doi:10.1038/nature12222

Contents

LIST OF SUPPLEMENTARY FIGURES

Batch effects

Copy number studies

Mutation analysis

Figure S24: Gap filling coverage improvement for VHL, SETD2 and KDM5C, and all 95 Exons taken together.

Figure S25: The allele fractions of the single nucleotide polymorphisms (SNPs) found jointly by both the Broad Institute (BI) and the Human Genome Sequencing Center (HGSC).

Figure S26: Validation rates (by combination of 454 and Ion Torrent), grouped by which center(s) called the mutation, the type of variant (SNP or INDEL), and sequencing platform.

Figure S27: Kaplan-Meier plot for BAP1 mutation versus wild-type.

Integrative analysis of mutation and copy number

Figure S28: Rarely altered oncogenes are associated with poor outcome.

Figure S29: Rarely altered tumor suppressors are associated with poor outcome.

Figure S30: Rarely altered tumor suppressors and oncogenes are associated with poor outcome.

RNA fusions

Figure S31: RT-PCR results for FAM172A-FHIT fusion validations for sample TCGA-B2-4101.

Figure S32: RT-PCR results for TFE3 fusion validations for sample TCGA-AK-3456.

Figure S33: RT-PCR results for fusion validations for sample TCGA-A3-3313.

Figure S34: RT-PCR results for fusion validations for samples TCGA-AK-3445 (SOGA2-LRRC41) and TCGA-B0-5095 (GORASP2-WIPF1).

Figure S35: FISH and IHC confirming TFE3 gene fusions.

Expression studies

Figure S36: CNMF clustering of 1,500 variably expressed genes and 417 TCGA ccRCC samples.

Figure S37: Distribution of stage and grade among the mRNA-based tumor subtypes.

Figure S38: Integrated visualization of gene set activation scores, genomic alterations significantly associated with mRNA subtypes, survival and tumor purity estimates.

Figure S39: Among the RNA-based subtypes, there is differential enrichment for genetic or genomic alternations in the PI3K pathway

miRNA studies

Figure S40: Sample groups identified by NMF consensus clustering of miRNA-seq abundance profiles.

Figure S41: Summary properties of the four miRNA-based sample groups.

Figure S42: Relationships between copy number alterations and miRNA abundance in the four miRNA-based sample groups.

Figure S43: DNA methylation and miRNA abundance for the most discriminatory miRNAs for each sample group.

Figure S44: Kaplan-Meier and box plots for VHL-associated genes and miRs.

HotNet

Figure S45: The subnetworks with 3 or more nodes identified by HotNet.

Paradigm

Figure S46: Consensus clustering reveals five clusters for PARADIGM-based subtypes.

Figure S47: PARDIGM integrated pathway levels and survival analysis.

Figure S48: PARADIGM analysis reveals multiple transcription factor hubs driving expression.

Figure S49: Differential Pathway Signature Correlation analysis reveals connections between genomic perturbations and clinical outcomes.

Figure S50: Genomic perturbations in kidney cancers are significantly associated with downstream transcriptional changes through known and novel pathway circuitry.

Figure S51: TieDIE identifies interlinking pathways connecting chromatin-related genes to downstream transcriptional hubs.

Figure S52: Chromatin-related TieDIE solution, alternate view showing the original, non-discriminant data and inference levels.

Integrative analysis of mutation and expression

Figure S53: Overlap of differentially expressed gene sets.

Mutual Exclusivity Modules (MeMO)

Figure S54: Candidate genes on 5q35.3 amplicon associated with mTOR signaling.

Figure S55: mTOR signaling pathway components display mutually exclusive pattern of alterations.

Figure S56: Frequent over-expression of EGFR, which correlates with increased phosphorylation of the receptor.

Survival correlates

Figure S57: Validation of mRNA prognostic signature in Zhao et al. mRNA dataset.

Figure S58: Kaplan-Meier plots for AMPK and ACC protein expression.

Figure S59: Kaplan-Meier and scatter plots for MIR21 and GRB10 RNA expression and DNA promoter methylation.

Figure S60: Heat maps of differential patterns for key metabolic-associated features.

Figure S61: Survival correlations for proteins involved in the PI3K pathway.

Figure S62: Survival correlations for MMP and TIMP genes.

LIST OF SUPPLEMENTARY TABLES

Integrative analysis of mutation and expression

Table S18: Chromatin-related genes containing mutations in ccRCC TCGA dataset.

Table S19: Number of differentially expressed genes when comparing tumors with mutations in the specified gene to tumors with no evidence of mutations in the specified gene.

Table S20: Most differentially expressed genes when comparing tumors with mutations in the specified gene to tumors with no evidence of mutations in the specified gene.

Table S21: Enriched classes in sets of differentially expressed genes.

Mutual Exclusivity Modules (MeMO)

Table S22: Gene modules displaying significant mutual exclusivity.

Survival correlates

Table S23: Multivariate Cox analysis results for prognostic molecular signatures.

Table S24: Table of methylation:mRNA survival anti-correlates.

LIST OF SUPPLEMENTAL DATA FILES

Data File S1: Sample list for Core and Extended sets.
Data File S2: Clinical dataset used in the study.
Data File S3: Significantly mutated genes.
Data File S4: mRNA fusion results.
Data File S5: mRNA subtype analysis results.
Data File S6: Analysis of known oncogenes and tumor suppressors in ccRCC.
Data File S7: Pathmark pathway enrichment results.
Data File S8: Molecular prognostic signatures.
Data File S9: mRNA and miRNA-based subtyping assignments by sample.
Data File S10: Complete Mutation Annotation File (MAF) for exome sequencing.

I. BIOSPECIMENS

Workgroup leaders: W. Kim Rathmell (rathmell@med.unc.edu) and Victor Reuter (reuterv@MSKCC.ORG)

Contributors: Chad J. Creighton, A. Ari Hakimi, James Hsieu, W. Marston Linehan, Kenna Shaw, Candace Shelton, Troy Shelton, Scott Morris, Robert Penny

Sample Acquisition and Data Freeze. Tumor samples were accrued as part of the Cancer Genome Atlas (TCGA) network. Briefly, flash-frozen samples of tumor resections were shipped to a centralized processing center (Biospecimen Core Resource, BCR) for additional pathological review and nucleic acids extraction. Aliquots of DNA and RNA were shipped to individual sites for all subsequent testing. Normal DNA samples were provided as processed DNA, or adjacent uninvolved normal kidney or blood aliquots from each patient, the latter simultaneously collected and shipped for DNA extraction. .

Biospecimens were collected from newly diagnosed patients with renal clear cell carcinoma undergoing surgical resection and had received no prior treatment for their disease, including chemotherapy or radiotherapy. All cases were collected regardless of surgical stage or histologic grade. Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Each frozen tumor specimen had a companion normal tissue specimen which could be blood components, adjacent normal tissue taken from greater than 2cm from the tumor, or previously extracted germline DNA from blood or nonmalignant tissue. Each frozen tumor specimen submitted to the BCR weighed at least 30 mg and was typically under 200 mg. Specimens were shipped overnight from one of 13 tissue source sites using a cryoport that maintained an average temperature of less then -180°C. The tissue source sites contributing biospecimens included: Catholic Health Initiative - Penrose St. Francis Health Services, Catholic Health Initiative - St. Joseph's Medical Center Cancer Institute, Christiana Care Health Services, Inc., Cureline, Inc., Fox Chase Cancer Center, Harvard University, International Genomics Consortium, Mayo Clinic, MD Anderson, MSKCC, National Cancer Institute Urologic Oncology Branch, University of North Carolina and University of Pittsburgh.

Clinical data definitions. Complete clinical data elements are compiled for all specimens included in the Freeze List in Table S1 and reflect current data as of April 13, 2012. Clinical/demographic data include: Sample code, primary site (Kidney for all specimens), gender, age at diagnosis, race, ethnicity, and year of tumor collection. Prior tumor and prior therapy are indicated (y/n). The surgical approach is documented, along with indication of neoadjuvant therapy if it was applied prior to surgical resection of the tumor. Samples of questionable authenticity as clear cell RCC tumors due to unexpected molecular analysis results were evaluated by secondary pathologic review as described below.

Tumor information recorded complete pathologic information regarding the tumor. All specimens included in this analysis are coded as kidney clear cell renal carcinoma. The table records: laterality (right/left), Fuhrman nuclear grade, maximum tumor dimension (cm), T stage, lymph node involvement (based on pathologic staging), and M stage (intended to be indicative of a review of clinical evidence for metastatic disease, but was often provided from available pathologic information only, so should be interpreted with caution). A compiled tumor stage using standard AJCC staging criteria using the Tumor Node Metastasis universal schema was reported.

Clinical status of patients at the point of enrollment, and as available at last follow-up was recorded. Sites were asked to indicate if patients following surgical resection were tumor free, or with tumor. We

also recorded vital status (living/deceased) at the time of enrollment. Follow-up data was requested for subjects out to a minimum of two years from the time of sample collection. The patient tumor status (tumor free/with tumor) was again recorded, along with vital status (living/deceased) from the most recent follow up data form completion at the time of data collection. Time to recurrence is recorded as the number of days to a new tumor event. We also recorded the days to last contact at the point of enrollment or most recent follow up. Finally, the days from diagnosis (sample collection) to death was recorded at both enrollment and in the most recent follow up forms. These data provide the information to explore survival-based outcomes and median follow-up for patients included in this study. Cause of death, from cancer or other causes, is not recorded.

Prognostic criteria for patients with non-metastatic disease are incorporated into a variety of algorithms to estimate risk for disease recurrence and survival. Common prognostic criteria of the UCLA Integrated Staging system include tumor T stage, Fuhrman nuclear grade, and clinical performance status. Pathologic and clinical evaluation of criteria associated with disease prognosis include tumor stage as detailed above and tumor grade, using the Fuhrman grading scale, G1-G4. Here we recorded the number of positive lymph nodes, data on performance status was recorded as the Karnofsky performance status or the Eastern Cooperative Group (ECOG) clinical performance status at the time of diagnosis, and laboratory prognostic criteria: lactate dehydrogenase (LDH), erythrocyte sedimentation rate (ESR), serum calcium, white cell count, hemoglobin, and platelet count.

Expanded prognostic criteria are utilized for estimating the duration of survival in the presence of metastatic disease have been extensively validated [1], and recently updated to reflect survival in the era of targeted therapy [2]. Standard negative prognostic criteria include presence of anemia, time from diagnosis to treatment for metastatic disease, performance status, serum calcium, and the variable inclusion of elevations in LDH, platelet count, white blood cell count, or erythrocyte sedimentation rate (ESR). Available data on this chemistry values as well as hemoglobin status, platelet count and white blood cell count at the time of nephrectomy are reported as normal, elevated, or low.

Limitations of this data are the large number of specimens for which the status of clinical metastatic disease at the time of nephrectomy is not known, and the uncertainty regarding the interval development of metastatic disease. In addition, complete availability of prognostic criteria for any of the commonly used algorithms is not present for the vast majority of individual specimens. Finally, the dataset is biased for the inclusion of biospecimens from patients with larger primary tumors (to have sufficient material available for tissue analysis) and patients with no evidence for or limited amounts of metastatic disease, who were by definition better candidates for surgical management of their disease. The samples, taken from primary tumor specimens, were reflective of patients from all disease stages who were fit for either definitive or cytoreductive nephrectomy. The extension of these findings to the genetics inherent to metastatic disease, or disease that presents so extensively that nephrectomy is not feasible, will require future studies for comparison to this primary dataset. In particular, future work detailing the genomic landscape of metastatic lesions, and the relationship to clinical responsiveness will be essential to place all of these molecular and genetic events into the context of patient outcomes.

Clinical parameters and Demographics. The following data elements were compiled and are provided in Data File S2. All patients included in this study were confirmed to display clear cell histology renal cell carcinoma. The complete set of samples included in the analysis represent 446 nephrectomy specimens from patients with histologically confirmed clear cell renal cell carcinoma collected between 1998-2010 (median year 2006). This complete freeze list includes all tumors for which at least one platform of data is available and quality verified. Demographics on this group are detailed in Table S1, but generally represent a median age at diagnosis of 60.9 years. The patients

included in this dataset were 65% male, 35% females, and represented 93% Caucasian, 3.4% African American/Black, and 1.6% Asian. Although this distribution is not exactly consistent with the US demographic, these ratios are consistent with the referral pattern of the several major centers which supplied the majority of samples. Considerable representation was also provided by European sites, additionally shifting the balance of tumor specimens in favor of Caucasian donors. The tumors represented a distribution of tumor stage and grade typical of the disease. Specifically, 372 tumors represent localized disease (range stage I-stage III), with 74 occurring with synchronous metastatic disease (stage IV). Clinical data regarding outcomes is current as of April 13, 2012.

A core list of 372 biospecimens reflects the subset of samples for which data is available across all platforms. Demographics on this group are detailed in Table S1 but are generally representative of the complete group. This group demonstrates a median age at diagnosis of 61 years. The patients included in this dataset were 65% male, 35% females, and represented 93% caucasion, 3.2% black, and 1.9% Asian, all consistent with the US representation of clear cell renal cell carcinoma. The tumors represented a distribution of tumor stage and grade typical of the disease. Specifically, 366 tumors represent localized disease, with 6 occurring with synchronous metastatic disease. No significant differences were observed between the core and extended biospecimen lists.

Verification of clear cell histology diagnosis. Tumors were selected meeting the criteria of clear cell histology lacking multifocality. Each tumor and adjacent normal tissue specimen were embedded in optimal cutting temperature (OCT) medium and histologic sections were obtained for review. Each H&E stained imaged taken from the frozen section was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with renal clear cell carcinoma and the adjacent normal specimen contained no tumor cells. The sections were required to contain an average of 60% tumor cell nuclei with less than 20% necrosis for inclusion in the study per TCGA protocol requirements. Independent pathologic review confirmed the histology of all of the tumors included in the analysis. A subset of the tumors which displayed outlier molecular characteristics (no 3p loss, or a chromophobe RCC-like genotype) underwent a second round of expert pathologic review to confirm histologic subtype based on a digital whole slide image of corresponding formalin fixed, paraffin embedded sections. Cases were excluded from the study by consensus, when at least 4 of 6 pathologists agreed the tumor was not a clear cell RCC.

Biospecimen processing. Our study sampled a single site of the primary tumor. All DNA and RNA were isolated from a co-isolation protocol where nucleic acids are isolated from the same piece of tissue, allowing for direct comparisons across all platforms. Tumor samples were generally from surgical resections, due to the requirement to process at least a 30mg portion of tissue. RNA and DNA were initially extracted from tumor specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The isolation methodology for each sample was noted in the Biospecimen XML uploaded to the DCC. A portion of the flow-through from the DNA column was processed according to the AllPrep RNA extraction instructions to produce RNA analytes >200 nt (designated ‘Allprep RNA Extraction’ in the biospecimen XML), while the other portion was precipitated after TRIzol separation (designated ‘Total RNA’ in the XML). Early in the project, the RNA extraction protocol was changed to utilize the mirVana miRNA Isolation Kit (Ambion). This protocol modification generated RNA preparations that included RNA <200 nt [designated ‘mirVana (Allprep DNA) RNA’] suitable for miRNA analysis. Following the protocol change, the BCR re-extracted all renal clear cell carcinoma cases where sufficient tissue remained, resulting in 94% of the cases within the data sent being extracted via the mirVana Isolation Kit. DNA was extracted from normal tissue using either the QiaAmp blood midi kit (Qiagen) or the QiaAmp tissue mini kit (Qiagen). Each specimen was quantified by measuring Abs260 with a UV spectrophotometer. DNA specimens were resolved by agarose gel electrophoresis to determine the range for fragment sizes. The AmpFISTR Identifiler (Applied Biosystems) or Sequenom SNP panel procedure was utilized to verify tumor DNA and germline DNA were derived from the same

patient. One μg each of tumor and normal DNA was sent to Qiagen for REPLI-g whole genome amplification using a 100 μg reaction scale. Only those specimens yielding a minimum of 6.9 g of tumor DNA, 4.9 μg of germline blood DNA, or 6.9 μg solid tissue normal DNA, and 5.15 μg of RNA, were included in this study. RNA was analyzed via the RNA6000 assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study. In addition, RNA extracted from 55 normal tissue biospecimens and exceeded a RIN >7.0 was also included in the study.

References

1. Motzer, RJ, M Mazumdar, J Bacik, W Berg, A Amsterdam, and J Ferrara, *Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma.* J Clin Oncol, 1999. **17**(8): p. 2530-40.
2. Heng, DY, W Xie, MM Regan, MA Warren, AR Golshayan, C Sahi, BJ Eigl, JD Ruether, T Cheng, S North, P Venner, JJ Knox, KN Chi, C Kollmannsberger, DF McDermott, WK Oh, MB Atkins, RM Bukowski, BI Rini, and TK Choueiri, *Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted agents: results from a large, multicenter study.* J Clin Oncol, 2009. **27**(34): p. 5794-9.

Table S1

Table S1. Clinical data summary of patients in the TCGA ccRCC dataset

	Core dataset	Extended dataset	Freeze List
Sample (n)	372	74	446
Age at dx (years)	61 (range 26-90)	60.4 (range 35-90)	60.9 (range 26-90)
Year of collection	2005 (1998-2010)	2005 (2001-2010)	2006 (1998-2010)
Gender			
Male	241 (65%)	49 (66%)	290 (65%)
Female	131 (35%)	25 (34%)	156 (35%)
Race			
White	348 (93%)	69 (93%)	417 (93%)
Black	12 (3.2%)	3 (4%)	15 (3.4%)
Asian	7 (1.9%)	0	7 (1.6%)
Not Available	5 (1.3%)	2 (2.7%)	7 (1.6%)
Ethnicity			
Hispanic	19 (5.1%)	4 (5.4%)	23 (5.1%)
Non-Hispanic	242 (65%)	40 (54%)	282 (63%)
Not Available	111 (29.8%)	30 (40.5%)	141 (32%)
Prior Tumor			
Yes	48 (12.9%)	15 (20%)	63 (14.1%)
Laterality			
Right	189 (50.8%)	47 (63.5%)	236 (53%)
Left	183 (49.2%)	27 (36.5%)	210 (47%)
Grade			
G1	6 (1.6%)	2 (2.7%)	7 (1.6%)
G2	150 (40.3%)	38 (51.3%)	177 (40.0%)
G3	158 (42.4%)	23 (31.1%)	179 (40.1%)
G4	57 (15.3%)	11 (14.9%)	68 (15.2%)
Tumor Size (cm)	6.54	6.49	6.53
Staging (TNM)			
T1	182 (49%)	37 (50%)	219 (49.1%)
T2	39 (10.5%)	13 (17.6%)	52 (11.7%)
T3	145 (39%)	24 (32.4%)	169 (38.0%)
T4	6 (1.6%)	0	6 (1.3%)
Nodes			
Node - (N0)	184 (49.5%)	23 (31%)	207 (46.4%)
Node +(N1)	11 (3%)	2 (2.7%)	13 (2.9%)
Node unknown (NX)	177 (47.6%)	49 (66.2%)	226 (50.7%)
Metastasis			
Mets - (M0)	317 (85.2%)	58 (78.3%)	375 (84.1%)
Mets + (M1)	55 (14.8%)	16 (21.6%)	71 (15.9%)
Clinical stage			
Stage I	180 (48.4%)	35 (47.3%)	215 (48.2%)
Stage II	31 (8.3%)	11 (14.9%)	42 (9.4%)
Stage III	103 (27.3%)	12 (16.2%)	115 (25.8%)
Stage IV	58 (15.6%)	16 (21.6%)	74 (16.6%)

Table S2

Table S2: Summary of data types (From Freeze 1.4.1)

Data Type	Platforms	Cases	Data access
Whole exome DNA sequence	Illumina and SOLiD	417	Controlled
Whole genome DNA sequence	Illumina	22	Controlled
DNA copy number/genotype	Affymetrix SNP 6	441	Controlled
mRNA expression	Illumina	417	Controlled - BAM files
			Open - expression files
miRNA expression	Illumina	414	Controlled - BAM files
			Open - expression files
CpG DNA methylation	Illumina 27K	192	Open
	Illumina 450K	253	Open
Protein expression	RPPA	411	Open
	All Platforms	372	
	At least one platform	74	
Total Cases		446	

II. BATCH EFFECTS

Workgroup leaders: Rehan Akbani (RAkbani@mdanderson.org) and Nianxiang Zhang (nzhang@mdanderson.org)

Contributors: Anna K. Unruh, Tod D. Casasent, Chris Wakefield, John N. Weinstein

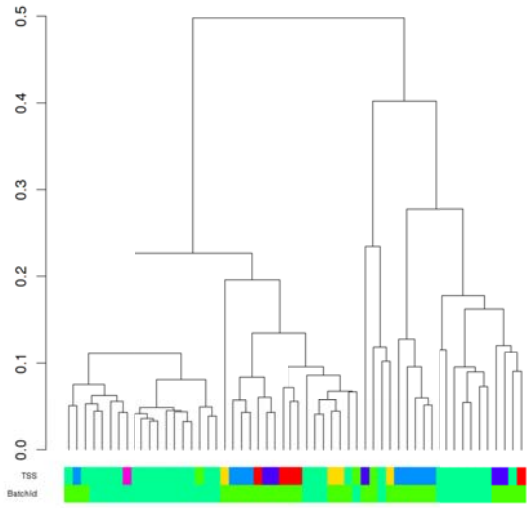
We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the TCGA Kidney renal clear cell carcinoma (KIRC) data sets. Five different data sets were analyzed: mRNA expression - genes (Agilent G4502A), mRNA sequencing (Illumina HighSeq), miRNA expression sequencing (RNA-seq Illumina GA and RNA-seq Illumina HighSeq), DNA methylation (Infinium HM27 and HM450 microarrays), and SNPs (GW SNP 6). All of the data sets were at TCGA level 3, since that's the level on 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).

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 the five data sets follow.

Figure S1

Figure S2

Figure S3



Legends

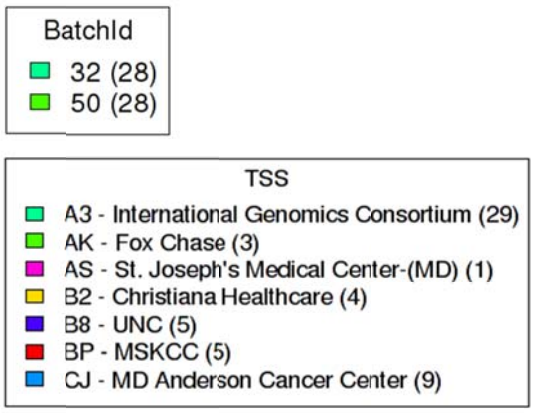


Figure S1. Hierarchical clustering plot for mRNA expression (Agilent microarray)

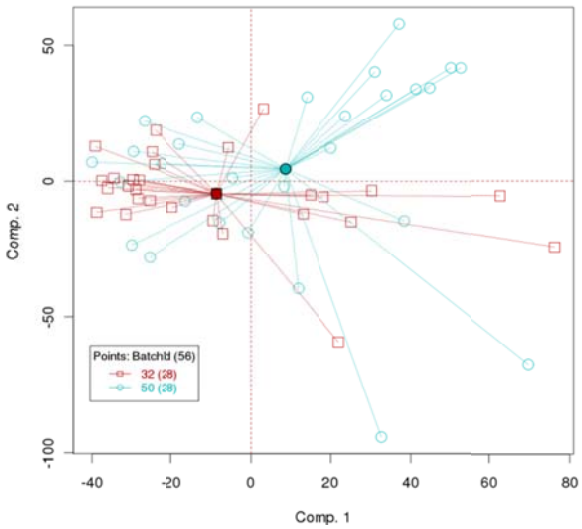


Figure S2. PCA: First two principal components for mRNA expression (Agilent microarray), with samples connected by centroids according to batch ID.

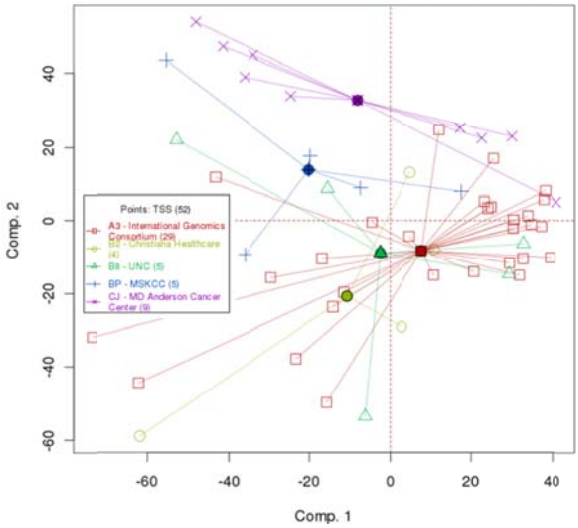


Figure S3. PCA: First two principal components for mRNA expression (Agilent microarray), with samples connected by centroids according to TSS.

mRNA expression – genes (Agilent G4502A microarray). Figures S1-S3 show clustering and PCA plots for the Agilent G4502A mRNA expression platform. The results do show some batch effects by

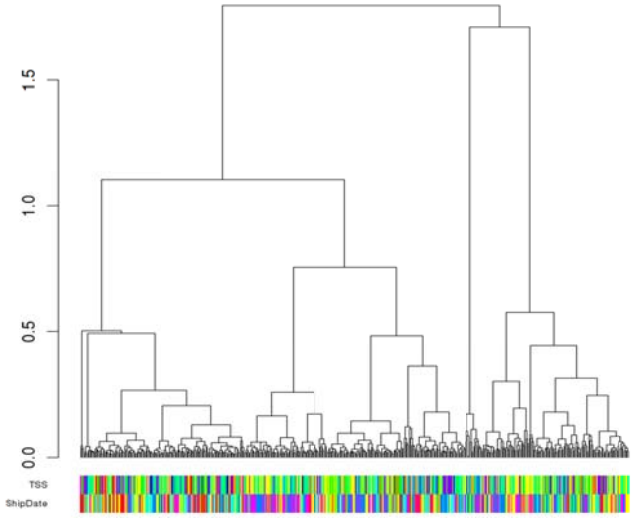
batch ID and tissue source site. However, gene expression data was not used in the analysis done in this paper. mRNA sequencing data was used instead. The results for batch effects in mRNA sequencing follow, and are much better. The differences in results hint at the superiority of sequencing over microarray technologies in terms of batch effects, although several other factors may also play a crucial role in mitigating those effects.

mRNA sequencing (RNA-seq Illumina HighSeq). Figures S4-S6 show the clustering and PCA plots for the mRNA sequencing (Illumina HighSeq) platform. The plots show that the batches are well mixed and none of them stand out from the rest. The PCA plots show some outliers, however, they don't group by batch ID or TSS, so it's unlikely for those outliers to be batch based. They may be due to biological differences.

Figure S4

Figure S5

Figure S6



Legends

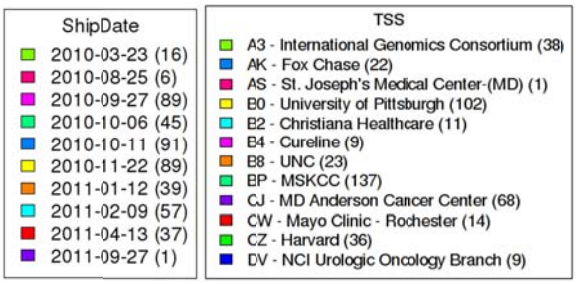


Figure S4. Hierarchical clustering plot for mRNA sequencing (RNASeq Illumina HighSeq)

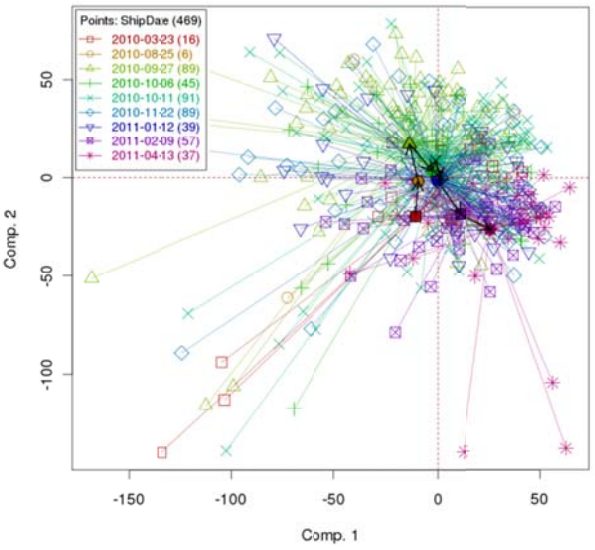


Figure S5. PCA: First two principal components for mRNA sequencing (RNASeq Illumina HighSeq), with samples connected by centroids according to batch ID.

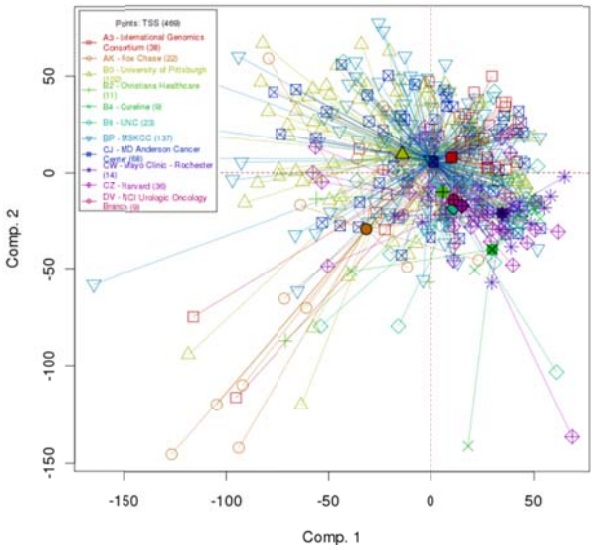


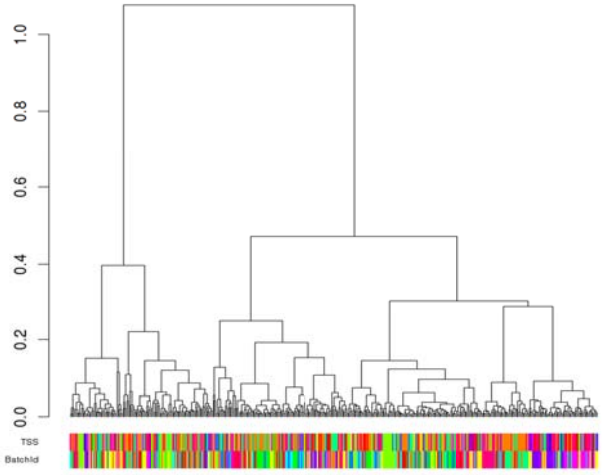
Figure S6. PCA: First two principal components for mRNA sequencing (RNASeq Illumina HighSeq), with samples connected by centroids according to TSS.

miRNA sequencing (RNA-seq Illumina). Figures S7-S9 show the clustering and PCA plots for the miRNA expression sequencing platform (RNA-seq Illumina). Both, Illumina GA and Illumina HighSeq platforms were combined together to generate the figures. The figures show no significant batch effects by either batch ID or TSS.

Figure S7

Figure S8

Figure S9



Legends

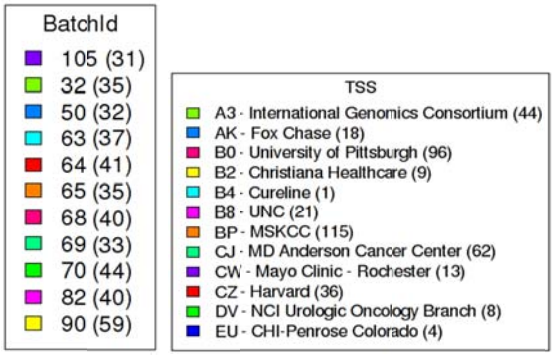


Figure S7. Hierarchical clustering for miRNA sequencing data.

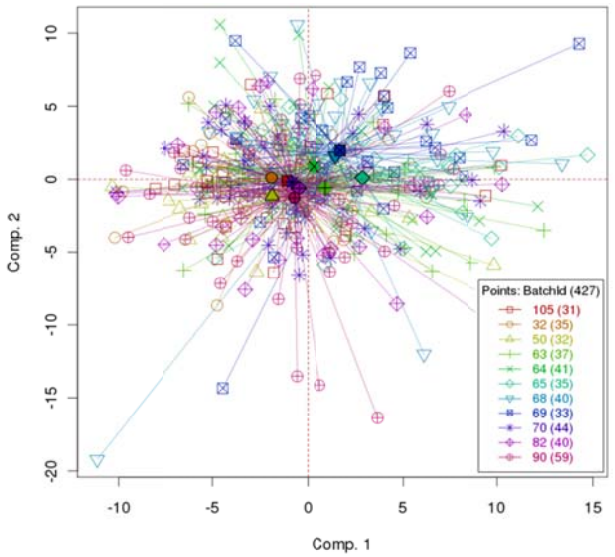


Figure S8. PCA: First two principal components for miRNA sequencing data, with samples connected by centroids according to batch ID.

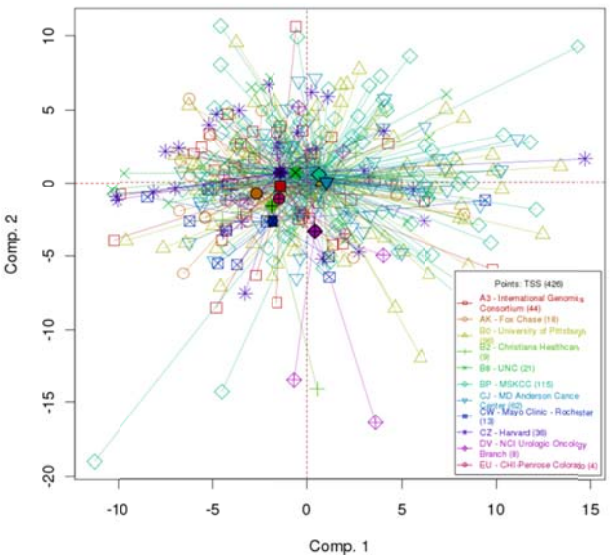
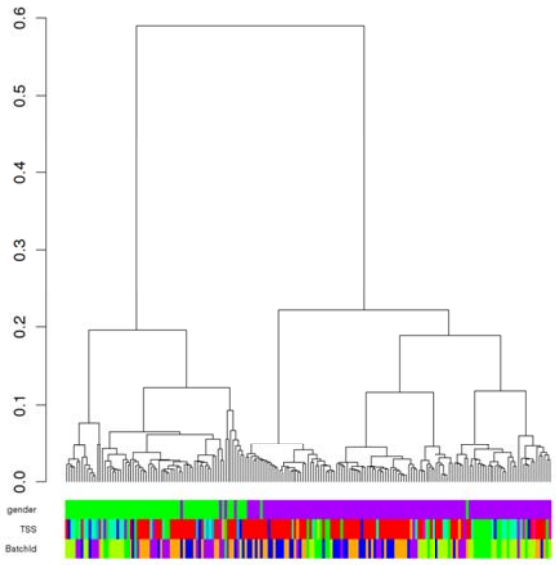


Figure S9. PCA: First two principal components for miRNA sequencing data, with samples connected by centroids according to TSS.

DNA methylation (Infinium HM27 microarray). Figures S10-S13 show clustering and PCA plots for the DNA methylation (Infinium HM27 microarray) using all probes. We see a clear dichotomy which is not related to batch or TSS but does appear to be related to gender. We suspected that that was due to the inclusion of sex chromosomes in the analysis, so we repeated the analysis after removing probes on sex chromosomes. The results are shown in Figures S14-S17.

Figure S10



Legends

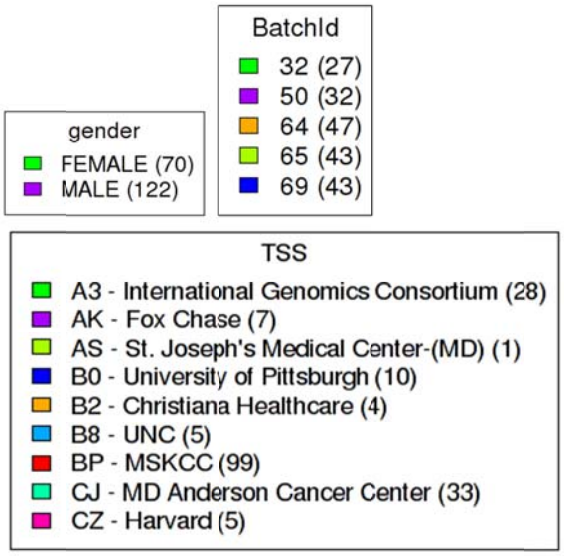


Figure S10. Hierarchical clustering for DNA methylation (Infinium HM27 microarray) data, using all the probes.

Figure S11

Figure S12

Figure S13

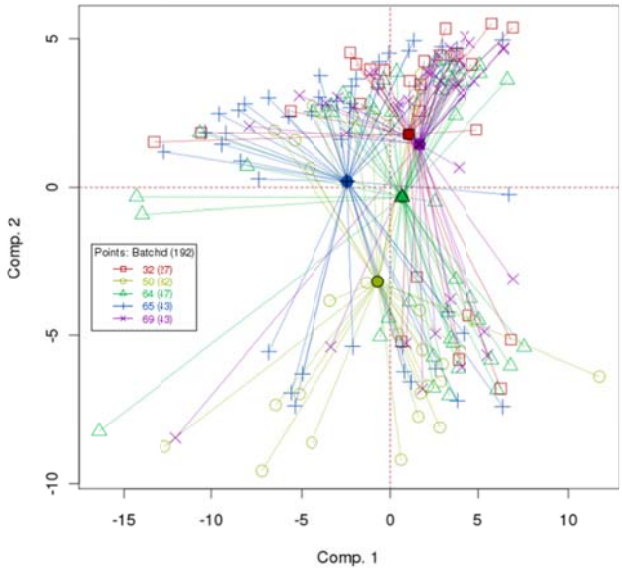


Figure S11. PCA: First two principal components for DNA methylation (Infinium HM27 microarray), with samples connected by centroids according to batch ID.

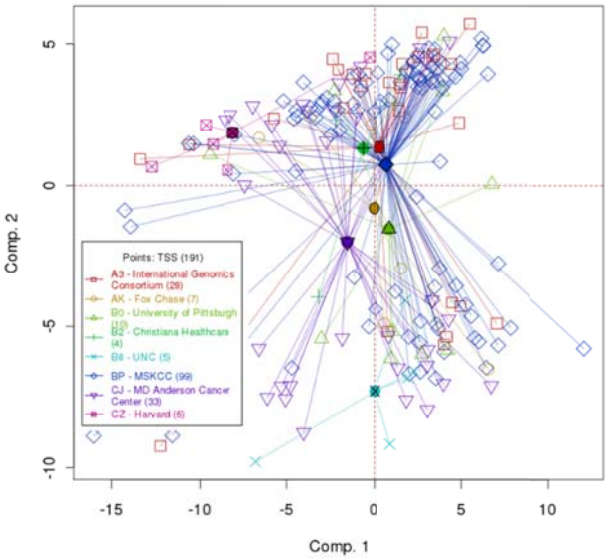


Figure S12. PCA: First two principal components for DNA methylation (Infinium HM27 microarray), with samples connected by centroids according to TSS.

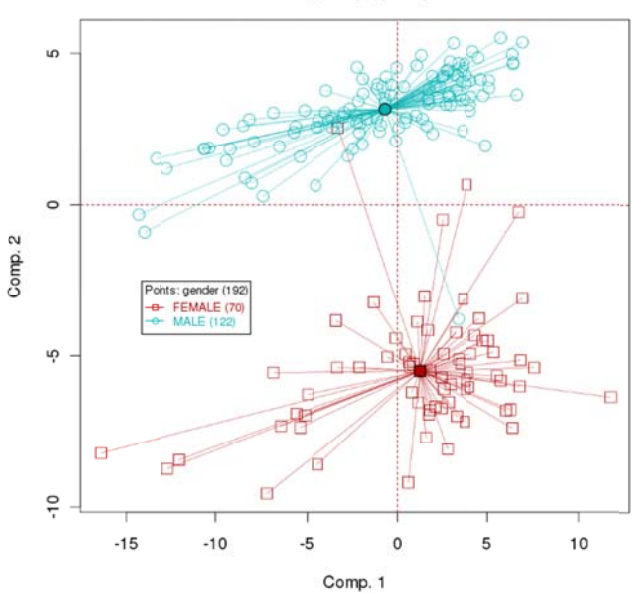
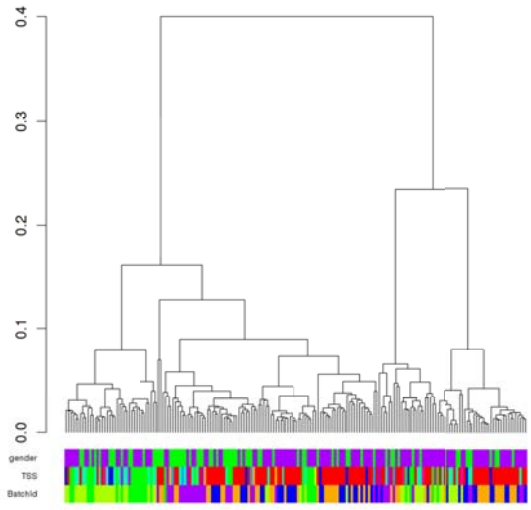


Figure S13. PCA: First two principal components for DNA methylation (Infinium HM27 microarray), with samples connected by centroids according to gender.

Figures S14-S17 show clustering and PCA plots for the DNA methylation (Infinium HM27 microarray) after removing probes on the X and Y chromosomes. The dichotomy disappears, indicating that it was indeed due to the inclusion of sex chromosomes in the analysis. Furthermore, no serious batch effects are seen by batch ID or TSS, although some minor effects by TSS may exist.

Figure S14



Legends

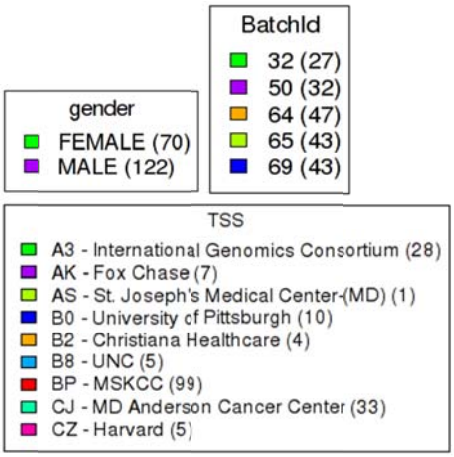


Figure S14. Hierarchical clustering for mRNA expression from DNA methylation (Infinium HM27 microarray) data with sex chromosomes removed.

Figure S15

Figure S16

Figure S17

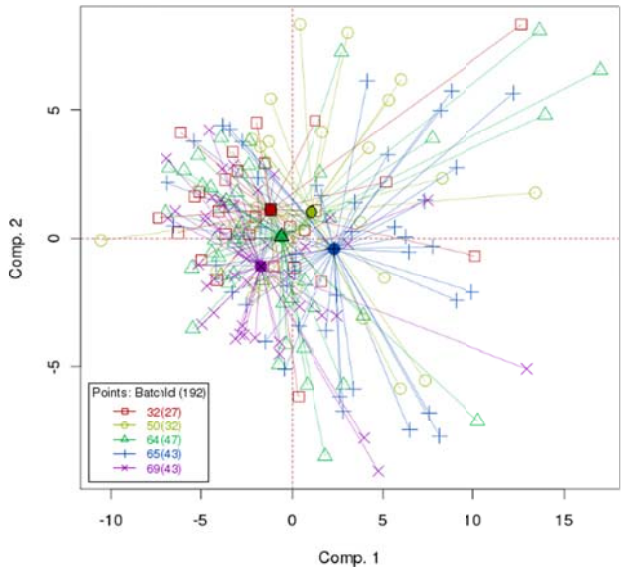


Figure S15. PCA: First two principal components for DNA methylation (Infinium HM27 microarray) without sex chromosome probes, with samples connected by centroids according to batch ID.

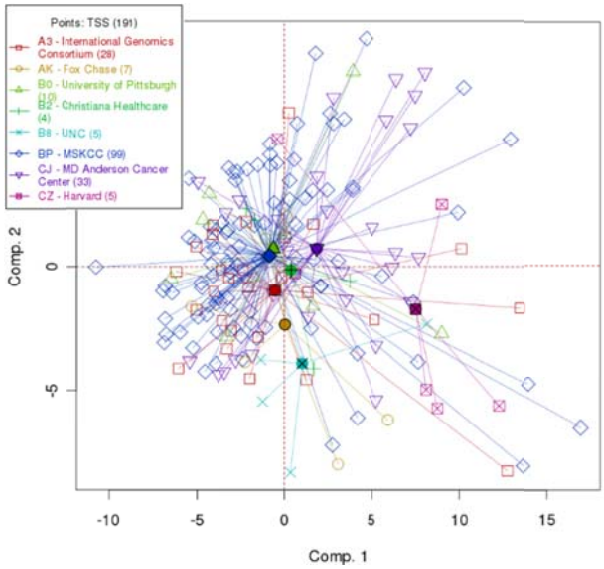


Figure S16. PCA: First two principal components for DNA methylation (Infinium HM27 microarray) without sex chromosome probes, with samples connected by centroids according to TSS.

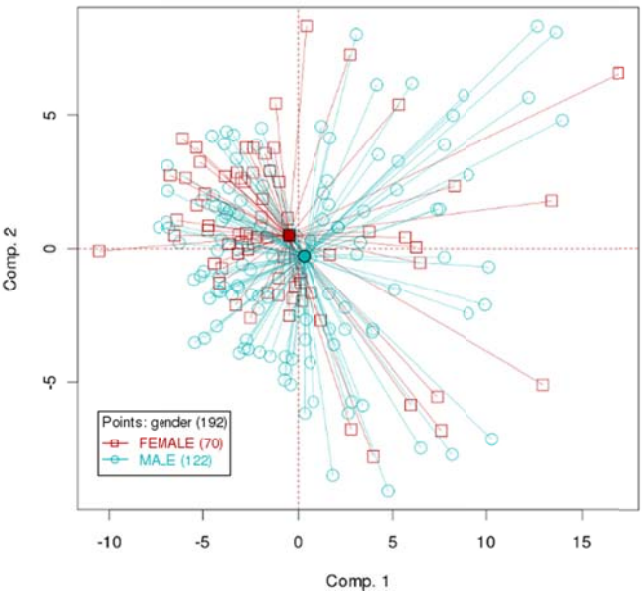


Figure S17. PCA: First two principal components for DNA methylation (Infinium HM27 microarray) without sex chromosome probes, with samples connected by centroids according to gender.

SNPs (GW SNP 6). Figure S18-S20 show clustering and PCA plots for the SNP platform. At level 3, the TCGA SNP data resemble copy number data when we use chromosomal segment counts (rather than actual SNP probes). We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in the figures. None of the batches or TSSs stand out from the rest.

Figure S18

Figure S19

Figure S20

Legends

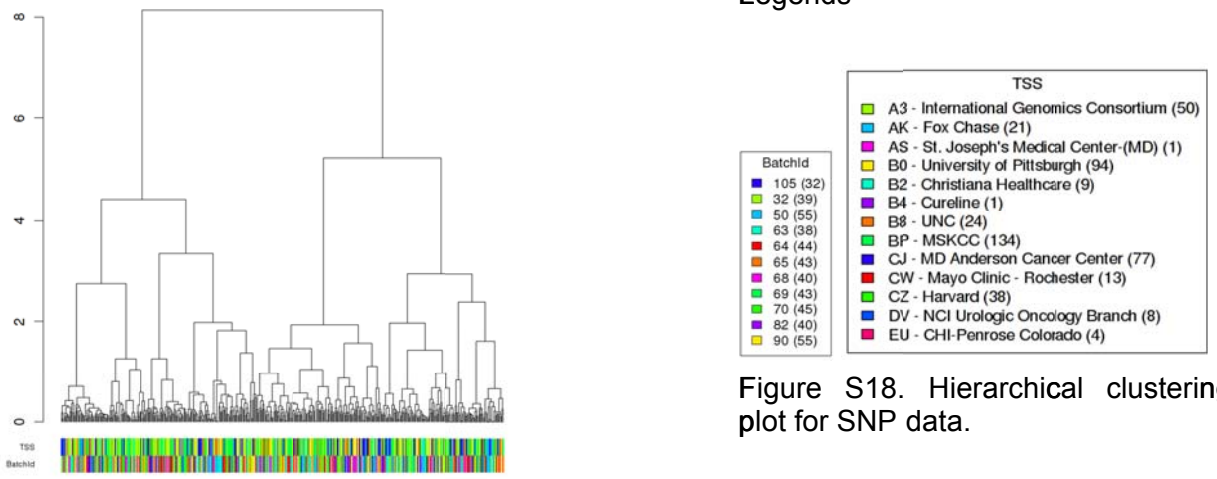


Figure S18. Hierarchical clustering plot for SNP data.

