol and oligonucleotide sequences are accessible from the HGSC website https://www.hgsc.bcm.edu/sites/default/files/documents/Illumina_Barcoded_Paired-End_Capture_Library_Preparation.pdf. Briefly, 1 ug of DNA in 100ul volume was sheared into fragments of approximately 300-400 base pairs in a Covaris plate with E210 system (Covaris, Inc. Woburn, MA) followed by end-repair, A-tailing and ligation of the Illumina multiplexing PE adaptors. Pre-capture Ligation Mediated-PCR (LM-PCR) was performed for 6-8 cycles of amplification using the 2X SOLiD Library High Fidelity Amplification Mix (a custom product manufactured by Invitrogen). Universal primer IMUX-P1.0 and a pre-capture barcoded primer IBC were used in the PCR amplification. In total, a set of 12 such barcoded primers were used on these samples. Purification was performed with Agencourt AMPure XP beads after enzymatic reactions. Following the final XP beads purification, quantification and size distribution of the pre-capture LM-PCR product was determined using the LabChip GX electrophoresis system (PerkinElmer).

Exome Capture

For the hybridization step, four pre-capture libraries were pooled together (~250 ng/sample). These pooled libraries were then hybridized in solution to the HGSC VCRome 2.1 design (42Mb, NimbleGen) according to the manufacturer's protocol *NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2)* with minor revisions.[8] Human COT1 DNA and full-length Illumina adaptor-specific blocking oligonucleotides were added into the hybridization to block repetitive genomic sequences and the adaptor sequences. Post-capture LM-PCR amplification was performed using the 2X SOLiD Library High Fidelity Amplification Mix with 14 cycles of amplification. After the final AMPure XP bead purification, quantity and size of the capture library was analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The efficiency of the capture was evaluated by performing a qPCR-based quality check on the four standard NimbleGen internal controls. Successful enrichment of the capture libraries was estimated to range from a 6 to 9 of ΔCt value over the non-enriched samples.

Sequencing

Library templates were prepared for sequencing using Illumina's cBot cluster generation system with TruSeq PE Cluster Generation Kits. Briefly, these libraries were denatured with sodium hydroxide and diluted to 6-9 pM in hybridization buffer in order to achieve a load density of ~800K clusters/mm. Each library pool was loaded in a single lane of a HiSeq flow cell, and each lane was spiked with 1% phiX control library for run quality control. The sample libraries then underwent bridge amplification to form clonal clusters, followed by hybridization with the sequencing primer. Sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 platform. Using the TruSeq SBS Kits, sequencing-by-synthesis reactions

were extended for 101 cycles from each end, with an additional 7 cycles for the index read. Sequencing runs generated approximately 300-400 million successful reads on each lane of a flow cell, yielding 9-11 Gb per sample. With these sequencing yields, samples achieved an average of 94% of the targeted exome bases covered to a depth of 20X or greater.

Sequence Alignment / Primary Data Analysis

Initial sequence analysis was performed using the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/content/mercury). First, the primary analysis software on the instrument produces .bcl files that are transferred off-instrument into the HGSC analysis infrastructure by the HiSeq Real-time Analysis module. Once the run is complete and all .bcl files are transferred, Mercury runs the vendor's primary analysis software (CASAVA), which demultiplexes pooled samples and generates sequence reads and base-call confidence values (qualities). The next step is the mapping of reads to the GRCh37 Human reference genome (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) using the Burrows-Wheeler aligner (BWA[9], http://bio-bwa.sourceforge.net/) and producing a BAM (binary alignment/map) file.[10] The third step involves quality recalibration (using GATK[11], http://www.broadinstitute.org/gatk/), and where necessary the merging of separate sequence-event BAMs into a single sample-level BAM. BAM sorting, duplicate read marking, and realignment to improve in/del discovery all occur at this step.

Mutation Detection

Mutation calling by multiple analysis centers was carried out, essentially as previously described.[12] Mutations were called using tumor and matched normal BAM files by the BCM HGSC, Broad Institute, University of California, Santa Cruz (UCSC), and British Colombia Genome Sequencing Center. The maf file is available from the following link, https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/kirp/gsc/hgsc.bcm.edu/illuminaga_dnaseq_curated/mutations/hgsc.bcm.edu_KIRP.IlluminaGA_DNASeq_curated.Level_2.1.0.0/. Candidate germline mutations in *FH* gene were called by BCM HGSC, with mutations that were either nonsence or insertions/deletions (n=4 cases) or other single nucleotide variants with no entry in dbSNP (n=1) being used for downstream analyses. Candidate germline mutations in the *MET* gene similarly were called by BCM HGSC, with known familial kidney cancer associated mutations (n=3) that being used for downstream analyses. For analysis of Significantly Mutated Genes (SMGs), MutSigCV algorithm was applied[13], using the set of mutation calls made concurrently by at least two of the four mutation calling centers. Four tumors, TCGA-A4-8516, TCGA-AL-3471, TCGA-B1-7332, TCGA-DW-7836, failed to produce sequencing data due to technical difficulties.

Sub-Clonal Analysis

Sub-clonal analysis was performed using the methodology described in Totoki et al.[14] In brief, the median allele fraction for somatic mutations was calculated for all subjects and considered to be the clonal allele fraction. For each recurrently mutated gene (mutated 5 times or more) we calculated the paired Wilcoxon p-value of the difference in allele fractions between the gene and the median for the matched sample. P-values and false discovery rate (FDR) adjusted values are shown in the table along with the counts for each gene and the number of times the gene was seen at sub-clonal levels.

DNA Methylation Analysis

Array-based DNA methylation assay

We used the Illumina Infinium BeadChips (Illumina, San Diego, CA) to obtain DNA methylation profiles of 172 TCGA papillary renal cell carcinoma samples and 50 normal (non-tumor adjacent) kidney samples (156 tumors and 45 normals on the HumanMethylation450 (HM450) platform, and 16 tumors and 5 normals on the Illumina Infinium HumanMethylation27 (HM27) platform). The HM27 array targets 27,578 loci near the Transcription Start Site of 14,475 consensus coding regions in the NCBI database. The Infinium HM450 array, which incorporates nearly all the HM27 probes, targets 482,421 CpG sites and covers 99% of RefSeq genes as well as intergenic regions, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. This platform covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them. The assay probe sequences and information for each interrogated CpG site by the HM450 and HM27 platforms can be found in the MAGE-TAB ADF (Array Design Format) file deposited on the TCGA Data Portal.

We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described.[15] All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified (WGA) and enzymatically fragmented prior to hybridization to the arrays. BeadArrays were scanned using the Illumina iScan technology, and the IDAT files (Level 1 data) were used to extract the intensities (Level 2 data) and calculate the beta value (Level 3 data) for each probe and sample with the R-based *methylumi* package. Dye-bias normalization and normalization described in Triche et al.[16] were performed in the same process.

The level of DNA methylation at each CpG locus is summarized as beta (β) value calculated as (M/(M+U)), ranging from 0 to 1, which represents the ratio of the methylated probe intensity to the overall intensity at each CpG locus. A p value comparing the intensity for each probe to the background level was calculated with the methylumi package at the same time, and data points with a detection p value >0.05 were deemed not significantly different from background measurements, and therefore were masked as "NA" in the Level 3 in both HM27 and HM450 data packages, as detailed below.

TCGA Data Packages

The three data levels are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). Please note that with continuing updates of genomic databases, data archive revisions become available at the TCGA Data Portal.

Level 1 - Level 1 data contain raw IDAT files. IDAT files are the direct output from the scanning program.

Level 2 - Level 2 data contain background corrected signal intensities of the M and U probes.

Level 3 - Level 3 data files contain β-value calculations and masked data points with "NA" from the probes that are annotated as having a SNP within 10 base pairs or repeat within 15 base pairs of the interrogated locus. The genomic characteristics for each probe are available for download via Illumina (www.illumina.com).

CDKN2A Epigenetic Silencing

Exon-level gene expression data were downloaded from the TCGA data, and reads per kilobase per million (RPKM) values for exon 1a (chr9:21975038-21974403) of *CDKN2A* was used to assess the expression of $p16^{INK4a}$. The DNA methylation level as interrogated by cg13601799 was used for *CDKN2A*.[17] The DNA copy number data were obtained from Firehose (gdac.broadinstitute.org). The expression level of exon 1a inversely correlates with DNA methylation at the interrogated CpG (Figure S7b). A beta value of 0.2 or above is considered evidence of epigenetic silencing. RPKM of Exon 1a is 0 or close to 0 for all eleven samples that met this criterion.

Unsupervised Clustering

The shared probe set between HM27 and HM450 platforms (N=25,978) were used for this analysis. We removed probes that contained any masked data due to detection p value, repeats and SNPs and non-uniquely mapped probes (n=23,381 remaining). The R package FDb.InfiniumMethylation.hg19 was used to obtain chromosomal locations of the probes and 961 X-linked probes were removed from the analysis. A standard deviation for each probe was calculated across the 161 tumors in the final data freeze. We chose probes that were unmethylated in the 50 normal samples (mean beta value <0.1) that had a standard deviation of greater than 0.15 (n=343) in the tumors for the clustering. Hierarchical clustering with Ward's method was used to cluster the 161 data freeze samples, and the clustering dendrogram was cut at k=3 to yield three clusters after assessing cluster strength and stability. One of the clusters (cluster number = 3) exhibited extensive hypermethylation across thousands of CpG loci and was renamed CpG island methylator phenotype (CIMP).

mRNA Analysis Methods

Sequencing and quantification

One µg of total RNA was converted to mRNA libraries using the lllumina mRNA TruSeq kit (RS-122-2001 or RS-122-2002) following the manufacturer's directions. Libraries were sequenced 48x7x48bp on the Illumina HiSeq 2000 as previously described.[17] FASTQ files were generated by CASAVA. RNA reads were aligned to the hg19 genome assembly using MapSplice 0.7.4.[18] Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1 (http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf), using RSEM[19] and normalized within-sample to a fixed upper quartile. For further details on this processing, refer to Description file at the DCC data portal under the V2_MapSpliceRSEM workflow (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/kirp/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KIRP.IlluminaHiSeq_RNASeqV2.mage-tab.1.17.0/DESCRIPTION.txt). FASTQ and BAM files can be found at CGHUB (https://cghub.ucsc.edu). Quantification of genes, transcripts, exons and junctions can be found at the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/).

Gene filtering and generation of tumor clusters

Level3 RNA-seq upper quartile normalized RSEM data for 161 tumor samples were retrieved. Gene expression values were log2 transformed. Genes that were present in more than 80% of samples were included for analysis. The maximum absolute deviation (maximum value minus the average) was calculated for each gene, and the top 10% (n=2050) were included for clustering. Clustering was done using the ConsensusClusterPlus package for R.[20] Clusters were generated using Pearson correlation, 1000 iterations, and 80% resampling of tumors as parameters. The three subtypes were selected by a combination of significant decrease in cophenetic correlation and visual inspection of the consensus clustering matrices (Figure S11a). Clustering was verified using non-negative matrix clustering (data not shown). The three clusters contained 77 (Cluster 1), 61(Cluster 2), and 23 (Cluster 3) samples, respectively.

Kaplan-Meier analysis of tumor clusters

The Kaplan-Meier method was used to determine overall survival. Overall survival was defined as the time from the nephrectomy to death of any cause. Log-rank test was used to assess the survival differences of the three mRNA clusters.

Fusion Gene Analysis

TCGA RNA sequencing data (RNA FastQ files) was downloaded for the entire KIRP set from TCGA cancer genomics hub (CGHub TCGA). We used deFuse[21] software version 0.6.1 with default settings to detect fusion genes. The defuse results were further filtered by removing identified read through fusions, selecting coding regions, selecting in-frame (ORF) genes and selecting samples with a defuse confidence score of >80%. Our sample set included 11 tissue adjacent normal (TAN) samples; any fusions that were also identified in the TAN sample set were removed from analysis. This filtering resulted in a list of candidate fusion genes. To

characterize these candidate fusion genes we took each spanning junction read and using the BLAT tool in UCSC genome browser examined where the reads mapped. The fusions that mapped with 100% identity to each part of the identified fusion (gene1 or gene2) were selected for further analysis. This filter removed genes that mapped to multiple locations. Next, each RNA BAM from candidate fusion genes was examined in IGV, looking for stacked soft clipped reads, changes in coverage, at the identified fusion breakpoints. The sequence of each soft clipped read was brought into the UCSC genome browser and mapped using BLAT. Only fusions that had reads that matched (100%) the identified fusion genes were considered further. Next, the IGV read surrounding the breakpoint, the fusion spanning read and the read continuing onto the partner gene were brought into EMBOSS sixpack, EMBOSS (http://www.ebi.ac.uk/Tools/st/emboss_sixpack/) and translated in 6 frames. The translated product that matched both identified fusion genes were considered further. The Cbio data portal (http://www.cbioportal.org/) was used to examine copy number data and gene expression data for each fusion identified.

MicroRNA (miRNA) Analysis Methods

miRNA library construction, sequencing and analysis

We generated microRNA sequence (miRNA-seq) data for 161 tumor and 32 adjacent normal samples using methods described previously.[22] We aligned reads to the GRCh37/hg19 reference human genome, and annotated miRNA read count abundance with miRBase v16. While we used only exact-match read alignments for this, the BAM files that are available from cgHUB (cghub.ucsc.edu[23]) include all sequence reads. We used miRBase v20 to assign 5p and 3p mature strand names to MIMAT accession IDs.

We identified groups of samples that had similar abundance profiles using unsupervised non-negative matrix factorization (NMF, v0.5.06) consensus clustering with default settings.[24] The input was a reads-per-million (RPM) data matrix for the ~300 (25%) most-variant 5p or 3p mature strands, which we parsed from the level 3 isomiR data files that are available from the TCGA data portal. After running a rank survey with 30 iterations per solution, we chose a preferred clustering solution from the cophenetic and average silhouette width score profiles, and then did a 500-iteration the main clustering run. Among other files, we wrote out a metagene (W) matrix that identified miRs that were discriminatory for NMF. We calculated a profile of silhouette widths from the NMF consensus membership matrix, and considered samples with relatively low widths within a cluster as atypical cluster members.

To generate a heatmap for the NMF results, we first identified miRs that were differentially abundant between the unsupervised miRNA clusters, using a multiclass analysis with SAMseq (samr 2.0[25]) in R 2.15.0, with a read-count input matrix and an FDR threshold of 0.05. For the heatmap displayed, we included the 40 miRs that had both the largest SAMseq scores, and median abundances greater than 25 RPM. The RPM filtering acknowledged potential sponge effects from competitive endogeneous RNAs (ceRNAs) that can make weakly abundant miRs less influential.[26,27] Adding the RPM data matrix for adjacent normal samples to the NMF-ordered 40 miR x 161 tumor RPM data matrix, we transformed each row of the matrix by $\log_{10}(\text{RPM} + 1)$, then used the pheatmap v0.7.7 R package to scale and then cluster only the rows, with a Euclidean distance measure.

For clinical and molecular covariates, we calculated contingency table association *P*-values with a Fisher exact test in R, setting the workspace size to $2 \cdot 10^9$.

To identify miRs that were differentially abundant between pairs of sample groups, we used unpaired two-class SAMseq analyses (FDR < 0.05). Given potential ceRNA effects (above), we support assessing fold change at the same time as absolute miR abundance by adding, to each fold change barplot, a boxwhisker plot that shows the distribution of miR abundance in the two sample groups, and we show only miRs that had a median abundance greater than 50 RPM in at least one of the two groups being compared.

We assessed potential miRNA targeting by calculating miR-mRNA and miR-RPPA Spearman correlations with the MatrixEQTL v2.1.1[28] R package, using an RNAseq (RSEM) and RPPA gene-level normalized abundance data matrices from Firehose (gdac.broadinstitute.org). Restricting the RSEM matrix to the 14732 genes with a mean RSEM above the 30th percentile, and the miRs to the 606 with a mean RPM above the 60th percentile (0.46 RPM), we calculated correlations with a P-value threshold of 0.05, and filtered the resulting anticorrelations at FDR<0.05. We then extracted miR-gene pairs that corresponded to a) functional validation publications reported by MiRTarBase v4.5[29], for stronger (luciferase reporter, qPCR, Western blot) vs. weaker experimental evidence types. For the genes satisfying the FDR and miRTarBase functional validation filters (strong evidence types), we identified enriched KEGG pathways (DAVID v6.7).[30,31] We displayed results with Cytoscape 2.8.3.[32]

Tumor sample purity was calculated by the Broad Institute using ABSOLUTE.[33]

We visualized the global relationship of somatic copy number alterations and miRNA unsupervised clustering by generating a heatmap from tumor samples for hg19 'seg' data, using IGV 2.3.40.[34] We identified miRNAs that were potentially dysregulated by somatic copy number alterations by calculating Spearman correlations with BH-adjusted P-values between the Gistic2 all_data_by_genes file and pre-miRNA RPM abundance file. We used the corresponding all_thresholded_by_genes file to compare integer sCNA levels across miRNA clusters.

Analysis of miRs targeting *CDKN2A*

CDKN2A is altered through mutation, copy number loss, or epigenetic silencing by DNA methylation in 23 samples, which were associated with type 2 pathology and were significantly associated with survival (p<1e-11, Fig. S7e). We assessed miR targeting as an additional mechanism that may alter *CDKN2A* transcript levels and so may represent an alteration that is functionally important in samples that lack a clear driver (i.e. dark matter samples).

We first used miRTarBase v4.5[29] to identify miRs that have been functionally validated as directly targeting *CDNK2A*, considering records with evidence from Luciferase reporters, qPCR, and/or Western blots. We determined which of these miRs had a strong Spearman correlation with *CDKN2A* in either the tumor dataset, or the tumor-and-normal dataset.

For each miR that was correlated to *CDKN2A* transcript levels, we identified the expression value that stratified patients into two groups that minimized the Kaplan-Meier log-rank p-value.[35] We used these expression thresholds to define the samples in which miR targeting was likely influencing *CDKN2A* transcript levels. We added these potential miR-targeting samples to the original 23 *CDKN2A*-altered samples, and compared the Kaplan-Meier results for this extended sample set to results for the 23 samples (Fig. S7e). Finally, we used bootstrapping to assess the extended sample sets. Starting with the 23 *CDKN2A*-altered cases, we added additional samples chosen at random 50,000 times, and calculated the frequency of obtaining a smaller log-rank p-value and larger hazard ratio than in we had seen from the actual extended sample sets.

Reverse Phase Protein Array (RPPA) Methods

RPPA sample preparation and analysis

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 nmol/L Hepes (pH 7.4), 150 nmol/L NaCl, 1.5 nmol/L MgCl2, 1 mmol/L EGTA, 100 nmol/L NaF, 10 nmol/L NaPPi, 10% glycerol, 1 nmol/L phenylmethylsulfonyl fluoride, 1 nmol/L Na_3VO_4, and aprotinin 10 Ag/mL) from human tumors and RPPA was performed as described previously.[36-40] Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually diluted in fivefold serial dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 193 validated primary antibodies (see table below) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation catalyzed system and DAB colorimetric reaction. Slides were scanned in CanoScan 9000F. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI[38,40], available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model.[36] During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[40] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[38,40,41] using median centering across antibodies (level 3 data). In total, 193 antibodies and 125 samples were used. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described.[42] These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and used with caution based on degree of validation by criteria previously described.[42]

Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

List of Antibodies Used for Sample Profiling in RPPA Analysis.

Full Slide Name (Ab Name + Slide ID)	Protein Name	Gene Name	Ab Validation Status	Ab Origin	Company	Catalog #
X14.3.3_beta.R.V_GBL11066140	14-3-3-beta	YWHAB	Valid	Rabbit	Santa Cruz	sc-628
X14.3.3_epsilon.M.C_GBL11066233	14-3-3-epsilon	YWHAE	Use with Caution	Mouse	Santa Cruz	sc-23957
X14.3.3_zeta.R.V_GBL11066141	14-3-3-zeta	YWHAZ	Valid	Rabbit	Santa Cruz	sc-1019
X4E.BP1.R.V_GBL11066045	4E-BP1	EIF4EBP1	Valid	Rabbit	CST	9452
X4E.BP1_pS65.R.V_GBL11066046	4E-BP1_pS65	EIF4EBP1	Valid	Rabbit	CST	9456
X4E.BP1_pT37_T46.R.V_GBL11066047	4E-BP1_pT37_T46	EIF4EBP1	Valid	Rabbit	CST	9459
X53BP1.R.E_GBL11066155	53BP1	TP53BP1	Valid	Rabbit	CST	4937
ACC_pS79.R.V_GBL11066049	ACC_pS79	ACACA ACACB	Valid	Rabbit	CST	3661
ACC1.R.E_GBL11066050	ACC1	ACACA	Under Evaluation	Rabbit	Epitomics	1768-1
ACVRL1.R.C_GBL11066174	ACVRL1	ACVRL1	Use with Caution	Rabbit	Abcam	ab108207
ADAR1.M.V_GBL11066236	ADAR1	ADAR	Valid	Mouse	Abcam	ab88574
Akt.R.V_GBL11066173	Akt	AKT1 AKT2 AKT3	Valid	Rabbit	CST	4691
Akt_pS473.R.V_GBL11066075	Akt_pS473	AKT1 AKT2 AKT3	Valid	Rabbit	CST	9271
Akt_pT308.R.V_GBL11066202	Akt_pT308	AKT1 AKT2 AKT3	Valid	Rabbit	CST	2965
AMPK_alpha.R.C_GBL11066051	AMPK-alpha	PRKAA1	Use with Caution	Rabbit	CST	2532
AMPK_pT172.R.V_GBL11066052	AMPK-alpha_pT172	PRKAA1	Valid	Rabbit	CST	2535
Annexin.1.M.E_GBL11066239	Annexin-I	ANXA1	Valid	Mouse	BD Biosciences	610066
Annexin_VII.M.V_GBL11066265	Annexin-VII	ANXA7	Valid	Mouse	BD Biosciences	610668
AR.R.V_GBL11066124	AR	AR	Valid	Rabbit	Epitomics	1852-1
A.Raf_pS299.R.C_GBL11066162	A-Raf	ARAF	Valid	Rabbit	CST	4432
ARHI.M.E_GBL11066273	ARHI	DIRAS3	Use with Caution	Mouse	MDACC Laboratory	Bast Lab
ATM.R.E_GBL11066053	ATM	ATM	Valid	Rabbit	CST	2873
Acetyl.a.Tubulin.Lys40.R.C_GBL11066205	Acetyl-a-Tubulin-Lys40	Several alpha tubulin genes	Use with Caution	Rabbit	CST	5335
alpha.Catenin.M.V_GBL11066255	alpha.Catenin	CTNNA1	Valid	Mouse	Calbiochem	CA1030
ASNS.R.V_GBL11066243	ASNS	ASNS	Valid	Rabbit	Sigma	HPA029318
B.Raf.M.C_GBL11066212	B-Raf	BRAF	Use with Caution	Mouse	Santa Cruz	sc-5284
Bad_pS112.R.V_GBL11066054	Bad_pS112	BAD	Valid	Rabbit	CST	9291
Bak.R.E_GBL11066055	Bak	BAK1	Use with Caution	Rabbit	Epitomics	1542-1
Bap1c.4.M.E_GBL11066238	BAP1	BAP1	Valid	Mouse	Santa Cruz	sc-28383
Bax.R.V_GBL11066056	Bax	BAX	Valid	Rabbit	CST	2772

