Cancer Cell, Volume *31*

Supplemental Information

Integrated Molecular Characterization of Uterine Carcinosarcoma

Andrew D. Cherniack, Hui Shen, Vonn Walter, Chip Stewart, Bradley A. Murray, Reanne Bowlby, Xin Hu, Shiyun Ling, Robert A. Soslow, Russell R. Broaddus, Rosemary E. Zuna, Gordon Robertson, Peter W. Laird, Raju Kucherlapati, Gordon B. Mills, The Cancer Genome Atlas Research Network, John N. Weinstein, Jiashan Zhang, Rehan Akbani, and Douglas A. Levine

Supplemental Data

Table S1. Clinical Data. Related to Figure 1, Provided as an Excel File
Table S2. Mutation counts from each center mutation callset and the combined ‘all’ calls. Related to Figure 1, Provided as an Excel File

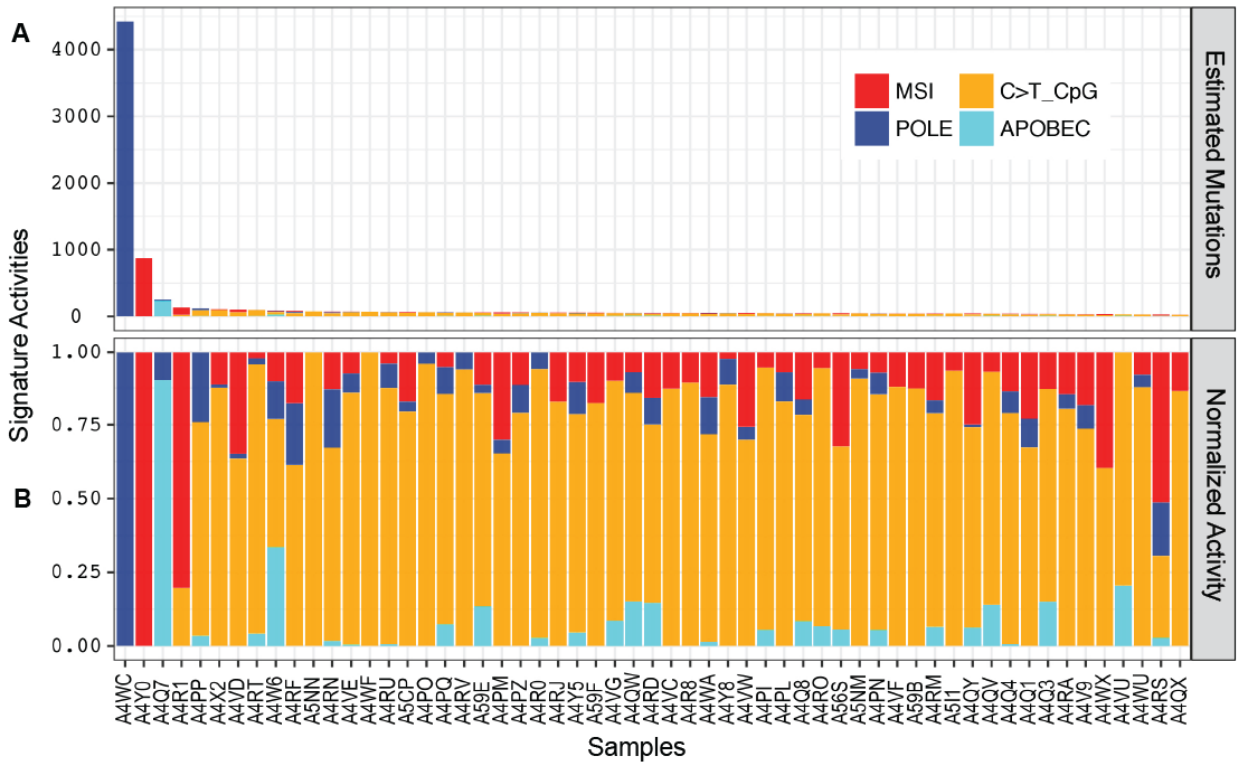


Figure S1. Mutation Signatures. Related to Figure 1. (**A)** Number of estimated mutations associated with identified signatures from Figure 1D. Single nucleotide variants were stratified by 96 tri-nucleotide base substitutions and indels were grouped by the number of inserted or deleted bases - one, two, three, four, and beyond-four bases indels – within the four identified mutational signatures (**B**) Normalized activities across samples measured by a fraction of mutation context assigned to each sample.

Table S3. Summary table of mutations, purity, ploidy across 57 samples. Related to Figure 1, Provided as an Excel File

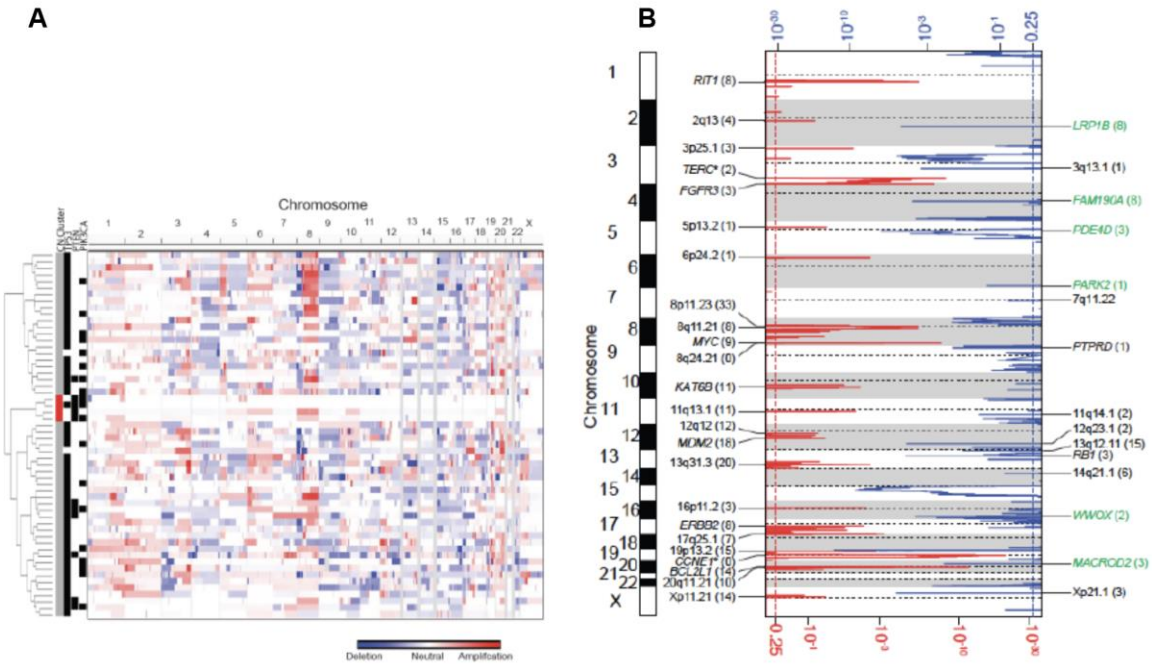
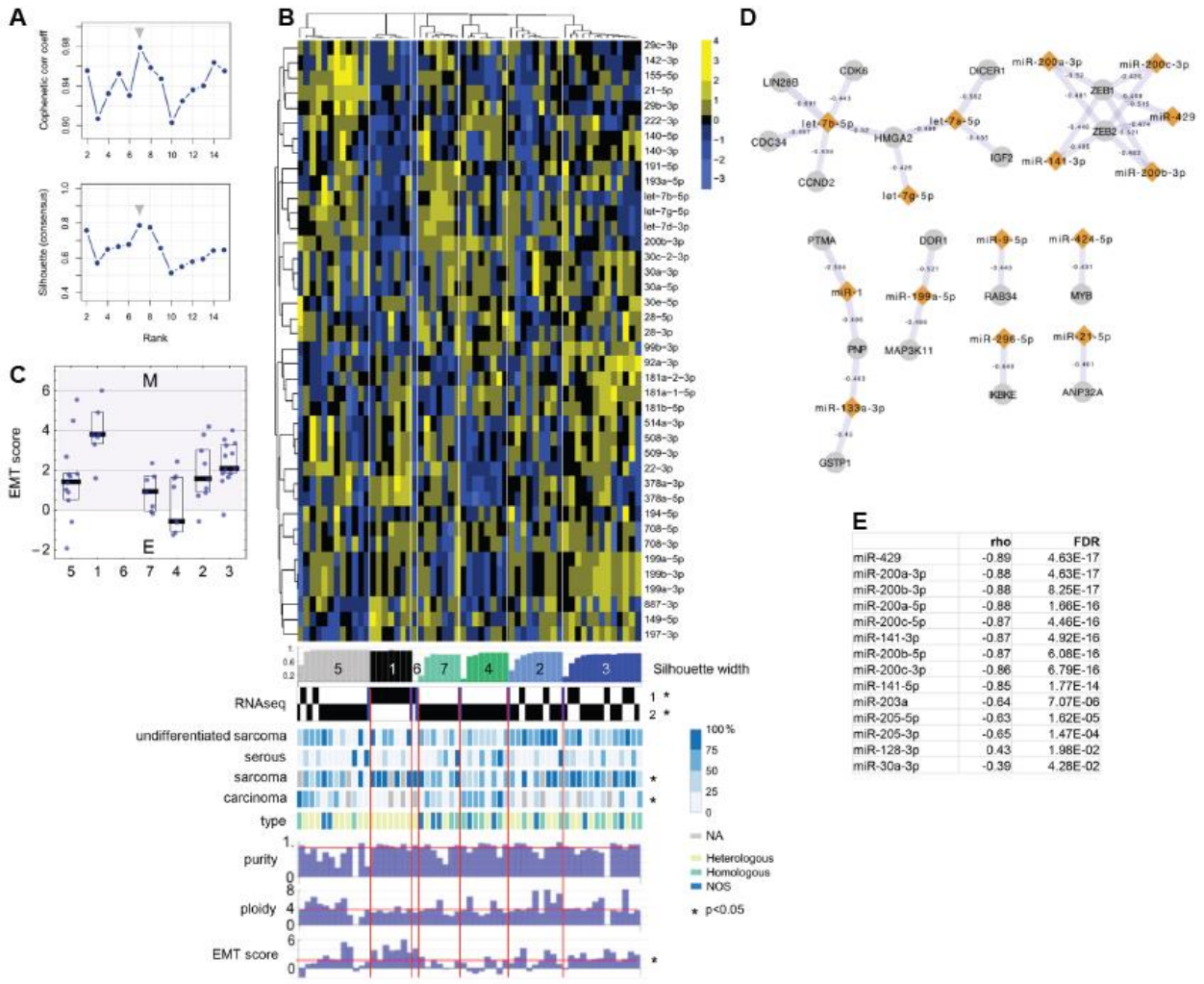


Figure S2. Somatic Copy Number Alterations (SCNA). **Related to Figure 1. (A)** SCNAs in tumors (vertical axis) are plotted by chromosomal location (horizontal axis). Tumors were hierarchical clustered by significantly reoccurring copy number alterations identified by GISTIC 2.0 analysis of the entire dataset. Vertical sidebars (left) show the location of the quiet copy number cluster (red), and tumors with TP53, PTEN and PIK3CA mutations (black). **(B)** GISTIC 2.0 amplifications and deletions. Chromosomal locations of peaks of significantly recurring focal amplifications (red) and deletions (blue) are plotted by false discovery rates. Annotated peaks have an FDR < .2 and encompass 35 or fewer genes. Peaks are annotated with candidate driver oncogenes, tumor suppressors, fragile site genes (green) or by cytoband. The number of genes within each peak is shown next to driver genes or cytobands. Genes marked with * are candidate focal drivers located outside the GISTIC 2.0 defined peak boundary of an alteration.