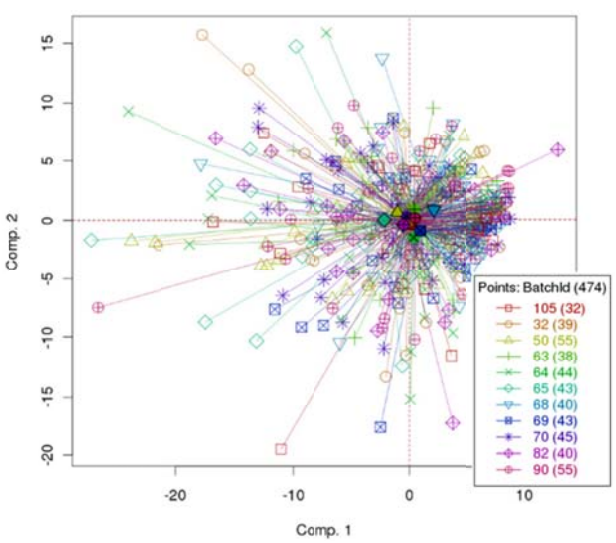


Figure S19. PCA for SNPs, with samples connected by centroids according to batch ID.

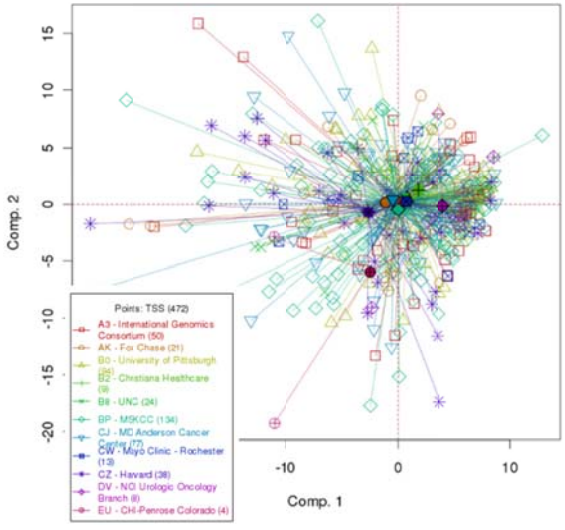


Figure S20. PCA for SNPs, with samples connected by centroids according to TSS.

Conclusions. We analyzed five different data sets for batch effects using batch ID and Tissue Source Site (TSS) as batch variables. We saw some batch effects in the mRNA expression Agilent arrays, however, those data were not used in any of the analysis in the main paper. mRNA sequencing, miRNA sequencing and SNPs showed no major batch effects. DNA methylation showed a dichotomy of samples, which was found to be gender based due to the inclusion of sex chromosomes. After removing probes on the sex chromosomes, the dichotomy disappeared, and no major batch effects were seen.

III. COPY NUMBER STUDIES

Workgroup leaders: Rameen Beroukhim (rameen@broadinstitute.org) and Sabina Signoretti (ssignoretti@partners.org)

SNP Array-Based Copy Number Analysis. Tumor- and germline-derived DNA samples were hybridized to Affymetrix SNP 6.0 arrays as previously described [1]. Preliminary copy numbers at each probe locus were inferred from raw .CEL files using Birdseed [2] and refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that is most similar to the tumor [3] (Tabak et al, manuscript in preparation). 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. Data were then segmented using Circular Binary Segmentation [4]. Similar analyses of all germline-derived data from TCGA were performed to identify regions frequently involved in germline copy number variations (Tabak et al, manuscript in preparation); probes within these regions were removed from the tumor copy number profile data.

All samples following processing were analyzed by several quality control metrics (Tabak et al, manuscript in preparation). Both tumor and germline samples were screened for noise, as evidenced by median absolute differences between $\log_2$ ratios greater than 0.6 or more than 1000 segments following Circular Binary Segmentation. Germline samples were also excluded if their Affymetrix FQC call rates were below 86% or their Birdseed call rates were below 95%, or if their copy number profiles suggested contamination with tumor DNA. Contamination was assessed by computing the mean absolute $\log_2$ ratio across all probes after segmentation; germline samples exceeding 0.073 on this metric were discarded. Five tumor samples (1%) and 252 germline samples (7%) were rejected on the basis of these criteria.

Segmented copy number profiles were analyzed using Ziggurat deconstruction [3,5] to determine the most likely set of events contributing to these profiles, and the lengths, amplitudes, and locations of these events. Absolute $\log_2$ ratios greater than 1.5 were capped to 1.5 to reduce hypersegmentation due to variations in dynamic range between probes, and events whose absolute amplitude was less than a $\log_2$ ratio of 0.1 were excluded from further analysis as likely to represent noise. Events whose length was greater than and less than 50% of the chromosome arm on which they resided were called arm-level and focal events, respectively, and these groups of events were analyzed separately using GISTIC 2.0 [5]. Regions were considered significant if assigned False Discovery Rate [6] q-values < 0.25. SNP array-based estimates of tumor purity or heterogeneity were made using the ABSOLUTE method [7].

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. 2008 Oct;40(10):1166-74.
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. 2008 Oct;40(10):1253-60.

3. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.
4. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007 Mar 15;23(6):657-63.
5. 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 Biol. 2011;12(4):R41.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 1995; 57(1): 289–300.
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. Nat Biotechnol. 2012 May;30(5):413-21.

Figure S21

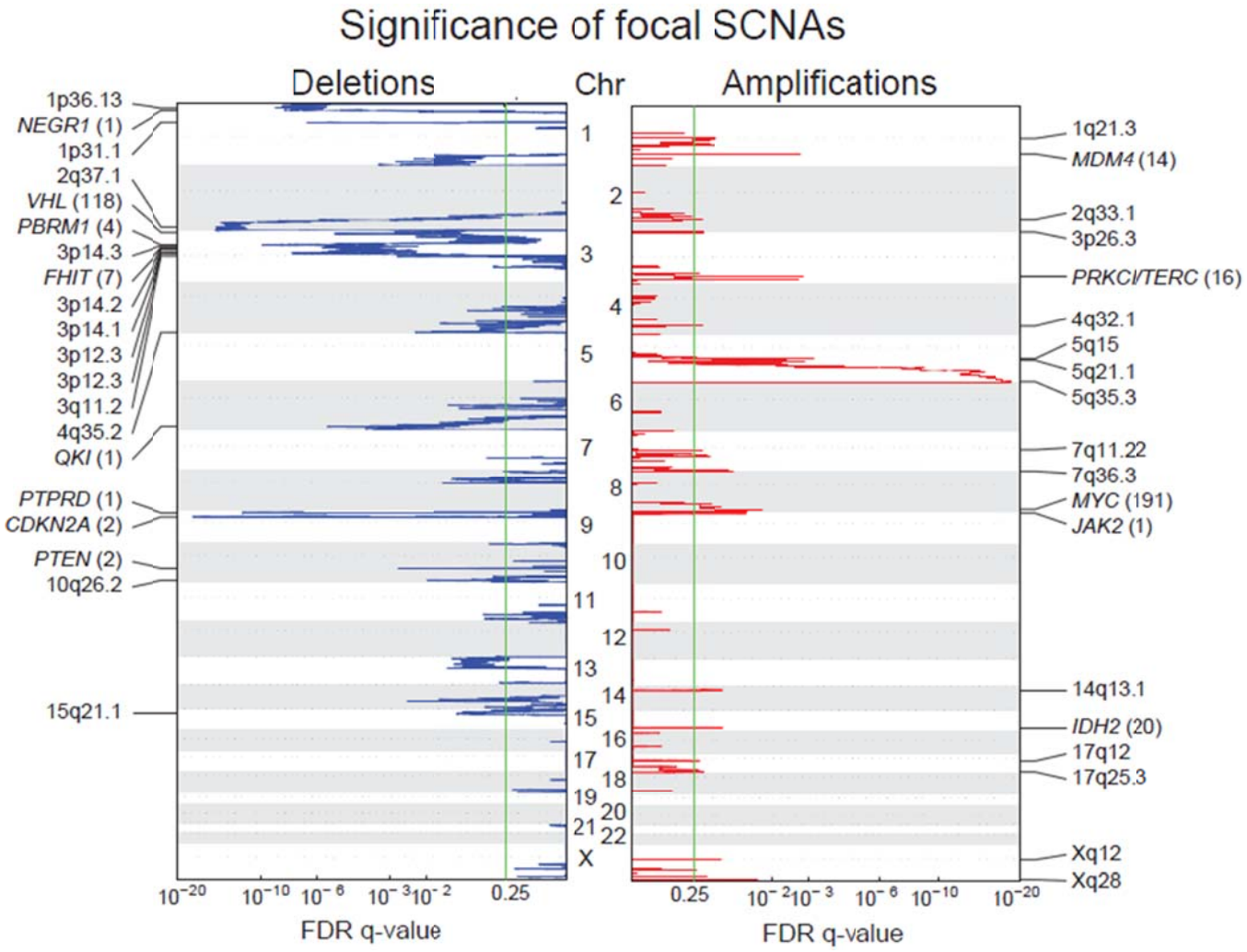


Figure S21. GISTIC results for Focal-level genomic gains and losses.

Figure S22

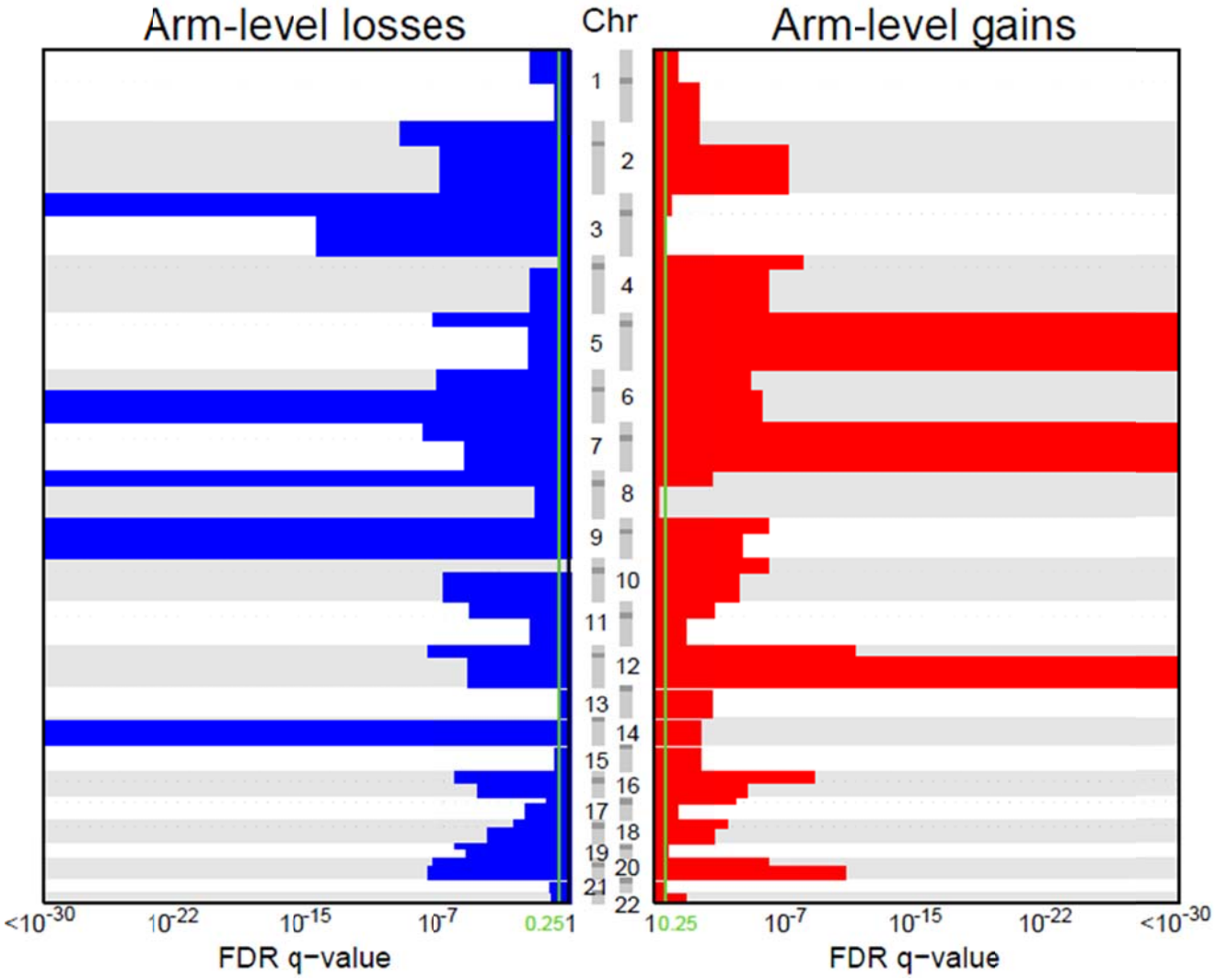


Figure S22. GISTIC results for Arm-level genomic gains and losses.

Table S3

Table S3. Peak regions of amplification and deletion

A) Amplification

Rank	Chromosome and band	Peak region	GISTIC q value*	Frequency** (%)	# genes	Candidate(s)
1	5q35.2	chr5:174209144-178285959	1.73E-19	67	60	
2	3q26.2	chr3:170542528-171791124	1.48E-03	15	16	*PRKCI, MECOM*
3	1q32.1	chr1:202117898-203315688	1.75E-03	14	14	*MDM4*
4	8q24.22	chr8:118645656-146274826	1.54E-02	15	191	*MYC*
5	Xq28	chrX:148431362-148952854	1.98E-02	7	10	
6	5q15	chr5:93674789-95079780	2.51E-02	44	10	
7	9p24.1	chr9:5027462-5087280	3.57E-02	3	1	*JAK2*
8	7q36.2	chr7:154601843-155095928	5.78E-02	34	3	
9	5q21.1	chr5:98449304-104685049	7.26E-02	50	12	
10	14q13.1	chr14:33505064-34060966	9.12E-02	3	2	
11	15q26.1	chr15:87959149-88798432	9.12E-02	5	20	*IDH2*
12	1q21.3	chr1:120325479-175321502	1.12E-01	14	557	
13	2q33.1	chr2:163524241-212418042	1.83E-01	16	253	
14	4q32.1	chr4:158901938-161466668	1.83E-01	3	9	
15	7q11.22	chr7:68848460-106992904	1.83E-01	34	313	
16	17q25.3	chr17:55285429-78774742	1.83E-01	8	317	
17	3p26.3	chr3:1-2192010	1.91E-01	2	3	
18	17q12	chr17:31417410-33664565	2.05E-01	6	29	
19	Xq12	chrX:67043093-67518992	2.05E-01	5	1	

B) Deletion

Rank	Chromosome and band	Peak region	GISTIC q value*	Frequency** (%)	# genes	Candidate(s)
1	9p21.3	chr9:21855498-21987722	8.69E-14	32	2	*CDKN2A*
2	2q37.1	chr2:231392094-231689020	2.76E-13	8	4	
3	9p24.1	chr9:7789607-12683402	1.44E-08	31	1	*PTPRD*
4	1p31.1	chr1:71313038-74273894	3.66E-05	12	1	*NEGR1*
5	3p21.1	chr3:52547488-52702296	8.57E-04	92	4	*PBRM1*
6	3p12.3	chr3:79897608-85095529	4.53E-03	60	1	
7	1p36.13	chr1:19523993-19798947	1.07E-02	16	1	
8	6q26	chr6:163662302-165618151	1.97E-02	28	1	*QKI*
9	10q23.31	chr10:89607138-90024018	1.64E-01	18	2	*PTEN*
10	3q11.2	chr3:96192501-98859899	2.86E-01	32	1	
11	4q35.2	chr4:169984149-191273063	5.21E-01	16	80	
12	3p26.1	chr3:1-21422837	6.25E-01	91	118	*VHL*
13	3p14.2	chr3:63085239-66096086	6.92E-01	78	9	
14	3p14.2	chr3:58626894-63067698	9.48E-01	80	7	*FHIT*
15	10q26.2	chr10:122337267-131156423	9.60E-01	18	51	
16	3p14.1	chr3:67781506-71088210	1.57E+00	71	9	

17	1p35.3	chr1:7751874-33753604	1.75E+00	14	383	
18	3p21.1	chr3:53349395-56568669	2.09E+00	89	11	
19	15q21.1	chr15:32830065-48264526	4.73E+00	8	155	
20	3p12.3	chr3:74651634-81622802	5.41E+00	65	6	
21	6q15	chr6:86444608-89853072	5.73E+00	24	17	
22	6q27	chr6:163913335-170899992	5.73E+00	28	40	
23	14q24.3	chr14:77467381-79403788	5.73E+00	45	1	*NRXN3*
24	4q24	chr4:88986033-120363308	7.54E+00	16	129	
25	14q24.3	chr14:55332533-106368585	8.73E+00	45	462	
26	1p36.22	chr1:1-30957443	9.47E+00	16	456	
27	1q42.3	chr1:209372793-247249719	1.01E+01	5	267	
28	3p12.1	chr3:81890234-87070815	1.01E+01	51	1	*CADM2*
29	8p11.21	chr8:42504810-42671789	1.27E+01	25	1	
30	11q23.3	chr11:102460302-134452384	1.27E+01	6	268	
31	7q31.1	chr7:109386704-111644293	1.28E+01	1	3	
32	10p15.3	chr10:1-15600404	1.47E+01	12	82	
33	Xq28	chrX:153391287-154913754	1.47E+01	10	33	
34	8p21.1	chr8:22065874-47872244	1.67E+01	30	151	
35	1q41	chr1:205382725-247249719	1.97E+01	5	294	
36	3p11.2	chr3:88286535-95217360	1.97E+01	47	2	
37	13q33.3	chr13:106016413-107662902	1.97E+01	14	2	
38	13q12.13	chr13:1-65781730	1.97E+01	13	229	
39	3q12.1	chr3:1-199501827	2.35E+01	31	1158	
40	8p23.2	chr8:1-146274826	2.35E+01	31	745	

* Residual q-value representing significance independent of neighboring peaks (see Supplementary Methods). Note that this usually differs from the overall q-values reported for individual genes or SNPs within a peak.

** These frequencies include both arm-level and focal events affecting the locus.

IV. MUTATION ANALYSIS

Workgroup leaders: David Wheeler (wheeler@bcm.edu) and Kristian Cibulskis (kcibul@broadinstitute.org)

Contributors: Caleb Davis, Liu Xi, Carrie Sougnez, Singer Ma, Anita Samantaray, David Haussler, Matthew Meyerson, Gad Getz, Richard Gibbs

Library construction: Illumina HiSeq. After QC, high molecular weight native DNA samples were constructed into Illumina PairEnd precapture libraries according to the manufacturer's protocol (Illumina Inc.) modified as follows: 0.5 - 1ug genomic DNA in 100ul volume were sheared into fragments of approximately 300 base pairs in Covaris plate with E210 system (Covaris, Inc. Woburn, MA). The setting was 10% Duty cycle, Intensity of 4, 200 Cycles per Burst, for 120 seconds. Fragment size was checked using a 2.2 % Flash Gel DNA Cassette (Lonza, Cat. No.57023). The Fragmented DNA was End-Repaired in 90ul total reaction volume containing sheared DNA, 9ul 10X buffer, 5ul END Repair Enzyme Mix and H2O (NEBNext End-Repair Module; Cat. No. E6050L) and then incubated at 20°C for 30 minutes. A-tailing was performed in a total reaction volume of 60ul containing End-Repaired DNA, 6ul 10X buffer, 3ul Klenow Fragment (NEBNext dA-Tailing Module; Cat. No. E6053L) and H_2O followed by incubation at 37°C for 30 minutes. Illumina multiplex adapter ligation (NEBNext Quick Ligation Module Cat. No. E6056L) was performed in a total reaction volume of 90ul containing 18ul 5X buffer, 5ul ligase, 0.5ul 100uM adaptor and H2O at room temperature for 30 minutes. After Ligation, PCR with Illumina PE 1.0 and modified barcode primers (manuscript in preparation) was performed in 170µl reactions containing 85 2x Phusion High-Fidelity PCR master mix, adaptor ligated DNA, 1.75ul of 50uM each primer and H2O. The standard thermocycling for PCR was 5' at 95°C for the initial denaturation followed by 6-10 cycles of 15 s at 95°C, 15 s at 60°C and 30 s at 72°C and a final extension for 5 min. at 72°C. Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) was used to purify DNA after each enzymatic reaction. After Beads purification, PCR product quantification and size distribution was determined using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).

DNA sequencing: Illumina HiSeq. Sequencing was performed in paired-end mode with Illumina HiSeq 2000. Illumina sequencing libraries were amplified by "bridge-amplification" process using Illumina HiSeq pair read cluster generation kits (TruSeq PE Cluster Kit v2.5, Illumina) according to the manufacturer's recommended protocol. Briefly, these libraries were denatured with sodium hydroxide and diluted to 3-4 pM in hybridization buffer for loading onto a single lane of a flow cell in order to achieve 600-700k clusters/mm. All lanes were spiked with 1% phiX control library. Cluster formation, primer hybridization were performed on the flow cell with illumina's cBot cluster generation system.

Sequencing reactions were extended for 202 cycles of SBS using TruSeq SBS Kit on an Illumina's Hiseq 2000 sequencing machine according to the manufacturer's instructions. The Illumina Sequence Control Software (SCS) control the reagent delivery and collect raw images. Real Time Analysis (RTA) software was used to process the image analysis and base calling. On average, about 80-100 million successful reads, consisting of 2 X100 bp, were generated on each lane of a flow cell.

Library construction: SOLiD 4. Whole genome amplified (WGA) DNA samples (5ug) were constructed into SOLiD precature libraries according to a modified version of the manufacturer's protocol (Applied Biosystems, Inc.). Briefly, The genomic DNA was sheared into fragments of approximately 120 base pairs with the Covaris S2 or E210 system as per manufacturer instruction (Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair (NEBNext End-

Repair Module; Cat. No. E6050L) and A-tailing (NEBNext dA-Tailing Module; Cat. No. E6053L). The resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit (Cat. No. M2200L). Solid Phase Reversible Immobilization (SPRI) bead cleanup (Beckman Coulter Genomics, Inc.; Cat. No. A29152) was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR (LM-PCR) was performed using Platinum PCR Supermix HIFi (Invitrogen; Cat. No.12532-016) and 6 cycles of amplification. After Beads purification, PCR products' quantification and their size distribution were analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). Primer sequences and a complete library construction protocol are available on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

Exome capture and sequencing: SOLiD. The precapture libraries (2 ug) were hybridized in solution to NimbleGen CCDS Solution Probes which targets ~36 Mbs of sequence from ~17K genes, according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture LM-PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over the background. The captured libraries were further processed for SOLiD sequencing. Primer sequences and a complete capture protocol are available on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

Exome capture and sequencing: Illumina. Precapture libraries (1 ug) were hybridized in solution to NimbleGen SeqCap EZ Exome 2.0 Solution Probes targeting ~44Mbs of sequence from ~30K genes, or VCRome 2.1 (HGSC design, NimbleGen) targeting 43 Mb of sequence from ~30K genes, according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos IHE1, IHE2 and IHE3 (manuscript in preparation) replaced oligos HE1.1 and HE2.1 and post-capture LM-PCR was performed using 14 cycles. Capture libraries were quantified using Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over background.

DNA Sequencing: SOLiD. Each captured library was hybridized to microbeads using Applied BioSystems' SOLiD platform-specific adapters) and submitted to an emulsion PCR to amplify the DNA fragments onto the beads (SOLiD ePCR Kit V2, Applied Biosystems). After amplification, the beads were recovered from the oil phase and the beads carrying amplified bound DNA were enriched (SOLiD Buffer and Bead Enrichment Kits, Applied Biosystems). The beads carrying amplified bound DNA were then modified to covalently adhere to a SOLiD coated slide (SOLiD Bead Deposition and Slide Kits, Applied Biosystems). The slides were loaded on the SOLiD v3 sequencing platform (SOLiD 3 Instrument Buffer Kit, Applied Biosystems) and sequenced over 8 days (SOLiD Fragment Library Sequencing Kit – MM50, Applied Biosystems).

Mapping Reads. *SOLiD.* Base and quality calling for SOLiD data was performed on-instrument using standard vendor software and settings. Upon completion of a run, read and quality data was copied into our data-center where individual sequence events are split into 10M read bundles and mapped in

parallel using BFAST (version 0.6.4). After read bundles are mapped their results are merged back into a single sequence-event-level BAM where read group tags are added. Where necessary, sample-level BAMs are generated by merging using Picard (version 1.7), and duplicate reads are marked at the library level using SAMtools (version 1.7). Variant calling is done using custom filters applied to pileups made at the sample level, also using SAMtools.

Illumina. The output of a Illumina HiSeq sequencer are binary bcl files that are processed using BCLConvertor 1.7.1. All reads from the prepared libraries that passed the illumina Chastity filter were formatted into fastq files. The fastq files are aligned to the genome using BWA (bwa-0.5.9rcl) against human reference genome build36 (NCBI). Default parameters are used for alignment except for a 40 bp seed sequence, 2 mismatches in the seed, and a total of 3 mismatches allowed. BAM files generated from alignment of Illumina sequencing reads were preprocessed using GATK [1] to recalibrate and locally realign reads. BAM files were deposited in CGHUB (https://cghub.ucsc.edu/) for controlled access distribution.

Mutation Detection. BAM files from all 460 patients were deposited into CGHUB (https://cghub.ucsc.edu/). Mutations were called using the tumor and matched normal BAM files by the BCM HGSC, Broad Institite and UCSC centers. Mutations at the Broad were discovered by the MuTect algorithm (http://www.broadinstitute.org/cancer/cga/MuTect).

At BCM, mutations in BAM files detected in SOLiD sequence data as follows: SamTools Pileup was run to list all variants found in multiple reads at a single locus. The variants were further filtered to remove all those observed fewer than 5 times or were present in less than 0.1 of the reads (variant allele fraction must be greater than 0.1). At least one variant had to be Q30 or better, and the variant had to lie in the central portion of the read, 15% from the 5' end of the read and 20% from the 3' end. In addition reads harboring the variant must have been observed in both forward and reverse orientations. Finally, the variant base was not observed in the normal tissue. Insertion or deletion variants ("indels") were discovered by similar processing except indels must have been observed in 10 of the reads. Mutations in Illumina were found by similar procesing, except that the initial processing was with AtlasSNP [3] instead of Pileup and the variant allele fraction threshold was 0.08.

UCSC's mutation caller, UCSC BamBam, is based on a novel bayesian model of joint tumor-normal genotypes. It infers the most likely joint genotype based on the sequencing data available taking into account the base quality of the bases. Certain low quality reads are ignored due to their propensity to produce false positive calls. Additionally, certain regions of the genome that are known to be difficult to sequence and/or align are removed from consideration based on data from the 1000 Genomes project. After the initial Single Nucleotide Variation calls are made, filtering is performed to reduce false positives using heuristics developed from examination of validated mutation calls. These seek to eliminate artifacts of the sequencer and mapping algorithm by removing mutations that exhibit significant strand bias or positional bias, as well as excluding mutations in the neighborhood of detected small insertions or deletions.

Target Coverage and Gap Filling. The regions from initial whole exome capture and sequencing were well covered to a sequence depth of at least 20X across and average of 80-90% of the target in the great majority of patients (Figure S23, also see coverage tables below for Solid and Illumina). However, exon 1 of VHL and two exons each of KDM5A and SETD2 were relatively poorly covered as a result of their high G+C content. Particularly in the case of VHL, these poorly covered exons could result in underestimations of the number of mutations in these genes. To fill these sequence gaps we designed PCR primers to amplify each of these exons plus an additional 90 low-coverage exons across all patients for further DNA sequencing. The 95 PCR reactions from each patient were barcoded and pooled in groups of 12 and run on MiSeq and PacBio sequencers using standard library generation procedures. This gap-filling sequencing increased the coverage over these exons by a variable amount, depending on the gene (see Figure S24). Mutations were detected at BCM in data from the

MiSeq sequencers as described above for whole exome data. Candidates were then visually validated in the data from the PacBio sequencers using IGV. Lastly, if no VHL mutations were observed by HiSeq or MiSeq in a sample, the PacBio sequencing data along the first exon of VHL was visualized in IGV. A mutation detected in this fashion was reported if the variant allele was also present in the MiSeq data.

The final mutation table composed of mutations from the initial coverage plus the gap filling coverage is shown in Data File S10.

Whole genome sequencing. Tumor and normal pairs, 22 patients were subjected to whole genome sequencing for an average of 90 Gb (30X coverage). Somatic mutations were called as described above for whole exome data. Large-scale rearrangements were called with BreakDancer. Mutation tallies for each patient are shown in Table S5.

The WGS sequencing of these 22 patients was used to calibrate the mutation callers from each center (Human Genome Sequencing Center, Broad Institute, UCSC; see above). Of 1264 mutations discovered by at least two centers in the whole exome data, 1045 were found in the WGS data (at least two reads with variant allele in tumor, and none in normal), for an overall validation rate of 83% (Table S6). The mutations could be classified by which center discovered them.

Inter-institutional Replicates. Four samples were subjected to whole exome sequencing independently by both the Broad Institute (BI) and the Human Genome Sequencing Center (HGSC). Mutations were then called independently by each center on their respective BAM files. There were 247 mutations detected by both centers, while 184 and 50 mutations were found uniquely by BI and HGSC, respectively. The allele fractions of the single nucleotide polymorphisms (SNPs) found jointly by both centers are shown in Figure S25.

Validation on 454 and Ion Torrent. We also validated our mutation calls from whole exome sequencing using alternative instrumentation and chemistries to avoid any systematic errors inherent to the processes described above. We submitted sample-sites -- the majority of which were in SMGs -- for amplification by PCR followed by analysis on 454 sequencing instruments as previously described [2]. Amplicons were also prepared for sequencing using the Life Technologies Ion Xpress and Ion OneTouch protocols and reagents. Briefly, amplicons were clonally amplified on Ion Sphere Particles (ISPs) through emulsion PCR and then enriched for template-positive ISPs. For Ion Torrent runs, approximately 35 million template-positive ISPs per run were deposited onto the Ion 318C chips (Life Technologies, 4466617) by a series of centrifugation steps that incorporated alternating the chip directionality. Sequencing was performed with the Ion Personal Genome Machine Sequencing Kit (Life Technologies, 4474004) using the 440 flow (“200bp”) run format. The instrumentation used for validation is indicated in the MAF file (Data File S10)

Resulting fastq files were aligned to the reference genome with BLAT and realigned at the site by cross_match. A validation status was assigned if at least 50 reads spanned a sample-site (depth of 50x). Of 2073 sample-sites analyzed, 2053 (99%) were covered to sufficient read depth. Of these, 1488 (72.5%) were validated as somatic, 454 (22.1%) were not observed in the tumor (wildtype), and 111 (5.4%) were observed in the normal (germline). The table below shows the allele fraction thresholds required during validation to determine the result.

Validation_Status	Tumor	Normal
Wildtype	<= 2% (SNV); <= 20% (indel)	NA
Valid (Germline)	> 2% (SNV); >20% (indel)	>=1.5% (SNV & indel)
Valid (Somatic)	> 2% (SNV); > 20% (indel)	< 1.5% (SNV & indel)

Of particular interest in our analysis was the validation rate grouped by which center(s) called the mutation, the type of variant (SNP or INDEL), and sequencing platform. Our results are summarized in Figure S26.

For the initial run of automated pathway analysis approaches (e.g. Paradigm, Memo, and Hotnet, see elsewhere), mutations calls made independently by two or three centers were used, as these were found to share high overlap with the orthogonal 454 and Ion Torrent results (Figure S26). The final Mutation Analysis File (MAF) includes all calls from BI and BCM as well as a handful of mutation events made by UCSC validation evidence (i.e., confirmation by Ion Torrent or 454).

Significantly mutated genes (SMGs). The ranking of genes in terms of estimated conferred selective advantage was performed using the mutation statistical analysis algorithm MutSig (v1.5, Lawrence et al., manuscript in preparation). The MutSig algorithm works with an aggregated list of mutations across the entire patient set and estimates background mutation rates. The p and q values for a certain gene, corresponding to raw probability, p, and the probability, q, corrected for multiple testing, are determined for the mutation rate observed in that gene in relation to the background model.

MutSig uses various factors to accurately estimate the background mutation rate, taking into account the background mutation rates of different mutation categories (e.g., transitions or transversions in different sequence contexts), as well as the fact that different samples have different background mutation rates. It then uses convolutions of binomial distributions to calculate the p and values for each gene, which represents the probability that we observe by chance a certain configuration of mutations in a gene, given the background model. It also takes into account the non-synonymous to synonymous mutation ratio for each gene in order to separate out the genes with a large number of non-synonymous events compared to synonymous ones.

Using the single nucleotide variants and indels listed in Data File S10, we tallied all somatic mutations that were called by two or more centers, plus all those from a single Center validated.

VHL mutation rates. DNA sequencing stuides by both Nickerson et al. [4] and Moore et al. [5] observed 82% of patients with a somatic mutation in VHL, whereas we observed 52%. Comparing each mutation type across the three studies (Table S7) we can see that our study had similar missense, nonsense and splice site substitution rates, but less than half the rate of insertion and deletion. We note that the Nickerson and Moore studies augmented their sequencing results with a sensitive endonuclease scanning procedure that relied on enzyme digestion of heteroduplexes formed between patient and wild-type DNA PCR products, which may account for some of the improved sensitivity to indels in their studies. We also recognize that indel discovery in next generation sequencing reads is somewhat less sensitive than substitution discovery.

References

1. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 9:1297-303.
2. The Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature. 487:330-337
3. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20:273-80.
4. Nickerson ML, Jaeger E, Shi Y, Durocher JA, Mahurkar S, Zaridze D, Matveev V, Janout V, Kollarova H, Bencko V, Navratilova M, Szeszenia-Dabrowska N, Mates D, Mukeria A, Holcatova I, Schmidt LS, Toro JR, Karami S, Hung R, Gerard GF, Linehan WM, Merino M, Zbar B, Boffetta P, Brennan P, Rothman N, Chow WH, Waldman FM, Moore LE. (2008). Improved identification

of von Hippel-Lindau gene alterations in clear cell renal tumors. Clin Cancer Res. 14:4726-34. PubMed PMID: 18676741; PubMed Central PMCID: PMC2629664.

5. Moore LE, Nickerson ML, Brennan P, Toro JR, Jaeger E, Rinsky J, Han SS, Zaridze D, Matveev V, Janout V, Kollarova H, Bencko V, Navratilova M, Szeszenia-Dabrowska N, Mates D, Schmidt LS, Lenz P, Karami S, Linehan WM, Merino M, Chanock S, Boffetta P, Chow WH, Waldman FM, Rothman N. (2011). Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: associations with germline VHL polymorphisms and etiologic risk factors. PLoS Genet. 2011 7:e1002312. PubMed PMID: 22022277; PubMed Central PMCID: PMC3192834.
6. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, Teague J, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Forbes S, Jia M, Jones D, Knott H, Kok CY, Lau KW, Leroy C, Lin ML, McBride DJ, Maddison M, Maguire S, McLay K, Menzies A, Mironenko T, Mulderrig L, Mudie L, O'Meara S, Pleasance E, Rajasingham A, Shepherd R, Smith R, Stebbings L, Stephens P, Tang G, Tarpey PS, Turrell K, Dykema KJ, Khoo SK, Petillo D, Wondergem B, Anema J, Kahnoski RJ, Teh BT, Stratton MR, Futreal PA.(2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature 463:360-3. PubMed PMID: 20054297; PubMed Central PMCID: PMC2820242.

Table S4

Table S4. Mutation frequencies of 50 top significantly mutated genes.

Symbol	Entrez ID	MAF count	MAF rate (%)	SMG count	SMG Q-value	Title
VHL	7428	234	52.3	177	<9.43e-12	von Hippel-Lindau tumor suppressor
PBRM1	55193	140	32.9	137	<9.43e-12	polybromo 1
SETD2	29072	53	11.5	51	2.58E-11	SET domain containing 2
KDM5C	8242	28	6.7	27	2.58E-11	lysine (K)-specific demethylase 5C
PTEN	5728	21	4.3	17	2.58E-11	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)
BAP1	8314	44	10.1	42	2.58E-11	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)
MTOR	2475	26	6	26	2.81E-10	mechanistic target of rapamycin (serine/threonine kinase)
TP53	7157	10	2.2	10	1.93E-06	tumor protein p53
PIK3CA	5290	12	2.9	12	7.33E-06	phosphoinositide-3-kinase, catalytic, alpha polypeptide
MSR1	4481	6	1.4	6	0.0559	macrophage scavenger receptor 1
TXNIP	10628	5	1.2	5	0.0566	thioredoxin interacting protein
TCEB1	6921	4	0.7	3	0.0566	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)
NFE2L2	4780	6	1.4	6	0.0655	nuclear factor (erythroid-derived 2)-like 2
BTNL3	10917	5	1.2	5	0.0655	butyrophilin-like 3
SLITRK6	84189	7	1.7	7	0.0655	SLIT and NTRK-like family, member 6
RHEB	6009	4	1	4	0.0655	Ras homolog enriched in brain
ARID1A	8289	12	2.9	12	0.0726	AT rich interactive domain 1A (SWI-like)
NPNT	255743	6	1.4	6	0.0911	nephronectin
CCNB2	9133	5	1.2	5	0.098	cyclin B2
ZNF800	168850	6	1.4	6	0.11	zinc finger protein 800
SLC27A6	28965	5	1.2	5	0.199	solute carrier family 27 (fatty acid transporter), member 6
COL6A6	131873	12	2.9	11	0.223	collagen, type VI, alpha 6
SPRED1	161742	5	1.2	5	0.278	sprouty-related, EVH1 domain containing 1
FBN2	2201	13	2.9	12	0.278	fibrillin 2 (congenital contractural arachnodactyly)
STAG2	10735	7	1.7	7	0.283	stromal antigen 2
SECISBP2L	9728	7	1.7	7	0.312	SECIS binding protein 2-like
TFDP2	7029	5	1.2	5	0.312	transcription factor Dp-2 (E2F dimerization partner 2)
HMCN1	83872	19	4.6	18	0.317	hemicentin 1
ATM	472	15	2.9	15	0.358	ataxia telangiectasia mutated
MAGEC1	9947	13	3.1	7	0.397	melanoma antigen family C, 1
GPM6A	2823	3	0.7	3	0.397	glycoprotein M6A
MS4A12	54860	3	0.7	3	0.397	membrane-spanning 4-domains, subfamily A, member 12
OR2L8	391190	4	1	4	0.42	olfactory receptor, family 2, subfamily L, member 8
ZFPM2	23414	7	1.7	7	0.42	zinc finger protein, multitype 2
NKAIN3	286183	3	0.7	3	0.42	Na+/K+ transporting ATPase interacting 3
PGLYRP3	114771	4	1	4	0.42	peptidoglycan recognition protein 3
OR10AG1	282770	3	0.7	3	0.42	olfactory receptor, family 10, subfamily AG, member 1
KIAA0174	9798	4	1	4	0.438	KIAA0174
FAM5B	57795	6	1.4	5	0.438	family with sequence similarity 5, member B
DIO2	1734	4	1	3	0.451	deiodinase, iodothyronine, type II
SFXN4	119559	4	1	4	0.451	sideroflexin 4
EMR3	84658	5	1.2	5	0.451	egf-like module containing, mucin-like, hormone receptor-like 3
HOXC8	3224	3	0.7	3	0.451	homeobox C8
ATF7IP2	80063	5	1.2	5	0.451	activating transcription factor 7 interacting protein 2
SCARB2	950	4	1	4	0.451	scavenger receptor class B, member 2
PCNA	5111	3	0.7	3	0.451	proliferating cell nuclear antigen
SLC17A6	57084	5	1.2	5	0.451	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 6
MS4A3	932	3	0.7	3	0.451	membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)
TSPAN19	144448	5	1.2	3	0.451	tetraspanin 19
DST	667	19	4.3	18	0.451	dystonin

Figure S23

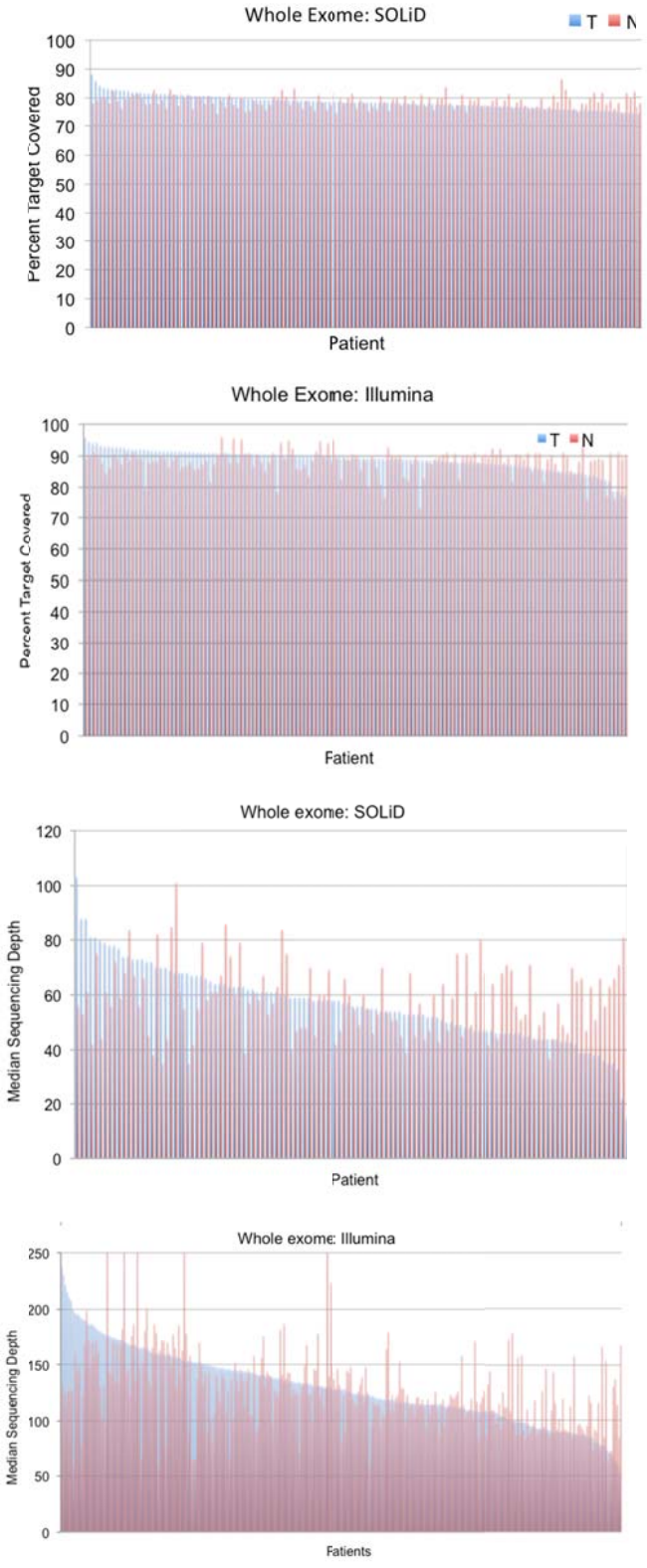


Figure S23. Target coverage and median coverages by sequencing platform. Only high-quality bases (phred score >Q20) are counted in median coverage calculation. T, tumor; N, normal.

Figure S24

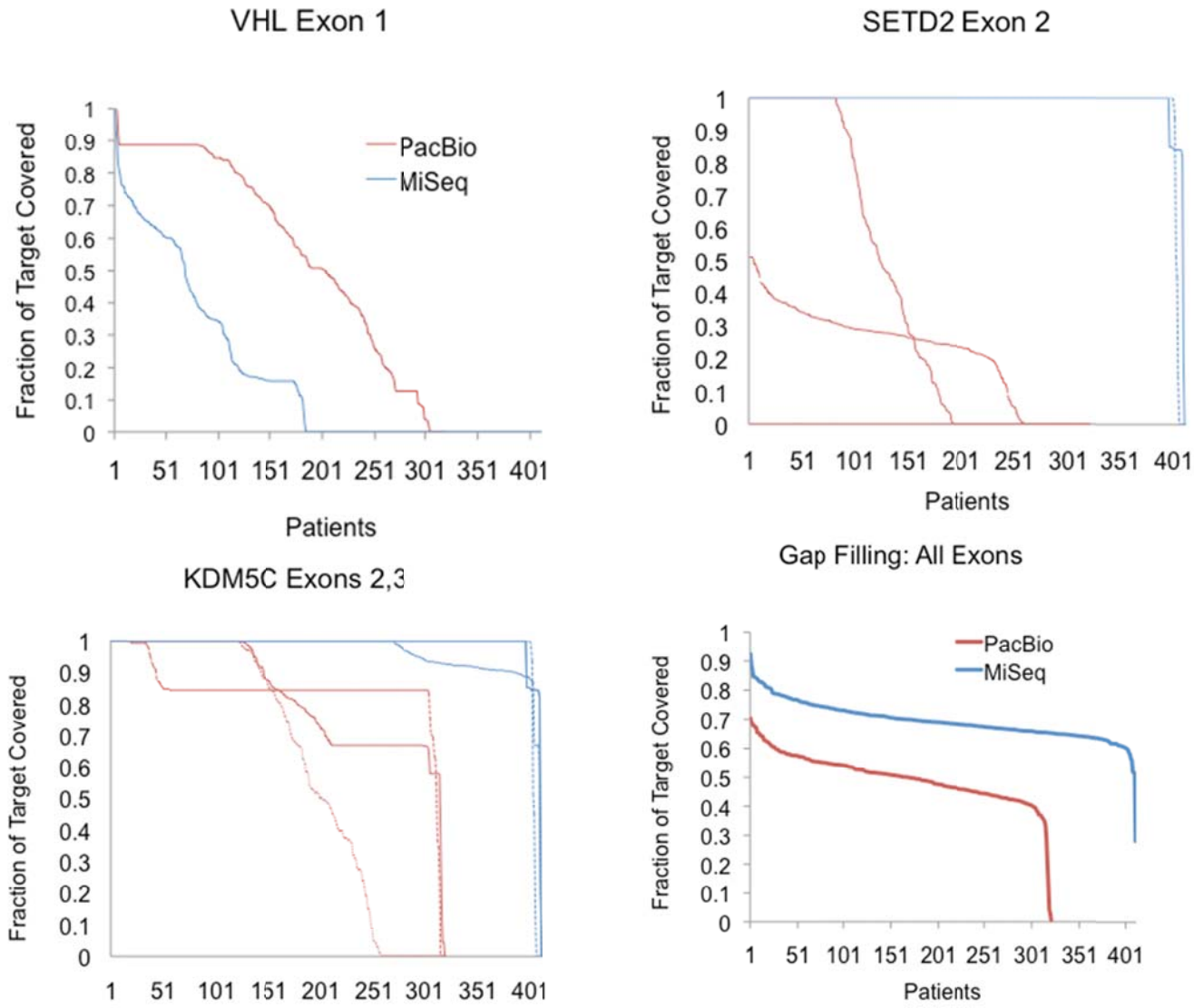


Figure S24. Gap filling coverage improvement for VHL, SETD2 and KDM5C, and all 95 Exons taken together. SETD2 and KDM5C exons were covered to nearly 100%. VHL was partially successful. In general MiSeq (blue lines) resulted in deeper coverage than PacBio (red lines). The list of all genes subjected to gap filling is the following: AKT1, CDKN2A, CROT, DOCK1, EGFR, EGLN2, FH, FNIP1, HRAS, IDH2, ITGA9, KDM5C, KDM6A, LINC00299, MDH2, MLL2, NF2, PLCL2, PRDM2, RASSF1, RET, RGS17, SEMA6D, SETD2, STK11, SUCLG1, SUCLG2, SYN3, TFE3, TFEB, TSC2, VEGFA, VHL, WT1.

