# Supplementary Materials and Methods

## 1. Sample Acquisition and Pathology Review

### 1.1 METHODS

#### 1.1.1 Sample Acquisition

MPM tumors were collected and shipped to the Biospecimen Core Resource (BCR) between September, 2012 and December, 2013. Qualifying tumor samples were obtained from patients who had received no prior treatment for their disease (chemotherapy or radiotherapy). Specimens were shipped overnight from 13 tissue source sites (TSS) using a cryoport that maintained an average temperature of less than -180°C. In addition to tumor samples, each frozen primary tumor specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the tissue source site). Adjacent non-tumor tissue was also submitted for a subset of cases.

Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Pathology quality control was performed on each tumor and adjacent normal tissue (if available) specimen from a frozen section slide prepared either by the BCR or by the TSS. Hematoxylin and eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumor specimen was histologically consistent with mesothelioma and the adjacent tissue specimen contained no tumor cells. The percent tumor nuclei, percent necrosis, and other pathology annotations were also assessed. Tumor samples with >60% tumor nuclei and ≤20% necrosis were submitted for nucleic acid extraction.

#### 1.1.2 Sample Processing

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

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

**1.1.3 Sample Qualification**

The BCR received tumor samples with germline controls from a total of 187 cases, of which 87 cases qualified and were sent for further genomic analysis. Of the 100 that failed to qualify, 21 cases were disqualified prior to processing, 15 failed for pathology screening, and 64 cases failed due to molecular criteria.

Of the 15 that failed pathologic criteria, 14 failed for absence of tumor cells, and one 1 failed for necrosis. The majority of the 64 cases that failed molecular screening had low normal DNA yields (40 cases). The remaining cases had insufficient tumor DNA (5 cases) or low RNA integrity (19 cases). The difference of 8 samples represents those we removed at the very beginning of the AWG as 4 had neoadjuvant therapy, 3 were unpaired (i.e. tumor and normal appeared unmatched), and 1 did not have sufficient DNA.

**1.1.4 Pathology Review**

Aperio© scanned H&E stained slides provided by the tissue source sites from 79 tumors were reviewed according to the 2015 WHO classification as definite MPM subtyped as epithelioid, biphasic or sarcomatoid(1). Where subtyping could not be achieved, tumors were classified as not otherwise specified (NOS). H&E stained slides from frozen sections were reviewed in all 79 cases. In 73 cases, an additional H&E slide was provided from a representative formalin fixed paraffin embedded (FFPE) tissue block. Whenever possible, immunohistochemical staining results were obtained from

primary pathology reports. Six pathologists participated in the pathology review and discrepant interpretations were resolved by consensus between two pathologists who reviewed all cases.

### 1.2 PATHOLOGY REVIEW RESULTS

Tumors were reclassified as summarized in Supplementary Figure 1A. A total of 74 tumors were accepted as MPM with 52 epithelioid (Supplementary Figure 1B&C), 13 biphasic (Supplementary Figure 2D&2E), 3 sarcomatoid (Supplementary Figure 2F&2G) and 6 NOS. Five tumors were excluded either because the diagnosis of MPM could not be confirmed (3 cases) or because the normal tissue submitted was not adequate (1 case), or because the frozen sample submitted as tumor contained only normal tissue (1 case).

The following immunohistochemical stains had been performed at the primary institutions. Keratin showed positive staining with the following antibodies: AE1/AE3 (n=25), CK7 (n=18), CAM5.2 (n=7) and keratin (not further specified, n=6). The following mesothelial stains were positive: calretinin (n=72), WT1 (n=55) and D2-40 (n=35). The following carcinoma markers were negative: CEA-NOS (n=25), CEA-polyclonal (n=7), CEA-monoclonal (n=7), CD15 (n=19, one focally positive tumor was excluded), BER-EP4 (n=25, 2 were focally positive), MOC-31 (n=17, 2 focally positive, 2 positive), and TTF-1 (n=52, 1 positive case was excluded).

## 2. Genome-wide LOH Validation Datasets

### 2.1 ICGC cohort description

Tumor-normal matched DNA samples from Japanese MPM cases were analyzed by whole exome sequencing. Data was analyzed using an in-house variant caller, Karkinos (http://sourceforge.net/projects/karkinos/), to detect single nucleotide variations (SNV) and allelic somatic copy number alterations (sCNA). Among 80 Japanese cases (48 frozen tissue samples and 32 primary cell lines from cancer patients within 20 passages), two cases showed genome-wide LOH (Figure 2B in the main manuscript).

SNV detection and estimation of tumor cellularity were previously described(2, 3). We detected sCNA in an allelic manner and generated an allelic copy number plot from exome sequencing data using karkinos, which detects sCNA by calculating the ratio of

allele specific reads between matched tumor and normal samples at the positions with heterozygous SNPs, by normalization of raw data, adjustment of GC contents, wavelet de-noising, and multi-state HMM.

Both genome-wide LOH cases showed loss of one copy of most chromosomes, whilst retaining two copies of chromosomes 5 and 7. Allelic copy number plots for sample K2F2-A45 (primary cell line at passages 9), and sample K2F2-H60 (frozen tissue sample) are shown in Figure 2B of the main manuscript. K2F2-A45 has a homozygous deletion of *SETDB1*, and *CDKN2A*, but doesn't have a driver mutation in *TP53*. K2F2-H60 has a stopgain SNV of *SETDB1*, a splice site mutation of *TP53*, and a frameshift deletion of *NF2*. No somatic mutations in *BAP1*, *SETD2*, and *PBRM1* were detected in either case.

### 2.2 Brigham and Women's genome-wide LOH cohort description

As cytogenetic data were not available on the TCGA MPM cases, we sought independent validation of this finding in a series of 916 MPM cases prospectively karyotyped at Brigham and Women's Hospital (BWH) in Boston, MA, between 1990 and 2013. Among these, 16 pleural MPM cases (1.7%) with a near-haploid karyotype (Supplementary Table S3) have been identified.

Whilst the mean age of the patients in the BWH MPM cohort was >65, the mean age of the near-haploid subset is 54. Moreover, this subgroup had a significantly higher percentage of female patients (10/16; 62.5%) compared to the overall BWH cohort (211/916; 23% female) (P = 0.005).

## 3. Whole Exome Sequencing and Multicenter Variant Calling

### 3.1 Exome enrichment and sequencing

Genomic libraries were prepared using the Illumina Paired End Sample Prep Kit following the manufacturer's instructions. Enrichment was performed as described previously(4) using the Agilent SureSelect Human All Exon 50Mb kit following the manufacturer's recommended protocol.

Each exome was sequenced using a 75bp paired-end protocol on an Illumina HiSeq DNA Analyzer, to produce approximately 10Gb of sequence per exome. Sequencing reads were aligned to the human genome (NCBI build 37) using the

Burrows-Wheeler Aligner (BWA) algorithm with default settings(5). Reads which were unmapped, PCR-derived duplicates, or outside the targeted region of the genome, were excluded from the analysis. The remaining uniquely mapping reads provided 70–90% coverage over the targeted exons at a minimum depth of 30x.

## 3.2. Whole Exome Sequencing and Multicenter mutation calling

### 3.2.1 Wellcome Trust Sanger Institute

The CaVEMan (Cancer Variants through Expectation Maximization) algorithm was used to call single-nucleotide substitutions(4). To call insertions and deletions, we used split-read mapping implemented as a modification of the Pindel algorithm(6, 7). Mutations were annotated to Ensembl version 58 using the VAGrENT algorithm(8). Post-processing of mutation calls was performed to remove recurrent artifacts from the set of initial variant calls as described elsewhere(9). Briefly, regions of recurrent mis-mapping or sequencing errors are removed through excluding regions near homopolymer tracts, germline indels, or with highly recurrent errors in a panel of normal exomes. All indels (n=548) and all putative driver substitutions (n=80) were reviewed by manual inspection of the sequence data. For the indels, 303 variants were true positive, 173 false positive, and 72 ambiguous. False positive indels were excluded from subsequent analyses. From the subset of inspected substitutions 77/80 were true positive.

### 3.2.2 Broad Institute

The Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose) performed quality control (QC) on the BAM files, point mutation calling, small insertion and deletion detection, annotation of detected mutations, filtering for OxoG artifacts and filtering by “panel-of-normals”. These steps are described in further detail below.

1. QC on BAM files: The sample cross-individual contamination levels were estimated using the ContEst program(10).
2. Somatic point mutation calling: The MuTect algorithm(11) was used to detect somatic single nucleotide variants (SNV).
3. Small insertion and deletion detection: The Indelocator algorithm (https://www.broadinstitute.org/cancer/cga/indelocator) was used to detect small insertions and deletions (InDel).
4. SNV and InDel annotations: variants detected by MuTect and Indelocator were annotated using Oncotator(12). Oncotator mapped somatic mutations to

respective genes, transcripts, and other relevant features. These annotations correspond to the fields in the TCGA Mutation Annotation Format (MAF) files version 2.4 (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification).

5. Filtering for OxoG artifacts: 1741 G>T/C>A transversions that are a consequence of heating, shearing, and oxidative damage to the DNA during genomic library preparation(12) were filtered out of the Broad call set.
6. Filtering by “panel-of-normals”: The sites of detected SNV and InDel were examined against a panel of 8313 normal samples (PoN). This identified any SNV or InDel that is a recurrent artifact and removed variants that were found outside coding regions. As a result of the filtering, 865 SNVs and 15639 InDels were removed, leaving a total 3361 mutations in the final MAF (3011 SNVs, 66 insertions and 284 deletions).

### 3.2.3 University of British Columbia Cancer Research Centre

Strelka(13) v1.0.6 was used to identify somatic SNVs, and short InDels from the WES dataset. 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 the dataset. Blood or normal tissue was used as matched normal specimen. The variants were subsequently annotated using SnpEff(14), and the COSMIC(15) v61 and dbSNP(16) v137 databases.

### 3.2.4 Multicenter Mutation Calling

SNV and InDels data from three centers, Broad Institute (BI; 4067 SNVs, 186 deletions, and 52 insertions), Sanger Institute (SI; 3687 SNVs, 434 deletions, and 114 insertions) and University of British Columbia (UBC; 8370 SNVs, 167 deletions, and 40 insertions) were used in the creation of the MCC MAF file. In order to create the set of SNVs, any SNV that appeared in at least two of the three centers was included in the MCC MAF file used for subsequent analysis. Also, any non-coding SNV or InDel was removed from the final set. In cases where coding status from SI was specified, that coding annotation was prioritized and used to determine the coding status of the SNV or InDel; otherwise, BI's annotations were used. Consequently, 3011 SNVs were included in the final MCC MAF file. For insertions and deletions, all SI's manually reviewed InDels

(278 deletions and 64 insertions) and 8 additional InDels (6 deletions and 2 insertions) that weren't manually reviewed by SI, but called by both BI and UBC and manually reviewed by BI, were included in the final set of InDel calls. Thus, 350 InDels (284 deletions and 66 insertions) were included in the final MCC MAF file.

Because *BAP1* alterations are of specific interest in MPM, RNA-Seq InDels (18 InDels) and exome calls in *BAP1* from BI, SI and UBC were closely analyzed by Analysis Working Group members at the University of North Carolina. Combining the calls across all centers yielded 26 *BAP1* SNVs and InDels (9 SNVs and 17 deletions). Three additional *BAP1* mutations (two SNVs and one deletion) that were deemed to be affecting cancer progression by the Analysis Working Group were included in the final set of calls as well. The three *BAP1* mutations were: intragenic SNV that affected *BAP1* expression level in ZN-A9VS, low allelic fraction SNV in 3U-A98G, and 48 basepair long intronic deletion in 3U-A98G. In all, 29 *BAP1* mutations (11 SNVs and 19 deletions) were in the MCC MAF file (See Supplementary Table S2A for details).