Table S4. Copy number at GISTIC peaks. Related to Figure 1, Provided as an Excel File



E

	rho	FDR
miR-429	-0.89	4.63E-17
miR-200a-3p	-0.88	4.63E-17
miR-200b-3p	-0.88	8.25E-17
miR-200a-5p	-0.88	1.66E-16
miR-200c-5p	-0.87	4.46E-16
miR-141-3p	-0.87	4.92E-16
miR-200b-5p	-0.87	6.08E-16
miR-200c-3p	-0.86	6.79E-16
miR-141-5p	-0.85	1.77E-14
miR-203a	-0.64	7.07E-06
miR-205-5p	-0.63	1.62E-05
miR-205-3p	-0.65	1.47E-04
miR-128-3p	0.43	1.98E-02
miR-30a-3p	-0.39	4.28E-02

Figure S3. miR subtypes, miR-gene targeting and miR-EMT score correlations. Related to Figure 4. (A) Rank survey profiles for cophenetic correlation coefficient and average silhouette width from the consensus membership. **(B)** For the seven-group clustering solution, top to bottom: a normalized abundance heatmap for the 40 5p or 3p mature strands that were highly ranked as differentially abundant; a silhouette width profile calculated from the consensus membership; covariate tracks highlight associations with the miRNA clusters and include the RNA-Seq clusters, 4 histologic types, and heterologous vs homologous type (NOS: not otherwise specified); profiles of ABSOLUTE purity and ploidy and EMT score. An asterisk identifies parameters with an association $p<0.05$. The scale bar shows row-scaled $\log_{10}(RPM+1)$ miR abundance. **(C)** Per-cluster distributions of the EMT score (Byers 2013). Boxplot center lines represent tumor medians, box limits are the inter-quartile range from 25% and 75%, blue dots indicate all data points. **(D)** miR-gene targeting relationships with FDR<0.05 that are supported by functional validation publications. Text on each edge shows the Spearman correlation coefficient between normalized abundances for a miR and gene. **(E)** Spearman correlations between EMT-associated miRs and the EMT score (FDR<0.05).

Table S5. Experimental results for miRNA-Seq data. Related to Figure 4, Provided as an Excel File

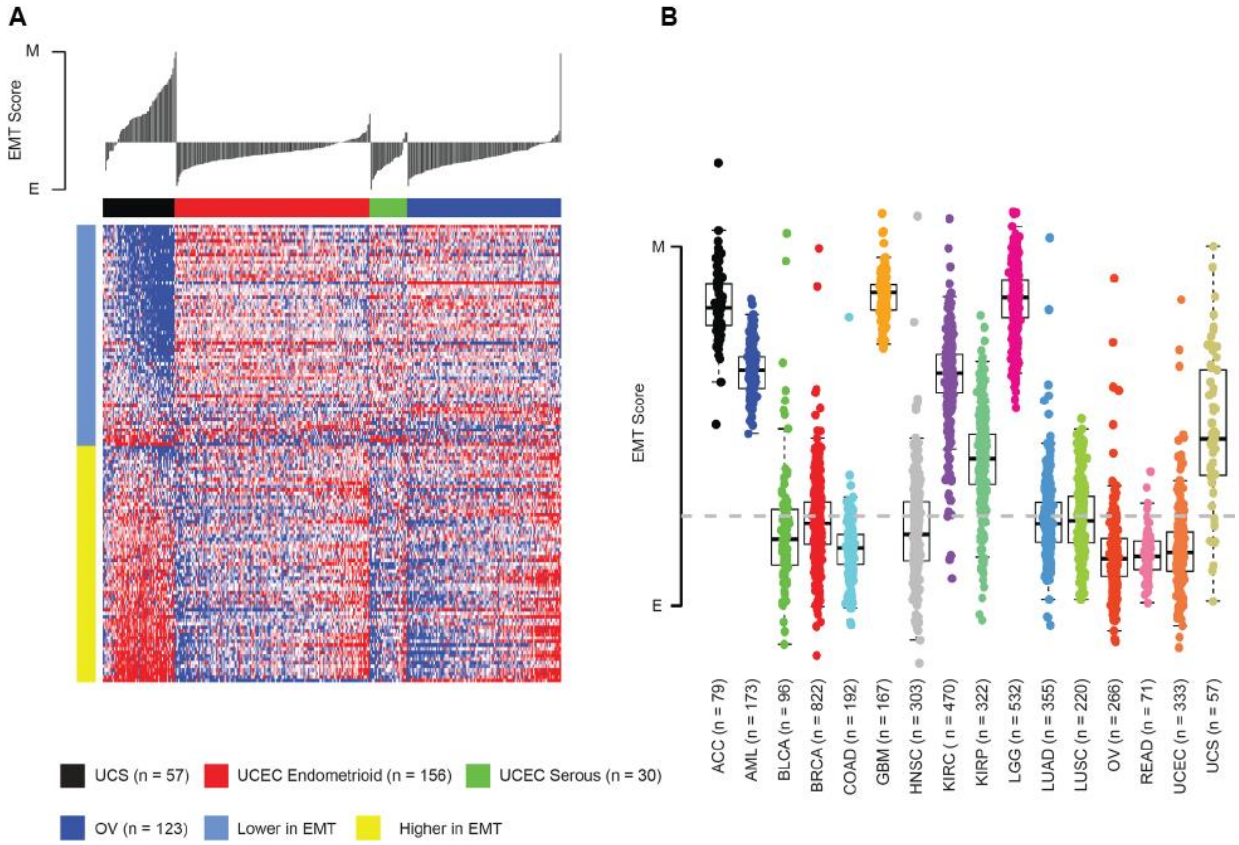


Figure S4. Expression and quantification of EMT across tumor types. Related to Figure 5. (A) The heatmap shows gene expression measurements of EMT-related genes in multiple gynecologic tumors. Samples appear in columns and are ordered by EMT score within four groups of tumors. Genes appear in rows and are ordered by mean expression value in UCS within two groups: those lower in EMT and those higher in EMT. Expression values are median-centered by gene. **(B)** The boxplot display shows EMT scores for 16 tumor types, and the dashed horizontal line corresponds to an EMT score of zero. Boxplot center lines represent tumor medians, box limits are the inter-quartile range from 25% and 75%, whiskers represent the extent of tumors out to 1.5 times the inter-quartile range. E, epithelial, M, mesenchymal; EMT, epithelial-to-mesenchymal transition; ACC, adrenocortical carcinoma; AML, acute myeloid leukemia; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, low grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous carcinoma; READ, rectum adenocarcinoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma.

Table S6. RPPA pathway analyses and antibodies. Related to Figure 7, Provided as an Excel File