Figure S25

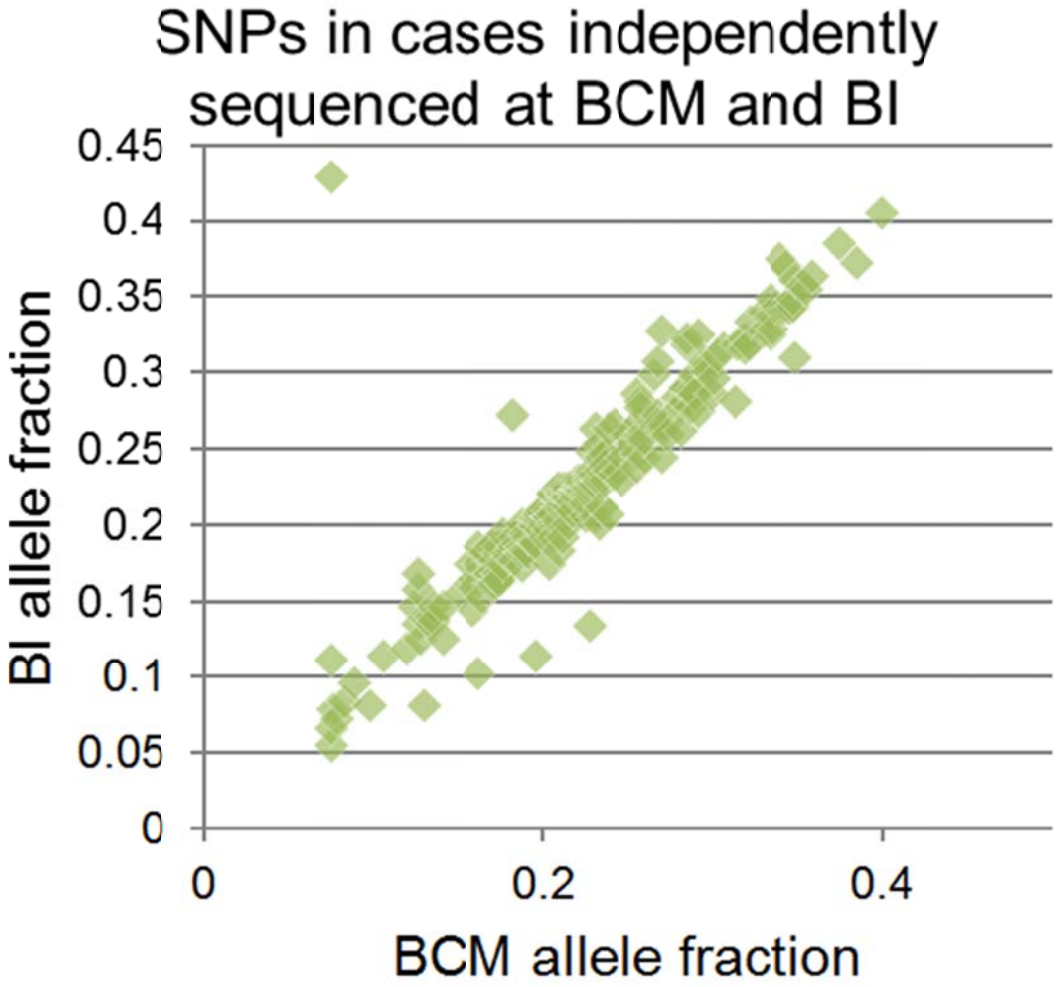


Figure S25. The allele fractions of the single nucleotide polymorphisms (SNPs) found jointly by both the Broad Institute (BI) and the Human Genome Sequencing Center (HGSC). Four samples were subjected to whole exome sequencing independently by both centers. There were 247 mutations detected by both centers, while 184 and 50 mutations were found uniquely by BI and HGSC, respectively.

Figure S26

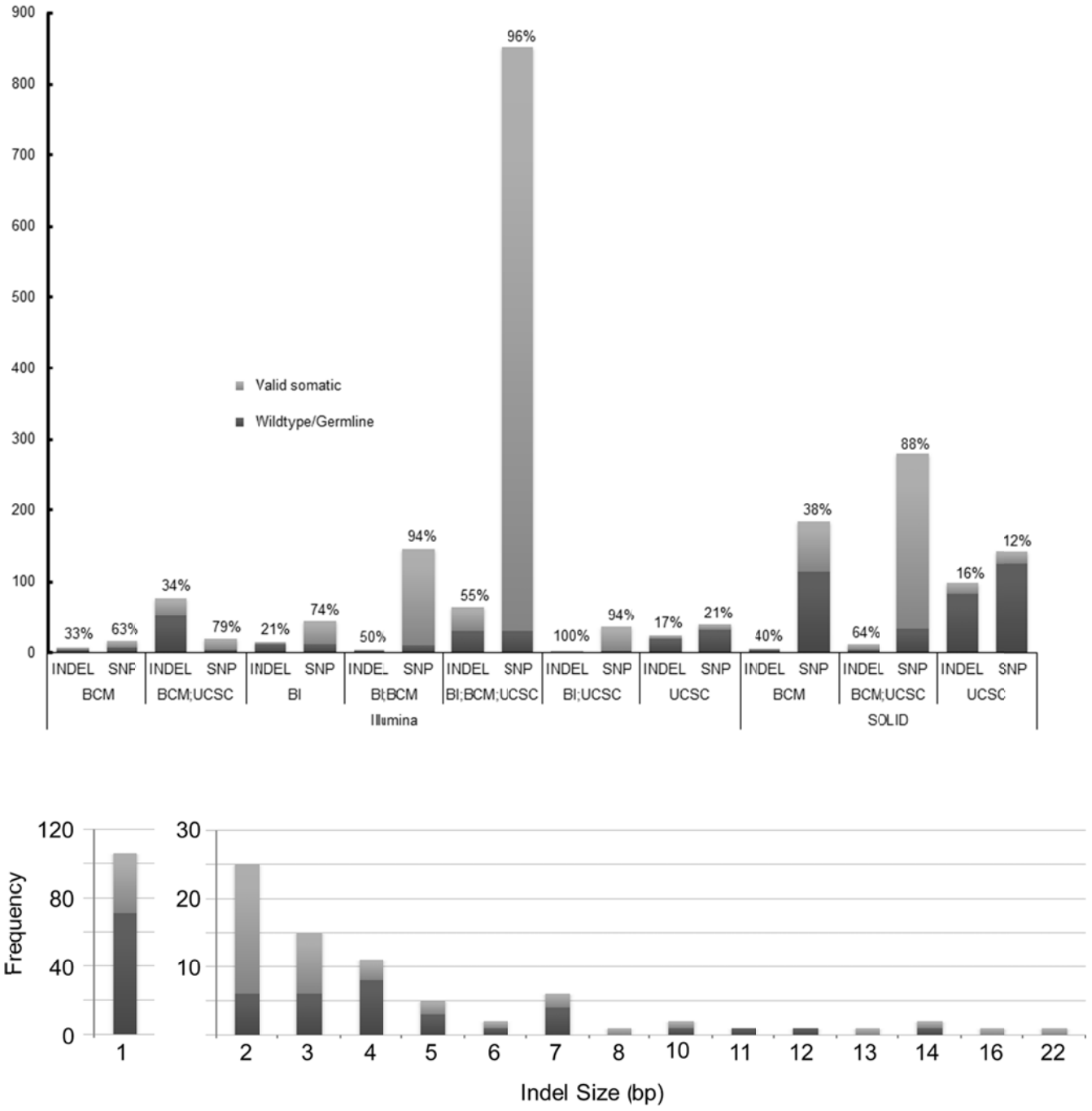


Figure S26. Validation rates (by combination of 454 and Ion Torrent). Y-axis represents numbers of mutation calls. In the top panel, the validation rates are grouped by which center(s) called the mutation, the type of variant (SNP or INDEL), and sequencing platform. After excluding calls by center(s) with less than a 20% validation rate, valid somatic (light grey) and wildtype/germline (dark grey) indels are reported in the bottom panel according to the sizes (in base pairs) of the mutations.

Table S5

Table S5. Mutation tallies for WGS for 22 renal clear cell carcinoma patients.

Tumor	Normal	Total	Novel	NovFrac
TCGA-BP-5168-01A-01D-2099-10	TCGA-BP-5168-11A-01D-2099-10	10837	8861	0.82
TCGA-BP-5010-01A-02D-2048-08	TCGA-BP-5010-11A-01D-2048-08	7059	5536	0.78
TCGA-BP-4781-01A-01D-2098-10	TCGA-BP-4781-11A-01D-2098-10	9873	7728	0.78
TCGA-CJ-5682-01A-11D-2103-10	TCGA-CJ-5682-11A-01D-2103-10	9173	7132	0.78
TCGA-CJ-4918-01A-01D-2101-10	TCGA-CJ-4918-11A-01D-2101-10	7486	5702	0.76
TCGA-DV-5566-01A-01D-2104-10	TCGA-DV-5566-10A-01D-2104-10	3391	2534	0.75
TCGA-CZ-5987-01A-11D-2104-10	TCGA-CZ-5987-11A-01D-2104-10	5740	4277	0.75
TCGA-A3-3308-01A-01D-2094-10	TCGA-A3-3308-11A-01D-2094-10	8074	6016	0.75
TCGA-AK-3455-01A-01D-2233-10	TCGA-AK-3455-10A-01D-2233-10	5493	4082	0.74
TCGA-A3-3372-01A-01D-2094-10	TCGA-A3-3372-11A-01D-2094-10	7348	5448	0.74
TCGA-CJ-6033-01A-11D-2104-10	TCGA-CJ-6033-11A-01D-2104-10	5755	4259	0.74
TCGA-BP-4968-01A-01D-2100-10	TCGA-BP-4968-11A-01D-2100-10	5229	3776	0.72
TCGA-AK-3454-01A-02D-2096-10	TCGA-AK-3454-10A-01D-2096-10	4921	3539	0.72
TCGA-CJ-4639-01A-01D-2095-10	TCGA-CJ-4639-11A-01D-2095-10	6872	4928	0.72
TCGA-A3-3387-01A-01D-2102-10	TCGA-A3-3387-11A-01D-2102-10	7899	5659	0.72
TCGA-CZ-5454-01A-01D-2102-10	TCGA-CZ-5454-11A-01D-2102-10	3693	2638	0.71
TCGA-CZ-4856-01A-02D-2101-10	TCGA-CZ-4856-11A-01D-2101-10	6031	4200	0.70
TCGA-B2-4101-01A-02D-2096-10	TCGA-B2-4101-11A-01D-2096-10	5686	3830	0.67
TCGA-A3-3370-01A-02D-2099-10	TCGA-A3-3370-11A-01D-2099-10	5041	3361	0.67
TCGA-B0-5693-01A-11D-2103-10	TCGA-B0-5693-11A-01D-2103-10	5805	3775	0.65
TCGA-CJ-4885-01A-01D-2098-10	TCGA-CJ-4885-11A-01D-2098-10	8424	5158	0.61
TCGA-CJ-4899-01A-01D-2100-10	TCGA-CJ-4899-11A-01D-2100-10	4795	2890	0.60

Table S6

Table S6. Accuracy of whole exome mutation calling.

Center*	Whole Exome	WGS	Percent Concordant
BI, HGSC	174	155	89
BI, HGSC,UCSC	762	687	90
BI, UCSC	51	42	82
HGSC, UCSC	277	161	58
TOTAL	1264	1045	83

*, BI, Broad Institute; HGSC, Human Genome Sequencing Center; UCSC, University of California, Santa Cruz.

Table S7

Table S7. VHL mutation rate in three studies.

	TCGA		Nickerson[a]		Moore[b]	
Category	Mutation	Pct	Mutation	Pct	Mutation	Pct
Deletion	53	12.7%	59	28.8%	145	30.9%
Insertion	18	4.3%	42	20.5%	62	13.2%
Missense	98	23.5%	42	20.5%	107	22.8%
Nonsense	38	9.1%	18	8.8%	39	8.3%
Splice Site[c]	18	4.3%	15	7.3%	30	6.4%
Patients	417		205		470	

a, see reference 5.

b, see reference 6.

c, Both Nickerson and Moore defined the splice site as 10 intron bases adjacent to the splice junction, TCGA used 2.

Table S8

Table S8. Nonsilent mutation rates for top mutated genes from Dalgliesh et al. study (Nature 2010) as observed in TCGA dataset.

Gene	Dangliesh %	TCGA %
SETD2	3.7%	11.5%
JARID1C	3.4%	6.5%
NF2	1.7%	1%
UTX	3%	1%
MLL2	4.2%	3.1%
HIF1A	0.7%	0.7%
NBN	0.5%	0.5%
ZUBR1	2.2%	1.7%
PMS1	0.7%	0.2%
WRN	0.7%	0.2%

Genes from Table 2 of Dalgliesh et al. [6]. TCGA data from final mutation analysis incorporating validation data.

Figure S27

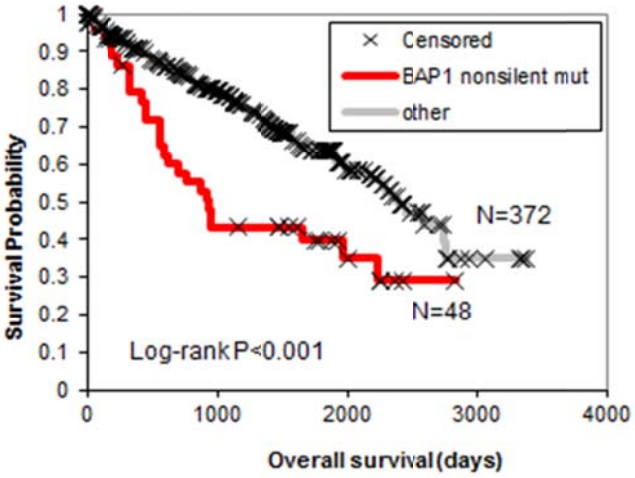


Figure S27: Kaplan-Meier plot for BAP1 nonsilent mutation versus wild-type, with patient overall survival as the outcome measure. Data from finalized mutation dataset (including validation results).

V. INTEGRATIVE ANALYSIS OF MUTATION AND COPY NUMBER

Workgroup leader: Boris Reva (borisr@mskcc.org)
Contributors: Nikolaus Schultz, Chris Sander

Rationale

Cancer is a disease driven by genomic alterations. A typical kidney tumor has ~60 missense and truncating mutations and about the same number of genes affected by copy number losses and amplifications. Most of these alterations are unique for a particular tumor. The enormous diversity of genomic alterations limits the applicability of statistical methods to determine concrete tumor-specific "driver" alterations among an overwhelming majority of "passengers". Only a few top genes that are frequently altered across a significant number of tumors can be nominated as drivers in a particular cancer. While the contribution of these genes to cancerogenesis is very important, there are many other cancer drivers that are mutated at lower frequencies. Therefore, an important practical task is to prioritize rarely and singly mutated genes as potential drivers by taking into account the total integrated impact of all alterations – here, we assess the impact of missense mutations, truncating mutations and copy number alterations.

Approach

To assess the impact of missense mutations, we used the functional impact score of MutationAssessor [1]; we used discretized copy number alterations predicted by GISTIC [2].

To prioritize genes by their role in cancer, we

- sum up different genomic alterations to determine a number of cases of potential inactivation of a given gene, as well as a number of cases of potential activation of a gene;

- determine statistical significance (Fisher test) of co-occurrence of heterozygous deletions and truncating and predicted functional mutations and statistical significance of co-occurrence of copy gains and amplifications with predicted functional mutations;

- use cancer annotations (e.g. oncogenes and tumor suppressors) [3,4];

The detailed results of this analysis are presented in **Data File S6**.

To sum up genomic alterations we used the following simple rules: 1. a gene is assumed to be inactivated if it is affected by a homozygous deletion or by truncating mutations; 2. a gene is assumed to be activated if it is affected by an amplification; 3. genes not annotated as oncogenes and affected by high-scoring functional mutations are predicted to be inactivated; 4. genes annotated as oncogenes are assumed to be activated by functional mutations.

The cancer annotations (tumor suppressors and oncogenes) are taken from the CancerGenes database from Memorial Sloan-Kettering Institute [3]. In this database, ~3,200 genes are annotated as

cancer genes; among them, 824 genes are as annotated tumor suppressors and 194 genes are annotated as oncogenes. CancerGenes [3] uses 25 "annotation sources", including major review papers and databases, which have information on the involvement of a particular gene in cancer. We integrated CancerGenes with the cancer genes census of Sanger Institute [4] and a list of cancer genes derived from the COSMIC database using the functional analysis of somatic mutations [1]. In total, we obtained a list of 3,761 genes that have at least one annotation source; 1225 genes have two or more annotation sources; 654 genes have 3 or more annotation sources. Using this information one can prioritize genomic alterations in rarely affected genes.

Summary of Results

A. Genes altered by mutations and copy number alterations were ranked by

1. Statistical significance of the enrichment of predicted functional mutations and truncating mutations in regions of heterozygous deletions. It is assumed that mutations statistically over-presented in regions of heterozygous deletions are selected in tumor evolution.

   We found that

   - the tumor suppressors **VHL, BAP1, SETD2, PBRM1, PTEN, and TP53,** have a significant enrichment of functional and truncating mutations in regions of heterozygous deletions (TP53 is rarely mutated in kidney cancer);
   - a few rarely mutated genes (RYR2, DYNC1H1, ESPL1, YLPM1, TCEB1, COL2A1, GPR179, ARSD, PLAC4) also have an enrichment of functional and truncating mutations in regions of heterozygous deletions
   - two oncogenes, **MTOR** and **MAP2K3,** also have significant enrichment of mutations in regions of heterozygous deletions.

2. A number of potentially inactivating alterations (homozygous deletions, truncating mutations, predicted functional mutations) in known tumor suppressors:

   The major tumor suppressors altered by homozygous deletions, truncating and predicted functional mutations in more than 5 tumors are: VHL, PBRM1, SETD2, BAP1, KAT2B, MLH1, PRKCD, KDM5C, PTEN, CDKN2A, ATM, TP53, SMARCA4, BRCA2, WRN, NF1.

   In total, there are **~364** annotated tumor suppressors altered in kidney cancer.

3. Statistical significance of the enrichment of predicted functional mutations and truncating mutations in regions of copy gains and amplifications. It is assumed that mutations statistically over-presented in regions of copy gains and amplification are selected in tumor evolution.
   - a few rarely mutated genes (RYR1, DNAH2, ITPR2, FMN2) have an enrichment of functional mutations in regions of copy gains;
   - a week sign of enrichment of mutations in regions of copy gains is detected for the oncogene **ERBB4** (P value ~0.09)

4. A number of potentially activating alterations (copy gains, amplifications, predicted functional mutations) in known oncogenes:

The major oncogenes altered by amplifications and predicted functional mutations in more than 3 tumors are: MAPK9, CDX1, MTOR, PIK3CA, ECT2, ERBB4, ABI2, EGFR, MET, PTK2, MAP2K3, PTPN12, PTGS2, PIK3CG, SMO, TET2, PDGFR;

In total, there are ~ **117** annotated OG altered in kidney cancer.

B. Rarely altered tumor suppressors and oncogenes are associated with poor outcome

To show the relevance of rarely mutated genes to cancer, we selected rarely mutated cancer genes (annotated oncogenes and tumor suppressors) from a "long tail" of genes altered by mutations and/or copy number amplifications. We selected only genes affected at least once by predicted functional mutations and we excluded genes located in regions of multiple amplification on chr5q and homozygous loss on chr3. Thus, 51 annotated oncogenes [3-4] were selected so that each of these genes has at least one predicted functional mutation and no more than two truncating mutations or homozygous deletions. Majority of the selected 51 genes are altered by mutations or copy number amplifications only in ~2 tumors. The maximal number of affected tumors - 7 is observed for genes EGFR and TET2. The main oncogenes, MTOR, PIK3CA were removed from this list. In the similar way, 71 tumor suppressors were selected so that each of them has no more than two genomic alterations resulted from predicted functional mutations, truncating mutations or homozygous deletions. All main tumor suppressors including VHL, SETD2, PBRM1, BAP1, TP53, CDKN2A, PTEN, ATM, SMARCA4, BRCA4, NF1 were removed from this list. Majority of the selected 71 genes are altered by mutations or homozygous deletions only in ~1-2 tumors. The maximal number of mutations - 7 is observed for ADAMTS18.

No survival information was used in selection of these genes. However, (i) 124 tumors (~30%) with mutations or amplifications in 51 top oncogenes have significantly poor outcome (Logrank P value=0.0023, Fig. X.iv-1); (ii) 132 tumors (~32%) affected by alterations in 71 selected tumor suppressors also show poor outcome (Logrank P value=0.02 Fig. X.iv-2); (iii) 210 tumors (~51%) are affected by the combined list of 122 top oncogenes and tumor suppressors (Logrank P value=0.0023, Fig.1C).

For the obtained survival classes, we compared the difference between distribution of tumors by size and by mutation count. The results of the comparison are given in a supplementary Table S9. There is a significant statistical difference between per tumor mutation counts in different survival classes (Table S9). Tumors with poor outcome have on average 42-44 missense mutations, while tumors in the class of better outcome have on average 35-37 mutations. Tumors of poor outcome have also more silent and truncating mutations as compared to tumors of better outcome, however the difference in a number of missense mutations per tumor is the most significant statistical factor that differentiate survival classes (Table S9). There is also a clear difference between average sizes of tumors in poor survival classes (tumors are bigger) and tumors in better survival class (tumors are smaller).

Table S9

Table S9. Statistical association between per tumor mutation counts and tumor sizes with the survival classes.

Survival class	Number of tumors	Average number per tumor				
		MM+TM+SM	MM	SM	TM	Size
better outcome (51 OG)	124	62.8	37.1	14.5	11.2	6.4
poor outcome (51 OG)	289	69.8	42.4	15.3	12.0	7.1
t-test P-value		**<E-06**	**<E-06**	7E-02	1.5E-01	E-02
better outcome (71 TS)	132	60.8	36.1	13.8	10.9	6.4
poor outcome (71 TS)	281	73.7	44.3	16.7	12.6	7.1
t-test P-value		**<E-06**	**<E-06**	**<E-06**	**2E-06**	8E-03
etter outcome (51 OG or 71 TS)	203	59.2	35.0	13.6	10.6	6.2
poor outcome (51 OG or 71 TS)	210	70.4	42.3	15.9	12.2	7.0
t-test P-value		**<E-06**	**<E-06**	**<E-06**	**<E-06**	3E-03

MM, SM, TM stand, respectively for a total number of missense, silent and truncating mutation per tumor; Size is “the maximum tumor dimension” in cm.

These results suggest that there is a “long-tail” of rarely mutated driver genes relevant to kidney cancer. Tumors with more missense mutations are enriched with such driver genes that contribute to tumor aggressiveness. Neither of these driver genes was detected by statistics analysis (MutSig [5]), however these genes can be prioritized by known oncogenic roles and by predicted functional mutations. Thus, rarely and singly mutated genes have to be taken into account in determining driver alterations in a particular tumor and in development of a personalized treatment of cancer.

C. The complete statistics of genomic alterations are presented in Data File S6. The worksheet “KIRC_StatistcsAllGenes” presents gene-based statistics of genomic alterations (missense and truncating mutations, predicted functional mutations, homozygous deletions, amplifications); the worksheets “KIRC_AnnotOG” and “KIRC_AnnotTS” present, respectively, genomic alterations in annotated oncogenes and tumor suppressors; the worksheets “KIRC_RarelyMutatedOG” and “KIRC_RarelyMutatedTS” present selected cancer genes used in survival analysis.

The Table uses the following abbreviations:

1. **Gene** – gene symbol
2. **Cytoband** – position on chromosome
3. **TS/OG** – 1 = annotated tumor suppressor; 2 = annotated oncogene; 3 = annotated as both tumor suppressor and oncogene; 0 = non-annotated gene;

4. **# cancer annot**. – number of annotation sources (major publications and databases where a given gene is assessed as a cancer gene [1,3-4])
5. **Samples** – total number of tumor samples used in derivation of statistics
6. **MM** - number of missense mutations
7. **TM** - number of truncating mutations
8. **SM** - number of silent mutations
9. **FIS≥2.0** - number of missense mutations scoring higher than 2.0 (predicted functional mutations [1]) ;
10. **FIS≥2.5** - number of missense mutations scoring higher than 2.5; (predicted high-scoring functional mutations [1])
11. **DD** - number of homozygous deletions
12. **DTFM** - number of truncating or predicted functional mutations affecting heterozygous deletions
13. **D** - number of heterozygous deletions
14. **A** - number of copy gains
15. **AA** - number of amplifications
16. **AFM** - number of predicted functional mutations in regions of copy gains
17. **NAFM** - number of predicted functional mutations in normal zygosity and regions of copy gains
18. **IndPD** =1/-1 , when P-value of over-representation of mutations in regions of heterozygous deletion is lower/higher than P-value of over-representation of mutations in copy-number-normal regions
19. **Pdtfm** - P value of statistical enrichment of truncating and predicted functional mutations in regions of heterozygous deletions
20. **IndPA** =1/-1 , when P-value of over-representation of mutations in regions of copy gains is lower/higher than P-value of over-representation in copy-number-normal regions.
21. **Pafm** - P value of statistical enrichment of predicted functional mutations in regions of copy gains

References

1. Reva B, Antipin Y, Sander C. **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Res*. 2011 Sep 1;39(17):e118.
2. Beroukhim *et al.*, **Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma**. *Proc Natl Acad Sci USA* 2007, **104:**20007-20012.
3. Higgins, M.E., Claremont, M., Major, J.E., Sander, C. and Lash, A.E. (2007) **CancerGene**s: **a gene selection resource for cancer genome projects**. *Nucleic Acids Res*, 35, D721-726.
4. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) **A census of human cancer genes**. *Nat Rev Cancer*, **4**, 177-183.
5. Getz G, Hoing H, Mesirov JP, Golub TR, Meyerson M, et al. (2007) **Comment on "the consensus coding sequences of human breast and colorectal cancers"**. Science 317: 1500.
6. Cerami *et al.,* **The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data**. Cancer Discovery. May 2012 2; 401.

Figure S28

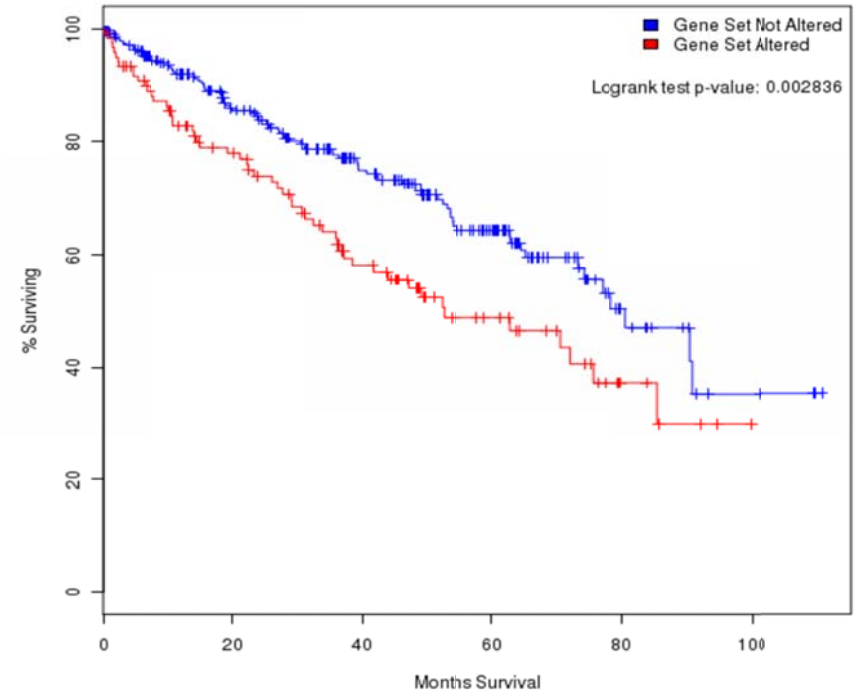


Figure S28. Rarely altered oncogenes are associated with poor outcome. The survival analysis and the survival plot are made using the Cancer Genomics Portal of MSKCC [6].

Figure S29

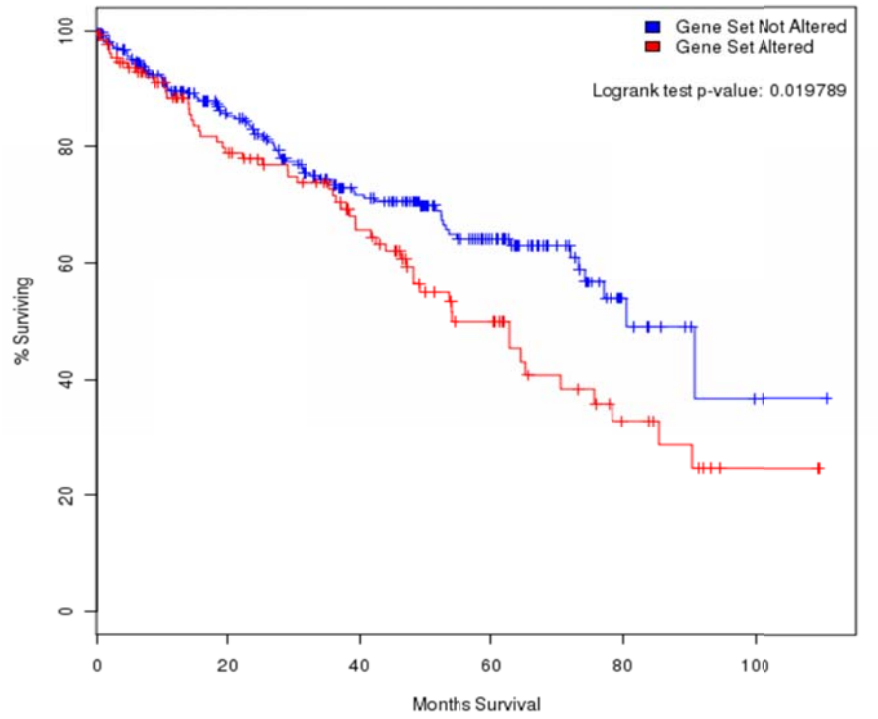


Figure S29. Rarely altered tumor suppressors are associated with poor outcome. The survival analysis and the survival plot are made using the Cancer Genomics Portal of MSKCC [6].

Figure S30

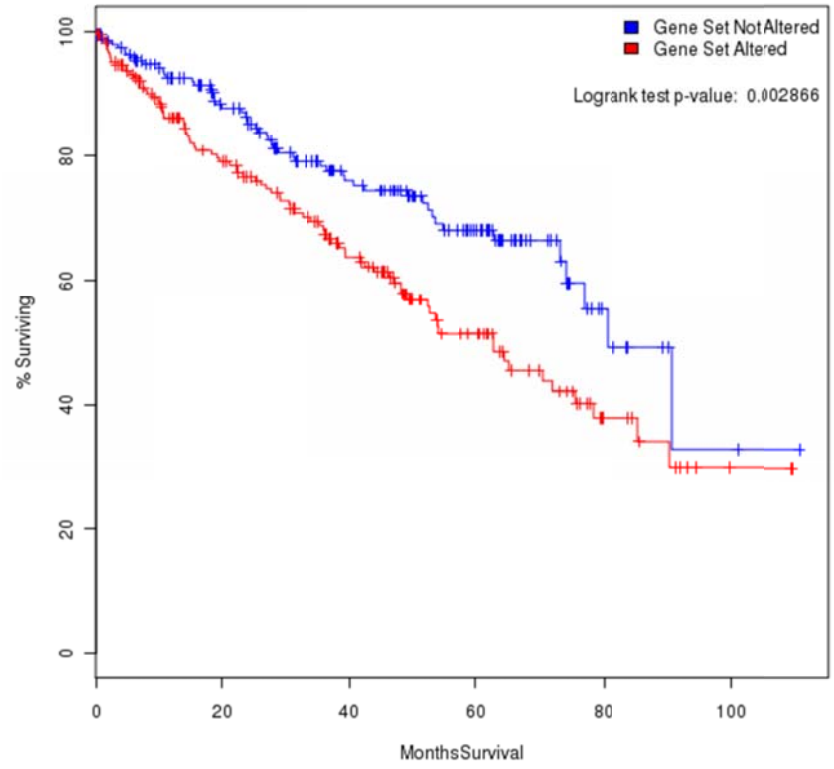


Figure S30. Rarely altered tumor suppressors and oncogenes are associated with poor outcome. The survival analysis and the survival plot are made using the Cancer Genomics Portal of MSKCC [6].

VI. RNA FUSIONS

Workgroup leader: Roel G.W. Verhaak (rverhaak@mdanderson.org)

Junior leader: Wandaliz Torres-Garcia

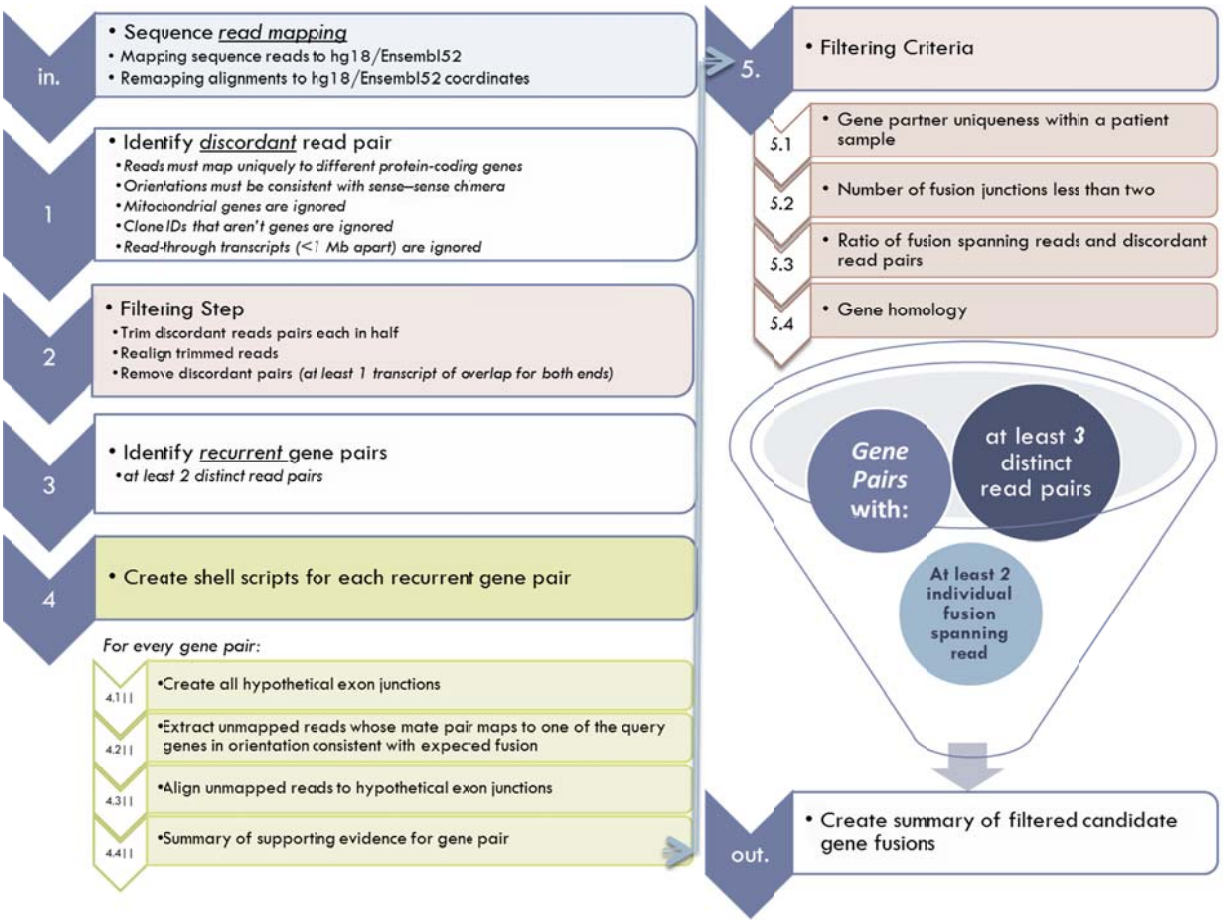
Contributors: Michael F. Berger, David Cogdell, Zhiyong Ding, Eric Jonasch, Hoon Kim, Victor Reuter, Rahul Vegesna, Wei Zhang

Methods for cDNA library construction. Total RNA for each sample was converted into a library of template molecules for sequencing on the Illumina Hiseq 2000 according to the protocol for the Illumina TruSeq RNA Sample Preparation Guide (Part # 15008136 Rev. A). In brief, poly-A mRNA was purified from total RNA (1 µg) and purified using poly-T oligo-attached magnetic beads. The mRNA was then fragmented and the first strand of cDNA was synthesized from the cleaved RNA fragments using reverse transcriptase and random primers. Following the synthesis of the second strand of cDNA using DNA Polymerase I and RNase H, end repair was performed on overhangs, followed by ligation of sequencing Adapters to the ends of the DNA fragments. The products were then purified and enriched with PCR to create the final cDNA library. The libraries were validated using a BioAnalyzer to assess size, purity and concentration of the purified cDNA libraries.

Methods for cDNA sequencing. The cDNA libraries were placed on an Illumina Cluster Station for single end cluster generation according to the protocol outlined in the Illumina Hiseq Analysis User Guide (Part# 11251649, RevA). The template cDNA libraries (1.5 µg) were hybridized to a flow cell, amplified and linearized and denatured to create a flow cell with ssDNA ready for sequencing. Each flow cell was sequenced on an Illumina HiSeq 2000 Genome Analyzer. Each sample underwent one lane of paired-end sequencing according to the protocol outlined in the Illumina HiSeq User Guide (Part# 11251649, RevA). After completion of the 50-cycle paired-end sequencing run (100 cycles total), bases and quality values were generated for each read by Bustard

Alignment and BAM file generation of RNA Sequencing Data. The "illumina2srf" tool from the DNA Sequence Read Toolkit (http://sourceforge.net/projects/sequenceread/) was used to convert the raw sequencing data from the vendor-specific format to standard SRF (sequence read format), which is a preferred compressed format for storing sequencing reads. Before data processing began, the "srf2fastq" tool from the Staden Package (http://staden.sourceforge.net/) was utilized to convert the data from SRF to the FASTQ format, using the "-C" parameter to simultaneously filter poor quality reads. A pre-processed lane of sequencing reads was then submitted to the BWA algorithm (http://bio-bwa.sourceforge.net/) for alignment against the reference transcript database using default parameters. The resulting SAM file is converted to BAM format using Picard tools (http://picard.sourceforge.net/).

Methods: PRADA: Pipeline for RNA Sequencing Data Analysis. Transcript fusions were identified from RNA sequencing data using the PRADA RNA-seq pipeline (http://sourceforge.net/p/prada/wiki/Home/). Paired end 50 bp mRNA sequence reads from 416 samples were aligned to a reference genome consisting of hg19 and the human transcriptome (Ensembl52 build) and hg18, using BWA with default settings [1]. Reads mapping to the transcriptome were remapping to hg18 genomic coordinates. This strategy allows the alignment of reads that span exon junctions while tolerating the detection of unannotated mRNAs.



PRADA: stringent approach for bona-fide fusion transcript candidates

The approach implemented to identify gene fusion candidates was previously discussed [2]. Fusion candidates were identified on the basis of (1) presence of at least three read pairs with the mates mapping to different genes with variable starting positions and no more than 1 mismatch (discordant read pairs); (2) identification of false positive discordant read pairs; (3) presence of at least two reads that span the putative fusion junction; (4) filtering of false positive fusion candidates using the following criteria: a) A gene cannot form a fusion transcript with multiple gene partners; b) No more than two possible fusion junctions were found; (c) Ratio of fusion spanning reads and discordant read pairs was less to 1.5; d) Maximum blastn homology bit score is less than 50.

STEP 1: Discordant read pairs

Discordant read pairs are identified as read pairs that maps uniquely (with mapping quality equal to 37) to different protein-coding genes with orientation consistent to sense-sense chimera; ignoring mitochondrial genes, clone IDs and possible read-through transcripts (subsequent gene pairs with the same transcription direction that are within 1Mb of each other). If a read maps to overlapping genes, these genes are split up as two different instances. Reads are not allowed to align with more than 1 mismatch and each of the discordant read pairs much map to different starting positions.

STEP 2: Mapping false positives

Discordant read pairs are split in half and realigned to the transcriptome as a measure to reduce false positives due to read sequences that are highly similar between the two genes annotated by the read pair or possible effects from poly-A sequences. If the split reads now overlap to at least one transcript the read pair is removed.

STEP 3: Fusion spanning reads

Aligning of unmapped reads to hypothetical exon junctions per recurrent gene pairs, using a custom, breakpoint specific sequence reference that includes all possible junctions matching the 3' end of gene A fused to the 5' end of gene B. All hypothetical exon junctions for a gene pair are created using Ensembl52 mapping information and transcript sequences as a customized fasta file for the specific recurrent gene pair. Furthermore, unmapped reads which mate pair maps to one of the recurrent genes in the pair are aligned to the hypothetical exon junctions' sequences using BWA with a maximum of four mismatches. This fourth step is performed for each recurrent gene pair in a parallel fashion to obtain fusion spanning reads supporting a fused gene pair.

STEP 4: Additional filtering criteria

To nominate gene fusion candidates for biological validation a series of rigorous filters are applied in this stage to reduce false positives. Transcript fusion candidates identified by at least three discordant read pairs, at least two fusion spanning reads and position consistent or partially consistent are further explored through a set of additional filters. Also, the positions of all discordant and fusion spanning reads are checked for position consistency. This step ensures that read pairs are outside the fusion junction or allowing one supporting read pair to flank one fusion spanning read. These filters are:

(a) Gene partner uniqueness within a sample

If any of the two genes in the possible fusion have multiple gene partners within the sample these are removed from further evaluation. For example, if a sample TCGA-XX-XXXX has two fusion candidate geneA-geneB and geneA-geneC; both candidates will be discarded.

Sample ID	GeneA	GeneB
TCGA-A3-3320-01A-02R-1325-07	ACLY	CD74
TCGA-B0-4846-01A-01R-1277-07	ACLY	HLA-B
TCGA-A3-3319-01A-02R-1325-07	ACLY	FN1
TCGA-A3-3319-01A-02R-1325-07	ACLY	GAPDH
TCGA-A3-3319-01A-02R-1325-07	ACLY	SPP1
TCGA-A3-3313-01A-02R-1325-07	ACLY	GPNMB
TCGA-A3-3320-01A-02R-1325-07	ACLY	ACTB
TCGA-A3-3320-01A-02R-1325-07	ACLY	FXYD2
TCGA-A3-3370-01A-02R-1420-07	ACLY	CD74
TCGA-CJ-4634-01A-02R-1325-07	ACLY	CA12
TCGA-A3-3319-01A-02R-1325-07	ACLY	ACTB
TCGA-A3-3319-01A-02R-1325-07	ACLY	AHNAK
TCGA-A3-3319-01A-02R-1325-07	ACLY	EEF2
TCGA-A3-3319-01A-02R-1325-07	ACLY	FAT1
TCGA-A3-3319-01A-02R-1325-07	ACLY	PROM2
	removed	

Example of gene fusion candidate that is discarded because of the gene partners uniqueness within a sample filter

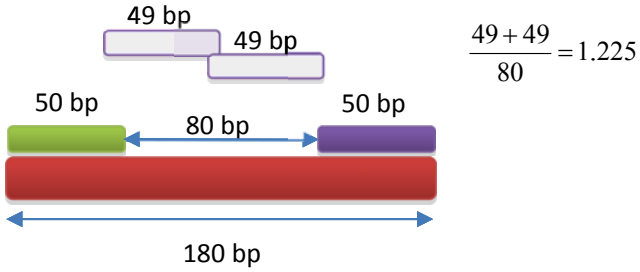
(b) Number of fusion junctions less than two

The number of fusion junctions is determined from the unmapped reads mapping to hypothetical exon junctions in the gene pair. Multiple fusion junctions could be a result of chaos behavior in these genes and not necessarily feature a specific fusion mechanism. The value of four is determined from preliminary analysis of RNAseq samples from Acute Myeloid Leukemia samples showing known fusion genes such as PML-RARA and BCR-ABL1 exhibiting values in the range of 1 to 4 (data not shown).

(c) Ratio of fusion spanning reads and discordant read pairs

The ratio of the number of fusion spanning reads over the number of discordant read pairs is dependent on the average size of the DNA fragments and the size of the sequence reads as shown in the figure below. With a median insert size of 180 bp and a read length of 50 bp, the expected ratio was

estimated to approximate 1.225. Based on this ratio the window of ratios that is accepted was set to [0, 1.5].



Ratio of fusion spanning reads and discordant read pairs.

(d) Gene homology

The homology of two genes is assessed using blastn and the bit-score indicating gene similarity cannot exceed 50.

Fusion transcript validation. To test the accuracy of the fusion identification pipeline, thirteen fusion transcripts from nine samples were selected for validation using other technologies. cDNA from seven of nine samples was obtained to perform validation experiments on putative fusions. One *SFPQ-TFE3* fusion transcript was included in validation from cDNA.

Reverse Transcription – Polymerase Chain Reaction (RT-PCR) based validation

Nested primers (Sigma-Genosys) were designed using breakpoint information collected through PRADA fusion spanning reads output and tools such as *PrimerBlast* (Table S10). RT-PCR was applied to validate the *FAM172A-FHIT* and *TFE3*-related fusion, as well as four additional fusions (*C6orf106-LRRC1*, *CYP39A1-LEMD2*, *ZNF193-MRPS18A* and *FTSJD2-GPX6*) potentially associated with a local chromotripsis event in TCGA-A3-3313.

The RT-PCR protocol performed for *FAM172A-FHIT* and *TFE3* related fusions used random hexamers to prime a Superscript II (cat 18064) based reverse transcription reaction using 250 ng RNA with 50 ng of random primers into cDNA as per Superscript II (18064-022) protocol (Invitrogen, Carlsbad, CA) as indicated by the manufacturer. 1 ul of cDNA was then used as template for hot-start PCR reaction using AmpliTaq Gold (N8080247) protocol (Applied Biosystems, Foster City, CA) using primers designed to the putative fusion genes and RNA from LN229 cell line (glioblastoma), HCT116-/- cells (colon cancer), and HEK-293 cells (kidney) was used as negative controls. Resulting PCR products were analyzed by 1% agarose gel electrophoresis with ethidium bromide staining. Additionally, Sanger sequencing was performed on positive PCR products using nested primers to confirm the fusion breakpoint.

The four candidate fusion transcripts, which suggested a local chromothrypsis event, were validated by the following RT-PCR protocol. cDNA was synthesized with 1µl of TCGA sample mRNA (150ng) as templates and oligo (dT)20 as primers using SuperScriptIII First-Strand Synthesis System from Invitrogen (Carlsbad, CA) according to the manufacturer's protocol. RNA from 786-0 cells (kidney cancer), HCT1187 cells (breast cancer), and OVISE cells (ovarian cancer) was used as negative controls. Touchdown PCR was performed for increased specificity and sensitivity using AccuPrime PCR Systems from Invitrogen. The touchdown phase of the PCR cycles started with 70°C as an annealing temperature for 45 seconds with a decrease of 0.5°C per cycle till 56°C was reached. Ten more cycles of amplification with 56°C as annealing temperature were included as the second phase of

the touchdown PCR. Denaturation at 94°C for 45 seconds and elongation at 70°C for 60 seconds were used for all cycles.

Primers designed for the fusion candidates

Sample ID	5' Gene	3' Gene	Primer	Primer	Fusion product size
TCGA-AK-3456	TFE3	SFPQ	GCAGTGCTAGCTCCATGGCT	AAGCCACTGCACTCCAGCCT	625 bp
TCGA-AK-3456	SFPQ	TFE3	CGTTTTGCCCAGCATGGCACG	GTTCCCGACGCTCACGCCTC	553 bp
TCGA-A3-3313	C6orf106	LRRC1	CTGGCCTCCAGGGGTTTGTC	GTCTCATTGTCTTCTTCATCTTTCTC	324 bp
			CTGGCCTCCAGGGGTTTGTC	CAGAAGGCAGCTGAGGAAGTAAG	347 bp
TCGA-A3-3313	CYP39A1	LEMD2	TCTGCTTCTGGAAGGTGCTGG	GCTATATATTCTTGGGCTTCC	283 bp
TCGA-B2-4101	FAM172A	FHIT	CGAAGTATGTATATGAGCTCCTGG	TCAAACTGGTTGGCAATAGC	404 bp
TCGA-AK-3445	SOGA2	LRRC41	GGGAGAGGGCACGCCAGTGAAG	TTGAGGGCAGTTTCCTCAATCCTCTC	281 bp
TCGA-B0-5095	GORASP2	WIPF1	GGCTCTCCGGCGGCAGCGAG	TGCAAACGTCGGGGGCGGCG	266 bp
TCGA-A3-3313	ZNF193	MRPS18A	GATGGTGGCCCACAGACACAGAC	ACCTGCTCGGTGGGCCATCTTC	276 bp
TCGA-A3-3313	FTSJD2	GPX6	CGATGAAGAGGAGAACTGACCCAG	CAGGATACTGAGCTGCCAAGCCTC	290 bp
TCGA-B0-4945	KIAA0427	GRM4	TGCTAGCCCCTGCCAGCCTAT	GGCCTCCACACCGCTCTCAC	660 bp
TCGA-B8-4143	SLC36A1	TTC37	TGCTTTTGGTTTGTTGGAAGGGACT	TGCCAAAGAGCCCCAGCTTTTGAT	769 bp

Results of RT-PCR validation

Nine of eleven fusion candidates were validated using RT-PCR (Table S10). Among the validated fusion transcripts, *FAM172A-FHIT* was associated with a partial deletion of *FHIT*. Both the sense and anti-sense product of the X(p11) associated *SFPQ-TFE3* fusion were confirmed. The four fusion transcripts that were predicted from a localized region on chromosome 6 in a single sample all validated, suggesting the applicability of RNA-sequencing to identify chromotripsis events.