In order to determine the alternate and reference allele counts for mutations, BI and SI computed allele counts. For SNVs, alternate and reference allele counts were determined by the BI's Mutation Validator tool. For InDels, SI calculated allele counts for all but 8 InDels in the MCC MAF file. The allele counts for these remaining InDels, which were called by BI and UBC, were determined by BI's Mutation Validator tool.

**3.2.5. Mutational spectrum**

The observed mutational spectrum of somatic SNVs for each sample in the cohort was extracted by considering each variant in its pyrimidine context (that is, C>A, C>G, C>T, T>A, T>C and T>G). Each of the six types of base substitution were further split into 16 subcategories by the reference base immediately 5' and 3' to the mutated base. For each patient, counts of SNVs in each of these 96 channels then represent the observed mutational spectrum. The dataset did not have sufficient information content for a formal signature extraction for two reasons. First, the low number of variants in each tumor (due to the low overall mutation burden combined with a footprint of only 30Mb in an exome) limits statistical power to detect signatures. Second, across patients (allowing for the aforementioned low numbers), there was very little evidence for variation in mutation spectrum, with the exception of the one hypermutated patient. Given that formal signature extraction relies on having individual signatures distributed unequally across patients, any such extraction would be unreliable. Although mutation

burden was too low for signature extraction at the level of individual patients, we felt that an analysis comparing the mutational spectra combined across asbestos-exposed patients versus asbestos-unexposed was well powered. This analysis was performed using a chi-squared test.

# 4. Copy number analysis

## 4.1 METHODS

Affymetrix SNP 6.0 arrays were used to hybridize genomic DNA from each tumor and normal sample using standard protocols at the Genome Analysis Platform of the Broad Institute(17). Briefly, from raw CEL files, Birdseed was used to infer preliminary copy number at each probe locus(18). For each tumor, tangent normalization was applied to estimate genome-wide copy number. Tangent normalization is based on the observation that the linear combination of all normal samples that are most similar to the tumor tends to match the noise profile of the tumor better than any set of individual normal samples; this linear combination is therefore used to divide the tumor signals (19) (http://www.broadinstitute.org/cancer/cga/copynumber_pipeline). Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation(20), during which regions corresponding to germline copy number alterations were removed. Ziggurat Deconstruction was then applied to assign a length and amplitude to each identified copy number change, in a way that accounts for different copy number values inferred across the locus from the heterogeneous cell population(21).

Allelic CN, whole genome doubling, subclonality, and purity and ploidy estimates were calculated using the ABSOLUTE algorithm(22) and CBS-derived segmented CN values were re-centered using the *In Silico* Admixture Removal (ISAR) procedure(23) and significant focal CN alterations were identified from using GISTIC 2.0.22(21).
CN clustering was based on total integer arm-level CN, normalized over 4 copies for tumors estimated to have been whole genome doubled once, and 8 copies for tumors estimated to have been whole genome doubled twice. Tumors were clustered based on thresholded CN at re-occurring alteration peaks from GISTIC analysis (all_lesions.conf_99.txt file). Clustering was done in R based on Manhattan distance using Ward's method. Allelic CN derived from ABSOLUTE was used along with visual

inspection of relative CN to determine regions of loss of heterozygosity (LOH) and homozygous deletions.

The FACETS algorithm(24) was used to perform allele-specific copy number and LOH analysis from WES data. Briefly, read count information was extracted from paired tumor-normal whole-exome sequencing BAM files. Total log-copy-ratio (logR) was computed from the total read count across germline SNP sites in tumor versus normal. GC-normalization was done using loess regression. Allelic imbalances were assessed using the variant allele log-odds-ratio (logOR) at heterozygous SNP sites. A joint segmentation analysis was applied to identify regions of the genome with copy number alterations by extending the Circular Binary Segmentation (CBS) algorithm to a bivariate change-point detection method based on the Hotelling T2 statistic. Integer copy number was estimated using a genotype-mixture model correcting for tumor purity, ploidy and clonal heterogeneity. Software is available at https://github.com/mskcc/facets.

**4.2 RESULTS**

The SCNA landscape of MPM was characterized by frequent recurring focal and arm-level deletions, but no recurrent focal amplifications. Using ABSOLUTE-inferred allelic CN information(22), we were able to determine if deletions at each locus were homozygous, hemizygous, or heterozygous. Several tumor suppressor genes previously found to be altered in MPM (25) were recurrently homozygously deleted, including *CDKN2A* (36/74), *BAP1* (12/74), *NF2* (6/74), *PBRM1* (3/74), *SETD2* (3/74), and *PTEN* (2/74). *PTPRD* and *RBFOX1* were also recurrently deleted, but both have previously been reported to be fragile sites in cancer genomes and their mRNA expression did not correlate with CN(26, 27). Additionally, many of the focally deleted tumor suppressors were also significantly recurrently mutated in the cohort (Fig. 1 in main manuscript).

While CN and mutational status of *BAP1* and *SETD2* were not correlated with overall survival, *NF2* status (stratified by homozygous deletion, mutation, 1-copy LOH, 2-copy LOH, and wild-type; Cox model P=0.024) and *CDKN2A* status (stratified by homozygous deletion and other; Cox model $P=7.3\times10^{-6}$) conferred significant differences in overall survival. Biallelic inactivation of *NF2* was also significantly enriched in tumors of patients with no history of asbestos exposure (Chi-square

P=0.0027). The 19/74 tumors harboring biallelic inactivation of *BAP1* were all epithelioid (Chi-square P=0.019).

Hierarchical clustering revealed 6 distinct CN clusters (Supplementary Figure 2A). One group (cluster 1) consisted of tumors with focal deletions in chromosome 3p around the *BAP1*/*PBRM1*/*SETD2* locus as well as partial or whole deletion of chromosome 22, but few other SCNAs. Cluster 4 consisted of tumors either lacking broad SCNAs entirely or exhibiting high levels of genome-wide LOH. With one exception, no homozygous deletions were found in this group. Cluster 2 consisted of 4 hyperdiploid/hypotriploid samples with more arm-level gains than losses. Cluster 3 was enriched for *CDKN2A* homozygous deletions but was otherwise generally chromosomally stable. Clusters 5 and 6 were both characterized by chromosomal instability, but tumors in cluster 5 had partial telomeric loss of chromosome 1p, whereas those in cluster 6 tended to have whole-arm or partial centromeric loss of 1p. The majority of 1p deletions affected only the telomeric or the centromeric ends of 1p, suggesting positive selection for deletion of either, or negative selection for deletion of the genomic region.

In GISTIC 2.0 analysis of MPM with wild type or 2-hit *BAP1* inactivation, we found no significant differences in focal SCNAs, other than 3p focal deletions containing BAP1 itself. However, there were more overall SCNAs in tumors without *BAP1* alterations. Total numbers of amplifications and deletions in chromosomal arms were significantly greater in tumors with wild type *BAP1* (median 15.5 vs. 9.5, P<0.01). All tumors in CN cluster 4, which contain few SCNAs apart from the 3p focal deletion containing *BAP1* and chromosome 22/6q deletion, had at least one allele with a *BAP1* alteration (Supplementary Figure 2D).

FACETS and ABSOLUTE analyses identified three cases with extensive loss of heterozygosity (Supplementary Figure 2E, 2F and 2G) and these cases were further analyzed using complementary methodologies to characterize their molecular features.

## 5. Gene Expression (RNA sequencing)

### 5.1 RNA library construction, sequencing, and analysis

One µg of total RNA was converted to mRNA libraries using the Illumina 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(28), with FASTQ files generated by CASAVA. RNA reads were aligned to the hg19 genome assembly using MapSplice(29) 0.7.4. Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1 (http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf), using RSEM(30) and normalized within-sample to a fixed upper quartile. Further details on this processing are available in the Description file at the DCC, under the V2_MapSpliceRSEM workflow (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/meso/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_MESO.IlluminaHiSeq_RNASeqV2.mage-tab.1.0.0/DESCRIPTION.txt) or at CGHUB (https://cghub.ucsc.edu/docs/tcga/UNC_mRNAseq_summary.pdf).

**5.2 RESULTS**

Unsupervised gene expression using variably and highly expressed genes was used to identify 4 mRNA-based clusters. Using ClaNC, we identified a set of 1,800 genes that strongly classified each of the four clusters and associated them with clinical and genomic features (Supplementary Figure 3A). Cluster 1 was predominantly epithelioid, and significantly less likely to have loss of *CDKN2A/B* (Fisher's exact test, p<0.001, Supplementary Figure 3B), contained 5/8 *SETD2* mutations (Fisher's Exact Test, P=0.03), had high expression of innate immune cells, *MSLN* (concordant with serum levels), and had the best overall survival (Supplementary Figure 3C). In contrast, cluster 3 predominantly contained non-epithelioid histologies and had significantly poorer outcomes compared to cluster 1 (HR 4.3, P=0.00009). Samples in this cluster expressed genes involved in NOTCH and WNT signaling and had high levels of cell cycle gene expression. This group of patients was also significantly older than the other three (t test, P=0.0008, Supplementary Figure 3D). Cluster 4, while epithelioid, was similar to 3 - with high expression of cell cycle genes and a poor overall survival compared to 1 (HR 6.8, P=0.00003). Cluster 2 was defined by high expression of *BAP1* and *CLDN* genes. This group was significantly less likely to have *BAP1* alterations (t test, P=0.001, Supplementary Figure 3A, 3B, 3E), and had similar survival to cluster 1, with a younger age at diagnosis. Using the gene signatures from Bueno(25) and de Reynies(31) for validation, samples clustered similarly to our TCGA analysis, with the highest concordance observed for non-epithelioid samples.

# 6. DNA Methylation Profiling

## 6.1 METHODS

The Illumina Infinium HM450 array(32) was used following standard protocols. Briefly, genomic DNA (1 µg) was treated with sodium bisulfite and recovered using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA). Bisulfite-converted DNA samples were amplified, fragmented and hybridized to BeadChips, followed by a locus-specific base extension with labeled nucleotides (cy3 and cy5). BeadArrays were scanned and the raw data imported into custom programs in the R computing language for pre-processing and calculation of DNA methylation β values for each probe and sample. Quality control and probe exclusions were performed using standard protocols as described(33).

Unsupervised consensus clustering was performed in the Bioconductor package ConsensusClusterPlus(34) v1.22.0, with Euclidean distance and partitioning around medoids (PAM) and was applied to the DNA methylation data using the most variable 1% of CpG probes. Fisher's exact test was used to test for associations of DNA methylation clusters with other platform clusters and significantly mutated genes.

Leukocyte fraction was estimated as described previously(22). As a source of leukocyte DNA methylation level, we used DNA methylation data of peripheral blood mononuclear cells from six healthy donors(35) (GSE35069).

To identify CpG probes associated with *BAP1* and *SETD2* status, we used empirical Bayes-modified t-tests as implemented in the limma package(36). The correlations between *SETD2* and *BAP1* status and DNA methylation was strong enough to support two criteria for probe selection - signature genes with FDR<0.01, and the mean difference of β values between altered and wild type samples of more than 0.3.