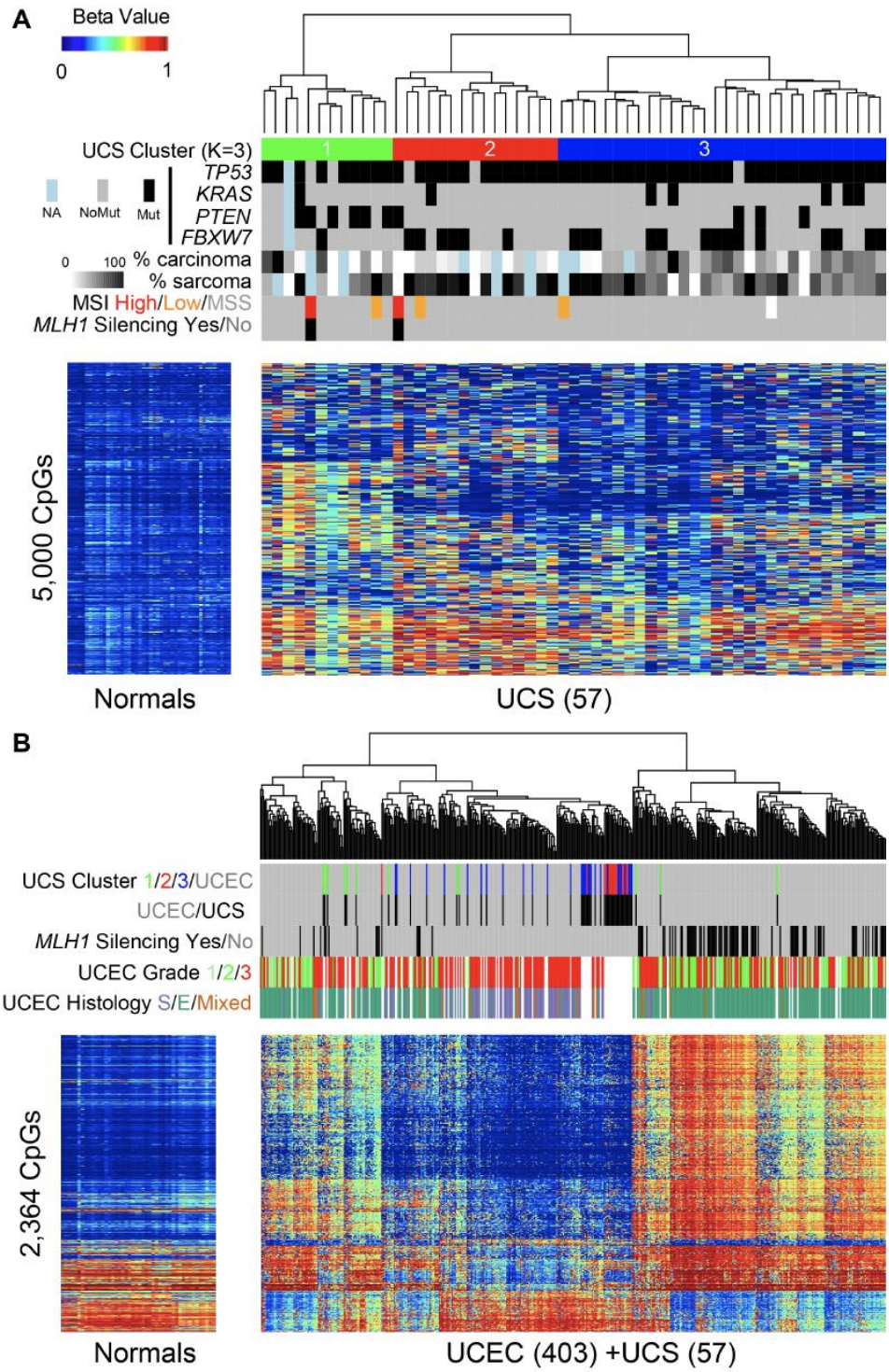


Figure S5. DNA Methylation Clusters. **Related to Figure 6. (A)** UCS tumors segregate into subgroups based on DNA methylation patterns. A random n=5000 probe subset was chosen from the top variably methylated (SD>0.15, n=10,464) autosomal probes unmethylated in normal tissues (left), and shown as rows of the heatmap. Top color bars show characteristics of the corresponding tumors, with black indicating mutation and gray indicating wild-type cases for bars denoting genes. (**B**) Joint clustering of UCS and UCEC tumors on 2,364 loci that distinguished serous-like (copy number high) and endometrioid UCEC tumors. Top color bars show characteristics of the corresponding tumors. S, serous, E, endometrioid.

Table S7. Differentially expressed genes in RNA subtypes. Related to Figure 6, Provided as an Excel File

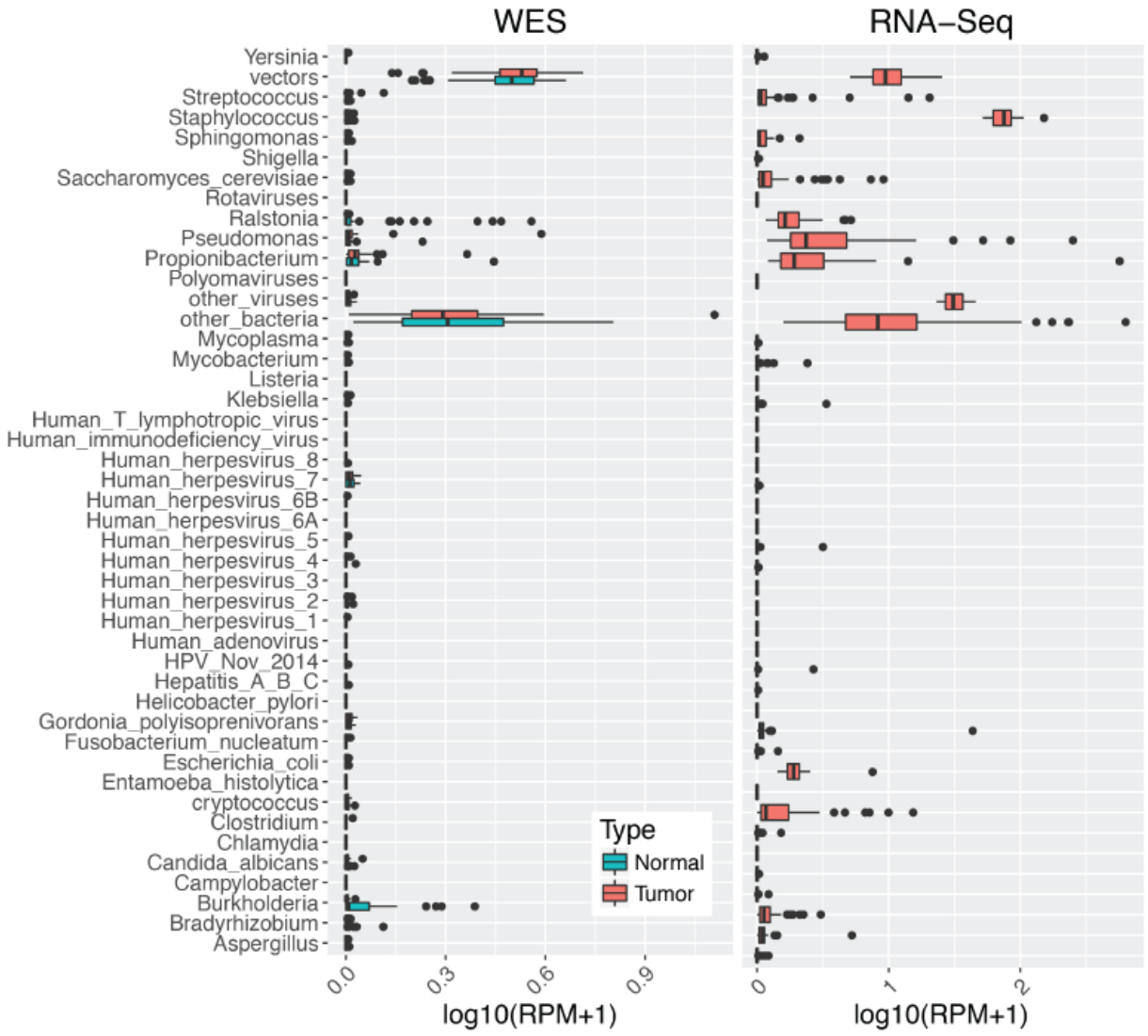


Figure S6. Distribution of RPM abundance for viruses, bacteria and fungi. Related to Figure 6. Microbe abundance (RPM) for mRNA-Seq data for 57 tumors and exome capture data for 50 tumors and 50 matched normal samples. Boxplot center lines represent tumor medians, box limits are the inter-quartile range from 25% and 75%, whiskers represent the extent of tumors out to 1.5 times the inter-quartile range, and circles are outliers beyond 1.5 times the inter-quartile range.

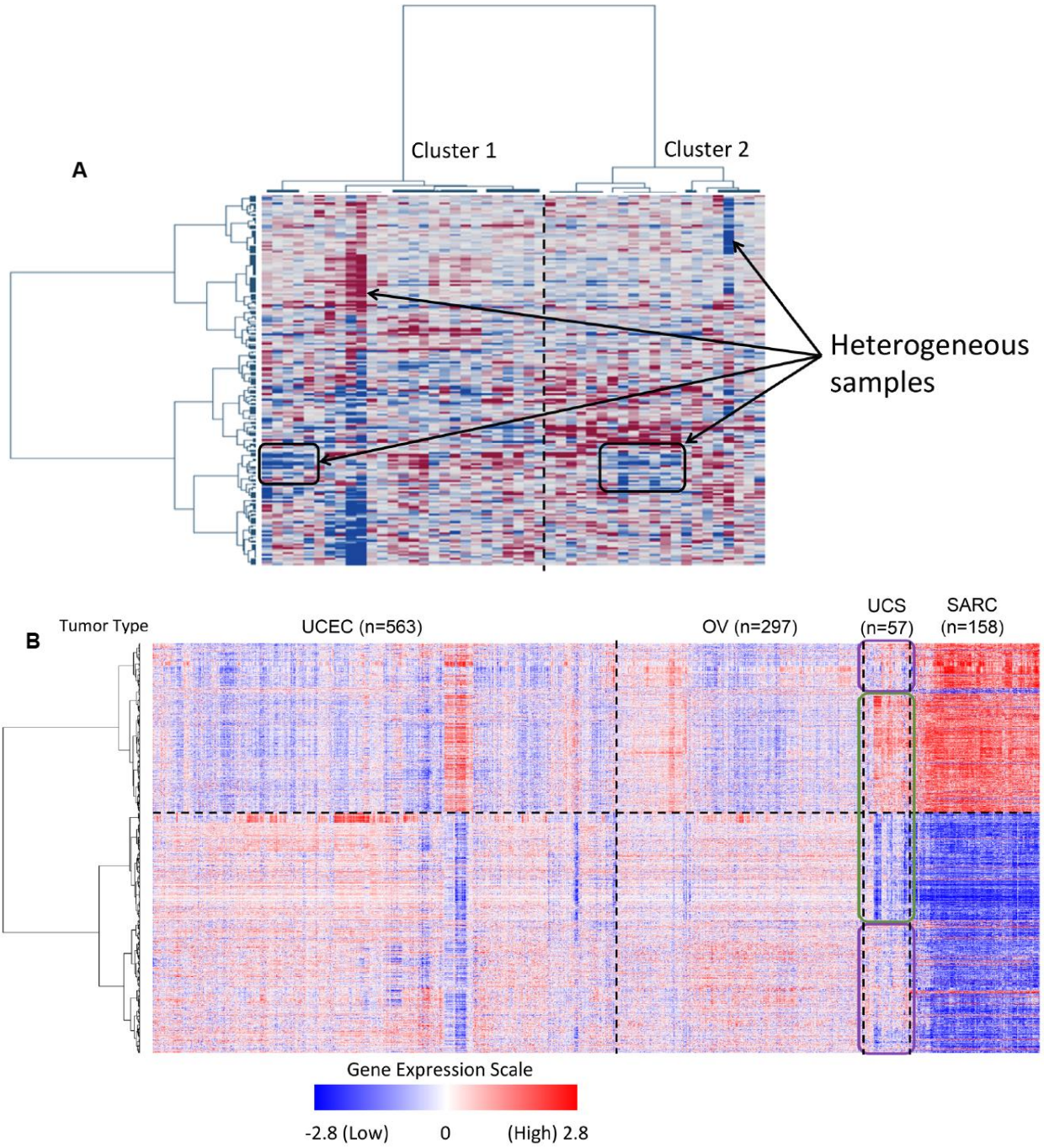


Figure S7. Related to Figure 7. (A) Unsupervised clustered heatmap of the protein expression data showing 48 samples (columns) across 200 proteins (rows). Arrows and boxes indicate samples within the clusters that have heterogeneous proteomic profiles. **(B)** Supervised clustered heatmap showing the top 2,000 differentially expressed genes from the mRNA platform between the gynecologic cancers (endometrial (UCEC) and ovarian (OV)) vs. sarcomas (SARC, without the leiomyosarcoma subtype). UCS was then added and the samples were clustered within each tumor type, but not across tumor types. The top part of the heatmap mostly shows genes that are highly expressed in sarcomas, whereas the bottom part mostly shows genes that are not highly expressed in sarcomas. Genes in UCS samples with similar profiles to the gynecologic cancers are outlined in purple and genes with profiles similar to sarcoma are outlined in green.