Fusion transcripts. Using the PRADA pipeline, 80 fusion transcripts in 62 samples (of 416 analyzed samples) were identified. Of 80 predicted fusions, 57 were intrachromosomal whereas 33 were fusions between different chromosomes. Four recurrent fusions were found: *SFPQ-TFE3* (n=5, chr1-chrX), *DHX33-NLRP1* (n=2, chr2), *TRIP12-SLC16A14* (n=2, chr17) and *TFG-GRP128* (n=4, chr3). *TFG-GRP128* has been previously reported to occur in lymphoma, as well as in normal tissues [10].

Recurrent fusions. Five samples with fusion of *SFPQ-TFE3* were identified, which has been previously been related to translocation-associated renal carcinoma. This subtype of renal carcinoma has been linked to differences in pathology and pediatric renal carcinoma. Slides from three SFPQ-TFE3 samples were selected for confirmatory pathology review using immunohistochemistry (Figure S35) which showed the presence of TFE3 protein in all three cases; a clear cell renal carcinoma histology in two of the three cases; and the presence of a genomic rearrangement in the single case that was analyzed using FISH. TFE3-associated renal cancers may thus represent a minor subset of clear cell renal cancer (~1%) that is missed during routine pathology review. The importance of the TFE3 translocation is suggested by the lack of mutations in the most significantly mutated genes in the five TFE3 cases (VHL: 1 of 5 samples; PBRM1, SETD2, BAP1, MTOR, PIK3CA, ARID1A, ATM, PTEN, KDM5C: 0 of 5 samples). Additionally, there was a relative lack of chromosome alterations (del(3p), amp(5q): 2 of 5 samples; del(14q), del(9p), del(6q): 0 of 5 samples).

The second most frequent recurrent fusion (four samples or 1% of the data set) was the result of an intrachromosomal inversion between the proximal genes TFG and GPR128. All four cases harbored a 3p deletion and a 5q amplification, but again a lack of driver mutations with three cases harboring no mutation in from the significantly mutated gene list.

Discussion. Through the study of genomic rearrangements of TCGA kidney clear cell carcinoma samples of RNA-seq data using PRADA, a pipeline for RNA-Sequencing data analysis, a total of 62 samples with RNA material were available for validation purposes. *SFPQ-TFE3* was found in five samples out of >400 TCGA kidney clear cell carcinoma dataset (0.7%). These have been reported in the kidney cancer literature [6,7] with association to *TFE3* translocation[8] as a rare subtype of kidney cancer. *TFE3* translocations have been linked to a rare subtype of renal cancer[8]. Identification of gene fusions in kidney clear cell carcinoma elucidates interesting patterns in the biology of this disease that may aid development of targeted treatment regimens.

References:

1. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]
2. Berger M.F., Levin, J. Z., Vijayendran, K. (2010) Integrative analysis of the melanoma transcriptome. Genome Res. 20:413-427.
3. Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews. Genetics, 11, 685–696.
4. Chin L., Andersen J.N., Futreal P.A. (2011) Cancer genomics: from discovery science to personalized medicine. Nat Med. 17(3):297-303.
5. Garber, M. , Grabherr, M. G., Guttman, M., and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods 8, 469–477.
6. Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12, 443-451.
7. Mitelman, F., Johansson, B., and Mertens, F. (2007) The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer 7, 233-245
8. Clark, J., Lu, YJ..; Sidhar, S.K., Parker, C., Gill, S.,et al. (1997) Fusion of splicing factor genes PSF and NonO (p54nrb) to the TFE3 gene in papillary renal cell carcinoma. Oncogene, 15(18):2233-9.
9. Malouf, G.G., Camparo, P., Molinie, V., et al. (2011) Transcription factor E3 and transcription factor EB renal cell carcinomas: clinical features, biological behavior and prognostic factors. J Urol.;185:24–29.
10. Chase A, Ernst T, Fiebig A, Collins A, Grand F, Erben P, Reiter A, Schreiber S, Cross NC. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. Haematologica. 2010 Jan;95(1):20-6

Table S10

Table S10. Eleven fusion candidates resulting from the gene fusion module with PRADA from the validation set.

Sample ID	5' Gene	3' Gene	Discordant Read Pairs	Fusion Span Reads	Fusion Junction (s)	5' Gene Chr	3' Gene Chr	Validated?
TCGA-AK-3456-01A-02R-1325-07	TFE3	SFPQ	175	129	1	chrX	chr1	Yes
TCGA-AK-3456-01A-02R-1325-07	SFPQ	TFE3	116	81	1	chr1	chrX	Yes
TCGA-A3-3313-01A-02R-1325-07	C6orf106	LRRC1	90	40	2	chr6	chr6	Yes
TCGA-A3-3313-01A-02R-1325-07	CYP39A1	LEMD2	37	9	1	chr6	chr6	Yes
TCGA-B2-4101-01A-02R-1277-07	FAM172A	FHIT	17	4	1	chr5	chr3	Yes
TCGA-AK-3445-01A-02R-1277-07	KIAA0802	LRRC41	14	6	1	chr18	chr1	Yes
TCGA-B0-5095-01A-01R-1420-07	GORASP2	WIPF1	14	2	1	chr2	chr2	Yes
TCGA-A3-3313-01A-02R-1325-07	ZNF193	MRPS18A	11	3	1	chr6	chr6	Yes
TCGA-A3-3313-01A-02R-1325-07	FTSJD2	GPX6	9	8	1	chr6	chr6	Yes
TCGA-B0-4945-01A-01R-1420-07	KIAA0427	GRM4	8	5	1	chr18	chr6	No
TCGA-B8-4143-01A-01R-1188-07	SLC36A1	TTC37	5	5	1	chr5	chr5	No

Figure S31

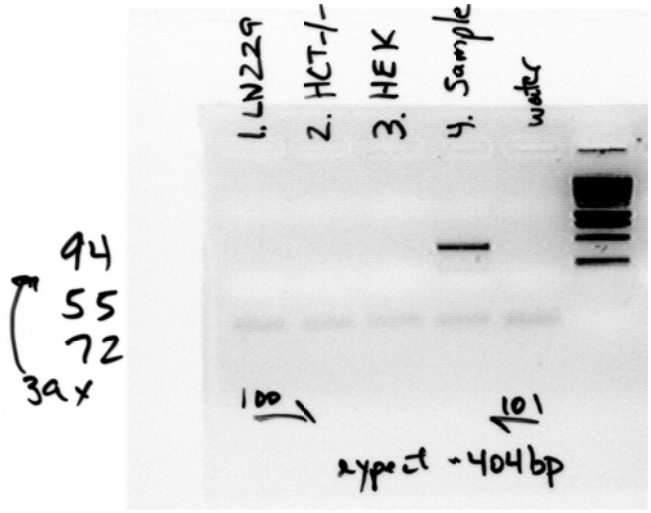
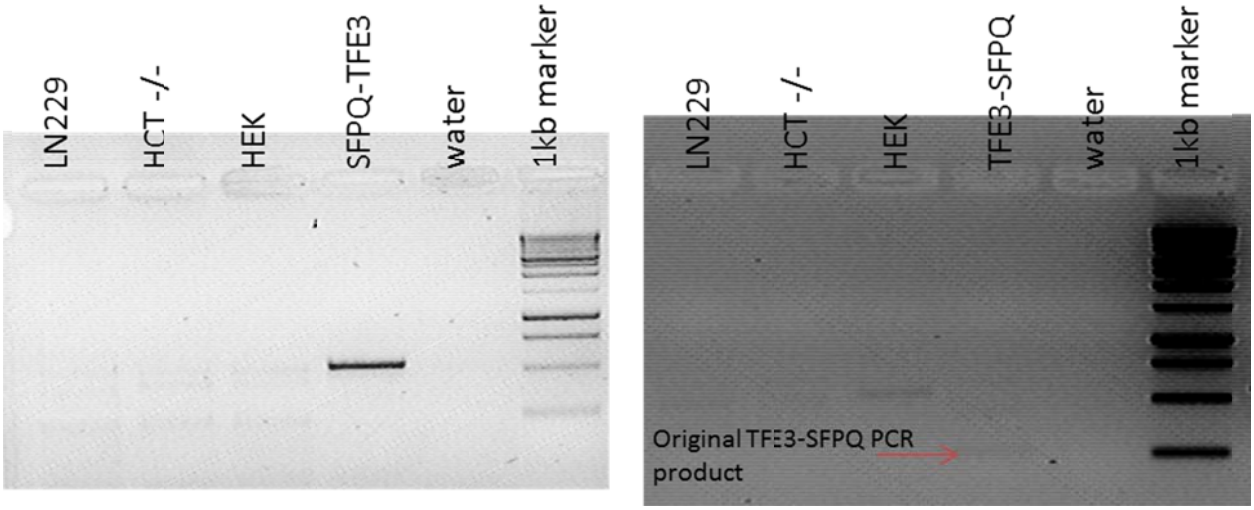


Figure S31. RT-PCR results for *FAM172A-FHIT* fusion validations for sample TCGA-B2-4101.

Figure S32



(a) (b)

Figure S32. RT-PCR results for *TFE3* fusion validations for sample TCGA-AK-3456.

Figure S33

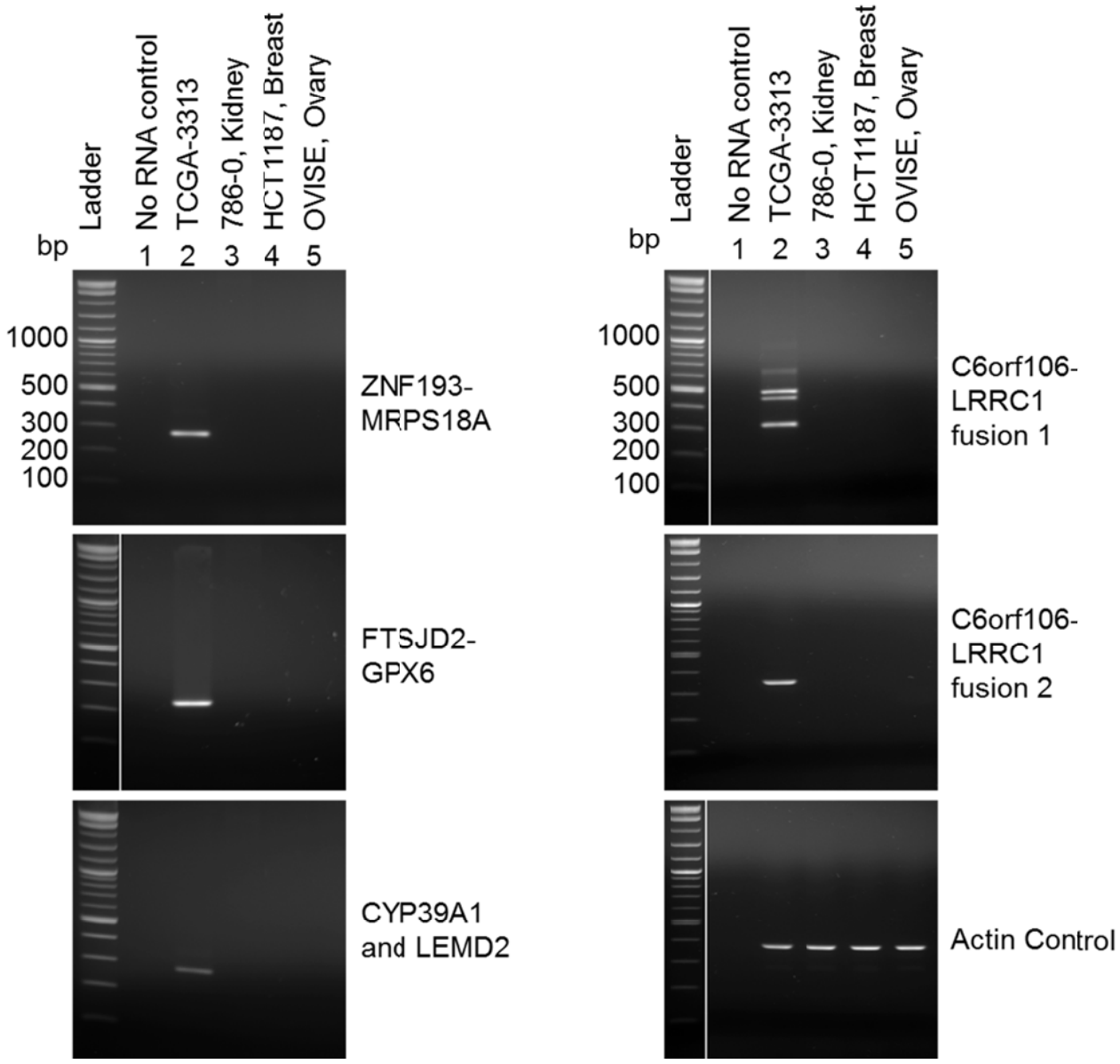


Figure S33. RT-PCR results for fusion validations for sample TCGA-A3-3313.

Figure S34

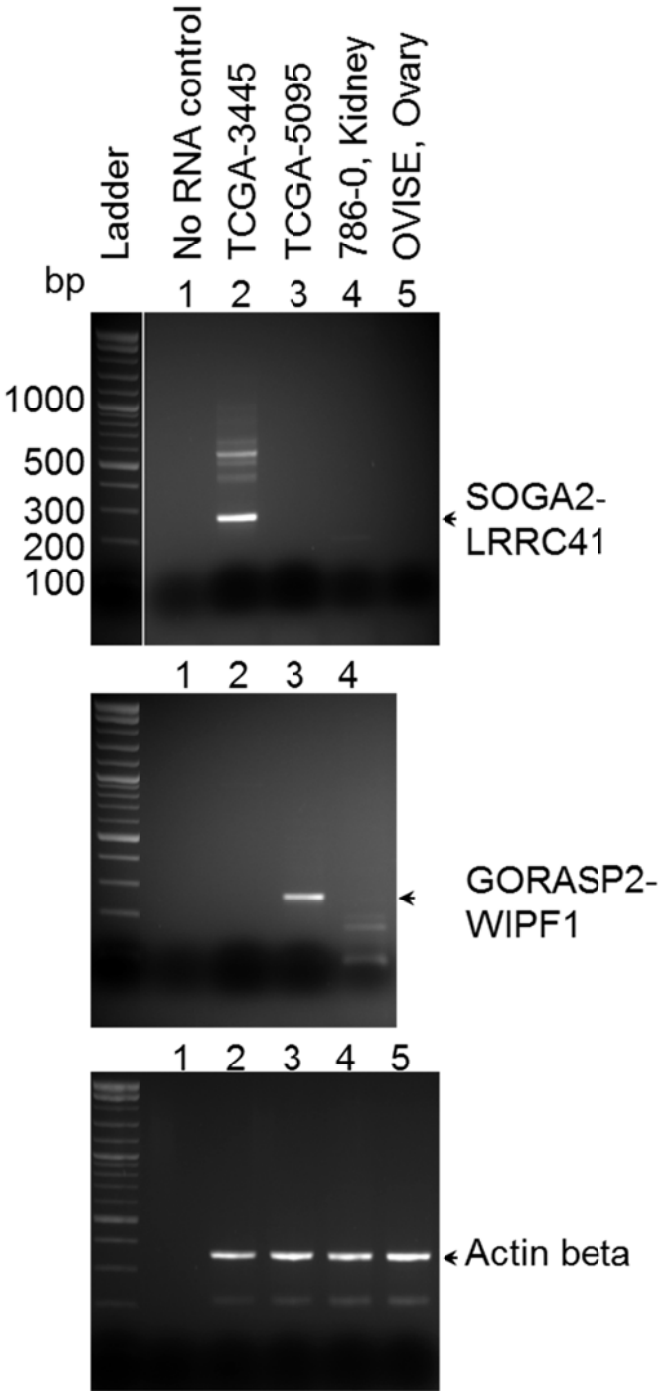


Figure S34. RT-PCR results for fusion validations for samples TCGA-AK-3445 (SOGA2-LRRC41) and TCGA-B0-5095 (GORASP2-WIPF1).

Figure S35

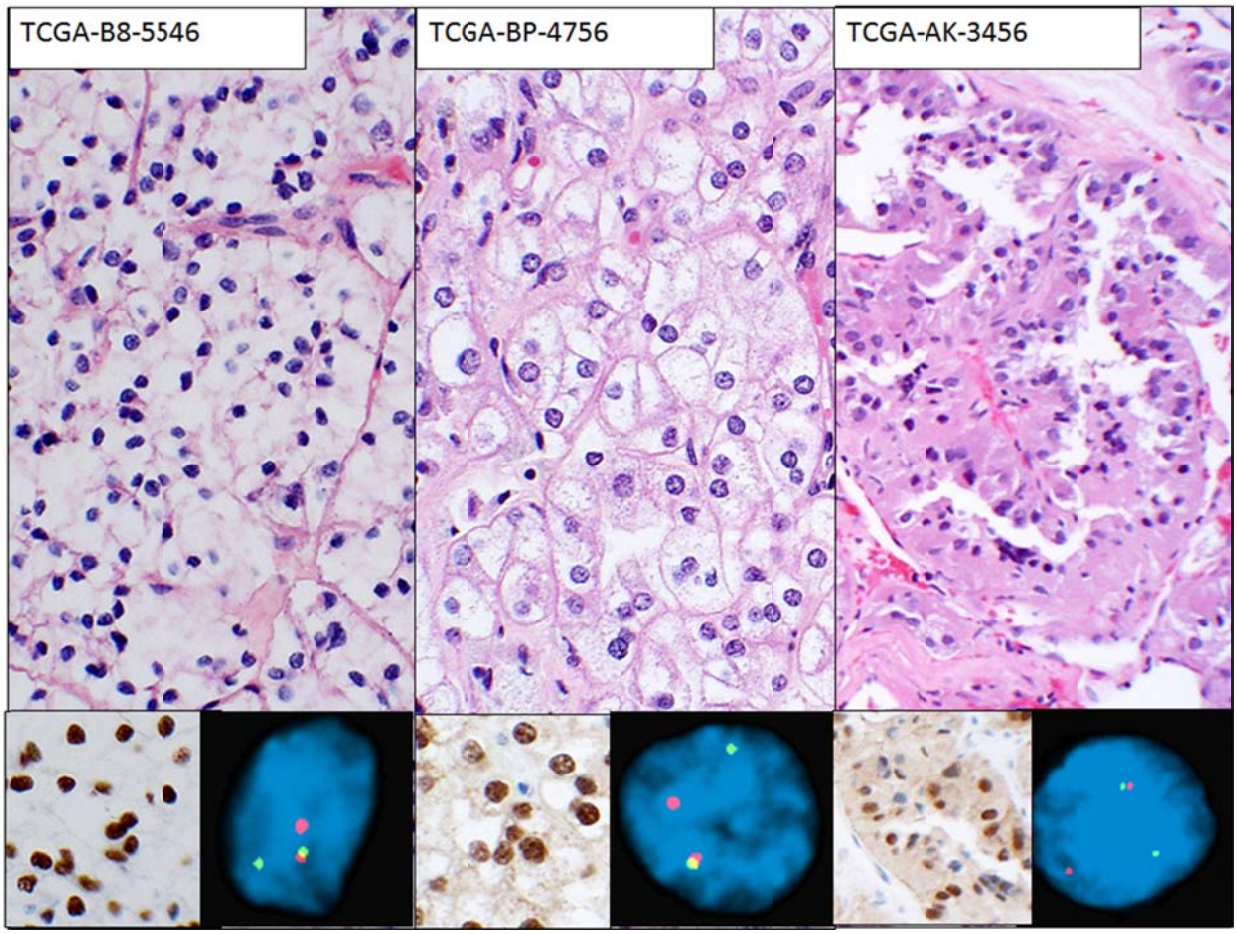


Figure S35. FISH and IHC confirming TFE3 gene fusions. Pathologic evaluation limited in 2 of 3 cases. Although morphologic variability is present, at least two cases have unequivocal clear cell/acinar areas with appropriate vascularity. TFE3 protein expression and rearrangement by FISH is confirmed in all cases.

VII. METHYLATION STUDIES

Workgroup leaders: Hui Shen (shenhui1986@gmail.com) and Peter Laird (plaird@usc.edu)

Array-based DNA methylation assay. We used two Illumina Infinium DNA methylation platforms, HumanMethylation27 (HM27) BeadChip and HumanMethylation450 (HM450) BeadChip (Illumina, San Diego, CA) to obtain DNA methylation profiles of 502 TCGA clear cell renal carcinoma samples and 359 adjacent non-tumor kidney tissue samples. Twelve technical replicates were also included in the assay to monitor technical variations, with six on the HM27 platform and six on the HM450 platform. We included 444 of the tumor samples and all normal samples in the 'extended' list used for analyses based on DNA methylation data only, and 373 of the 444 tumor samples in the 'core' list used for cross-platform comparisons. The Infinium HM27 array targets 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36). The Infinium HM450 array targets 482,421 CpG sites through out the genome and covers 99% of RefSeq genes. It 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 on both Infinium DNA methylation platforms can be found in the MAGE-TAB ADF (Array Design Format) file available through the TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga/).

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 described previously [1]. 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.

The level of DNA methylation at each CpG locus is summarized as a beta (β) value calculated as (M/(M+U)), ranging from 0 to 1, which represents the 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 also calculated with the methylumi package, 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 2 and 3 in HM27 and Level 3 in 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.

HM27: *Level 1* - Level 1 data packages contain the non-background corrected signal intensities of the M and U probes and the mean negative control cy5 (red) and cy3 (green) signal intensities. A detection p-value for each data point, the number of replicate beads for M and U probes as well as the standard error of M, U, and control probe signal intensities are also provided. It is important to note that for some CpG targets, both M and U measurements will be cy3, and for others both will be cy5. To resolve ambiguities regarding this subtlety of the Infinium DNA Methylation assay, we have labeled the cy3 and cy5 values deposited to the DCC as “Methylated Signal Intensity” and “Unmethylated Signal Intensity”. The information of the color channel for each CpG locus is contained in the MAGE-TAB ADF file deposited in the DCC. *Level 2* - Level 2 data files contain the β-value calculations for each probe and sample. Data points with detection p-values >0.05 were not considered to be significantly different from background, and were masked as “NA”. *Level 3* - Level 3 data contain β-value calculations, HUGO gene symbol, chromosome number and genomic coordinate for each targeted CpG site on the array. In addition, we masked data points with "NA" from the probes that 1) contain known single nucleotide polymorphisms (SNPs) after comparison to the dbSNP database (Build 132), 2) contain repetitive sequence elements that cover the targeted CpG locus in each 50 bp probe sequence, 3) are not uniquely aligned to the human genome (NCBI build 36.1) at 20 nucleotides at the 3’ terminus of the probe sequence, 4) span known regions of small insertions and deletions (indels) in the human genome (dbSNP build 130).

HM450: *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 of the interrogated locus (HM27 carryover or recently discovered). The genomic characteristics for each probe are available for download via Illumina (www.illumina.com).

The following data archives were used for the analyses described in this manuscript.

[1] "jhu-usc.edu_KIRC.HumanMethylation27.Level_3.1.3.0"
[2] "jhu-usc.edu_KIRC.HumanMethylation27.Level_3.2.3.0"
[3] "jhu-usc.edu_KIRC.HumanMethylation27.Level_3.3.3.0"
[4] "jhu-usc.edu_KIRC.HumanMethylation27.Level_3.4.3.0"
[5] "jhu-usc.edu_KIRC.HumanMethylation27.Level_3.5.3.0"
[6] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.1.4.0"
[7] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.2.4.0"
[8] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.3.4.0"
[9] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.4.4.0"
[10] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.5.4.0"
[11] "jhu-usc.edu_KIRC.HumanMethylation450.Level_3.6.4.0"

Merging HM27 and HM450 Data. The shared probe set between HM27 and HM450 platforms (n=25,978) were used for the analysis. Out of the 25,978 probes, 887 probes were masked due to detection p-value, repeats and SNPs and non-uniquely mapped probes (n=25,091 remaining). We observed batch and platform specific effects with the

technical replicates. To alleviate systematic platform-specific effects (dye bias, background level, etc.) we fitted a LOESS regression model between the two platforms using the twelve technical replicates. We normalized the HM450 data against the HM27 data with this fitted model on the M values, stratified by the number of CpGs in the probe (CpG=1,2,3,4,5,6+). M value is the log 2 ratio of Methylated (M) intensity and Unmethylated (U) intensity and better satisfies the linearity assumption. The M values were then transformed back to beta values with the equation Beta=2^M/(2^M+1). In order to further remove probes that were not fitted well with the LOESS model, as well as probes prone to technical variation other than platform-specific effects, we calculated probe-wise standard deviation across the twelve technical replicates and masked any probe with a probe-wise standard deviation of greater than 0.1 (n=24,383 remaining).

Global Hypermethylation and Clinical Stage/Grade. We investigated all CpG loci (n=16,123) assayed on both HM27 and HM450 platforms that are unmethylated in the normal adjacent kidney tissue (average DNA methylation beta value <0.2) for cancer-specific hypermethylation. We further excluded 1,021 loci methylated in the normal white blood cells (average DNA methylation beta value > 0.2) to avoid 'passive' hypermethylation signature due to blood contamination or lymphocyte infiltration at those loci. We calculated the percentage of hypermethylated (beta value>0.2) loci. Boxplots of this percentage for normal and tumors of different stages were used to visualize the trend of increasing DNA hypermethylation with advancing stages and grade.

Epigenetic Silencing Calls. Again, only intersect probes (n=25,091) of HM27 and HM450 were used for this analysis. For each gene, we chose DNA methylation probes that satisfy the following criteria:

1. *The locus studied should be unmethylated in the normal kidney tissue*: 95th percentile for methylation in normals <0.2; we use the 95th percentile instead of the maximum to allow for field effects in 5% of the normal;
2. *DNA methylation at the locus studied should be inversely correlated with expression level of the gene*: correlation coefficient with log2(RPKM+1) < -0.2, and adjusted p-value testing for correlation < 0.005;
3. *The locus studied should be methylated in some of the tumors*: 95th percentile for DNA methylation beta value in tumor > 0.2;
4. *The hypermethylation level in the tumor should be considerable*: maximum DNA methylation beta value in the tumor > 0.5.

Any gene with at least one such CpG locus detected was called epigenetic silenced. Then, for each sample, we looked at all probes that satisfy the above criteria for each silenced gene. A sample is called silencing if it satisfies the following criteria:

1. *Overall hypermethylation:* the mean methylation across those loci > 0.2;
2. *Consistency across various loci*: DNA methylation at each CpG locus uniformly > 95th quantile (normal).

DNA Methylation Pattern Changes Associated with *SETD2* mutation. We used a univariate two-sample t-test to evaluate whether DNA methylation level at each CpG locus investigated at the HM450 platform was different in the *SETD2* mutants (n=32) and wildtype tumors (n=192). The p-values were then corrected for multiple comparisons using the Benjamini-Hochberg procedure. The adjusted p-value was plotted against the difference between the mean beta value in *SETD2* mutants and mean beta value in *SETD2* wildtype tumors (volcano plot). A heatmap was used to visualize a subset of the

loci (absolute difference in beta value > 0.1, adjusted p-value <0.001) at which *SETD2* mutants were significantly differently methylated from the wildtype tumors. Roadmap (http://nihroadmap.nih.gov/epigenomics/) human adult kidney H3K36me3 ChIP-Seq data were downloaded from GEO (Accession: GSM773000) and the number of reads overlapping with each of the loci in the heatmap was plotted as a rowside color bar.

Statistics. All statistical analyses were conducted in R version 2.15.0 (2012-03-30). All *p*-values reported were two-sided.

References

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

VIII. EXPRESSION STUDIES

Workgroup leader: Roel Verhaak (rverhaak@mdanderson.org)
Junior leader: A. Rose Brannon (brannona@mskcc.org)
Contributors: Wandaliz Torres Garcia, Suzanne S. Fei, Chad Creighton, W. Kim Rathmell

Methods for cDNA library construction and sequencing. Identical to the above (RNA fusions section).

RNA Sequencing Expression Workflow. Gene expression was quantified based on the gene models defined in the TCGA Gene Annotation File (GAF) [1]. Gene expression was quantified by counting the number of reads overlapping each gene model's exons and converted to Reads per Kilobase Mapped (RPKM) values by dividing by the transcribed gene length, defined in the GAF and by the total number of reads aligned to genes as previously described [2]. In parallel, each lane of sequencing was assessed for a variety of pre- and post- alignment quality control measures as previously described [2].

Gene filtering and NMF clustering. Level3 RNA-seq RPKM data for samples in the extended set of the 1.4 data freeze (n=417) were retrieved. Gene expression values were Z-score transformed by subtracting the average, then dividing by standard deviation. The maximum absolute deviation (maximum value minus the average) was calculated for each gene and the top 1,500 genes were selected for clustering. The data was then transformed into a non-negative matrix and clustered using non-negative matrix factorization (NMF)[3]. Primary clustering of the tumors was run in NMF for 200 iterations of ranks 2-8, with default settings of method brunet and seed random. Rank estimates were calculated using 50 iterations of ranks 2-8. The number of subtypes was selected by a combination of a) the major increase in cophenetic coefficient between k=3 and k=4, and the subsequent drop-off (Figure S36) b) visual inspection of the consensus clustering matrices and c) visual inspection of the correlation matrices (Figure S36). The four subtypes contained respectively 147 (m1), 90 (m2), 94 (m3) and 86 (m4) samples.

Selection of top differential gene features for heat map display. For Figure 3A, an example set of top differential genes distinguishing the subtypes was selected. Given the four mRNA-based tumor subtypes, we computed the two-sided t-test for each gene, comparing each subtype with the rest of the tumors; this was carried out four times, once for each subtype. For each gene, the p-value selected was for the subtype having highest expression compared to others (as denoted by lowest p-value), and the top 500 genes with the lowest p-value were selected for the expression heat map.

Single sample gene set enrichment analysis. Gene sets were downloaded from the MSigDb (http://www.broadinstitute.org/msigdb), collections C2 and c5 (version 3.0). These two collections were augmented with additional "combined" signatures for the ones that have both up-regulated (UP) and down-regulated (DN) versions produced a total of 5,525 gene sets. Single sample gene set enrichment analysis was performed as described previously [4].

Subclass Signature Gene and Gene Set Identification. The significance analysis of microarrays (SAM) method was used to identify marker genes and gene sets of each subtype. Each class was compared to the other three classes combined [5]. Both rank order and test statistic for all of these analyses are provided to allow independent confirmation of the findings on future analyses and data sets. For marker gene set identification, SAM was applied on gene set activation scores. The top 200 most strongly and uniquely associated genes were included in the class signature. Signature genes could be either specifically down- or upregulated. Fold change, SAM F-score and q-value for each gene

and each mRNA subtype are included in Data File S5. SAM F-score and q-value for each MSigDb gene set and each mRNA subtype are included in Data File S5.

Correlation with Copy Number. Using GISTIC2.0, 62 lesions with significant copy number loss or gain were identified. Each lesion was associated with a specific set of samples harboring the copy number change. Association of copy number alterations with subtype was determined by comparing each subtype versus the remaining three subtypes using a two-tailed Fisher's exact test and using the Hochberg method implemented in p.adjust (R Development Core Team, 2008) for controlling the Family-wise Error rate. The frequencies of GISTIC lesions per mRNA subtype and p-values are shown in Data File S5.

Subtype specific mutation analysis. Somatic mutations in 428 samples (RPKM expression data available for 374 of 428) were established using methods described in the Mutations section (see above); all mutations identified by at least two different methods and occurring in five or more samples were included in the expression subtype analysis (n=3,013). Association of somatic mutations with subtype was determined by comparing each subtype versus the remaining three subtypes using a two-tailed Fisher's exact test and using the Hochberg method implemented in p.adjust (R Development Core Team, 2008) for controlling the Family-wise Error rate. The frequencies of GISTIC lesions per mRNA subtype and p-values are shown in Data File S5.

References

1. http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf
2. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012, 487:330-7.
3. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010, 11:367.
4. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010, 17:98-110.
5. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001, 98:5116-21.
6. Brannon AR, Reddy A, Seiler M, Arreola A, Moore DT, Pruthi RS, Wallen EM, Nielsen ME, Liu H, Nathanson KL, Ljungberg B, Zhao H, Brooks JD, Ganesan S, Bhanot G, Rathmell WK. Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. Genes Cancer. 2010 Feb 1;1(2):152-163.

Figure S36

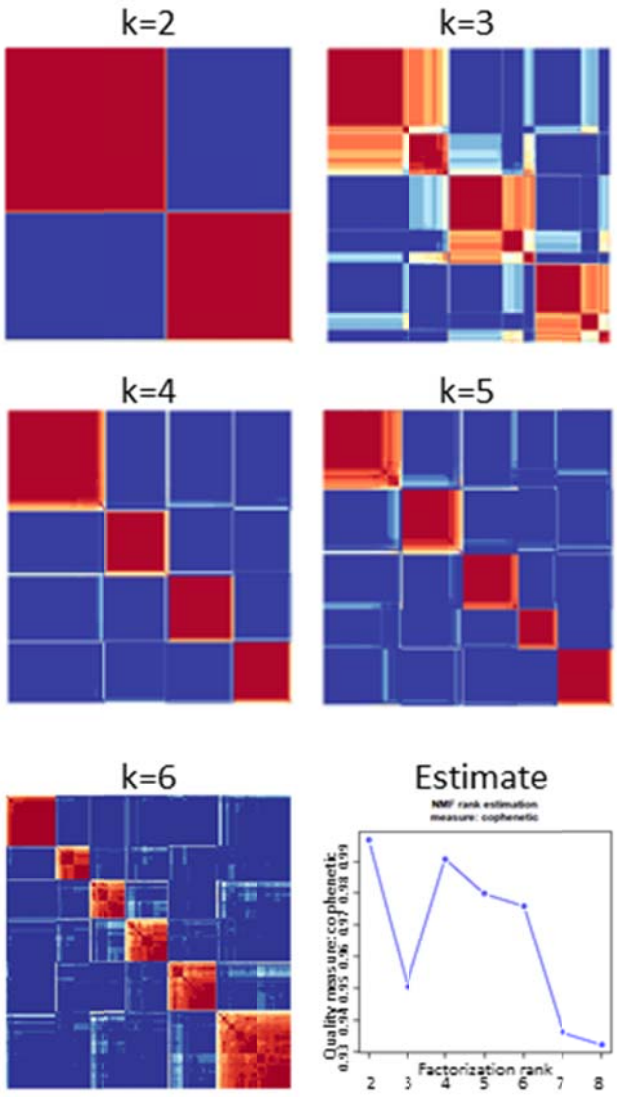


Figure S36. CNMF clustering of 1,500 variably expressed genes and 417 TCGA ccRCC samples. Consensus matrices are shown for clustering with k=2 to k=6. The cophenetic coefficient shows high values for both k=2 and k=4. Clustering with k=4 shows four robust clusters with limited overlap between clusters.

Figure S37

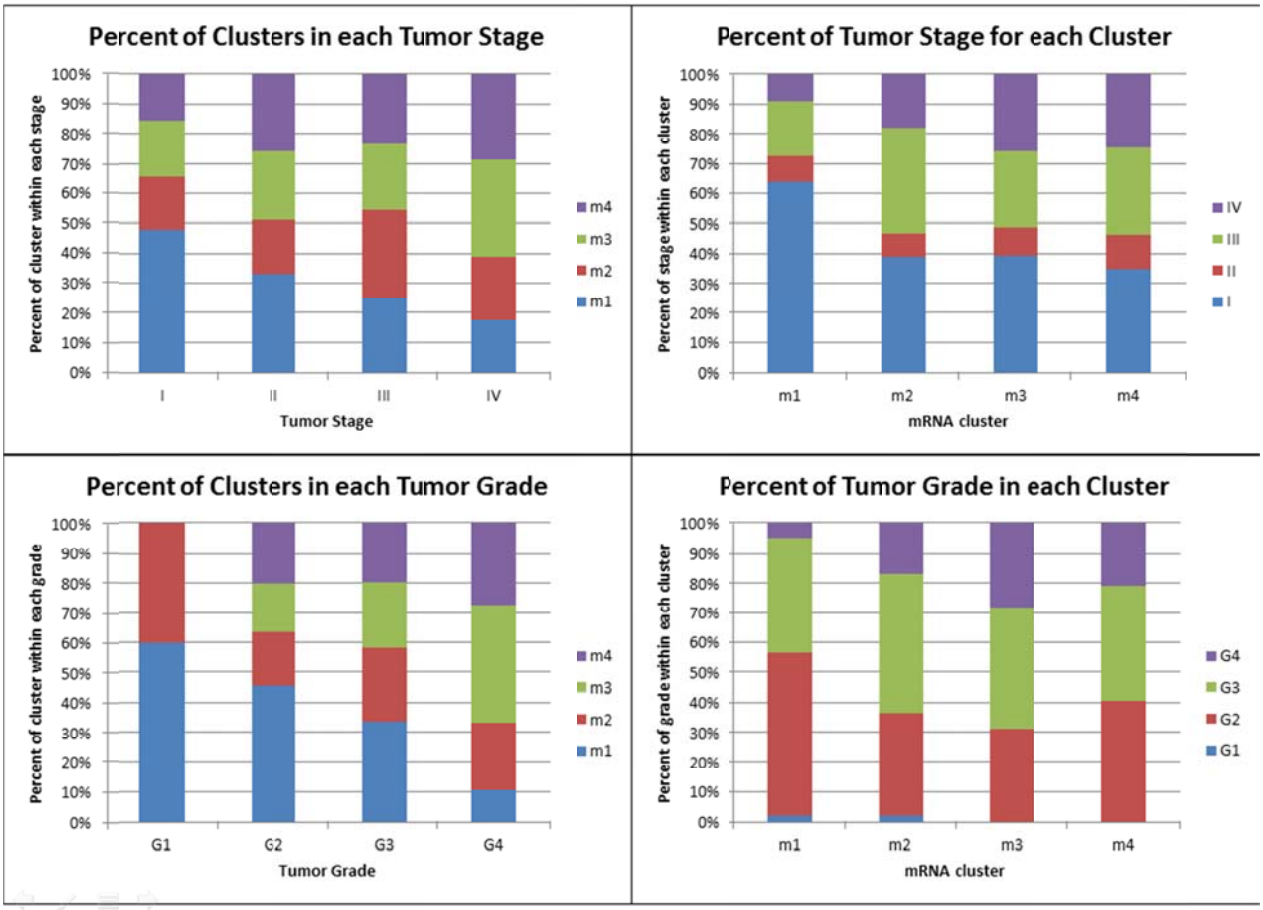


Figure S37. Distribution of stage and grade among the mRNA-based tumor subtypes.

Figure S38

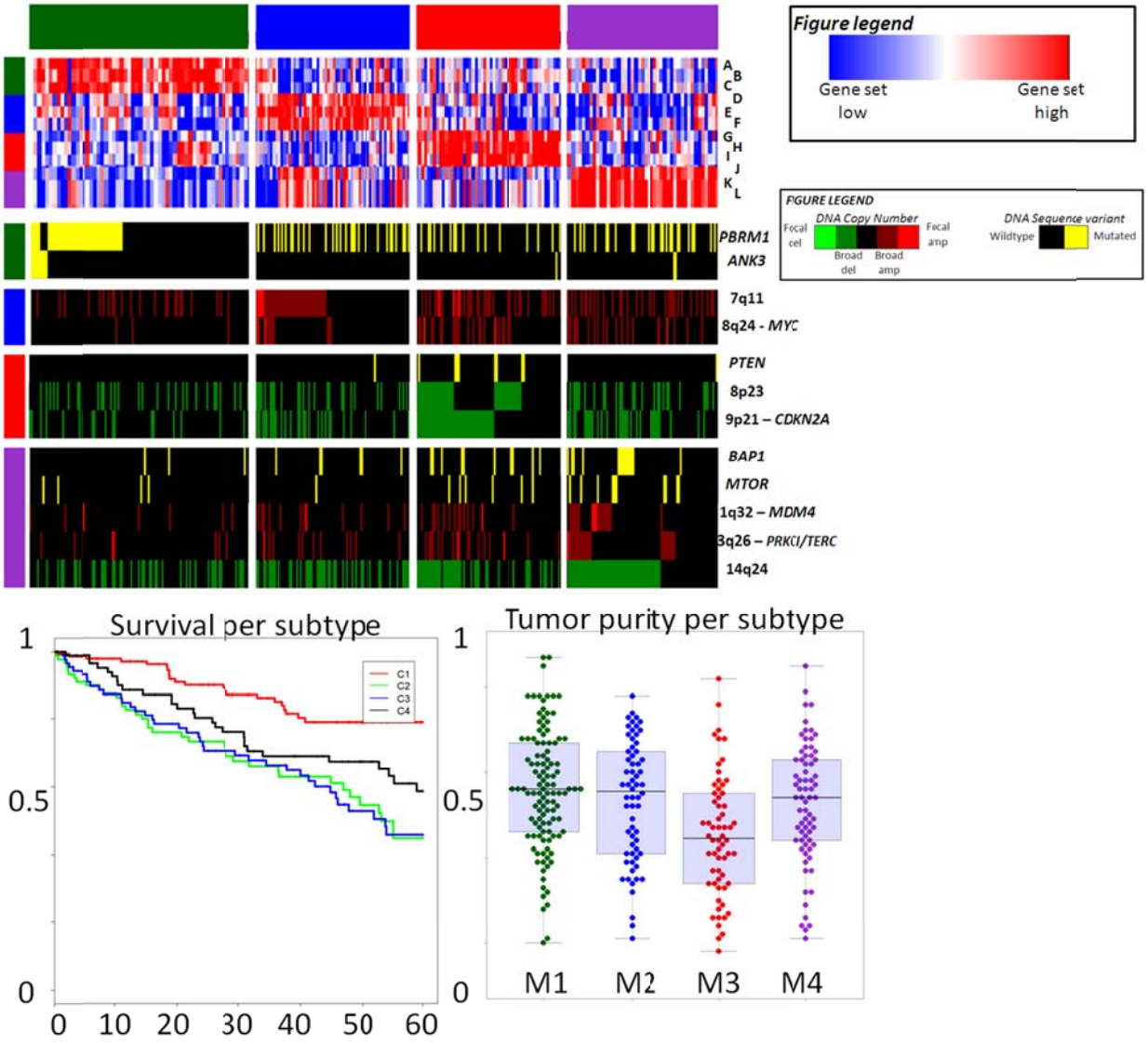


Figure S38. Integrated visualization of gene set activation scores, genomic alterations significantly associated with mRNA subtypes, survival and tumor purity estimates. Gene sets were obtained from MSigDb and the SAM algorithm was applied to find the most significantly activated gene sets per subtype. Tumor purity was estimated by ABSOLUTE. Note that m4 is associated with a number of genomic alterations (i.e. 3q26 amp, 14q del, BAP1/MTOR), suggesting that, while this subtype was not discovered previously using smaller sample sets, it does not represent a spurious group.

Gene sets: (A) REACTOME_GLUCOSE_AND_OTHER_SUGAR_SLC_TRANSPORTERS

(B) SHEN_SMARCA2_TARGETS (C) ACETYLTRANSFERASE_ACTIVITY

(D) REACTOME_BOTULINUM_NEUROTOXICITY (E) BECKER_TAMOXIFEN_RESISTANCE

(F). INTEGRATOR_COMPLEX (G) REACTOME_CELLEXTRACELLULAR_MATRIX_INTERACTIONS (H) SMID_BREAST_CANCER_BASAL (I) EXTRACELLULAR_REGION

(J) REACTOME_BASE_EXCISION_REPAIR (K) ORGANELLAR_RIBOSOME

(L) DACOSTA_UV_RESPONSE_VIA_ERCC3

Figure S39

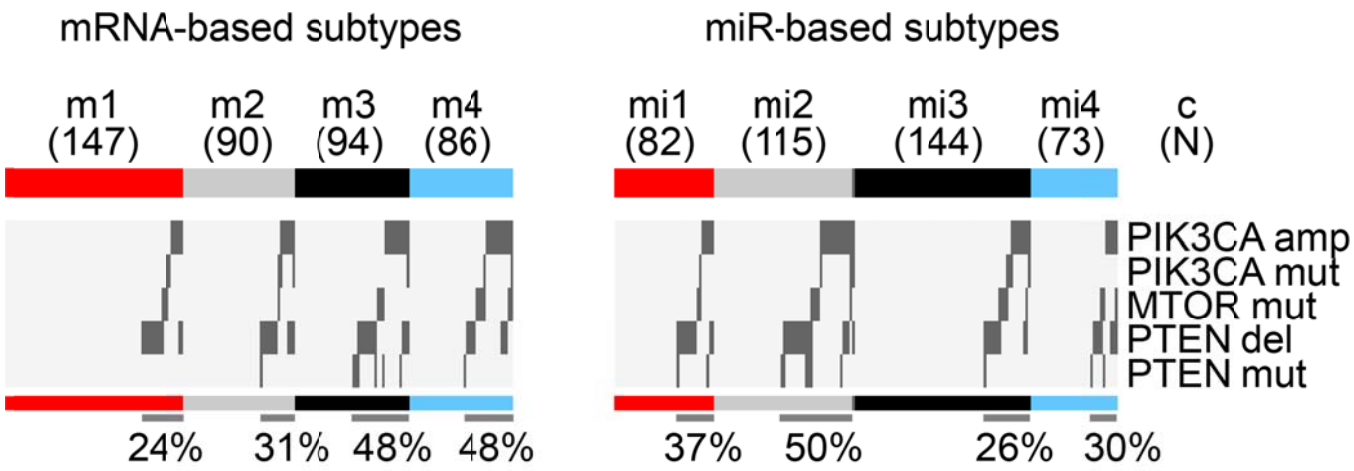


Figure S39. Among the RNA-based subtypes, there is differential enrichment for genetic or genomic alternations in the PI3K pathway, with higher frequencies in mRNA-based m3/m4 and miR-based mi2 (P<0.01, two-sided chi-square, each dataset).

Table S11

Table S11. Overlap between TCGA mRNA-based clusters and previously described ccA/ccB clusters (from Brannon et al.[6]). Red = significant overlap.

		TCGA clusters			
		m1	m2	m3	m4
ccA/ccB biomarkers	ccA	**140**	21	15	44
	ccB	**7**	**69**	**79**	42

IX. MIRNA STUDIES

Workgroup leader: Gordon Robertson (grobertson@bcgsc.ca)

Contributors: Andy Chu, Chad Creighton, Preethi Gunaratne, Anders Jacobsen, Chandra Lebovitz, Sheila Reynolds, Hui Shen, Weimin Xiao

Library construction and sequencing. Two micrograms of total RNA per sample are arrayed into 96-well plates, with controls as described below. RNA entering library construction is required to have at least a minimum quality on the BCR submission documentation. Total RNA is mixed with oligo(dT) MicroBeads and loaded into a 96-well MACS column, which is then placed on a MultiMACS separator (Miltenyi Biotec, Germany). The separator's strong magnetic field allows beads to be captured during washes. From the flow-through, small RNAs, including miRNAs, are recovered by ethanol precipitation. Flow-through RNA quality is checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip.

miRNA-Seq libraries are constructed using an plate-based protocol developed at the British Columbia Genome Sciences Centre (BCGSC). Negative controls are added at three stages: elution buffer is added to one well when the total RNA is loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR brew mix to a final well just before PCR. A 3' adapter is ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation of 1 hour at 22°C. This adapter is adenylated, single-strand DNA with the sequence 5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and is incubated at 37°C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation is completed, 1st strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). This is the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix is made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and DMSO. The mix is distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) is used to transfer the PCR template (1st strand cDNA) and indexed 5' PCR primers into the brew mix plate. Each indexed 5' PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3', contains a unique six-nucleotide 'index' (shown here as N's), and is added to each well of the 96-well PCR brew plate. PCR is run at 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a 5 min incubation at 72°C. Quality is then checked across the whole plate using a Caliper LabChipGX DNA chip. PCR products are pooled, then are size selected to remove larger cDNA fragments and smaller adapter contaminants, using a 96-channel automated size selection robot that was developed at the BCGSC. After size selection, each pool is ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters are generated, and lanes are sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

Preprocessing, alignment and annotation. Briefly, the sequence data are separated into individual samples based on the index read sequences, and the reads undergo an initial QC assessment. Adapter sequence is then trimmed off, and the trimmed reads for each sample are aligned to the NCBI GRCh37-lite reference genome. Below we describe these steps in more detail.

Routine QC assesses a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error is estimated by a method originally developed for SAGE (Khattra 2007). Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, typically they are ~21 bp long, and so are shorter than the 31-bp read length. Given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the reference genome, 3' adapter sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5') 15 bases of the 3' adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.