Bcl.2.M.V_GBL11066211	Bcl2	BCL2	Valid	Mouse	Dako	M0887
Bcl.xL.R.V_GBL11066058	Bcl-xL	BCL2L1	Valid	Rabbit	CST	2762
Beclin.G.C_GBL11066241	Beclin	BECN1	Use with Caution	Goat	Santa Cruz	sc-10086
beta.Catenin.R.V_GBL11066057	b-Catenin	CTNNB1	Valid	Rabbit	CST	9562
Bid.R.C_GBL11066059	Bid	BID	Use with Caution	Rabbit	Abcam	ab32060
Bim.R.V_GBL11066060	Bim	BCL2L11	Valid	Rabbit	Abcam	ab32158
BRCA2.R.C_GBL11066126	BRCA2	BRCA2	Use with Caution	Rabbit	CST	9012
c.Jun_pS73.R.V_GBL11066066	c-Jun_pS73	JUN	Valid	Rabbit	CST	9164
c.Kit.R.V_GBL11066067	c-Kit	KIT	Valid	Rabbit	Abcam	ab32363
c.Met_pY1235.R.V_GBL11066121	c-Met_pY1234_Y1235	MET	Valid	Rabbit	CST	3129
c.Myc.R.C_GBL11066248	c-Myc	MYC	Use with Caution	Rabbit	Santa Cruz	sc-764
C.Raf.R.V_GBL11066133	C-Raf	RAF1	Valid	Rabbit	Millipore	04-739
C.Raf_pS338.R.E_GBL11066069	C-Raf_pS338	RAF1	Valid	Rabbit	CST	9427
Caspase.7_cleavedD198.R.C_GBL11066061	Caspase-7-cleaved	CASP7	Use with Caution	Rabbit	CST	9491
Caveolin.1.R.V_GBL11066062	Caveolin-1	CAV1	Valid	Rabbit	CST	3238
CD20.R.C_GBL11066063	CD20	MS4A1	Use with Caution	Rabbit	Epitomics	1632
CD31.M.V_GBL11066250	CD31	PECAM1	Valid	Mouse	Dako	M0823
CD49b.M.V_GBL11066252	CD49b	ITGA2	Valid	Mouse	BD Biosciences	611016
CDK1.R.V_GBL11066247	CDK1	CDC2-CDK1	Valid	Rabbit	CST	9112
Chk1.M.C_GBL11066237	Chk1	CHEK1	Use with Caution	Mouse	CST	2360
Chk1_pS345.R.C_GBL11066146	Chk1_pS345	CHEK1	Use with Caution	Rabbit	CST	2348
Chk2.M.E_GBL11066214	Chk2	CHEK2	Valid	Mouse	CST	3440
Chk2_pT68.R.E_GBL11066209	Chk2_pT68	CHEK2	Use with Caution	Rabbit	CST	2197
cIAP.R.V_GBL11066151	cIAP	BIRC2	Valid	Rabbit	Millipore	07-759
Claudin.7.R.V_GBL11066138	Claudin-7	CLDN7	Valid	Rabbit	Novus Biologicals	NB100-91714
Collagen_VI.R.V_GBL11066068	Collagen-VI	COL6A1	Valid	Rabbit	Santa Cruz	sc-20649
Cyclin_B1.R.V_GBL11066070	Cyclin-B1	CCNB1	Valid	Rabbit	Epitomics	1495-1
Cyclin_D1.R.V_GBL11066071	Cyclin-D1	CCND1	Valid	Rabbit	Santa Cruz	sc-718
Cyclin_E1.M.V_GBL11066215	Cyclin-E1	CCNE1	Valid	Mouse	Santa Cruz	sc-247
Cyclin_E2.R.C_GBL11066072	Cyclin-E2	CCNE2	Use with Caution	Rabbit	Epitomics	1142
DJ.1.R.E_GBL11066144	DJ1	PARK7	Valid	Rabbit	Abcam	ab76008
Dvl3.R.V_GBL11066153	Dvl3	DVL3	Valid	Rabbit	CST	3218
E.Cadherin.R.V_GBL11066175	E-Cadherin	CDH1	Valid	Rabbit	CST	3195
eEF2.R.C_GBL11066163	eEF2	EEF2	Use with Caution	Rabbit	CST	2332

eEF2K.R.V_GBL11066164	eEF2K	EEF2K	Valid	Rabbit	CST	3692
EGFR.R.V_GBL11066177	EGFR	EGFR	Valid	Rabbit	CST	2232
EGFR_pY1068.R.C_GBL11066073	EGFR_pY1068	EGFR	Use with Caution; also sees pHer2	Rabbit	CST	2234
EGFR_pY1173.R.V_GBL11066074	EGFR_pY1173	EGFR	Valid	Rabbit	Abcam	ab32578
eIF4E.R.V_GBL11066120	eIF4E	EIF4E	Valid	Rabbit	CST	9742
eIF4G.R.C_GBL11066180	eIF4G	EIF4G1	Use with Caution	Rabbit	CST	2498
ER.alpha.R.V_GBL11066076	ER-alpha	ESR1	Valid	Rabbit	Lab Vision	RM-9101-S
ER.alpha_pS118.R.V_GBL11066077	ER-alpha_pS118	ESR1	Valid	Rabbit	Epitomics	1091-1
ERCC1.M.V_GBL11066272	ERCC1	ERCC1	Valid	Mouse	Santa Cruz	sc-17809
ERK2.R.E_GBL11066078	ERK2	MAPK1	Under Evaluation	Rabbit	Santa Cruz	Sc-154
ETS.1.R.V_GBL11066206	Ets-1	ETS1	Valid	Rabbit	Bethyl	A303-501A
FASN.R.V_GBL11066203	FASN	FASN	Valid	Rabbit	CST	3180
Fibronectin.R.V_GBL11066079	Fibronectin	FN1	Valid	Rabbit	Epitomics	1574-1
FoxM1.R.V_GBL11066179	FoxM1	FOXM1	Valid	Rabbit	CST	5436
FOXO3a.R.C_GBL11066178	FoxO3a	FOXO3	Use with Caution	Rabbit	CST	2497
FOXO3a_pS318_S321.R.C_GBL11066080	FoxO3a_pS318_S321	FOXO3	Use with Caution	Rabbit	CST	9465
G6PD.M.V_GBL11066266	G6PD	G6PD	Valid	Mouse	Santa Cruz	sc-373887
Gab2.R.V_GBL11066154	Gab2	GAB2	Valid	Rabbit	CST	3239
GAPDH.M.C_GBL11066217	GAPDH	GAPDH	Use with Caution	Mouse	Life Technologies	AM4300
GATA3.M.V_GBL11066230	GATA3	GATA3	Valid	Mouse	BD Biosciences	558686
GSK3.alpha.beta.M.V_GBL11066218	GSK-3ab	GSK3A GSK3B	Valid	Mouse	Santa Cruz	sc-7291
GSK3.alphabeta_pS21_S9.R.V_GBL11066081	GSK-3ab_pS21_S9	GSK3A GSK3B	Valid	Rabbit	CST	9331
GSK3_pS9.R.V_GBL11066171	GSK-3b_pS9	GSK3B	Valid	Rabbit	CST	9336
HER2.M.V_GBL11066260	HER2	ERBB2	Valid	Mouse	Lab Vision	MS-325-P1
HER2_pY1248.R.C_GBL11066168	HER2_pY1248	ERBB2	Use with Caution; likely sees pEGFR	Rabbit	R&D Systems	AF1768
HER3.R.V_GBL11066149	HER3	ERBB3	Valid	Rabbit	Santa Cruz	sc-285
HER3_pY1289.R.C_GBL11066122	HER3_pY1289	ERBB3	Use with Caution	Rabbit	CST	4791
Heregulin.R.V_GBL11066143	Heregulin	NRG1	Valid	Rabbit	CST	2573
HSP70.R.C_GBL11066082	HSP70	HSPA1A	Use with Caution	Rabbit	CST	4872
IGFBP2.R.V_GBL11066083	IGFBP2	IGFBP2	Valid	Rabbit	CST	3922
INPP4B.R.V_GBL11066167	INPP4b	INPP4B	Valid	Rabbit	CST	4039
IRS1.R.V_GBL11066132	IRS1	IRS1	Valid	Rabbit	Millipore	06-248
JNK_pT183_pY185.R.V_GBL11066142	JNK_pT183_Y185	MAPK8	Valid	Rabbit	CST	4668

JNK2.R.C_GBL11066084	JNK2	MAPK9	Use with Caution	Rabbit	CST	4672
Ku80.R.C_GBL11066147	Ku80	XRCC5	Use with Caution	Rabbit	CST	2180
Lck.R.V_GBL11066085	Lck	LCK	Valid	Rabbit	CST	2752
LKB1.M.E_GBL11066219	LKB1	STK11	Under Evaluation	Mouse	Abcam	Ab15095
MAPK_pT202_Y204.R.V_GBL11066086	MAPK_pT202_Y204	MAPK1 MAPK3	Valid	Rabbit	CST	4377
MEK1.R.V_GBL11066087	MEK1	MAP2K1	Valid	Rabbit	Epitomics	1235-1
MEK1_pS217_S221.R.V_GBL11066169	MEK1_pS217_S221	MAP2K1 MAP2K2	Valid	Rabbit	CST	9154
MIG.6.M.V_GBL11066261	MIG6	ERRFI1	Valid	Mouse	Sigma-Aldrich	WH0054206M1
Mre11.R.C_GBL11066088	Mre11	MRE11A	Use with Caution	Rabbit	CST	4847
MSH2.M.V_GBL11066231	MSH2	MSH2	Valid	Mouse	CST	2850
MSH6.R.C_GBL11066165	MSH6	MSH6	Use with Caution	Rabbit	Novus Biologicals	22030002
mTOR.R.V_GBL11066089	mTOR	MTOR	Valid	Rabbit	CST	2983
mTOR_pS2448.R.C_GBL11066090	mTOR_pS2448	MTOR	Use with Caution	Rabbit	CST	2971
MYH11.R.V_GBL11066198	Myosin-11	MYH11	Valid	Rabbit	Novus Biologicals	1139
Myosin.IIa.pS1943.R.V_GBL11066204	Myosin-IIa_pS1943	MYH9	Valid	Rabbit	CST	5026
N.Cadherin.R.V_GBL11066091	N-Cadherin	CDH2	Valid	Rabbit	CST	4061
N.Ras.M.V_GBL11066254	N-Ras	NRAS	Valid	Mouse	Santa Cruz	sc-31
NDRG1_pT346.R.V_GBL11066182	NDRG1_pT346	NDRG1	Valid	Rabbit	CST	3217
NF.kB.p65_pS536.R.C_GBL11066092	NF-kB-p65_pS536	RELA	Use with Caution	Rabbit	CST	3033
NF2.R.C_GBL11066161	Merlin	NF2	Use with Caution	Rabbit	Novus Biologicals	1046
Notch1.R.V_GBL11066166	Notch1	NOTCH1	Valid	Rabbit	CST	3268
P.Cadherin.R.C_GBL11066245	P-Cadherin	CDH3	Use with Caution	Rabbit	CST	2130
p21.R.V_GBL11066064	p21	CDKN1A	Valid	Rabbit	Santa Cruz	sc-397
p27.R.V_GBL11066145	p27-Kip-1	CDKN1B	Valid	Rabbit	Abcam	ab32034
p27_pT157.R.C_GBL11066137	p27_pT157	CDKN1B	Use with Caution	Rabbit	R&D Systems	AF1555
p27_pT198.R.V_GBL11066139	p27_pT198	CDKN1B	Valid	Rabbit	Abcam	ab64949
p38.R.V_GBL11066093	p38	MAPK14	Valid	Rabbit	CST	9212
p38_pT180_Y182.R.V_GBL11066094	p38_pT180_Y182	MAPK14	Valid	Rabbit	CST	9211
p53.R.E_GBL11066244	p53	TP53	Use with Caution	Rabbit	CST	9282
p62.LCK.ligand.M.C_GBL11066267	p62-LCK-ligand	SQSTM1	Use with Caution	Mouse	BD Transduction Lab	610833
p70S6K.R.V_GBL11066096	p70-S6K1	RPS6KB1	Valid	Rabbit	Epitomics	1494-1
p70S6K_pT389.R.V_GBL11066097	p70-S6K_pT389	RPS6KB1	Valid	Rabbit	CST	9205
p90RSK.R.C_GBL11066125	P90RSK	RPS6KA1	Use with Caution	Rabbit	CST	9347
p90RSK_pT359_S363.R.C_GBL11066127	p90RSK_pT359_S363	RPS6KA1	Use with Caution	Rabbit	CST	9344

PAI.1.M.E_GBL11066220	PAI-1	SERPINE1	Valid	Mouse	BD Biosciences	612024
Paxillin.R.C_GBL11066098	Paxillin	PXN	Use with Caution	Rabbit	Epitomics	1500-1
PCNA.M.C_GBL11066222	PCNA	PCNA	Use with Caution	Mouse	Abcam	ab29
PDCD4.R.C_GBL11066135	Pdcd4	PDCD4	Use with Caution	Rabbit	Rockland	600-401-965
PDK1.R.V_GBL11066100	ATR	ATR	Caution	Rabbit	CST	2790
PDK1_pS241.R.V_GBL11066101	PDK1_pS241	PDPK1	Valid	Rabbit	CST	3061
PEA15.R.V_GBL11066157	PEA-15	PEA15	Valid	Rabbit	CST	2780
PEA15_pS116.R.V_GBL11066158	PEA-15_pS116	PEA15	Valid	Rabbit	Invitrogen	44-836G
PI3K.p110.alpha.R.C_GBL11066134	PI3K-p110-alpha	PIK3CA	Use with Caution	Rabbit	CST	4255
PI3K.p85.R.V_GBL11066102	PI3K-p85	PIK3R1	Valid	Rabbit	Millipore	06-195
PKC.alpha.M.V_GBL11066223	PKC-alpha	PRKCA	Valid	Mouse	Millipore	05-154
PKC.alpha_pS657.R.C_GBL11066103	PKC-alpha_pS657	PRKCA	Use with Caution	Rabbit	Millipore	06-822
PKC.delta_pS664.R.V_GBL11066152	PKC-delta_pS664	PRKCD	Valid	Rabbit	Millipore	07-875
PKC.pan_BetaII_pS660.R.V_GBL11066196	PKC-pan_BetaII_pS660	PRKCA, PRKCB PRKCD, PRKCE PRKCH, PRKCQ	Valid	Rabbit	CST	9371
PR.R.V_GBL11066104	PR	PGR	Valid	Rabbit	Abcam	ab32085
PRAS40_pT246.R.V_GBL11066123	PRAS40_pT246	AKT1S1	Valid	Rabbit	Life Technologies	441100G
PRDX1.R.V_GBL11066249	PRDX1	PRDX1	Valid	Rabbit	Sigma	HPA007730
PREX1.R.E_GBL11066207	PREX1	PREX1	Valid	Rabbit	Abcam	ab102739
PTEN.R.V_GBL11066106	PTEN	PTEN	Valid	Rabbit	CST	9552
Rab11.R.E_GBL11066172	Rab11	RAB11A RAB11B	Under Evaluation	Rabbit	CST	3539
Rab25.R.V_GBL11066201	Rab25	RAB25	Valid	Rabbit	CST	4314
Rad50.M.V_GBL11066258	Rad50	RAD50	Valid	Mouse	Millipore	05-525
Rad51.R.V_GBL11066210	Rad51	RAD51	Valid	Rabbit	CST	8875
Raptor.R.V_GBL11066183	Raptor	RPTOR	Valid	Rabbit	CST	2280
Rb.M.QC_GBL11066224	Rb	RB1	**Used for QC**	Mouse	CST	9309
Rb_pS807_S811.R.V_GBL11066105	Rb_pS807_S811	RB1	Valid	Rabbit	CST	9308
RBM15.R.V_GBL11066197	RBM15	RBM15	Valid	Rabbit	Novus Biologicals	21390002
Rictor.R.C_GBL11066184	Rictor	RICTOR	Use with Caution	Rabbit	CST	2114
Rictor_pT1135.R.V_GBL11066195	Rictor_pT1135	RICTOR	Valid	Rabbit	CST	3806
S6.R.E_GBL11066108	S6	RPS6	Under Evaluation	Rabbit	CST	2217
S6_pS235_S236.R.V_GBL11066109	S6_pS235_S236	RPS6	Valid	Rabbit	CST	2211
S6_pS240_S244.R.V_GBL11066110	S6_pS240_S244	RPS6	Valid	Rabbit	CST	2215
SCD1.M.V_GBL11066253	SCD	SCD	Valid	Mouse	Santa Cruz	sc-58420

SF2.M.V_GBL11066263	SF2	SRSF1	Valid	Mouse	Invitrogen	32-4500
Shc_pY317.R.E_GBL11066159	Shc_pY317	SHC1	Valid	Rabbit	CST	2431
Smad1.R.V_GBL11066150	Smad1	SMAD1	Valid	Rabbit	Epitomics	1649-1
Smad3.R.V_GBL11066131	Smad3	SMAD3	Valid	Rabbit	Abcam	ab40854
Smad4.M.V_GBL11066234	Smad4	SMAD4	Valid	Mouse	Santa Cruz	sc-7966
Src.M.V_GBL11066227	Src	SRC	Valid	Mouse	Millipore	05-184
Src_pY416.R.C_GBL11066111	Src_pY416	SRC LYN FYN LCK YES1 HCK	Use with Caution	Rabbit	CST	2101
Src_pY527.R.V_GBL11066112	Src_pY527	SRC YES1 FYN FGR	Valid	Rabbit	CST	2105
STAT3_pY705.R.V_GBL11066113	Stat3_pY705	STAT3	Valid	Rabbit	CST	9131
STAT5.alpha.R.V_GBL11066114	Stat5a	STAT5A	Valid	Rabbit	Abcam	ab32043
Stathmin.R.V_GBL11066119	Stathmin-1	STMN1	Valid	Rabbit	Abcam	ab52630
Syk.M.V_GBL11066259	Syk	SYK	Valid	Mouse	Santa Cruz	sc-1240
TAZ.R.V_GBL11066128	TAZ	WWTR1	Valid	Rabbit	CST	2149
TFRC.R.V_GBL11066199	TFRC	TFRC	Valid	Rabbit	Novus Biologicals	22500002
TIGAR.R.V_GBL11066176	TIGAR	C12ORF5	Valid	Rabbit	Abcam	ab137573
Transglutaminase.M.V_GBL11066232	Transglutaminase	TGM2	Valid	Mouse	Lab Vision	MS-224-P1
TSC1.R.C_GBL11066181	TSC1	TSC1	Use with Caution	Rabbit	CST	4906
TTF1.R.V_GBL11066170	TTF1	NKX2-1	Valid	Rabbit	Abcam	ab76013
Tuberin.R.E_GBL11066115	Tuberin	TSC2	Valid	Rabbit	Epitomics	1613-1
Tuberin_pT1462.R.V_GBL11066116	Tuberin_pT1462	TSC2	Valid	Rabbit	CST	3617
VEGFR2.R.V_GBL11066117	VEGFR-2	KDR	Valid	Rabbit	CST	2479
VHL.M.C_GBL11066228	EPPK1 (This antibody actually recognizes & binds to EPPK1, validated by MS, siRNA & correlation to RNA expression)	EPPK1	Use with Caution	Mouse	BD Pharmingen	556347
X4E.BP1_pT70.R.V_GBL11066048	4EBP1_pT70	EIF4EBP1	Valid	Rabbit	CST	9455
XBP1.G.C_GBL11066242	XBP1	XBP1	Use with Caution	Goat	Santa Cruz	sc-32136
XRCC1.R.E_GBL11066148	XRCC1	XRCC1	Use with Caution	Rabbit	CST	2735
YAP.R.E_GBL11066129	YAP	YAP1	Under Evaluation	Rabbit	Santa Cruz	sc-15407
YAP_pS127.R.E_GBL11066130	YAP_pS127	YAP1	Under Evaluation	Rabbit	CST	4911
YB.1.R.V_GBL11066118	YB1	YBX1	Valid	Rabbit	Novus Biologicals	17250002
YB.1_pS102.R.V_GBL11066136	YB1_pS102	YBX1	Valid	Rabbit	CST	2900

Pathway Analysis Methods

In-silico Ingenuity-based pathway analysis

Pathway Analysis was performed on the mRNA expression data comparing the 60 Type2 PRCC tumors with the 75 Type 1 PRCC tumors. The genes that were on average up-regulated by 2-fold or greater in Type 2 PRCC compared Type 1 PRCC with a t-test p-value < 0.000001 were selected to represent a stringent, statistically significant gene list (n=353). This gene list was analyzed using the Ingenuity Systems Interactive pathway analysis of complex 'omics data software (IPA - http://www.ingenuity.com/) using the core analysis workflow. This provided data on the statistical enrichment of genes associated with disease or molecular and cellular function or a known canonical pathway.

Metabolic pathway analysis

The mRNA expression data was extracted for a specific list of genes associated with several metabolic pathways, including glycolysis, the Krebs cycle and the AMPK complex, that are listed in the following table. This data was used to perform unsupervised clustering of the 161 PRCC tumors. The CIMP PRCC tumors produced a specific cluster and further clustering was performed using selected genes that demonstrated strong differential expression (highlighted in gray within the following table) and only assessing the Type 2 PRCC tumors to demonstrate that the CIMP PRCC still represented a specific cluster.

Hierarchical cluster analysis was performed based on Euclidean distance using the Gene Cluster 3.0 software (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) and visualized using java TreeView 3.0 (http://jtreeview.sourceforge.net/).

Metabolic Function	Gene Name	Entrez No.	Full Gene name
Glycolysis	*HK1*	3098	hexokinase 1
Glycolysis	*HK2*	3099	hexokinase 2
Glycolysis	*HK3*	3101	hexokinase 3 (white cell)
Glycolysis	*GPI*	2821	glucose phosphate isomerase
Glycolysis	*PFKFB1*	5207	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
Glycolysis	*PFKFB2*	5208	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2
Glycolysis	*PFKFB3*	5209	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
Glycolysis	*PFKFB4*	5210	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4
Glycolysis	*PFKL*	5211	phosphofructokinase, liver
Glycolysis	*PFKM*	5213	phosphofructokinase, muscle
Glycolysis	*PFKP*	5214	phosphofructokinase, platelet
Glycolysis	*ALDOA*	226	aldolase A, fructose-bisphosphate
Glycolysis	*ALDOB*	229	aldolase B, fructose-bisphosphate
Glycolysis	*ALDOC*	230	aldolase C, fructose-bisphosphate
Glycolysis	*TPI1*	7167	triosephosphate isomerase 1
Glycolysis	*GAPDH*	2597	glyceraldehyde-3-phosphate dehydrogenase
Glycolysis	*PGK1*	5230	phosphoglycerate kinase 1
Glycolysis	*PGK2*	5232	phosphoglycerate kinase 2
Glycolysis	*PGAM1*	5223	phosphoglycerate mutase 1 (brain)
Glycolysis	*PGAM2*	5224	phosphoglycerate mutase 2 (muscle)
Glycolysis	*PGAM4*	441531	phosphoglycerate mutase family member 4
Glycolysis	*PGAM5*	192111	phosphoglycerate mutase family member 5
Glycolysis	*ENO1*	2023	enolase 1, (alpha)
Glycolysis	*ENO2*	2026	enolase 2 (gamma, neuronal)
Glycolysis	*ENO3*	2027	enolase 3 (beta, muscle)
Glycolysis	*ENOPH1*	58478	enolase-phosphatase 1
Glycolysis	*ENOSF1*	55556	enolase superfamily member 1
Glycolysis	*PKLR*	5313	pyruvate kinase, liver and RBC
Glycolysis	*PKM2*	5315	pyruvate kinase, muscle
Glycolysis	*LDHA*	3939	lactate dehydrogenase A
Glycolysis	*LDHB*	3945	lactate dehydrogenase B
Glycolysis	*LDHC*	3948	lactate dehydrogenase C
Glycolysis	*LDHD*	197257	lactate dehydrogenase D

Pyruvate Metabolism	*PDK1*	5163	pyruvate dehydrogenase kinase, isozyme 1
Pyruvate Metabolism	*PDK2*	5164	pyruvate dehydrogenase kinase, isozyme 2
Pyruvate Metabolism	*PDK3*	5165	pyruvate dehydrogenase kinase, isozyme 3
Pyruvate Metabolism	*PDK4*	5166	pyruvate dehydrogenase kinase, isozyme 4
Pyruvate Metabolism	*PDP1*	54704	pyruvate dehyrogenase phosphatase catalytic subunit 1
Pyruvate Metabolism	*PDP2*	57546	pyruvate dehyrogenase phosphatase catalytic subunit 2
Pyruvate Metabolism	*PDPR*	55066	pyruvate dehydrogenase phosphatase regulatory subunit
Krebs Cycle	*CS*	1431	citrate synthase
Krebs Cycle	*ACO1*	48	aconitase 1, soluble
Krebs Cycle	*ACO2*	50	aconitase 2, mitochondrial
Krebs Cycle	*IDH1*	3417	isocitrate dehydrogenase 1 (NADP+), soluble
Krebs Cycle	*IDH2*	3418	isocitrate dehydrogenase 2 (NADP+), mitochondrial
Krebs Cycle	*OGDH*	4967	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)
Krebs Cycle	*SUCLA2*	8803	succinate-CoA ligase, ADP-forming, beta subunit
Krebs Cycle	*SUCLG1*	8802	succinate-CoA ligase, alpha subunit
Krebs Cycle	*SUCLG2*	8801	succinate-CoA ligase, GDP-forming, beta subunit
Krebs Cycle	*SDHAF1*	644096	succinate dehydrogenase complex assembly factor 1
Krebs Cycle	*SDHAF2*	54949	succinate dehydrogenase complex assembly factor 2
Krebs Cycle	*SDHA*	6389	succinate dehydrogenase complex, subunit A, flavoprotein (Fp)
Krebs Cycle	*SDHB*	6390	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
Krebs Cycle	*SDHC*	6391	succinate dehydrogenase complex, subunit C, integral membrane protein
Krebs Cycle	*SDHD*	6392	succinate dehydrogenase complex, subunit D, integral membrane protein
Krebs Cycle	*FH*	2271	fumarate hydratase
Krebs Cycle	*MDH1*	4190	malate dehydrogenase 1, NAD (soluble)
Krebs Cycle	*MDH2*	4191	malate dehydrogenase 2, NAD (mitochondrial)
Krebs Cycle	*ME1*	4199	malic enzyme 1, NADP(+)-dependent, cytosolic
Krebs Cycle	*ME2*	4200	malic enzyme 2, NAD(+)-dependent, mitochondrial
Krebs Cycle	*ME3*	10873	malic enzyme 3, NADP(+)-dependent, mitochondrial
Krebs Cycle	*PC*	5091	pyruvate carboxylase
Krebs Cycle	*GOT1*	2805	glutamic-oxaloacetic transaminase 1, soluble
Krebs Cycle	*GOT2*	2806	glutamic-oxaloacetic transaminase 2, mitochondrial
Krebs Cycle	*NNT*	23530	nicotinamide nucleotide transhydrogenase
Ribose Sugar Metabolism	*TKTL1*	8277	transketolase-like 1
Ribose Sugar Metabolism	*TKTL2*	84076	transketolase-like 2

Ribose Sugar Metabolism	*TKT*	7086	transketolase
Ribose Sugar Metabolism	*TALDO1*	6888	transaldolase 1
Ribose Sugar Metabolism	*G6PD*	2539	glucose-6-phosphate dehydrogenase
Ribose Sugar Metabolism	*PGD*	5226	phosphogluconate dehydrogenase
Ribose Sugar Metabolism	*PGLS*	25796	6-phosphogluconolactonase
Ribose Sugar Metabolism	*RPIA*	22934	ribose 5-phosphate isomerase A
Ribose Sugar Metabolism	*RPE*	6120	ribulose-5-phosphate-3-epimerase
AMPK Complex	*PRKAA1*	5562	protein kinase, AMP-activated, alpha 1 catalytic subunit
AMPK Complex	*PRKAA2*	5563	protein kinase, AMP-activated, alpha 2 catalytic subunit
AMPK Complex	*PRKAB1*	5564	protein kinase, AMP-activated, beta 1 non-catalytic subunit
AMPK Complex	*PRKAB2*	5565	protein kinase, AMP-activated, beta 2 non-catalytic subunit
AMPK Complex	*PRKAG1*	5571	protein kinase, AMP-activated, gamma 1 non-catalytic subunit
AMPK Complex	*PRKAG2*	51422	protein kinase, AMP-activated, gamma 2 non-catalytic subunit
AMPK Complex	*PRKAG3*	53632	protein kinase, AMP-activated, gamma 3 non-catalytic subunit
Fatty Acid Synthesis	*ACACA*	31	acetyl-CoA carboxylase alpha
Fatty Acid Synthesis	*ACACB*	32	acetyl-CoA carboxylase beta
Fatty Acid Synthesis	*ACLY*	47	ATP citrate lyase
Fatty Acid Synthesis	*FASN*	2194	fatty acid synthase
Fatty Acid Synthesis	*SLC1A5*	6510	solute carrier family 1 (neutral amino acid transporter), member 5

Batch Effect Analysis Methods

Assessment of batch effects

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the papillary renal cell carcinoma data sets. Five different data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), copy number variations (GW SNP 6), and protein expression (RPPA). All of the data sets were at TCGA level 3, since that's the level at which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for all five data sets are shown in supplementary figure S21.