Supplemental Experimental Procedures

I. Biospecimen collection and clinical data

A. Sample inclusion criteria and pathology review

Biospecimens were collected at diagnosis from patients with uterine carcinosarcoma according to consent provided by the relevant institutional review boards. Patients were selected only if their treatment plan required surgical resection and had received no prior treatment for their disease, including chemotherapy or radiotherapy. The targeted accrual was 50 cases. Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Each frozen primary tumor specimen had a companion normal tissue specimen, who could be blood/blood components (including DNA extracted at the tissue source site), adjacent normal tissue taken from greater than 2 cm from the tumor, or both. No cases had qualifying metastatic tumor in addition to the primary tumor. Each tumor specimen was shipped overnight from one of the tissue source sites using a cryoport that maintained an average temperature of less than -180°C. Tumor and adjacent normal tissue specimens (if available) were embedded in optimal cutting temperature (OCT) medium and a histologic section was obtained for review. Pathologic diagnoses were made at local tissue source sites using diagnostic formalin-fixed and paraffin-embedded (FFPE) sections. A gynecologic specialty pathologist reviewed a representative FFPE H&E stained sections centrally and confirmed that the tumor specimen was consistent with the diagnosis. The adjacent normal specimen (when provided) was also confirmed to contain no tumor cells. Per TCGA protocol requirements, the sections were required to contain at least 60% tumor cell nuclei with less than 20% necrosis for inclusion in the study.

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

Each specimen was quantified by measuring Abs260 with a UV spectrophotometer or by PicoGreen assay. Analytes were resolved by 1% agarose gel electrophoresis (DNA) or Bioanalyzer RNA6000 nano assay (RNA) to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred (500) nanograms each of tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100-µg-reaction scale. Only those specimens yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study. In addition, DNA specimens with fragmentation resulting in low molecular weight smears or RNA with RIN < 7.0 were excluded in this study.

At the time of study closure, 82 uterine carcinosarcoma cases were received by the BCR and 69.5% passed pathology and molecular quality control. The biospecimens included in this report come from 57 uterine carcinosarcomas.

B. Microsatellite Instability Testing

Microsatellite instability (MSI) status of uterine carcinosaromas was evaluated in the Biospecimen Core Resource at Nationwide Children's Hospital. A panel of four mononucleotide repeat loci (polyadenine tracts BAT25, BAT26, BAT40, and transforming growth factor receptor type II) and three dinucleotide repeat loci (CA repeats in D2S123, D5S346, & D17S250) was used including the recommended markers from the National Cancer Institute Workshop on MSI in 2002 (Umar et al., 2004). Two additional pentanucleotide loci (Penta D & Penta E) were included in this assay to confirm sample identity. Electrophoretic mobility in these microsatellites from tumor and matched non-neoplastic tissue or mononuclear blood cells was compared after multiplex fluorescent-labeled PCR and capillary electrophoresis to identify variation in the number of repeats. Equivocal or failed markers were re-evaluated by singleplex PCR or through re-analysis of the entire MSI panel. Tumor DNA was classified as microsatellite-stable (MSS) if zero markers were altered, low level MSI (MSI-L) if one to two markers (less than 40%) were altered and high level MSI (MSI-H) if three or more markers (greater than 40%) were altered. Penta D and E markers were scored in the same manner as the MSI markers; however, they did not contribute to MSI class calculation.

Individual markers were assigned a value of 0 through 6 based on the presence or absence of a MSI shift, homo/heterozygosity in the normal sample, and loss of heterozygosity (LOH) if observed in the tumor. LOH for a

marker was assigned if the ratio of allele peak heights between tumor and matched normal control was less than 0.7 or greater than 1.6. Markers were classified as follows: 0= Marker not evaluable. 1= MSI; homozygous in Normal. 2= MSI; heterozygous in Normal with discernible LOH. 3= MSI; heterozygous in Normal where LOH was either not present or could not be calculated due to MSI interference with peak heights. 4= No MSI; homozygous in Normal. 5= No MSI; heterozygous in Normal with discernible LOH. 6= No MSI; heterozygous in Normal where LOH is not present. A single marker found to be "not evaluable" was allowed in MSI cases if the marker would not influence the overall call for the case.

C. Clinical data analyses

Clinical data included in this report were downloaded from the TCGA Data Portal on April 24, 2015. Age was recorded at initial pathologic diagnosis as reported by tissue source sites (TSSs). International Federation of Gynecology and Obstetrics (FIGO) stage was provided by TSSs. Body mass index (BMI) was calculated using the following formula: BMI = weight (kg) / [height (m) * height (m)]. Overall survival was calculated from date of pathologic diagnosis to date of death or last follow-up. Progression interval was censored for patients who had a status of "with tumor" in the data field for person_neoplasm_cancer_status, but did not have a date of progression listed in the data field for "days_to_new_tumor_event_after_initial_treatment". A summary of key clinical data is provided in Table S1. The 57 patients had a median age of 68 years (range, 51-90 years).

Specimen and assay summary are listed below:

Assay	**Number of specimens**
Exome sequencing	57 pairs
RNA sequencing	57 samples
miRNA sequencing	57 samples
DNA methylation (Infinium HM450)	57 samples
DNA copy number (Affymetrix SNP6.0)	56 samples
Reverse phase protein arrays	48 samples

II. Copy number analysis

SNP-based copy number analysis

DNA from each tumor or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described (McCarroll et al., 2008). Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus (Korn et al., 2008). For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor (Cancer Genome Atlas Research, 2011). This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation (Olshen et al., 2004). As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile (Mermel et al., 2011). SNP 6.0 derived somatic copy number alterations number was successfully derived from 56 of 57 tumors in this cohort. The SNP array from the one remaining tumor did not meet minimal QC standards to produce reliable somatic copy number data. Analysis of broad copy-number alterations was then conducted as previously described (Mermel et al., 2011). Significant focal copy number alterations were identified from segmented data using GISTIC 2.0 (Mermel et al., 2011). For copy number based clustering, tumors were clustered based on thresholded copy number at reoccurring alteration peaks from GISTIC analysis (Table S4). Clustering was done in R based on Euclidean distance using Ward's method. Allelic copy number, regions of homozygous deletions, whole genome doubling and purity and ploidy estimates were calculated using the ABSOLUTE algorithm (Carter et al., 2012).

III. Mutational Analysis.

A. DNA sequencing and data processing

From each sample, 0.5–3 micrograms of DNA were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. Whole exome capture was performed using Agilent SureSelect Human All Exon (http://www.genomics.agilent.com/en/Exome-Sequencing/SureSelect-Human-All-Exon-Kits) protocol according to the manufacturers' instructions. Exome capture regions were based on CCDS and RefSeq genes (http://www.ncbi.nlm.nih.gov/projects/CCDS/ and http://www.ncbi.nlm.nih.gov/RefSeq/) representing 188,260 exons from ~18,560 genes (93% of known, non-repetitive protein coding genes) and spanning ~1% of the genome (32.7 Mb). Whole exome (WES, n=57 tumor/normal sample pairs) and whole genome (WGS, n=49 tumor/normal sample pairs) sequencing was performed on the Illumina HiSeq 2000 platform using the V3 Sequencing Kits (http://support.illumina.com/sequencing/sequencing_instruments) and the Illumina 1.3.4 pipeline to produce paired-end sequenced data (2 x 101 bp for WGS to roughly 30x read coverage and 2 x 76 bp for WES to roughly 100x read coverage). Basic alignment and sequence QC was done on the "Picard" and "Firehose" pipelines at the Broad Institute.

B. Sequencing data-processing pipeline ("Picard pipeline")

The "Picard" pipeline (http://picard.sourceforge.net/) generates a BAM file (http://samtools.sourceforge.net/SAM1.pdf) for each sample and was developed by the Sequencing Platform at the Broad Institute. Picard pipeline aggregates data from multiple libraries and flow cell runs into a single BAM file for a given sample. This file contains reads aligned to the human genome with quality scores recalibrated using the TableRecalibation tool from the Genome Analysis Toolkit. Reads were aligned to the Human Genome Reference Consortium build 37 (GRCh37) using BWA v0.5.9 (Li and Durbin, 2010) (http://bio-bwa.sourceforge.net/). Unaligned reads were also stored in the BAM file such that all reads that passed the Illumina quality filter (PF reads) were kept. Duplicate reads were marked such that only unique sequenced DNA fragments were used in subsequent analysis. Sequence reads corresponding to genomic regions that may harbor small insertions or deletions (indels) were jointly realigned to improve detection of indels and to decrease the number of false positive single nucleotide variations caused by misaligned reads, particularly at the 3' end. In order to improve the efficiency of this step, we performed a joint local-realignment of all samples from the same individual ("co-cleaning"). All sites potentially harboring small insertions or deletions in either the tumor or the matched normal were realigned in all samples. Finally, the Picard pipeline provided summary QC metrics for each BAM that were used in subsequent processing.

Broad Institute "Firehose" analysis pipeline

The Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose) performed additional QC on the bams, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations, and identified significantly recurring mutated genes. The Firehose pipeline is a series of tools for analyzing massively parallel cancer sequencing data for both tumor DNA samples and their patient-matched normal DNA samples. The pipeline contains the following steps:

1. Quality control – confirms identity of individual tumor and normal to avoid mix-ups between tumor and normal data for the same individual. Sample contamination levels were also estimated using the ContEst program (Cibulskis et al., 2011). In subsequent analysis steps, tumor normal pairs of samples were required to have contamination less than 4% in order to ensure accurate somatic variant detection. There were no samples with significant contamination in this cohort.
2. Identification of somatic single nucleotide variations (SSNVs) – Mutect algorithm (Cibulskis et al., 2013) – candidate SSNVs were detected using a statistical analysis of the bases and qualities in the tumor and normal BAMs.
3. Identification of somatic small insertions and deletions – Indelocator algorithm – putative somatic events were first identified within the tumor BAM file and then filtered out using matched normal data as well a filter rejecting recurrent indels appearing in a panel of 254 normal samples.
4. Somatic Copy Number Alterations (SCNAs) were detected from read depth over each exon target region in the WES data using the ReCapSeg algorithm. Allelic Copy ratios were then assigned to each copy number segment using allele fractions of het sites provided by the GATK Haplotypecaller (https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php) and resegmented into regions of SCNA and LOH by Allelic Capseg. Allelic copy ratios from Allelic Capseg were subsequently used by one version ABSOLUTE (below) to make purity and ploidy estimates.
5. Tumor purities and mutation cancer cell fractions were assessed using two versions of ABSOLUTE program (Carter et al., 2012) (see Mutation Clonality Analysis).