## 6.2 RESULTS

Three robust DNA methylation clusters were identified based on the most variable CpG loci on the Illumina array (Supplementary Figure 4A). Cluster 3 had a higher leukocyte fraction and lower purity than the other two and was associated with miRNA cluster 5, lncRNA cluster 4, PARADIGM cluster 4, and iCluster 4 (Supplementary Figure 4A and 4B). Cluster 2 was enriched with *SETD2* mutations, which have been associated with altered DNA methylation in cancer(33, 37). To explore further the

correlation of *SETD2* mutations and DNA methylation, we looked for CpG sites that associated with *SETD2* status and found more than 200 differentially methylated CpG sites (Supplementary Table S6A). The association between *BAP1* status and DNA methylation was strong enough to support two criteria for probe selection, signature genes were selected to have FDR<0.01 and the mean difference of beta values between altered and wild type samples was more than 0.3. We found 84 CpG probes that had significantly higher methylation in *BAP1* inactivated samples (Supplementary Figure 4C). For most of the probes, higher methylation was associated with low expression for the corresponding gene (Supplementary Table S6B).

# 7. Noncoding RNA (lncRNAs and miRNAs) expression analyses

## 7.1 METHODS

### 7.1.1 RNA-seq read mapping

RNAseq FASTQ files for TCGA data (n=74) were downloaded from CGHub(38), while those for from the validation cohort (Bueno(25), n=211) were downloaded from the European Genome-phenome Archive (EGAS00001001563). For both datasets, FASTQ files were processed using STAR(39) v2.3.0e with the following parameters: minimum/maximum intron size set to 30 and 500,000, respectively, noncanonical, unannotated junctions were removed, the maximum of tolerated mismatches set to 10, and the outSAMstrandField intron motif option enabled. The Cuffdiff command included with Cufflinks(40) v2.0.2 was used to calculate the fragments per kilobase of exon per million fragments mapped (FPKM) with upper quartile normalization, fragment bias correction, and multiread correction enabled. All other options were set to default. Ensembl v82 gene annotations were used.

We generated microRNA sequence (miRNA-seq) data using methods described previously(41, 42) with miRBase v16 annotations, and assigned 5p and 3p mature strand names using miRBase v20.

### 7.1.2 Unsupervised consensus clustering

For the TCGA cohort, we extracted 347 expressed (mean FPKM ≥1) and highly variable (95th percentile, FPKM variance) lncRNAs from a normalized abundance matrix of 8167 lncRNAs (7671 lincRNA and 496 processed transcripts). We identified groups

of samples with similar abundance profiles by unsupervised consensus clustering with ConsensusClusterPlus v1.36.0. Calculations were performed using Pearson correlations, hierarchical clustering, 20,000 iterations, and a random gene fraction of 0.975 in each iteration.

For the Bueno validation cohort, we clustered 420 expressed, high-variance lncRNA profiles using Pearson correlations, partitioning around mediods, 5000 iterations, and a random 0.95 gene fraction in each iteration. We selected a four-cluster solution for both cohorts.

For miRNA mature strand data, we selected an input file containing reads-per-million (RPM) data for the 303 (25% of 1212) most-variant 5p or 3p mature strands. We used unsupervised consensus clustering with ConsensusClusterPlus v1.36.0, using a Pearson distance, hierarchical clustering, 20000 iterations and a gene fraction of 0.95.

To generate a heatmap for subtypes, we first used a SAM multiclass analysis(43) to identify differentially abundant lncRNAs or miRNAs (FDR<0.05), which we filtered to retain lncRNAs or miRs that were expressed at least a mean of 5 FPKM or 25 RPM respectively(44). We transformed each row of the matrix by log10(RPM+1), then used the pheatmap R package (v1.0.2) to scale and cluster only the rows, and to generate the heatmap.

An FDR threshold of 0.05 was set for both miRNA and lncRNA.

#### 7.1.3 Maximal-search associations with survival

For both the TCGA and the validation cohorts, we identified miRNAs and lncRNAs that were associated with overall survival. We used R functions from CutoffFinder(45) v2.1 to determine the RPM or FPKM value that stratified samples into two groups(46), then adjusted the corrected p values for multiple testing with the Benjamini-Hochberg method.

#### 7.1.4 Correlations with EMT scores

We used MatrixEQTL(47) to calculate Spearman correlations (FDR<0.05) between RNAseq-based EMT scores (Supplementary Section 10) and miR RPMs or lncRNA FPKMs.

### 7.2 RESULTS

We assessed two types of noncoding RNAs that may offer molecular insights into MPM(48): microRNAs (miRNAs) and long noncoding RNAs (lncRNAs).

miRNA mature strands (miRs) are associated with MPM (48-51). Here, we used unsupervised consensus clustering to identify five miR subtypes (Figure 6H). These were associated with purity ($P=1.0x10^{-7}$), leukocyte fraction ($P=1.1x10^{-6}$), EMT scores ($P=1.7x10^{-7}$), and 5-year survival ($P=6.6x10^{-4}$). They were concordant with subtypes for lncRNAs (see below, $P=1.2x10^{-12}$), mRNA ($P=1.2x10^{-14}$), iCluster ($P=1.3x10^{-12}$), PARADIGM ($P=3.0x10^{-13}$), and were associated with subtypes for SCNA ($P=2.6x10^{-4}$), and DNA methylation ($P=1.6x10^{-4}$).

LncRNA expression can be more specific for cell type than coding gene expression(52-54). To our knowledge, only one expression lncRNA profiling report, based on NCode long noncoding microarrays, is available for MPM(55). Recently, multiplatform data that included transcriptome sequencing, and analyses focused on protein-coding genes, were reported for a large MPM cohort(25); this 'Bueno' transcriptome data served as an independent lncRNA data source.

For the TCGA cohort, we used unsupervised consensus clustering with transcriptome sequence data to identify 4 lncRNA subtypes (Figure 6A). These were associated with leucocyte fraction ($P=3.2x10^{-4}$), EMT scores ($P=1.8x10^{-4}$), and 5-year survival (log-rank $P=1.4x10^{-4}$) (Figures 6A-D). They were strongly concordant with subtypes from mRNA ($P=3.5x10^{-17}$), iCluster ($P=7.7x10^{-14}$), PARADIGM ($P=2.1x10^{-21}$), and miRNA ($P=1.4x10^{-12}$), and were associated with subtypes for SCNA ($P=0.011$), and DNA methylation ($P=0.025$). Samples in lncRNA cluster 1 (n=20, 27%) had better survival, while cluster 4 had a relatively high leukocyte fraction and low purity, many of the non-epithelioid samples, and relatively shorter survival.

We noted that the iCluster, PARADIGM, miRNA and lncRNA subtypes that had the best and worst survival were strongly concordant.

For the Bueno validation dataset(25), we identified 4 lncRNA subtypes that were concordant with the four RNA-seq-based subtypes reported in that work ($P=3.4x10^{-28}$) and were associated with 5-year survival ($P=1.1x10^{-3}$) (Figure 6E). For both the TCGA and the Bueno cohorts, lncRNAs that were differentially abundant between the better-survival subtype and other samples included those known to be associated with cancers in general (e.g. H19, LINC00152, MEG3) or with MPM in particular: NEAT1 and SNHG8(55), and GAS5(56).

EMT is a factor in MPM(31, 57). Many miRNAs(58-60) and lncRNAs(61) have been associated with EMT. We noted that the miR-based and lncRNA-based unsupervised clusters that had better survival had the lowest median EMT scores. In the TCGA cohort, we identified miRs and lncRNAs that were statistically correlated (FDR < 0.05) with RNA-seq-based EMT scores (Table S5A). As expected, miR-200 family members, members of an Xq27.3 genomic miRNA cluster(62), and miR-29 family members(63) were correlated with the scores, as was the oncogenic LINC00152(64) ($\rho$=0.41).

Next, we assessed miRs and lncRNAs that were differentially abundant in the subtypes that had better survival (Figures 6F,G,L; Table S5B). For the TCGA cohort, miR-126, 143-3p and 145-5p(65, 66) were less abundant in miR cluster 1, while miR-193a-3p(67) was more abundant. Among the differential lncRNAs in both the TCGA (lncRNA cluster 1) and the Bueno (lncRNA cluster 4) cohorts (Figures 6F,G), the most highly differential lncRNA RP11−263K4.5 showed a positive fold-change of 63.6 in TCGA and 9.1 in the validation datasets, suggesting that this lncRNA may be a tumor suppressor in MPM.

Finally, for all samples and then for only epithelioid samples, we identified miRNAs that were statistically associated with survival in the TCGA cohort, and lncRNAs that were that were associated with survival in both TCGA and Bueno cohorts (Table S5C). For the full TCGA cohort, miRs with significant adjusted p-values (e.g. miR-514a-3p, 508-3p, 29b-2-5p, 101-3p, Hazard Ratio<1) included those that were negatively correlated to EMT scores, and those that were relatively abundant in the miR cluster with the best survival. For the epithelioid cases in the TCGA cohort, miR-148b-3p ($P=5.7x10^{-4}$) and miR-148a-5p ($P=9.9x10^{-4}$) were statistically significant, though miR-148b-3p was not statistically significant for the full cohort. miR-148b-3p is well known in lung and other cancers(68). A number of lncRNAs were statistically associated with survival in the full TCGA and validation cohorts. For example, GS1-600G8.5 ($P=1.1x10^{-3}$ and 0.036, respectively, Hazard Ratio>1) was less abundant in the good-survival lncRNA subtypes in both cohorts, and was positively associated with TCGA EMT scores. Taken together, the above results suggest that both miRNAs and lncRNAs may be important regulators of EMT and of survival in MPM.

## 8. Microbial sequences

### 8.1 METHODS

We screened RNA and DNA sequence reads with a microbial detection pipeline based on BioBloomTools (BBT), v1.2.4.b1, a Bloom filter-based method for rapidly classifying RNA-seq or DNA-seq read sequences(69). We ran BBT in paired-end mode to screen FASTQ files from 74 tumor RNA-seq libraries, 74 tumor and their respective normal whole exome libraries within the TCGA MPM cohort. In a single-pass scan for each library, BBT categorized each read pair as matching the human, a unique microbial, more than one (multi-match), or no-match filters. We then calculated a reads-per-million (RPM) abundance for each filter.

To detect integration of human herpes virus (HHV) and human papilloma virus (HPV), we performed de novo assembly with ABySS v1.3.4. We assembled only the reads classified by BBT, then merged the sets with Trans-ABySS(70) 1.4.8 to generate a contig working set. We identified breakpoint locations by using BLAT(71) v34 to align to the GRCh37/hg19 and to either 109 HHV or 268 HHV reference sequences. We retained contig alignments in which: a) the aligned human and viral sequences summed to at least 90% of the contig length, and b) the human and viral aligned overlapped by less than 50%. Human breakpoint coordinates were annotated against RefSeq and UCSC gene annotations(72). Breakpoints that had at least 3 spanning mate-pair reads or 5 flanking mate-pair reads were considered potential integration sites.

### 8.2 RESULTS

We assessed microbes using RNAseq and WES data for 74 MPM samples (Supplementary Table S7). The RNAseq libraries returned signals for common microbial contaminants. We found no evidence for genomic integration in 9 libraries that returned positive RPMs: one RNAseq library that showed weak signals for HPV and 8 tumor WES libraries that showed weak signals for HHV4 (EBV). Notably, there were no hits for polyomavirus (including SV40) sequences in any of the libraries.

## 9. Reverse-Phase Protein Array analysis

### 9.1 METHODS

Protein was extracted using RPPA lysis buffer from tumors and RPPA was performed as described previously(73-77). SDS-reduced lysates were adjusted to 1 µg/µL and manually serially diluted. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs), which were probed with 219 validated primary antibodies followed by corresponding secondary antibodies.