Supplementary Figures S1-S21

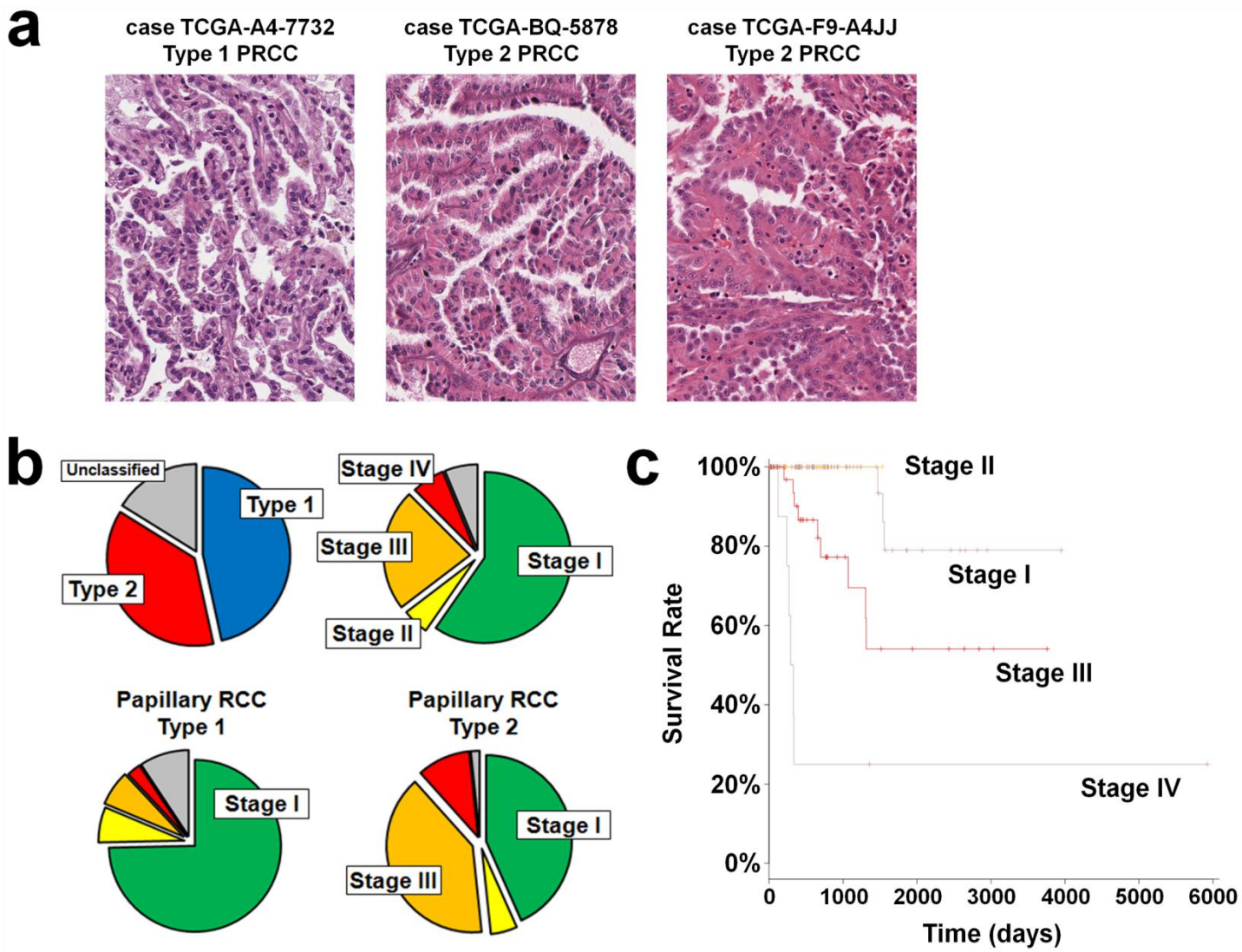


Figure S1: Type and Tumor Stage Analysis of the 161 Papillary Renal Cell Carcinoma (PRCC) Tumors in the TCGA KIRP Cohort

(a) Representative histological sections of Type 1 and Type 2 PRCC cases, with Type 2 showing more heterogeneity at the level of histopathology.

(b) The breakdown of PRCC type and tumor stage for all the tumors within the cohort and the tumor stage breakdown for PRCC Type 1 and Type 2 tumors alone.

(c) Kaplan-Meier survival analysis for the 161 PRCC tumors dependent upon tumor stage.

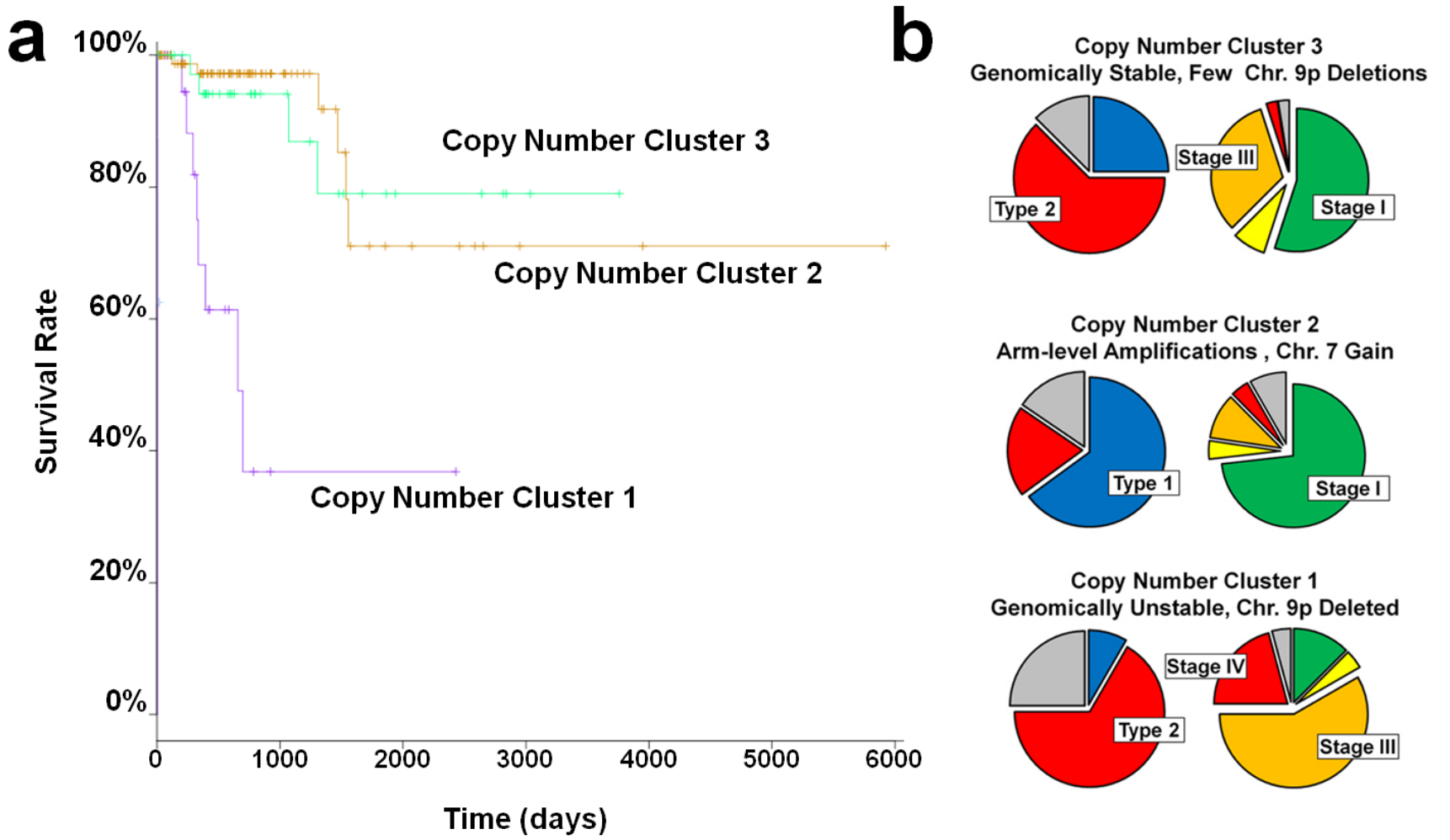


Figure S2: Type, Tumor Stage and Survival Analysis of the 161 PRCC Tumors dependent upon Copy Number Cluster Analysis

(a) Kaplan-Meier survival analysis for the 161 PRCC tumors dependent upon the copy number clusters.

(b) The breakdown of PRCC type and tumor stage for each of the copy number clusters generated from the SNP-array based profiling of somatic copy number alterations (SCNAs).

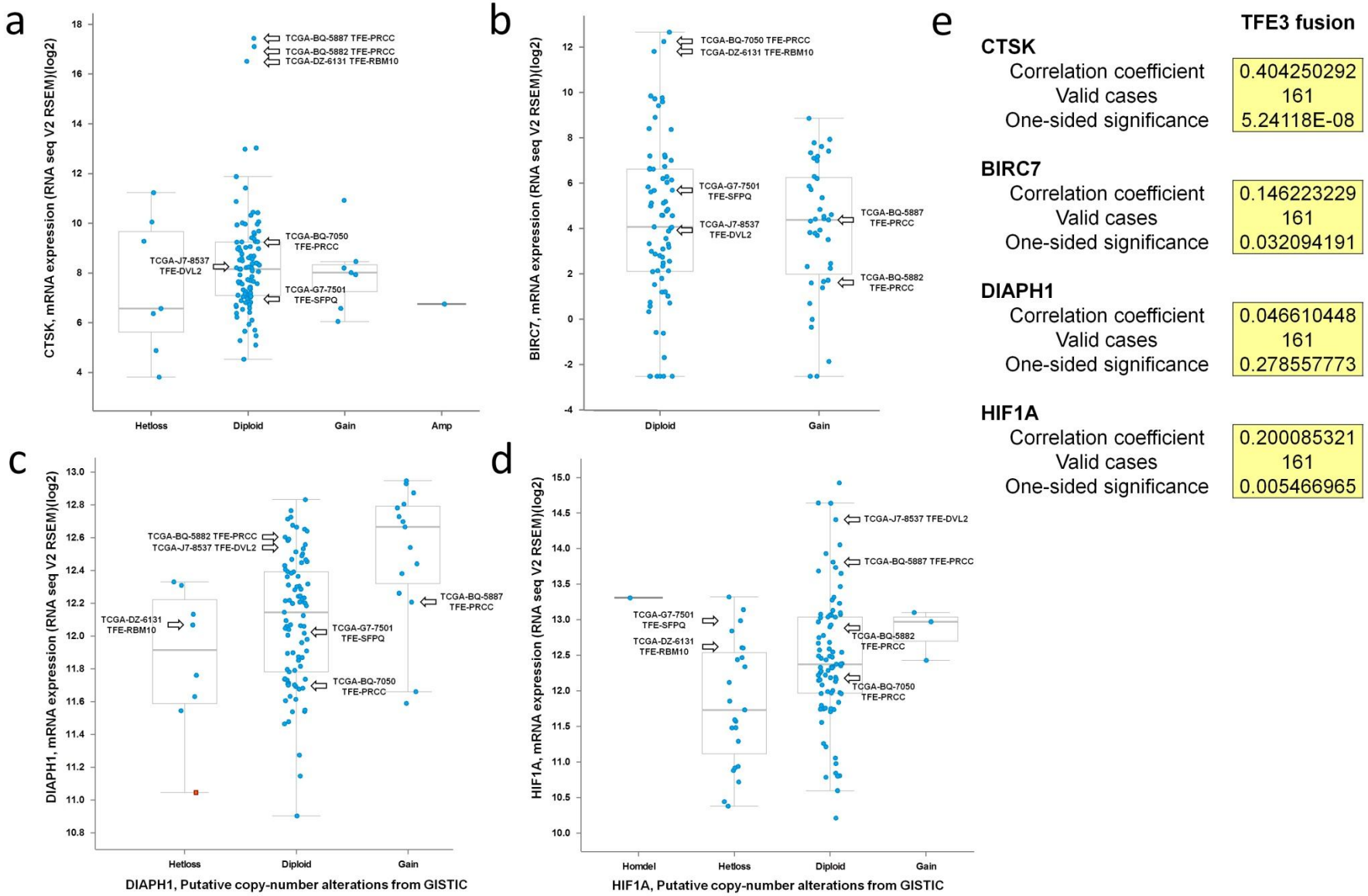


Figure S3: Expression Analysis of *TFE3* Target Genes

Expression analysis of known target genes for the *TFE3* transcription factor was performed using the RNA-Seq data and combined with putative copy-number alteration provided by GISTIC. The target gene assessed were **(a)** *CTSK*, **(b)** *BRIC7*, **(c)** *DIAPH1*, and **(d)** *HIF1A*. The arrows designate the expression markers for the individual *TFE3*-fusion tumors. The correlation between the presence of the TFE3 fusions and increased mRNA expression were calculated for each gene **(e).**

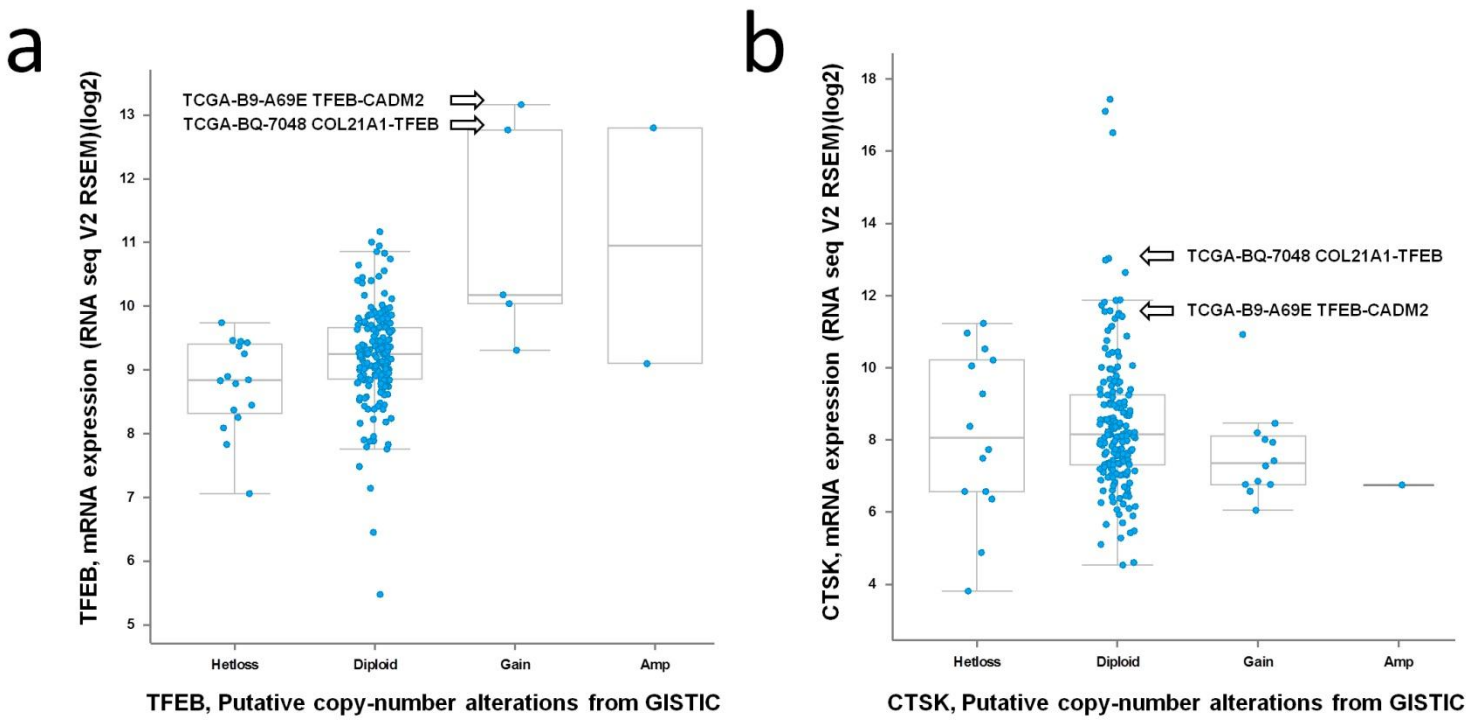


Figure S4: Expression Analysis of *TFEB* and a Known Target Gene

Expression analysis for **(a)** the transcription factor *TFEB* and its known target gene, **(b)** *CTSK*, was performed using the RNA-Seq data and combined with putative copy-number alteration provided by GISTIC. The arrows designate the expression markers for the individual *TFEB*-fusion tumors.

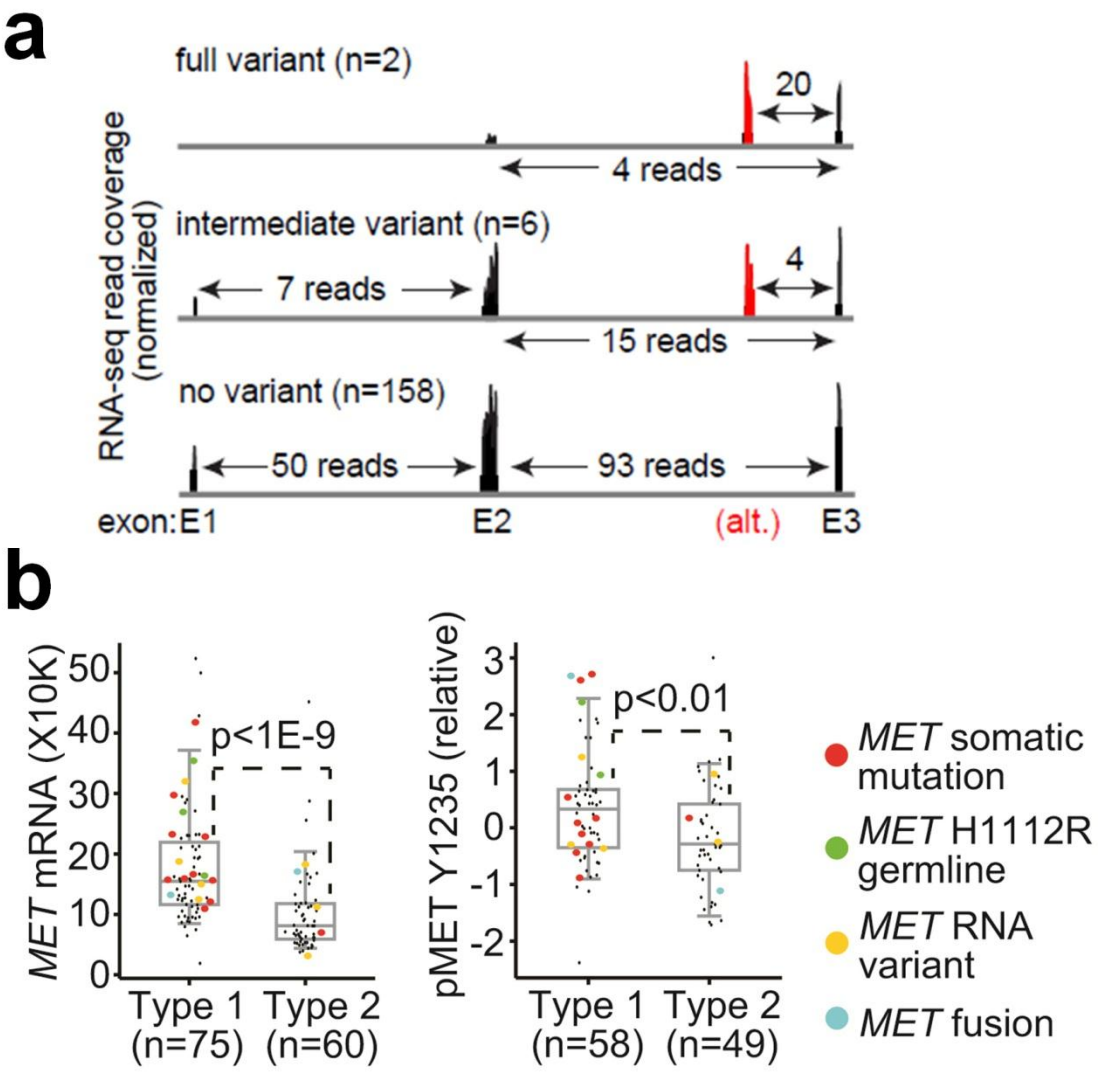


Figure S5: Novel RNA transcript variant of *MET* and Expression of the *MET* gene

(a) Novel RNA transcript variant of *MET* lacking the canonical exons 1 and 2 but containing a novel exon 1 that splices to the canonical exon 3. This variant is represented in the majority of transcripts in two PRCC cases and in a fraction of the transcripts in six other PRCC cases. Numbers of sequence reads spanning the given exon junctions are indicated.

(b) Differential expression of *MET* mRNA RPKM levels (left) and phospho-Met (pY1235) levels (right), between Type 1 and Type 2 PRCC. Cases harboring mutation or RNA transcript variant for MET are indicated. P-values by t-test on log-transformed data.

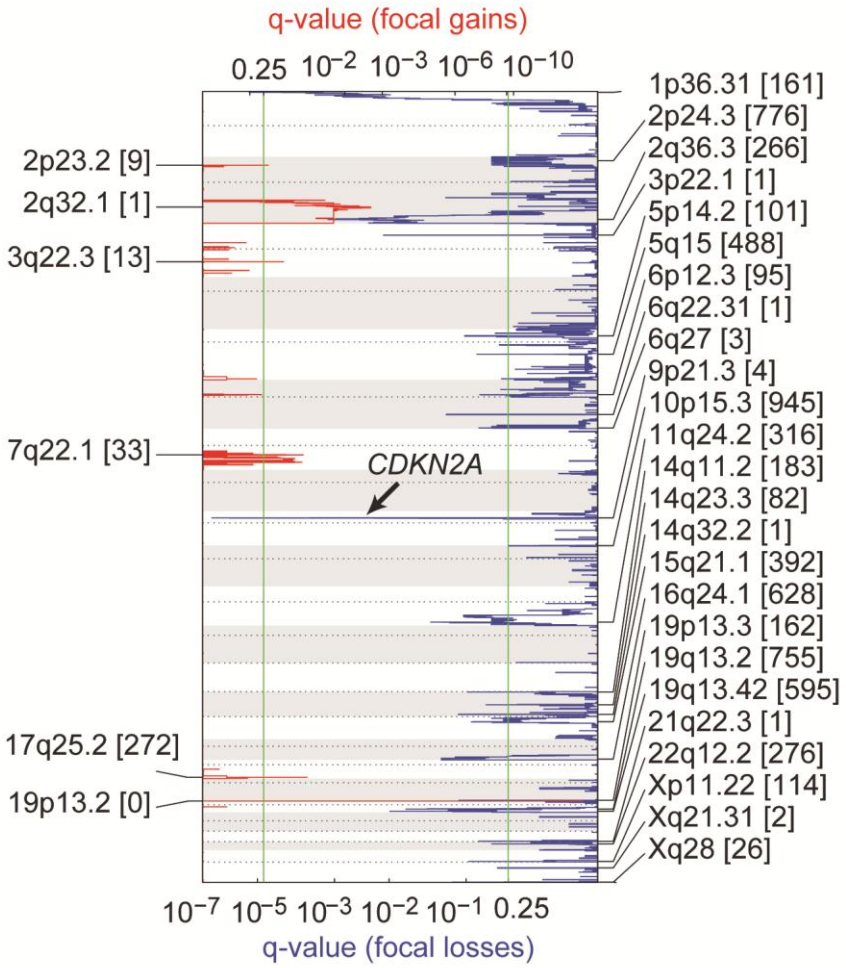


Figure S6: GISTIC2.0 Analysis of Focal Copy-Number Alteration

GISTIC2.0 analysis of focal copy-number alteration was performed on the 161 PRCCs and demonstrated both deletions and amplifications. The regions of deletion or amplification are designated with their chromosomal cytogenetic band and the number of genes within each region is shown within square-brackets. Two significant deletions were observed at 1p36.31 and 9p21.3 that contained 136 genes and 4 genes respectively.