6. Point mutations and indels as described above were also annotated using publicly available databases. In brief, a local database of human genome build hg19-derived annotations compiled from multiple different public resources was used to map genomic variants to specific genes, transcripts, and other relevant features. These annotations correspond to fields in the Mutation Annotation Format (MAF) files version 2.4 (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification). MAF files for somatic mutations are publically available from the NCI TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=UCS&diseaseName=Uterine%20Carcinosarcoma) or the Broad Institute GDAC (Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for Uterine Carcinosarcoma (Primary solid tumor) - 17 October 2014. Broad Institute of MIT and Harvard. doi:10.7908/C1Q23Z6P).

University of California Santa Cruz / RADIA mutation calls

Single nucleotide somatic mutations were identified by RADIA (RNA AND DNA Integrated Analysis), a method that combines the patient matched normal and tumor DNA whole exome sequencing (DNA-WES) with the tumor RNA sequencing (RNA-Seq) for somatic mutation detection (Radenbaugh et al., 2014) (software available at: https://github.com/aradenbaugh/radia/). The inclusion of the RNA-Seq data in RADIA increases the power to detect somatic mutations, especially at low DNA allelic frequencies. By integrating the DNA and RNA, mutations that would be missed by traditional mutation calling algorithms that only examine the DNA can be rescued back. RADIA classifies somatic mutations into 3 categories depending on the read support from the DNA and RNA: 1) DNA calls – mutations that had high support in the DNA, 2) RNA Confirmation calls – mutations that had high support in both the DNA and RNA, 3) RNA Rescue calls – mutations that had high support in the RNA and weak support in the DNA. Here RADIA identified 8,776 DNA mutations, 2,621 RNA Confirmation mutations, and 211 RNA Rescue mutations.

Baylor College of Medicine mutation calls

Atlas-SNP of the Atlas2 Suite was run to list all variants found in multiple reads at a single locus; and variants were annotated with dbSNP by ANNOVAR and COSMIC (Catalogue Of Somatic Mutations In Cancer) (Challis et al., 2012; Shen et al., 2010; Wang et al., 2010a). The variants were further filtered to remove all those observed fewer than 4 times or were present in less than 0.04 of the reads. Normal variant ratio must be less than 1% of tumor variant ratio. At least one variant had to be mapping quality of Q20 or better, and the variant had to lie in the central portion of the read. In addition, at least one variant must appear in both forward and reverse orientations. COSMIC variants were exempted from above filters. Insertion or deletion variants ("indels") were discovered by similar processing except that the initial processing was with Atlas-Indel of the Atlas2 Suite, and indels must have been observed in 10 of the reads with ratio of 0.15. All the variants were compared to a population of normal genomes and any matching variant was removed; then the file were further filtered by removing variants with normal sample coverage less than 2 or tumor variant coverage less than 0.05 or genes with greater than 2 variants for the same sample.

BCGSC / Strelka variant caller

Strelka (v1.0.6) was used to identify somatic single nucleotide variants and short insertions and deletions from the exome sequencing data (Saunders et al., 2012). All parameters were set to defaults, with the exception of "isSkipDepthFilters", which was set to 1 in order to skip depth filtration, given the higher coverage in exome datasets. When a blood sample was available, it served as the matched normal specimen; otherwise, the matched normal tissue was used. The variants were subsequently annotated using SnpEff, and the COSMIC (v61) and dbSNP (v137) databases (Cingolani et al., 2012; Forbes et al., 2015; Smigielski et al., 2000).

Washington University in St. Louis Mutation calls

WUSTL Read Realignment

Imported data were realigned to GRCh37-lite with bwa v0.5.9. Defaults are used in both bwa aln and bwa sampe (or bwa samse if appropriate) with the exception that for bwa aln four threads are utilized (-t 4) and bwa's built in quality-based read trimming (-q 5). ReadGroup entries were added to resulting SAM files using gmt sam add-read-group-tag. This SAM file was converted to a BAM file using Samtools v0.1.16, name sorted (samtools sort -n), mate pairings assigned (samtools fixmate), resorted by position (samtools sort), and indexed using gmt sam index-bam.

WUSTL Read Duplication Marking and Merging

Duplicate reads from the same sequencing library were merged using Picard v1.46 MergeSamFiles and duplicates are then marked per library using Picard MarkDuplicates v1.46. Lastly, each per-library BAM with duplicates marked is merged together to generate a single BAM file for the sample. For MergeSamFiles we run with SORT_ORDER=coordinate and MERGE_SEQUENCE_DICTIONARIES=true. For both tools, ASSUME_SORTED=true and VALIDATION_STRINGENCY=SILENT are specified. All other parameters are set to defaults. Samtools flagstat is run on each BAM file generated (per-lane, per-library, and final merged).

WUSTL Somatic Mutation Calling
We detected somatic point mutations using Samtools v0.1.16 (samtools pileup –cv -A -B), SomaticSniper v1.0.2 (bam-somaticsniper -F vcf -q 1 -Q 15), Strelka v1.0.10 (with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 –p value 0.10 --somatic-p value 0.05 --strand-filter 1).

We detected somatic indels using the GATK 1.0.5336 (-T IndelGenotyperV2 --somatic --window_size 300 -et NO_ET), retaining only those which were called as Somatic, Pindel v0.2.2 (-w 10; with a config file generated to pass both tumor and normal BAM files set to an insert size of 400), Strelka v1.0.10 (with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 –p value 0.10 --somatic-p value 0.05 --strand-filter 1).

C. Cross Center mutation filtering.
Mutations from each calling center were combined into a single Mutation Annotation Format list (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification) and labeled according to the centers that detected each mutation with additional fields counting reference and alternate supporting allele counts from each center. At this point there were 59677 separate coding+noncoding mutations in the list. The mutation selection criteria varied widely across centers so a series of filtering steps was applied uniformly across all mutations to minimize the contribution of artifacts and germline mutations to the final somatic mutation list.

Post-processing (Consensus and Panel of Normals filtering)
Not all the center's mutation calls included reference and alternate supporting allele counts, so each site's allele depth was extracted ("Force-called") from the tumor and normal bam files to provide a common allele counting metric and a common annotation framework to distinguish coding from non-coding mutations. First all non-coding mutations were removed the list of candidate mutations. One TP53 somatic mutation from sample id TCGA-ND-A4WF at hg19 position 17:7579307 was 5 bases from the exon and was not originally labeled as a splice site mutation according to TCGA convention (within 2bp of the exon edge). This sample showed an unusually low TP53 RNA expression which is typically associated with a biallelic deactivation of TP53, so it was retained with the list of coding mutations. Any mutations with less than 3 reads supporting the alternate allele was removed from the candidate mutation list. We then used a representative panel of 4513 normal WES bams to model a wide range of sequencing or alignment artifacts, or rare germline mutations, that might be misidentified as candidate somatic mutations. The Panel of Normals (PoN) filter removed any mutation with a corresponding alternate allele appearing normal according to several criteria:

a) Remove mutations at sites in which more than 0.2% of the normal bams had an alternate allele fraction exceeding 20%. This criterion primarily removes common germline variants.
b) Remove mutations with less than 3 supporting alternate allele reads in the tumor. This criterion removes candidate mutations with insufficient evidence.
c) Remove mutations at sites with fewer than 4 reads from the tumor supporting the alternate allele and with read depth less than 12 reads. This criterion also removes sites with insufficient evidence.
d) Remove mutations at sites with normal or tumor depth is less than 5 reads. This criterion removes sites with insufficient coverage.
e) Remove mutations at sites with normal depth less than 20 reads and in which more than 25% of the normal bams had coverage less than 8 reads. This criterion primarily removes sites with low coverage in the normal that could confirm that the mutation is somatic.
f) Remove mutations at sites with at least one alternate allele read in the normal and a tumor alternate allele fraction less than 16 times the normal allele fraction. This criterion primarily removes sites consistent with a germline variant or sequencing artifact.

g) Remove mutations at sites with the alternate allele fraction in the tumor less than 0.3 times the fraction of panel of normal samples with allele fraction between 0.1% and 20%. This criterion primarily removes sites consistent with recurring alignment or sequencing artifacts.

These criteria were devised and adjusted to compensate for the wide range of mutation properties across the five center calls. Overall, the PoN filter removed 16% of the coding mutation calls (Table S2).

Each candidate mutation site was also assessed in the matched RNA-seq tumor data to identify candidate mutations with independent confirmation from RNA. 35.6% of the candidate mutations occurred at expressed sites in the RNA-seq with sufficient coverage for >90% power to detect the mutation observed in the tumor DNA. Generally any mutation with two or more supporting reads in the matched tumor RNA was considered to be validated by RNA-seq, although the validation threshold at each site was separately calculated based on the sequencing noise level observed in the normal DNA.

The final list candidate mutations required that two or more centers made the call or that the mutation was supported by RNA-seq. In order to ensure that no candidate driving mutations were mistakenly removed by a filter, TP53, PTEN, and other candidate mutations were manually reviewed using IGV (Thorvaldsdottir et al., 2013).

D. Significantly Mutated Genes

Genes with a significant excess of the number of non-synonymous mutations relative to the estimated density of background mutations were identified using the MutSig algorithm (Lawrence et al., 2014; Lawrence et al., 2013). MutSig has been previous used to identify significantly mutated genes (SMGs) in several previous TCGA tumor sequencing projects (Cancer Genome Atlas Research, 2011; Cancer Genome Atlas Research et al., 2013) and has undergone a development path starting from the most basic approach implemented in MutSig 1.0 to the current version MutSig 2CV. This study made use of MutSig 1.5, MutSig 2.0, and MutSig 2CV to produce a robust list of significantly mutated genes (Getz et al., 2007; Lawrence et al., 2014; Lawrence et al., 2013). All three versions of MutSig identified TP53, FBXW7, PPP2R1A, PTEN, PIK3CA, KRAS, ZBTB7B, and SPOP, with false discovery rates < 0.1 while PIK3R1, RB1, ARHGAP35, CHD4, ARID1A, and U2AF1 were identified by at least one of the MutSig versions with multiple test corrected FDR q<0.1 (Benjamini and Hochberg, 1995).