Signal was captured using a colorimetric, DAB-based DakoCytomation-catalyzed system. 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). SuperCurveGUI(75, 77) was used to estimate $EC_{50}$ values of the proteins in each dilution series(73). A QC metric(77) was returned to help determine the quality of each slide: if the score was less than 0.8 on a 0-1 scale, the slide was omitted. In most cases, the staining was repeated and the highest QC scoring slide was used for analysis (Level 2 data).

Protein measurements were corrected for loading(75, 77, 78) using median centering across antibodies (level 3 data). Final selection of antibodies was also driven by the availability of high quality antibodies, as assessed by specificity, sensitivity and dynamic range for quantification(79). In total, 219 antibodies and 52 MPM samples were used for the analysis. RPPA arrays were quantitated and processed as described previously(73, 75). Raw data (level 1), SuperCurve nonparameteric model fitting (level 2), and loading corrected data (level 3) were deposited at the DCC.

**9.2 RESULTS**

MPM samples were consensus clustered using 1-Pearson correlation as the distance metric and Wards' linkage algorithm. We identified five robust clusters with differential pathway expression (Supplementary Figure 5A and 5B).

The role of cell signaling networks in MPM was illustrated by computing 12 pathway scores described previously(80). The analysis showed that there were differences among the RPPA clusters for many of these pathways (Supplementary Figure 5C).

RPPA cluster 1 (n=11) showed higher EMT, Hormone receptor, RAS/MAPK, Breast reactive and Core reactive pathway expressions. Significantly higher FIBRONECTIN and COLLAGEN VI, and lower E-CADHERIN and BETA CATENIN expression levels contributed to higher EMT in 1, which also showed high apoptosis

activity; lower expression in PCNA associated with Cell cycle pathway; CHK1PS345, CHK2PT68, MRE11 and RCC1 associated with high DNA damage response pathway activity; and high expression of MYH11 from the Breast reactive pathway activity. Cluster 2 showed relatively high Hormone signaling (breast) pathway activity, which was due to a higher expression of BCL2. Cluster 3 showed relatively low apoptosis, EMT, Hormone receptor and Core reactive activity, and high PI3K/AKT and TSC/mTOR pathway activity, whilst cluster 4 showed high Cell cycle, DNA damage response and EMT activity, coupled with low RAS/MAPK and Breast reactive activity, and had the worst prognosis. Cluster 5 (n=11) had the best prognosis. It showed high activity of DNA damage response pathway, and low activity of EMT, RAS/MAPK, TSC/mTOR, Breast reactive and Core reactive pathways. The most differentially expressed proteins between clusters 5 (best prognosis) and 4 (worst prognosis) were PAI1, CYCLINB1, CAVEOLIN1, EPPK1 and EEF2K. The worst prognosis cluster had significantly higher expression of PAI1, has been associated with poor prognosis in previous studies, and could be a potential therapeutic target in MPM(81, 82). CYCLINB1(83), CAVEOLIN1, EPPK1 and EEF2K(84) could also be prognostically and/or therapeutically relevant.

## 10. Epithelial-Mesenchymal Transition Analysis

### 10.1 METHODS

74 MPM samples were scored based on expression of epithelial-mesenchymal transition (EMT) signature genes using a method previously developed by our group(85). Briefly, the EMT score for each sample is calculated as the mean expression of epithelial markers subtracted from the mean expression of mesenchymal markers. Higher EMT scores correlate with a more mesenchymal profile.

### 10.2 RESULTS

EMT is known to be common in MPM(31, 57). Across multiple tumor types, MPM possessed the second highest overall EMT score after sarcoma (Fig. 7B). Although the vast majority of MPM tumors had EMT scores greater than 0, indicating higher expression of mesenchymal versus epithelial genes, the individual scores varied significantly. Histology correlated with EMT score, with epithelioid MPM possessing the lowest EMT score (Figure 7A). However, a few epithelioid tumors possessed high EMT

scores), atypical for this histology and suggesting distinct epithelial and mesenchymal biologies exist within otherwise similar histologies that may predict distinct responses to targeted therapies. Across all histologies, increasing EMT scores were significantly correlated with higher leukocyte fraction (r=0.30; P=0.008). EMT is known to contribute to immune escape[1] and we have previously observed an association between EMT score and increased expression of immune checkpoint and other potentially targetable immune genes. In light of these data and recent clinical data suggesting a subset of MPM patients respond to immune checkpoint blockade, we investigated whether EMT was associated with expression of immune genes in MPM. As in other solid tumors(86), EMT score was significantly associated with the expression of many potentially targetable immune checkpoint genes, including OX40 ligand (*OX40L*), Transforming Growth Factor Beta 1 (*TGFB1*), *B7-H3* (aka Cluster of Differentiation 276, *CD276*), OX40 receptor (*OX40*) and Programmed Death-Ligand 2 (*PD-L2*) (P<0.001), while the V-domain Ig suppressor of T cell activation (*VISTA*), a negative regulator of T cell proliferation and T-cell cytokine production(87) was strongly associated with low EMT score (r=-0.476, P=$1.56 \times 10^{-5}$). However, *VISTA* expression was higher in MPM compared to all other tumor types available in the TCGA due to high *VISTA* expression in most epithelioid MPM samples. Thus, EMT score could be used to predict enrichment of specific immune checkpoint targets for selection from already available immunotherapy agents. The correlation between EMT and immune checkpoint gene expression also supports combinatorial strategies to target both immune checkpoint and EMT simultaneously. Additionally, these data suggest that epithelioid histology, with its relatively low EMT score, may predict response to targeting of VISTA, a negative immune checkpoint molecule that, when blocked or removed, appears to slow tumor growth in a mechanism distinct from PD-1 signaling in preclinical models(88, 89). As previously observed(90), we found that EMT scores were also significantly associated with subgroup classification via unsupervised analysis across multiple platforms including integrative platforms such as iCluster and PARADIGM, as well as individual platforms like mRNA, miRNA, lncRNA, methylation and RPPA (Fig. 7A).

## 11. iCluster Supplementary Information

### 11.1 METHODS

We utilized iCluster(91), which formulates the problem of subgroup discovery as a joint multivariate regression of multiple data types with reference to a set of common latent variables that represent the underlying tumor subtypes 1-4.

Five molecular assay platforms - SCNA, DNA methylation, mRNA expression, lncRNA and miRNA expression were provided as input. Data were pre-processed using the following procedures. CBS-segmented SCNA data was further reduced to a set of non-redundant regions as described(92). For methylation data, the median absolute deviation was employed to select the top 4000 most variable CpG sites after β-mixture quantile normalization(93). Methylation probes with >20% or more missing data and those corresponding to SNP and autosomal chromosomes were removed. For mRNA, lncRNA, and mature strand miRNA sequence data, poorly expressed genes were excluded based on median-normalized counts, and variance filtering led to a list of reduced features for clustering. mRNA, lncRNA and miRNA expression features were log2 transformed, normalized and scaled before using as an input to iCluster.

## 11.2. VALIDATION COHORTS

### 11.2.1 Bueno Cohort

Normalized gene expression profiles for 211 MPM samples previously described(25) (Bueno cohort) were log-transformed and row-centered. For each sample, Pearson correlation was computed between the profile and the centroids of the four iClusters across the detected 2,606 of the 2,807 signature genes; the sample was assigned to the iCluster with the highest correlation. Kaplan-Meier plots were generated and multivariable Cox regression analysis was performed on the assigned tumor samples, where iCluster assignment, histology, and age were considered as covariates. We performed a similar analysis for the 141 epithelioid samples, employing 2,123 of the 2,292 signature genes for this histological subtype. Kaplan-Meier survival curves were generated and Cox regression analysis was performed on the assigned epithelioid tumor samples, where iCluster assignment and age were considered as covariates.

No significant predictive value of iCluster assignments was confirmed for the whole 211-case cohort (Supplementary Figure 6A and 6B), which could be attributed to the vastly different histological type distributions between the TCGA and the Bueno cohorts. However, when the analysis was restricted to the epithelioid cases, cases

assigned to epithelioid iCluster 1 had a significant survival advantage, even when adjusted for age (Fig. 4D, Supplementary Figure 6C)

**11.2.2 Lopez-Rios Cohort:**

We log-transformed and row-centered the gene expression data (Affymetrix U133A arrays) from 52 previously published MPM cases(94) (Lopez-Rios) for 1,995 genes from the 2,807 signature genes present on the Affymetrix chip. The cohort size was too small to allow histology-based stratification, so we did not perform an epithelioid-only analysis. Pearson correlation coefficients were computed between each sample's profile and the centroids of the four iCluster groups. Each sample was assigned to the iCluster with the highest correlation (Supplementary Figures 6D and 6E). Log rank test was performed on the 49 samples with survival data, and demonstrated a significant survival difference between clusters (Supplementary Figure 6F). Multivariable Cox regression analysis was performed, with histology and age as covariates (Supplementary Figure 6G).

# 12. PARADIGM supplementary Information

**12.1 METHODS**

We used median centered, log scaled mRNA expression and SCNA GISTIC2(21) data to calculate inferred pathway activity levels using PARADIGM(95). Integrative PARADIGM analysis of the 74 MPM cases clustered using ConsensusClusterPlus(34) identified 4 distinct clusters. To compare Cluster 4 (worst prognosis) to Cluster 1 (best prognosis), we ran PATHMARK(37) on the statistically significant differential activities obtained from SAM to extract connected components of the global PARADIGM regulatory network. Activities that fall outside 2 standard deviations outside of the empirical distribution of the statistically significant differentials are included the final result. A network connection is extracted if both vertices in that connection pass the filter. Networks are then visualized using Cytoscape(96) and CircleGraph(97).

**12.2 RESULTS**

Cluster 4 patients have significantly worse survival and show upregulation in AURKA, E2F targets, G2M checkpoints, as well as PI3K and mTOR pathways. This cluster faithfully recapitulates platform-specific clustering. Cluster 1 patients have the best prognosis, and have upregulation in EGFR signaling, whilst kinase subnetworks are downregulated. There is an enrichment of epithelioid patients in this cluster, however, even upon correction, the group remains distinct. Patients are also more likely to have undergone pneumonectomy. Patients in clusters 2 and 3 have similar prognosis, but are genomically distinct.

# 13. Regulome Explorer

## 13.1. METHODS

### 13.1.1 Integrated Analysis and Interactive Exploration

We have integrated all of the data types produced by TCGA and described in this paper into a single "feature matrix" for MPM. From this comprehensive dataset, significant pairwise associations have been inferred and can be visually explored using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). This application allows interactive exploration of significant associations between molecular features, between molecular features and derived numeric features, and between molecular and categorical features, such as clinical parameters or cluster assignments. In addition to associations inferred directly from the TCGA data, additional sources of information are integrated into the visualization (e.g., NCBI Gene, miRBase, UCSC Genome Browser, etc).

### 13.1.2 Feature Matrix Construction and Pairwise Statistical Significance

A feature matrix was constructed using all available clinical, sample, and molecular data for 74 unique patient/tumor samples. The molecular data includes all analytical platforms described here. For mRNA and miRNA expression, quantification files were log2 transformed, and filtered to remove low-variability targets (bottom 25%). For methylation data, probes were filtered to remove the bottom 25%. For somatic mutations, several binary features indicating the presence or absence of a mutation in each sample were generated.

Statistical association among data types was evaluated by pairwise comparisons within the feature matrix. P values for the associations between and among clinical and molecular data were computed using appropriate statistical tests for each pair. To account for multiple-testing bias, the p value was adjusted using the Bonferroni correction.