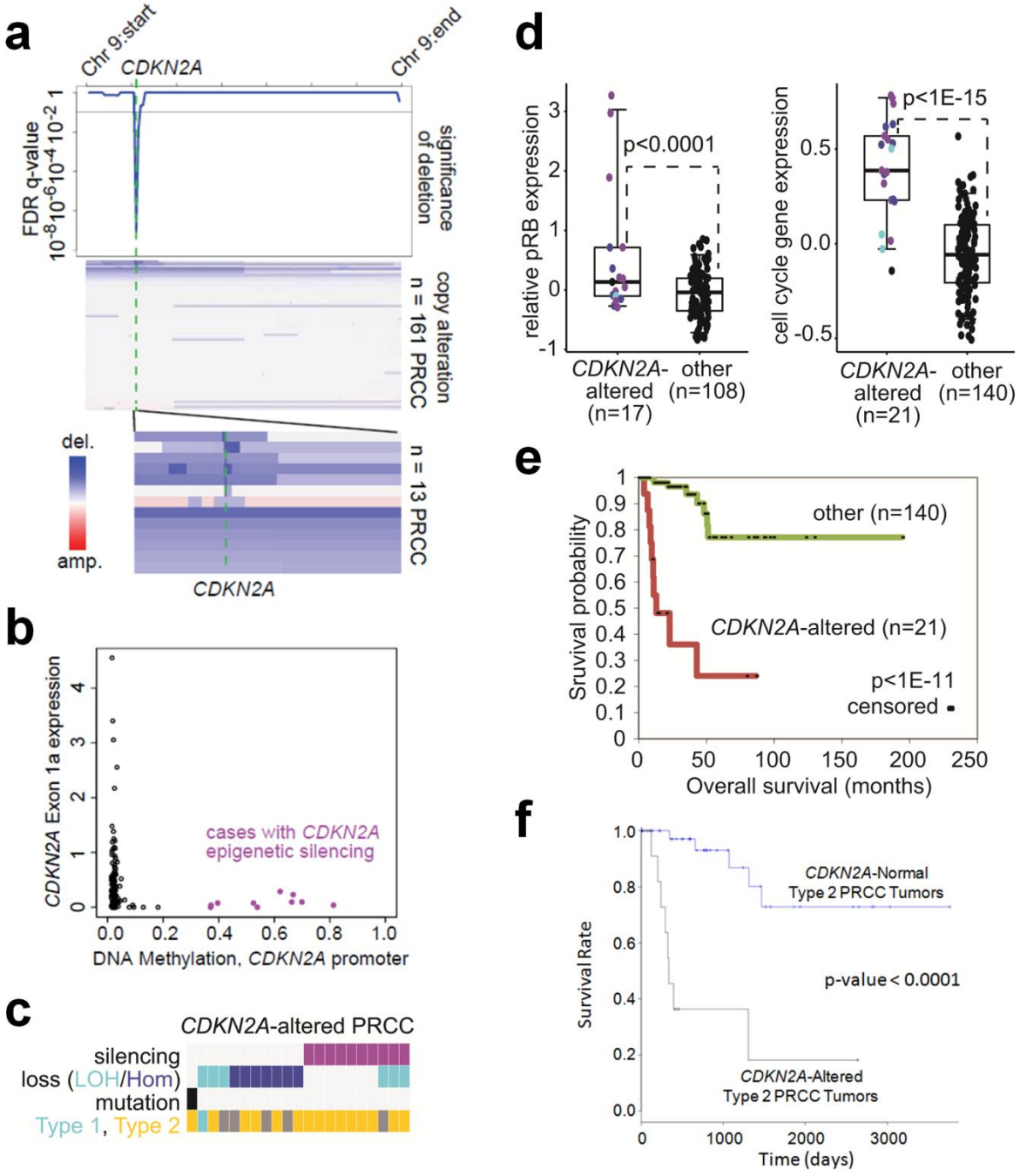
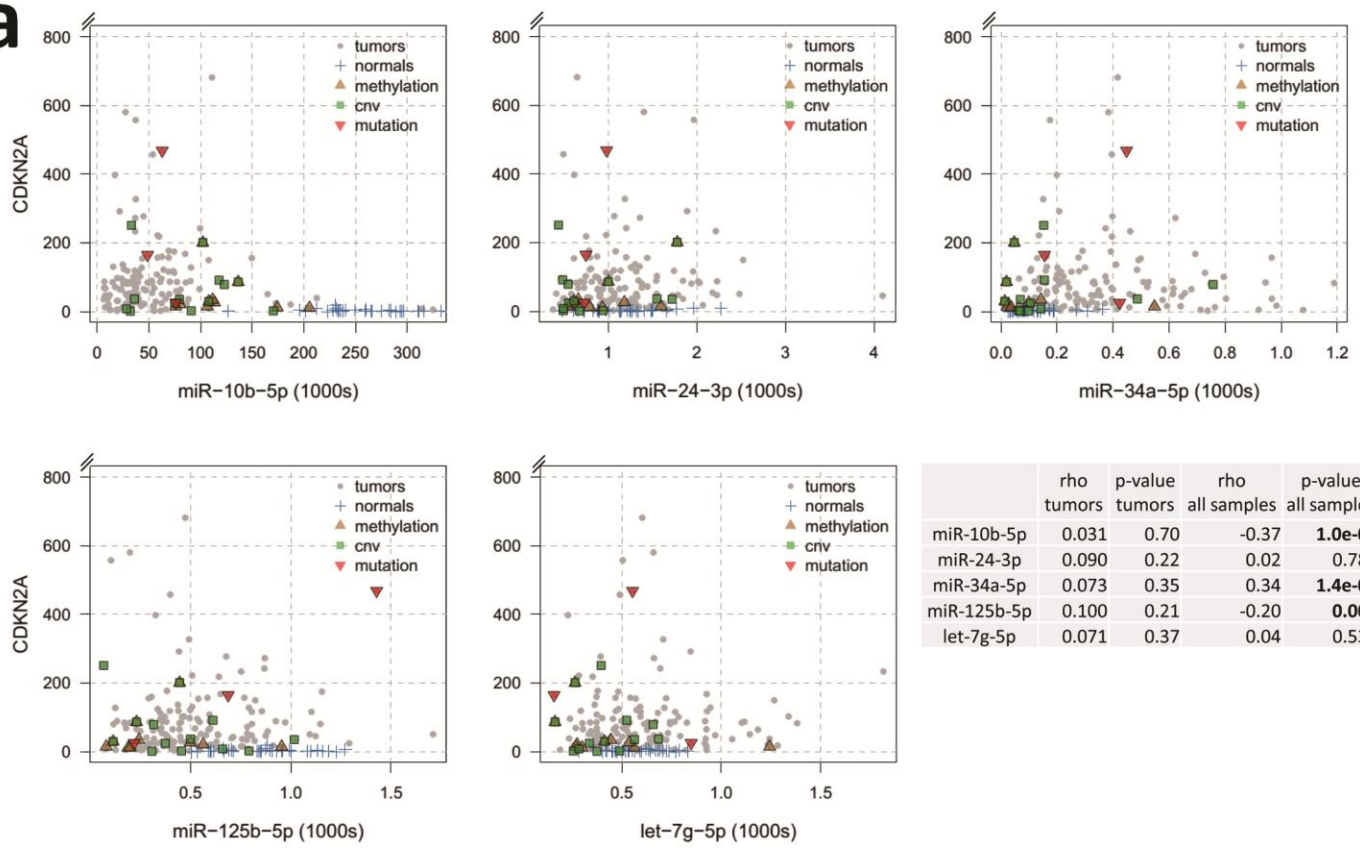


Figure S7: *CDKN2A* Alterations in PRCC

(a) GISTIC analysis identified a region of chromosome 9 as representing a clear target of focal copy deletion (q-value<1E-6) in PRCC resulting in loss of genomic regions specifically surrounding CDKN2A. **(b)** Epigenetic silencing of CDKN2A locus in ten PRCC cases. Exon 1a expression corresponds to p16INK4a isoform. **(c)** PRCC cases with *CDKN2A* genetic (black = mutation, light blue = partial focal deletion, dark blue = near complete focal deletion) or epigenetic alterations (purple) (n=21 cases). **(d)** Comparing PRCC cases with and without *CDKN2A* alterations for differences in phospho-Rb levels (left, p<0.0002, t-test) and average expression of cell cycle-related genes[43] (right, p<1E-10). **(e)** Differences in patient overall survival, comparing PRCC cases with and without *CDKN2A* alterations (p<1E-10, log-rank test). **(f)** Kaplan-Meier survival analysis was performed comparing the PRCC tumors with *CDKN2A* alterations verses the Type 2 PRCC tumors without *CDKN2A* alterations. The *CDKN2A* altered tumors still demonstrated a significantly worse predicted survival. This demonstrated that the poorer survival rate observed when all PRCC tumors with *CDKN2A* alterations was not simply due to the increased percentage of Type 2 PRCC samples in the *CDKN2A* altered cohort compared to the *CDKN2A* normal cohort.

a

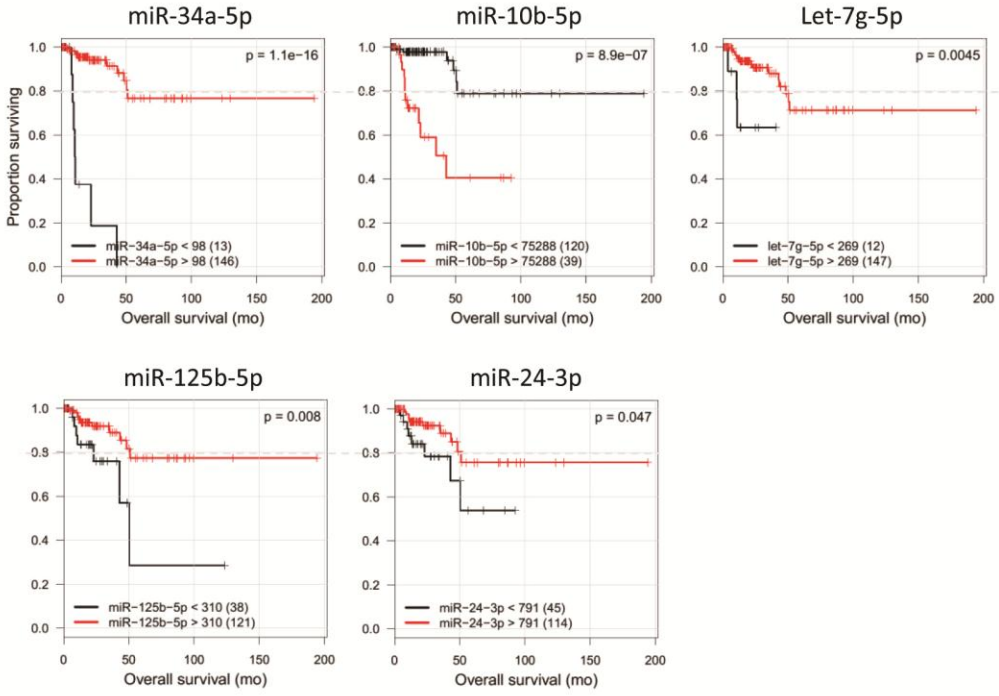


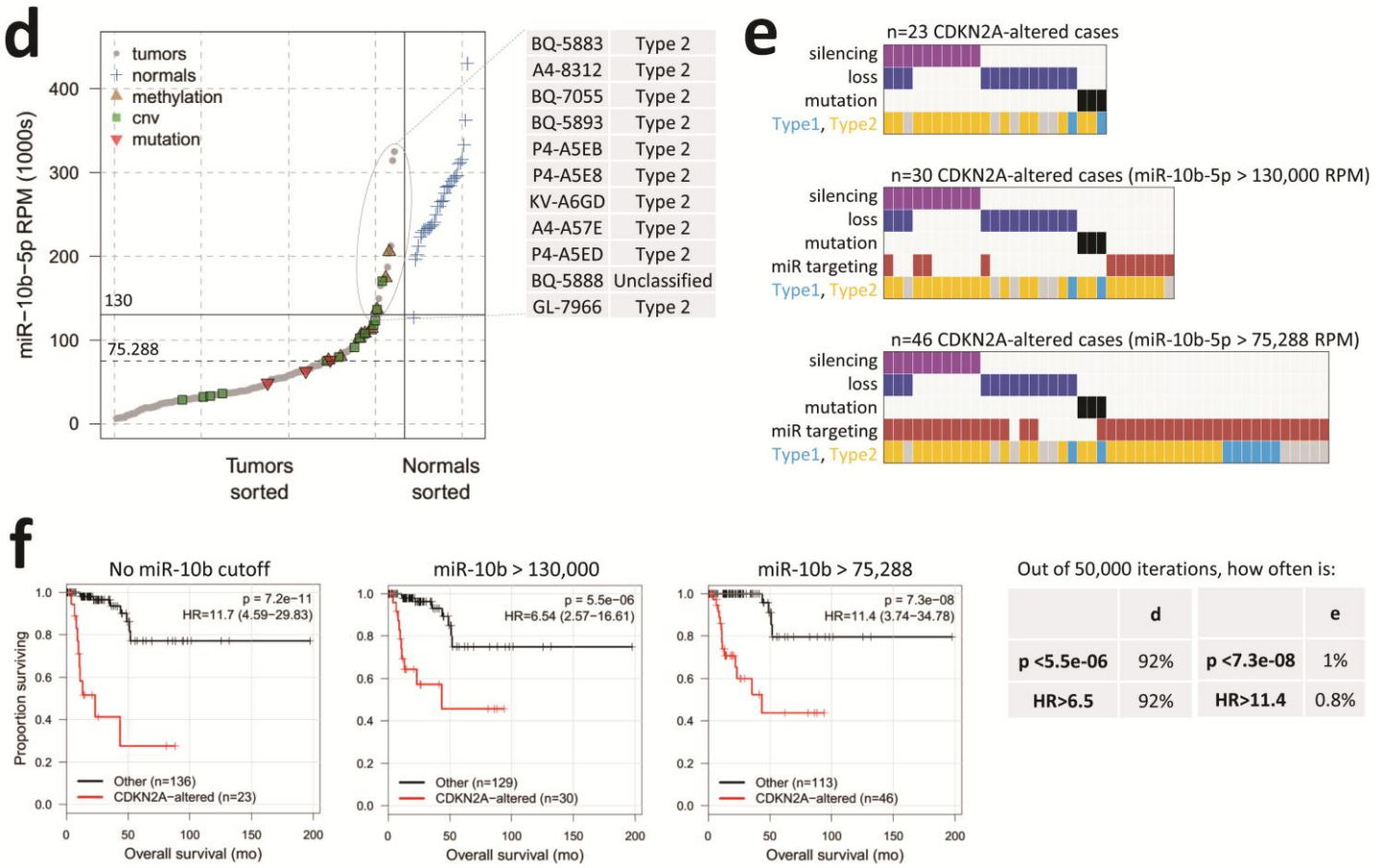
	rho tumors	p-value tumors	rho all samples	p-value all samples
miR-10b-5p	0.031	0.70	-0.37	**1.0e-07**
miR-24-3p	0.090	0.22	0.02	0.780
miR-34a-5p	0.073	0.35	0.34	**1.4e-05**
miR-125b-5p	0.100	0.21	-0.20	**0.005**
let-7g-5p	0.071	0.37	0.04	0.530

b

	Optimal RPM cutoff	log rank p-value	Hazard ratio	Lower 95% CI	Upper 95% CI	num below cutoff	num above cutoff	High expression of this miR is:
miR-34a-5p	98.43	1.11E-16	0.04	0.01	0.13	13	146	Protective
miR-10b-5p	75287.59	8.88e-07	8.03	3.00	21.52	120	39	Anti-protective
let-7g-5p	268.64	0.0045	0.19	0.05	0.68	12	147	Protective
miR-125b-5p	309.97	0.0080	0.29	0.11	0.77	38	121	Protective
miR-24-3p	791.33	0.0470	0.40	0.16	1.02	45	114	Protective

c





Out of 50,000 iterations, how often is:

	d		e
p <5.5e-06	92%	**p <7.3e-08**	1%
HR>6.5	92%	**HR>11.4**	0.8%

Figure S8: Assessment of *CDKN2A* Targeting microRNAs (miRs)

(a) *CDKN2A* vs. miR abundance in tumor (gray dots) and normal (blue plus sign) samples for the five miRs that are functionally validated as targeting *CDKN2A*: miR-10b-5p, 24-3p, 34a-5p, 125b-5p and let-7g-5p. Colored markers indicate the 23 tumor samples with a *CDKN2A* alteration. The table lists the Spearman coefficients (rho) and p-values for both the tumor and tumor-and-normal datasets. Of these miRs, miR-10b-5p is the most likely to influence *CDKN2A* transcript levels, as it is both highly abundant and significantly anti-correlated with *CDKN2A*.
(b) Kaplan-Meier results for the miR expression value that stratifies patients into the two groups that minimize the Kaplan-Meier log-rank p-value. The optimal threshold for miR-10b-5p is 75,288 RPM, which results in a log-rank p-value of 8.9e-07 and a hazard ratio of 8.0 (3.0 – 21.5). 39 patients have a miR-10b-5p expression level above this threshold and have poorer overall survival.
(c) Kaplan-Meier plots for each miR, with patients stratified by expression above or below the optimal cutoff. Red indicates the group of samples that have abundance above the cutoff.
(d) Tumor and adjacent normal samples sorted by miR-10b-5p abundance. Two horizontal lines indicate the optimal cutoff of 75,288 RPM, and a more conservative cutoff of 130,000 RPM, which we chose because the samples with higher miR-10b-5p include 10 Type 2, one unclassified, and no Type 1.
(e) Oncoprints of the 23 samples with *CDKN2A* genomic alterations, plus the samples in which miR-10b-5p RPM was above 130,000 or 75,288 RPM. The 39 patients with miR-10b-5p > 75,288 RPM are dominated by type 2 pathology, with 7 Type 1, 25 Type 2, and 7 unclassified. 70% of the *CDKN2A*-altered cases also have high miR-10b-5p abundance.
(f) Kaplan-Meier plots demonstrated the original 23 *CDKN2A*-altered samples (log-rank p=7.2e-11) compared with either the original *CDKN2A*-altered samples plus samples with miR-10b-5p expression above 130,000 RPM (n=30, log-rank p=5.5e-06) or the original *CDKN2A*-altered samples plus samples with miR-10b-5p expression above 75,288 RPM (n=46, log-rank p=7.0e-08) . Both options remained statistically significant, but to test if this was simply due to the original 23 samples bootstrap assessment of Kaplan-Meier was performed by adding either 7 or 23 random additional samples chosen at 50,000 times. While the addition of 7 random samples would be equally significant 92% of the time, addition of 23 random samples would only be equally significant 1% of the time. These results are consistent with miR-10b-5p directly targeting *CDKN2A* potentially being functionally important in Kidney Papillary Cancer.

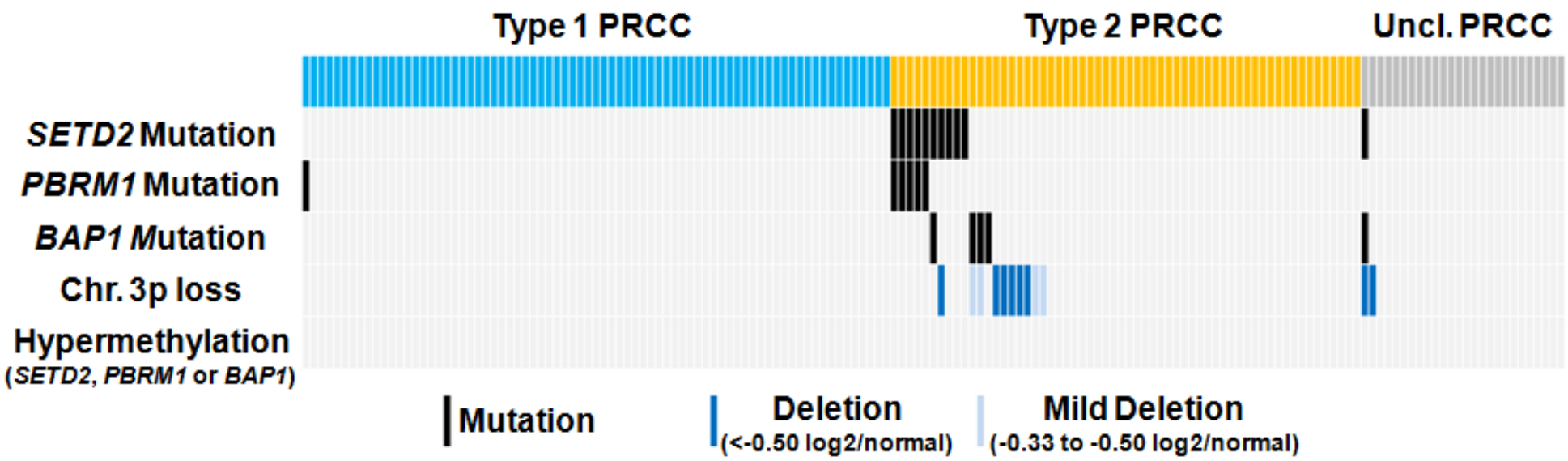


Figure S9: Oncoprint of the Chromosome 3p-Encoded Chromatin Remodeling Tumor Suppressor Genes.

The mutations identified in the chromosome 3p-encoded chromatin remodeling genes, *SETD2*, *PBRM1* and *BAP1*, within the PRCC samples that were analyzed by the whole exome sequencing were mapped on an oncoprint and ordered by PRCC tumor type. This demonstrates that the majority of mutation occur within the Type 2 PRCC samples and that *SETD2* and *PBRM1* mutations appear to co-occur while *PBRM1* and *BAP1* mutations are mutually exclusive. Heterozygous or mild deletion of chromosome 3p was highlighted in dark blue and light blue respectively (no homozygous loss was observed) and hypermethylation of the CpG islands associated with each of the three genes was assessed, but provided no positive hits. Loss of chromosome 3p seems to associate with Type 2 PRCC but there was only a partial overlap with mutations of SETD2, PBRM1 or BAP1 suggesting that haploinsufficiency of these genes may be important to tumorigenesis.

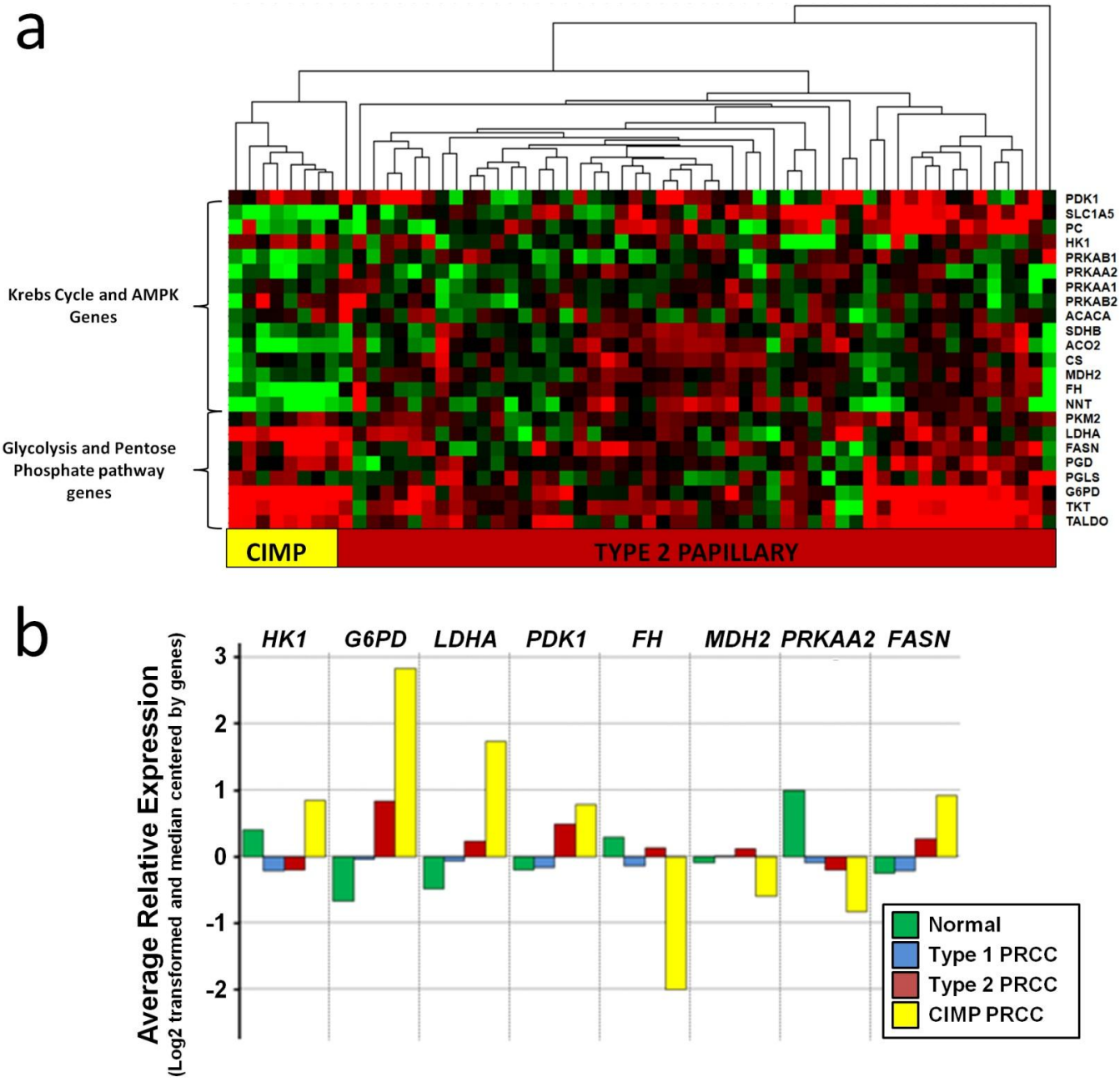


Figure S10: Metabolic Analysis of the CIMP Associated PRCC tumors

Analysis of the mRNA expression data by unsupervised clustering of all the genes involved in several major metabolic pathways, consisting of glycolysis, the pentose phosphate pathway, the Krebs cycle and fatty acid synthesis, demonstrated the CIMP associated PRCC clustered separately from the rest of the PRCC tumors. Due to the majority of CIMP associated tumors being Type 2 PRCCs, this separation was confirmed by assessing all Type 2 PRCCs using a selection of metabolic genes.

(a) This heatmap demonstrates the unsupervised clustering of the all the Type 2 PRCC tumors (n=60), including the CIMP associated Type 2 PRCC tumors (n=8) for a selection of glycolysis genes (*HK1*, *PKM2*, *LDHA*), pentose phosphate pathway genes (*PGD*, *PGLS*, *G6PD*, *TKT*, *TALDO*), Krebs cycle genes (*PDK1*, *CS*, *ACO2*, *SDHB*, *FH*, *MDH2*, *PC*, *NNT*), fatty acid synthesis genes (*FASN*, *ACACA*), AMPK genes (*PRKAA1*, *PRKAA2*, *PRKAB1*, *PRKAB2*) and the glutamine transporter gene, *SLC1A5*. The CIMP associated Type 2 PRCC tumors produced a distinct cluster.

(b) The average relative mRNA expression for 8 key genes within the glycolysis, Krebs cycle and fatty acid synthesis pathways (*HK1*, *G6PD*, *LDHA*, *PDK1*, *FH*, *MDH2*, *PRKAA2* and *FASN*) were calculated for either the associated normal kidney samples (green), the Type 1 PRCCs (blue), the Type 2 PRCCs (red) or the CIMP associated PRCCs (yellow).

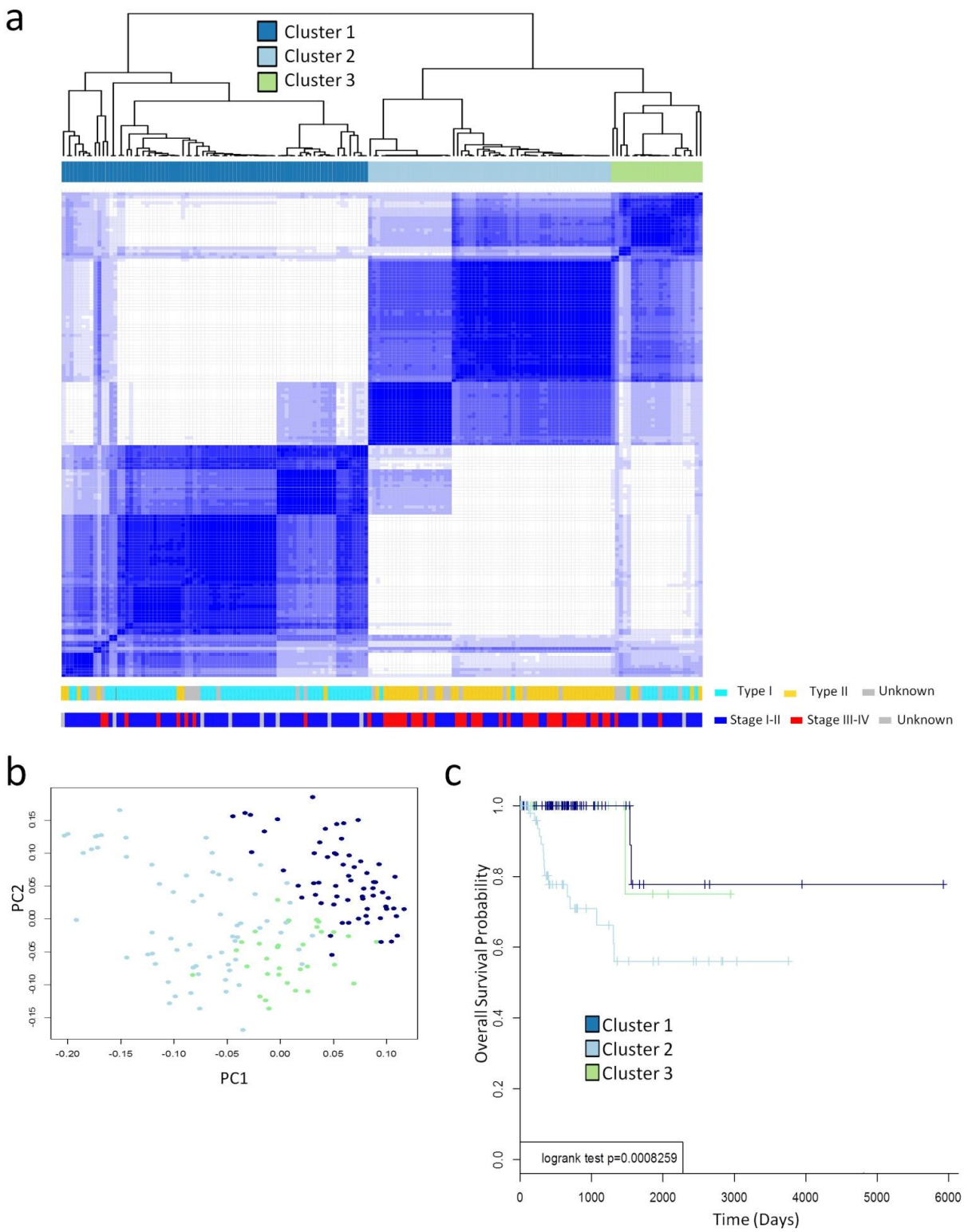


Figure S11: mRNA Expression Analysis in 161 Papillary Renal Cell Carcinomas

(a) Results of unsupervised consensus clustering of RNA-Seq data for 161 pRCC tumor samples. 2050 genes were used to identify three distinct clusters within tumor samples, which are illustrated by consensus matrices. No divergent groups were identified beyond k mean 3. The PRCC type and tumor stage were indicated.