The algorithm first determines whether a read sequence should be discarded as an adapter dimer by checking whether the 3' adapter sequence occurs at the start of the read. For reads passing this stage, the algorithm then tries to identify an exact 15-bp match anywhere within the read sequence. If it cannot, it then retries, starting from the 3' end, and allowing up to 2 mismatches. If the full 15bp is not found, decreasing lengths of adapter are checked, down to the first 8 bases, allowing one mismatch. If a match is still not found, from 7 bases down to 1 base is checked, with an exact match required. Finally, the algorithm will trim 1 base off the 3' end of a read if it happens to match the first base of the adapter. This is based on two considerations. First, it is preferable to get a perfect alignment than an alignment that has a potential one-base mismatch. Second, if only 1 base of adapter was found in the read sequence, the read is likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample's overall miRNA profiling result.

After each read has been processed, a summary report is generated containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v16 is 15 bp, any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome. BWA (Li 2009) alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches are used. Based on comparing expression profiles of test libraries (data not shown), reads that fail the Illumina basecalling chastity filter are retained, while reads that have soft-clipped CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases (see table below), and requiring a minimum 3-bp overlap between the alignment and an annotation. In annotating reads we address two potential issues. First, a single read alignment can overlap feature annotations of different types; second, a read can have up to three alignment locations, and each alignment location can overlap a different type of feature annotation. By considering heuristically determined priorities, we resolve the first issue by giving each alignment a single annotation. We resolve the second by collapsing multiple annotations to a single annotation, as follows.

For small RNA sequencing, annotation priorities that are used to resolve multiple database matches for a single alignment location and multiple alignment locations for a read.

Priority	Annotation type	Database
1	mature strand	miRBase v16
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	"unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	
6	snoRNA	UCSC small RNAs, RepeatMasker
7	tRNA	
8	rRNA	
9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	
13	coding exons with zero annotated CDS region length	UCSC knownGenes
14	3' UTR	
15	5' UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker Low complexity	
24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

If a read has more than one alignment location, and the annotations for these are different, we use the priorities from the above table to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, the read is flagged as cross-mapped (de Hoon 2010), and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

From the profiling results for a tumor type, for a minimum of approximately 100 samples, we identify the depth of sequencing required to detect the miRNAs that are expressed in a sample by considering a

graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs. For the current work, a library from a sequenced pool was required to have at least 750,000 reads mapped to miRBase annotations. For any sequencing run that fails to meet this threshold, we sequence the sample again to achieve at least the minimum number of miRNA-aligned reads. Finally, for each sample, the reads that correspond to particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Quantification files include information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

Unsupervised consensus clustering. Normalized read count data for 414 tumor samples were extracted from Level 3 data archives on the TCGA Data Portal website (http://tcga.cancer.gov/dataportal). The set of isoform.quantification.txt files, which give read counts at base pair resolution, was processed to sum up read counts at mature and star strand resolution (corresponding to miRBase v16 MIMAT accessions). Read counts for each sample were normalized to RPM, i.e. to reads per million reads aligned to miRBase mature or star strands. Strands corresponding to miRNAs that had been removed from v18 miRBase (miRNA.dead) were eliminated. Mature and star strands were ranked by RPM variance across the samples, and the most variant 25% (306 MIMATs) were input to the NMF v0.5.02 R package (Gaujoux and Seoighe 2010) in R v2.12.0 for unsupervised consensus clustering, using the default Brunet algorithm and 200 iterations, with the rank survey using 50 iterations. A four-cluster result was selected by considering profiles of cophenetic score and average silhouette width for clustering solutions having between 3 and 15 clusters; comparing silhouette plots for solutions with high cophenetic and width scores; and comparing core/non-core cluster members with clinical covariate tracks. Silhouette results were generated from the consensus membership matrix using the R 'cluster' package v1.14.1. Silhouette width profiles were generated for samples ordered as in the NMF heatmap, and atypical, or 'non-core' members in each cluster were identified using a silhouette width threshold set to a fraction (e.g. 0.95) of the maximum width in each cluster. Asymptotic association p-values for covariate contingency tables were calculated using R's chi-square test. KM-curves were calculated with the R 'survival' package v2.36-12.

Discriminatory miRNAs. RF-ACE (www.systemsbiology.org/rf-ace) calculations on pre-miRNA data are outlined below. To complement these results, for each unsupervised sample group we identified discriminatory mature and star strands ('MIMATs') by generating a random forest classifier for samples in that group vs. all other samples (Mehrian-Shai 2007). Classifiers used R v2.15.1 and randomForest v4.6-6, typically with 50000 trees and mtry=100. For each classifier, using Gini variable importances, we profiled the estimated out-of-bag (OOB) error as a function of the number of most-important MIMATs, and identified the smallest set of miRNA strands that minimized the OOB error (data not shown).

miRNA:mRNA correlation analysis. Correlations between miRNA and mRNA were determined using Pearson's or Spearman's rank coefficient as indicated (Pearson's using log-transformed data). Predicted targeting relationships for miRNA:mRNA correlations were identified using miRanda (microRNA.org, August 2010, conserved set), though similar results, in terms of overall enrichment associations, could be seen using TargetScan 6.0. For a given miRNA, overall enrichment (or anti-enrichment) for its set of predicted mRNA targets within the top expression correlates, was determined using Spearman's rank test; for Figure 3d, the entire set of top expressed mRNAs (5153 with average RPM>10 and at least one predicted interaction for the top 26 differential miRNAs) were ranked by Pearson's correlation with the given miRNA, and the overall positions of the predicted targets within the ranked list was determined by Spearman's.

Unsupervised clustering identifies four prognostic clusters that are discriminated by miR-10b, miR-21, miR-30a and miR-143. Unsupervised consensus clustering (Gaujoux and Seoighe 2010) of miRNA-seq abundance profiles for 414 tumor samples identified four sample groups (Figure S40a-c). Per-group average silhouette widths calculated from the consensus matrix were at least 0.9, indicating that groups were distinct and relatively homogeneous. Sample purity from SNP6 microarray data was somewhat higher in groups 1 and 3, while groups 1 and 2 included more aneuploid samples (Figure S41a,b).

Differences between groups in overall survival were significant, and survival was poorer for group 2 (Figure S40d). For survival times up to approximately five years, survival was relatively insensitive to VHL mutation status for all sample groups; for longer times, survival appeared poorer for samples in groups 1 to 3 that had VHL mutations.

We used two approaches to identify miRNAs that discriminated the unsupervised groups. First, from the normalized RPM values for pre-miRNAs from mirna.quantification.txt Level 3 data archives, we identified the most statistically significant miRNAs using RF-ACE (Figure S41c). For pairs of sample groups, applying a p-value threshold of 1.0e-6, miR-10b discriminated group 1 from all other groups; miR-143 and let-7a discriminated group 4 from all other groups; miR-30a discriminated group 3 from groups 1 and 4, and miR-21 discriminated group 2 from groups 1 and 3. Then, from the isomer quantification data archives, we identified discriminatory mature and star strands using random forest classifiers for samples in each group vs. all other samples (Figure S41d,e). The most statistically significant pre-miRNAs from RF-ACE were consistent with the highest-ranked mature and star strands, and the highly ranked miR-10b, 21, 30a and 204 had been reported in a set of 35 miRNAs that distinguished ccRCC tumor from normal kidney tissue (Liu 2010a).

Group 1 had relatively high levels of miR-10b, consistent with metastatic renal cell cancer (RCC) being linked with low expression of this miRNA (White 2011). Group 2 had relatively high levels of miR-21, consistent with (Faragalla 2012 and Zaman 2012), and low levels of miR-204, consistent with (Mikhaylova 2012). Group 3 had high levels of miR-30a, which is part of a signature that distinguishes metastatic ccRCC (Heinzelmann 2011), and inhibits cell migration and invasion in breast cancer (Cheng 2012).

Group 4 had high levels of miR-143 and let-7a. In colorectal cancer, miR-143 acts as a tumor suppressor by directly targeting MACC1, which transactivates the metastatis-related HGF/MET signaling pathway (Zhang 2012). In renal cell carcinoma, low abundance of this miRNA and of -10b has been correlated with tumor relapse after nephrectomy (Slaby 2012). Let-7 family members are downregulated in aggressive primary metastatic ccRCC tumours (Heinzelmann 2011).

We assessed whether the relatively high or low abundance of the discriminatory miRNAs were due to copy number or DNA methylation differences between the sample groups (Figure S41, Table S14). Using Bonferroni-corrected p=0.01 thresholds, concordance between copy number (CN) and miRNA abundance was statistically significant for only a small number of miRNAs in group 2, with miR-204 losses showing the strongest concordance (Figure S42). For group 3, miR-30a gains were on the threshold of significance. For DNA methylation, we considered the relationship of β values and miRNA abundance for the two most discriminatory miRNAs for each sample group (Figure S43), and, for each miRNA used the probe with the best inverse correlation to miRNA abundance. We noted that certain normal samples had DNA methylation levels typical of tumor samples.

For samples in group 1, the relatively high abundance of both miR-10b and mir-30a was more likely

due to DNA methylation changes than to copy number changes. For group 2, miR-21's high abundance was likely due to DNA methylation but not copy number, while miR-204's low abundance was likely due to copy number but not DNA methylation. For group 3, mir-30a's relatively high abundance was likely due to both DNA methylation and copy number, while neither factor likely contributed to mir-30c-2's relatively high abundance. For samples in group 4, the relatively high abundance of mir-143 and let-7a were unlikely due to changes in either DNA methylation or copy number.

miR-21 mediated repression of predicted target genes iron chaperones PCPB1/PCPB2, hypoxia response factor PURA and VEG2FR antagonist TIMP3 and may play a role in the accumulation of HIF1 in RCC. A key function of the VHL complex is to promote ubiquitin-mediated degradation of the protein products of hypoxia-inducible HIF1α and HIF2α (EPAS1), which can heterodimerize with ARNT to form activating transcription factors (Shay 2012, Dondeti 2012, Pal 2010, Hasse 2006). RPKM was more variable across the sample groups for EPAS1 (2p21) than for HIF1A (14q23.2) (Figure S40f); group 2, which had the poorest survival, had the lowest EPAS1 RPKM and EPAS1-to-HIF1A RPKM ratio.

HIF1A's abundance was correlated with copy number (r=0.35, q=1.4e-14), and the 14q arm showed losses in 43% of cases and gains in only 5%. As 14q and HIF1A losses occurred most frequently in clusters 2 and 3 (Figure S40a), high miR-21 abundance and chromosome 14q loss frequently occurred together and were not mutually exclusive.

EPAS1's abundance was comparable in cluster 2 and in tissue normal samples (Figure S40f). Its relatively high abundance in the three other tumor clusters was unlikely due to changes in copy number or to DNA methylation. The chromosome 2p arm was amplified in only 15% of cases (data not shown), consistent with EPAS1-specific copy number gains (Figure S40a), and the gene's abundance was only moderately correlated with copy number changes (r=0.13, q=2.4e-3). EPAS1 was not DNA hypomethylated in any of the clusters, relative to normals (data not shown).

As noted above, in RCC miR-21 is over-expressed and is associated with poor survival. In the current data it appears to be intimately connected to the VHL-HIF1 axis, as its abundance was negatively correlated with mRNA or RPPA data for predicted target genes that are upstream mediators or activators (PTEN, TSC2, PCBP1, PCBP2) and downstream mediators (PURA, PDGFD, TIMP3) of the VHL-mediated degradation of HIF1 (Tables S12 and S13). High expression of each of these predicted miR-21 targets was associated with better survival (Figure S44a), and the RPKM was lowest in group 2 (Figure S44b).

It is important to note that emerging evidence suggests that HIF2α and not HIF1α is most likely the driver of ccRCC. HIF1α has been confirmed as having a tumor suppressor role (Shen et al. 2012) and is a target of 14q loss in kidney cancer, suggesting that the relationship between these factors is more complicated than earlier envisioned.

VHL-mediated ubiquitinylation of HIF1 is critically dependent on the hydroxylation of two key HIF prolyl residues by the HIF prolyl hydroxylase PHDs (Linehan 2010, Baldewijns 2010). This in turn is dependent on PHD activation by the iron chaperones PCBP1 and PCBP2. Depletion of PCBP1 or PCBP2 was recently reported to result in reduced prolyl hydroxylation of HIF1α, and inhibition of the degradation of HIF1α through the VHL/proteasome pathway (Nadal 2011). This was rescued by the addition of excess Fe(II), or purified Fe-PCBP1, and PCBP1 bound to PHD2 in vivo (Nadal 2011).

TIMP3 is a validated miR-21 target in RCC cell lines, and has relatively low abundance in RCC (Zhang 2011, Masson 2010). Further, TIMP3 is a powerful angiostatic agent that blocks the binding of VEGF to its receptor VEGFR2, downstream of HIF1 (Fogarasi 2008). Decorin, a biological ligand of EGFR, can suppress tumors in the colon by inducing genes that include TIMP3 (Moscatello 1998, Santra 1995).

HIF-mediated tumorigenesis in the kidney results from enhanced signaling via mTOR pathway, which is a target of the drug rapamycin (Linehan 2010). In RPPA data for RCC tumors, the predicted target PTEN was negatively correlated to miR-21 (-0.34, Figure S40e, Table S13). While miR-21 was only weakly anti-correlated to PTEN mRNA (-0.21, rank < 4000 in Table S12), PTEN's mRNA abundance was lowest in group 2 (Figure S44b). PTEN targeting by miR-21 has been associated with TORC1 activation in renal carcinoma cell proliferation and invasion (Dey 2012).

We noted that miR-21 was also negatively correlated with mRNA data for two genes that are downstream targets of HIF1 and predicted targets of miR-21: hypoxia response factors PDGFD and PURA, which are required for driving angiogenesis and inflammation (below). This suggests that miR-21 may counteract downstream impacts of HIF-1 accumulation.

Platelet-derived growth factor-D is typically expressed at high levels in the podocytes. Overexpression of PDFG has been shown to result in glomerulonephritis (van Roeyen 2011). Inhibiting PDGFD has been shown to result in decreased pathological angiogenesis (Kumar et al. 2010).

Surface expression of the β2-integrin is critical to hypoxia-mediated inflammation. PURA is a transcription factor that acts in conjunction with HIF-1 to coordinately express all the members of the heterodimeric β2-integrin family (Kong 2007). PURA has been shown to bind to the promoters of E2F1 and AR (androgen receptor) to inhibit transcription and suppress cell proliferation (Gallia 2000) and to increase sensitivity of patients with advanced prostate cancer to hormone replacement therapy (Liu 2010b). Furthermore, deletions of both PURA and PURB, which function as a heterodimer, are associated with an increased incidence of progression of MDS to AML (Lezon-Geyda 2001).

Together, these results suggest that miR-21, and likely other microRNAs, may alter the regulatory relationships between HIF transcription factors and the genes that they activate (Tanimoto 2010). While in the majority of RCC tumors accumulation of HIF1 is a direct result of mutations in VHL (Linehan 2010), the current results also suggest that miR-21 over-expression can also contribute to RCC through the accumulation of HIF-1, due either to a failure to undergo proly-hydroxylation via PCBP1 and PCBP2, or a failure to inhibit the mTOR pathway via increased HIF1 accumulation. Drugs that target individual genes and pathways that are in the VHL-HIF1 axis have been found to be only partially effective, leading researchers to suggest that therapies that target upstream effectors of HIF or multiple downstream pathways may substantially improve survival of patients with RCC (Linehan 2010). In that context, targeted therapies for miR-21 inhibition may offer the combination of pleiotropic effects that are needed for strong tumor suppression of RCC. For example, therapeutic agents that inhibit miR-21 may have the effect of combination therapies using VEGFR/VEGF inhibitors (sorafenib, sunitinib, pazopanib and bevacizumab) and mTOR inhibitors (temsirolimus and everolimus).

References

Baldewijns MM, van Vlodrop IJ, Vermeulen PB, Soetekouw PM, van Engeland M, de Bruïne AP. VHL and HIF signalling in renal cell carcinogenesis. J Pathol. 2010; 221(2):125-38.

Cheng CW, Wang HW, Chang CW, Chu HW, Chen CY, Yu JC, Chao JI, Liu HF, Ding SL, Shen CY. MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. Breast Cancer Res Treat. 2012; 134(3):1081-93.

de Hoon MJ, Taft RJ, Hashimoto T, Kanamori-Katayama M, Kawaji H, Kawano M, Kishima M, Lassmann T, Faulkner GJ, Mattick JS, Daub CO, Carninci P, Kawai J, Suzuki H, Hayashizaki Y. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. Genome Res. 2010; 20(2):257-64.

Dey N, Das F, Ghosh-Choudhury N, Mandal CC, Parekh DJ, Block K, Kasinath BS, Abboud HE, Choudhury GG. microRNA-21 Governs TORC1 Activation in Renal Cancer Cell Proliferation and Invasion. PLoS One. 2012; 7(6):e37366.

Dondeti VR, Wubbenhorst B, Lal P, Gordan JD, D'Andrea K, Attiyeh EF, Simon MC, Nathanson KL. Integrative genomic analyses of sporadic clear cell renal cell carcinoma define disease subtypes and potential new therapeutic targets. Cancer Res. 2012; 72(1):112-21.

Faragalla H, Youssef YM, Scorilas A, Khalil B, White NM, Mejia-Guerrero S, Khella H, Jewett MA, Evans A, Lichner Z, Bjarnason G, Sugar L, Attalah MI, Yousef GM. The Clinical Utility of miR-21 as a Diagnostic and Prognostic Marker for Renal Cell Carcinoma. J Mol Diagn. 2012; 14(4):385-92.

Fogarasi M, Janssen A, Weber BH, Stöhr H. Molecular dissection of TIMP3 mutation S156C associated with Sorsby fundus dystrophy. Matrix Biol. 2008;27(5):381-92.

Gallia GL, Johnson EM, Khalili K. Puralpha: a multifunctional single-stranded DNA- and RNA-binding protein. Nucleic Acids Res. 2000; 28(17):3197-205.

Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010; 11:367.

Haase VH. The VHL/HIF oxygen-sensing pathway and its relevance to kidney disease. Kidney Int. 2006; 69(8):1302-7.

Heinzelmann J, Henning B, Sanjmyatav J, Posorski N, Steiner T, Wunderlich H, Gajda MR, Junker K. Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. World J Urol. 2011; 29(3):367-73.

Khattra J, Marra MA. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells and cell lines. Genome Res. 2007; 17(1):108-16.

Kong T, Scully M, Shelley CS, Colgan SP. Identification of Pur alpha as a new hypoxia response factor responsible for coordinated induction of the beta 2 integrin family. J Immunol. 2007; 179(3):1934-41.

Kumar A, Hou X, Lee C, Li Y, Maminishkis A, Tang Z, Zhang F, Langer HF, Arjunan P, Dong L, Wu Z, Zhu LY, Wang L, Min W, Colosi P, Chavakis T, Li X. Platelet-derived growth factor-DD targeting arrests pathological angiogenesis by modulating glycogen synthase kinase-3beta phosphorylation. J Biol Chem. 2010; 285(20):15500-10.

Lezon-Geyda K, Najfeld V, Johnson EM. Deletions of PURA, at 5q31, and PURB, at 7p13, in myelodysplastic syndrome and progression to acute myelogenous leukemia. Leukemia. 2001; 15(6):954-62.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754-60.

Linehan WM, Srinivasan R, and Schmidt LS. The genetic basis of kidney cancer: a metabolic disease. Nat Rev Urol. 2010; 7(5): 277–285.

Liu H, Brannon AR, Reddy AR, Alexe G, Seiler MW, Arreola A, Oza JH, Yao M, Juan D, Liou LS, Ganesan S, Levine AJ, Rathmell WK, Bhanot GV. Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. BMC Syst Biol. 2010; 4:51.

Liu X, Gomez-Pinillos A, Liu X, Johnson EM, Ferrari AC. Induction of bicalutamide sensitivity in prostate cancer cells by an epigenetic Puralpha-mediated decrease in androgen receptor levels. Prostate. 2010a; 70(2):179-89.

Masson D, Rioux-Leclercq N, Fergelot P, Jouan F, Mottier S, Théoleyre S, Bach-Ngohou K, Patard JJ, Denis MG. Loss of expression of TIMP3 in clear cell renal cell carcinoma. Eur J Cancer. 2010; 46(8):1430-7.

Mehrian-Shai R, Chen CD, Shi T, Horvath S, Nelson SF, Reichardt JK, Sawyers CL. Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. Proc Natl Acad Sci U S A. 2007; 104(13):5563-8.

Mikhaylova O, Stratton Y, Hall D, Kellner E, Ehmer B, Drew AF, Gallo CA, Plas DR, Biesiada J, Meller J, Czyzyk-Krzeska MF. VHL-Regulated MiR-204 Suppresses Tumor Growth through Inhibition of LC3B-Mediated Autophagy in Renal Clear Cell Carcinoma. Cancer Cell. 2012; 21(4):532-46.

Moscatello D. K., Santra M., Mann D. M., McQuillan D. J., Wong A. J., Iozzo R. V. (1998) Decorin suppresses tumor cell growth by activating the epidermal growth factor receptor. J. Clin. Investig. 101:406–412.

Nandal A, Ruiz JC, Subramanian P, Ghimire-Rijal S, Sinnamon RA, Stemmler TL, Bruick RK, Philpott CC. Activation of the HIF prolyl hydroxylase by the iron chaperones PCBP1 and PCBP2. Cell Metab. 2011; 14(5):647-57.

Pal SK, Kortylewski M, Yu H, Figlin RA. Breaking through a plateau in renal cell carcinoma therapeutics: development and incorporation of biomarkers. Mol Cancer Ther. 2010; 9(12):3115-25.

Santra M., Skorski T., Calabretta B., Lattime E. C., Iozzo R. V. De novo decorin gene expression suppresses the malignant phenotype in human colon cancer cells. Proc. Natl. Acad. Sci. U.S.A. 1995; 92:7016–7020.

Shay JE, Celeste Simon M. Hypoxia-inducible factors: Crosstalk between inflammation and metabolism. Semin Cell Dev Biol. 2012; 23(4):389-94.

Shen C, Beroukhim R, Schumacher SE, Zhou J, Chang M, Signoretti S, Kaelin WG Jr. Genetic and functional studies implicate HIF1α as a 14q kidney cancer suppressor gene. Cancer Discov. 2011; 1(3):222-35.

Figure S40

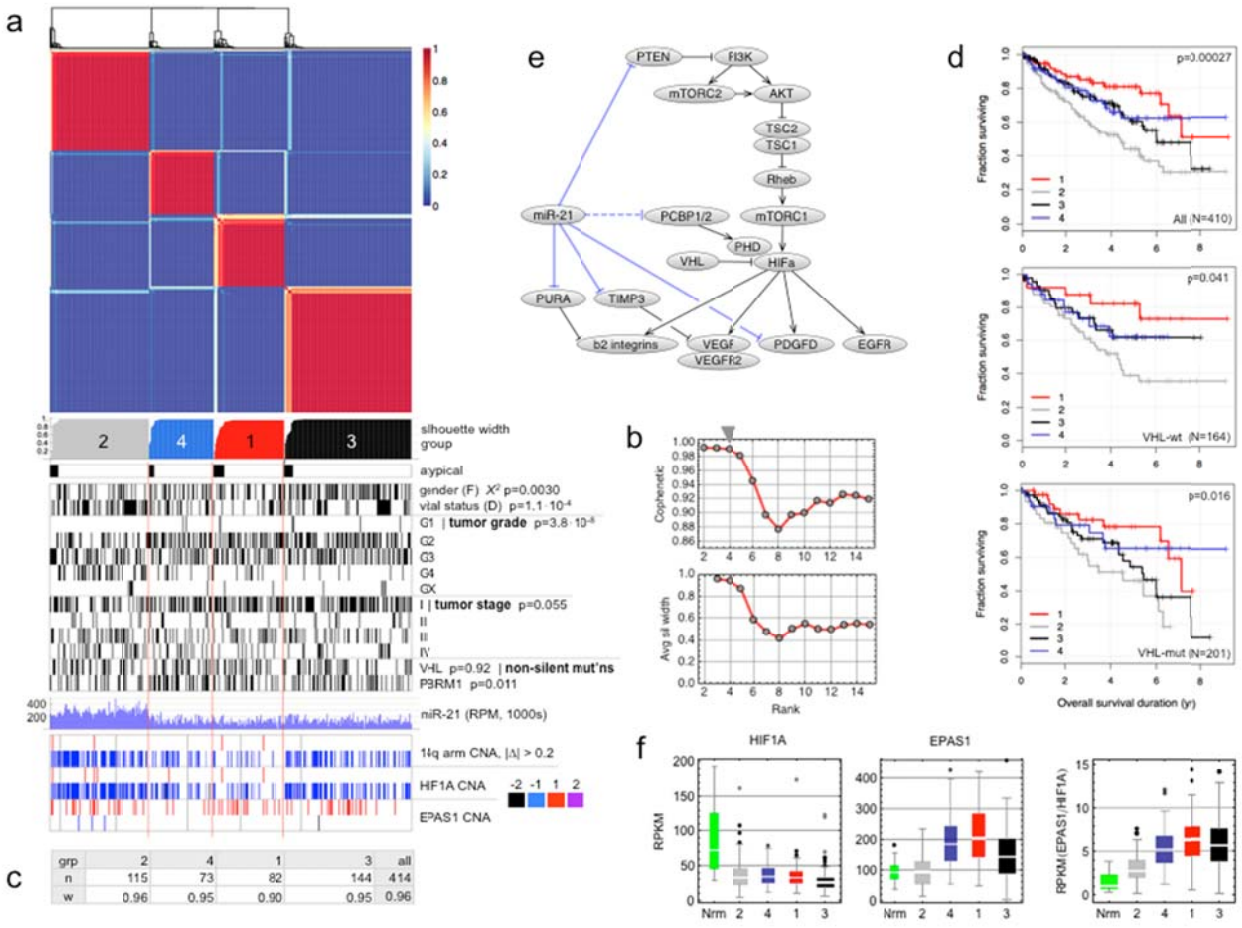


Figure S40. Sample groups identified by unsupervised NMF consensus clustering of miRNA-seq abundance profiles for 414 tumor samples. **a)** Consensus membership heatmap for four sample groups. Below the heatmap are (top to bottom): silhouette width profile calculated from the consensus matrix; atypical members of each group (black) based on an f=0.95 per-cluster silhouette width threshold; gender, vital status, tumor grade and stage, and somatic mutation calls for VHL and PBRM1, with associated Chi-square p-values; miR-21 abundance profile; copy number gains and losses for the 14q arm, then for HIF1A and EPAS1. **b)** Profiles of cophenetic correlation coefficient and average silhouette widths for NMF clustering solutions with up to 15 sample groups. A grey triangle marks the four-cluster solution. **c)** The number of samples (n) and average silhouette width (w) for each group and for all samples. **d)** Kaplan Meier curves for overall survival. Follow-up time was taken as days-to-last-followup vs. days-to-death for samples with a living vs. deceased status. Top to bottom: all samples, samples with wild-type VHL, and samples with at least one non-silent mutation in VHL. **e)** Schematic of relationships between miR-21 and genes related to the VHL-HIF1α-integrin axis (adapted from Fig. 3 of Linehan et al. 2010). Blue lines show negative correlations between miR-21 abundance and RPPA or mRNA-seq data that are consistent with validated (for PTEN) or predicted (TargetScan) targeting relationships. The dashed horizontal link indicates that HIF1a's abundance can reflect a complex set of regulatory influences. **f)** mRNA abundance (RPKM) for HIF1A, EPAS1 (HIF2A) and VHL in 71 normal samples and each of the four tumor sample groups.

Figure S41

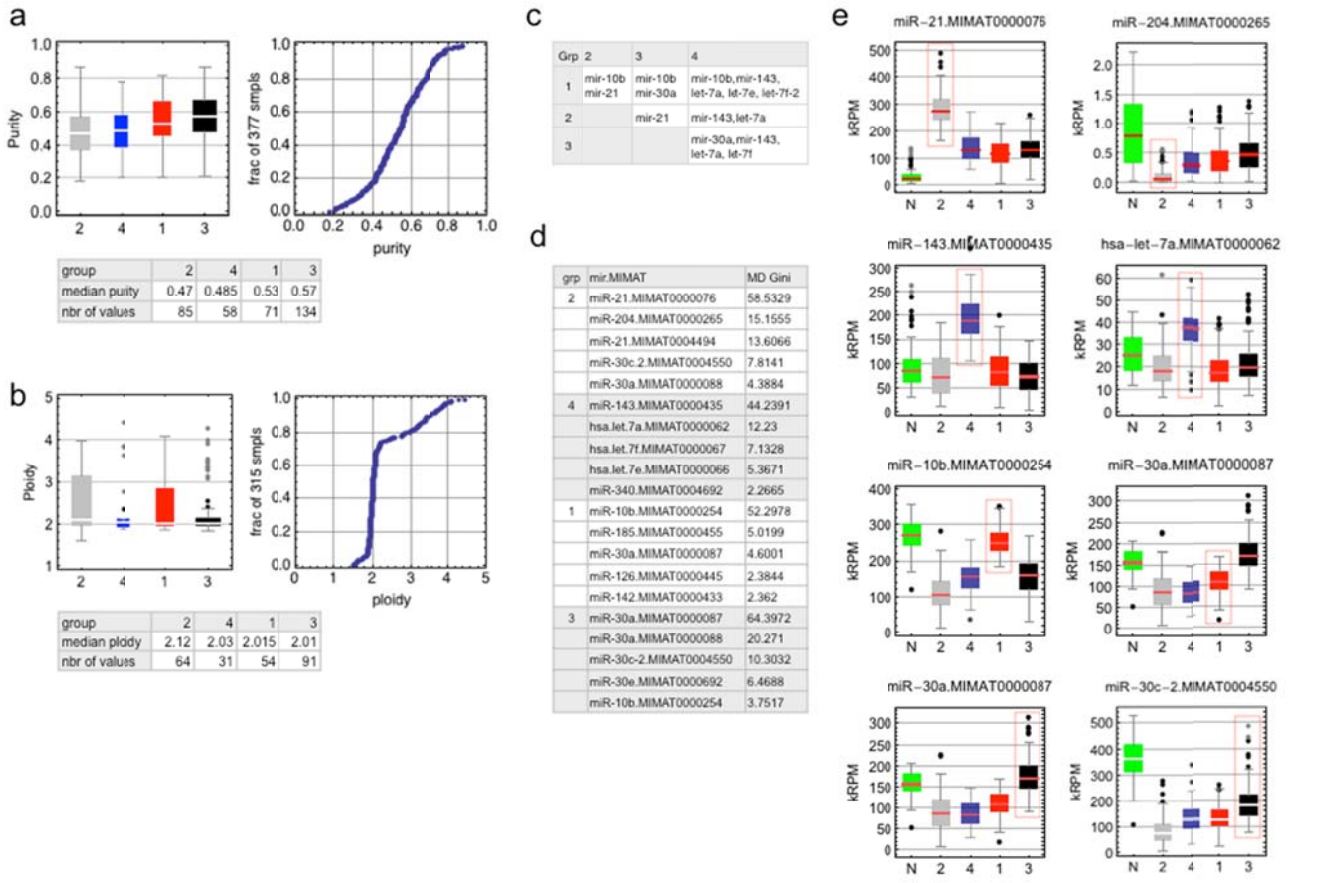


Figure S41. Summary properties of the four miRNA-based sample groups. **a,b)** Summary of purity and ploidy results from SNP6 microarray data. **c)** miRNAs that significantly discriminate one cluster group from another as determined using RF-ACE from pre-miRNA data. **d)** The five most discriminatory miRNA mature or star strands for each group-specific random forest classifier, with Gini-based variable importances. **e)** Distribution of RPM abundance for the two most discriminatory mature or star strands for each sample group in **(d)**. MIMATs are miRBase v16 identifiers. Red rectangles highlight sample groups for which MIMATs were discriminatory.

Figure S42

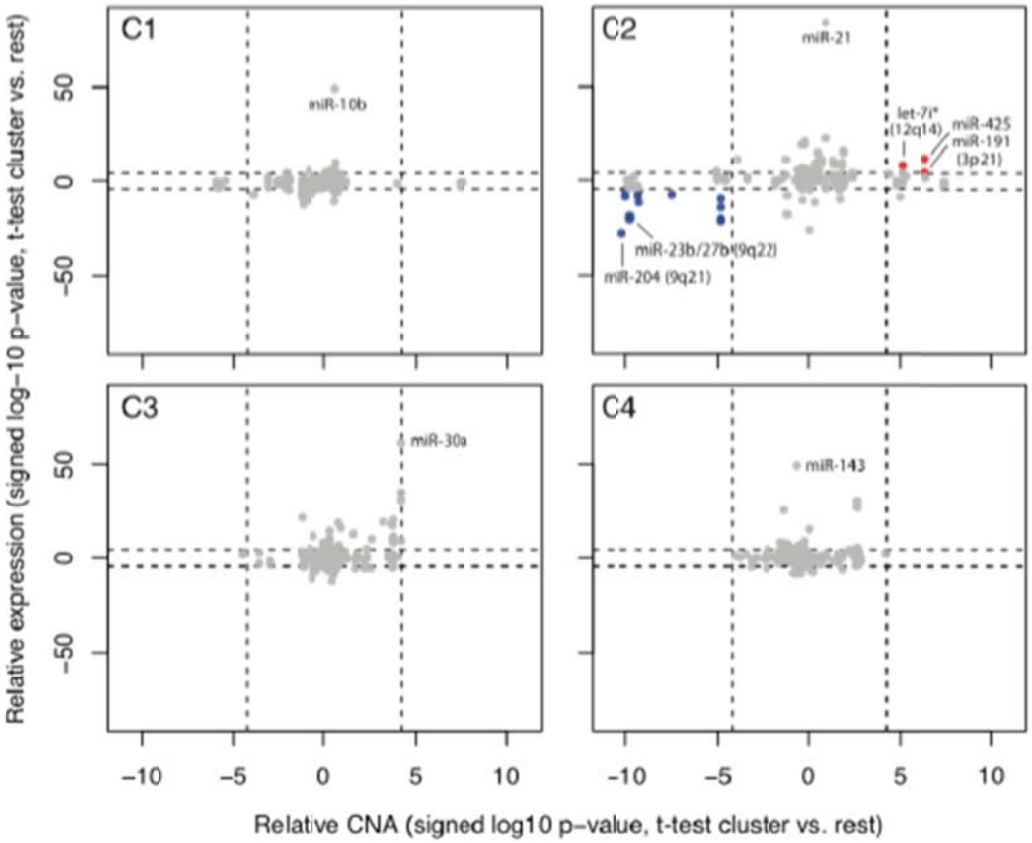


Figure S42. Relationships between copy number alterations and miRNA abundance in the four miRNA-based sample groups. Dashed lines show Bonferroni p=0.01 thresholds.

Figure S43

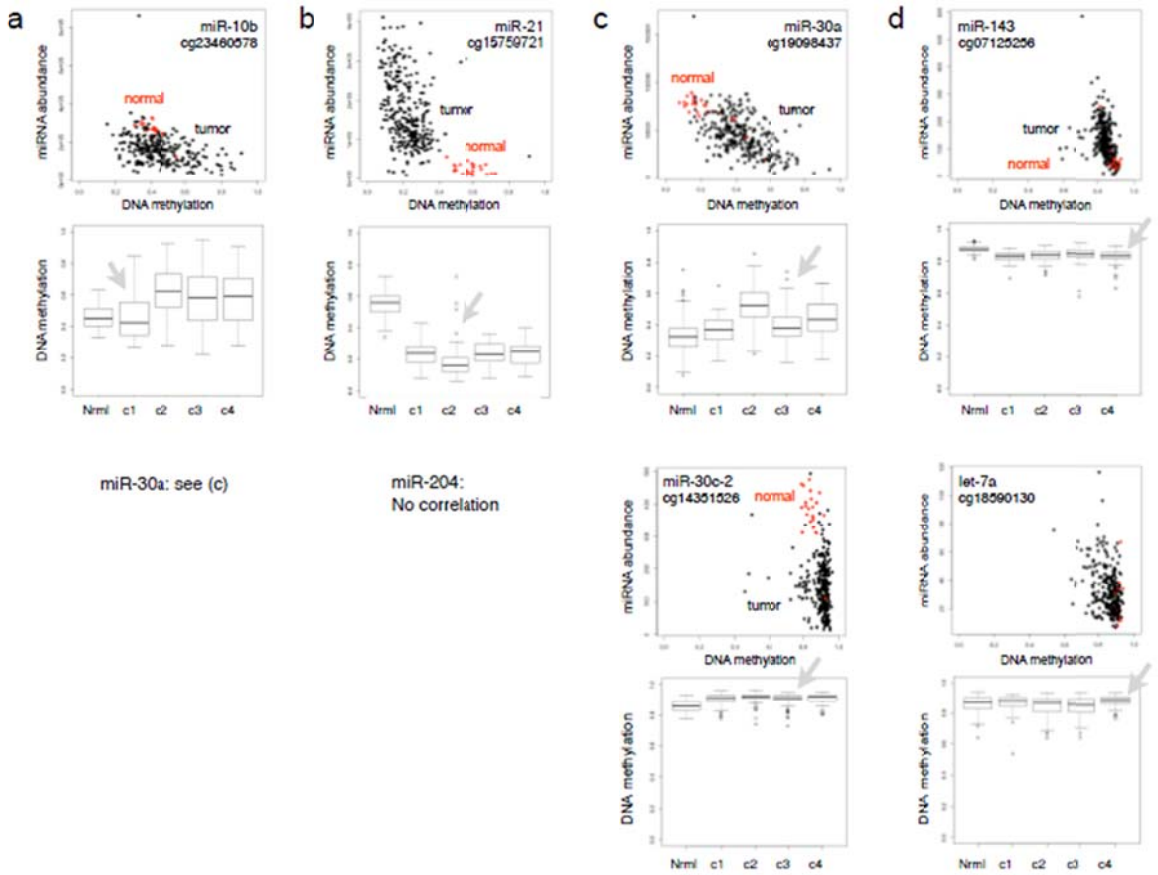


Figure S43. DNA methylation and miRNA abundance for the most discriminatory miRNA for each sample group (Figures S40, S41). For each miRNA, results are shown for the probe that had the best inverse correlation with miRNA abundance. Probe ‘cg’ identifiers are given. DNA methylation is reported as beta values. Scatterplots compare normal samples to all tumor samples. Boxwhisker plots summarize distributions of beta values for normal samples, then for each tumor sample group. Gray arrows mark sample groups for which miRNAs were discriminatory.

Figure S44

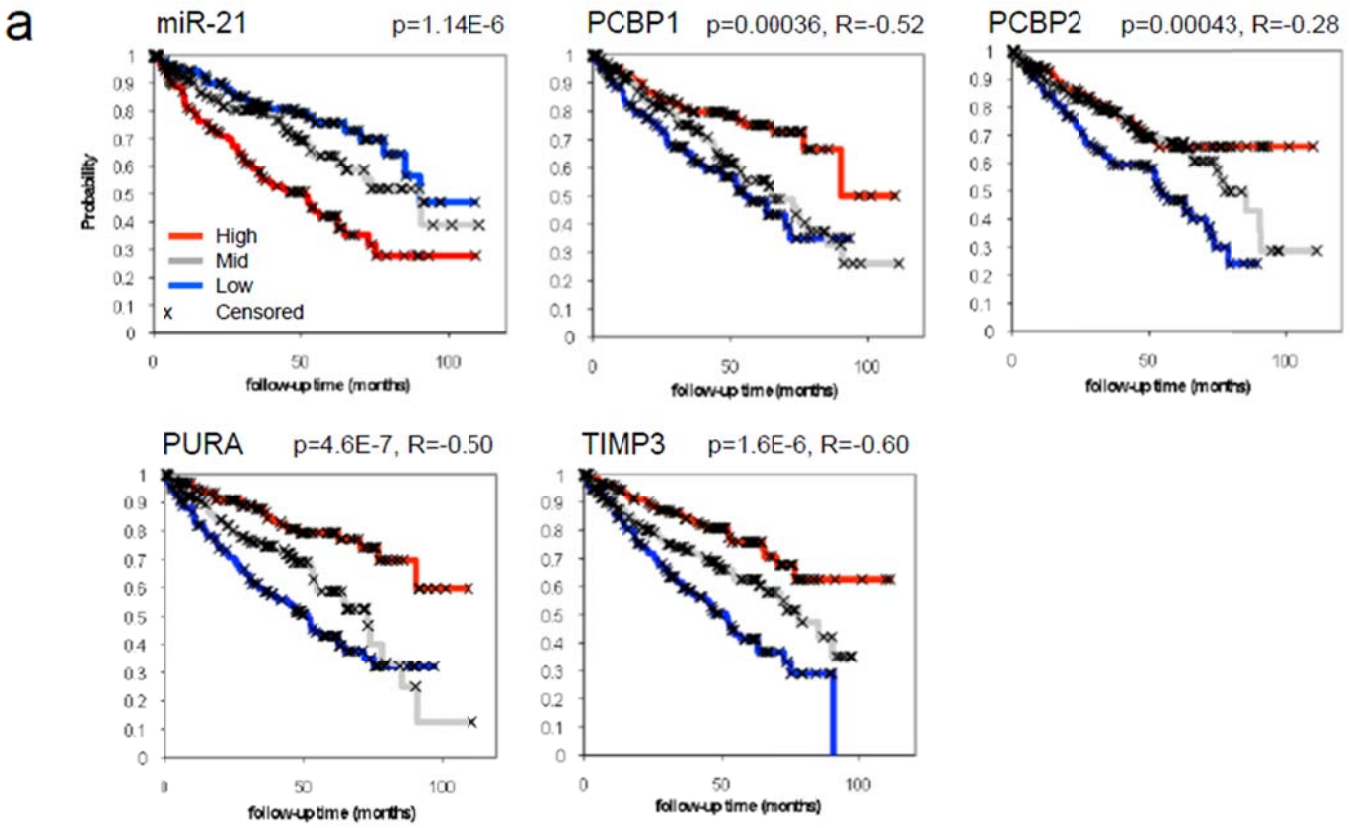


Figure S44. **a)** Kaplan Meier results for miR-21, and for genes associated with the VHL complex and HIF1A (Figure S40e). P-values are from a Cox-Mantel log-rank test. One-sided Pearson correlation coefficients (R) are relative to miR-21.

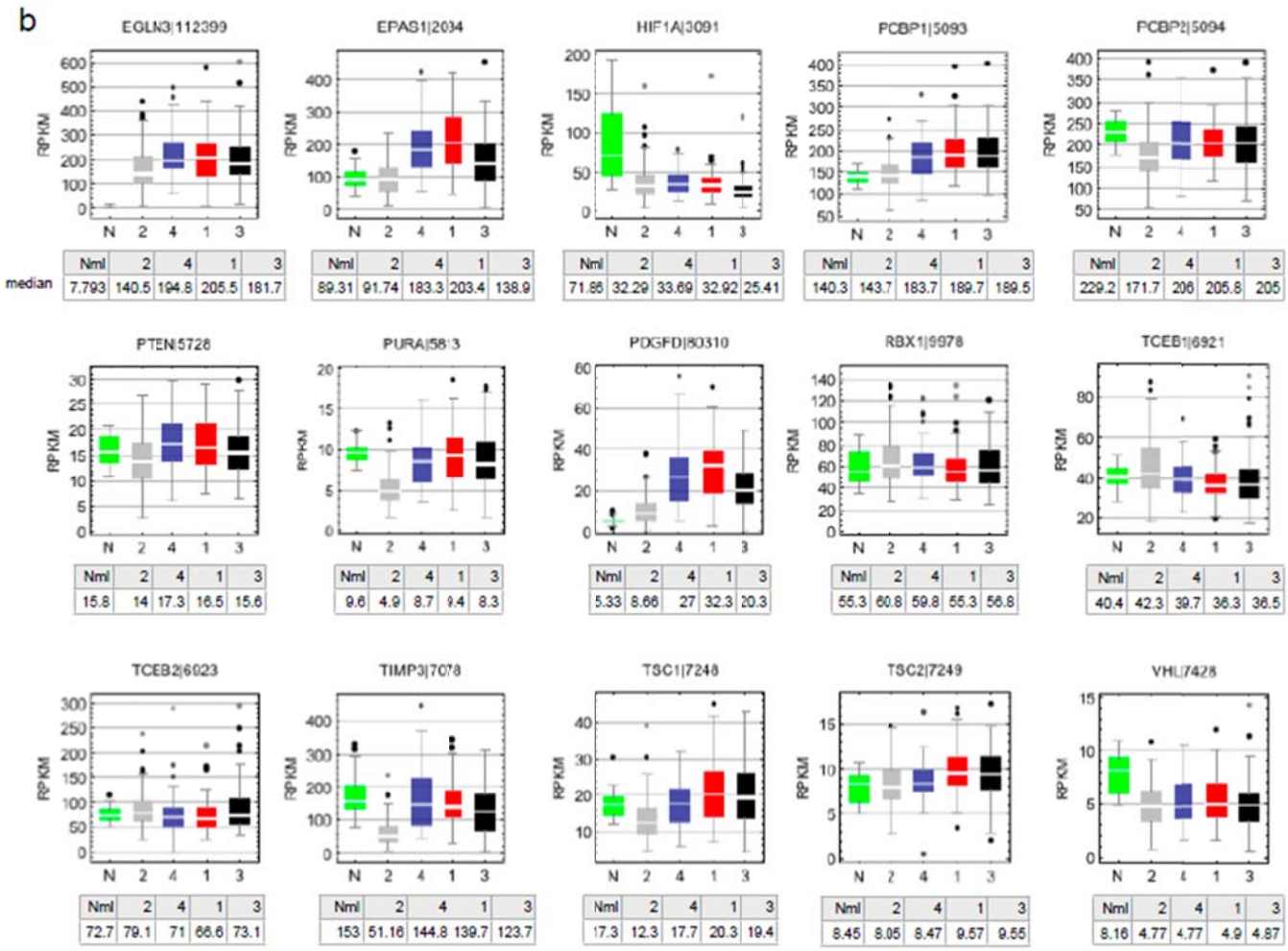


Figure S44. **b)** RPKM abundance of VHL-associated genes in adjacent normals and the four miRNA-based tumor sample groups.

Table S12

Table S12. Predicted miR-21 targets that have strong negative correlations to mRNA RPKM data.

Gene	Correlation gene rank (N = 18,053)	Spearman correlation	miRanda-miRSVR score	TargetScan contextscore	Targetscan conservation score
PDGFD	1	-0.68	0.35	0.11	0
EMCN	8	-0.65	0.18	0.02	0
TIMP3[1]	12	-0.64	0.56	0.09	0.81
SDPR	17	-0.64	0.63	0.25	0.058
RMND5A	37	-0.60	0.43	0.32	0.40
DDAH1	40	-0.59	0.20	0.18	0
RCAN2	45	-0.59	0.70	0.07	0.06
WDR72	55	-0.58	0.31	0.09	0
FRMD3	60	-0.58	1.22	0.33	0
ZNF189	69	-0.57	0.64	0.17	0
GBA3	73	-0.57	0.42	0.12	0
EPHA4	79	-0.57	1.05	0.16	0.01
TSHZ1	84	-0.56	0.19	0.10	0.06
PTPRB	95	-0.56	0.14	0.02	0
PAIP2B	100	-0.55	0.20	0.24	0.35
ANKRD46[2]	104	-0.55	0.85	0.23	0.11
RAVER2	109	-0.55	0.89	0.15	0.13
PCBP1[3]	111	-0.55	1.17	0.25	0.36
TMEM27	117	-0.55	1.18	0.29	0
PURA	130	-0.54	0.42	0.12	0.13
...					
PCBP2	2124	-0.32	0.878	0.212	0.47
PDCD4[4-6]	2373	-0.30	1.211	0.296	0.67

1. Zhang, A., Liu, Y., Shen, Y., Xu, Y. & Li, X. miR-21 Modulates Cell Apoptosis by Targeting Multiple Genes in Renal Cell Carcinoma. *Urology* **78**, 474.e13–474.e19 (2011).
2. Yan, L. X. *et al.* Knockdown of miR-21 in human breast cancer cell lines inhibits proliferation, in vitro migration and in vivo tumor growth. *Breast Cancer Res* **13**, R2 (2011).
3. Yang, Y., Chaerkady, R., Beer, M. A., Mendell, J. T. & Pandey, A. Identification of miR-21 targets in breast cancer cells using a quantitative proteomic approach. *Proteomics* **9**, 1374–1384 (2009).
4. Asangani, I. A. *et al.* MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene* (2007).
5. Frankel, L. B. *et al.* Programmed Cell Death 4 (PDCD4) Is an Important Functional Target of the MicroRNA miR-21 in Breast Cancer Cells. *Journal of Biological Chemistry* **283**, 1026 –1033 (2008).

Table S13

Table S13. Pairwise negative correlations between mir-21 abundance and RPPA data for 166 antibodies for 129 unique gene symbols. Regulome Explorer (explorer.cancerregulome.org), p<=1e-6, N=437 samples were considered for each of the antibodies listed.