# 14. VISTA Immunohistochemistry

## 14.1. Methods

Two epithelioid MPM cases contributed to the TCGA cohort by MSKCC (TCGA-SC-A6LQ-01 and TCGA-SC-A6LM-01) were selected for VISTA immunohistochemical studies based on availability of additional FFPE tumor tissue. Blocks were sectioned, de-paraffinized, and stained with a fully automated system (Benchmark ULTRA; Ventana Medical Systems, Tucson, AZ), using the rabbit monoclonal anti-VISTA antibody, clone D1L2G, 0.1 µg/mL (Cell Signaling Technology, Danvers, MA, USA). Additionally, VISTA expression was assessed in normal mesothelial lining from pleura and benign pleuritis with reactive mesothelial proliferation (derived from the Department of Pathology at MSKCC). Spleen and colon FFPE tissue was used as positive control for antibody specificity.

REFERENCES

1. Travis WD, Brambilla, E., Burke, A.P., Marx A., Nicholson A.G. WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart. 4 ed. Lyon: International Agency for Research on Cancer; 2015.

2. Kakiuchi M, Nishizawa T, Ueda H, Gotoh K, Tanaka A, Hayashi A, et al. Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nature genetics. 2014;46:583-7.

3. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nature genetics. 2014;46:1267-73.

4. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011;469:539-42.

5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589-95.

6. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865-71.

7. Raine KM, Hinton J, Butler AP, Teague JW, Davies H, Tarpey P, et al. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End

Sequencing. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2015;52:15 7 1-2.

8. Menzies A, Teague JW, Butler AP, Davies H, Tarpey P, Nik-Zainal S, et al. VAGrENT: Variation Annotation Generator. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2015;52:15 8 1-1.

9. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. The New England journal of medicine. 2011;365:1384-95.

10. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27:2601-2.

11. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013;31:213-9.

12. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: cancer variant annotation tool. Human mutation. 2015;36:E2423-9.

13. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811-7.

14. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. 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. 2012;6:80-92.

15. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic acids research. 2010;38:D652-7.

16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29:308-11.

17. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature genetics. 2008;40:1166-74.

18. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature genetics. 2008;40:1253-60.

19. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609-15.

20. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557-72.

21. 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:R41.

22. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413-21.

23. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nature genetics. 2013;45:1134-40.

24. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.

25. Bueno R, Stawiski EW, Goldstein LD, Durinck S, De Rienzo A, Modrusan Z, et al. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. Nat Genet. 2016;48:407-16.

26. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463:893-8.

27. Rajaram M, Zhang J, Wang T, Li J, Kuscu C, Qi H, et al. Two Distinct Categories of Focal Deletions in Cancer Genomes. PloS one. 2013;8:e66264.

28. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519-25.

29. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research. 2010;38:e178.

30. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011;12:323.

31. de Reynies A, Jaurand MC, Renier A, Couchy G, Hysi I, Elarouci N, et al. Molecular classification of malignant pleural mesothelioma: identification of a poor prognosis subgroup linked to the epithelial-to-mesenchymal transition. Clin Cancer Res. 2014;20:1323-34.

32. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288-95.

33. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543-50.

34. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572-3.

35. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS one. 2012;7:e41361.

36. Smyth GK. Limma: Linear Models for Microarray Data. In: Gentleman R, Carey, V., Dudoit, S., Irizarry, R., Huber, W., editor. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer; 2005. p. 397-420.

37. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499:43-9.

38. Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database : the journal of biological databases and curation. 2014;2014.

39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.

40. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010;28:511-5.

41. Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic acids research. 2016;44:e3.

42. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61-70.

43. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Statistical methods in medical research. 2013;22:519-36.

44. Mullokandov G, Baccarini A, Ruzo A, Jayaprakash AD, Tung N, Israelow B, et al. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nature methods. 2012;9:840-6.

45. Budczies J, Klauschen F, Sinn BV, Gyorffy B, Schmitt WD, Darb-Esfahani S, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PloS one. 2012;7:e51862.

46. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829-35.

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

48. Quinn L, Finn SP, Cuffe S, Gray SG. Non-coding RNA repertoires in malignant pleural mesothelioma. Lung Cancer. 2015;90:417-26.

49. Ramirez-Salazar EG, Salinas-Silva LC, Vazquez-Manriquez ME, Gayosso-Gomez LV, Negrete-Garcia MC, Ramirez-Rodriguez SL, et al. Analysis of microRNA expression signatures in malignant pleural mesothelioma, pleural inflammation, and atypical mesothelial hyperplasia reveals common predictive tumorigenesis-related targets. Experimental and molecular pathology. 2014;97:375-85.

50. Reid G. MicroRNAs in mesothelioma: from tumour suppressors and biomarkers to therapeutic targets. Journal of thoracic disease. 2015;7:1031-40.

51. Truini A, Coco S, Alama A, Genova C, Sini C, Dal Bello MG, et al. Role of microRNAs in malignant mesothelioma. Cellular and molecular life sciences : CMLS. 2014;71:2865-78.

52. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5' ends. Nature. 2017;543:199-204.

53. Mele M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Res. 2017;27:27-37.

54. Nguyen Q, Carninci P. Expression Specificity of Disease-Associated lncRNAs: Toward Personalized Medicine. Curr Top Microbiol Immunol. 2016;394:237-58.

55. Wright CM, Kirschner MB, Cheng YY, O'Byrne KJ, Gray SG, Schelch K, et al. Long non coding RNAs (lncRNAs) are dysregulated in Malignant Pleural Mesothelioma (MPM). PloS one. 2013;8:e70940.

56. Renganathan A, Kresoja-Rakic J, Echeverry N, Ziltener G, Vrugt B, Opitz I, et al. GAS5 long non-coding RNA in malignant pleural mesothelioma. Mol Cancer. 2014;13:119.

57. Fassina A, Cappellesso R, Guzzardo V, Dalla Via L, Piccolo S, Ventura L, et al. Epithelial-mesenchymal transition in malignant mesothelioma. Mod Pathol. 2012;25:86-99.

58. Abba ML, Patil N, Leupold JH, Allgayer H. MicroRNA Regulation of Epithelial to Mesenchymal Transition. Journal of clinical medicine. 2016;5.

59. Diaz-Martin J, Diaz-Lopez A, Moreno-Bueno G, Castilla MA, Rosa-Rosa JM, Cano A, et al. A core microRNA signature associated with inducers of the epithelial-to-mesenchymal transition. The Journal of pathology. 2014;232:319-29.

60. Khanbabaei H, Teimoori A, Mohammadi M. The interplay between microRNAs and Twist1 transcription factor: a systematic review. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2016;37:7007-19.

61. Dhamija S, Diederichs S. From junk to master regulators of invasion: lncRNA functions in migration, EMT and metastasis. International journal of cancer. 2016;139:269-80.

62. Ma N, Zhang W, Qiao C, Luo H, Zhang X, Liu D, et al. The Tumor Suppressive Role of MiRNA-509-5p by Targeting FOXM1 in Non-Small Cell Lung Cancer. Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology. 2016;38:1435-46.

63. Cicchini C, de Nonno V, Battistelli C, Cozzolino AM, De Santis Puzzonia M, Ciafre SA, et al. Epigenetic control of EMT/MET dynamics: HNF4alpha impacts DNMT3s through miRs-29. Biochimica et biophysica acta. 2015;1849:919-29.

64. Zhao J, Liu Y, Zhang W, Zhou Z, Wu J, Cui P, et al. Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. Cell cycle. 2015;14:3112-23.

65. Andersen M, Grauslund M, Ravn J, Sorensen JB, Andersen CB, Santoni-Rugiu E. Diagnostic potential of miR-126, miR-143, miR-145, and miR-652 in malignant pleural mesothelioma. The Journal of molecular diagnostics : JMD. 2014;16:418-30.

66. Andersen M, Trapani D, Ravn J, Sorensen JB, Andersen CB, Grauslund M, et al. Methylation-associated Silencing of microRNA-126 and its Host Gene EGFL7 in Malignant Pleural Mesothelioma. Anticancer research. 2015;35:6223-9.

67. Williams M, Kirschner MB, Cheng YY, Hanh J, Weiss J, Mugridge N, et al. miR-193a-3p is a potential tumor suppressor in malignant pleural mesothelioma. Oncotarget. 2015;6:23480-95.

68. Sui C, Meng F, Li Y, Jiang Y. miR-148b reverses cisplatin-resistance in non-small cell cancer cells via negatively regulating DNA (cytosine-5)-methyltransferase 1(DNMT1) expression. J Transl Med. 2015;13:132.

69. Chu J, Sadeghi S, Raymond A, Jackman SD, Nip KM, Mar R, et al. BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics. 2014;30:3402-4.

70. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nature methods. 2010;7:909-12.

71. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002;12:656-64.

72. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Briefings in bioinformatics. 2013;14:144-61.

73. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, 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. 2006;5:2512-21.

74. Liang J, Shao SH, Xu ZX, Hennessy B, Ding Z, Larrea M, et al. The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. Nature cell biology. 2007;9:218-24.

75. Hu J, He X, Baggerly KA, Coombes KR, Hennessy BT, Mills GB. Non-parametric quantification of protein lysate arrays. Bioinformatics. 2007;23:1986-94.

76. Hennessy BT, Lu Y, Poradosu E, Yu Q, Yu S, Hall H, et al. Pharmacodynamic markers of perifosine efficacy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2007;13:7421-31.

77. Coombes K, Neely, S., Joy, C. et al. SuperCurve Package. R Package Version 1.4.1. 1.4.1 ed. Houston, Texas: MD Anderson Cancer Center; 2011.

78. Gonzalez-Angulo AM, Hennessy BT, Meric-Bernstam F, Sahin A, Liu W, Ju Z, et al. Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clinical proteomics. 2011;8:11.

79. Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, 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. Clinical proteomics. 2010;6:129-51.

80. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nature communications. 2014;5:3887.

81. Bajou K, Noel A, Gerard RD, Masson V, Brunner N, Holst-Hansen C, et al. Absence of host plasminogen activator inhibitor 1 prevents cancer invasion and vascularization. Nat Med. 1998;4:923-8.

82. Takayama Y, Hattori N, Hamada H, Masuda T, Omori K, Akita S, et al. Inhibition of PAI-1 Limits Tumor Angiogenesis Regardless of Angiogenic Stimuli in Malignant Pleural Mesothelioma. Cancer Res. 2016;76:3285-94.

83. Hayashi M, Madokoro H, Yamada K, Nishida H, Morimoto C, Sakamoto M, et al. A humanized anti-CD26 monoclonal antibody inhibits cell growth of malignant mesothelioma via retarded G2/M cell cycle transition. Cancer Cell Int. 2016;16:35.

84. Liu R, Proud CG. Eukaryotic elongation factor 2 kinase as a drug target in cancer, and in cardiovascular and neurodegenerative diseases. Acta Pharmacol Sin. 2016;37:285-94.

85. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res. 2016;22:609-20.

86. Chen L, Gibbons DL, Goswami S, Cortez MA, Ahn YH, Byers LA, et al. Metastasis is regulated via microRNA-200/ZEB1 axis control of tumour cell PD-L1 expression and intratumoral immunosuppression. Nature communications. 2014;5:5241.

87. Lines JL, Pantazi E, Mak J, Sempere LF, Wang L, O'Connell S, et al. VISTA is an immune checkpoint molecule for human T cells. Cancer research. 2014;74:1924-32.