(b) Principal component analysis using the same genes identified in **(a)** illustrated the three distinct clusters.

(c) Kaplan-Meier curves were used to calculate overall survival for each mRNA cluster, with a long-rank p-value.

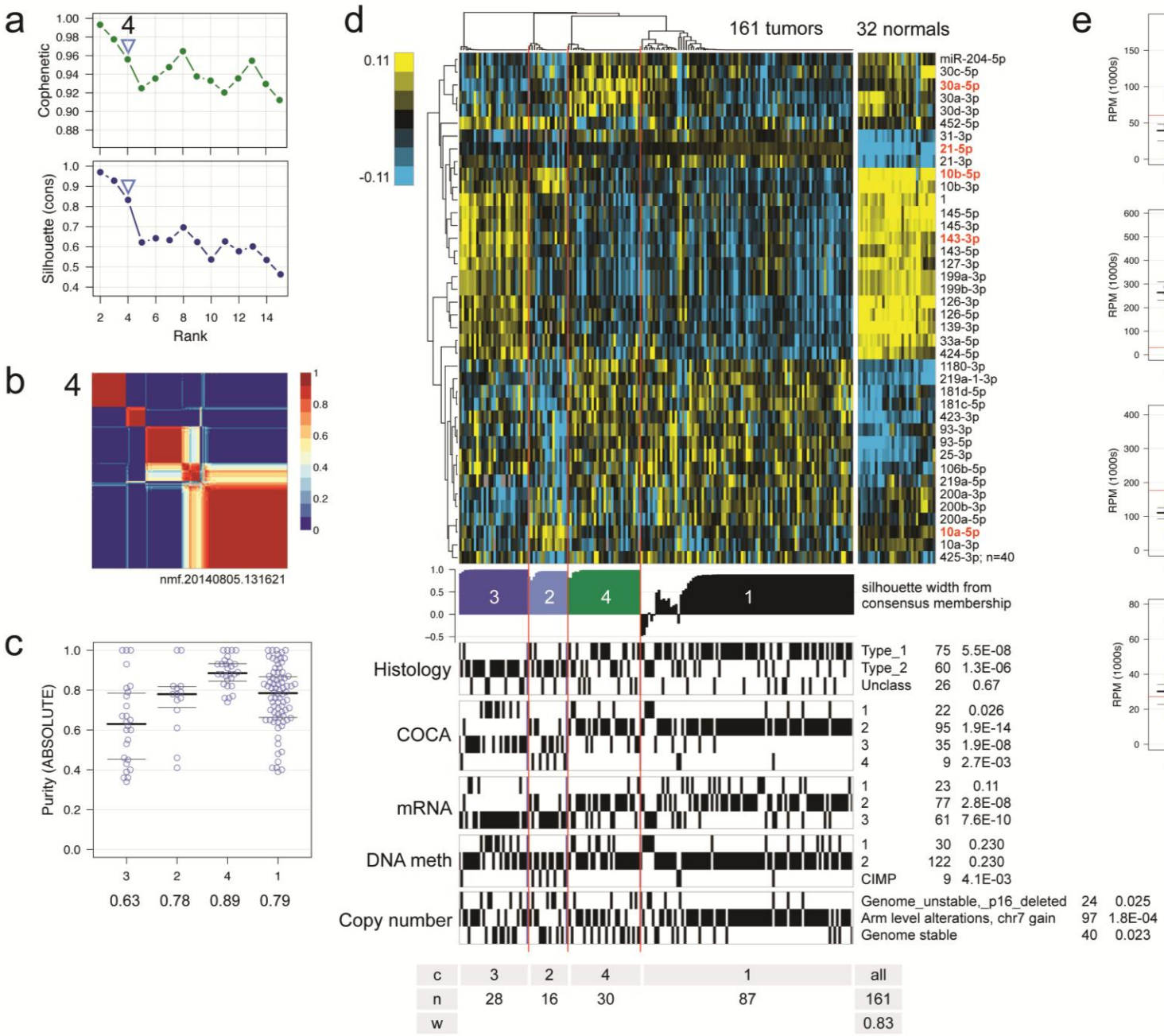

a
Cophenetic
Silhouette (cons)
Rank
b
nmf.20140805.131621
c
Purity (ABSOLUTE)
0.63 0.78 0.89 0.79
d
161 tumors
32 normals
silhouette width from consensus membership
Histology
COCA
mRNA
DNA meth
Copy number
Type_1 75 5.5E-08
Type_2 60 1.3E-06
Unclass 26 0.67
CIMP 9 4.1E-03
Genome_unstable,_p16_deleted 24 0.025
Arm level alterations, chr7 gain 97 1.8E-04
Genome stable 40 0.023
c 3 2 4 1 all
n 28 16 30 87 161
w 0.83
e
miR-10a-5p
miR-10b-5p
miR-21-5p
miR-22-3p
miR-30a-5p
miR-30a-3p
miR-99b-5p
miR-143-3p
RPM (1000s)
f
Fraction surviving
161 T
P=0.13
Overall survival duration (mo)

Figure S12: Unsupervised NMF Consensus Clustering of miR-Seq Data for 161 PRCC Tumor Samples

(a) Rank survey profiles for cophenetic correlation coefficient and average silhouette width.

(b) Consensus membership heatmap for a four-cluster solution, with yellow-white indicating samples that are less 'typical' cluster or group members.

(c) Per-cluster distributions of ABSOLUTE (Carter 2012) tumor sample purity. The median purity per cluster was at least 0.78 for clusters 1, 2 and 4, and was 0.63 for cluster 3.

(d) For the four-group solution, top to bottom: a normalized abundance heatmap for the 40 5p or 3p mature strands that were highly ranked as differentially abundant; silhouette width profile calculated from the consensus membership; covariates with Fisher exact association *P*-values; and a summary table of cluster number, number of samples in each cluster, and the overall average silhouette width. The scale bar shows row-scaled $\log_{10}$(RPM+1) miR abundance. In the covariate tracks under the heatmap, Types 1 and 2 were strongly associated with the miR clusters (P=5.5E-8 and 1.3E-6 respectively), while unclassified samples were statistically independent of the miR clusters (P=0.67). Clusters 2 (n=16) and 3 (n=28) were largely Type 2 samples, while cluster 1 (n=87) was enriched in type 1 samples. Cluster 4 (n=30) contained both type 1 and 2 samples. Of miR cluster 1's 87 samples, 8 samples (21%) were considered atypical cluster members, because they had silhouette widths below 0.6, which is approximately 0.7 of the cluster's maximum silhouette width, 0.884. The miR clusters were associated with COCA clusters 2 to 4 (P<2.7E-3), mRNA clusters 2 and 3 (P<2.8E-8), the CIMP DNA methylation cluster (P=4.1E-3), and the arm level/chromosome 7 gain copy number cluster (P=1.8E-4) (Fig. S1d). We noted that both of the type 2-enriched miR clusters (2 and 3) were enriched in COCA cluster 3 and mRNA cluster 3. In contrast, miR cluster 2, but not miR cluster 3, was enriched in COCA cluster 4 and DNA methylation CIMP samples. Both the type 1-enriched miR cluster 1 and the mixed type miR cluster 4 were enriched in COCA cluster 2 and mRNA cluster 2. miR cluster 1 was enriched in arm level/chromosome 7 gain copy number alterations.

(e) Per-cluster distributions of normalized (RPM) abundance for a subset of miRs that were highly scored by NMF as discriminatory. Black horizontal bars indicate median RPMs. The Type 1-enriched miR cluster 1 was discriminated by very high miR-21-5p levels, and had less abundant miR-10b, 99b-5p and 143-3p than adjacent normal samples. miR cluster 2 was discriminated from other tumor clusters by high miR-10a-5p and only moderately less abundant miR-10b-5p, and had high miR-21-5p and moderately lower miR-143-3p than adjacent normal samples. miR cluster 3 was discriminated from other tumor clusters by high miR-22-3p, low miR-30a-5p, and near-normal levels of miR-143-3p, while having a low miR-10b-5p level comparable to type 1 clusters 1 and 4, moderately high miR-21-5p levels that were comparable to type 2 cluster 2 and type 1 cluster 4, and miR-143-3p levels comparable to adjacent normal. miR cluster 4, which contains both type 1 and 2 samples, was discriminated by high miR-30a-5p and -3p, and had higher miR-21-5p and lower miR-143-3p than adjacent normal samples.

(f) Kaplan-Meier plot for overall survival for the four unsupervised clusters, with a log-rank P-value. The two type 2 miR clusters trended to poorer outcomes than the type 1 clusters (log-rank P=0.13).

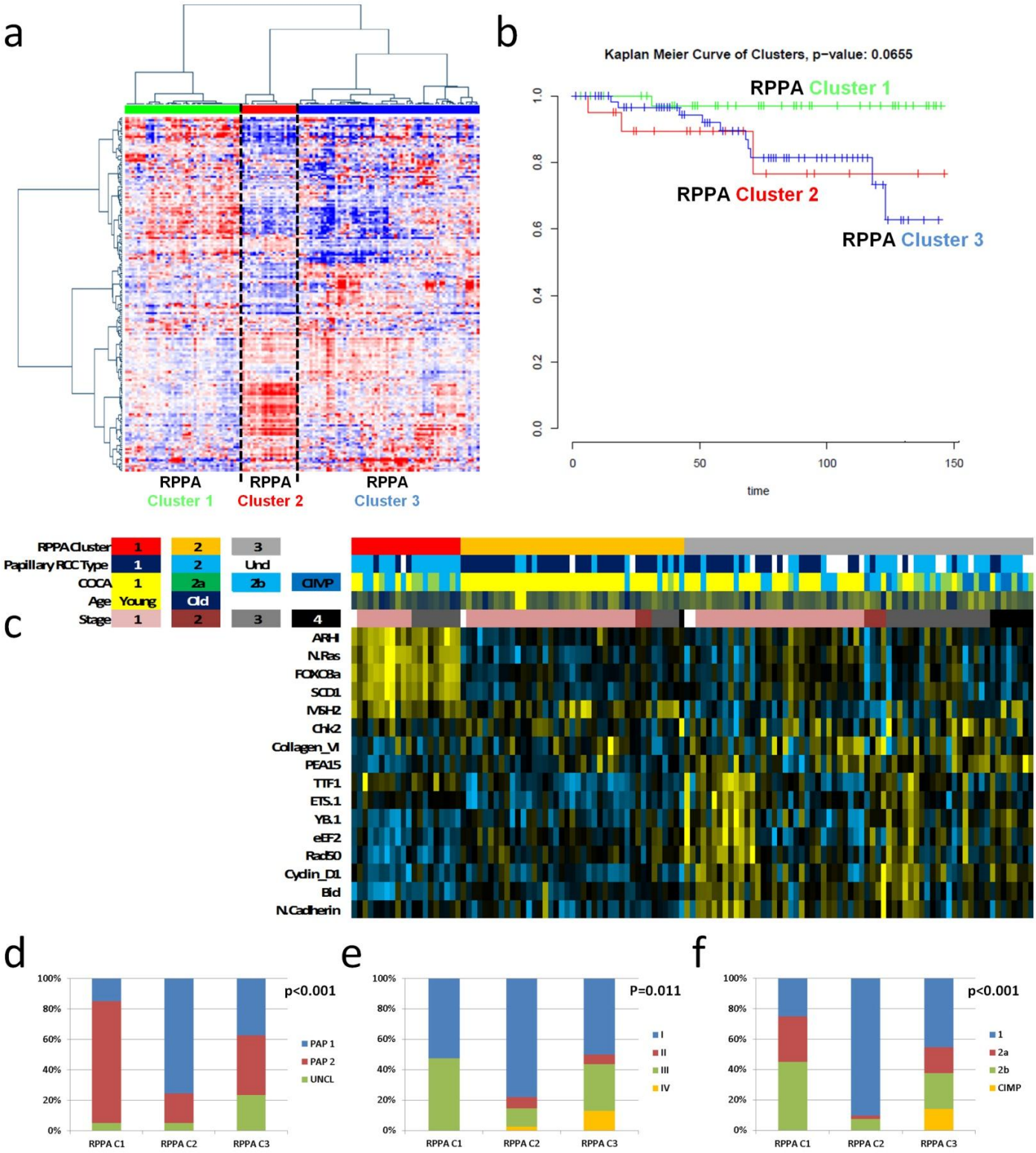


Figure S13: Clustering of papillary RCC tumors based on RPPA data reveals distinct subtypes

(a) Consensus Cluster analysis of 193 RPPA probes from 125 papillary RCC tumors reveals three distinct clusters. **(b)** A trend towards superior overall survival for RPPA cluster 2 tumors relative to other tumors was observed. **(c)** A heatmap for the 16 most differentially expressed proteins annotated for Papillary type, COCA cluster, age and stage. The three RPPA clusters have distinct clinical and molecular features: **(d)** RPPA cluster 1 is contains mostly papillary type 2 tumors while RPPA cluster 2 contains predominantly papillary type 1 tumors. Meanwhile, RPPA cluster 3 tumors contain more histologically unclassified papillary tumors, **(e)** stage IV tumors, and **(f)** CIMP tumors.

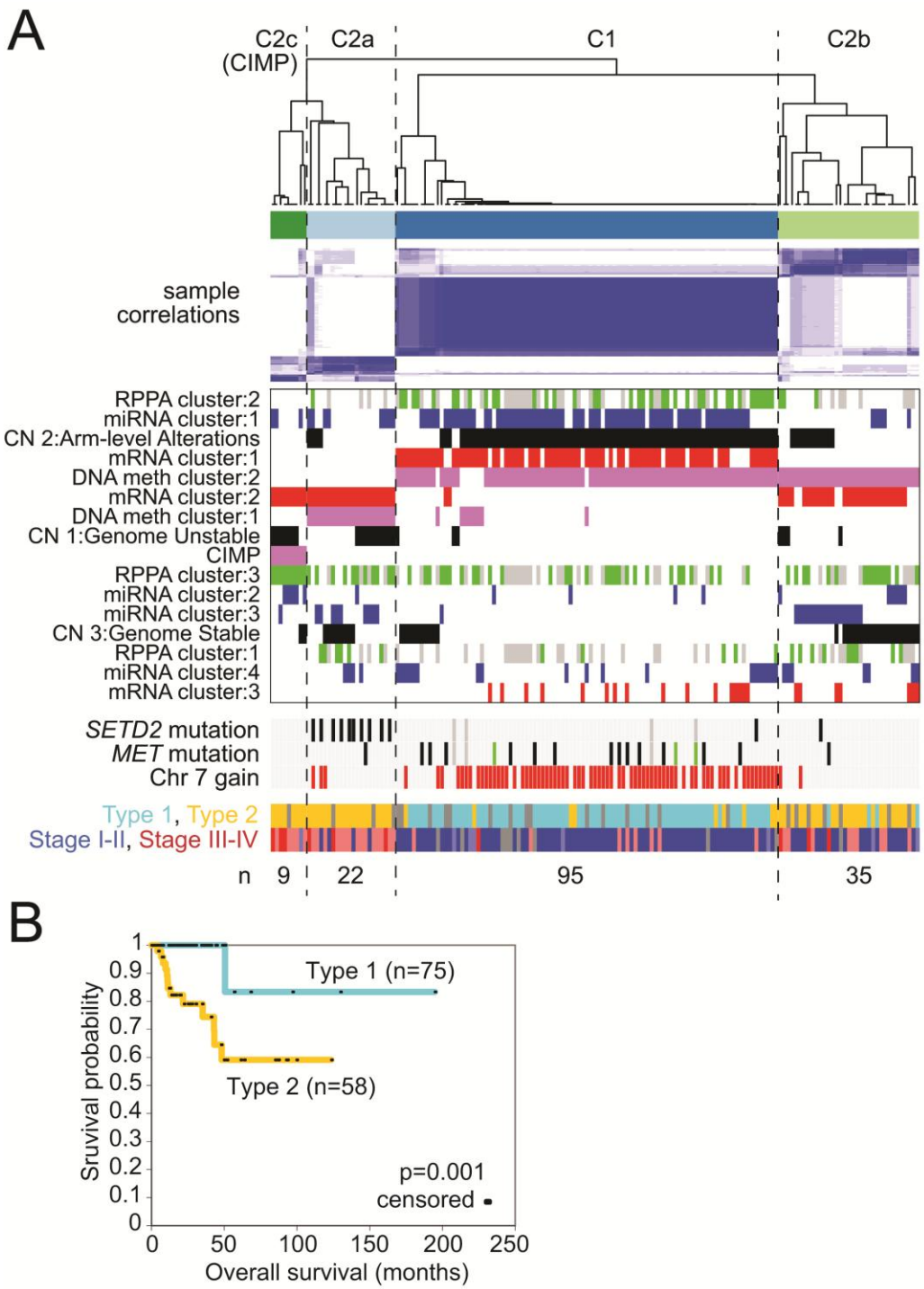


Figure S14: Multi-platform-based Cluster of Cluster Analysis (COCA) in PRCC

(a) Integration of subtype classifications from five "omic" data platforms identified four major PRCC groups: C1, C2a, C2b, and C2c (representing the CIMP PRCCs). The blue and white heat map displays sample consensus, below which a second heat map displays the subtypes defined independently by DNA methylation (Pink), Chromosomal copy number (CN)(Black), miRNA expression (Blue), mRNA expression (Red), and protein (RPPA) expression (Green, Gray represents samples missing RPPA expression data). Somatic (Black) and Germline (Green) mutations for *MET* and *SETD2* and gain of chromosome 7 (Red) was indicated for each tumor. Tumor type and stage were designated with Type 1 in light blue, Type 2 in yellow, and those not classified in gray, while stages I and II were shown in dark blue or light blue respectively, and stages III and IV were shown in light red or dark red respectively. **(b)** Differences in patient overall survival between histology-based subtypes Type 1 and Type 2 (left, p=0.001, log-rank).

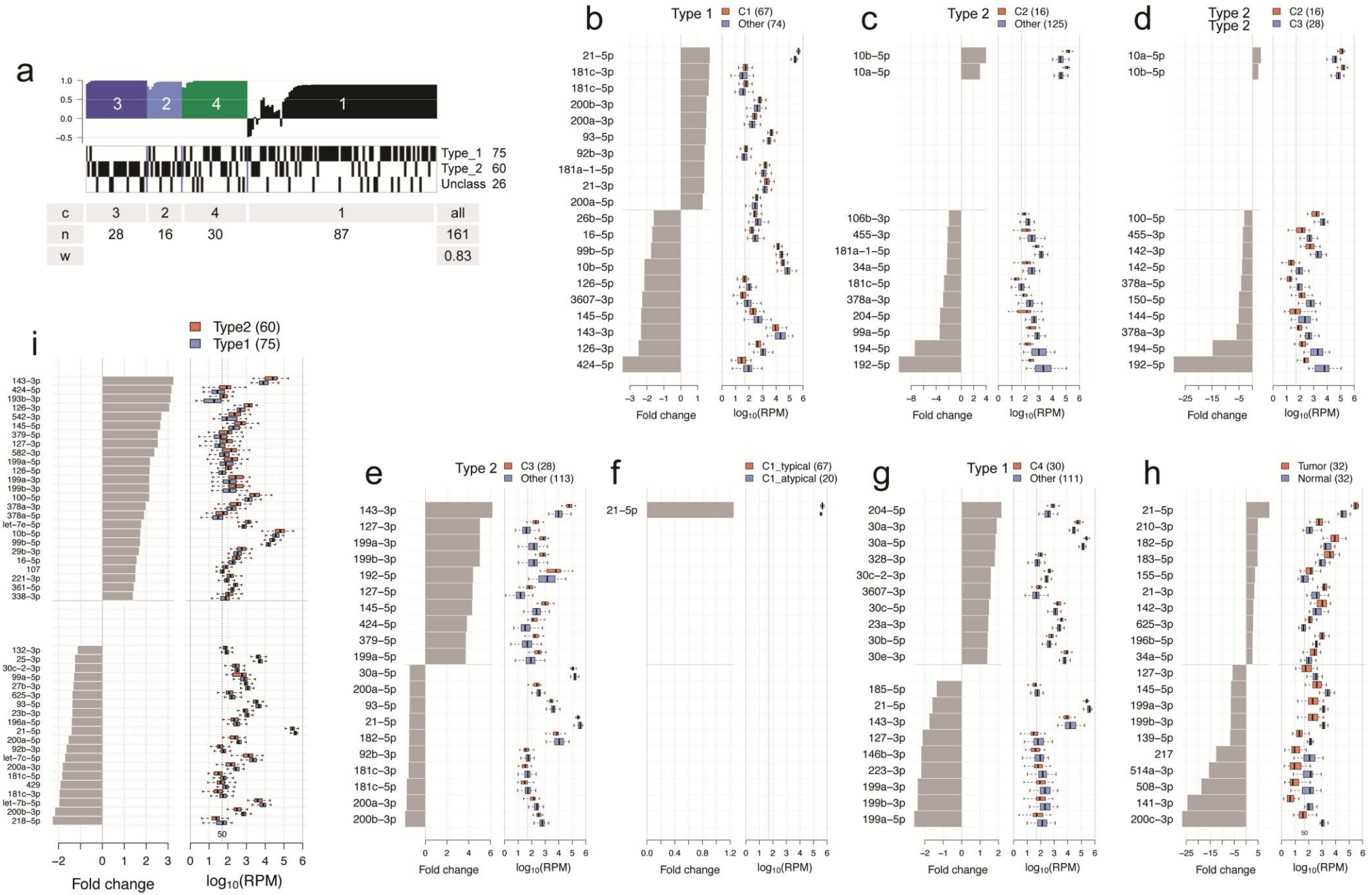
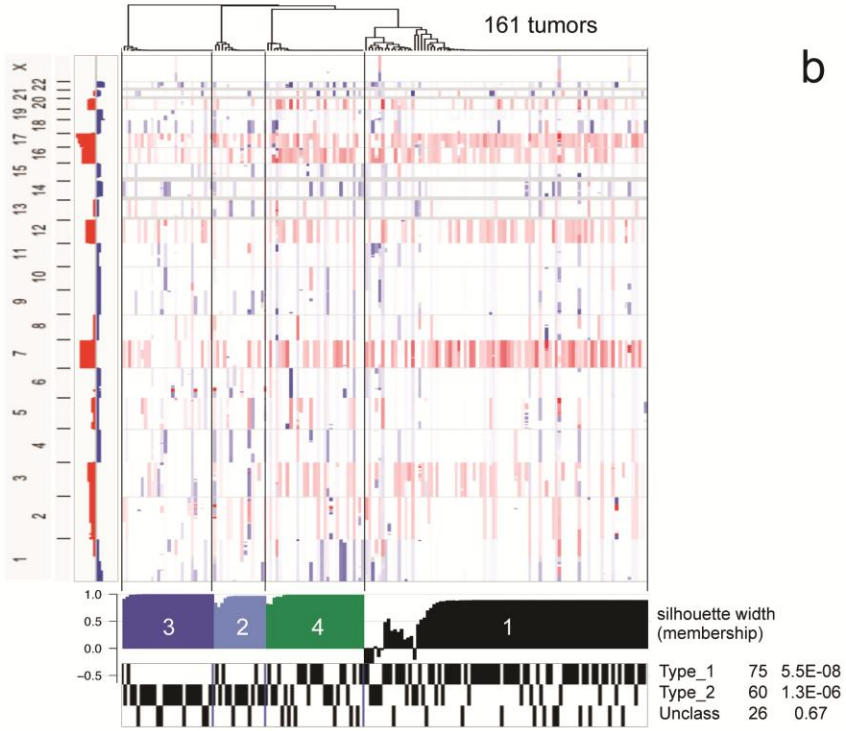

a
1.0
0.5
0.0
−0.5
3
2
4
1
Type_1 75
Type_2 60
Unclass 26
c 3 2 4 1 all
n 28 16 30 87 161
w 0.83
b
Type 1
C1 (67)
Other (74)
21–5p
181c–3p
181c–5p
200b–3p
200a–3p
93–5p
92b–3p
181a–1–5p
21–3p
200a–5p
26b–5p
16–5p
99b–5p
10b–5p
126–5p
3607–3p
145–5p
143–3p
126–3p
424–5p
Fold change
log10(RPM)
c
Type 2
C2 (16)
Other (125)
10b–5p
10a–5p
106b–3p
455–3p
181a–1–5p
34a–5p
181c–5p
378a–3p
204–5p
99a–5p
194–5p
192–5p
Fold change
log10(RPM)
d
Type 2
Type 2
C2 (16)
C3 (28)
10a–5p
10b–5p
100–5p
455–3p
142–3p
142–5p
378a–5p
150–5p
144–5p
378a–3p
194–5p
192–5p
Fold change
log10(RPM)
i
Type2 (60)
Type1 (75)
143–3p
424–5p
193b–3p
126–3p
542–3p
145–5p
379–5p
127–3p
582–3p
199a–5p
126–5p
199a–3p
199b–3p
100–5p
378a–3p
378a–5p
let–7e–5p
10b–5p
99b–5p
29b–3p
16–5p
107
221–3p
361–5p
338–3p
132–3p
25–3p
30c–2–3p
99a–5p
27b–3p
625–3p
93–5p
23b–3p
196a–5p
21–5p
200a–5p
92b–3p
let–7c–5p
200a–3p
181c–5p
429
181c–3p
let–7b–5p
200b–3p
218–5p
50
Fold change
log10(RPM)
e
Type 2
C3 (28)
Other (113)
143–3p
127–3p
199a–3p
199b–3p
192–5p
127–5p
145–5p
424–5p
379–5p
199a–5p
30a–5p
200a–5p
93–5p
21–5p
182–5p
92b–3p
181c–3p
181c–5p
200a–3p
200b–3p
Fold change
log10(RPM)
f
C1_typical (67)
C1_atypical (20)
21–5p
Fold change
log10(RPM)
g
Type 1
C4 (30)
Other (111)
204–5p
30a–3p
30a–5p
328–3p
30c–2–3p
3607–3p
30c–5p
23a–3p
30b–5p
30e–3p
185–5p
21–5p
143–3p
127–3p
146b–3p
223–3p
199a–3p
199b–3p
199a–5p
Fold change
log10(RPM)
h
Tumor (32)
Normal (32)
21–5p
210–3p
182–5p
183–5p
155–5p
21–3p
142–3p
625–3p
196b–5p
34a–5p
127–3p
145–5p
199a–3p
199b–3p
139–5p
217
514a–3p
508–3p
141–3p
200c–3p
50
Fold change
log10(RPM)

Figure S15: Differentially Abundant miRs

(a) Silhouette width profile, histology and summary table for the four-cluster NMF solution.