The median proportion of synonymous mutations is 20.5% (mean 21.2% +/- 1%). The median number of covered bases is 29.6 Mb (mean 29.5 +/- 0.06 Mb, range 27.9 to 30.24 Mb). The total targeted exome region was 32.76 Mb, so 90% of the targeted exome region was covered on average. The mutation density is calculated as the count of mutations divided by the number of covered bases. The median non-synonymous mutation density was 0.71 per Mb (mean 1.76 +/0.88 per Mb, range 0.30 to 50.2 per Mb).

Significantly mutated gene list:

Gene	Npat	Mutsig 2CV rank	MutSig 2CV q	MutSig 1.5 rank	Mutsig 1.5 q	MutSig 2.0 rank	Mutsig 2.0 q
TP53	51	1	9.12E-13	1	2.01E-11	4	3.62E-12
FBXW7	22	2	9.12E-13	3	3.07E-11	1	0
PPP2R1A	16	3	5.40E-12	2	3.07E-11	2	0
KRAS	7	4	1.01E-11	7	1.69E-07	5	3.62E-12
PTEN	11	5	2.59E-11	4	3.07E-11	6	1.29E-10
PIK3CA	20	6	2.79E-10	5	3.50E-11	3	3.62E-12
PIK3R1	6	7	3.09E-04	12	0.054	17	0.123
RB1	5	8	0.0011	66	0.670	133	1
ZBTB7B	6	9	0.0084	8	2.53E-05	9	7.58E-05
ARHGAP35	6	13	0.0909				
SPOP	4	14	0.0928	9	7.21E-05	8	1.71E-05
CHD4	10	15	0.1525	6	1.00E-08	7	3.57E-09
U2AF1	2	17	0.3373	18	0.170	12	0.0201

ARID1A	7	29	0.8744	10	2.97E-04	10	0.00305

Twenty other sites were mutated more than once in the UCS cohort:

mutation	Count
TP53:p.R248Q:Missense_Mutation@17:7577538	6
FBXW7:p.R465C:Missense_Mutation@4:153249385	4
PIK3CA:p.E545K:Missense_Mutation@3:178936091	4
PIK3CA:p.H1047R:Missense_Mutation@3:178952085	4
PPP2R1A:p.P179R:Missense_Mutation@19:52715971	4
FBXW7:p.R479Q:Missense_Mutation@4:153247366	3
KRAS:p.G12V:Missense_Mutation@12:25398284	3
TP53:p.R273H:Missense_Mutation@17:7577120	3
TP53:p.S241Y:Missense_Mutation@17:7577559	3
FBXW7:p.R465H:Missense_Mutation@4:153249384	2
FBXW7:p.R505G:Missense_Mutation@4:153247289	2
FBXW7:p.R689W:Missense_Mutation@4:153244092	2
KRAS:p.G12D:Missense_Mutation@12:25398284	2
PIK3CA:p.G106V:Missense_Mutation@3:178916930	2
PIK3CA:p.R108H:Missense_Mutation@3:178916936	2
PPP2R1A:p.R183W:Missense_Mutation@19:52715982	2
PPP2R1A:p.S219L:Missense_Mutation@19:52716212	2
TP53:p.H179R:Missense_Mutation@17:7578394	2
TP53:p.H193R:Missense_Mutation@17:7578271	2
TP53:p.R175H:Missense_Mutation@17:7578406	2

E. Mutation Clonality Analysis

To assess whether mutations are clonal (i.e. present in all cancer cells), we estimated the cancer cell fraction (CCF) of each mutation, as described (Carter et al., 2012; Landau et al., 2013). Mutations for which the CCF is close to 1 are considered clonal. Those mutations with lower probable CCFs are considered subclonal. To determine the CCF we first estimated the sample purity (i.e. the percentage of tumor cells in the tumor sample) and ploidy (the average copy number across the tumor genome) using two versions of ABSOLUTE, one (version A) used allelic copy ratios detected from SNP6 array data as well as mutation allele fractions from WES data, and the other (version B) used alleleic copy number segments from WES data using ReCapSeg and Allelic Capseq from the Broad Firehose framework as well as mutation allele fractions from WES data. Purity and ploidy for each sample was manually reviewed from both ABSOLUTE versions and is listed in Table S3. In most cases the two estimated values of purity and ploidy were consistent between the two methods. The few discrepant cases were resolved by manual review. For each mutation ABSOLUTE provided a cancer cell fraction (CCF) probability distribution. Mutations with more than 50% of the CCF probability distribution at CCF>=0.95 are considered to be clonal.

F. Mutation Signature Analysis

A variant of non-negative matrix factorization (Bayesian NMF) to the mutation count matrix across samples (101 by 57), where single nucleotide variants (SNVs) were stratified by 96 tri-nucleotide base substitutions and indels were grouped by the number of inserted or deleted bases - one, two, three, four, and beyond-four bases indels, identified mutational signatures; APOBEC signature characterized by C>T/G at TCW (W=A/T), POLE signature with a predominance of C>A at TCT and C>T at TCG, microsatellite instability (MSI) signature with a significant enrichment of one-base indels, and spontaneous deamination signature dominated by C>T at CpG di-nucleotides.

G. Mutation counts

Mutation counts from each center mutation callset and the combined ‘all’ calls are listed below. The column labeled ‘Mutation calls’ is the count of all mutations submitted by each center. ‘Coding mutations’ is the subset of mutations limited to the coding region of the genome according to GENCODE v18. ‘PoN_Filter’ counts the subset of

mutations passing the panel of normal filter, and 'Consensus_RNA' is the final count of mutations meeting the criteria that the mutation has validation evidence from RNA-seq or that two or more centers made the call.

Center	Mutation calls	Coding mutations	PoN_Filter	Consensus_RNA
BCGSC	17582	8901	8645	8636
BCM	8066	7793	7686	7670
BI	22562	9949	9367	8897
WU	48927	9976	8873	8808
UCSC	8683	7717	7423	7419
All	59677	11560	9713	9149

IV. DNA methylation

A. Array-based DNA methylation assay

We used the Illumina Infinium HumanMethylation450 (HM450) BeadChips (Illumina, San Diego, CA) to obtain DNA methylation profiles of 57 uterine carcinosarcoma (UCS) samples.The Infinium HM450 array, which incorporates nearly all the HM27 probes, targets 482,421 CpG sites and covers 99% of RefSeq genes as well as intergenic regions, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. This platform covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them. The assay probe sequences and information for each interrogated CpG site by the HM450 and HM27 platforms can be found in the MAGE-TAB ADF (Array Design Format) file deposited on the TCGA Data Portal.

We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described (Campan et al., 2009). All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified (WGA) and enzymatically fragmented prior to hybridization to the arrays. BeadArrays were scanned using the Illumina iScan technology, and the IDAT files (Level 1 data) were used to extract the intensities (Level 2 data) and calculate the beta value (Level 3 data) for each probe and sample with the R-based *methylumi* package. Dye-bias normalization and normalization described in (Triche et al., 2013) were performed in the same process.

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

B. TCGA Data Packages

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

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

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

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

C. DNA Methylation Clusters

Data on 403 endometrial carcinoma and 46 normal tissue controls were obtained TCGA Uterine Endometrioid Carcinoma (UCEC) project (The Cancer Genome Atlas 2013). To focus on cancer-specific hypermethylation we excluded probes that showed any methylation in the 46 normal endometrial tissues (median beta value <0.2). We also excluded probes that had a SNP within 10 base pairs or repeats within 15 base pairs of the interrogated locus. Sex-linked (Chromosome X and Chromosome Y linked) probes were also removed. A random 5,000 (seed = 12345) were chosen from the remaining 18,147 probes, and used for clustering. The R function *hclust* was used for hierarchical clustering of these tumors with Ward's method, and the resulting dendrogram tree was cut at different levels to evaluate cluster stability and biological relevance, and k=3 was chosen, as it best separate the samples in this space.

We used the carcinoma (UCEC) data to find loci differentially methylated in serous/endometrioid histologies. 2,364 probes had an absolute difference of greater than 0.3 in the mean beta value of the two histologies. The UCS tumors

were jointly clustered with the UCEC tumors on this space, and Cluster 3 (green) clustered with the endometrioid UCEC tumors, both exhibiting more cancer-specific hypermethylation, while the rest are more similar to the serous histology of UCEC, featuring minimal DNA hypermethylation. For the analyses of *BRCA1* alterations, we had at least 98% power to detect *BRCA1* inactivation if the frequency was 5% or more in this cohort at an α=0.05 significance level.

V. RNA sequencing

A. RNA library construction, sequencing, and analysis

One μg of total RNA was converted to mRNA libraries using the lllumina mRNA TruSeq kit (RS-122-2001 or RS-122-2002) following the manufacturer's directions. Libraries were sequenced 48x7x48bp on the Illumina HiSeq 2000 as previously described (TCGA, 2012). FASTQ files were generated by CASAVA. RNA reads were aligned to the hg19 genome assembly using MapSplice 0.7.4 (Wang et al., 2010). Further details on this processing and expression quantification files can be found at the Genomic Data Commons (https://gdc-portal.nci.nih.gov/legacy-archive). FASTQ and BAM files can be found at CGHUB (https://cghub.ucsc.edu). Quantification of genes, transcripts, exons and junctions can be found at the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/).

B. Unsupervised clustering

Gene expression measurements were obtained by replacing all RSEM values (Li and Dewey, 2011) identically equal to zero with the smallest non-zero RSEM value, applying a $\log_2$ transformation, and median centering by gene. This produced expression values for 57 subjects and 20500 genes after removing genes with repeated or missing HUGO gene symbols. Using the median absolute deviation, the 2500 most variably expressed genes were identified. Consensus clustering was then performed using R 2.15.1 (R Core Team) and the ConsensusClusterPlus package (Wilkerson and Hayes, 2010). Gene expression heatmaps, principal components analysis, and silhouette plots suggested the presence of two expression subtypes, class one (n = 17) and class two (n = 40). The statistical significance of the differences in gene expression patterns present in the subtypes was assessed with the SigClust R package (Liu et al., 2008) using 1000 permutations, the default covariance estimation method, and all 20500 genes. The resulting permutation p value was equal to zero, which indicates that the differences in the gene expression patterns between the two subtypes were statistically significant.