Gene	Antibody	rho score	log10(p)
CTNNA1	alpha-Catenin-M-V	-0.476	-25.4
IGF1R	IGF-1R-beta-R-C	-0.443	-21.7
PECAM1	CD31-M-V	-0.430	-20.4
CTNNB1	beta-Catenin-R-V	-0.374	-15.3
SRC	Src_pY527-R-V	-0.374	-15.2
BCL2	Bcl-2-M-V	-0.373	-15.2
PTEN	PTEN-R-V	-0.342	-12.7
ERBB2	HER2-M-V	-0.340	-12.5
TSC2	Tuberin-R-C	-0.329	-11.8
PRKAA1	AMPK_pT172-R-V	-0.319	-11.1
WWTR1	TAZ_pS89-R-C	-0.309	-10.4
PRKAA1	AMPK_alpha-R-C	-0.304	-10.1
SHC1	Shc_pY317-R-NA	-0.289	-9.1
CDH1	E-Cadherin-R-V	-0.288	-9.1
NOTCH1	Notch1-R-V	-0.285	-8.9
AR	AR-R-V	-0.270	-8.0
AKT1	Akt-R-V	-0.254	-7.1
MAPK1	ERK2-R-NA	-0.252	-7.0
COL6A1	Collagen_VI-R-V	-0.251	-7.0
EGFR	EGFR_pY1068-R-V	-0.250	-6.9
PDK1	PDK1_pS241-R-V	-0.250	-6.9
RAD50	Rad50-M-C	-0.242	-6.5
ERBB3	HER3-R-V	-0.237	-6.3
CDKN1B	p27-R-V	-0.233	-6.1

Table S14

Table S14. Summary of relationships between relative abundance (Figure S41e), copy number (Figure S42) and DNA methylation (Figure S43) for two discriminatory microRNAs per sample group.

Group	miRNA	Relative abundance	CNV, per-group	DNA methylation
1	mir-10b	Weakly below normals in this tumor group, but much lower in other tumor groups.	Changes not statistically significant (NSS).	Hypermethylated in other groups. The median beta is lowest and comparable to normals in this group.
	mir-30a	Intermediate among tumor groups. Less abundant than normals in groups 1, 2 and 4, with group 3 comparable to normals.	NSS	Weakly to clearly hypermethylated in tumor groups. Median betas are lower and closer to normals in groups 1 and 3.
2	mir-21	High in all tumor groups, and highest in this tumor group.	NSS	Hypomethylated in tumor samples. The median beta is lowest in this group.
	mir-204	Low in all tumor groups, and lowest in this tumor group.	Statistically significant losses	No probes were correlated to RPM.
3	mir-30a	Highest, and comparable to normals, in this tumor group. See group 1 comments.	Gains are on threshold of significance	See comments above for group 1.
	mir-30c-2	Well below normals for all tumor groups, and highest in this group.	NSS	Tumor samples appear only marginally hypermethylated, but have low abundances. Betas are typically >0.8 for tumors and normals, and beta differences between groups are small.
4	mir-143	High in this group, but comparable to normals in other tumor groups.	NSS	Tumor samples are marginally hypomethylated, but beta is typically >0.8 for tumors and normals, and differences in beta between tumor groups are small.
	let-7a	High in this group, but comparable to or weakly below normals in other tumor groups.	NSS	Beta is typically >0.75 for tumors and normals, and there are only small beta differences between normals and tumor groups, and between tumor groups.

X. REVERSE PHASE PROTEIN ARRAY (RPPA)

Workgroup leader: Gordon B. Mills (gmills@mdanderson.org)

Contributors: Dimitra Tsavachidou and Yiling Lu

Methods. 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 Na3VO4, and aprotinin 10 Ag/mL) from human tumors and RPPA was performed as described previously[1-5]. 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 five-fold 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 172 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[3,5], 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[1]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[5] 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[3,5,6] using median centering across antibodies (level 3 data). In total, 172 antibodies and 454 samples were used (411 of which were represented in the extended sample set and 344 of which were represented in the core sample set). Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described[7]. 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 use with caution based on degree of validation by criteria previously described[7].

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.

References

1. Tibes R, Qiu Y, Lu Y, et al: 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

2. Liang J, Shao SH, Xu Z-X, et al: The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol 9:218-224, 2007

3. Hu J, He X, Baggerly KA, et al: Non-parametric quantification of protein lysate arrays. Bioinformatics 23:1986-1994, 2007

4. Hennessy BT, Lu Y, Poradosu E, et al: Pharmacodynamic Markers of Perifosine Efficacy. Clinical Cancer Research 13:7421-7431, 2007

5. Coombes K, Neeley S, Joy C, et al: SuperCurve: SuperCurve Package. R package version 1.4.1. 2011

6. Gonzalez-Angulo A, Hennessy B, Meric-Bernstam F, et al: Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics 8:11

7. Hennessy B, Lu Y, Gonzalez-Angulo A, et al: 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

List of Antibodies used for sample profiling by RPPA.

Full Slide Name (Antibody Name + Slide ID)	Protein Name	Gene Name	Antibody validation status	Antibody Origin	Antibody Source (Company)	Catalog Number
14-3-3_epsilon-M-C_GBL9016851	14-3-3_epsilon	YWHAE	Use with Caution	Mouse	Santa Cruz	sc-2395
4E-BP1-R-V_GBL9016651	4E-BP1	EIF4EBP1	Validated	Rabbit	CST	9452
4E-BP1_pS65-R-V_GBL9016652	4E-BP1_pS65	EIF4EBP1	Validated	Rabbit	CST	9456
4E-BP1_pT37-R-V_GBL9016653	4E-BP1_pT37	EIF4EBP1	Validated	Rabbit	CST	9459
53BP1-R-C_GBL9016765	53BP1	TP53BP1	Use with Caution	Rabbit	CST	4937
A-Raf_pS299-R-NA_GBL9016776	A-Raf_pS299	ARAF	NA	Rabbit	CST	4431
ACC_pS79-R-V_GBL9016655	ACC_pS79	ACACA ACACB	Validated	Rabbit	CST	3661
ACC1-R-C_GBL9016656	ACC1	ACACA	Use with Caution	Rabbit	Epitomics	1768-1
AIB1-M-V_GBL9016799	AIB1	NCOA3	Validated	Mouse	BD Biosciences	611105
Akt-R-V_GBL9016657	Akt	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9272
Akt_pS473-R-V_GBL9016658	Akt_pS473	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9271
Akt_pT308-R-V_GBL9016659	Akt_pT308	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9275
alpha-Catenin-M-V_GBL9016803	alpha-Catenin	CTNNA1	Validated	Mouse	Calbiochem	CA1030
AMPK_alpha-R-C_GBL9016660	AMPK_alpha	PRKAA1	Use with Caution	Rabbit	CST	2532
AMPK_pT172-R-V_GBL9016661	AMPK_pT172	PRKAA1	Validated	Rabbit	CST	2535
Annexin_I-R-V_GBL9016745	Annexin_I	ANXA1	Validated	Rabbit	Invitrogen	71-3400
AR-R-V_GBL9016741	AR	AR	Validated	Rabbit	Epitomics	1852-1
ATM-R-C_GBL9016662	ATM	ATM	Use with Caution	Rabbit	Abcam	ab32420
B-Raf-M-NA_GBL9016813	B-Raf	BRAF	NA	Mouse	Santa Cruz	sc-5284
Bak-R-C_GBL9016663	Bak	BAK1	Use with Caution	Rabbit	Epitomics	1542-1
Bax-R-V_GBL9016664	Bax	BAX	Validated	Rabbit	CST	2772

Bcl-2-M-V_GBL9016815	Bcl-2	BCL2	Validated	Mouse	Dako	Dako M0887
Bcl-2-R-NA_GBL9016666	Bcl-2	BCL2	NA	Rabbit	Epitomics	1017-1
Bcl-X-R-C_GBL9016667	Bcl-X	BCL2L1	Use with Caution	Rabbit	Epitomics	1018-1
Bcl-xL-R-V_GBL9016668	Bcl-xL	BCL2L1	Validated	Rabbit	CST	2762
Beclin-G-V_GBL9016868	Beclin	BECN1	Validated	Rabbit	Santa Cruz	sc-10086
beta-Catenin-R-V_GBL9016665	beta-Catenin	CTNNB1	Validated	Rabbit	CST	9562
Bid-R-C_GBL9016669	Bid	BID	Use with Caution	Rabbit	Epitomics	1008-1
Bim-R-V_GBL9016670	Bim	BCL2L11	Validated	Rabbit	Epitomics	1036-1
c-Jun_pS73-R-C_GBL9016678	c-Jun_pS73	JUN	Use with Caution	Rabbit	CST	9164
c-Kit-R-V_GBL9016679	c-Kit	KIT	Validated	Rabbit	Epitomics	1522
c-Met-M-C_GBL9016800	c-Met	MET	Use with Caution	Mouse	CST	3127
c-Met_pY1235-R-C_GBL9016861	c-Met_pY1235	MET	Use with Caution	Rabbit	CST	3129
c-Myc-R-C_GBL9016680	c-Myc	MYC	Use with Caution	Rabbit	CST	9402
C-Raf-R-V_GBL9016748	C-Raf	RAF1	Validated	Rabbit	Millipore	05-739
C-Raf_pS338-R-C_GBL9016681	C-Raf_pS338	RAF1	Use with Caution	Rabbit	CST	9427
Caspase-3_active-R-C_GBL9016671	Caspase-3_active	CASP3	Use with Caution	Rabbit	Epitomics	1476-1
Caspase-7_cleavedD198-R-C_GBL9016673	Caspase-7_cleavedD198	CASP7	Use with Caution	Rabbit	CST	9491
Caspase-8-M-C_GBL9016805	Caspase-8	CASP8	Use with Caution	Mouse	CST	9746
Caspase-9_cleavedD330-R-C_GBL9016764	Caspase-9_cleavedD330	CASP9	Use with Caution	Rabbit	CST	9501
Caveolin-1-R-V_GBL9016674	Caveolin-1	CAV1	Validated	Rabbit	CST	3238
CD20-R-C_GBL9016816	CD20	CD20	Use with Caution	Rabbit	Epitomics	1632
CD31-M-V_GBL9016838	CD31	PECAM1	Validated	Mouse	Dako	M0823
CD49b-M-V_GBL9016804	CD49b	CD49	Validated	Mouse	BD	611016
CDK1-R-V_GBL9016766	CDK1	CDC2	Validated	Rabbit	CST	9112
Chk1-R-C_GBL9016676	Chk1	CHEK1	Use with Caution	Rabbit	CST	2345
Chk1_pS345-R-C_GBL9016756	Chk1_pS345	CHEK1	Use with Caution	Rabbit	CST	2348
Chk2-M-C_GBL9016789	Chk2	CHEK2	Use with Caution	Mouse	CST	3440
Chk2_pT68-R-C_GBL9016677	Chk2_pT68	CHEK2	Use with Caution	Rabbit	CST	2197
cIAP-R-V_GBL9016760	cIAP	BIRC2	Validated	Rabbit	Millipore	07-759
Claudin-7-R-V_GBL9016752	Claudin-7	CLDN7	Validated	Rabbit	Novus	NB100-91714
Collagen_VI-R-V_GBL9016779	Collagen_VI	COL6A1	Validated	Rabbit	Santa Cruz	SC-20649
COX-2-R-C_GBL9016740	COX-2	PTGS2	Use with Caution	Rabbit	Epitomics	2169-1
Cyclin_B1-R-V_GBL9016682	Cyclin_B1	CCNB1	Validated	Rabbit	Epitomics	1495-1
Cyclin_D1-R-V_GBL9016780	Cyclin_D1	CCND1	Validated	Rabbit	Santa Cruz	SC-718
Cyclin_E1-M-V_GBL9016850	Cyclin_E1	CCNE1	Validated	Mouse	Santa Cruz	SC-247

Cyclin_E2-R-C_GBL9016683	Cyclin_E2	CCNE2	Use with Caution	Rabbit	Epitomics	1142
DJ-1-R-C_GBL9016754	DJ-1	PARK7	Use with Caution	Rabbit	Abcam	ab76008
Dvl3-R-V_GBL9016844	Dvl3	DVL3	Validated	Rabbit	CST	3218
E-Cadherin-R-V_GBL9016684	E-Cadherin	CDH1	Validated	Rabbit	CST	4065
eEF2-R-V_GBL9016771	eEF2	EEF2	Validated	Rabbit	CST	2332
eEF2K-R-V_GBL9016772	eEF2K	EEF2K	Validated	Rabbit	CST	3692
EGFR-R-C_GBL9016781	EGFR	EGFR	Use with Caution	Rabbit	Santa Cruz	SC-03
EGFR_pY1068-R-V_GBL9016685	EGFR_pY1068	EGFR	Validated	Rabbit	CST	2234
EGFR_pY1173-R-C_GBL9016686	EGFR_pY1173	EGFR	Use with Caution	Rabbit	Epitomics	1124
EGFR_pY992-R-V_GBL9016687	EGFR_pY992	EGFR	Validated	Rabbit	CST	2235
eIF4E-R-V_GBL9016736	eIF4E	EIF4E	Validated	Rabbit	CST	9742
ER-alpha-R-V_GBL9016782	ER-alpha	ESR1	Validated	Rabbit	Lab Vision	RM-9101-S
ER-alpha_pS118-R-V_GBL9016688	ER-alpha_pS118	ESR1	Validated	Rabbit	Epitomics	1091-1
ERCC1-M-C_GBL9016839	ERCC1	ERCC1	Use with Caution	Mouse	Lab Vision	MS-671-PO
ERK2-R-NA_GBL9016832	ERK2	MAPK1	NA	Rabbit	Santa Cruz	sc-154
FAK-R-C_GBL9016689	FAK	PTK2	Use with Caution	Rabbit	Epitomics	1700-1
Fibronectin-R-C_GBL9016690	Fibronectin	FN1	Use with Caution	Rabbit	Epitomics	1574-1
FOXO3a-R-C_GBL9016691	FOXO3a	FOXO3	Use with Caution	Rabbit	CST	9467
FOXO3a_pS318_S321-R-C_GBL9016692	FOXO3a_pS318_S321	FOXO3	Use with Caution	Rabbit	CST	9465
GAB2-R-V_GBL9016763	GAB2	GAB2	Validated	Rabbit	CST	3239
GATA3-M-V_GBL9016801	GATA3	GATA3	Validated	Mouse	BD Biosciences	558686
GSK3-alpha-beta-M-V_GBL9016840	GSK3-alpha-beta	GSK3A GSK3B	Validated	Mouse	Santa Cruz	SC-7291
GSK3-alpha-beta_pS21_S9-R-V_GBL9016693	GSK3-alpha-beta_pS21_S9	GSK3A GSK3B	Validated	Rabbit	CST	9331
GSK3_pS9-R-V_GBL9016775	GSK3_pS9	GSK3B	Validated	Rabbit	CST	9336
HER2-M-V_GBL9016807	HER2	ERBB2	Validated	Mouse	Lab Vision	MS-325-P1
HER2_pY1248-R-NA_GBL9016694	HER2_pY1248	ERBB2	NA	Rabbit	R&D	AF1768
HER3-R-V_GBL9016788	HER3	ERBB3	Validated	Rabbit	Santa Cruz	sc-285
HER3_pY1298-R-C_GBL9016862	HER3_pY1298	ERBB3	Use with Caution	Rabbit	CST	4791
HSP70-R-C_GBL9016695	HSP70	HSPA1A	Use with Caution	Rabbit	CST	4872
IGF-1R-beta-R-C_GBL9016697	IGF-1R-beta	IGF1R	Use with Caution	Rabbit	CST	3027
IGFBP2-R-V_GBL9016696	IGFBP2	IGFBP2	Validated	Rabbit	CST	3922
INPP4B-G-C_GBL9016846	INPP4B	INPP4B	Use with Caution	Goat	Santa Cruz	SC-12318
IRS1-R-V_GBL9016747	IRS1	IRS1	Validated	Rabbit	Upstate (Millipore)	06-248
JNK2-R-C_GBL9016698	JNK2	MAPK9	Use with Caution	Rabbit	CST	4672

K-Ras-M-C_GBL9016837	K-Ras	KRAS	Use with Caution	Mouse	Santa Cruz	sc-30 (F234)
Ku80-R-C_GBL9016757	Ku80	XRCC5	Use with Caution	Rabbit	CST	2180
Lck-R-V_GBL9016818	Lck	LCK	Validated	Rabbit	CST	2752
LKB1-M-NA_GBL9016790	LKB1	STK11	NA	Mouse	Abcam	ab15095
MAPK_pT202_Y204-R-V_GBL9016700	MAPK_pT202_Y204	MAPK1 MAPK3	Validated	Rabbit	CST	4377
MEK1-R-V_GBL9016701	MEK1	MAP2K1	Validated	Rabbit	Epitomics	1235-1
MEK1_pS217_S221-R-V_GBL9016702	MEK1_pS217_S221	MAP2K1	Validated	Rabbit	CST	9154
MIG-6-M-V_GBL9016808	MIG-6	ERRFI1	Validated	Mouse	Sigma	WH0054206M1
Mre11-R-C_GBL9016819	Mre11	MRE11A	Use with Caution	Rabbit	CST	4847
MSH2-M-C_GBL9016802	MSH2	MSH2	Use with Caution	Mouse	CST	2850
MSH6-R-C_GBL9016773	MSH6	MSH6	Use with Caution	Rabbit	SDI	2203.00.02
mTOR-R-V_GBL9016820	mTOR	FRAP1	Validated	Rabbit	CST	2983
N-Cadherin-R-V_GBL9016705	N-Cadherin	CDH2	Validated	Rabbit	CST	4061
NF-kB-p65_pS536-R-C_GBL9016706	NF-kB-p65_pS536	NFKB1	Use with Caution	Rabbit	CST	3033
NF2-R-C_GBL9016769	NF2	NF2	Use with Caution	Rabbit	SDI	2271.00.02
Notch1-R-V_GBL9016774	Notch1	NOTCH1	Validated	Rabbit	CST	3268
Notch3-R-C_GBL9016785	Notch3	NOTCH3	Use with Caution	Rabbit	Santa Cruz	sc-5593
P-Cadherin-R-C_GBL9016713	P-Cadherin	CDH3	Use with Caution	Rabbit	CST	2130
p21-R-C_GBL9016784	p21	CDKN1A	Use with Caution	Rabbit	Santa Cruz	SC-397
p27-R-V_GBL9016755	p27	CDKN1B	Validated	Rabbit	Epitomics	1591-1
p27_pT157-R-C_GBL9016863	p27_pT157	CDKN1B	Use with Caution	Rabbit	R&D	AF1555
p38_MAPK-R-C_GBL9016707	p38_MAPK	MAPK14	Use with Caution	Rabbit	CST	9212
p38_pT180_Y182-R-V_GBL9016708	p38_pT180_Y182	MAPK14	Validated	Rabbit	CST	9211
p53-R-V_GBL9016709	p53	TP53	Validated	Rabbit	CST	9282
p70S6K-R-V_GBL9016710	p70S6K	RPS6KB1	Validated	Rabbit	Epitomics	1494-1
p70S6K_pT389-R-V_GBL9016711	p70S6K_pT389	RPS6KB1	Validated	Rabbit	CST	9205
p90RSK_pT359_S363-R-C_GBL9016742	p90RSK_pT359_S363	RPS6KA1	Use with Caution	Rabbit	CST	9344
PARP_cleaved-M-C_GBL9016791	PARP_cleaved	PARP1	Use with Caution	Mouse	CST	9546
Paxillin-R-V_GBL9016712	Paxillin	PXN	Validated	Rabbit	Epitomics	1500-1
PCNA-M-V_GBL9016792	PCNA	PCNA	Validated	Mouse	Abcam	ab29
PDK1_pS241-R-V_GBL9016714	PDK1_pS241	PDK1	Validated	Rabbit	CST	3061
PEA-15-R-V_GBL9016767	PEA-15	PEA15	Validated	Rabbit	CST	2780
PI3K-p110-alpha-R-C_GBL9016749	PI3K-p110-alpha	PIK3CA	Use with Caution	Rabbit	CST	4255
PI3K-p85-R-V_GBL9016715	PI3K-p85	PIK3R1	Validated	Rabbit	Upstate (Millipore)	06-195

PKC-alpha-M-V_GBL9016793	PKC-alpha	PRKCA	Validated	Mouse	Upstate (Millipore)	05-154
PKC-alpha_pS657-R-V_GBL9016716	PKC-alpha_pS657	PRKCA	Validated	Rabbit	Upstate (Millipore)	06-822
PKC-delta_pS664-R-V_GBL9016761	PKC-delta_pS664	PRKCA	Validated	Rabbit	Millipore	07-875
PR-R-V_GBL9016810	PR	PGR	Validated	Rabbit	Epitomics	1483-1
PRAS40_pT246-R-V_GBL9016857	PRAS40_pT246	AKT1S1	Validated	Rabbit	Biosource	441100G
PTCH-R-C_GBL9016770	PTCH	PTCH1	Use with Caution	Rabbit	SDI	2113.00.02
PTEN-R-V_GBL9016719	PTEN	PTEN	Validated	Rabbit	CST	9552
Rab25-R-C_GBL9016720	Rab25	RAB25	Use with Caution	Rabbit	Covance Custom	Covance Custom
Rad50-M-C_GBL9016806	Rad50	RAD50	Use with Caution	Mouse	Millipore	05-525
Rad51-M-C_GBL9016814	Rad51	RAD51	Use with Caution	Mouse	Chem Biotech	na 71
Rb-M-V_GBL9016834	Rb	RB1	Validated	Mouse	CST	9309
Rb_pS807_S811-R-V_GBL9016821	Rb_pS807_S811	RB1	Validated	Rabbit	CST	9308
S6-R-NA_GBL9016721	S6	RPS6	NA	Rabbit	CST	2217
S6_pS235_S236-R-V_GBL9016722	S6_pS235_S236	RPS6	Validated	Rabbit	CST	2211
S6_pS240_S244-R-V_GBL9016723	S6_pS240_S244	RPS6	Validated	Rabbit	CST	2215
SETD2-R-NA_GBL9016768	SETD2	SETD2	NA	Rabbit	Abcam	ab69836
Shc_pY317-R-NA_GBL9016777	Shc_pY317	SHC1	NA	Rabbit	CST	2431
Smac-M-V_GBL9016795	Smac	DIABLO	Validated	Mouse	CST	2954
Smad1-R-V_GBL9016759	Smad1	SMAD1	Validated	Rabbit	Epitomics	1649-1
Smad3-R-V_GBL9016843	Smad3	SMAD3	Validated	Rabbit	Epitomics	1735-1
Smad4-M-V_GBL9016841	Smad4	SMAD4	Validated	Mouse	Santa Cruz	sc-7866
Snail-M-C_GBL9016796	Snail	SNAI2	Use with Caution	Mouse	CST	3895
Src-M-V_GBL9016797	Src	SRC	Validated	Mouse	Upstate (Millipore)	05-184
Src_pY416-R-C_GBL9016724	Src_pY416	SRC	Use with Caution	Rabbit	CST	2101
Src_pY527-R-V_GBL9016725	Src_pY527	SRC	Validated	Rabbit	CST	2105
STAT3_pY705-R-V_GBL9016822	STAT3_pY705	STAT3	Validated	Rabbit	CST	9131
STAT5-alpha-R-V_GBL9016729	STAT5-alpha	STAT5	Validated	Rabbit	Epitomics	1289-1
Stathmin-R-V_GBL9016735	Stathmin	STMN1	Validated	Rabbit	Epitomics	1972-1
Syk-M-V_GBL9016836	Syk	SYK	Validated	Mouse	Santa Cruz	sc-1240
Tau-M-C_GBL9016812	Tau	MAPT	Use with Caution	Mouse	Upstate (Millipore)	05-348
TAZ-R-C_GBL9016743	TAZ	WWTR1	Use with Caution	Rabbit	Abcam	ab3961
TAZ_pS89-R-C_GBL9016786	TAZ_pS89	WWTR1	Use with Caution	Rabbit	Santa Cruz	sc-17610
Tuberin-R-C_GBL9016730	Tuberin	TSC2	Use with Caution	Rabbit	Epitomics	1613-1
VASP-R-C_GBL9016825	VASP	VASP	Use with Caution	Rabbit	CST	3112
VEGFR2-R-V_GBL9016826	VEGFR2	KDR	Validated	Rabbit	CST	2479

XBP1-G-C_GBL9016853	XBP1	XBP1	Use with Caution	Goat	Santa Cruz	sc-32136
XIAP-R-C_GBL9016827	XIAP	XIAP	Use with Caution	Rabbit	CST	2042
XRCC1-R-C_GBL9016758	XRCC1	XRCC1	Use with Caution	Rabbit	CST	2735
YAP-R-V_GBL9016833	YAP	YAP1	Validated	Rabbit	Santa Cruz	sc-15407
YAP_pS127-R-C_GBL9016744	YAP_pS127	YAP1	Use with Caution	Rabbit	CST	4911
YB-1-R-V_GBL9016866	YB-1	YBX1	Validated	Rabbit	SDI	1725.00.02
YB-1_pS102-R-V_GBL9016750	YB-1_pS102	YBX1	Validated	Rabbit	CST	2900
JNK_pT183_pT185_GBL9016891	JNK_pT183_pY185	MAPK8	Use with Caution	Rabbit	CST	4668
Pal_1_GBL9016890	PAI-1	PAI-1	NA	Mouse	BD Biosciences	612024
ARID1A-M-V_GBL9016830	ARID1A	ARID1A	Validated	Mouse	Abgent	AT1188a
mTOR_pS2448-R-C_GBL9016892	mTOR_pS2448	FRAP1	Use with Caution	Rabbit	CST	2971
ASNS-R-NA_GBL9016894	ASNS	ASNS	Validated	Rabbit	Sigma	HPA029318
VHL-R-NA_GBL9016893	VHL	VHL	NA	Rabbit	BD Pharmingen	556347 (Lot 21483)

XI. HOTNET

Workgroup leaders: Fabio Vandin (vandinfa@cs.brown.edu) and Ben Raphael (braphael@cs.brown.edu)

Contributors: Suzanne S. Fei, Andrew Stout, and Hsin-Ta Wu

We used HotNet [1] to identify subnetworks of a large protein-protein interaction network that contain genes with significant numbers of single nucleotide mutations (or indels) and copy number alterations (CNAs). HotNet considers each mutation or CNA in each sample as a unit heat source, and uses a diffusion process to derive “hot" subnetworks that contain more alterations than expected by chance. Therefore the significance of a subnetwork is determined by both the frequency of alteration of genes in the subnetwork and the local topology of the subnetwork. HotNet returns a list of subnetworks, each containing at least *s* genes, and employs a two-stage statistical test to assess the significance of the list of subnetworks. The first stage of the test computes a *p*-value for the number of subnetworks in the list, for different values of *s*, under a suitable null hypothesis. The second stage estimates the false discovery rate (FDR) of the list of subnetworks, providing a bound on the number of subnetworks in the list that are expected to be significant. Finally, we assess the significance of each individual subnetwork in the list by comparing to known pathways and protein complexes (see below).

We analyzed the combined mutation and copy number data for the both the core set of 372 ccRCC samples and the extended set of 446 ccRCC samples. The subnetworks obtained from the two analyses are similar. Here we describe the results obtained from the extended set of samples, since with a larger number of samples we obtain a better characterization of the aberrations in each subnetwork. For each sequenced gene, we defined the gene as altered in a sample if the gene had a non-silent somatic mutation, or if the gene was present in a focal aberration (heterozygous or homozygous) according to GISTIC analysis. We used the wide peaks output by GISTIC to identify the genes present in a focal aberration, discarding segments that spanned more than half of the chromosome arm. We also discarded aberrations for which the mutations and copy number segmentation did not provide enough supporting evidence to identify a strict target region for the aberration. In particular, we discarded an aberration if the region common to the segments defining the aberration contained fewer than 5 segments unique to the aberration or fewer than 2 non-silent somatic mutations in genes in the aberration. Moreover, we discarded CNAs for which the “sign” of the aberration (i.e. amplification or deletion) was not the same in at least 90% of altered samples and also discarded CNAs in regions that were (concordantly) altered in less than 2% of the samples.

The resulting alteration data on ccRCC 446 samples was input to HotNet. We used the interaction network derived from the Human Protein Reference Database (HPRD) [2]. We also ran HotNet using the interaction network derived from Pathway Commons [3], obtaining results similar to the ones obtained from the HPRD network; in this section we describe the results obtained using the HPRD network. For the HotNet statistical test, we generated random datasets in the following manner. We simulated mutations using the estimated background mutation rate (1.06×10^{-6}). We simulated CNAs using the observed distribution of CNA lengths and permuting their positions. The latter minimizes potential artifacts resulting from functionally related genes that are both neighbors on the interaction network and close enough on the genome that they are affected by the same CNA. To further reduce such artifacts, we removed candidate subnetworks returned by HotNet that contain 2 or more genes in the same focal CNA in at least one of the samples. We also removed genes that are potentially biased toward a higher number of silent mutations than expected (because of their length or higher background mutation rate).

Using this approach HotNet identified 25 candidate subnetworks containing at least 2 genes ($p \leq 0.007$) with a corresponding FDR = 0.68 for the list of subnetworks. The FDR is a conservative estimate of the ratio of false positives among all subnetworks reported by HotNet, and implies that (around) one third of

the subnetworks reported by HotNet are significant. All 25 subnetworks remained after CNA filtering (Table S15 and Figure S45).

To gain additional support for individual subnetworks and to focus attention on subnetworks with known biological function, we computed the overlap between the genes in candidate subnetworks and: (i) known pathways from the KEGG database [4]; (ii) protein complexes from PINdb [5]. Of those 25 subnetworks returned by HotNet, 4 had statistically significant (corrected $p \leq 0.05$) overlap with at least one KEGG pathway or PINdb protein complex (Table S15). Among the most significant subnetworks, are: a subnetwork containing VHL and many of its interacting partners in the VHL complex (CUL2, Elongin C, Ubiquitin gene USP33) and the hypoxia-inducible factor (HIF) genes; and a subnetwork containing three genes (PBRM1, ARID1A, SMARCA4) in the SWI/SNF chromatin remodelling complex. In particular, the chromatin remodeling subnetwork was altered in 39% of samples (main text). ARID1A is a tumor suppressor in gynecologic cancers [8], and has been identified as frequently mutated in ovarian clear cell carcinoma [9]. Another subnetwork identified by HotNet contains the interacting genes PTPRD and MTSS1, both identified as tumor suppressors in other cancers type [6,7].

References

1. Vandin, F., E. Upfal, and B. Raphael, Algorithms for Detecting Significantly Mutated Pathways in Cancer, J Comput Biol. 2011 Mar;18(3):507-22.

2. Keshava Prasad, T.S., et al., Human Protein Reference Database--2009 update. Nucleic Acids Res, 2009. 37(Database issue): p. D767-72.

3. Cerami et al. Pathway Commons, a web resource for biological pathway data. Nucl. Acids Res. (2010).

4. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.

5. Luc, P.V. and P. Tempst, PINdb: a database of nuclear protein complexes from human and yeast. Bioinformatics, 2004. 20(9): p. 1413-5.

6. Veeriah S et al. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9435-40.

7. Fan H. et al. MTSS1, a novel target of DNA methyltransferase 3B, functions as a tumor suppressor in hepatocellular carcinoma. Oncogene. 2012 May 3;31(18):2298-308.

8. Guan B, Wang TL, Shih IeM. ARID1A, a factor that promotes formation of SWI/SNF-mediated chromatin remodeling, is a tumor suppressor in gynecologic cancers. Cancer Res. 2011 Nov 1;71(21):6718-27.

9. Jones S, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science. 2010 Oct 8;330(6001):228-31.

SUBNETWORK	KEGG PATHWAYS ENRICHMENTS		PROTEIN COMPLEXES (PINdb) ENRICHMENTS	
	Name	*p*-value	Name	*p*-value
HIF3A RPS6KA1 EPAS1 PHF17 VBP1 PPP1R3A TOB2 NR4A3 TCEB1 HIF1A VHL POLR2G RNF139 USP33 CUL2	Renal cell carcinoma	3.63E-04		
NRG1 NRG3 NRG2 ERBB4	ErbB signaling pathway	5.66E-05		
FGFR2 FGFR4 CDH2 FGF3 FGF2	MAPK signaling pathway Regulation of actin cytoskeleton	8.89E-03 8.32E-03		
ARID1A SMARCA4 MLLT1 PBRM1			NUMAC	4.87E-02
EPHA6 EFNA4				
PTPRN2 SPTBN4 CKAP5				
DST CELSR3 KIAA1549 KRT5 COL17A1				
NUDC PAFAH1B1				
NR1I3 NR0B2 NR5A2				
RARA NSD1 ZNF496 THRA				
MFAP5 FBN2 MFAP2 LTBP1 FBN1				
MAX MAGEA11 MXD3				
SEC13 SEC16B SEC31A				
C3orf10 NCKAP1				
TESC SLC9A1 MAP4K4				
DHDDS LOX				
COL4A3 USH2A				
TRAPPC2 CLIC2				
PCSK6 FLG				
SFN CHST1				
PTPRD PPFIA2 PPFIA3 MTSS1 PPFIA1				
CIC SETD2				
XPO5 ZNF346 ILF3				
ACE2 GHRL				
PKHD1 CAMLG				

Table S15

Table S15. The 25 subnetworks identified by HotNet, and corresponding enriched KEGG pathways or PINdb protein complexes. For the enrichments, the name of the pathway (or protein complex) the subnetwork is enriched for and the (multiple hypothesis corrected) p-value of the hypergeometric enrichment test are reported.

Figure S45

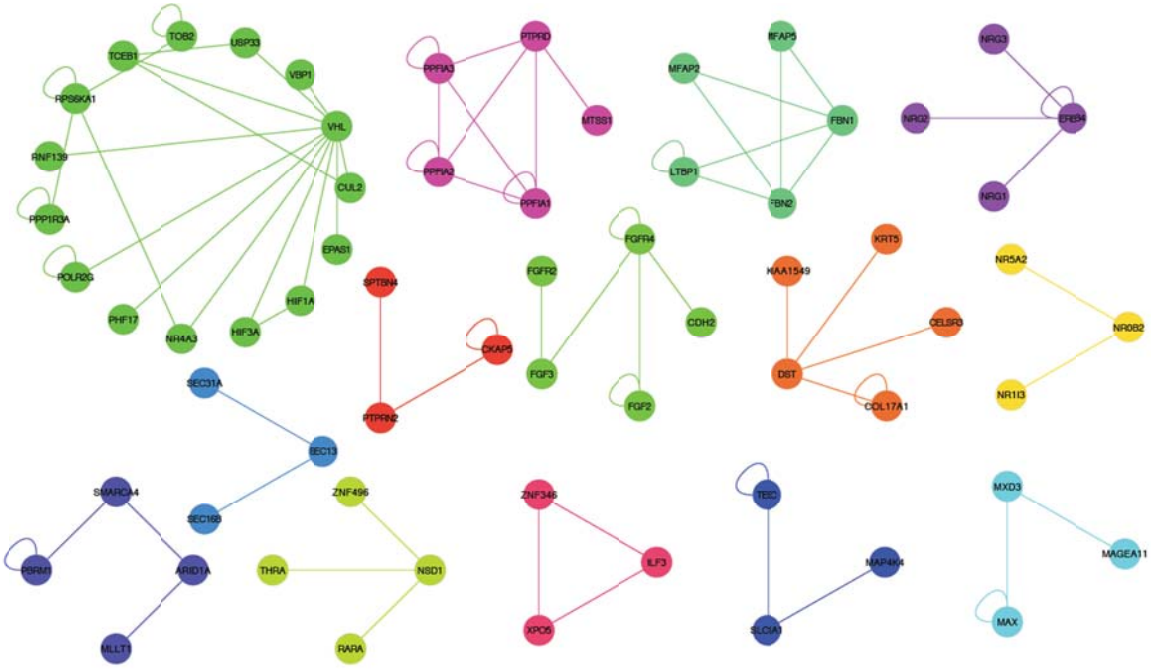


Figure S45. The subnetworks with 3 or more nodes identified by HotNet. Nodes correspond to proteins, and are colored using a different color for each subnetwo rk. Edges correspond to interations in HPRD [2].

XII. PARADIGM

Workgroup leader: Josh Stuart (jstuart@soe.ucsc.edu)
Contributors: Evan Paull, Peter Waltman, Ted Goldstein, Sam Ng

Inferring gene activity from pathway analysis of copy number and expression data. Integration of copy number, mRNA expression and pathway interaction data was performed on the 416 samples using the PARADIGM software [1]. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. The mRNA data was converted to relative mRNA expression levels by subtracting each gene's median computed over 32 tumor-adjacent normal controls from its level observed in each patient sample. Level 3 copy number data (segmented and normalized to reflect the difference in copy number between a gene's level detected in tumor versus normal blood) was mapped to the genome using the UCSC hg19 Knowngenes track. Gene-level copy number estimates were then derived by taking the median of all segments falling within the length of the gene. Both expression and gene-level copy number data were then rank transformed before use by the PARADIGM analysis.

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov, the Reactome database from http://reactome.org, and the set of signaling and metabolic pathways in the last public release of the KEGG database. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway).

Enzymatic reaction pathway information was included into the SuperPathway so that it would be possible to detect changes in metabolism in the tumor samples. To this end, human global metabolic pathway maps covering approximately 120 metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [2,3] were downloaded. The metabolic pathway is described in the KEGG Markup Language (KGML), an XML representation [4]. We parsed the KGML pathway file to extract pathway entities and their interactions to generate a PARADIGM pathway file to fold into the merged SuperPathway.

Briefly, the metabolic global pathway map in KGML is represented as a graph, comprising nodes (entry elements) and arcs (relation and reaction elements) [4]. An entry element in the KGML file can represent a gene or a family of genes, a chemical compound, or another pathway map, as specified in the "type" attribute. We extracted all gene entities. We parsed KEGG-specific human gene IDs (prefix: hsa) from the name attribute of the entry. When more than one gene of the same gene family was listed in the name attribute of an entry, we considered that the genes comprised a family. To map KEGG-specific gene IDs to standardized gene names, we downloaded a table of 26,222 mappings from KEGG gene IDs to gene symbols using the KEGG genes LinkDB search tool on April 10, 2012. We also downloaded a table of 19,026 protein-coding gene symbols and their synonyms from the HUGO Gene Nomenclature Committee on April 10, 2012 [5]. KEGG gene IDs were mapped uniquely to HUGO gene symbols. Small molecules were also extracted and represented as features in the SuperPathway. We extracted all Protein-Protein interaction ("PPrel") relations. The different types of the "PPrel" relations included binding/association, activation, expression, inhibition, which were represented as a complex, activation, transcriptional activation, and inhibition in the PARADIGM pathway file. The reaction element in KGML represents enzymatic reactions, which are unique in metabolic pathways. For example, in a simple reaction in which A is the enzyme, B substrate, and C the product, we considered A and B as a complex AB, and C to be "activated by" AB.

Genes, complexes, and abstract processes (e.g. “cell cycle” and “apoptosis”) were retained and referred to collectively as pathway *features*. Before merging gene features, all gene identifiers were translated into HUGO standard identifiers. All interactions, even those introducing cycles and conflicting paths were retained as PARADIGM’s inference procedure has been shown to be robust to both circular and contradictory regulatory logic that may reside within pathway databases or as a result from the merging of databases. A breadth-first traversal starting from the feature with the highest number of interactions was performed to build one single component. The resulting pathway structure contained a total of 20,713 concepts, representing 7706 proteins, 8998 complexes, 1700 families, 55 RNAs, 15 miRNAs and 582 processes.

To reveal pathway signatures associated with molecular subtypes of disease, we clustered the PARADIGM results both in gene and sample dimensions. For the gene-dimension clustering, we used the hierarchical agglomerative procedure implemented in Cluster 3.0 [6] with uncentered Pearson correlation and average linkage as clustering parameters. To limit the amount of redundant pathway features that may heavily influence visualization and calculation of clinical correlates, we restricted the clustering to single protein level features, excluding features such as complexes and families.

For clustering in the patient sample dimension, ConsensusClustering (CC) revealed that the optimum number of clusters was five based on identifying the maximum number of clusters yielding an appreciable increase in the area of the cumulative density function of CC’s association matrix between consecutive values of the cluster number parameter (Figure S46). The resulting heatmap revealed several subtypes with heightened activities in several major sub-pathways driven by transcription factor hubs including AKT and PI3-kinase signaling, MYC/Max complex activity, AP-1, ErbB1 signaling, TP53 effectors, FOXA1, and p38-alpha signaling. In this cohort of kidney cancer, HIF1-alpha appears to have heightened activity across most of the samples (Figure S47).

We asked whether the subtypes implied by the clustering of PARADIGM-inferred activities revealed a clinically meaningful division of the patient samples. Kaplan-Meier analysis on the 5-cluster result using overall survival as a clinical outcome variable revealed that sample clustering based on PARADIGM IPLs indeed correlated with days to death ($P < 0.003$; Figure S47b).

PathMark identification of connected mechanisms of activity. In order to identify pathway features most altered in tumor samples compared to normal, pathway features were scored based on the average value of the unsigned PARADIGM IPLs. A sub-network was constructed in which only those interactions connecting two features with at least one standard deviation above the average absolute IPL score were retained. This procedure revealed a network containing 1218 pathway features connected by 2398 interactions, of which 645 were proteins (8.4% of the SuperPathway). The largest connected component (LCC) contained virtually the entire solution with 1204 features (637 proteins) connected by 2390 interactions.

To gauge the significance of the resulting network map, we repeated this procedure using PARADIGM IPLs produced from a set of random control samples where each was created by permuting genes in the SuperPathway diagram, thus associated random data tuples throughout the SuperPathway. Five random controls were created for each patient sample, providing 2080 total random controls. From these random controls we sub-sampled a random 416 to create a simulated random cohort from which a new PathMark solution could be calculated. We repeated this sub-sampling 100 times, providing a background distribution against which the observed PathMark could be contrasted. In this case, the size and interconnectivity was highly significant with the average network size from the background set

averaging around 50+/-10 features, more than 10-times smaller than the observed solution (data not shown).

Visual inspection of the PathMark solution revealed several large hubs of altered activity including JUN/FOS, FOXM1, AKT1, and HIF1A. Along with HIF1A, a significant number of genes involved in glycolysis / gluconeogenesis were identified from this analysis (GAPDH, PFKM, PFKL, PGM1, PGK1, ALDOA) (Figure S48).

To identify overrepresented pathways in the solution, a hypergeometric test was performed by overlapping each constituent pathway used to construct the SuperPathway with either the entire PathMark solution or the largest connected component (LCC) within the PathMark solution. We tabulated the top 25 enriched pathways found with this procedure (Table S16) and made the full list available (Table S17).

The analysis highlighted the putative role of the hypoxia inducing factors HIF1A and HIF2A as both transcription factors were found to regulate several targets extracted in the PathMark solution. Interestingly, HIF1A itself is not upregulated but is inferred by the PARADIGM algorithm as highly activated in most samples due to the upregulation of many of the targets of the HIF1-alpha/ARNT transcription factor complex. The targets of this complex impinge on various transformational pathways involving hypoxic response, angiogenesis, and metabolic rewiring that may characterize these tumors. For example, a PathMark diagram centered on the HIF1A/ARNT complex (Figure S48b) illustrates several examples including the egl 9 homologs, EGLN1 and EGLN3, involved in oxygen sensing and response that influence the stabilization of HIF1, and the enzymes aldolase A (ALDOA) and enolase 1 (ENO1) involved in glycolysis. Interesingly, the PathMark solution also includes AKT1 that post-translationally activates HIF1/ARNT, which provides a link between the PI3-kinase growth pathway, oxygen homeostasis, and glycolytic shift.

Differential pathway signature correlation analysis (DiPSC) to identify connections between genomic events and clinical outcomes. The pathway activities inferred by PARADIGM were used to construct “signatures” indicating the presence or absence of a particular tumor attribute of interest. Tumor attributes were either genomic perturbations, such as a mutation in VHL, or clinical outcomes such as tumor stage. Signatures were derived by classifying samples into dichotomous sets based on the presence versus the absence of a tumor feature. For example, each significantly mutated gene was used to dichotomize the samples into two groups; for example, those with a VHL mutation and those without the mutation. The IPL values for all proteins were then provide as input to the Significance Analysis of Microarrays (SAM) procedure [8] to determine the Differential IPLs (DPLs). SAM uses a dichotomization of the samples to calculate a t-test-like measure of significance but using a variance correction to avoid inflated significance due to low-variance effects. The DPL scores are then collected into a single vector and referred to as the DiPSC signature.

Comparing the similarity in dichotomy signature then allows the relationships between the clinical and genomic characteristics to be assessed. The Pearson correlation between all pairwise clinical- and mutation-derived DPL signatures were calculated. However, because two dichotomies may share an overlapping set of samples their correlation could be merely due to high sample overlap. While sample overlap may be one indicator of a statistical association between two events it would also trivially induce the same pathway signatures. To compensate for this effect and derive a measure of correlation due only to the similarity in pathway signatures not influenced by sample overlap, a statistical sub-sampling procedure was performed to disentangle the correlation present in non-overlapping samples from overlapping samples. Briefly, 1000 bootstrap iterations are conducted, each time dividing the

cohort into two random, mutually-exclusive groups of samples A and B. A signature for the first dichotomy is calculated using only the samples in A; a signature for the second dichotomy is calculated using only the samples in B. The Pearson correlation between the two signatures are then recorded for each bootstrap replication. The mean and standard deviation across bootstrap replications is then recorded for each dichotomy pair and used as the similarity measure in the DiPSC dichotomy-by-dichotomy plot (Figure S49).

For DiPSC analysis we included all 39 significantly mutated genes from the MutSig results on the pre-validated MAFs, focal copy number events, and several clinical variables such as tumor grade, tumor stage, lymph node status, and patient survival information. Also included were cluster assignments from the mRNA and miRNA analyses. The DiPSC analysis reveals several large groups of related events sharing the same signatures. The upper-left sub-matrix of the DiPSC plot reveals a good prognosis subtype characterized by a cluster of patients with overall better outlook as indicated by the lack of lymph node spread, grade 1 and stage 1 tumors. Interestingly, PBRM1 and VHL mutations were found to be associated with this subtype, which may be consistent with previous findings. Also correlated with this subtype are ARID1A mutations the fourth mRNA cluster, first miRNA cluster, and normal blood work.

In contrast, DiPSC also reveals a subtype of more advanced disease identified as the middle and bottom right overlapping sub-matrices containing patients that are deceased or have tumor samples detected in at least one lymph node, higher tumor stage and grade, and prior presentation of the disease. The moderately advanced subtype defined by the middle sub-matrix associates several genomic events and implicates them with stage II disease including PTEN deletions and mutations in MTOR. The common signature for PTEN and MTOR are expected given the well-known roles of these genes in PI3K-AKT-MTOR signaling pathways. The lower right sub-matrix appears to correlated with advanced stage and grade and implicates several genomic events that may serve as indicators of advanced disease. Among the genomic events are deletions in the well-known p16 tumor suppressor CDKN2A, mutations in the chromatin-related genes BAP1 and JMJD6.

Several genomic events are of particular interest and deserve further investigation. First, deletions of PTEN are here implicated in moderate disease while, on the other hand, mutations in PTEN are associated with advanced disease. Thus, the nature of genomic disruption in PTEN may lead to very different cellular states compared to mutations in PTEN. While this result is not surprising given the known array of pathways and interactions both at the cell surface and in the nucleus for this tumor suppressor [9], cataloging the different pathway signatures associated with PTEN may provide relevant treatment information for triaging patients with either form of a disruption in PTEN's normal function.

The DiPSC results, coupled with the observations from the mutation frequency, suggest that early gateway mutations in either VHL or PBRM1 provide a genetic background in which several equivalent events can take place and aggravate disease. For example, only six samples harbored mutations in NFE2L2, an oncogene frequently found with gain-of-function mutations in lung squamous carcinomas that interacts with KEAP1. The DiPSC analysis suggests mutations of NFE2L2 in kidney also may be oncogenic as in lung, sharing signatures with CDKN2A deletions and MTOR mutations. Finally, the DiPSC associations now provide information about mutations in more rarely mutated genes such as EME1 and C6orf146, which may also give insight into the variety of mechanisms by which this tumor type progresses.