(b-i) miRs that were differentially abundant between pairs of unsupervised miR sample groups. **(b)** Cluster 1 vs. all other samples. **(c)** Cluster 2 vs. all other samples. **(d)** Cluster 2 vs. cluster 3. **(e)** Cluster 3 vs. all other samples. **(f)** Within cluster 2, typical vs. atypical samples, separating samples by the silhouette width profile. **(g)** Cluster 4 vs. all other samples. **(h)** Tumor samples vs. matched adjacent normals, n=32. **(i)** All Type 1 vs. all type 2 samples. Each panel has (left) a barplot of median-based fold change, and (right) boxplots showing distributions of normalized (RPM) abundance, with black vertical lines indicating medians. Up to 15 (25 in panel i) of the largest fold changes in each direction are shown. The numbers of samples in each group are in parentheses. Only miRs that have a mean abundance of at least 50 RPM are shown. Results from two-class comparisons were consistent with NMF results (**Fig. S12d** and **S12e**). For example, miR-21-5p had both the largest positive fold change and the highest absolute abundance, and, of miRs with negative fold changes, miRs 99-5p, 10b-5p and 143-3p were the most abundant. Of miRs that were differentially abundant between 32 matched tumors and adjacent normals, the largest negative fold changes (i.e. less abundant in tumor samples) were miR-200c-3p and 141-3p from the miR-200-family, followed by miR-508-3p and miR-514a-3p, two miRs from an Xq27.3 genomic cluster that includes 15 miRNAs. Many miRs were differentially expressed between Type 2 and 1 samples. Of miRs that were more abundant in Type 2 (i.e. positive fold changes), miR-143-3p and miR-10a-5p were the most abundant, while for negative fold changes, miR-21-5p was the most abundant, followed by let-7b-5p, 25-3p, 93-5p, let-7c-5p, and 200b-3p. The only miR that was differentially abundant between the typical and atypical samples in miR cluster 1 was miR-21.

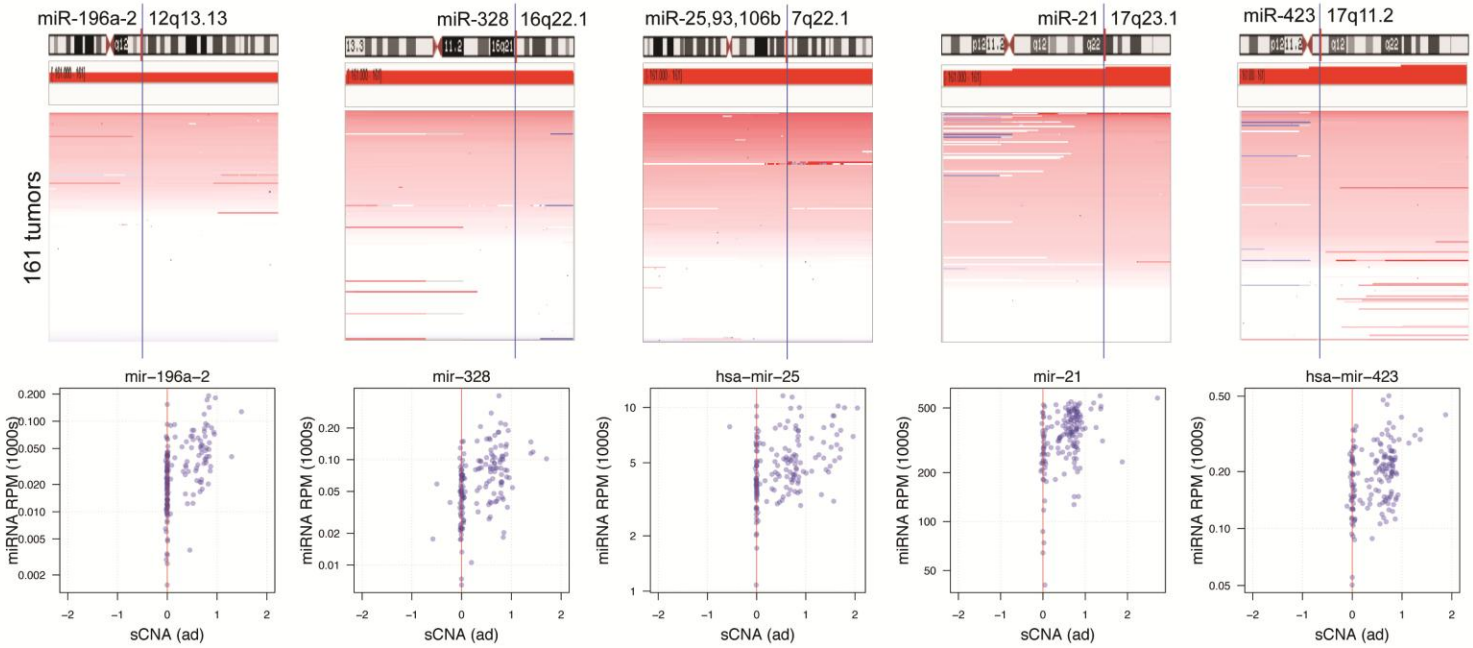
a



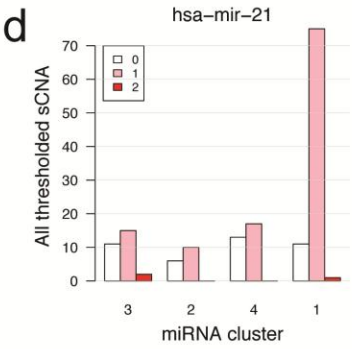
b

miRNA	Cytoband	mn(RPM)	rho	P.adj	Comment
hsa-mir-196a-2	12q13.13	37.9	0.55	0.0	HOXC cluster, HOTAIR
hsa-mir-429	1p36.33	72.9	0.47	1.1E-07	200a,b, 429 cluster
hsa-mir-328	16q22.1	75.2	0.47	1.2E-07	
hsa-mir-200a	1p36.33	597.8	0.46	3.8E-07	200a,b, 429 cluster
hsa-mir-93	7q22.1	4192.4	0.45	1.3E-06	25,93,106b cluster
hsa-mir-1180	17p11.2	45.8	0.43	6.8E-06	
hsa-mir-193b	16p13.12	52.9	0.43	7.4E-06	
hsa-mir-200b	1p36.33	586.8	0.42	9.5E-06	200a,b, 429 cluster
hsa-mir-339	7p22.3	39.3	0.40	3.8E-05	
hsa-mir-25	7q22.1	5139.6	0.39	1.1E-04	25,93,106b cluster
hsa-mir-21	17q23.1	350532.3	0.38	2.1E-04	
hsa-mir-148b	12q13.13	141.0	0.34	2.5E-03	
hsa-mir-423	17q11.2	203.5	0.33	6.1E-03	
hsa-let-7i	12q14.1	413.4	0.32	0.013	
hsa-mir-589	7p22.1	79.7	0.31	0.018	
hsa-mir-23b	9q22.32	1046.0	0.30	0.026	23b,27b,24-1,3074 cluster
hsa-mir-185	22q11.21	70.0	0.29	0.051	
hsa-mir-31	9p21.3	73.7	0.28	0.072	

c



d



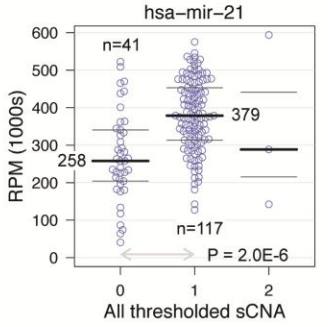


Figure S16: Somatic Copy Number Alterations (sCNA) and miRNAs

(a) sCNA heatmap, ordered by the miRNA unsupervised clustering solution (**Fig. S12d**). The Type 2-enriched miR clusters 2 and 3 had fewer amplifications than the other two miR clusters in chromosomes 3, 7, 12, 16 and 17.

(b) pre-miRNAs that are statistically associated with sCNA (Spearman correlation, BH-corrected $P < 0.1$). Samples shown have mean RPM > 25. The relatively few miRNAs had were strongly associated with sCNA included the miR-200a,b/429 genomic cluster, and the miR-25~106b genomic cluster, which is paralogous to the miR-17~92a cluster.[27] **(c)** Per-chromosome sCNA heatmaps and sCNA-RPM scatterplots for a subset of pre-miRNAs in **(b)**. Vertical lines show the locations of the miRNAs. These details help interpret the correlation results in **(b)**. **(d)** Above: Number of samples with thresholded sCNA of 0, 1 or 2 across the miR unsupervised clusters. Below: the RPM abundance of miR-21 as a function of thresholded sCNA. The P-value between sCNA=0 and 1 is from a one-sided KS test. Of the sCNA-associated miRNAs, miR-21 on 17q23.1 was by far the most abundant **(b)**, and so is likely influential.[27] sCNA appears to contribute to miR-21 being relatively abundant in miR cluster 1, as copy number gains were enriched in this miR cluster and the miRNA was more abundant with such gains (one-sided KS test, P=2.0E-6).

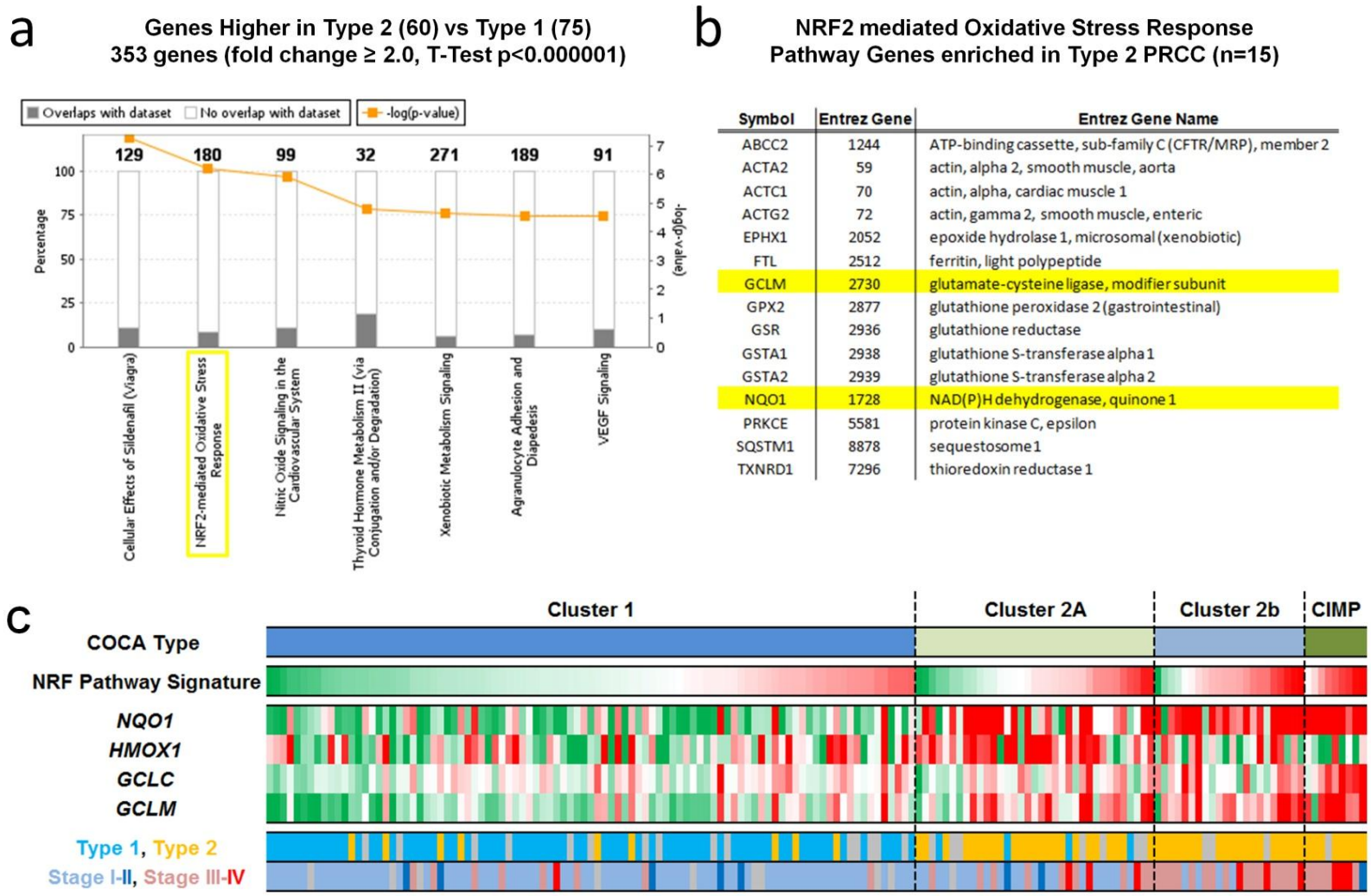


Figure S17: mRNA Pathway Analysis Comparing Type 2 PRCC to Type 1 PRCC

The RNA-Seq data for the 60 Type 2 PRCC tumors was compared to the 75 Type 1 PRCC tumors and the genes that demonstrated a 2 fold increase in Type 2 PRCC compared to Type 1 PRCC with a t-test p-value of <0.00001 were selected to represent the most differentially expressed genes (n=353). **(a)** This selection of 353 genes was assessed using the Ingenuity Core Analysis Software (https://analysis.ingenuity.com/pa/) to identify enriched biological pathways. This demonstrated that increased expression of the NRF2 mediated oxidative stress response pathway genes was enriched in Type 2 PRCC. **(b)** 15 genes within the NRF2 mediated pathway were identified within the selected 353 genes including two genes, *NQO1* and *GCLM* (highlighted in yellow), commonly associated with NRF2 activation. **(c)** A heatmap of the relative expression these two genes as well as a previously published NRF2 pathway activation mRNA signature score[44] was generated and ordered firstly by COCA subtype and secondly by NRF2 pathway mRNA signature score from low to high.

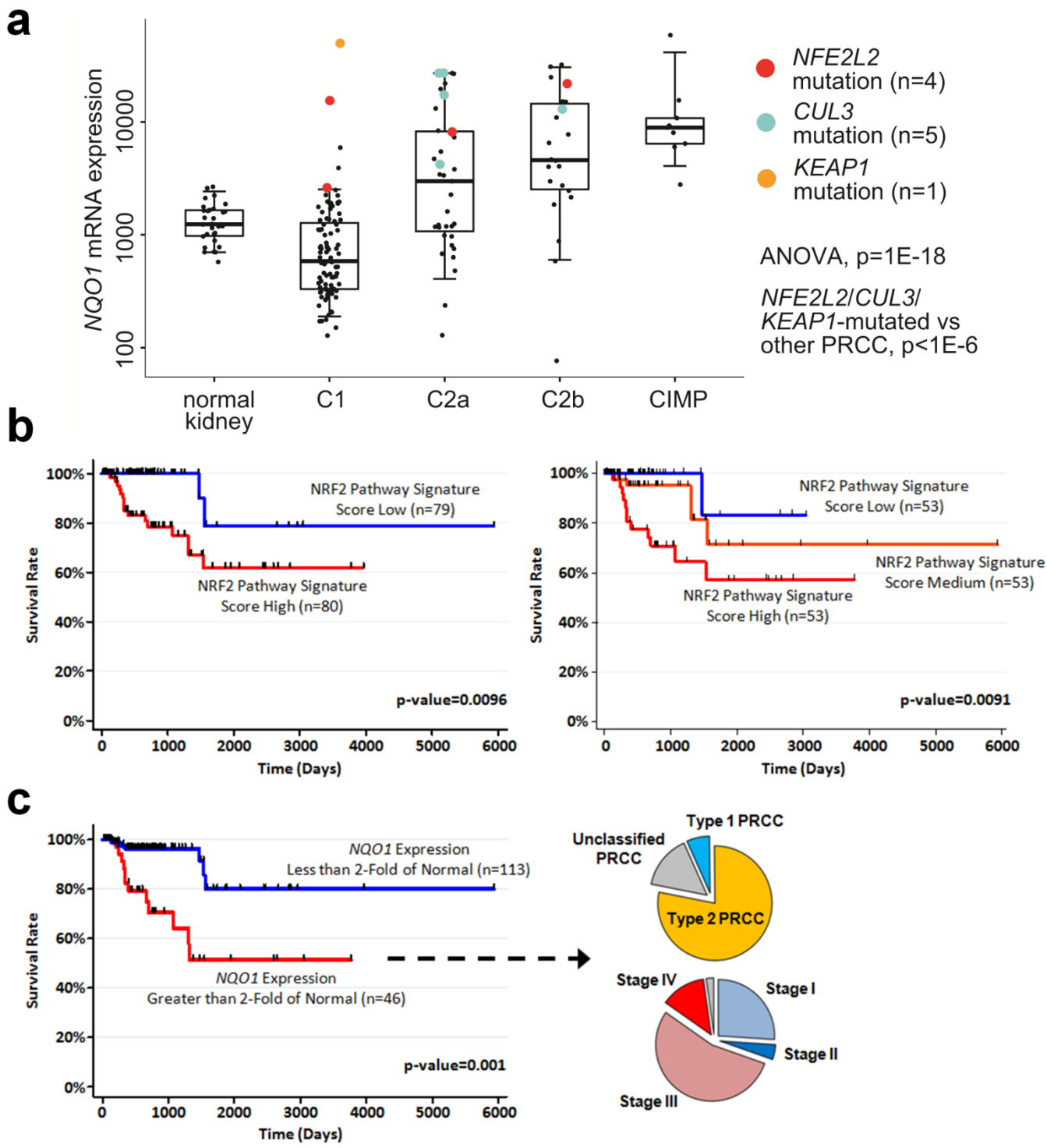


Figure S18: Survival Analysis of the NRF2-ARE Pathway and its Critical Marker Gene, *NQO1*

(a) Differential mRNA expression of NRF2-ARE pathway marker *NQO1* among the multi-platform-based subtypes and normal kidney, with cases harboring somatic mutation in NRF2-ARE pathway members *NFE2L2*, *KEAP1*, or *CUL3* being highlighted (p<1E-6, t-test, comparing PRCC harboring *NFE2L2*, *KEAP1*, or *CUL3* mutations with other PRCC).
(b) Kaplan-Meier survival analyses were performed comparing the PRCC tumors based on a NRF2 pathway activation mRNA signature score[44] with the samples either split equally into high and low score or split three ways. (2 samples lacked survival data thus n=159).
(c) Kaplan-Meier survival analysis was performed comparing the PRCC tumors based on the expression of a critical and well-studied NRF2-ARE pathway activated gene, *NQO1*. Increased *NQO1* expression was defined being greater than 2-fold of the average expression calculated from the normal kidney samples (n=32). The PRCC type and tumor stage breakdown in shown for the 46 tumors with increased *NQO1* expression.

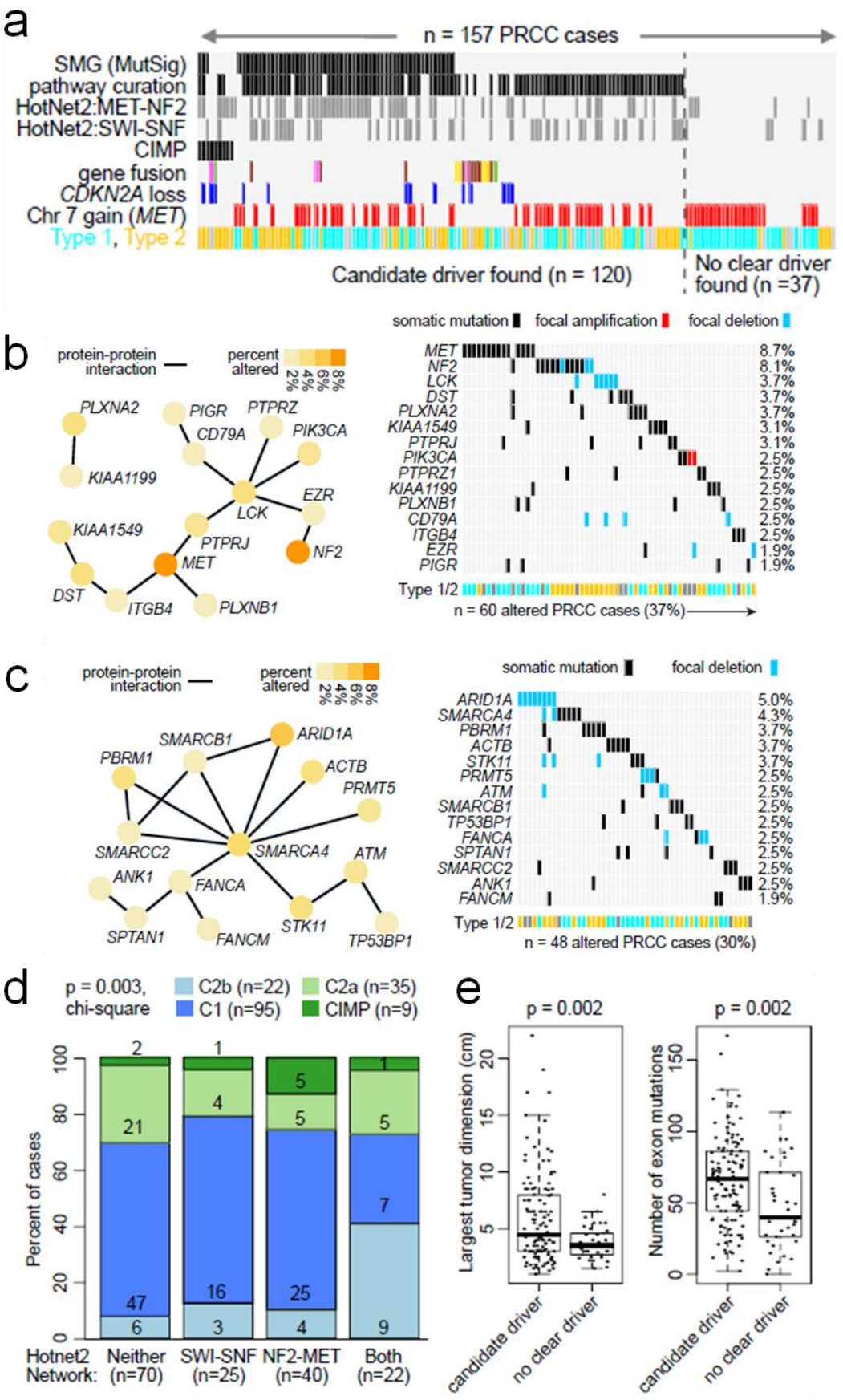

a
n = 157 PRCC cases
SMG (MutSig)
pathway curation
HotNet2:MET-NF2
HotNet2:SWI-SNF
CIMP
gene fusion
CDKN2A loss
Chr 7 gain (MET)
Type 1, Type 2
Candidate driver found (n = 120)
No clear driver found (n =37)
b
protein-protein interaction
percent altered
2%
4%
6%
8%
PLXNA2
PIGR
PTPRZ
CD79A
PIK3CA
KIAA1199
EZR
LCK
KIAA1549
PTPRJ
NF2
MET
DST
ITGB4
PLXNB1
somatic mutation
focal amplification
focal deletion
MET 8.7%
NF2 8.1%
LCK 3.7%
DST 3.7%
PLXNA2 3.7%
KIAA1549 3.1%
PTPRJ 3.1%
PIK3CA 2.5%
PTPRZ1 2.5%
KIAA1199 2.5%
PLXNB1 2.5%
CD79A 2.5%
ITGB4 2.5%
EZR 1.9%
PIGR 1.9%
Type 1/2
n = 60 altered PRCC cases (37%)
c
protein-protein interaction
percent altered
2%
4%
6%
8%
SMARCB1
ARID1A
PBRM1
ACTB
PRMT5
SMARCC2
SMARCA4
ATM
ANK1
FANCA
SPTAN1
FANCM
STK11
TP53BP1
somatic mutation
focal deletion
ARID1A 5.0%
SMARCA4 4.3%
PBRM1 3.7%
ACTB 3.7%
STK11 3.7%
PRMT5 2.5%
ATM 2.5%
SMARCB1 2.5%
TP53BP1 2.5%
FANCA 2.5%
SPTAN1 2.5%
SMARCC2 2.5%
ANK1 2.5%
FANCM 1.9%
Type 1/2
n = 48 altered PRCC cases (30%)
d
p = 0.003, chi-square
C2b (n=22)
C2a (n=35)
C1 (n=95)
CIMP (n=9)
Percent of cases
0
20
40
60
80
100
2
21
47
6
1
4
16
3
5
5
25
4
1
5
7
9
Hotnet2 Network:
Neither (n=70)
SWI-SNF (n=25)
NF2-MET (n=40)
Both (n=22)
e
p = 0.002
Largest tumor dimension (cm)
5
10
15
20
candidate driver
no clear driver
p = 0.002
Number of exon mutations
0
50
100
150
candidate driver
no clear driver

Figure S19: Candidate Driver Mutations and HotNet2 Network analysis of PRCCs

(a) Candidate driver analysis in TCGA PRCC cases reveals a subset of tumors with no obvious drivers. Green bars in the SMG row indicate germline mutations of the MET gene. Colors in the gene fusion row indicate TFE3/TFEB fusions in brown, MET fusions in green, RPL11-TCEB3 fusions in pink and other remaining single instance gene fusions in yellow. The CDKN2A loss row represents samples with CDKN2A mutation, focal deletion or epigenetic silencing.

(b-c) By integrating mutation and focal copy number data with databases of protein-protein interactions, the HotNet2 algorithm identified two significant sub-networks ($p<0.03$): **(b)** one network involving MET and NF2 and associated genes, and **(c)** a second network involving SWI-SNF complex genes. For **(b)** and **(c)**, heat maps shown on the right display nonsilent mutation and focal deletion or amplification events in the altered PRCC cases implicated in the network.

(d) Association between HotNet2 subnetwork and multi-platform-based subtype assignment.

(e) Largest tumor dimension and number of exonic mutations was lower in PRCC cases for which no driver candidate driver was found compared to cases with candidate drivers. P-values by t-test.