88. Le Mercier I, Chen W, Lines JL, Day M, Li J, Sergent P, et al. VISTA Regulates the Development of Protective Antitumor Immunity. Cancer research. 2014;74:1933-44.

89. Liu J, Yuan Y, Chen W, Putra J, Suriawinata AA, Schenk AD, et al. Immune-checkpoint proteins VISTA and PD-1 nonredundantly regulate murine T-cell responses. Proc Natl Acad Sci U S A. 2015;112:6682-7.

90. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576-82.

91. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906-12.

92. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences of the United States of America. 2013;110:4245-50.

93. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.

94. Lopez-Rios F, Chuai S, Flores R, Shimizu S, Ohno T, Wakahara K, et al. Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction. Cancer Res. 2006;66:2970-9.

95. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237-45.

96. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498-504.

97. Wong CK, Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, et al. The UCSC Interaction Browser: multidimensional data views in pathway context. Nucleic Acids Res. 2013;41:W218-24.

## Supplementary Figures and Figure Legends

**Supplementary Figure S1**: **Pathology Review. A**: Reclassification based on pathology review. **B-G**: Histologic patterns of MPM: Epithelioid with tubules of uniform tumor cells showing abundant eosinophilic cytoplasm (**B**); epithelioid consisting of solid sheets of epithelioid cells (**C**); biphasic with bundles of spindle-shaped tumor cells surrounded by nests of epithelioid tumor cells (**D**); biphasic with malignant spindled tumor cells (left) and epithelioid cells (right) (**E**); sarcomatoid with cytologically malignant spindle cells showing hyperchromatic nuclei (**F**); sarcomatoid with malignant spindle cells infiltrating among fat cells (**G**).

**Supplementary Figure S2: Copy Number Analyses. A:** Unsupervised hierarchical clustering of DNA copy number profiles of Mesotheliomas into six SCNA cluster groups. In the heatmap, SCNAs in tumors (x axis) are plotted by genomic location (y axis). Red bars designate regions of amplification, while blue bars designate deleted regions. **B:** The heatmap shows SCNAs in tumors grouped by either those with biallelic inactivation of *BAP1* or those that have two active wild type *BAP1* alleles. **C**: CDKN2A vs. MTPA mRNA expression. Co-deleted samples tend to exhibit lower levels of expression in both genes. **D:** Number of inactivated *BAP1* alleles in copy number clusters in A. For a $\mathbf{X}^2$ test across all cluster groups, p = 0.028. **E:** Allele-specific copy number analysis of TCGA-MQ-A6BR whole-exome sequencing data. Top panel shows total copy log-ratio, middle panel shows allelic log-odds-ratio revealing allelic imbalances. Red lines are joint segmentation using the FACETS algorithm. Integer copy number estimates (total and minor) along with cellular fraction (cf-em) estimates are shown in the bottom panel. **F:** Allele-specific copy number analysis of TCGA-SC-A6LP whole-exome sequencing data. **G:** Allele-specific copy number analysis of TCGA-UD-AAC1 whole-exome sequencing data.

**Supplementary Figure S3: Gene Expression Analyses.** Distinct expression subtypes of MPM associated with histology, clinical, and genomic features. **A**:

Unsupervised gene expression clustering defined four distinct mRNA-based subtypes of MPM. A predictive gene list of 1,800 genes are displayed in the heatmap with selected genes/pathways that are expressed in each group listed. Annotation tracks are displayed for the four expression clusters (K1-K4) and histology classification. K3 is enriched in non-epithelioid histologies. **B**: Selected genomic features significantly associated with expression subtypes. BAP1 alterations annotation track includes single nucleotide variants (SNV), small and large insertions and deletions (INDEL), and structural variants (SV), loss of heterozygosity (LOH) events, and copy number (CN) loss events (blue). CN, GISTIC scores – pink, gain; light blue, loss; dark blue, deep loss. **C**: Kaplan-Meier survival curves of probability of overall survival with overall significance testing by log rank test and cox proportional hazards analysis using the good outcome group K4 as the reference. **D**: Patient age of diagnosis by expression subtype. Median and mean ages per subtype displayed below plot. Asterisk denotes a pvalue of 0.0008 for a t-test comparing ages of K3 vs all other subtypes. **E**: Normalized *BAP1* expression levels across subtypes. *BAP1* alterations are indicated by a red box and wildtype *BAP1* by black circles. **F and G**: Using gene sets from two previously published datasets: Bueno et al. (**F**) and de Reynies et al. (**G**), we observe samples clustering similar to our classification. **H**: Gene expression significantly associated with *BAP1* inactivation.

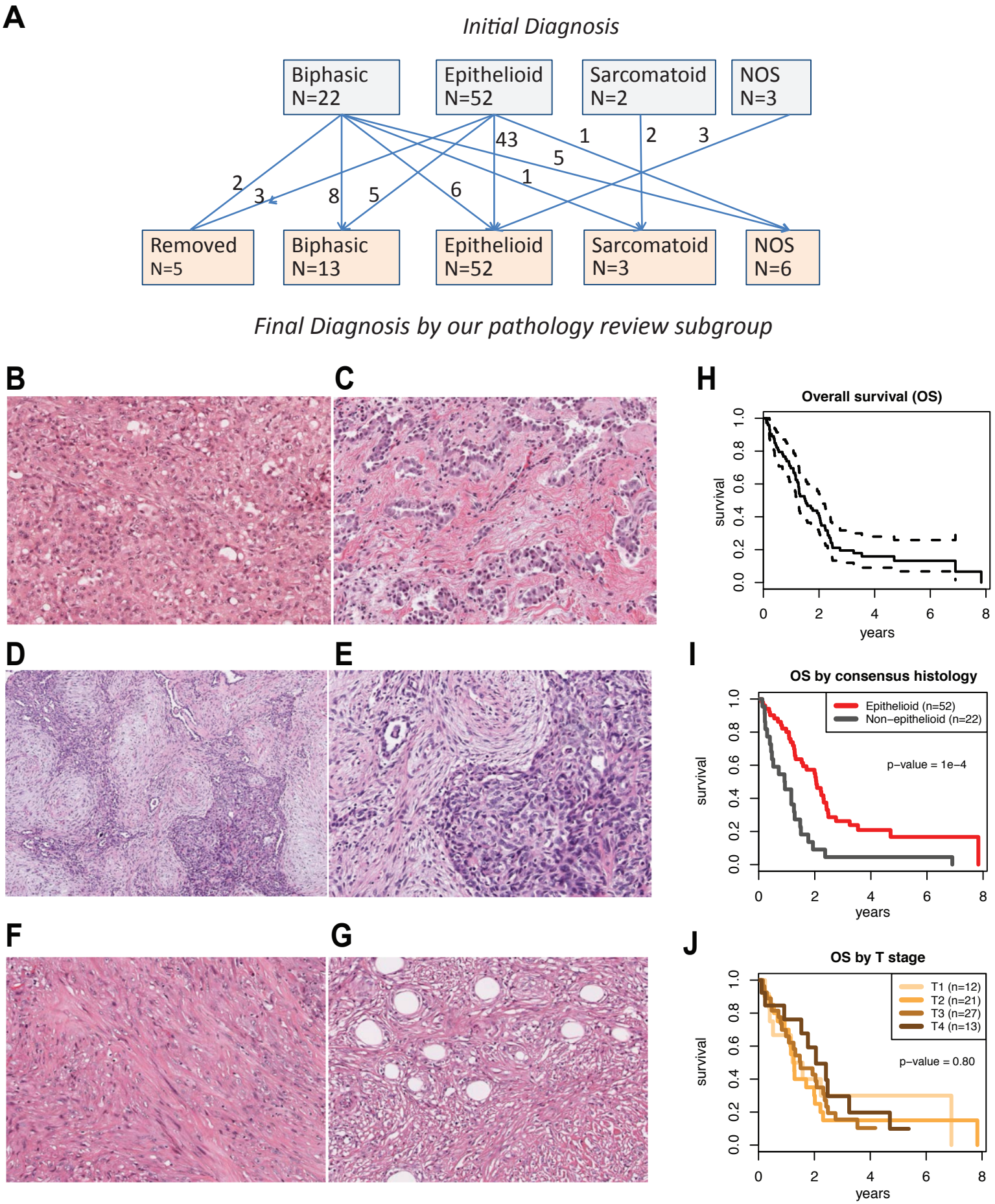
**Supplementary Figure S4: DNA Methylation. A**: Heatmap showing beta values of 74 samples ordered by unsupervised clusters of DNA methylation data. Samples are presented in columns and the CpG loci are presented in rows. An annotation panel on the right of the heatmap indicates location of CpG locus relative to CpG island status and gene. Annotation bars at the top of the heatmap show genomic variables associated with the methylation clusters. Features marked with (*) are significantly associated with DNA methylation clusters (Fisher's Exact test $p<0.01$). **B**: Box plots of the leukocyte fraction and tumor purity by DNA methylation clusters. **C**: Heatmap showing differentially methylated probes between *BAP1* altered and WT samples. An annotation panel on the right

of the heatmap indicates location of CpG locus relative to CpG island status and gene. An annotation bar at the top corresponds to the number of *BAP1* alterations in a sample.