TieDIE identification of connections between genomic perturbations and transcriptional changes. *TieDIE Method.* We asked whether the genomic perturbations were significantly associated

with the transcriptional hubs identified by the PARADIGM analysis. To this end, we developed an integrative approach called Tied Diffusion through Interacting Events (TieDIE) to search for significant interconnections between genomic perturbations and downstream transcriptional changes. Like HotNet, TieDIE uses a heat diffusion process to identify relevant pathways. TieDIE can be distinguished from HotNet in that it takes as input two distinct sets and searches for interlinking pathways connecting the genes in the two sets to one another. It uses a set of *sources*, in this case the mutations in significantly mutated genes, and a set of *targets*, in this case transcriptional hubs, whose state in the tumor cells is assumed influenced by one or more of the upstream sources. TieDIE then diffuses heat from both the sources and targets to determine a *linker* set of genes as those that gain more heat from the two diffusion processes than would be gained from diffusion from only the sources or the targets alone. The method then identifies a spanning tree connecting the source, target, and linker sets using a fast algorithm to approximate a solution to the Prize Collecting Steiner Tree problem available as part of the BioNet package [7]. For each solution network, the TieDIE algorithm computes an *influence score* measuring the degree to which the proportion of diffused heat ends up on a common intersecting set of genes between sources and targets (manuscript in preparation).

TieDIE Significance Analysis. We determined if the TieDIE solutions were significant by performing a constrained permutation analysis to evaluate the significance of the resulting influence scores. One random simulation was generated by permuting the set of sources while maintaining the given set of targets. A source set that contains a lot of *hubs*, genes with a large number of connections, could produce a significantly interlinked network due trivially to the fact that many targets are more likely to be reached from paths emanating from hubs. The random simulation therefore needs to control for the degree distribution represented among the sources. We therefore performed a constrained permutation of the sources such that random genes selected to be the *i*th source had approximately the same number of neighbors. To do this, we sorted all of the genes by their degree. We then created non-overlapping bins by collecing *K* consecutive genes from the sorted list and putting them into the same bin together. Note that it is possible to include multiple sources in the same bin using this procedure, which makes the overall random model more conservative. The bin size, *K*, was chosen to be *n**10, where *n* was the number of sources supplied. In this case, *n*=19 so bins of 190 genes were created. Permutations were performed by permuting within each bin only to create swaps among genes of approximately the same size. Once all genes were swapped with another gene in the same bin, the TieDIE algorithm was repeated and a random influence score was recorded. The influence score of the network determined for the original dataset could then be compared to the background distribution obtained from this permutation analysis.

TieDIE Application to KIRC dataset. For the sources we selected any MutSig gene with a q-value of 0.05 or smaller, which resulted in 19 genes (out of the original 39 significant at the 0.10 level found by MutSig using the pre-validation MAF) for use as sources to TieDIE. All transcription factors with an average unsigned IPL greater than unity that also had at least one transcriptional target with an IPL above unity were selected as “active hubs.” An IPL above unity corresponds to having an activity of at least one standard deviation more extreme than levels seen in normal controls. These selection criteria allowed active transcription factors with relatively few targets to be included while controlling for well connected "hub" transcription factors found to be inactive given the pathway context. This resulted in the selection of 115 transcription factors. The TieDIE solution was found to be highly significant using a conservative background model determined with constrained permutations (Figure S50). The resulting network contained 529 genes connected by 10,707 interactions (3396 HPRD-PPI, 4052 regulatory, 3259 component; $p < 0.017$). In this network, 14 (74%) of the sources were connected by some path to 115 (100%) of the targets involving 400 interconnecting linking genes.

Pathway enrichment analysis revealed that 5 genes (PBRM1, SETD2, BAP1, KDM5C, and ARID1A) participating in chromatin remodeling were overrepresented beyond chance expectation in the list of 19

significantly mutated genes based on MutSig analysis. To elucidate the pathways that may be disrupted due to mutations in these chromatin genes we extracted a chromatin-related sub-network of the TieDIE solution. A graph traversal was used to search for paths linking mutated chromatin genes to those genes with gene expression levels significantly correlated or anti-correlated with mutations in any of the chromatin genes. Correlation was determined by performing a t-test in which samples having a mutation in one of the 5 chromatin gene were grouped into one set and those without a mutation in any of the 5 were grouped into a second set. A two-sided t-test was then calculated for each gene using either the gene's expression levels or its IPLs from PARADIGM and links connecting genes with t-statistics with and FDR <= 0.10 were retained. A depth-first search was then used to find all paths connecting the chromatin-related genes to IPL-correlated "signaling layer" genes, through up to one "linking" gene. Similarly, we connected the IPL-correlated signaling genes to expression-correlated "output" genes through active transcriptional hubs. The final subnetwork was defined as the union of all complete paths connecting the chromatin-related genes to "output" genes through the linker, signaling and transcriptional-hub layers.

The TieDIE solution gives clues into the various pathways affected by modulation in the chromatin-related genes (Figures S51-S52). Of particular interest is the finding that the chromatin complex made up of PBRM1, ARID1A, and several SMARCA proteins was found to interact with NFKB1. In addition, central to the network are genes involved in TGF-beta and Wnt-related signaling. For example, beta-catenin (CTNNB1) has higher activity in non-chromatin mutants. This suggests that the more advanced disease stages are driven by pathways involving beta-catenin activation, which in turn would then activate such targets as MYC leading to well-known de-differentiation programs seen in many aggressive cancers.

Several interlinking genes in the TieDIE solution are of particular interest because they may be overlooked when the data is analyzed without consultation of the known and/or predicted pathway interaction logic. For example, JUN, FOS, and SP1 are major transcriptional regulators that are inferred by PARADIGM analysis to be active in many of the samples (see outer rings of these genes in Figure S51). However, neither the inferred activities nor the expression of these genes is associated significantly with chromatin-specific genomic perturbations. However, these transcription factors together interconnect several genes that are associated with chromatin mutations from the signaling layer, such as COBRA1, to genes in the transcriptional output layer, such as estrogen receptor (ESR1), TGF-beta, and IL6 differential expression. The existence of such connections suggest that mutations in chromatin modifiers enable particular transcription factor linkers such as JUN and FOS to express sets of growth factor receptors, cyclins (e.g. CCNB1) and interleukins leading to a global turn-over in the signaling circuitry of tumor cells.

References

1. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.Bioinformatics. 2010;26:i237–i245.
2. Okuda S, Yamada T, Hamajima M, et al. KEGG atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 2008;36:W423–26.
3. KEGG pathway database. http://www.genome.jp/kegg/pathway.html. KEGG. 2012
4. KEGG markup language. http://www.genome.jp/kegg/xml/docs/. KEGG. 2012.
5. HUGO gene nomenclature committee. http://www.genenames.org/cgi-bin/hgnc_stats.pl. April 2012.

6. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95:14863-8.
7. Daniela Beisser, Gunnar W. Klau, Thomas Dandekar, Tobias Müller, Marcus T. Dittrich: BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics 2010 26(8): 1129-1130.
8. Tusher, Tibshirani and Chu. "Significance analysis of microarrays applied to the ionizing radiation response." *PNAS* 2001 98: 5116-5121.
9. Song MS, Salmena L, Pandolfi PP. "The functions and regulation of the PTEN tumour suppressor." *Nat Rev Mol Cell Biol.* 2012 Apr 4;13(5):283-96

Pathway[a]	P value[b]	Proportion[c]
Direct p53 effectors	1.9E-40	76/164
C-MYB transcription factor network	9.5E-35	74/186
HIF-1-alpha transcription factor network	3.1E-21	47/110
AP-1 transcription factor network	8.6E-20	39/80
E2F transcription factor network	1.0E-18	56/176
ATF-2 transcription factor network	8.9E-18	32/63
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	1.5E-16	33/67
FOXM1 transcription factor network	1.1E-15	36/86
HIF-2-alpha transcription factor network	1.2E-15	24/36
ErbB1 downstream signaling	1.9E-13	47/167
p73 transcription factor network	9.7E-13	37/109
IL12-mediated signaling events	1.2E-12	41/129
Glucocorticoid receptor network	9.4E-12	39/124
Validated targets of C-MYC	2.1E-11	32/87
Validated targets of deltaNp63	2.2E-11	34/98
Regulation of nuclear SMAD2/3 signaling	2.3E-11	38/121
Regulation of retinoblastoma protein	2.4E-11	36/113
PDGFR-beta signaling pathway	3.9E-11	62/292
IL6-mediated signaling events	5.1E-11	30/80
keratinocyte differentiation	1.0E-10	36/118
FOXA1 transcription factor network	2.0E-10	24/57
LPA receptor mediated events	2.3E-10	45/176
IL4-mediated signaling events	2.3E-10	34/105
p38 MAPK signaling pathway	1.2E-09	32/101
Angiopoietin receptor Tie2-mediated signaling	3.1E-09	28/79

Table S16. Pathways enriched in the PathMark solution.

[a] Pathways were taking from those used to build the SuperPathway.

[b] P-values calculated using Hypergeometric test with Bonferroni correction.

[c] Denominator is the number of genes in the pathway; numerator is the number also in the PathMark solution.

Table S16

Table S17

Pathway[a]	P value[b]	proportion[c]
Glycolysis / Gluconeogenesis	5.2E-04	12/63
Fructose and mannose metabolism	2.0E-01	6/31
Pentose phosphate pathway	4.5E-01	5/25
Glutathione metabolism	5.1E-01	6/37

Table S17. Pathways enriched in the PathMark solution when restricting to metabolism-related genes.

[a] Pathways were taken only from the KEGG genes added to the SuperPathway.

[b] P-values calculated using Hypergeometric test with Bonferroni correction.

[c] Denominator is the number of genes in the pathway; numerator is the number also in the PathMark solution.

Figure S46

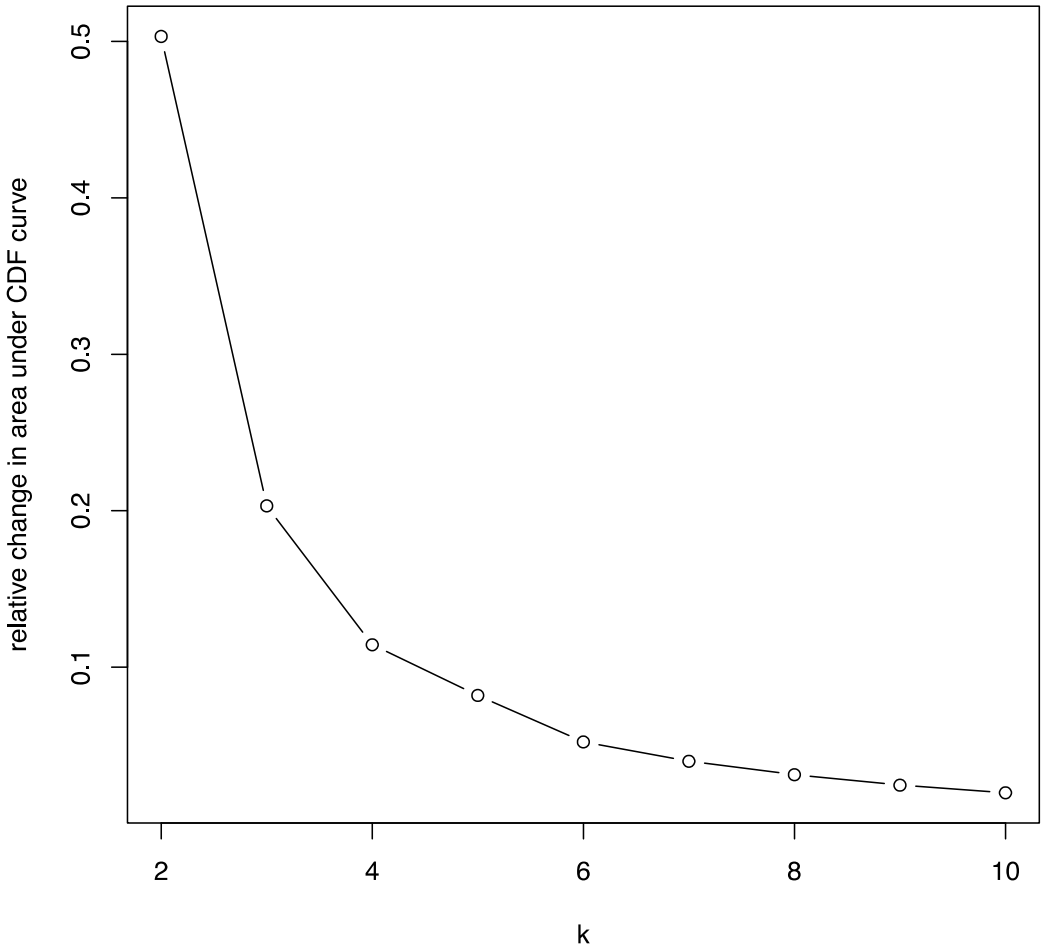


Figure S46. Consensus clustering reveals five clusters for PARADIGM-based subtypes. Consensus clustering analysis using K-means as the baseline clustering revealed that five clusters produced an appreciable increase in the cluster structure as revealed by the drop in the association-matrix CDF plot (*y*-axis) as a function of the number of clusters (*x*-axis).

Figure S47

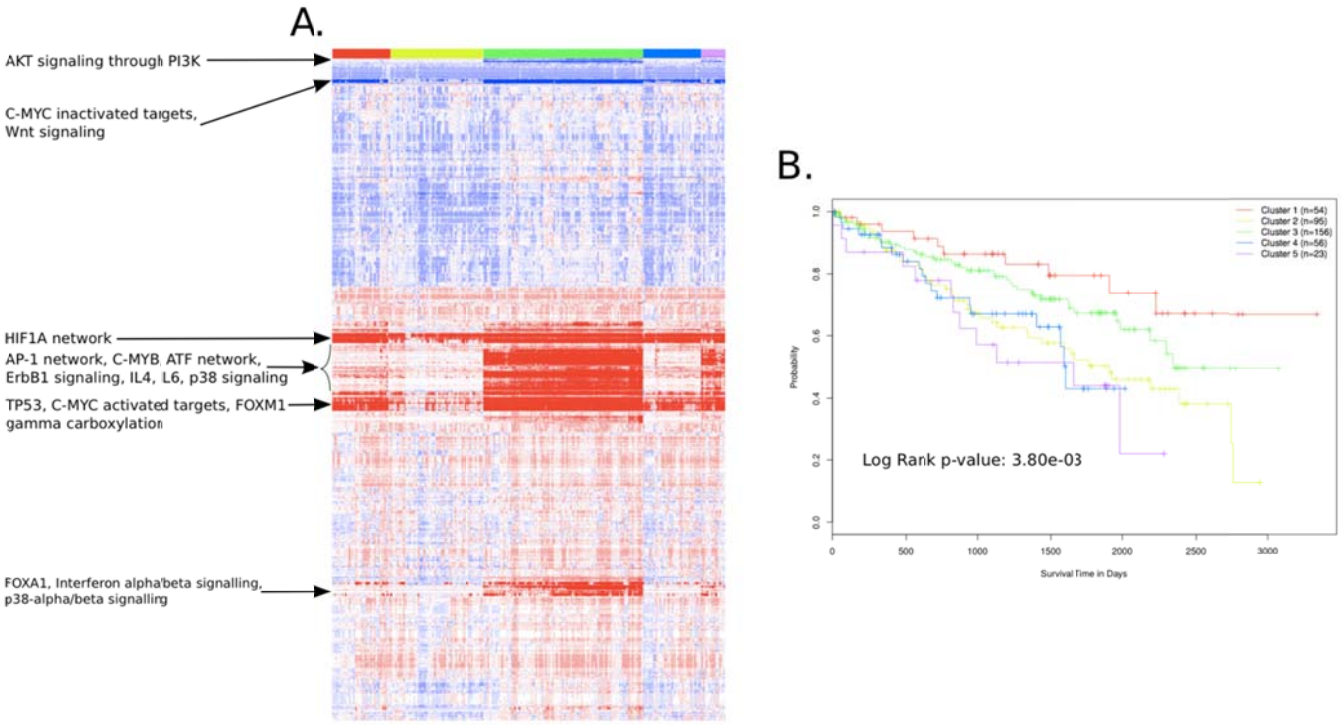


Figure S47. **A. PARADIGM integrated pathway levels.** PARADIGM inferred pathway levels (IPLs) for all protein level features (rows) were clustered in both the protein dimension (rows) and in the sample dimension (columns); higher inferred activity in tumor compared to normal, red; lower activity in tumor compared to normal, blue. Several subnetworks under the control of a diverse set of hubs show distinct patterns of activity some of which are labeled. Consensus clustering revealed five sample clusters indicated with red, yellow, green, blue, and purple bars above heatmap. **B. PARADIGM-based subtypes correlated with overall survival.** Kaplan-Meier analysis revealed that the first sample cluster (red curve) had better overall survival than the other clusters, especially the blue and purple clusters.

Figure S48

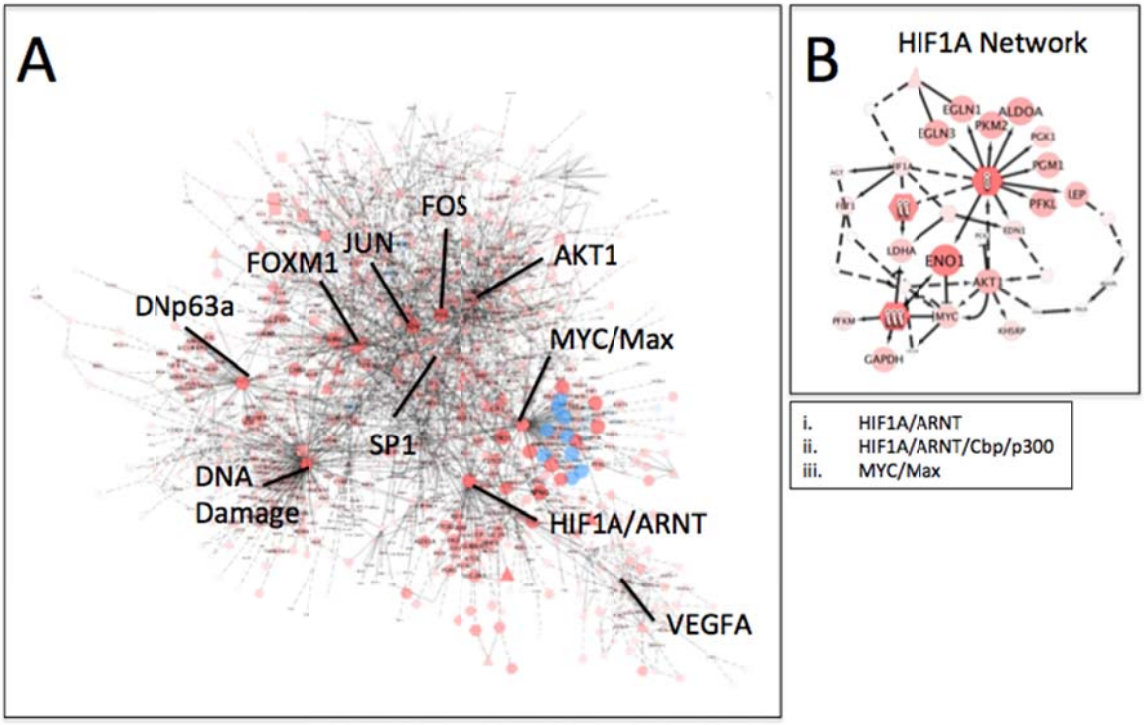


Figure S48. PARADIGM analysis reveals multiple transcription factor hubs driving expression. **A. PathMark solution overview.** PathMark was used to search for subnetworks of interconnected genes with activities either elevated (red) or inactivated (blue) in tumors compared to normal samples. A differential activity for each pathway feature was computed by taking the difference between the average level of each feature in normal samples from the average observed in tumor samples. Interactions in the SuperPathway were retained if they connected any two features with better than average absolute differential activity. Proteins involved in several regulatory connections (hubs) are labeled to provide a general overview. The full Cytoscape session of the network is available as part of the supplement. **B. HIF1A-related subnetwork.** To illustrate some pathways with expression levels heightened in tumor samples under the control of the HIF1A/ARNT complex showing connections between the AKT pathway signaling and metabolism (e.g. ENO1 and ALDOA).

Figure S49

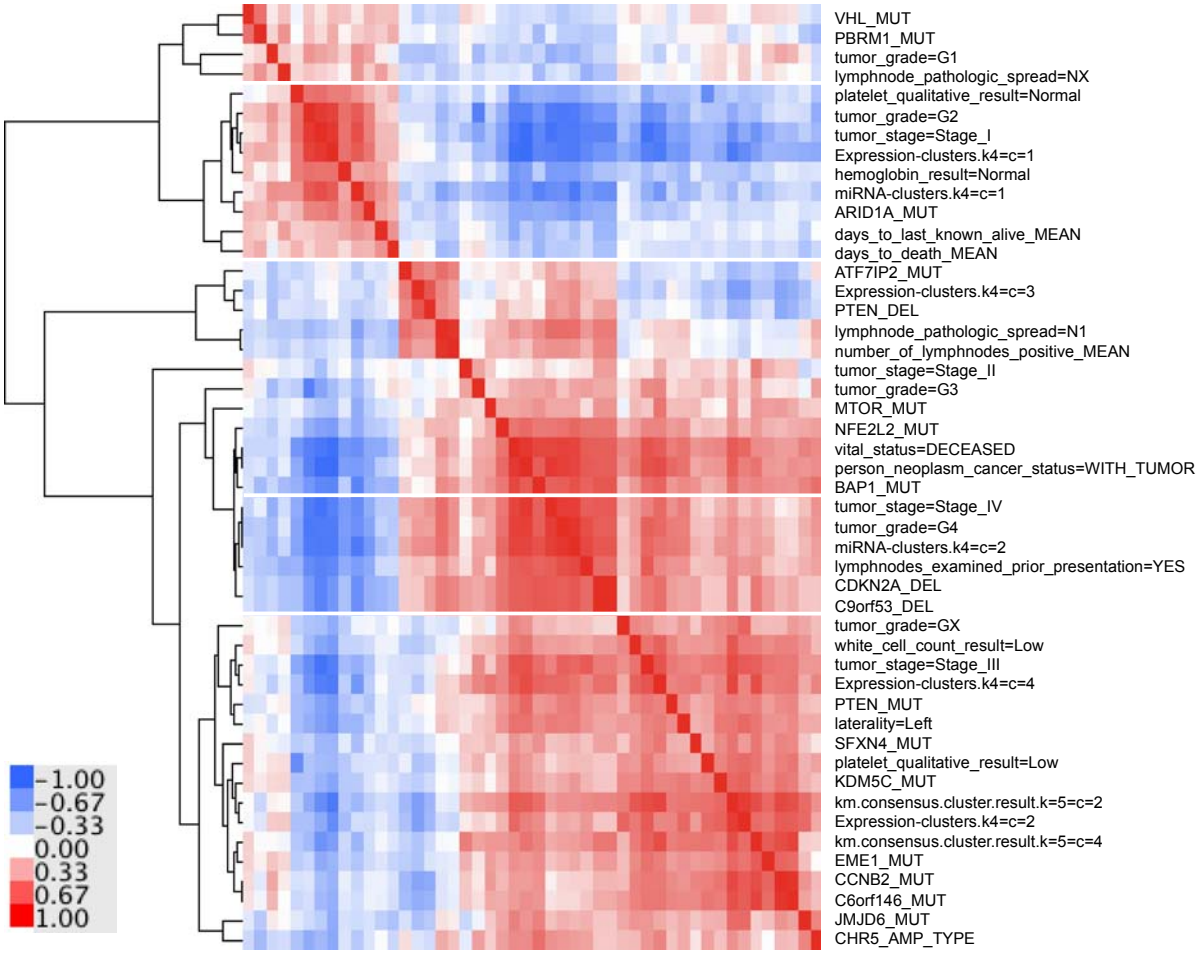


Figure S49. Differential Pathway Signature Correlation analysis reveals connections between genomic perturbations and clinical outcomes. Samples were dichotomized into two groups according to several genomic or clinical variables. From each dichotomy a signature was derived by calculating differential activity of PARADIGM inferred pathway levels for each feature in the SuperPathway representing the activity of the feature in one group of samples compared to samples outside of the group. The signature of a dichotomy (e.g. tumor stage equal to 3 vs. tumor stage not equal to 3) could then be compared to the signature of another dichotomy (e.g. VHL mutation present vs. VHL mutation absent) using Pearson correlation. High Pearson correlations (red squares) indicate genomic events or clinical outcomes associated with similar pathway activities while anti-correlated pairings (blue squares) indicate events are associated with highly different molecular pathway states.

Figure S50

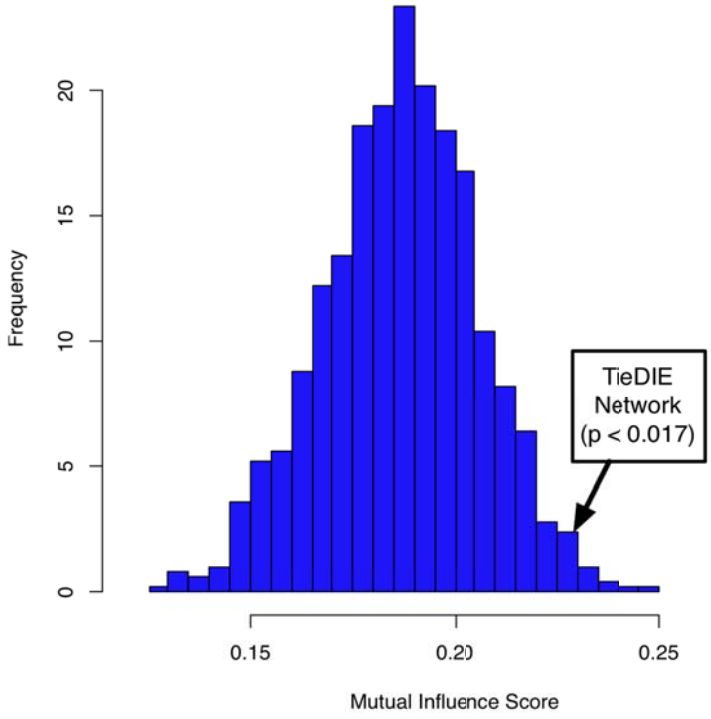


Figure S50. Genomic perturbations in kidney cancers are significantly associated with downstream transcriptional changes through known and novel pathway circuitry. The TieDIE algorithm was used to identify a network connecting the top 28 significantly mutated genes to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls.

Figure S51

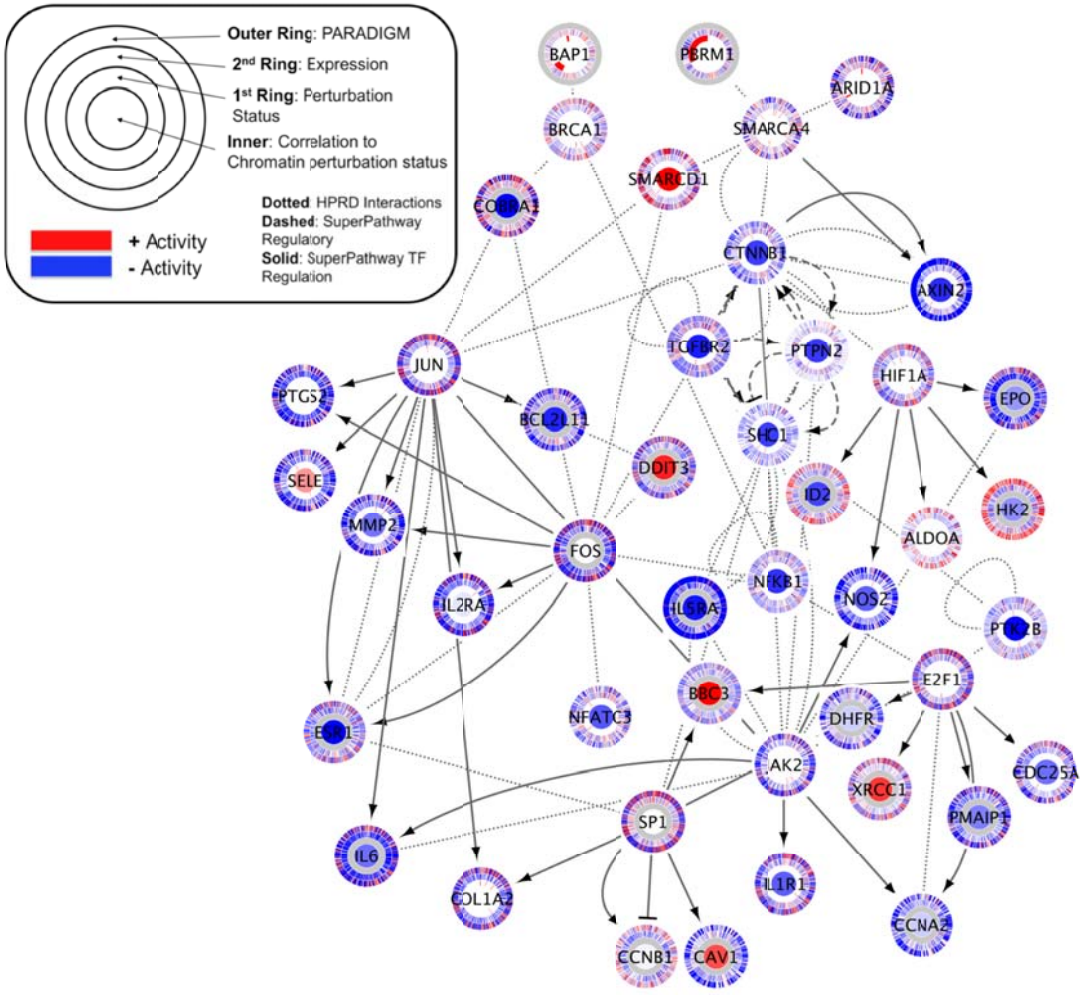


Figure S51. TieDIE identifies interlinking pathways connecting chromatin-related genes to downstream transcriptional hubs. Each circle illustrates a protein in the SuperPathway network. Each tick-mark in each circle represents a single patient sample. Circles represent genomic perturbations, expression, and PARADIGM inferred activities for different proteins. Outer ring represents the inferred pathway level produced by the PARADIGM algorithm. Second-most outer ring represents the gene expression level of the gene. Inner ring represents whether the gene has a mutation or not. Central circle color and shade reflects correlation of either the inferred level or expression of the gene to the mutation status in chromatin-related genes (red, higher activity/expression in mutants compared to non-mutants; blue higher expression in non-mutants). All rings are sorted in the same order according to the presence/absence of a mutation in one of the chromatin-related genes. In the counter-clockwise direction starting from the “noon” position samples are ordered by those with mutations in PBRM1, followed by those with mutations in BAP1, then by those with mutations in ARID1A, and finally samples without mutations in any of these three genes.

Figure S52

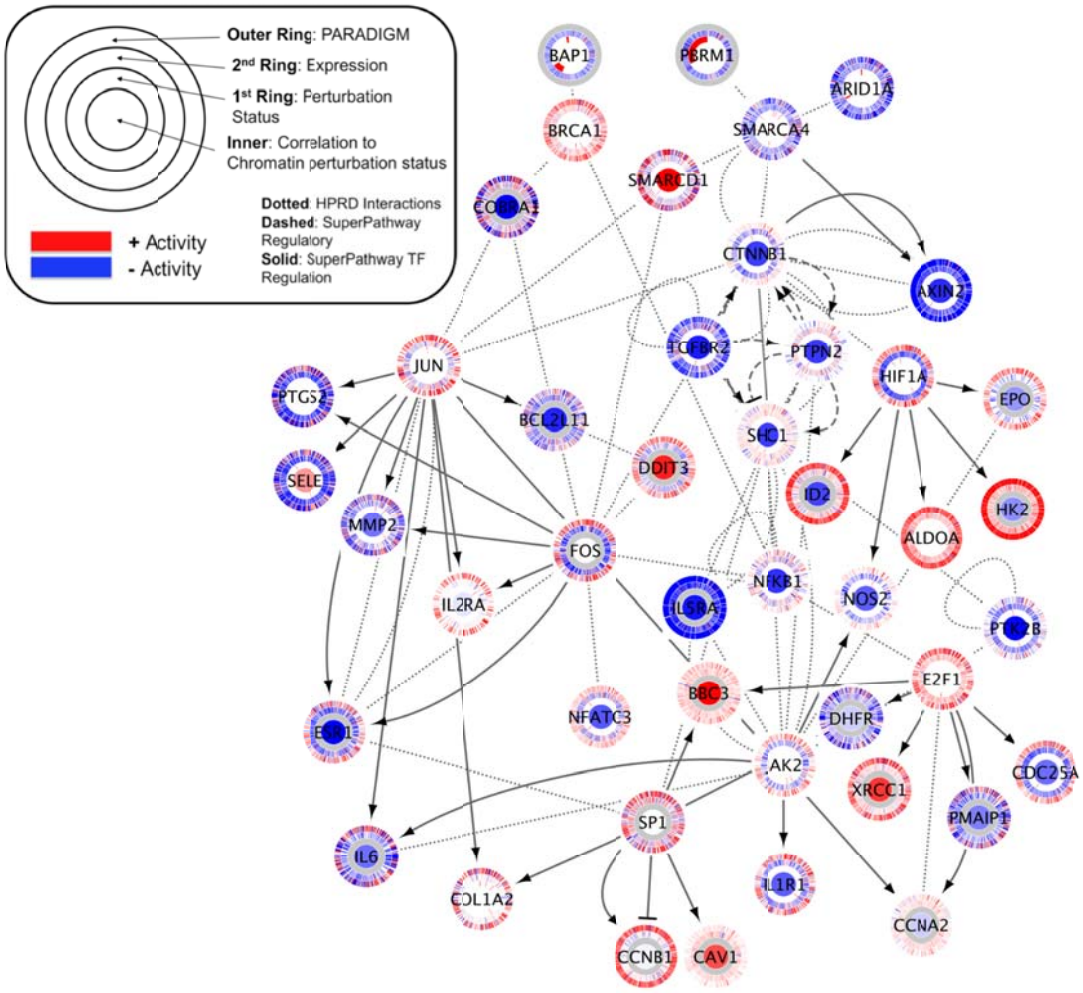


Figure S52. Chromatin-related TieDIE solution, alternate view showing the original, non-discriminant data and inference levels. Same plot as in above except that the original pathway inferences and gene expression levels are shown instead of the differential levels.

XIII. INTEGRATIVE ANALYSIS OF MUTATION AND EXPRESSION

Workgroup leaders: Suzanne S. Fei (feis@ohsu.edu) and Paul T. Spellman (spellmap@ohsu.edu)

Summary. The TCGA clear cell Renal Cell Carcinoma (ccRCC) working group surveyed 446 tumors. Out of the 19 genes determined to be significantly recurrently mutated (MutSig algorithm, q<0.10), five were involved in chromatin remodeling and histone modification. This represents a significant enrichment in chromatin regulators (q=0.041), suggesting a role for chromatin dysregulation in ccRCC.

Due to widespread loss of one copy of chromosome 3p, a number of tumor suppressor genes residing on 3p are more easily inactivated through mutation. Mutations in the most frequently mutated gene, VHL, lead to stabilization of hypoxia inducible factors and increased angiogenesis. After VHL, the three most highly recurrently mutated genes on 3p were PBRM1, SETD2, and BAP1, which participate in chromatin-remodeling and histone modification. It is unclear how or if these chromatin-related genes are contributing to oncogenesis.

To investigate how these mutations may be affecting the cells, we compared gene expression levels between tumors with mutations in these genes to tumors with no evidence of mutations in these genes. Specifically, we identified differentially expressed genes in four separate comparisons: PBRM1 mutants vs. nonmutants, BAP1 mutants vs. nonmutants, SETD2 mutants vs. nonmutants, and, for comparison, VHL mutants vs. nonmutants. We found that the expression levels of many more genes were disrupted in the chromatin-regulator mutants compared to the VHL mutants. These global effects on expression suggest that altered chromatin remodeling can dramatically alter gene expression patterns by leading to major shifts in accessible DNA elements.

We then compared the sets of genes that were differentially expressed in each of the four comparisons. In PBRM1 and BAP1 mutants, similar classes of genes were affected (primarily signaling and glycoproteins) and some of the targets were shared; however, most of the genes were only differentially expressed in either the PBRM1 or the BAP1 mutants. SETD2 mutations, on the other hand, affected several zinc finger and KRAB domain proteins. This reveals that each chromatin regulator appears to have a very distinct set of downstream effects.

Introduction. Due to widespread loss of one copy of chromosome 3p, a number of tumor suppressor genes residing on 3p are more easily inactivated through mutation. Three such genes, PBRM1, SETD2, and BAP1, participate in chromatin-remodeling and histone modification. In addition to the genes on 3p, there are also a number of other chromatin-related genes recurrently mutated in ccRCC (Data File S3 and Table S18). It is unclear how or if these chromatin-related genes are contributing to oncogenesis; however, the list of significantly mutated genes is enriched in chromatin regulators (q=0.041), suggesting a more pervasive a role for chromatin dysregulation in ccRCC. To determine the impact of having mutations in these genes, gene expression in the tumors with the genes mutated was compared to expression in the tumors with no evidence of a mutation in the genes.

Methods. Because RNA-seq data consists of read counts that are used to quantify the expression of a gene, we utilized the edgeR package in Bioconductor (v2.10) (Robinson, McCarthy, & Smyth, 2010) to determine differential expression. This package uses a negative binomial model well-suited for count data. We selected the tagwise method for estimating dispersion. The extended ccRCC TCGA dataset contains both mutation and expression data for 417 tumors. Only genes that had at least one count-per-million in at least 30% of the samples were analyzed. The total number of genes with adequate counts to quantify was 14,686. If any one of the three centers that called mutations on the dataset

found a mutation in the gene, that tumor was included in the mutated group. Tumors with no evidence of mutations in the gene were included in the non-mutated group (Table S18). VHL was also included in the analysis for comparison. For VHL, both mutations called in the mutation calling pipelines and mutations found by subsequent gap filling were included.

Results. The BAP1 mutants vs. non-mutants and PBRM1 mutants vs. non-mutants comparisons had the greatest number of differentially expressed genes. SETD2 had a moderate number of expression differences, and VHL had the smallest number of expression differences (Table S19). The genes that passed the strictest filtering criteria are listed in Table S20. Using the moderate filtering approach in Table S19, we compared the sets of differentially expressed genes for the genes. We found that some genes are found in more than one set; however, most genes are found in only one set (Figure S53). However, considering that 14,686 genes were quantified in total, the sets of differentially expressed genes all overlap significantly more than expected by chance ($p<5E-10$ for all pairwise comparisons). The overlap between gene sets is particularly striking between VHL and PBRM1, the two most frequently mutated ccRCC genes.

To verify that the differences seen between the sets of mutated and non-mutated tumors were greater than expected by chance, we performed 700 permutations of the mutated/non-mutated labels. In the three BAP1 and PBRM1 comparisons, the observed number of differentially expressed genes exceeded the number found in all of the permutations, highlighting that there are many real differences between the mutated and non-mutated tumor sets. Since these mutations may be correlated with variables such as stage and grade, future iterations of this analysis will include these variables as covariates. The results were still highly significant in the SETD2 and VHL comparisons; however, several of the permutations (0.4% and 1.7%, respectively) found a greater number of differentially expressed genes than in the true groups. Although this could apply to all of the mutated genes, the number of differentially expressed genes in the VHL mutated vs. nonmutated comparison, particularly, may be artificially low because VHL is known to also be deactivated through other mechanisms such as hypermethylation, leading to heterogeneity in the VHL non-mutated group.

Enrichments in the sets of differentially expressed genes were determined in DAVID (Huang, Sherman, & Lempicki, 2009a, 2009b) using the 14,686 total quantified genes as the background (Table S21). To account for the potential confounding effects of the other three mutant genes, we looked at enrichments in the set exclusive to only one mutant gene (see Figure S53). The similarities in enrichments between PBRM1 and BAP1 gene sets reveal that they influence similar gene classes (glycoproteins and signaling); however, the actual genes they influence are often different. As expected, the genes they share in common are also strongly enriched in glycoproteins and signaling, indicating they share some common targets as well. SETD2, on the other hand, has weaker enrichments for different classes of genes, such as those containing KRAB domains and various zinc finger proteins.

Conclusions. Several chromatin regulators are recurrently mutated in ccRCC, suggesting the importance of chromatin dysregulation in ccRCC oncogenesis. Three of the regulators reside on chromosome 3p. When comparing tumors with mutations in these three genes to tumors with no evidence of mutations in these genes, we found that the expression levels of many more genes were disrupted in the chromatin-regulator mutants compared to VHL mutants. In PBRM1 and BAP1, similar classes of genes were affected and some of the targets were shared, but most genes were only affected by one of the regulators. The number of genes that are affected may indicate that these genes affect a wide range of targets across the genome. To test this hypothesis, it would be necessary to obtain genome-wide histone location and modification data, preferably in matched tumor-normal samples with and without mutations in these genes.

References

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., Dhir, R., et al. (2011). Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353), 609–615. doi:10.1038/nature10166

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1–13.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44–57.

Oliveros, J. C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams. Retrieved from http://bioinfogp.cnb.csic.es/tools/venny/index.html

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139–40.

Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. Journal of computational biology : a journal of computational molecular cell biology, 18(3), 507–22. doi:10.1089/cmb.2010.0265

Gene	Location	# tumors with this gene mutated (in extended, 1+ centers list)	# tumors with this gene mutated (in extended, 2+ centers list)	Significant by MutSig?	Function
PBRM1	3p21	143	134	yes, #2, q<0.0001	SWI/SNF chromatin remodeling (binding domain of PBAF complex)
SETD2	3p21	49	48	yes, #3, q<0.0001	Histone methyltransferase for H3K36
KDM5C	Xp11	27	27	yes, #4, q<0.0001	Histone demethylase for H3K4
BAP1	3p21	41	40	yes, #6, q<0.0001	Deubiquitinating enzyme for histone H2A and HCFC1
ARID1A	1p35	16	12	yes, #17, q=0.0726	SWI/SNF chromatin remodeling (binding domain of BAF complex)
SMARCA4	19p13	9	6	no, #820, q=0.84	SWI/SNF chromatin remodeling (ATPase of BAF and PBAF complex)

Table S18. Chromatin-related genes containing mutations in ccRCC TCGA dataset. Out of the 19 genes determined to be significantly mutated in the extended (n=446 tumors) dataset (q<0.1), five were involved with chromatin remodeling and histone modification. They are listed in order of significance as determined by the MutSig algorithm (Broad Institute). The final gene on the list, SMARCA4, although not significant as determined by MutSig, still contains a number of mutations that are potentially functional. The HotNet algorithm (Bell et al., 2011; Vandin, Upfal, & Raphael, 2011) found SMARCA4 because it directly interacts with PBRM1 as part of the PBAF complex.

Table S18

Table S19

Significance Threshold	PBRM1	SETD2	BAP1	VHL
q<0.05	3,665	1,164	6,270	384
q<0.005 and |Fold Change| > 1.5	484	166	1,136	105
q<5E-10 and |Fold Change| > 1.5	64	8	117	25

Table S19. Number of differentially expressed genes when comparing tumors with mutations in the specified gene to tumors with no evidence of mutations in the specified gene. Three filtering thresholds are shown. The total number of genes quantified was 14,686.

Table S20

Mutated gene	Differentially expressed genes
VHL	B4GALNT1 BAAT CATSPERG CDH3 CILP CRYGS ENO3 ERP27 FXYD3 GPR143 GPRC5A HSPA6 IGF2 KRT17 MFI2 MMP1 PKD1L2 PLXNA4 PROM2 SCUBE1 SILV SV2B TFF3 THBS4 TMEM158
PBRM1	ADD2 APOL1 ATRNL1 B4GALNT1 BAAT BAG2 BAI1 C1QTNF3 C2 C20orf132 CCDC109B CHN2 CLSTN2 CMYA5 CPN2 CPS1 CYB5R2 DIRAS3 ERP27 FAM40B FBXO41 FMO5 GALNT5 IGF2 IGFBP2 ITIH3 KLHDC7B KRT17 KRT7 LEPREL2 LPAR1 LTF MACROD2 MEST MRGPRF MSMP MUC13 PAPPA PDPN PIGR PROM2 PRR15L PTPRN2 RAMP1 RIC3 RIMS3 SCNN1B SDR42E1 SILV SLC2A10 SLC44A4 SPTBN2 STK33 SV2B SYN1 TACSTD2 TCHH THPO TMC4 TMEM158 TMPRSS3 WDR17 ZDHHC2 ZNF280B
SETD2	ARC CD274 CYP2E1 FAM57A MMP1 PDE1A PRSS12 RCL1
BAP1	AIFM3 AK7 AKAP6 ALOX5 APBA2 AQP4 ARSK ASXL3 ATP1A3 ATP5G1 B4GALT5 BRP44 C19orf33 C2orf7 CACNB3 CCDC78 CISD1 COX7B CXorf57 DCBLD2 DCDC2 EML1 EPB41 F10 FA2H FAM109A FAM196B FAM40B FBXL19 FHL1 FOXP4 GGT7 GPR153 HAGHL HGF HMGN3 HOMER3 HOXA4 HOXD8 IGSF1 IL17RB IL1RL2 IL28RA IL34 ISM1 ITIH3 KCNMB4 KIAA1522 LEAP2 LHPP LOC283392 LOC644538 LRP8 LSM4 MANSC1 MAP6D1 MBOAT7 MDK MEGF6 MFSD10 MGAT4B MIPOL1 MRPS12 MSC MSMP MTSS1 MXD3 NCS1 NEFL NEIL3 NIT2 NKAIN4 NOVA1 NOX4 NPEPL1 NUMBL OSR2 PABPC4L PCYT2 PLEKHA5 PLXNA1 PPIF PRELID1 PRKAR1B PWWP2B RAB6B RAG1 RGS9 RNF43 RPS10P7 RUFY3 SDK2 SEMA4G SFRS13B SFXN5 SLC13A1 SLC25A10 SLC45A3 SNAP25 ST14 STEAP2 STX1A THRB TMEM116 TMEM97 TNFRSF19 TSPO TTC39A TXNDC16 UBE2CBP UNC119 UQCRQ VANGL2 YDJC ZFHX4 ZFP37 ZNF433

Table S20. Most differentially expressed genes when comparing tumors with mutations in the specified gene to tumors with no evidence of mutations in the specified gene. Only the genes with q<5E-10 and |Fold Change|>1.5 are shown.

Figure S53

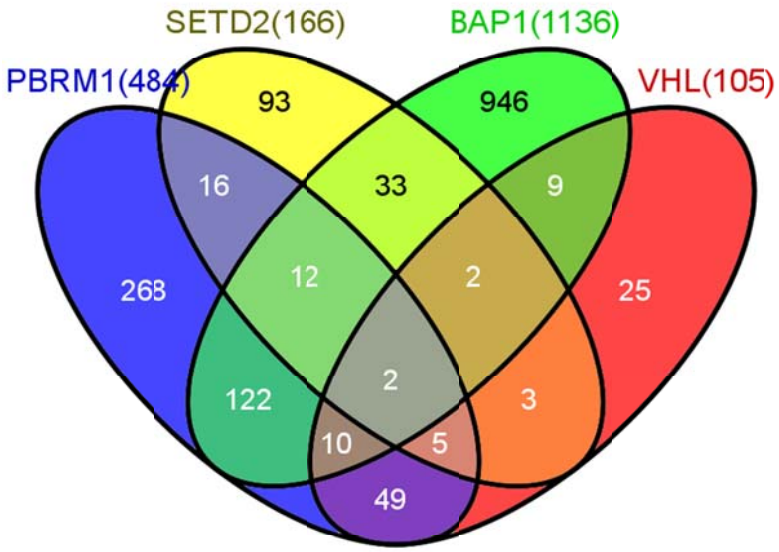


Figure S53. Overlap of differentially expressed gene sets. Lists of differentially expressed genes were obtained by comparing tumors with mutations in the gene to tumors with no evidence of mutations in the gene. The results were then filtered to genes with q<0.005 and |Fold Change|>1.5. The numbers shown in parentheses are the total number of differentially expressed genes. For example, 105 genes are differentially expressed when you compare tumors with VHL mutations to tumors with no evidence of VHL mutations. This was repeated for PBRM1, SETD2, and BAP1. The overlap of the differentially expressed gene sets is shown. (Figure made using (Oliveros, 2007)).

Table S21

Mutated Gene	Enriched class in differentially expressed set	Number of genes in class / total number of differentially expressed in set	q-value
VHL	Disulfide bonds	14/25	6.3E-6
	Signaling	13/25	3.7E-4
	Extracellular Region	12/25	5.2E-4
	Various EGF-like domains	5/25	5.8E-4
SETD2	Prothrombin Activation Pathway	3/93	0.13
	퇀RAB domain proteins	9/93	0.21
	Various zinc finger regions	11/93	~0.40
PBRM1	Glycoproteins	138/268	2.8E-33
	Signaling	121/268	1.6E-33
	Secreted	74/268	1.5E-26
	Extracellular region	83/268	9.5E-23
BAP1	Cell adhesion	57/946	5.8E-10
	Glycoproteins	286/946	3.1E-9
	Signaling	222/946	8.7E-8
	Membrane proteins	374/946	1.8E-7

Table S21. Enriched classes in sets of differentially expressed genes. Gene sets found to be differentially expressed when comparing tumors with a mutation in the listed gene and tumors with no evidence of a mutation in the listed gene. Only genes found exclusively in one gene set were analyzed for enrichments (see Figure S53).