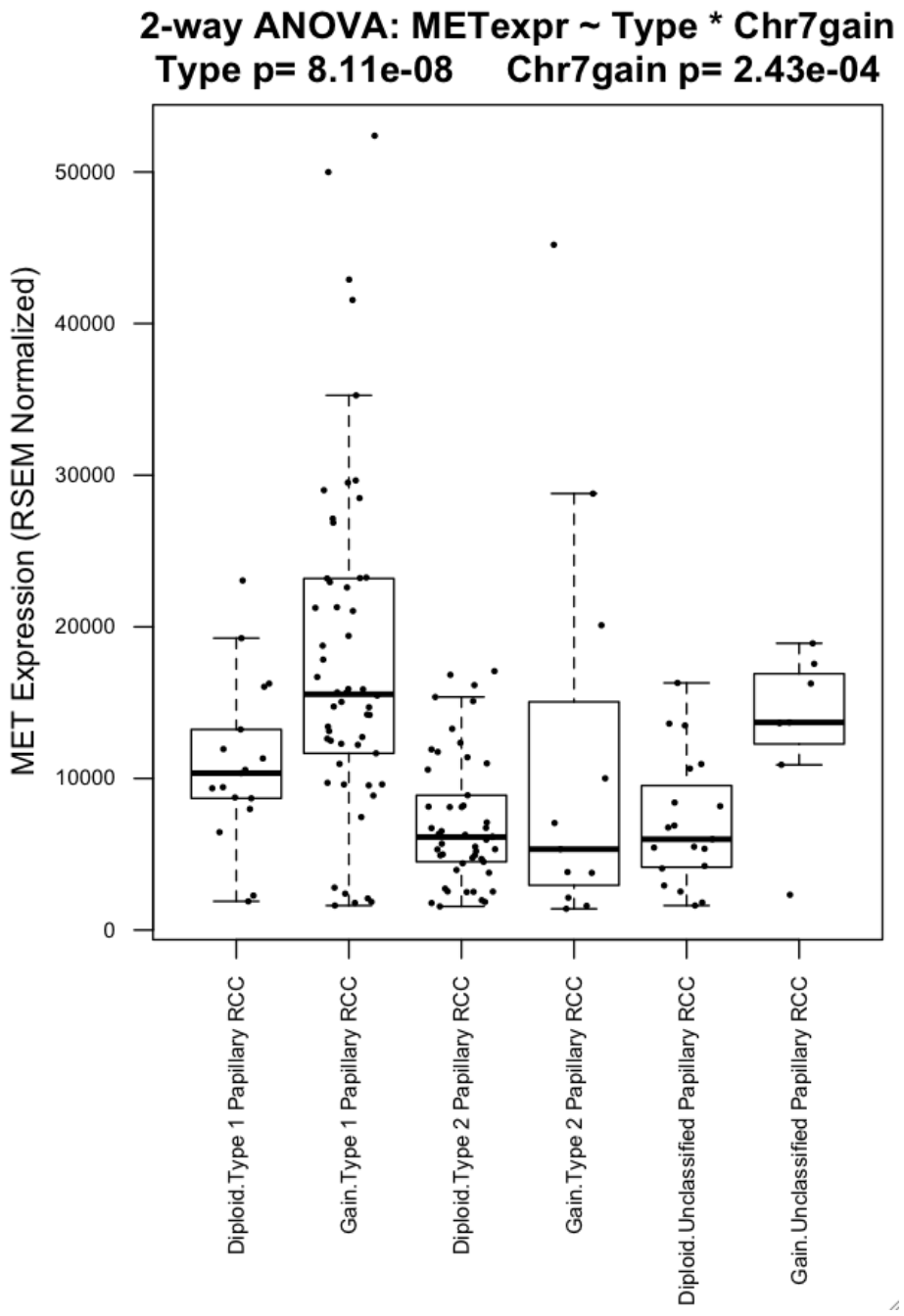
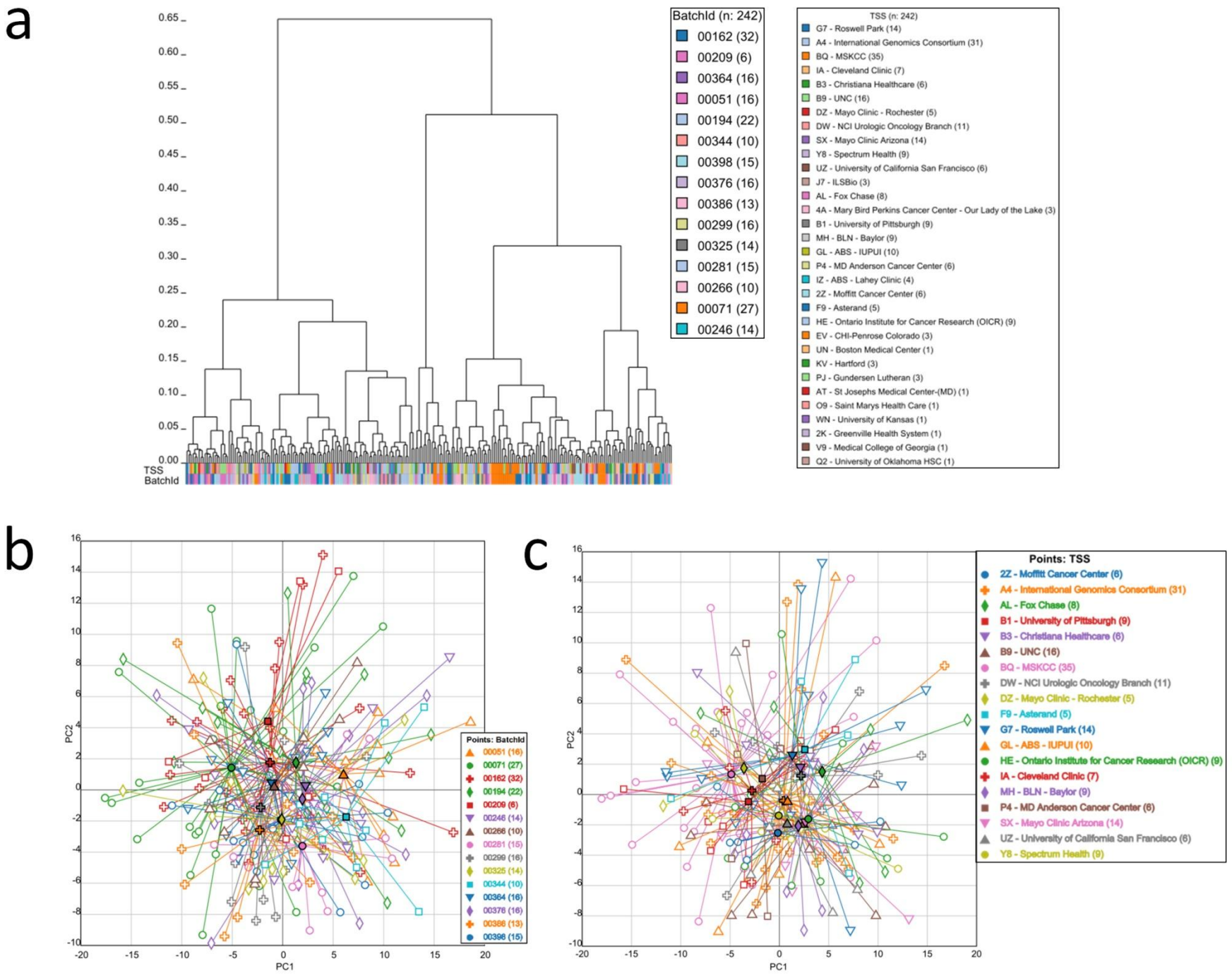
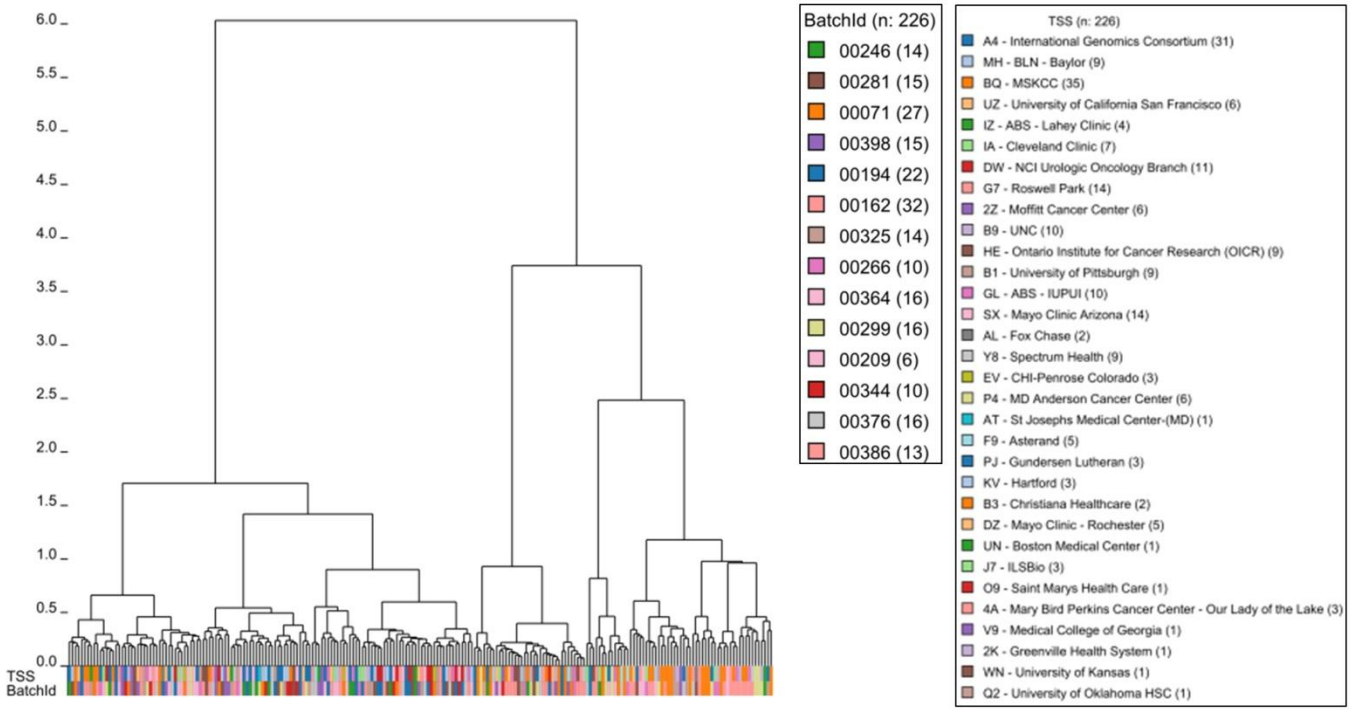


Figure S20: *MET* Expression Dependent Upon PRCC Histological Type and *MET* Copy-Number

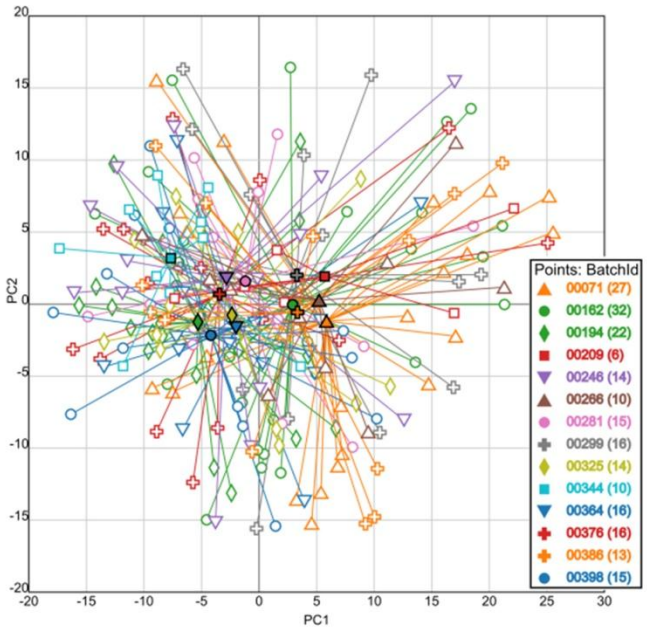
The RNA-Seq data for the *MET* gene was plotted dependent upon the histological PRCC Type of each individual tumor and the presence of copy-number gain of the *MET* gene. The Type 1 PRCC and unclassified PRCC tumors demonstrated increased *MET* expression in association with *MET* copy-number gain, while the Type 2 PRCC tumors demonstrated no obvious changes in *MET* expression when associated with *MET* copy-number gain.


a
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
TSS
BatchId
BatchId (n: 242)
00162 (32)
00209 (6)
00364 (16)
00051 (16)
00194 (22)
00344 (10)
00398 (15)
00376 (16)
00386 (13)
00299 (16)
00325 (14)
00281 (15)
00266 (10)
00071 (27)
00246 (14)
TSS (n: 242)
G7 - Roswell Park (14)
A4 - International Genomics Consortium (31)
BQ - MSKCC (35)
IA - Cleveland Clinic (7)
B3 - Christiana Healthcare (6)
B9 - UNC (16)
DZ - Mayo Clinic - Rochester (5)
DW - NCI Urologic Oncology Branch (11)
SX - Mayo Clinic Arizona (14)
Y8 - Spectrum Health (9)
UZ - University of California San Francisco (6)
J7 - ILSBio (3)
AL - Fox Chase (8)
4A - Mary Bird Perkins Cancer Center - Our Lady of the Lake (3)
B1 - University of Pittsburgh (9)
MH - BLN - Baylor (9)
GL - ABS - IUPUI (10)
P4 - MD Anderson Cancer Center (6)
IZ - ABS - Lahey Clinic (4)
2Z - Moffitt Cancer Center (6)
F9 - Asterand (5)
HE - Ontario Institute for Cancer Research (OICR) (9)
EV - CHI-Penrose Colorado (3)
UN - Boston Medical Center (1)
KV - Hartford (3)
PJ - Gundersen Lutheran (3)
AT - St Josephs Medical Center-(MD) (1)
O9 - Saint Marys Health Care (1)
WN - University of Kansas (1)
2K - Greenville Health System (1)
V9 - Medical College of Georgia (1)
Q2 - University of Oklahoma HSC (1)
b
PC2
PC1
Points: BatchId
00051 (16)
00071 (27)
00162 (32)
00194 (22)
00209 (6)
00246 (14)
00266 (10)
00281 (15)
00299 (16)
00325 (14)
00344 (10)
00364 (16)
00376 (16)
00386 (13)
00398 (15)
c
PC2
PC1
Points: TSS
2Z - Moffitt Cancer Center (6)
A4 - International Genomics Consortium (31)
AL - Fox Chase (8)
B1 - University of Pittsburgh (9)
B3 - Christiana Healthcare (6)
B9 - UNC (16)
BQ - MSKCC (35)
DW - NCI Urologic Oncology Branch (11)
DZ - Mayo Clinic - Rochester (5)
F9 - Asterand (5)
G7 - Roswell Park (14)
GL - ABS - IUPUI (10)
HE - Ontario Institute for Cancer Research (OICR) (9)
IA - Cleveland Clinic (7)
MH - BLN - Baylor (9)
P4 - MD Anderson Cancer Center (6)
SX - Mayo Clinic Arizona (14)
UZ - University of California San Francisco (6)
Y8 - Spectrum Health (9)

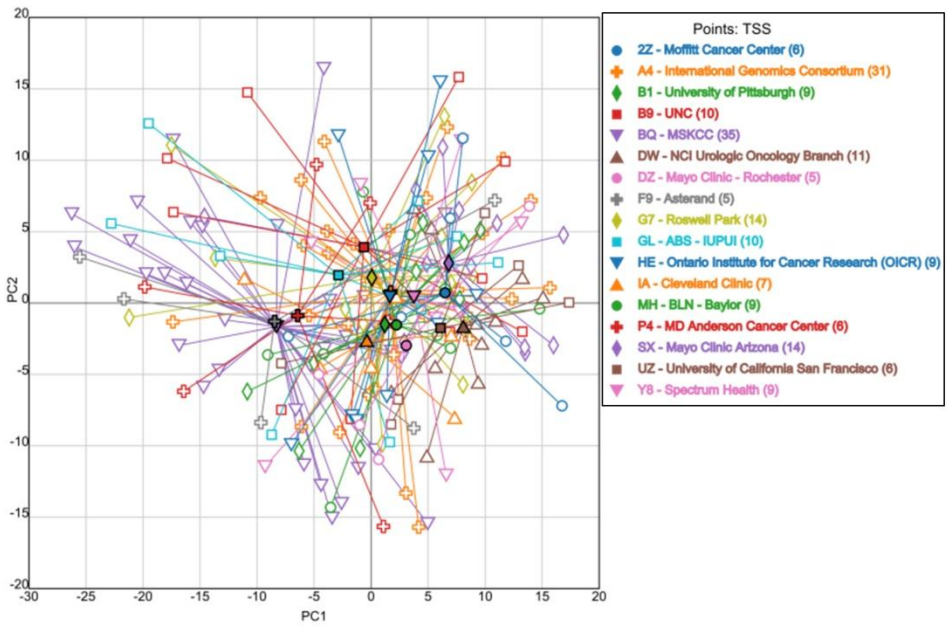
d


6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
TSS
BatchId
BatchId (n: 226)
00246 (14)
00281 (15)
00071 (27)
00398 (15)
00194 (22)
00162 (32)
00325 (14)
00266 (10)
00364 (16)
00299 (16)
00209 (6)
00344 (10)
00376 (16)
00386 (13)
TSS (n: 226)
A4 - International Genomics Consortium (31)
MH - BLN - Baylor (9)
BQ - MSKCC (35)
UZ - University of California San Francisco (6)
IZ - ABS - Lahey Clinic (4)
IA - Cleveland Clinic (7)
DW - NCI Urologic Oncology Branch (11)
G7 - Roswell Park (14)
2Z - Moffitt Cancer Center (6)
B9 - UNC (10)
HE - Ontario Institute for Cancer Research (OICR) (9)
B1 - University of Pittsburgh (9)
GL - ABS - IUPUI (10)
SX - Mayo Clinic Arizona (14)
AL - Fox Chase (2)
Y8 - Spectrum Health (9)
EV - CHI-Penrose Colorado (3)
P4 - MD Anderson Cancer Center (6)
AT - St Josephs Medical Center-(MD) (1)
F9 - Asterand (5)
PJ - Gundersen Lutheran (3)
KV - Hartford (3)
B3 - Christiana Healthcare (2)
DZ - Mayo Clinic - Rochester (5)
UN - Boston Medical Center (1)
J7 - ILSBio (3)
O9 - Saint Marys Health Care (1)
4A - Mary Bird Perkins Cancer Center - Our Lady of the Lake (3)
V9 - Medical College of Georgia (1)
2K - Greenville Health System (1)
WN - University of Kansas (1)
Q2 - University of Oklahoma HSC (1)

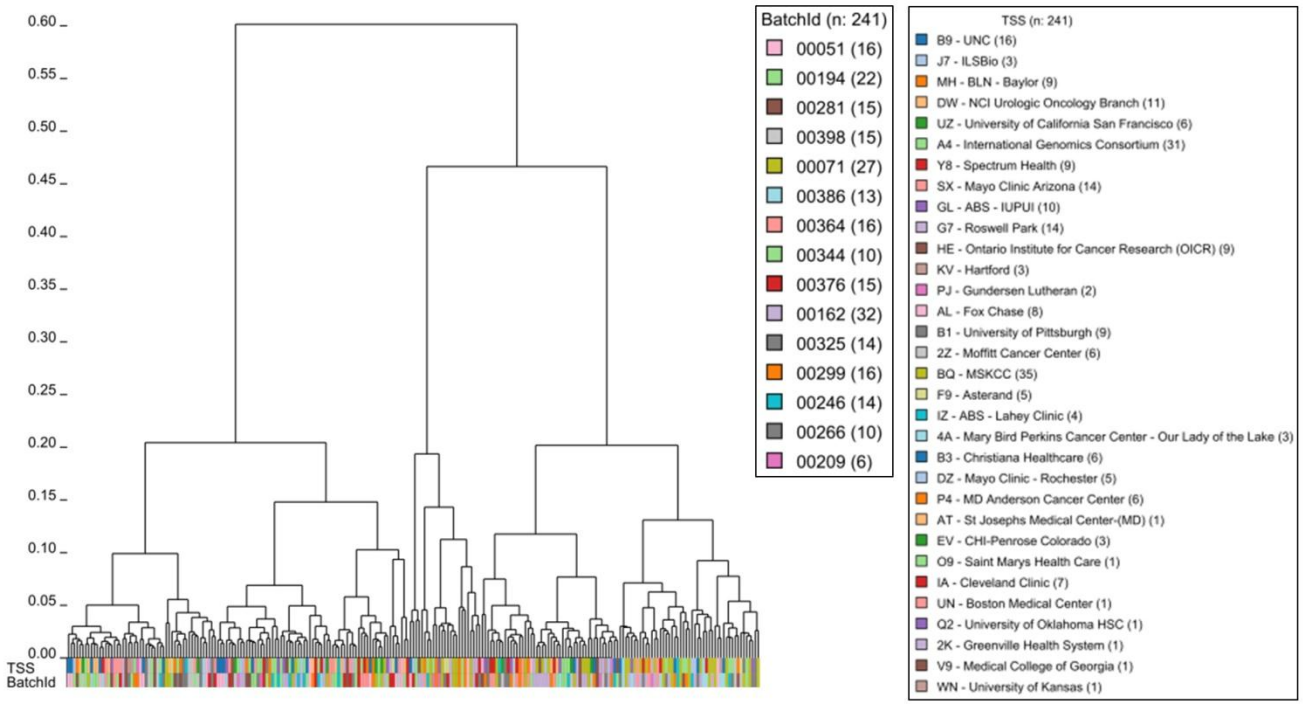
e


PC2
PC1
Points: BatchId
00071 (27)
00162 (32)
00194 (22)
00209 (6)
00246 (14)
00266 (10)
00281 (15)
00299 (16)
00325 (14)
00344 (10)
00364 (16)
00376 (16)
00386 (13)
00398 (15)

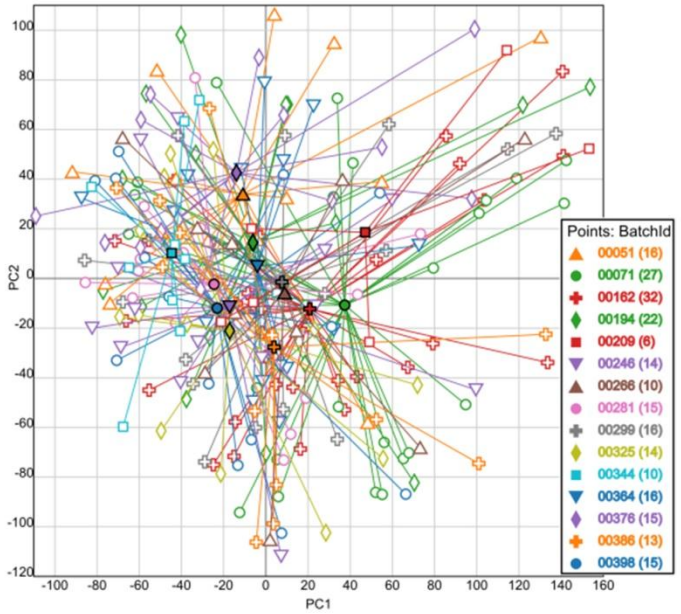
f


PC2
PC1
Points: TSS
2Z - Moffitt Cancer Center (6)
A4 - International Genomics Consortium (31)
B1 - University of Pittsburgh (9)
B9 - UNC (10)
BQ - MSKCC (35)
DW - NCI Urologic Oncology Branch (11)
DZ - Mayo Clinic - Rochester (5)
F9 - Asterand (5)
G7 - Roswell Park (14)
GL - ABS - IUPUI (10)
HE - Ontario Institute for Cancer Research (OICR) (9)
IA - Cleveland Clinic (7)
MH - BLN - Baylor (9)
P4 - MD Anderson Cancer Center (6)
SX - Mayo Clinic Arizona (14)
UZ - University of California San Francisco (6)
Y8 - Spectrum Health (9)

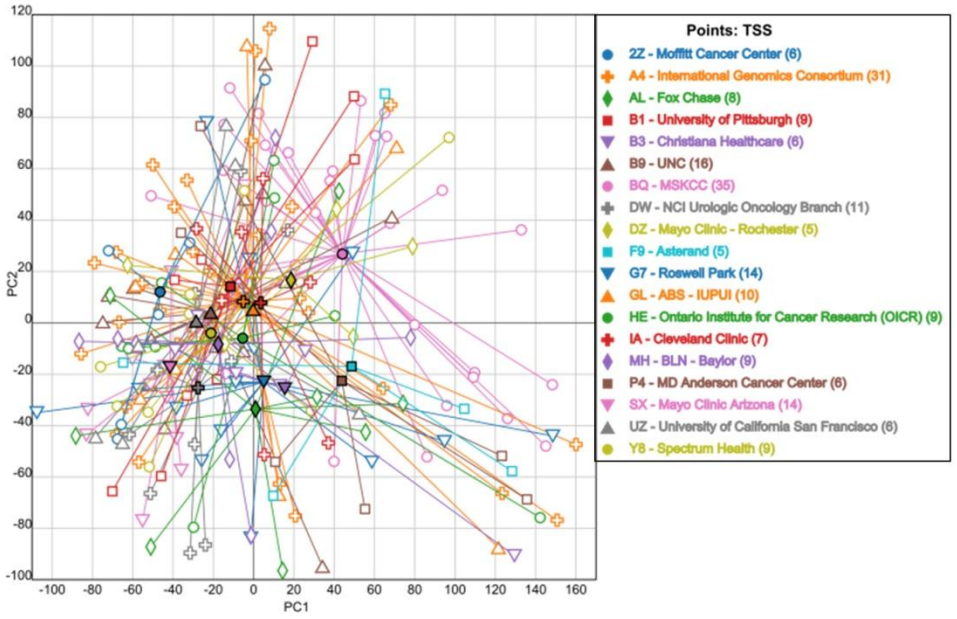
g


0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
TSS
BatchId
BatchId (n: 241)
00051 (16)
00194 (22)
00281 (15)
00398 (15)
00071 (27)
00386 (13)
00364 (16)
00344 (10)
00376 (15)
00162 (32)
00325 (14)
00299 (16)
00246 (14)
00266 (10)
00209 (6)
TSS (n: 241)
B9 - UNC (16)
J7 - ILSBio (3)
MH - BLN - Baylor (9)
DW - NCI Urologic Oncology Branch (11)
UZ - University of California San Francisco (6)
A4 - International Genomics Consortium (31)
Y8 - Spectrum Health (9)
SX - Mayo Clinic Arizona (14)
GL - ABS - IUPUI (10)
G7 - Roswell Park (14)
HE - Ontario Institute for Cancer Research (OICR) (9)
KV - Hartford (3)
PJ - Gundersen Lutheran (2)
AL - Fox Chase (8)
B1 - University of Pittsburgh (9)
2Z - Moffitt Cancer Center (6)
BQ - MSKCC (35)
F9 - Asterand (5)
IZ - ABS - Lahey Clinic (4)
4A - Mary Bird Perkins Cancer Center - Our Lady of the Lake (3)
B3 - Christiana Healthcare (6)
DZ - Mayo Clinic - Rochester (5)
P4 - MD Anderson Cancer Center (6)
AT - St Josephs Medical Center-(MD) (1)
EV - CHI-Penrose Colorado (3)
O9 - Saint Marys Health Care (1)
IA - Cleveland Clinic (7)
UN - Boston Medical Center (1)
Q2 - University of Oklahoma HSC (1)
2K - Greenville Health System (1)
V9 - Medical College of Georgia (1)
WN - University of Kansas (1)