C. Gene fusion detection

In addition to quantifying gene expression, RNA sequencing can detect structural variants, including alternate splicing, intra-chromosomal fusions, and inter-chromosomal fusions. Two algorithms were used to identify gene fusions:

A. MapSplice (Wang et al., 2010b),
B. PRADA (Torres-Garcia et al., 2014)

The PRADA pipeline was implemented separately at the Broad Institute and the M.D. Anderson Cancer Center. In both cases, paired-end reads were aligned to a combined genome and Ensemble transcriptome using the BWA aligner (Li and Durbin, 2010). Discordant read pairs were identified if each of the reads mapped to distinct genes. Junction spanning read pairs were identified if one read mapped to the putative fusion junction and the other read mapped to one of the two genes in the fusion. Filtering pipelines were then employed to reduce the number of false positive fusion transcript candidates. The M.D. Anderson team employed a different methodology that included filtering gene pairs with high sequence similarity (as determined by blastn), fusion pairs with a large variety of partner genes, and fusion pairs detected in normal samples (Yoshihara et al., 2015). Because of the differences in the filtering procedures, it is not surprising that the two implementations of PRADA produced different results. A total of 1195 unique fusion genes were identified by at least one of the three approaches. These findings were then filtered based on the level of evidence supporting each fusion as well as its functional relevance, as described below. Fusion genes were required to have (i) at least two discordant read pairs, (ii) at least one junction spanning read pair, and (iii) a unique fusion partner. This resulted in 774 fusion genes, 198 of which were identified by MapSplice and at least one of the two implementations of PRADA. No fusion was observed in more than one sample. 54 of the 774 fusions involved a gene with a kinase domain, and seven of these fusion genes produced an in-frame protein. Three in-frame fusion genes – *NUP210>MAST1*, *DDX6>ALK*, and *CENPP>WNK2* – that have the form [promoter of donor gene]>[kinase domain of acceptor gene], which is the most functionally relevant for fusions involving kinase domains. None of these gene fusions were previously described in Stransky et al. (Stransky et al., 2014).

D. Differential expression analysis

The SAMR package was used to identify genes that were differentially expressed in the RNA subtypes using 1000 permutations and a q value threshold of .05 (Tusher et al., 2001). The results of the analysis can be found in Table S6. We then used the DAVID annotation database to identify pathways that were enriched for differentially expressed genes (Huang da et al., 2009; Huang et al., 2007). Class one exhibited elevated levels of a number of genes belong to muscle development and contraction pathways, including *TTN*, *TNNC1/2*, and several myosin heavy chain genes. A number of genes associated with an epithelial-to-mesenchymal transition were differentially expressed, and class one exhibited higher expression levels of *ZEB1/2* and *FLNC*, as well as decreased expression of *CDH1*. Numerous cell adhesion genes were upregulated in class two, such as *CDH3/6* and *CLDN3*. In addition, class two exhibited elevated expression of immune-response gene, including *HLA-A/B/C* and *TAP1*.

E. Endometrioid/serous gene classifier

Gene expression measurements based on RSEM measurements from 25 tumor types were normalized together, as described in Hoadley et al. (Hoadley et al., 2014). Restricting to endometrial carcinoma samples with endometrioid or serous histological type resulted in gene expression data for 16,105 genes and 472 samples (374 endometrioid and 98 serous). ClaNC was then used to identify a set of 320 classifier genes whose expression patterns distinguish the endometrioid and serous samples (Dabney, 2006).

VI. miRNA Sequencing

A. miRNA library construction, sequencing and analysis

We generated microRNA sequence (miRNA-seq) data for 57 tumor samples using methods described previously except that 1ug of total RNA (at 250ng/uL) was used as input instead of messenger RNA-depleted RNA (Cancer Genome Atlas, 2012). We aligned reads to the GRCh37/hg19 reference human genome, and annotated miRNA read count abundance with miRBase v16. While we used only exact-match read alignments for quantifying miRNA abundance, BAM files are available from CGHub (cghub.ucsc.edu) that include all sequence reads (Wilks et al., 2014). We used miRBase v20 to assign 5p and 3p mature strand (miR) names to MIMAT accession IDs.

We identified groups of samples that had similar abundance profiles using unsupervised non-negative matrix factorization (NMF) consensus clustering with default settings (v0.5.06) (Gaujoux and Seoighe, 2010). The input was a reads-per-million (RPM) data matrix for the ~300 (25%) most-variant 5p or 3p mature strands, which we parsed from the level 3 isomiR data files that are available from the TCGA data portal. After running a rank survey with 30 iterations per solution, we chose a preferred clustering solution from profiles of the cophenetic correlation coefficient and the average silhouette width calculated from the consensus membership matrix, and performed a 500-iteration run to generate the final clustering solution. We calculated a profile of silhouette widths from the final NMF consensus membership matrix, to support identifying samples with relatively low widths within a cluster, which are less-typical cluster members.

To generate a heatmap for the NMF results, we first identified miRs that were differentially abundant between the unsupervised miRNA clusters, using a multiclass analysis with SAMseq (samr 2.0) in R 2.15.0, with a read-count input matrix and an FDR threshold of 0.05 (Li and Tibshirani, 2013). For the heatmap displayed, we included the 40 miRs that had both the largest SAMseq scores, and median abundances greater than 25 RPM. The RPM filtering acknowledged potential sponge effects from competitive endogeneous RNAs (ceRNAs) that can make weakly abundant miRs less influential (Mullokandov et al., 2012; Tay et al., 2014). We transformed each row of the matrix by $\log_{10}(\text{RPM} + 1)$, then used the pheatmap v0.7.7 R package to scale and cluster only the rows, with a Euclidean distance measure.

For clinical and molecular covariates, we calculated contingency table association p values with a Fisher exact test for categorical data, and a one-way ANOVA with a Bartlett test for equal variance for continuous data. Tumor sample purity and ploidy were calculated by the Broad Institute using ABSOLUTE (Carter et al., 2012). EMT scores were calculated by an RNAseq-based gene signature (Byers et al., 2013).

To identify miRs that were differentially abundant between UCEC endometrioid (n=182) and UCEC serous (n=38) samples, we used unpaired two-class SAMseq v2.0 analyses with FDR < 0.05 and removed miRs for which the median abundance in both groups was less than 25 RPM (Table S5).

We assessed potential miRNA targeting by calculating miR-mRNA and miR-RPPA Spearman correlations with the MatrixEQTL v2.1.1 R package (Shabalin, 2012), using RNAseq (RSEM) and RPPA gene-level normalized abundance data matrices from Firehose (gdac.broadinstitute.org). We calculated correlations with a p value threshold of 0.05, and filtered the resulting anticorrelations at FDR<0.05. We then extracted miR-gene pairs that corresponded to functional validation publications reported by MiRTarBase v4.5, for stronger (luciferase reporter, qPCR, Western blot) and weaker experimental evidence types (Hsu et al., 2014). We displayed the results with Cytoscape 2.8.3 (Table S5).

B. Results

The cophenetic coefficients and silhouette widths from the NMF survey support a seven-cluster solution, with the large number of clusters suggesting that our cohort was molecularly heterogeneous. We noted that mRNA cluster 1 had relatively high EMT scores and low miR-200b abundances. No associations were found with the DNA methylation or sCNA clusters, purity or ploidy ($p<0.05$). Percent sarcoma and carcinoma were statistically

associated ($p<0.05$) with the miRNA clusters, while percent undifferentiated sarcoma, percent serous and homologous vs. heterologous type were not.

Given the small cohort, relatively few functionally validated targeting relationships from miR-mRNA anticorrelations passed the FDR threshold, and those that did had large negative correlation coefficients. Among these potential targeting relationships, we noted that multiple members of the miR-200 family were associated with the EMT transcription factors ZEB1 and ZEB2.

While no significant (FDR<0.05) miR-RPPA Spearman anticorrelations were supported by functional validations in miRTarBase v4.5, we noted large negative correlation coefficients (<-0.6) between miR-503-3p and miR-499a-5p and ERRFI1 (MIG-6); between miR-135-5p and PDK1; and between miR-23a-5p and Bap1 (Table S5). ERRFI1 (1p36.23, Mig-6) is a tumor suppressor that interacts with and inhibits the intracellular kinase domains of EGFR (Zhang et al., 2007; Zhang et al., 2012). PDK1 (2q31.1) is a kinase that acts downstream of PI3K, and is a master regulator of the AGC kinase family, which includes Akt (Calleja et al., 2014; Fyffe and Falasca, 2013). Bap1 (3p21.1) is a tumor suppressor deubiquitylase that binds to BRCA1, and is associated with protein complexes that regulate important cellular pathways like DNA damage response (Carbone et al., 2013; Ismail et al., 2014; Murali et al., 2013).

Having noted miR-200 family members in our NMF clustering and in miRNA targeting results, we identified 62 miRs that have been associated with EMT in the literature (Carstens et al., 2014; Ceppi and Peter, 2014; D'Amato et al., 2013; Diaz-Martin et al., 2014; Kiesslich et al., 2013; Tam and Weinberg, 2013) (Table S5). We then determined which of these miRs were correlated with the EMT score using FDR-thresholded Spearman correlations. Anticorrelated miRs passing FDR<0.05 are listed in Figure S3E. The miRs most strongly anticorrelated with the EMT score were used in Figure 4.

VII. Reverse-Phase protein array

A. RPPA experiments and data processing

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mmol/L Hepes (pH 7.4), 150 mmol/L NaCl, 1.5 mmol/L MgCl2, 1 mmol/L EGTA, 100 mmol/L NaF, 10 mmol/L NaPPi, 10% glycerol, 1 mmol/L phenylmethylsulfonyl fluoride, 1 mmol/L Na3VO4, and aprotinin 10 ug/mL) from human tumors and RPPA was performed as described previously (Hennessy et al., 2007; Hu et al., 2007; Liang et al., 2007; Tibes et al., 2006). 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 serial diluted in two-fold of 5 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 200 validated primary antibodies (Table S7) 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 a CanoScan 9000F. Spot intensities were analyzed and quantified using Array-Pro Analyzer (Media Cybernetics Washington DC) to generate spot signal intensities (Level 1 data). The software SuperCurveGUI, 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) (Neeley et al., 2009). 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(Tibes et al., 2006). During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric 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 (Neeley et al., 2009). 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 using median centering across antibodies (level 3 data). In total, 200 antibodies and 48 samples were used. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described (Hennessy et al., 2010). 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 (Hennessy et al., 2010).