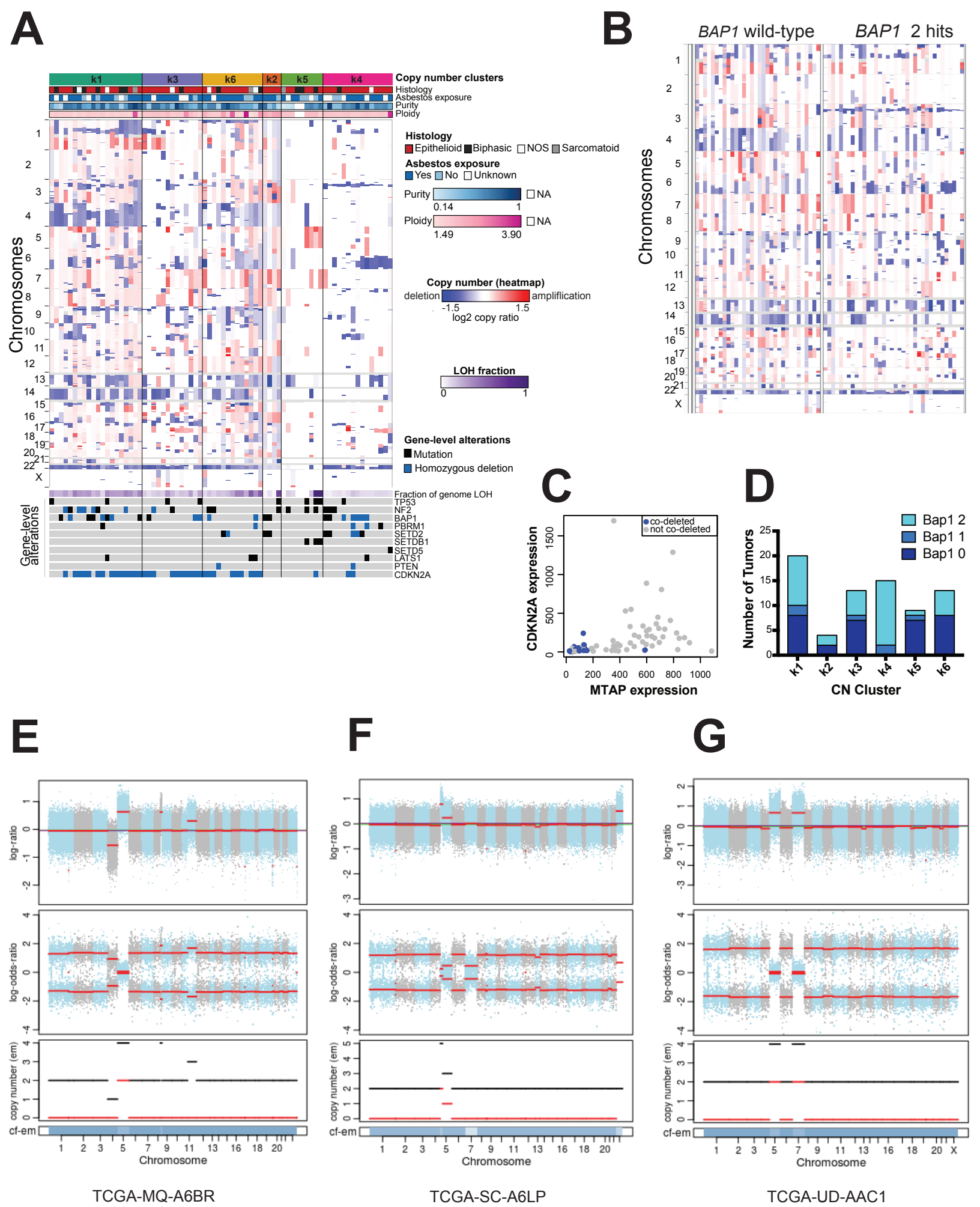
**Supplementary Figure S5: RPPA analysis and mRNA immune signatures**. **A**: Consensus matrix (k=5) for clustering 52 MPMs, based on 1-Pearson correlation as distance metric and Wards' linkage algorithm. 80% samples used in each of the 1000 iterations. **B**: Heatmap of consensus clusters from RPPA Data. The columns and rows represent the 52 samples and 219 cancer-associated antibodies respectively. Five clusters identified with significant differential protein expression and pathway activity. The annotation bars shown above the heatmap were not used for clustering. **C**: Box plots of pathway scores for each of the five RPPA clusters. P-values based on the ANOVA test. Pathways differentially expressed between the clusters include Apoptosis, Cell cycle, EMT, Hormone receptor, RAS/MAPK, TSC/mTOR, Breast reactive and Core reactive pathways. **D**: Kaplan-Meier survival curves for the five RPPA sample clusters. P-values based on the G-rho family of Harrington and Fleming [13] tests to evaluate the difference between two or more survival curves. The proteins differentially expressed between good prognosis cluster C5 and the poor prognosis cluster C4 were PAI1, CYCLINB1, CAVEOLIN1, EPPK1 and EEF2K as shown in panel B. **E**: GSEA-based mRNA immune signatures across iCluster clusters.

**Supplementary Figure S6: iCluster validation analyses. A**: Correlation of Bueno cohort (n=211, all histologies) to iCluster centroids. **B**: Kaplan-Meier plot of survival analysis by Bueno iClusters across all histologies. **C**: Multivariate Cox regression of OS by Bueno iClusters adjusted by histology and age. **D**: Silhouette plot of Lopez-Rios cohort (n=52) iClusters. **E**: Breakdown of histology in Lopez-Rios iClusters. **F**: Kaplan-Meier plot of survival analysis by Lopez-Rios iClusters (n=49). **G**: Multivariate Cox regression of OS by Lopez-Rios iClusters adjusted by histology, age. **H:** Multivariate Cox regression of OS by iCluster in our cohort.

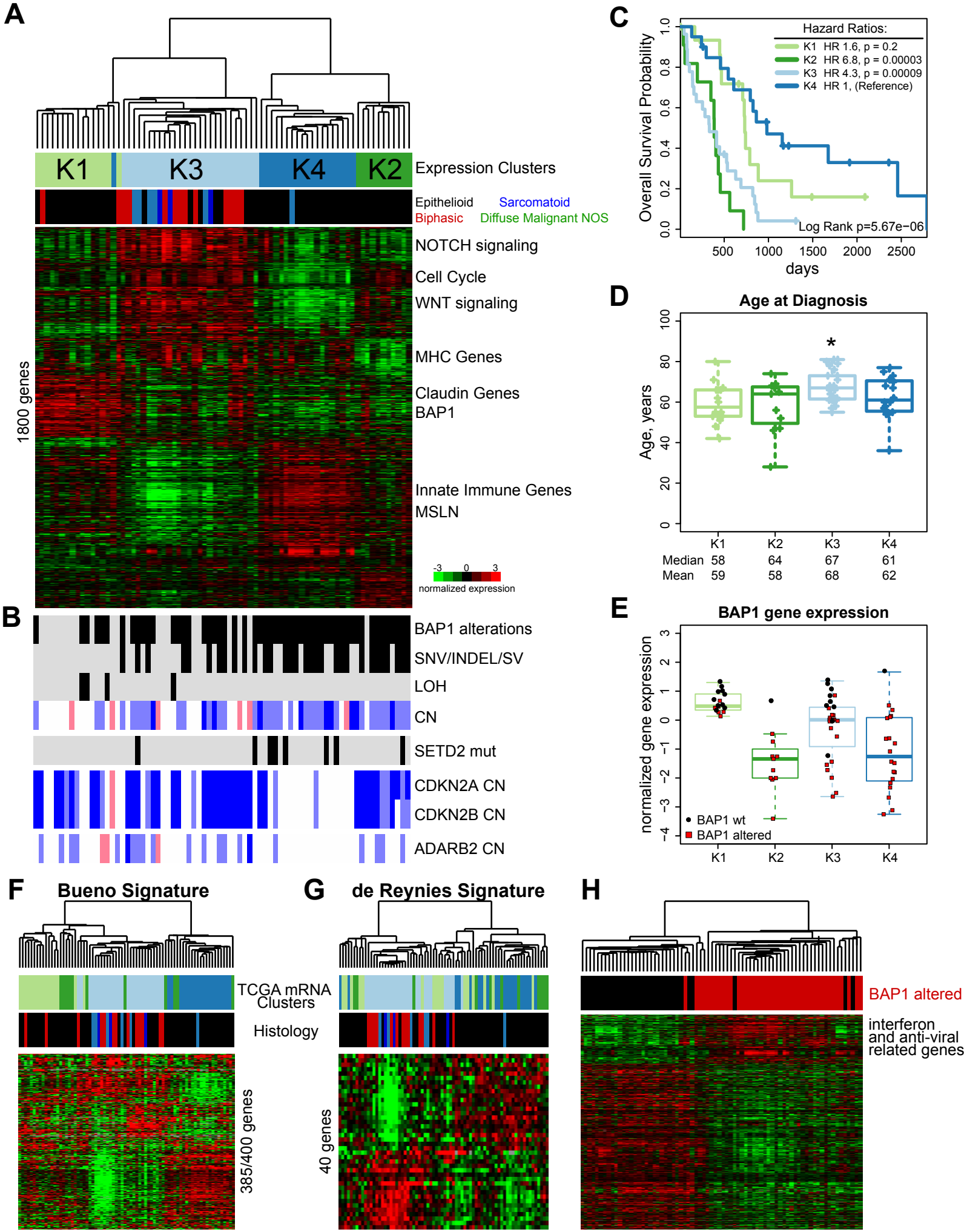
**Supplementary Figure S7: Integrated mRNA and SCNA PARADIGM analysis.** **A:** Top heatmap shows molecular subtypes derived through unsupervised analysis of PARADIGM results. Major pathway groups characterizing these subtypes are annotated to the right of the heatmap. Middle and bottom heatmaps show mRNA expression and SCNA within PARADIGM-derived subtypes. **B**: Survival analysis of the PARADIGM-derived molecular subtypes. Kaplan–Meier plot shows significant difference in survival of the best-surviving and the worst-surviving clusters. **C**: Differential activity pathway for *BAP1* wild-type vs. inactivated samples based on PARADIGM results. Red nodes indicate genes that are active in *BAP1* wild-type MPM. Blue nodes indicate genes with low activity in the same samples. The size of the nodes is inversely proportional to the differential p-value (larger nodes indicate higher significance). **D**: Top differentially active pathways between samples with *BAP1* inactivation vs. wild-type. PARADIGM IPLs for the genes in each pathway are aggregated per-patient and ANOVA test is used to compute significance in distributions of aggregated IPLs between the two groups (*BAP1* wild-type and inactivated) of samples.