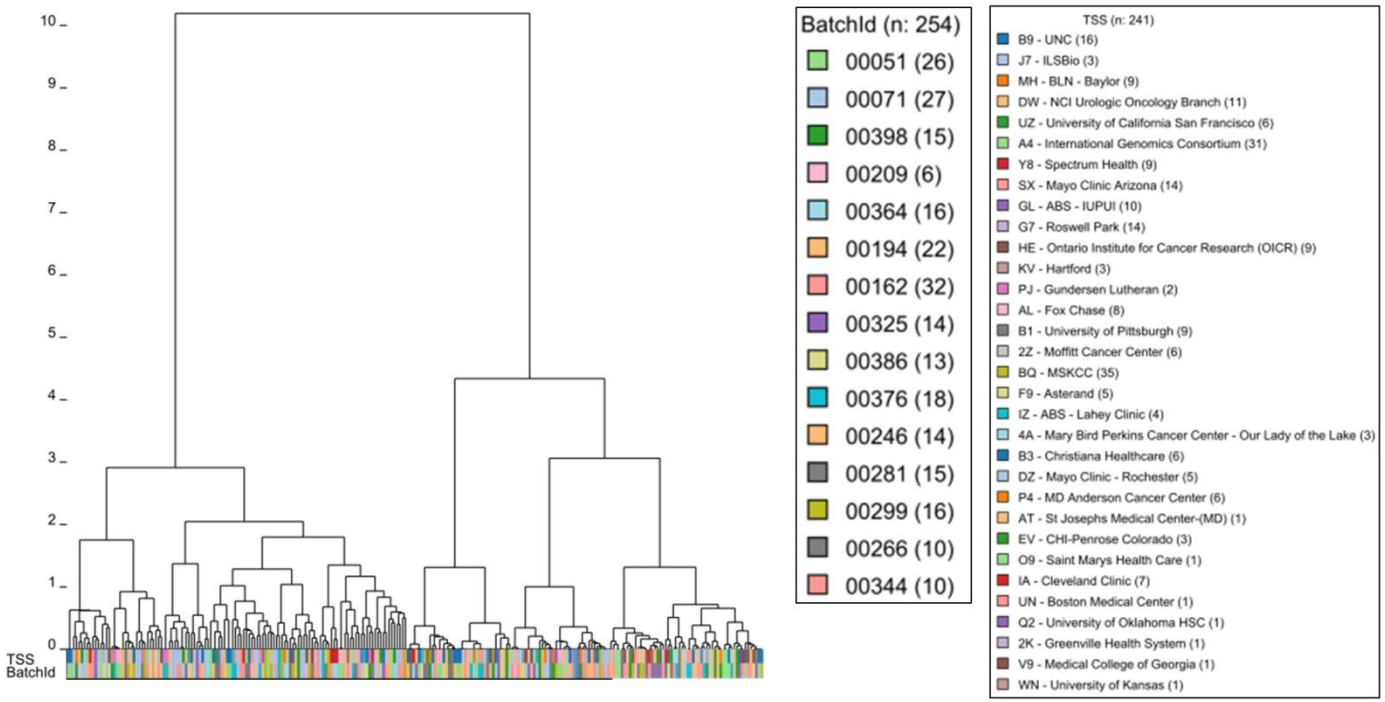
h


PC2
PC1
Points: BatchId
00051 (16)
00071 (27)
00162 (32)
00194 (22)
00209 (6)
00246 (14)
00266 (10)
00281 (15)
00299 (16)
00325 (14)
00344 (10)
00364 (16)
00376 (15)
00386 (13)
00398 (15)

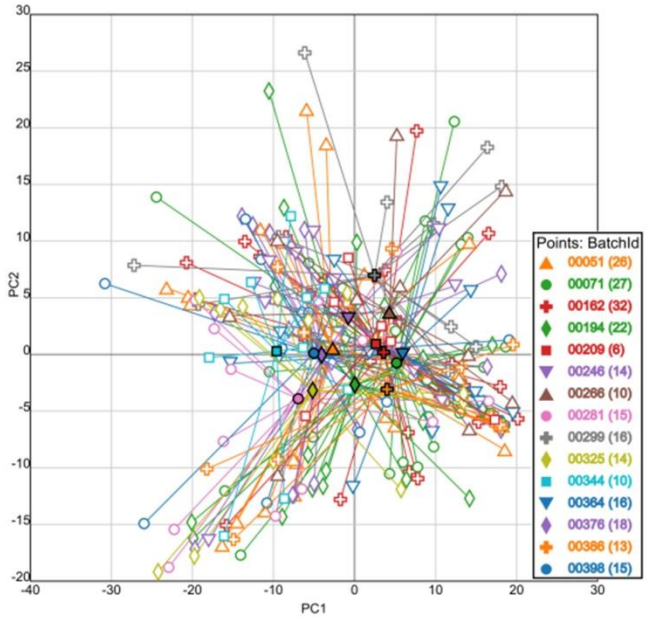
i


PC2
PC1
Points: TSS
2Z - Moffitt Cancer Center (6)
A4 - International Genomics Consortium (31)
AL - Fox Chase (8)
B1 - University of Pittsburgh (9)
B3 - Christiana Healthcare (6)
B9 - UNC (16)
BQ - MSKCC (35)
DW - NCI Urologic Oncology Branch (11)
DZ - Mayo Clinic - Rochester (5)
F9 - Asterand (5)
G7 - Roswell Park (14)
GL - ABS - IUPUI (10)
HE - Ontario Institute for Cancer Research (OICR) (9)
IA - Cleveland Clinic (7)
MH - BLN - Baylor (9)
P4 - MD Anderson Cancer Center (6)
SX - Mayo Clinic Arizona (14)
UZ - University of California San Francisco (6)
Y8 - Spectrum Health (9)

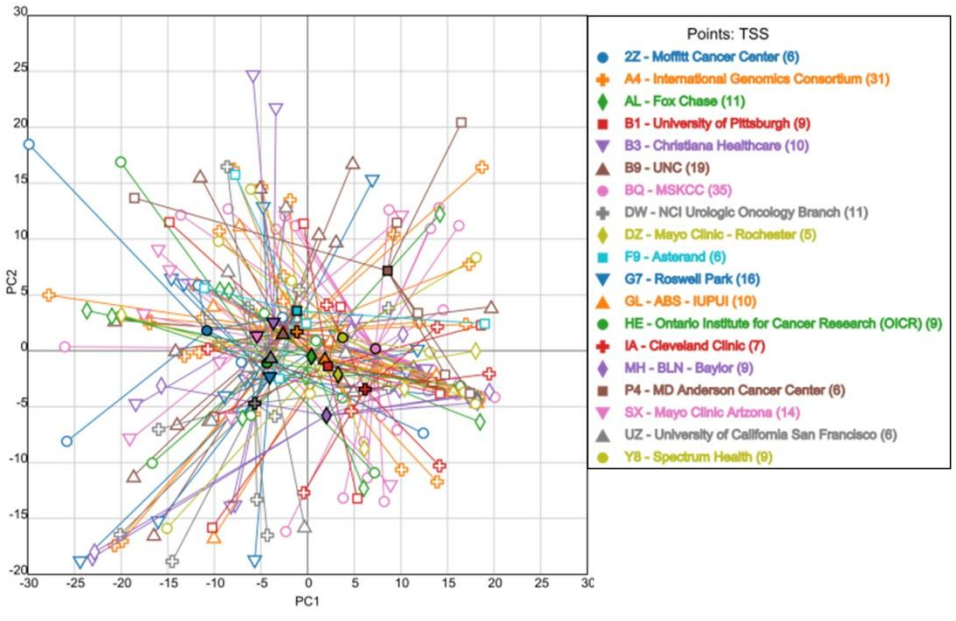
j


10
9
8
7
6
5
4
3
2
1
0
TSS
BatchId
BatchId (n: 254)
00051 (26)
00071 (27)
00398 (15)
00209 (6)
00364 (16)
00194 (22)
00162 (32)
00325 (14)
00386 (13)
00376 (18)
00246 (14)
00281 (15)
00299 (16)
00266 (10)
00344 (10)
TSS (n: 241)
B9 - UNC (16)
J7 - ILSBio (3)
MH - BLN - Baylor (9)
DW - NCI Urologic Oncology Branch (11)
UZ - University of California San Francisco (6)
A4 - International Genomics Consortium (31)
Y8 - Spectrum Health (9)
SX - Mayo Clinic Arizona (14)
GL - ABS - IUPUI (10)
G7 - Roswell Park (14)
HE - Ontario Institute for Cancer Research (OICR) (9)
KV - Hartford (3)
PJ - Gundersen Lutheran (2)
AL - Fox Chase (8)
B1 - University of Pittsburgh (9)
2Z - Moffitt Cancer Center (6)
BQ - MSKCC (35)
F9 - Asterand (5)
IZ - ABS - Lahey Clinic (4)
4A - Mary Bird Perkins Cancer Center - Our Lady of the Lake (3)
B3 - Christiana Healthcare (6)
DZ - Mayo Clinic - Rochester (5)
P4 - MD Anderson Cancer Center (6)
AT - St Josephs Medical Center-(MD) (1)
EV - CHI-Penrose Colorado (3)
O9 - Saint Marys Health Care (1)
IA - Cleveland Clinic (7)
UN - Boston Medical Center (1)
Q2 - University of Oklahoma HSC (1)
2K - Greenville Health System (1)
V9 - Medical College of Georgia (1)
WN - University of Kansas (1)

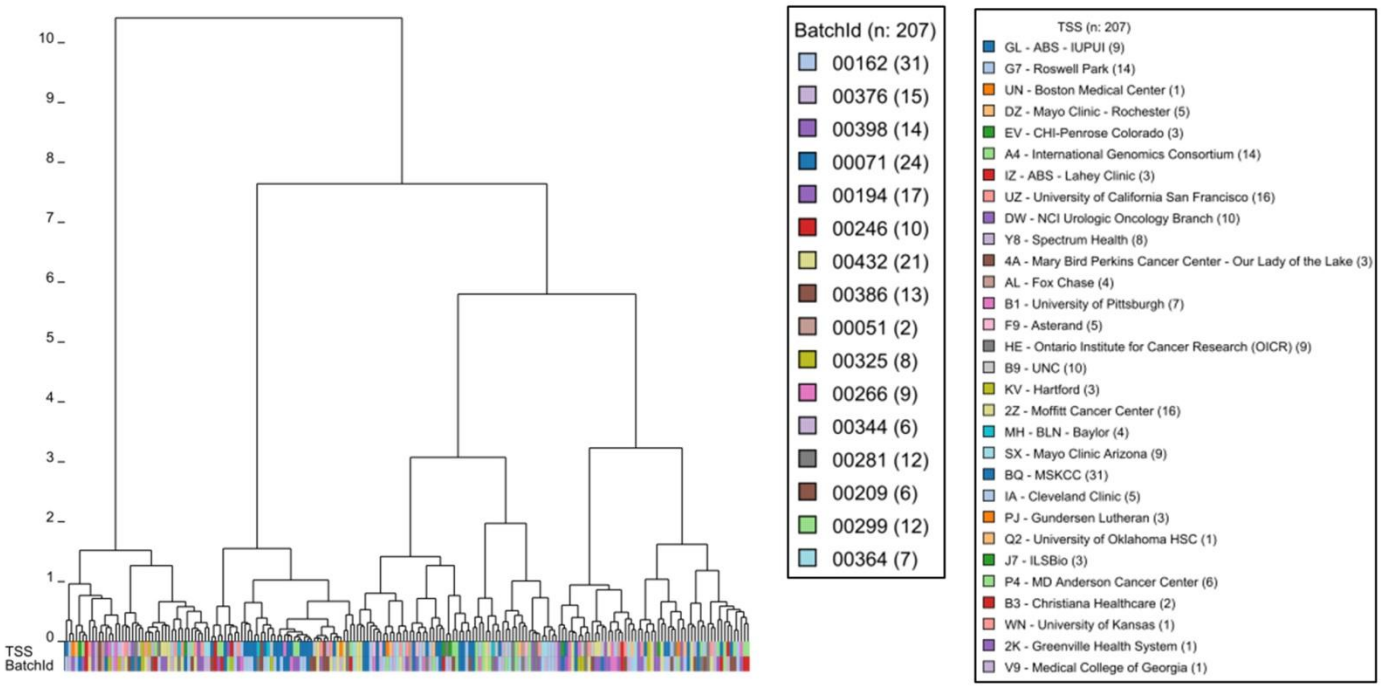
k


PC2
PC1
Points: BatchId
00051 (26)
00071 (27)
00162 (32)
00194 (22)
00209 (6)
00246 (14)
00266 (10)
00281 (15)
00299 (16)
00325 (14)
00344 (10)
00364 (16)
00376 (18)
00386 (13)
00398 (15)

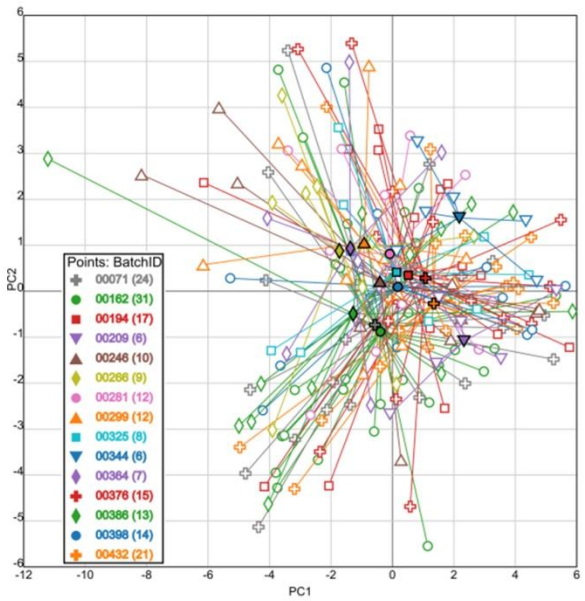
l


PC2
PC1
Points: TSS
2Z - Moffitt Cancer Center (6)
A4 - International Genomics Consortium (31)
AL - Fox Chase (11)
B1 - University of Pittsburgh (9)
B3 - Christiana Healthcare (10)
B9 - UNC (19)
BQ - MSKCC (35)
DW - NCI Urologic Oncology Branch (11)
DZ - Mayo Clinic - Rochester (5)
F9 - Asterand (6)
G7 - Roswell Park (16)
GL - ABS - IUPUI (10)
HE - Ontario Institute for Cancer Research (OICR) (9)
IA - Cleveland Clinic (7)
MH - BLN - Baylor (9)
P4 - MD Anderson Cancer Center (6)
SX - Mayo Clinic Arizona (14)
UZ - University of California San Francisco (6)
Y8 - Spectrum Health (9)

m


10
9
8
7
6
5
4
3
2
1
0
TSS
BatchId
BatchId (n: 207)
00162 (31)
00376 (15)
00398 (14)
00071 (24)
00194 (17)
00246 (10)
00432 (21)
00386 (13)
00051 (2)
00325 (8)
00266 (9)
00344 (6)
00281 (12)
00209 (6)
00299 (12)
00364 (7)
TSS (n: 207)
GL - ABS - IUPUI (9)
G7 - Roswell Park (14)
UN - Boston Medical Center (1)
DZ - Mayo Clinic - Rochester (5)
EV - CHI-Penrose Colorado (3)
A4 - International Genomics Consortium (14)
IZ - ABS - Lahey Clinic (3)
UZ - University of California San Francisco (16)
DW - NCI Urologic Oncology Branch (10)
Y8 - Spectrum Health (8)
4A - Mary Bird Perkins Cancer Center - Our Lady of the Lake (3)
AL - Fox Chase (4)
B1 - University of Pittsburgh (7)
F9 - Asterand (5)
HE - Ontario Institute for Cancer Research (OICR) (9)
B9 - UNC (10)
KV - Hartford (3)
2Z - Moffitt Cancer Center (16)
MH - BLN - Baylor (4)
SX - Mayo Clinic Arizona (9)
BQ - MSKCC (31)
IA - Cleveland Clinic (5)
PJ - Gundersen Lutheran (3)
Q2 - University of Oklahoma HSC (1)
J7 - ILSBio (3)
P4 - MD Anderson Cancer Center (6)
B3 - Christiana Healthcare (2)
WN - University of Kansas (1)
2K - Greenville Health System (1)
V9 - Medical College of Georgia (1)

n


PC2
PC1
Points: BatchID
00071 (24)
00162 (31)
00194 (17)
00209 (6)
00246 (10)
00266 (9)
00281 (12)
00299 (12)
00325 (8)
00344 (6)
00364 (7)
00376 (15)
00386 (13)
00398 (14)
00432 (21)

o

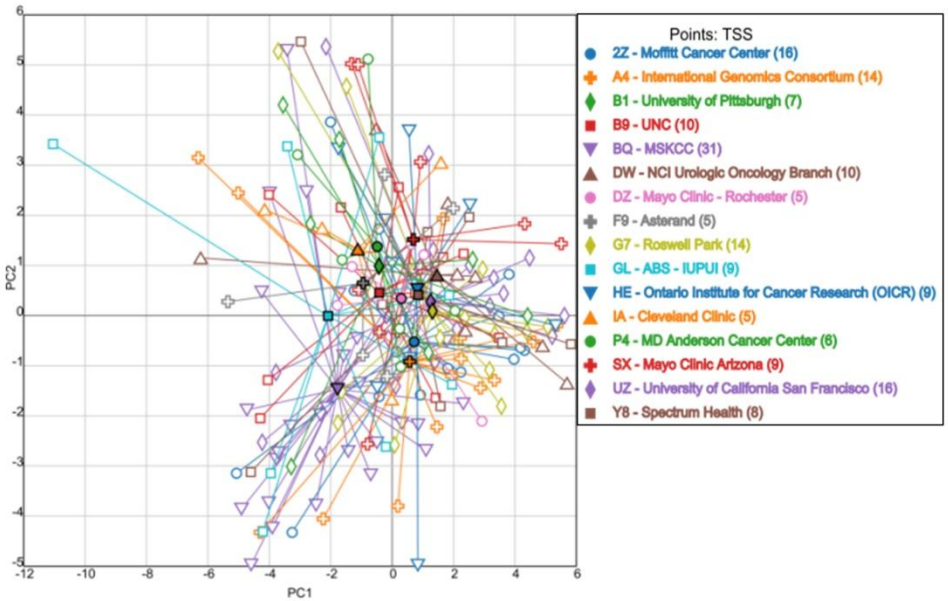

PC2
PC1
Points: TSS
2Z - Moffitt Cancer Center (16)
A4 - International Genomics Consortium (14)
B1 - University of Pittsburgh (7)
B9 - UNC (10)
BQ - MSKCC (31)
DW - NCI Urologic Oncology Branch (10)
DZ - Mayo Clinic - Rochester (5)
F9 - Asterand (5)
G7 - Roswell Park (14)
GL - ABS - IUPUI (9)
HE - Ontario Institute for Cancer Research (OICR) (9)
IA - Cleveland Clinic (5)
P4 - MD Anderson Cancer Center (6)
SX - Mayo Clinic Arizona (9)
UZ - University of California San Francisco (16)
Y8 - Spectrum Health (8)

Figure S21: Batch Effect Analysis for 161 PRCCs in the KIRP Cohort

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. The samples were clustered and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. **(a-c)** show clustering and PCA plots for miRNA-Seq data. miRNAs with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. **(a)** Hierarchical clustering for miRNA expression from miRNA-seq data. **(b)** PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to batch ID. **(c)** PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to TSS. Although section **(a)** shows a small batch effect by batch #71, the PCA plots in **(b)** and **(c)** don't show the same batch effect. Therefore, the observed batch effect is considered minor, not warranting any special batch effects correction.
(d-f) show clustering and PCA plots for the Infinium DNA methylation platform. **(d)** Hierarchical clustering plot for DNA methylation data. **(e)** PCA for DNA methylation, with samples connected by centroids according to batch ID. **(f)** PCA for DNA methylation, with samples connected by centroids according to TSS. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.
(g-i) show clustering and PCA plots for the RNA-Seq platform. **(g)** Hierarchical clustering plot for mRNA expression from RNA-seq data. **(h)** PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID. **(i)** First two principal components for RNA-seq, with samples connected by centroids according to TSS. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.
(j-l) show clustering and PCA plots for the copy number variations using the SNP 6 platform. **(j)** Hierarchical clustering plot for mRNA expression from RNA-seq data. **(k)** PCA: First two principal components for copy number, with samples connected by centroids according to batch ID. **(l)** First two principal components for copy number, with samples connected by centroids according to TSS. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.
(m-o) show clustering and PCA plots for RPPA data. **(m)** Hierarchical clustering plot for mRNA expression from RNA-seq data. **(n)** PCA: First two principal components for RPPA, with samples connected by centroids according to batch ID. **(o)** First two principal components for RPPA, with samples connected by centroids according to TSS. There seems to be a small batch effect by the tissue source site MSKCC, but not enough to warrant any special batch effects correction.
miRNA and RPPA data showed small batch effects, but not enough to justify any kind of batch effects correction. mRNA expression, DNA methylation, and CNV data didn't show any major batch effects.

List of Supplementary Tables

References

1. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Integrated detection and population genetic analysis of SNPs and copy number variation. Nat Genet. 40: 1166-1174 (2008).
2. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 40: 1253-1260 (2008).
3. The Cancer Genome Atlas Research Network, Integrated genomic analyses of Ovarian Carcinoma. Nature 474: 609-615 (2011).
4. Tabak B. and Beroukhim R. Manuscript in preparation.
5. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array based DNA copy number data. Biostatistics 5: 557-572 (2004).
6. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy number alteration in human cancers. Genome Bio. l12: R41 (2011).
7. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nature 47: 609-615 (2011).
8. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, Albert TJ, Burgess DL, Gibbs RA. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 12: R68 (2011).
9. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754-60 (2009).
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-9 (2009).
11. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D and Daly, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43: 491-498 (2011).
12. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43-9 (2013).
13. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505: 495-501 (2014).

14. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, Tsuji S, Donehower LA, Slagle BL, Nakamura H, Yamamoto S, Shinbrot E, Hama N, Lehmkuhl M, Hosoda F, Arai Y, Walker K, Dahdouli M, Gotoh K, Nagae G, Gingras MC, Muzny DM, Ojima H, Shimada K, Midorikawa Y, Goss JA, Cotton R, Hayashi A, Shibahara J, Ishikawa S, Guiteau J, Tanaka M, Urushidate T, Ohashi S, Okada N, Doddapaneni H, Wang M, Zhu Y, Dinh H, Okusaka T, Kokudo N, Kosuge T, Takayama T, Fukayama M, Gibbs RA, Wheeler DA, Aburatani H, Shibata T. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat Genet. 46:1267-73 (2014).

15. Campan M, Weisenberger DJ, Trinh B, and Laird PW. Methylight. Methods in molecular biology (Clifton, NJ), 507: 325 (2009).

16. Triche TJ Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41: e90 (2013).

17. The Cancer Genome Atlas. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489: 519–525 (2012).

18. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research 38: e178 (2010).

19. Li B and Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics 12: 323 (2011).

20. Wilkerson M and Waltman P. ConsensusClusterPlus. R package version 1160. (2013).

21. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 5: e1001138 (2011)

22. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature. 490: 61-70 (2012).

23. Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, Murphy D, Pierce H, Black J, Nelson D, Litzinger B, Hatton T, Maltbie L, Ainsworth M, Allen P, Rosewood L, Mitchell E, Smith B, Warner J, Groboske J, Telc H, Wilson D, Sanford B, Schmidt H, Haussler D, Maltbie D. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford). 2014 (2014).

24. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC *Bioinformatics* 11: 367 (2010).

25. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 22: 519-36 (2013).

26. Mullokandov G, Baccarini A, Ruzo A, Jayaprakash AD, Tung N, Israelow B, Evans MJ, Sachidanandam R, Brown BD. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat Methods 9: 840-6 (2012).

27. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature 505: 344-52 (2014).

28. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28: 1353-8 (2012).

29. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, Huang HD. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res 42: D78-85 (2014).

30. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4: 44-57 (2009).

31. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37: 1-13 (2009).

32. Smoot M, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432 (2011).

33. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 30: 413-21 (2012).

34. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative Genomics Viewer. Nature Biotechnology 29: 24–26 (2013).

35. Budczies J, Klauschen F, Sinn BV, Győrffy B, Schmitt WD, Darb-Esfahani S, Denkert C. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One 7: e51862 (2012).

36. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, Kornblau SM. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Molecular Cancer Therapeutics 5: 2512-2521 (2006).

37. Liang J, Shao SH, Xu ZX, Hennessy B, Ding Z, Larrea M, Kondo S, Dumont DJ, Gutterman JU, Walker CL, Slingerland JM, Mills GB. The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol 9: 218-224 (2007).

38. Hu J, He X, Baggerly KA, Coombes KR, Hennessy BT, Mills GB. Non-parametric quantification of protein lysate arrays. Bioinformatics 23: 1986-1994 (2007).

39. Hennessy BT, Lu Y, Poradosu E, Yu Q, Yu S, Hall H, Carey MS, Ravoori M, Gonzalez-Angulo AM, Birch R, Henderson IC, Kundra V, Mills GB : Pharmacodynamic Markers of Perifosine Efficacy. Clinical Cancer Research 13: 7421-7431 (2007).

40. Coombes K, Neeley S, Joy C, Hu J, Baggerly K, Roebuck P. SuperCurve: SuperCurve Package. R package version 1.4.1. (2011).

41. Gonzalez-Angulo AM, Hennessy BT, Meric-Bernstam F, Sahin A, Liu W, Ju Z, Carey MS, Myhre S, Speers C, Deng L, Broaddus R, Lluch A, Aparicio S, Brown P, Pusztai L, Symmans WF, Alsner J, Overgaard J, Borresen-Dale AL, Hortobagyi GN, Coombes KR, Mills GB. Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics 8: 11 (2011).

42. Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, Davies MA, Liu W, Coombes K, Meric-Bernstam F, Bedrosian I, McGahren M, Agarwal R, Zhang F, Overgaard J, Alsner J, Neve RM, Kuo WL, Gray JW, Borresen-Dale AL, Mills GB. A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected. Human Breast Cancers. Clin Proteomics 6: 129-151 (2010).

43. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13: 1977-2000 (2002)

44. Abazeed ME, Adams DJ, Hurov KE, Tamayo P, Creighton CJ, Sonkin D, Giacomelli AO, Du C, Fries DF, Wong KK, Mesirov JP, Loeffler JS, Schreiber SL, Hammerman PS, Meyerson M. Integrative radiogenomic profiling of squamous cell lung cancer. Cancer Res 73: 6289-6298 (2013).