The RPPA arrays were quantitated and processed (including normalization and load controlling) as described previously, using Array-Pro Analyzer (Media Cybernetics L.P., Silver Spring, MD) and the R package SuperCurve (version-1.3), available at http://bioinformatics.mdanderson.org/OOMPA [1,3]. Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the TCGA portal

B. Data normalization

We performed median centering across all the antibodies for each sample to correct for sample loading differences. Those differences arise because protein concentrations are not uniformly distributed per unit volume. That may be due to several factors, such as differences in protein concentrations of large and small cells, differences in the amount of proteins per cell, or heterogeneity of the cells comprising the samples. By observing the expression levels across many different proteins in a sample, we can estimate differences in the total amount of protein in that sample vs. other samples. Subtracting the median protein expression level forces the median value to become zero, allowing us to compare protein expressions across samples.

Surprisingly, processing similar sets of samples on different slides of the same antibody may result in datasets that have very different means and variances. Neely et al. processed clinically similar ALL samples in two batches and observed differences in their protein data distributions (Neeley et al., 2009). There were additive and multiplicative effects in the data that could not be accounted by biological or sample loading differences. We observed similar effects when we compared the batches of TCGA Pan-Cancer protein expression data. A new algorithm, replicates-based normalization (RBN), was therefore developed using replicate samples run across multiple batches to adjust the data for batch effects. The underlying hypothesis is that any observed variation between replicates in different batches is primarily due to linear batch effects plus a component due to random noise. Given a sufficiently large number of replicates, the random noise is expected to cancel out (mean=zero by definition). Remaining differences

are treated as systematic batch effects. We can compute those effects for each antibody and subtract them out. Many samples were run in all the batches as replicates. One batch was arbitrarily designated the "anchor" batch and was to remain unchanged. We then computed the means and standard deviations of the common samples in the anchor batch, as well as the other batches. The difference between the means of each antibody in the anchor batch vs. the other batches and the ratio of the standard deviations provided an estimate of the systematic effects between the batches for that antibody (both location-wise and scale-wise). Each data point in the non-anchor batch was adjusted by subtracting the difference in means and multiplying by the inverse ratio of the standard deviations to cancel out those systematic differences. Our normalization procedure significantly reduced technical effects, thereby allowing us to merge the datasets from different batches.

C. Consensus hierarchical clustering

We performed consensus hierarchical clustering on the UCS RPPA data. Pearson correlation was used as distance metric and Ward was used as a linkage algorithm in the unsupervised hierarchical clustering analysis. No reliable subtypes could be identified in the UCS RPPA data.

D. Supervised clustering

We used LIMMA software (bioconductor package in R, https://bioconductor.org/packages/release/bioc/html/limma.html) to determine differentially expressed proteins between gynecologic carcinomas and sarcomas. The top 100 out of 200 proteins with the smallest p values were used for the analysis. We used the same procedure to determine the top 2,000 differentially expressed genes using mRNA data for Supplementary Figure S7B. We then used the differentially expressed proteins (or genes) to perform supervised clustering based on tumor type, after adding in the UCS samples. Hierarchical clustering with 1-Pearson correlation dissimilarity measure and Wards' linkage was used for the clustering.

E. Pathway analysis

For Figure 7B, we computed pathway scores for each sample using the method described previously (Akbani et al., 2014). The scores were then used to construct the box plots. P values were computed using the ANOVA test. Correlations between pathway scores and EMT scores were performed using Pearson's correlation coefficients and standard tests of significance.

VIII. Microbial Detection

Microbial detection methods

Our microbial detection pipeline is based on BioBloomTool (BBT, v1.2.4.b1), which is a Bloom filter-based method for rapidly classifying RNA-seq or DNA-seq read sequences (Chu et al., 2014). We generated 43 filters from 'complete' NCBI genome reference sequences of bacteria, viruses, fungi and protozoa, using 25 bp k-mers and a false positive rate of 0.02. We ran BBT in paired-end mode with a sliding window to screen fastq files from 57 RNA-seq libraries (49-bp PE reads; all tumors), and 100 whole exome libraries (49-bp PE reads, 50 tumors and 50 normals). In a single-pass scan for each library, BBT categorized each read pair as matching the human filter, matching a unique microbial filter, matching more than one filter (multi-match), or matching neither human nor microbe (no-match). For each filter, we then calculated a reads-per-million (RPM) abundance metric as:

$$Abundance metric = \left(\frac{\#\, reads\, mapped\, to\, a\, microbe\, filter}{\#\, chastity\, passed\, reads\, in\, the\, sample} \bullet 10^{6} \right)$$

Supplemental References:

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Calleja, V., Laguerre, M., de Las Heras-Martinez, G., Parker, P. J., Requejo-Isidro, J., and Larijani, B. (2014). Acute regulation of PDK1 by a complex interplay of molecular switches. Biochem Soc Trans *42*, 1435-1440.

Campan, M., Weisenberger, D. J., Trinh, B., and Laird, P. W. (2009). MethyLight. Methods Mol Biol *507*, 325-337.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Carbone, M., Yang, H., Pass, H. I., Krausz, T., Testa, J. R., and Gaudino, G. (2013). BAP1 and cancer. Nat Rev Cancer *13*, 153-159.

Carstens, J. L., Lovisa, S., and Kalluri, R. (2014). Microenvironment-dependent cues trigger miRNA-regulated feedback loop to facilitate the EMT/MET switch. The Journal of clinical investigation *124*, 1458-1460.

Ceppi, P., and Peter, M. E. (2014). MicroRNAs regulate both epithelial-to-mesenchymal transition and cancer stem cells. Oncogene *33*, 269-278.

Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R. A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.

Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., Mohamadi, H., Butterfield, Y. S., Robertson, A. G., and Birol, I. (2014). BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics *30*, 3402-3404.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology *31*, 213-219.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601-2602.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

D'Amato, N. C., Howe, E. N., and Richer, J. K. (2013). MicroRNA regulation of epithelial plasticity in cancer. Cancer Lett *341*, 46-55.

Dabney, A. R. (2006). ClaNC: point-and-click software for classifying microarrays to nearest centroids. Bioinformatics *22*, 122-123.

Diaz-Martin, J., Diaz-Lopez, A., Moreno-Bueno, G., Castilla, M. A., Rosa-Rosa, J. M., Cano, A., and Palacios, J. (2014). A core microRNA signature associated with inducers of the epithelial-to-mesenchymal transition. The Journal of pathology *232*, 319-329.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., *et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res *43*, D805-811.

Fyffe, C., and Falasca, M. (2013). 3-Phosphoinositide-dependent protein kinase-1 as an emerging target in the management of breast cancer. Cancer Manag Res *5*, 271-280.

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

Getz, G., Hofling, H., Mesirov, J. P., Golub, T. R., Meyerson, M., Tibshirani, R., and Lander, E. S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". Science *317*, 1500.

Hennessy, B. T., Lu, Y., Poradosu, E., Yu, Q., Yu, S., Hall, H., Carey, M. S., Ravoori, M., Gonzalez-Angulo, A. M., Birch, R., *et al.* (2007). Pharmacodynamic markers of perifosine efficacy. Clinical cancer research : an official journal of the American Association for Cancer Research *13*, 7421-7431.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., *et al.* (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929-944.

Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., Chu, C. F., Huang, H. Y., Lin, C. M., Ho, S. Y., *et al.* (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res *42*, D78-85.

Hu, J., He, X., Baggerly, K. A., Coombes, K. R., Hennessy, B. T., and Mills, G. B. (2007). Non-parametric quantification of protein lysate arrays. Bioinformatics *23*, 1986-1994.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.
Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res *35*, W169-175.
Ismail, I. H., Davidson, R., Gagne, J. P., Xu, Z. Z., Poirier, G. G., and Hendzel, M. J. (2014). Germline mutations in BAP1 impair its function in DNA double-strand break repair. Cancer research *74*, 4282-4294.
Kiesslich, T., Pichler, M., and Neureiter, D. (2013). Epigenetic control of epithelial-mesenchymal-transition in human cancer. Mol Clin Oncol *1*, 3-11.
Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., *et al.* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature genetics *40*, 1253-1260.
Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., *et al.* (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714-726.
Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.
Li, J., and Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res *22*, 519-536.
Liang, J., Shao, S. H., Xu, Z. X., Hennessy, B., Ding, Z., Larrea, M., Kondo, S., Dumont, D. J., Gutterman, J. U., Walker, C. L., *et al.* (2007). The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol *9*, 218-224.
Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008). Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data. J Am Stat Assoc *103*, 1281-1293.
McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., *et al.* (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature genetics *40*, 1166-1174.
Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A. D., Tung, N., Israelow, B., Evans, M. J., Sachidanandam, R., and Brown, B. D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat Methods *9*, 840-846.
Murali, R., Wiesner, T., and Scolyer, R. A. (2013). Tumours associated with BAP1 mutations. Pathology *45*, 116-126.
Neeley, E. S., Kornblau, S. M., Coombes, K. R., and Baggerly, K. A. (2009). Variable slope normalization of reverse phase protein arrays. Bioinformatics *25*, 1384-1389.
Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.
Radenbaugh, A. J., Ma, S., Ewing, A., Stuart, J. M., Collisson, E. A., Zhu, J., and Haussler, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One *9*, e111516.
Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811-1817.
Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353-1358.
Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., and Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res *20*, 273-280.
Smigielski, E. M., Sirotkin, K., Ward, M., and Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res *28*, 352-355.
Stransky, N., Cerami, E., Schalm, S., Kim, J. L., and Lengauer, C. (2014). The landscape of kinase fusions in cancer. Nature communications *5*, 4846.
Tam, W. L., and Weinberg, R. A. (2013). The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med *19*, 1438-1449.
Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344-352.
Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178-192.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther *5*, 2512-2521.
Torres-Garcia, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M. F., Weinstein, J. N., Getz, G., and Verhaak, R. G. (2014). PRADA: pipeline for RNA sequencing data analysis. Bioinformatics *30*, 2224-2226.
Triche, T. J., Weisenberger, D., Van Den Berg, D., Laird, P., and Siegmund, K. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res *41*, e90.
Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.
Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Ruschoff, J., Fishel, R., Lindor, N. M., Burgart, L. J., Hamelin, R., *et al.* (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst *96*, 261-268.
Wang, K., Li, M., and Hakonarson, H. (2010a). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164.
Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., *et al.* (2010b). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res *38*, e178.
Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572-1573.
Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., *et al.* (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford) *2014*.
Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R. G. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845-4854.
Zhang, X., Pickin, K. A., Bose, R., Jura, N., Cole, P. A., and Kuriyan, J. (2007). Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface. Nature *450*, 741-744.
Zhang, Y. W., Staal, B., Dykema, K. J., Furge, K. A., and Vande Woude, G. F. (2012). Cancer-type regulation of MIG-6 expression by inhibitors of methylation and histone deacetylation. PLoS One *7*, e38955.