XIV. MUTUAL EXCLUSIVITY MODULES (MEMO)

Workgroup leaders: Anders Jacobsen (jacobsen@cbio.mskcc.org) and Giovanni Ciriello (ciriello@cbio.mskcc.org)

Contributors: Boris Reva, Nikolaus Schultz, Chris Sander

We performed unsupervised analysis of genomic alterations in a pathway context using the MEMo algorithm [1]. MEMo (Mutual Exclusivity Modules) identifies mutually exclusive alterations targeting frequently altered genes that are likely to belong to the same pathway. The algorithm constructs a global network of pathways from several sources (A,B,C,D) and analyzes patterns of pathway alterations based on a general abstraction of gene alteration per sample. Gene alterations belong to one of three categories:

· Category 1: Gene is altered by mutations.

· Category 2: Gene is altered by copy number alterations, where mRNA expression levels correlate with copy number changes.

· Category 3: "Wild-card" events (e.g. gene shows aberrant mRNA expression and/or methylation status correlated with mRNA expression).

Gene mutations were defined by the list of recurrently mutated genes identified by the MutSig algorithm (39 genes significant with FDR < 0.1), and in the initial discovery run we only used somatic mutations which were called by two out of three sequencing center mutation detection pipelines. DNA copy number alterations were defined as amplification (gain of >= 2 copies) or homozygous deletion of genes in the frequently amplified and deleted Regions of Interest (ROI) as identified by the GISTIC algorithm (see description of these algorithms elsewhere in the supplementary text). We only included DNA copy number alterations for a given gene when the gene also displayed significant differential mRNA expression in altered versus diploid cases (Fisher's exact test, $p < 0.05$). We used MEMo to analyze the core freeze set of tumors (N=368), which had all types of data available (mutation, copy number and mRNA expression data).

In a first discovery run of MEMo we observed that the top mutually exclusive patterns included several known components of the mTOR pathway (such as PTEN, PIK3CA, MTOR, adj. $P < 0.001$). To analyze in more detail the extent of possible mTOR pathway activation alterations, we included somatic mutations called by any single sequencing center and added the following gene alterations to the analysis:

- *EGFR* mutations and over-expression (5 mutated cases and 17 cases with mRNA expression >2 standard deviations from the mean)
- Akt mutations (grouping 7 mutually exclusive mutations for *AKT1*, *AKT2*, and *AKT3*)
- *RHEB* mutations (4 mutated cases, 3 of which occurred at same position, Y35)
- *TSC1/TSC2* mutations (grouping 7 mutually exclusive for *TSC1* and *TSC2*)
- *GNB2L1* and *SQSTM1* over-expressed cases, independently of CNA status (11 additional cases: 6 for *GNB2L1* and 5 for *SQSTM1*)

This analysis resulted in the discovery of 8 statistically significant modules (FDR < 0.1; Table S22) of which the top altered modules all comprised different mutually exclusive configurations of known mTOR pathway components (*EGFR, PTEN, PIK3CA, AKT, TSC1/2, RHEB, MTOR*). We also found evidence for mutual exclusivity of a module associated with DNA damage response (*ATM, CDKN2A, TP53*).

Three additional genes also showed mutual exclusivity with mTOR pathway components: GNB2L1, MAPK9, SQSTM1. These three genes were all located on the distal region of the 5q chromosome arm which showed significant recurrent amplifications (5q35.3, GISTIC q = 7.0e-09). The inferred amplification peak, 5q35.3, contained more than 60 genes. In principle any single gene or combination of the genes could be the target of the 5q35.3 amplification, but the three genes specifically highlighted by MEMo were selected by the algorithm because they were connected to the mTOR pathway in the reference network and they showed correlation of DNA copy number and mRNA expression levels.

We further analyzed possible driver genes on the 5q35.3 amplicon in a manner not biased by the pathway reference network used by MEMo. We hypothesized that a driver gene on 5q35.3 capable of activating mTOR signaling, or a factor downstream of mTOR signaling, through genomic amplification would display two features: a) strong correlation of DNA copy number and mRNA expression levels in the set of samples with gain of 5q35.3, b) high mRNA expression levels specifically in the samples without other types of mTOR pathway alterations identified by MEMo. The latter feature relies on the hypothesis that mutual exclusivity between genomic alterations indicate functional relationship. We screened all genes (N = 64, having both copy number and mRNA expression data) in the 5q35.3 region for these two associations: we use Pearson correlation coefficient to compare mRNA and DNA copy number levels, and a Kolmogorov-Smirnov test to evaluate differential expression of 5q35.3 genes with respect of mTOR pathway status. We identified only four genes significant (p < 0.05) in both tests (Figure S54A).

Strikingly, both GNB2L1 (showing the most significant association of all genes in both tests) and SQSTM1 displayed significant associations (Figure S54A,B,C). Another MEMo candidate gene, MAPK9, showed no change in expression in samples without mTOR pathway alterations. Interestingly, the two other genes that showed significant associations in both tests, C5orf45 and TRIM7, were neighbouring genes of SQSTM1 and GNB2L1 (Figure S55D), further increasing the evidence that specifically these two genomic regions on the 5q35.3 amplicon could be linked to mTOR signaling. The mutual exclusivity module also includes frequent over-expression and rare, recurrent kinase domain mutations of EGFR (Figure S56).

References

1. Ciriello, Giovanni, Ethan Cerami, Chris Sander, and Nikolaus Schultz. “Mutual Exclusivity Analysis Identifies Oncogenic Network Modules.” *Genome Research* 22, no. 2 (February 2012): 398–406.

Table S22

Module	Genes [# altered samples]	Total Altered Cases	*P*-value	FDR
M1	Akt [7], EGFR [18], MTOR [22], PIK3CA [16], PTEN [19]	21.53%	<1E-03	< 0.01
M2	Akt [7], MTOR [22], PIK3CA [16], PTEN [19], TSC1/2 [7]	18.53%	<1E-03	< 0.01
M3	Akt [7], EGFR [18], GNB2L1 [26], PIK3CA [16]	18.20%	<1E-03	< 0.01
M4	Akt [7], MAPK9 [22], MTOR [22]	13.62%	<1E-03	< 0.01
M5	Akt [7], EGFR [18], PTEN [19], TP53 [13]	14.44%	0.003	< 0.01
M6	ATM [11], CDKN2A [12], TP53 [13]	9.54%	0.02	0.08
M7	Akt [7], PIK3CA [16], SQSTM1 [26]	13.32%	0.032	0.08
M8	Akt [7], MTOR [22], RHEB [4], TSC1/2 [7]	10.35%	0.043	0.08

Table S22. Gene modules displaying significant mutual exclusivity. Significant (FDR < 0.1) gene modules identified by MEMo. For each module the following information is shown: genes included in the module along with the number of alterations across all cases/samples for each gene (N =368); overall percentage of altered cases; Mutual exclusivity *P*-value evaluated for the pattern of gene alterations; *P*-value corrected for False Discovery Rate.

Figure S54

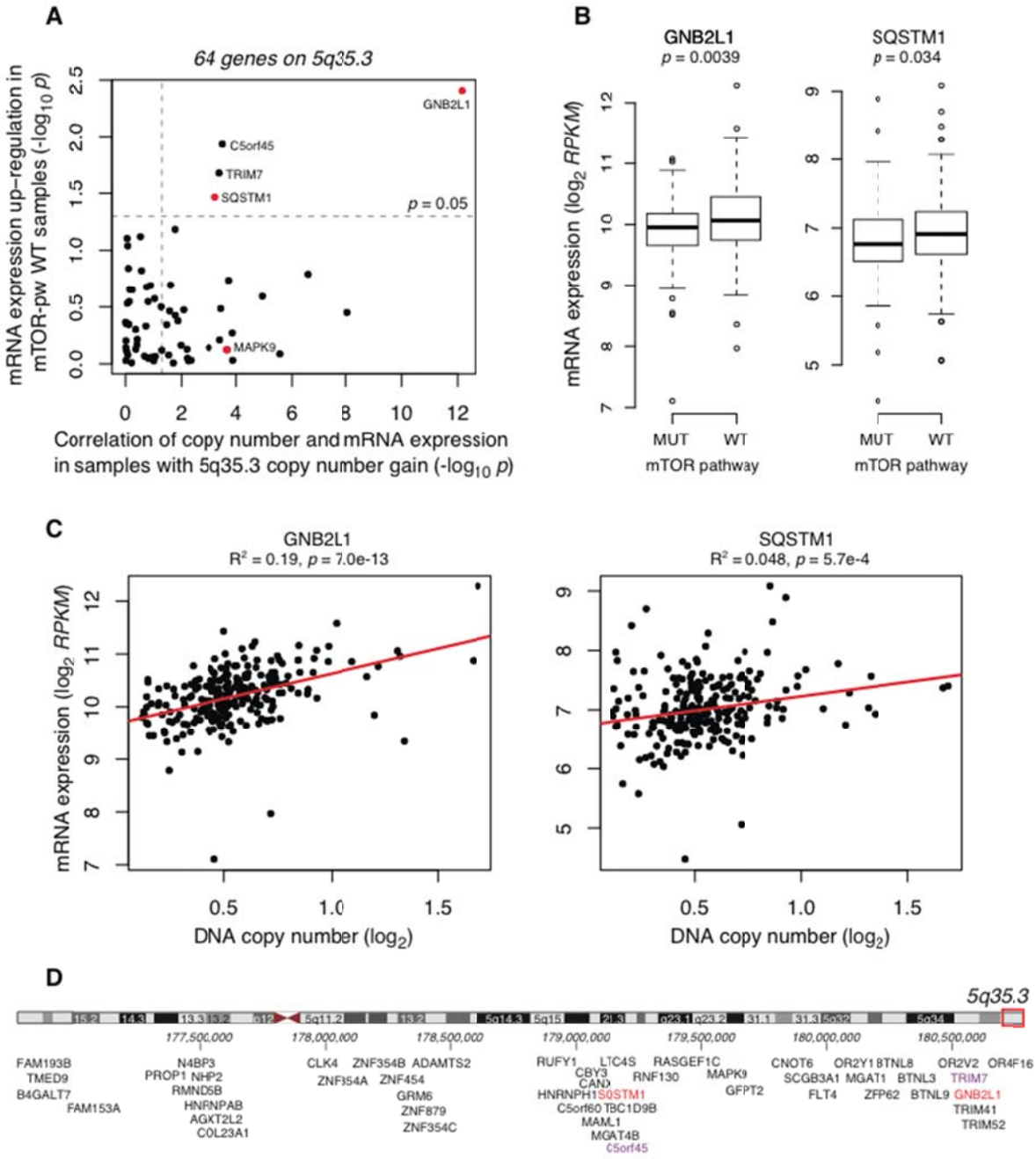


Figure S54. Candidate genes on 5q35.3 amplicon associated with mTOR signaling. **A**) We analyzed all genes with both copy number and mRNA expression data (N =64) in the 5q35.3 region for a possible association with mTOR signaling. First (x-axis), we tested the correlation (Pearson's) of DNA copy number and mRNA expression levels for each gene in the set of samples with 5q35.3 gain (gain of more than one copy of DNA). Secondly (y-axis), we evaluated the extent a gene was up-regulated in samples without mTOR pathway alteration (Kolmogorov Smirnov test, one-tailed). **B**) Expression levels of GNB2L1 and SQSTM1 in samples with (MUT) and without (WT) mTOR pathway alterations. This plot visualizes the data for the test on the y-axis in figure A. **C**) Correlation of DNA copy number and mRNA expression levels for GNB2L1 and SQSTM1 in samples with 5q35.3 gain. This plot visualizes the data for the test on the x-axis in figure A. **D**) Genes in the 5q35.3 region, showing that C5orf45 and TRIM7 are neighbouring genes to SQSTM1 and GNB2L1, respectively.

Figure S55

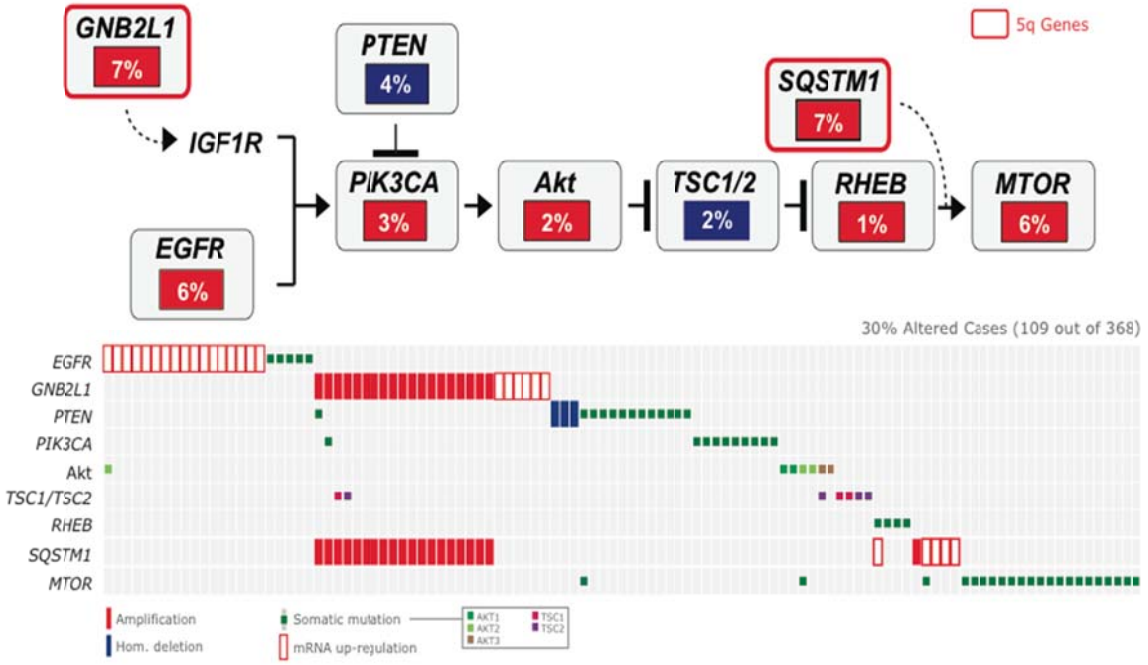


Figure S55. mTOR signaling pathway components display mutually exclusive pattern of alterations. Analysis of mutually exclusive gene modules by the MEMo algorithm identified a pattern of mutually exclusive gene alterations (somatic mutations, CNAs and aberrant mRNA expression) targeting multiple components of the mTOR pathway, including 2 genes from the recurrent amplicon on 5q35.3. Alterations were present in 30% of the analyzed tumors. The alteration frequency and inferred alteration type (blue for inactivation, and red for activation) is shown for each gene in the pathway diagram.

Figure S56

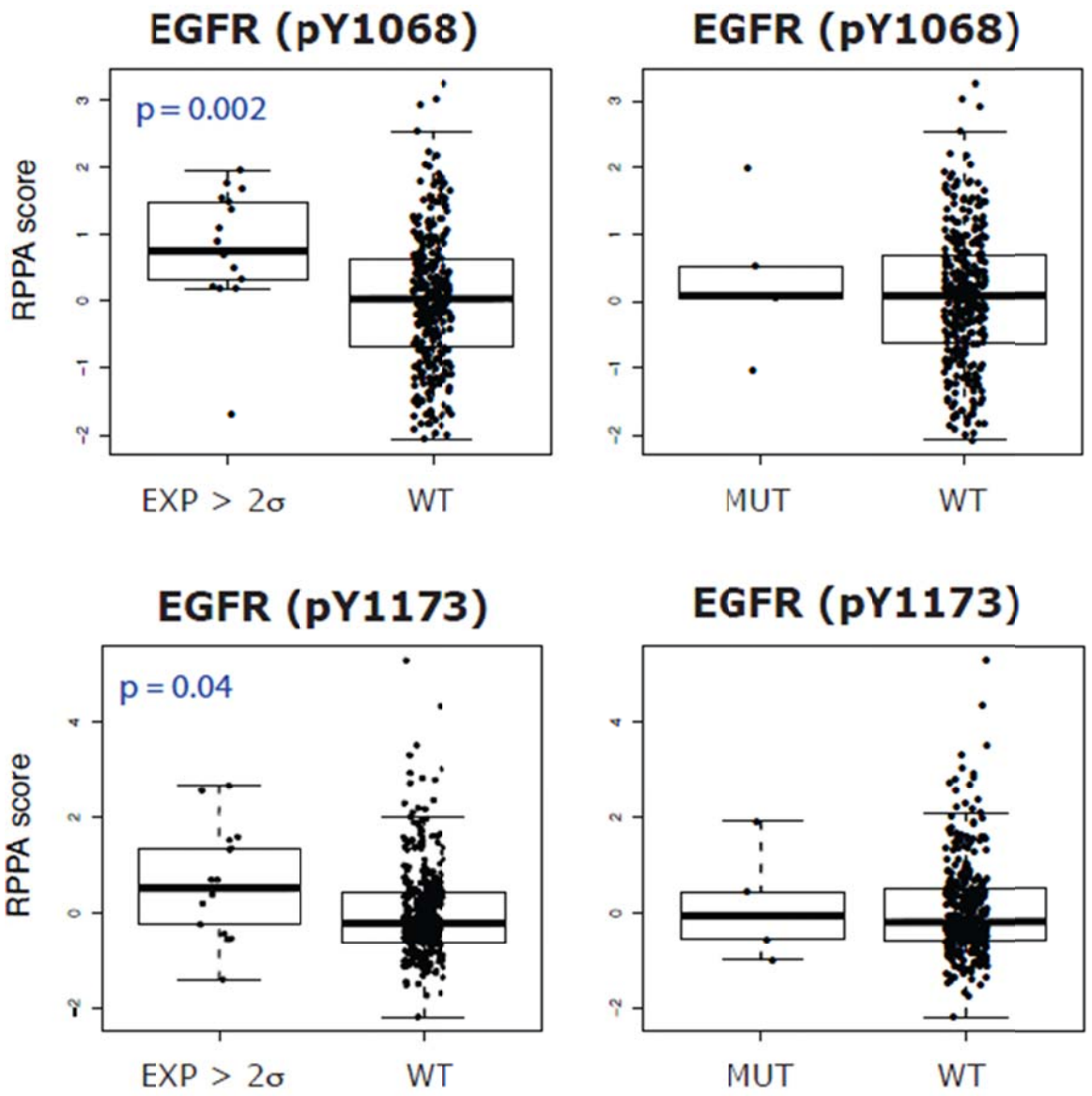


Figure S56. Frequent over-expression of EGFR, which correlates with increased phosphorylation of the receptor.

XV. SURVIVAL CORRELATES

Workgroup leader: Chad J. Creighton (creighto@bcm.edu)

Contributors: W. Marston Linehan, Chris Ricketts, Roel Verhaak, Gordon Robertson, Preethi Gunaratne, Hui Shen, Rehan Akbani, W. Kim Rathmell

Methods. The 446 patient samples (Core+Extended) were first separated into discovery and validation subsets (making use of a natural break in the dataset, namely profiling on methylation 27K versus 450K array), with all platforms using the same split. Therefore, the discovery subset consisted of batches 32-50-64-65-69 (profiled on 27K array, N=193) and the validation subset consisted of batches 63-68-70-82-90-105 (450K array, N=253). Level 3 data were analyzed for all the platforms. Using the discovery subset, profiled features were pre-filtered in the following manner: mRNA-seq, genes with average RPKM>10 (leaving 5944 genes); miR-seq: miRNAs with average "average_reads_per_million_mapped">10 (217 miRNAs); DNA methylation arrays: Illumina probes common to both 27K and 450K platforms, with data for at least 90% of the samples profiled, with batch effects "Deviation Index"<0.09 as assessed by the batch effects analysis group (see elsewhere), with beta values in 90th-% beta above 0.5, and with the difference between 90th and 10th percentile above 0.2, based on 27K platform (4296 probes remaining); RPPA: all 190 features used (which included duplicates for some antibodies). For methylation arrays, beta values were first logit-transformed, with missing values inferred using the median.

For survival analysis, patient death was the endpoint, with follow-up time defined using "days_to_last_followup" field if the patient was alive and "days_to_death" field if the patient was deceased. For each pre-filtered discovery dataset, the top survival correlates were determined by univarate Cox. Each prognostic signature was defined as the set of features most correlated with poor (worse) prognosis or with good (better) prognosis (RPPA and miRNA datasets, using Cox P<0.05; mRNA and methylation datasets, using Cox P<0.01, two-sided). For each validation dataset, feature values were normalized across samples to standard deviations from the median. The prognostic signature score was our previously described "t-score" metric [1,2], defined as the two-sided t-statistic comparing, within each tumor profile, the average of the poor prognosis features with the average of the good prognosis features (e.g. the t-score for a given tumor being high when both the "poor prognosis" features in the signature were high and the "good prognosis" features were low). For Cox analysis, the prognostic t-scores for each data type were normalized to standard deviations from the median across samples (the mRNA t-scores first being capped at 30, and the methylation t-scores being capped at 100). While samples were found to vary somewhat in tumor purity (as assessed by either pathology-based methods or by SNP array analysis), each prognostic signature was found to predict outcome independently of tumor purity estimates, as determined by multivariate Cox.

For the top survival molecular correlates, False Discovery Rates (FDR), as estimated using the Storey and Tibshirani method [3], were sufficiently low (indicative of the vast majority of the observed correlations not being due to chance expected from multiple testing). For the prognostic signatures based on the discovery set, FDRs were 10% for mRNA (0.01*5944/606), 25% for miRNA, 23% for protein, and 11% for DNA methylation. For the analysis focusing on the metabolic pathways, the global FDRs were even lower, as all the data (both discovery and validation sets) were used to compute univariate Cox, thereby increasing power (e.g. for mRNA, 7514 with nominal Cox P<0.01 out of 20532 tested, FDR=0.01*20532/7514<3%); in addition, our narrowing the focus to genes and proteins

involved in the AMPK/metabolic pathways (Figure 5B) allowed us to avoid multiple testing of all features; and subsequent mining of the miRNA/methylation correlates to complement the above mRNA/protein patterns was even more restricted to those with established metabolism roles and patterns of anti-correlation between expression and promoter methylation (essentially, mir-21 and GRB10).

We also examined the list of significantly mutated genes for survival correlations, of the genes tested, only BAP1 nonsilent somatic mutation showed a nominally significant correlation with worse outcome ($P<0.04$ log-rank test, Figure S27).

References

1. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 474:609-15.
2. Creighton CJ, et al. Insulin-like growth factor-I activates gene transcription programs strongly associated with poor breast cancer prognosis. J Clin Oncol. 2008 26:4078–4085
3. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003 100:9440-5.

Figure S57

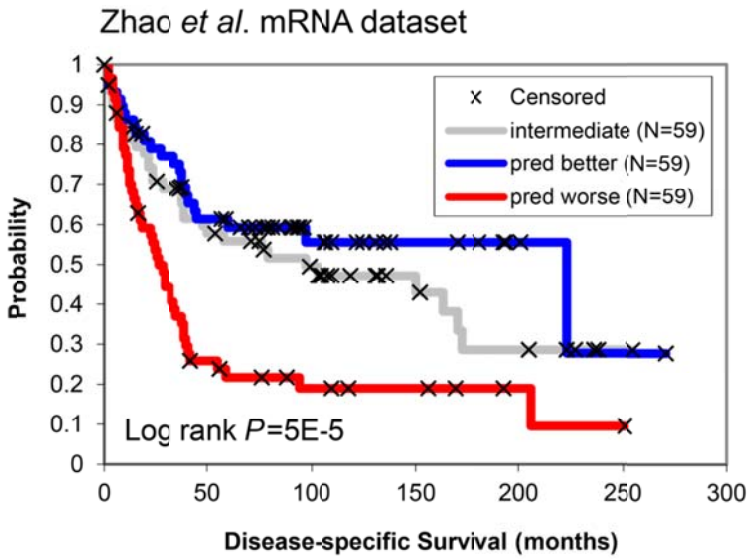


Figure S57. TCGA mRNA prognostic signature for kidney clear cell predicts survival in an independent dataset. Kaplan-Meier analysis of the TCGA mRNA prognostic signature, as applied to the previously published mRNA profile dataset from Zhao *et al*. (PLoS Med. 2006 3:e13), comparing survival for patients with predicted higher risk (red, top third of signature scores), lower risk (blue, bottom third), or intermediate (gray, middle third).

Figure S58

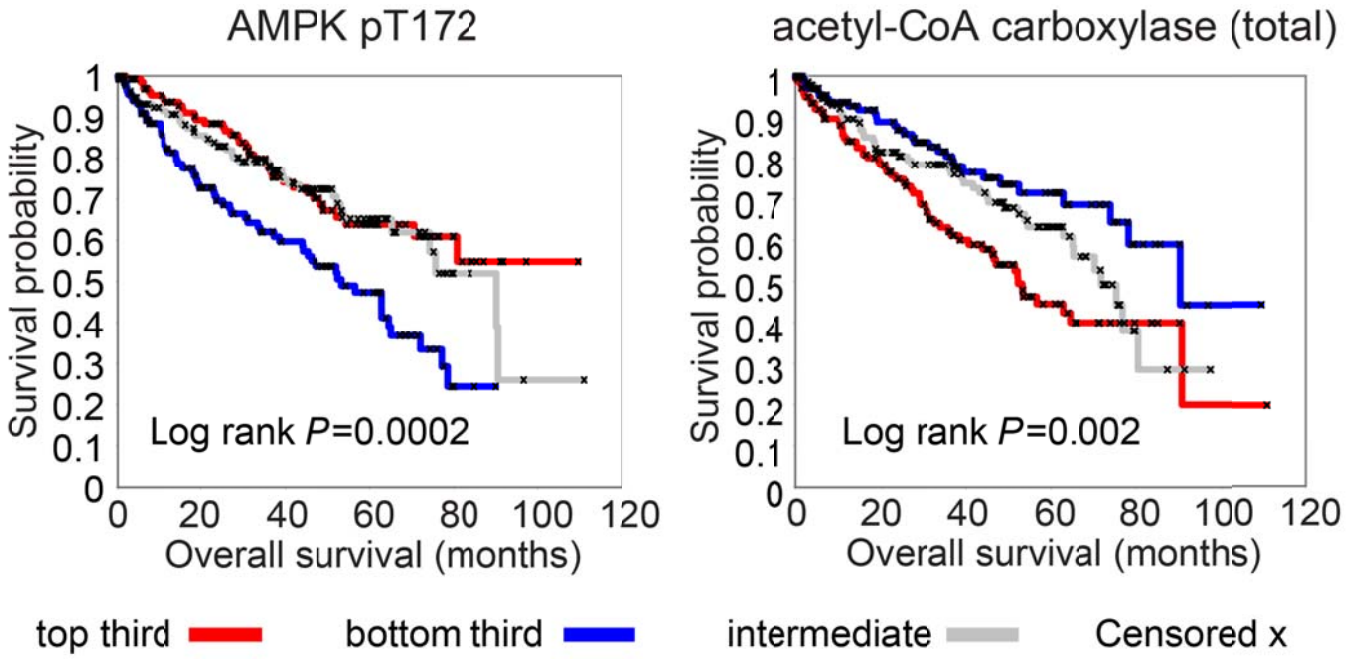


Figure S58. Protein survival correlates include AMPK and acetyl-CoA carboxylase (ACC). Using the full dataset (N=411 tumors), samples were binned by expression into thirds (red line, top third; blue line, bottom third; gray line, middle third); Kaplan-Meier plots show time of overall survival for each group. Proteins were also significantly correlated by univariate Cox analysis (AMPK: P<1E06 for pT172, P=0.01 for total levels; ACC: P<0.0003 for pS79, P<1E-6 for total levels).

Figure S59

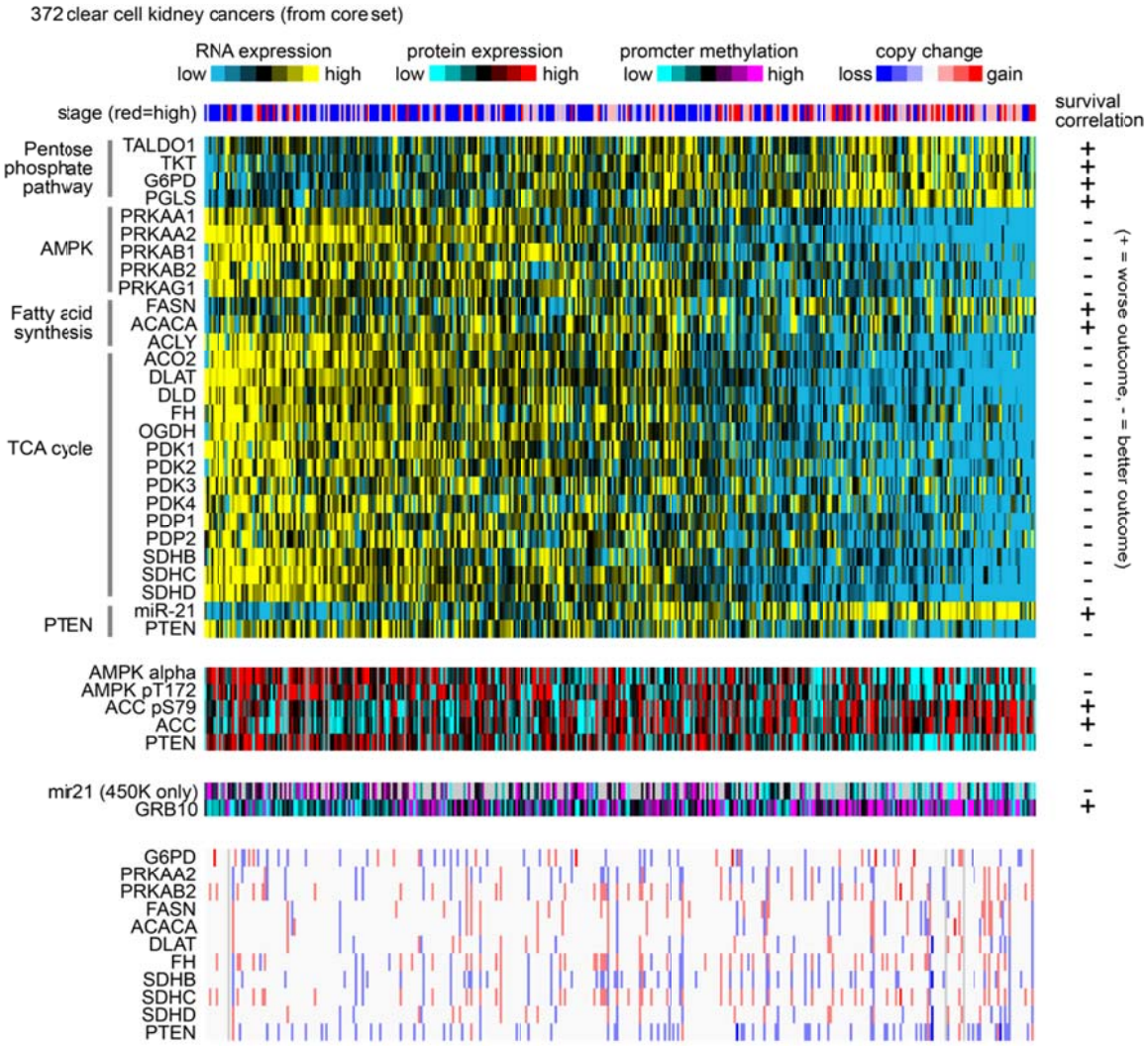


Figure S59. Genes in core metabolic pathways correlated with patient outcome are coordinately expressed in a sizeable proportion of kidney cancers, but are not uniformly driven by copy number alterations. Heat map of selected features from main Figure 5b, representing the “Core” set of kidney cancer profiles. Tumors are ordered by the sumproduct of the tumor molecular patterns with the survival correlation (“1” for worse outcome and “-1” for better outcome). Tumors towards the right show an expression profile more indicative of a glycolytic shift. Stage is represented by the red/blue color bar along the top (blue/light blue/light red/red for stages I/II/III/IV, respectively). For genes found to be in regions of significant gain or loss, as assessed by the GISTIC algorithm, as well as significant correlation between mRNA and copy, corresponding copy values are shown.

Figure S60

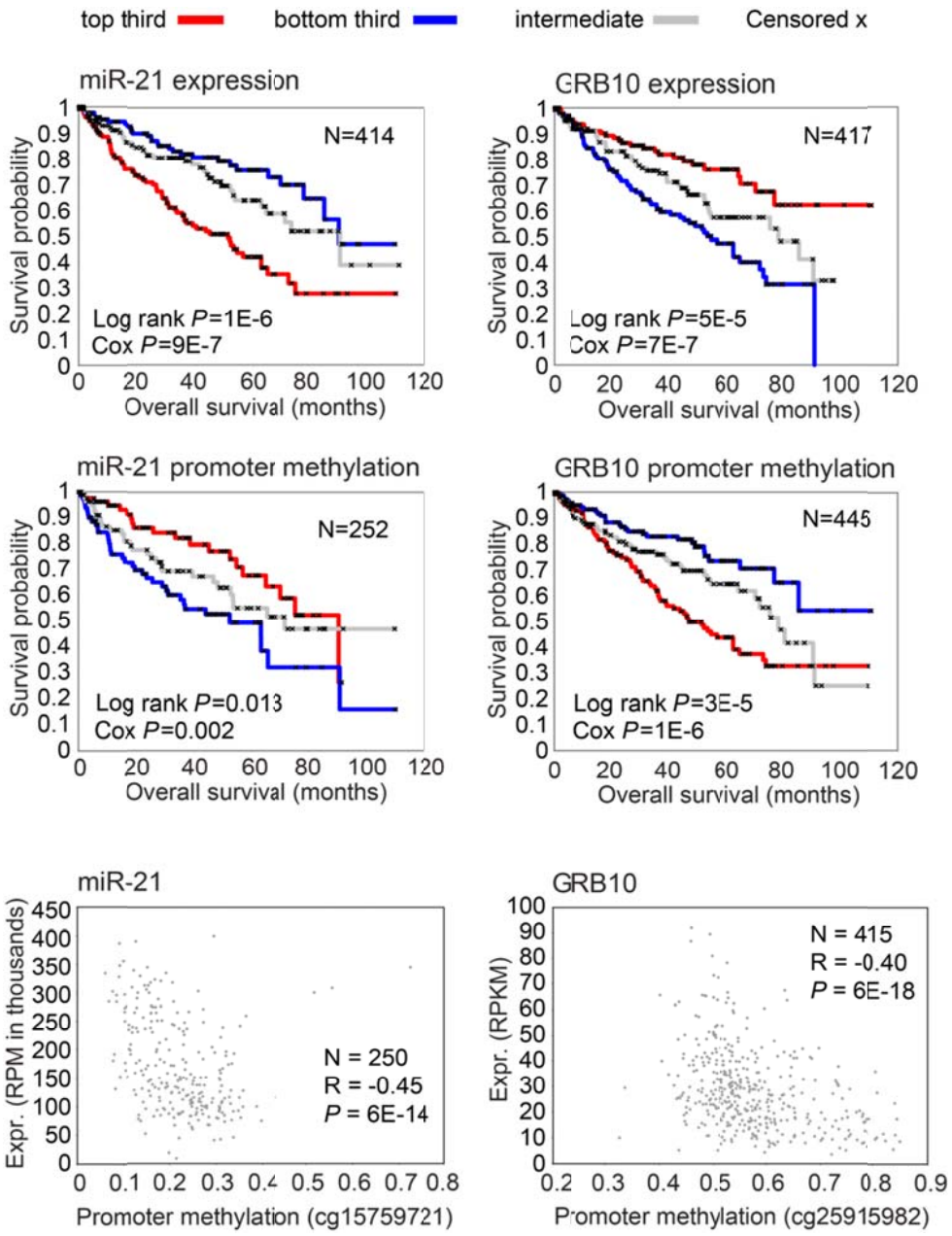


Figure S60. Promoter methylation and expression of miR-21 and GRB10 are each correlated with patient outcome. For Kaplan-Meier plots and log-rank statistics, samples were binned by measured feature levels into thirds (red line, top third; blue line, bottom third; gray line, middle third). Log-rank statistic evaluates for any significant differences between the three arms; univariate Cox evaluates the feature as a continuous variable (log-transform for expression data, logit transform for methylation data). For scatter plots (bottom panels), correlations by Spearman's rank test. GRB10 methylation values from the two platforms (27K and 450K) were first set to a common median before analysis.

Figure S61

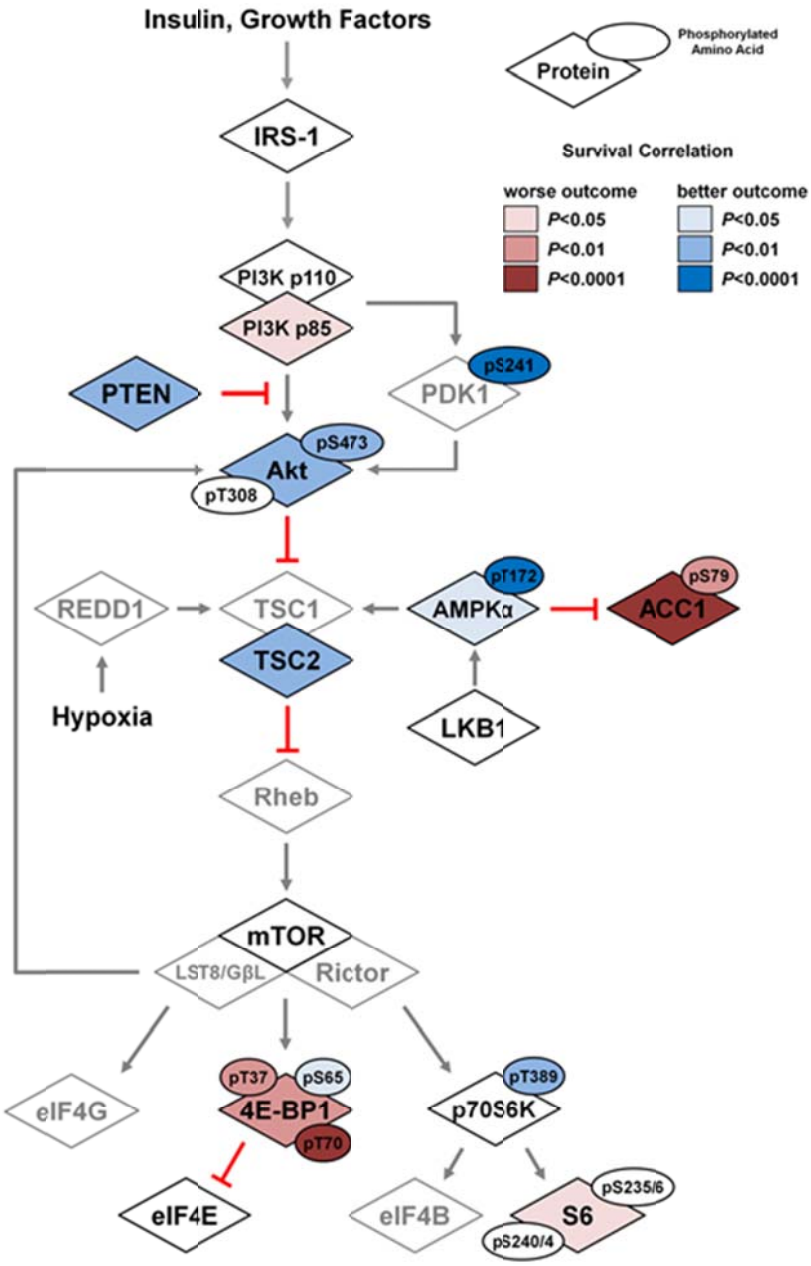


Figure S61. Survival correlations for proteins involved in the PI3K pathway, based on RPPA data (red/blue, correlation with worse/better survival, univariate Cox based on entire extended dataset).

Figure S62

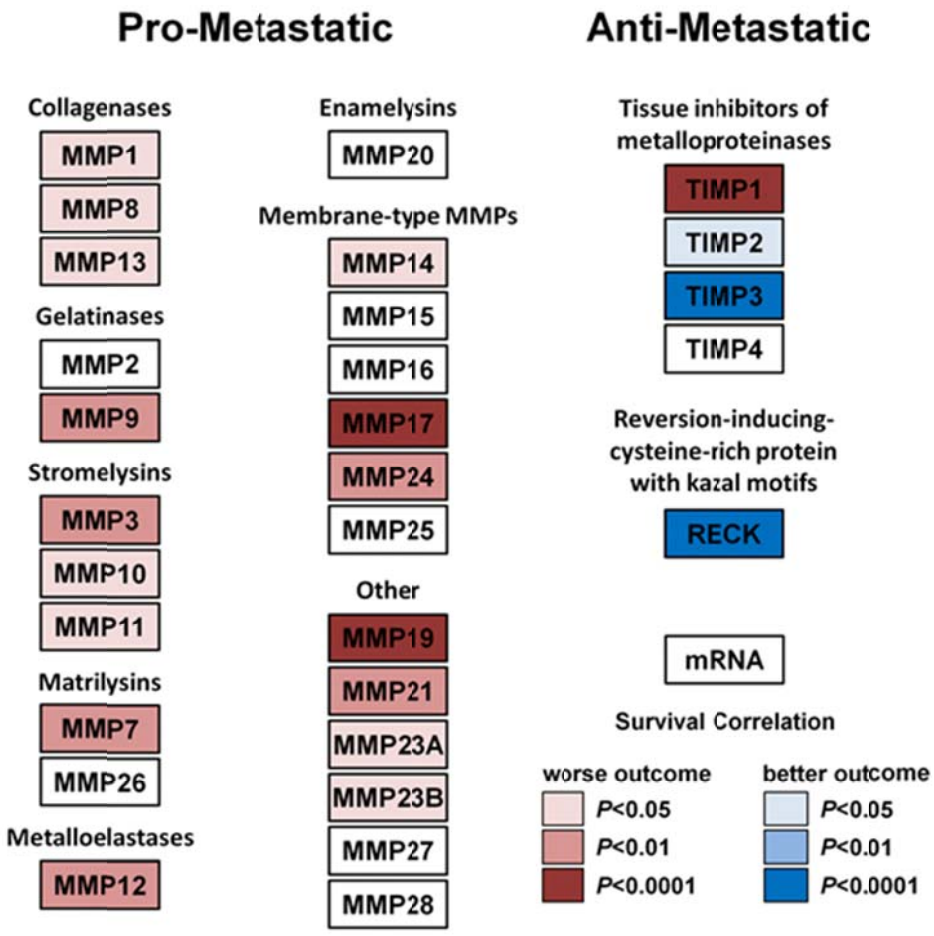


Figure S62. Survival correlations for mRNAs encoding matrix metalloproteinases (MMPs) or tissue inhibitors of metalloproteinases (TIMPs) (red/blue, correlation with worse/better survival, univariate Cox based on entire extended dataset).

Table S23

Table S23. Cox analyses for the molecular prognostic signatures scores.				
	N	**P**	**HR**	**95% CI**
Univariate Cox analysis				
mRNA signature	233	*7.E-07*	2.08	1.68 to 2.44
microRNA signature	251	*3.E-07*	1.95	1.56 to 2.44
protein signature	243	*4.E-07*	1.90	1.56 to 2.31
DNA meth signature	252	*1.E-06*	1.53	1.29 to 1.81
Multivariate Cox analysis				
mRNA signature	232	*6.E-05*	1.67	1.30 to 2.15
age (continuous)		*3.E-04*	1.04	1.02 to 1.07
stage (1-4)		*0.001*	1.73	1.25 to 2.40
grade (1-4)		0.40	1.16	0.82 to 1.62
metastasis (yes/no)		0.34	1.39	0.71 to 2.74
microRNA signature	250	*0.01*	1.36	1.07 to 1.72
age (continuous)		*5.E-04*	1.04	1.02 to 1.06
stage (1-4)		*3.E-04*	1.83	1.32 to 2.52
grade (1-4)		0.11	1.31	0.94 to 1.84
metastasis (yes/no)		0.45	1.30	0.66 to 2.55
protein signature	242	*0.003*	1.40	1.12 to 1.75
age (continuous)		*9.E-04*	1.04	1.01 to 1.06
stage (1-4)		*0.001*	1.70	1.23 to 2.35
grade (1-4)		0.13	1.29	0.93 to 1.80
metastasis (yes/no)		0.26	1.47	0.75 to 2.86
DNA meth signature	251	0.87	1.02	0.82 to 1.26
age (continuous)		*4.E-04*	1.04	1.02 to 1.06
stage (1-4)		*2.E-04*	1.84	1.33 to 2.55
grade (1-4)		*0.009*	1.60	1.13 to 2.25
metastasis (yes/no)		0.37	1.36	0.70 to 2.67
Based on validation subset (batches 63-68-70-82-90-105).				

Table S24

Table S24. mRNA:Promoter methylation pairs for which each are significantly correlated with survival (P<0.05) but in opposite directions to each other.

			univariate Cox, Methylation		univariate Cox, mRNA	
methylation array probe	**Gene**	**Entrez ID**	**beta**	**p-value**	**beta**	**p-value**
Methylation correlated with WORSE outcome, mRNA correlated with BETTER outcome						
cg24765005	BEX2	84707	0.197	0.0362	-0.272	0.0056
cg13853761	C11orf2	738	0.522	0.0088	-0.959	0.0019
cg09637363	CCND1	595	1.009	0.0014	-0.325	0.0173
cg21057494	CLEC3B	7123	0.550	0.0156	-0.265	0.0075
cg03032025	CPEB4	80315	0.571	0.0020	-0.624	0.0042
cg10707565	CUBN	8029	0.514	0.0078	-0.293	0.0001
cg10238818	CYYR1	116159	0.457	0.0001	-0.457	0.0022
cg14706739	EPB49	2039	0.473	0.0195	-0.503	0.0077
cg10134939	FAM176A	84141	0.677	0.0078	-0.390	0.0304
cg25645462	GPR56	9289	0.593	0.0018	-0.427	0.0237
cg04001668	GPR56	9289	0.457	0.0008	-0.427	0.0237
cg25915982	GRB10	2887	0.671	0.0095	-0.384	0.0318
cg15441973	HNMT	3176	0.512	0.0081	-0.683	0.0009
cg18248112	KCNJ15	3772	0.489	0.0056	-0.328	0.0003
cg09113530	MALL	7851	0.798	0.0028	-0.387	0.0043
cg01476044	MGAM	8972	0.599	0.0139	-0.296	0.0008
cg22545356	MMRN2	79812	0.788	0.0428	-0.415	0.0077
cg22289115	MUPCDH	53841	0.992	0.0008	-0.232	0.0093
cg12611860	PIK3C2A	5286	1.149	0.0046	-0.593	0.0131
cg03894103	PREPL	9581	0.456	0.0412	-0.587	0.0045
cg18149207	RORC	6097	0.565	0.0239	-0.387	0.0066
cg06236276	SLC22A2	6582	0.664	0.0021	-0.147	0.0190
cg12302621	SLC28A1	9154	0.541	0.0470	-0.278	0.0071
cg00668685	SLC39A5	283375	0.612	0.0187	-0.299	0.0010
cg20544605	SORBS2	8470	0.537	0.0093	-0.416	0.0006
cg11761535	TM4SF18	116441	0.698	0.0034	-0.352	0.0064
cg01211097	USP10	9100	0.828	0.0097	-0.726	0.0370
Methylation correlated with BETTER outcome, mRNA correlated with WORSE outcome						
cg15484375	SAA1	6288	-0.649	0.0469	0.107	0.0070
cg06190732	SERPINA3	12	-0.783	0.0269	0.197	0.0026
cg22242539	SERPINF1	5176	-0.746	0.0127	0.266	0.0075

Univariate Cox analysis based on training subset.

miRNAs correlated with survival and anti-correlated with DNA promoter methylation included miR-21 and miR-10b (associations identified by the miRNA group).

XVI. Acknowledgements

We wish to thank all patients and families who contributed to this study. This work was supported by the following grants from the USA National Institutes of Health (NIH): 5U24CA143799 (P. Spellman, PI), 5U24CA143835 (I. Shmulevich), 5U24CA143840 (C. Sander), 5U24CA143843 (D. Wheeler), 5U24CA143845 (L. Chin), 5U24CA143848 (C. Perou), 5U24CA143858 (D. Haussler), 5U24CA143866 (M. Marra), 5U24CA143867 (M. Meyerson), 5U24CA143882 (P. Laird), 5U24CA143883 (J. Weinstein), 5U24CA144025 (R. Kucherlapati), U54HG003067 (E. Lander), U54HG003079 (R. Wilson), U54HG003273 (R. Gibbs), U54CA143798 (R. Beroukhim), Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research (W.M. Linehan, L.S. Schmidt, C.J. Ricketts and R.A. Worrell, M.J. Merino), with Federal funds from the Frederick National Lab, NIH, under contract HHSN261200800001E (L.S.Schmidt), 5P50CA101942 (S. Signoretti), and P30CA16672 (G. Mills, MD Anderson CCGS RPPA Core, for technical support).