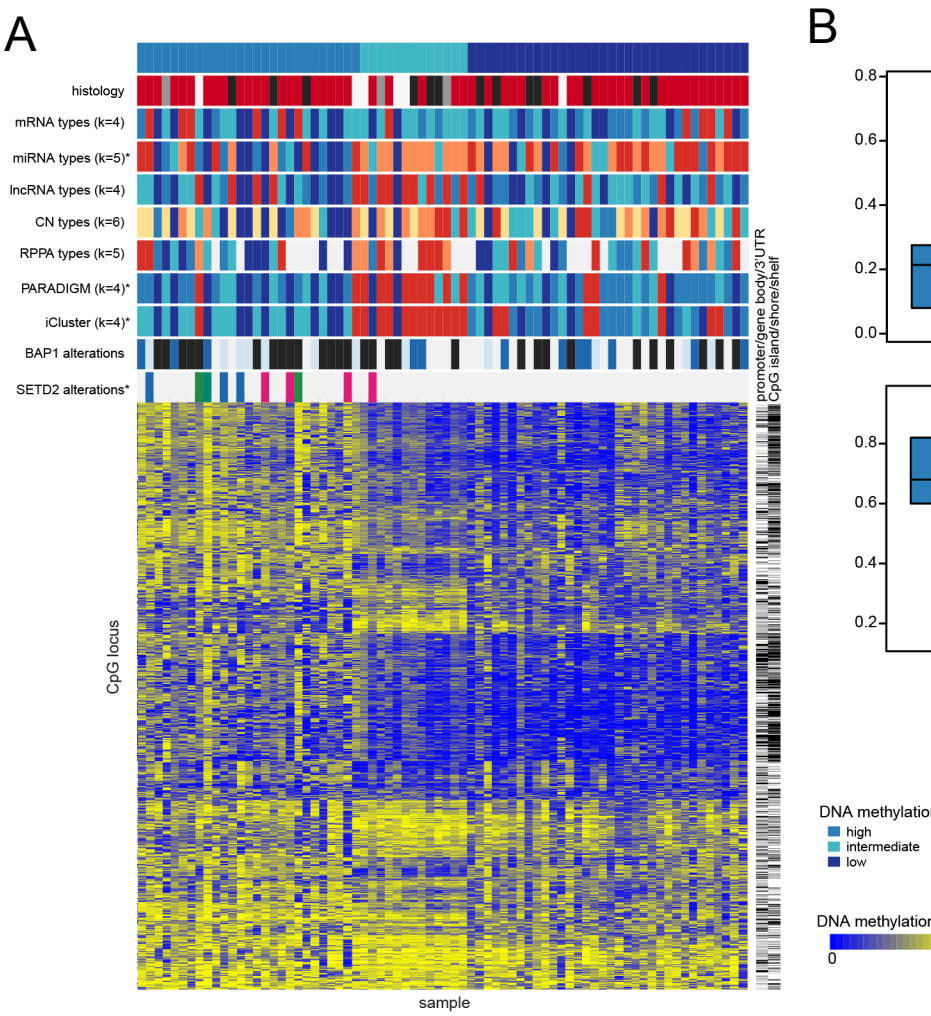

Supplementary Figure 1

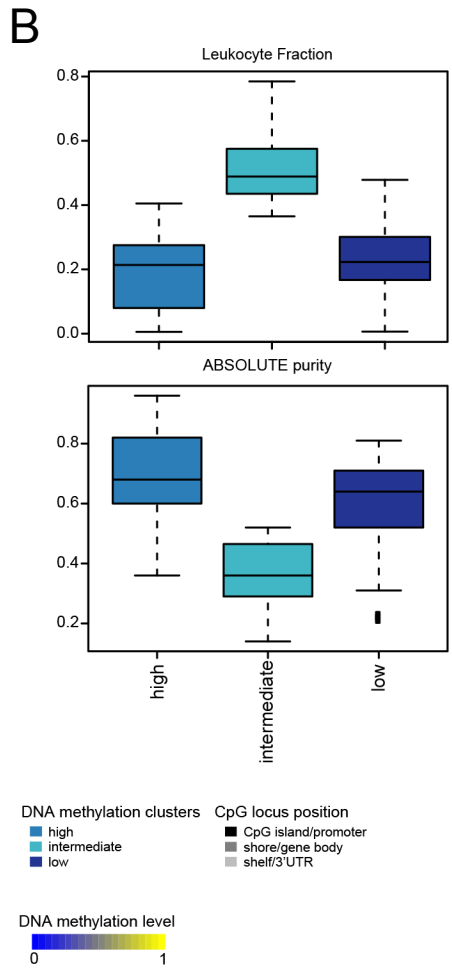


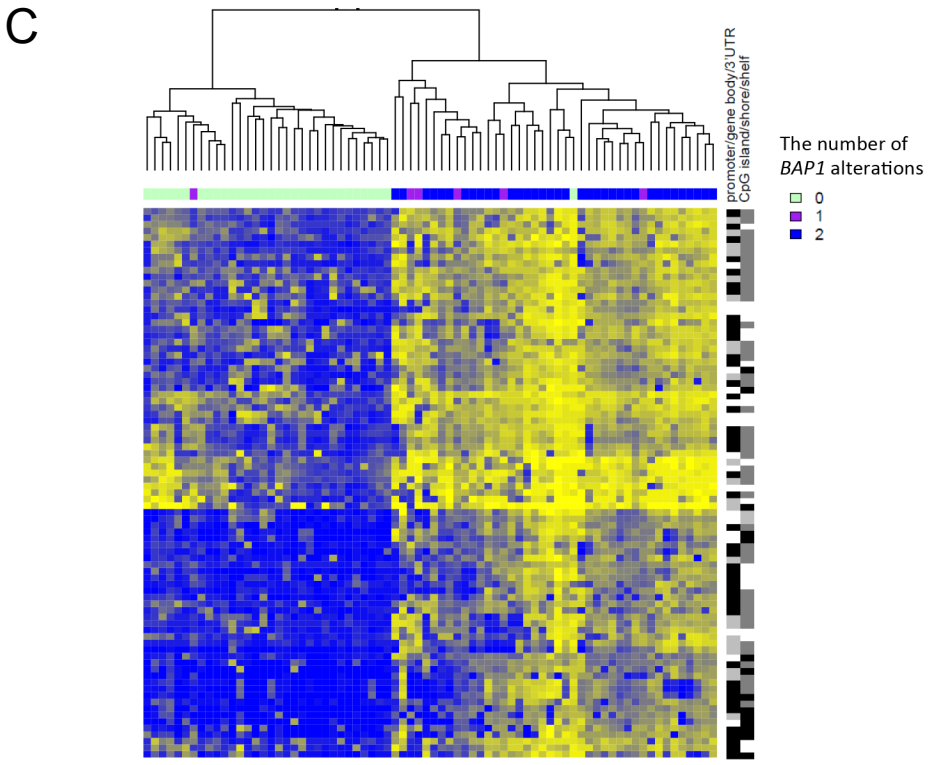

Supplementary Figure 2



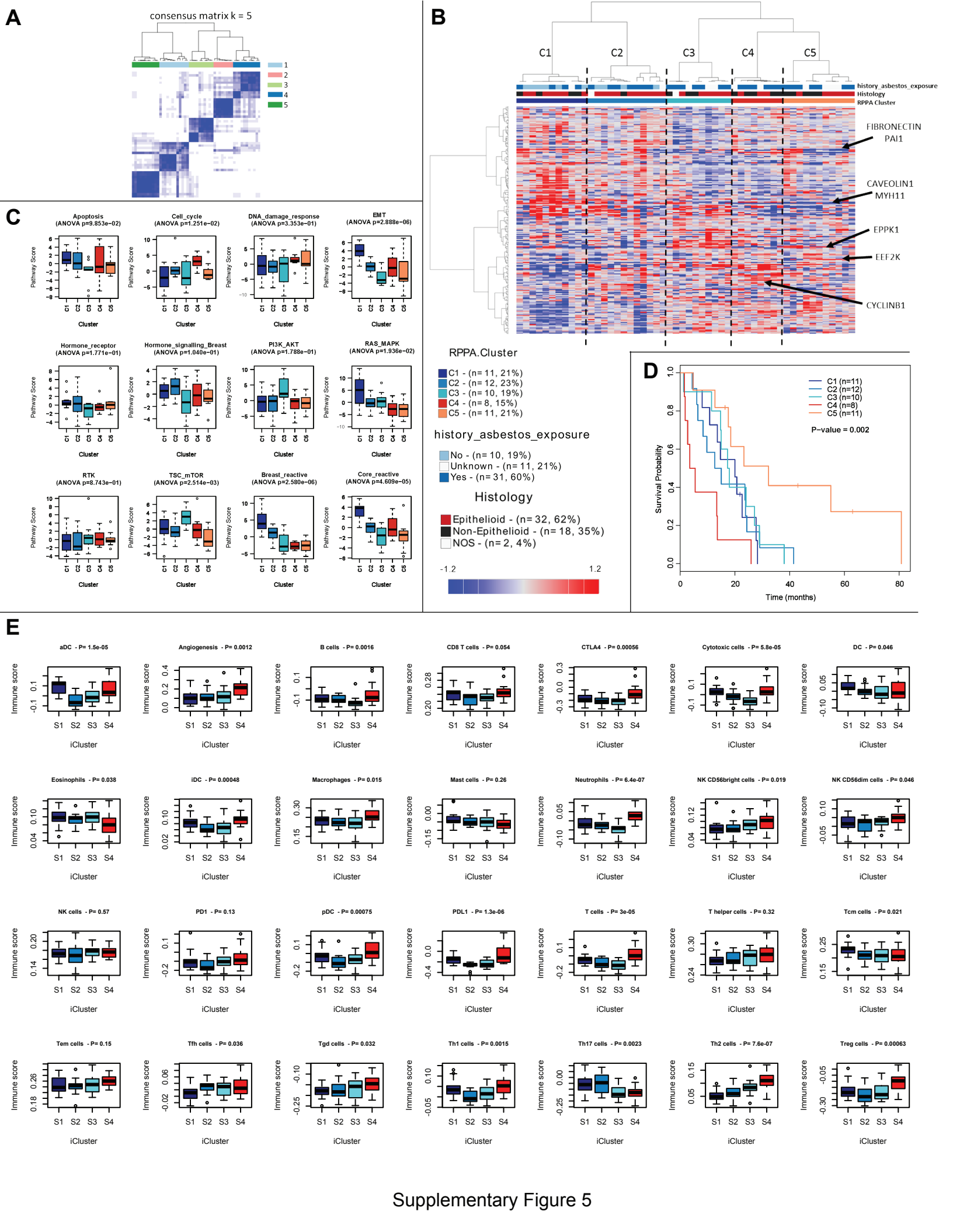

Supplementary Figure 3



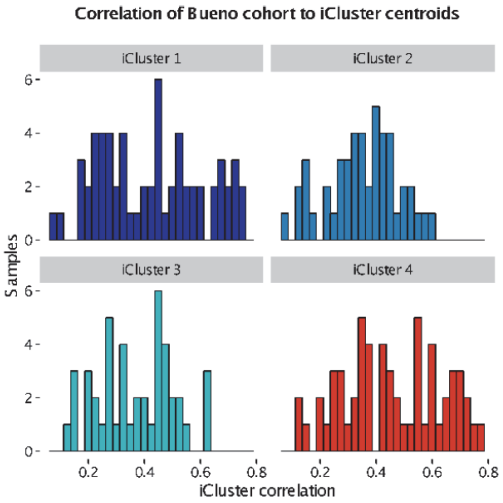




Supplementary Figure 4

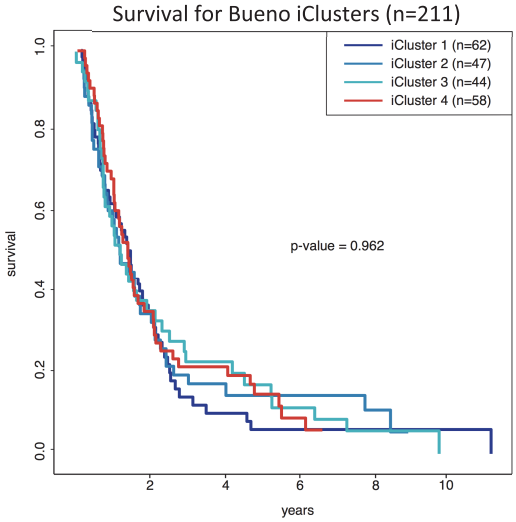



Supplementary Figure 5

**A**




**B**




**C**

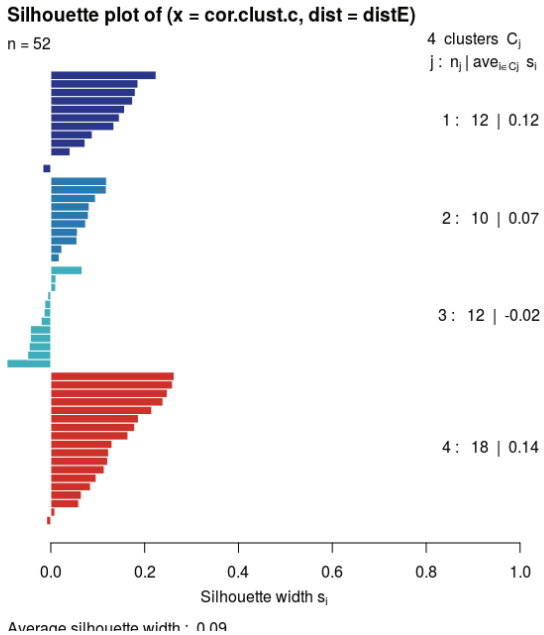
Cox Regression with iCluster, Histology and Age (n=211)

| Characteristic | HR (95% CI) | P |
|---|---|---|
| iCluster group (ref iCluster 1) | | |
| iCluster2 | 1.06 (0.70, 1.60) | |
| iCluster3 | 0.94 (0.62, 1.43) | 0.93 |
| iCluster4 | 0.94 (0.63, 1.38) | |
| Histology (ref Epithelioid) | | |
| Non-Epithelioid | 1.45 (1.06, 2.00) | **0.023** |
| Age (continuous) | 1.03 (1.01, 1.04) | **2.92e-4** |

Cox Regression with iCluster and Age restricted to epithelioid histology (n=141)

| Characteristic | HR (95% CI) | P |
|---|---|---|
| iCluster group (ref iCluster 1) | | |
| iCluster2 | 2.36 (1.40, 4.00) | |
| iCluster3 | 2.75 (1.64, 4.61) | **1.24e-4** |
| iCluster4 | 2.43 (1.47, 4.04) | |
| Age (continuous) | 1.03 (1.02, 1.05) | **8.98e-5** |

**D**

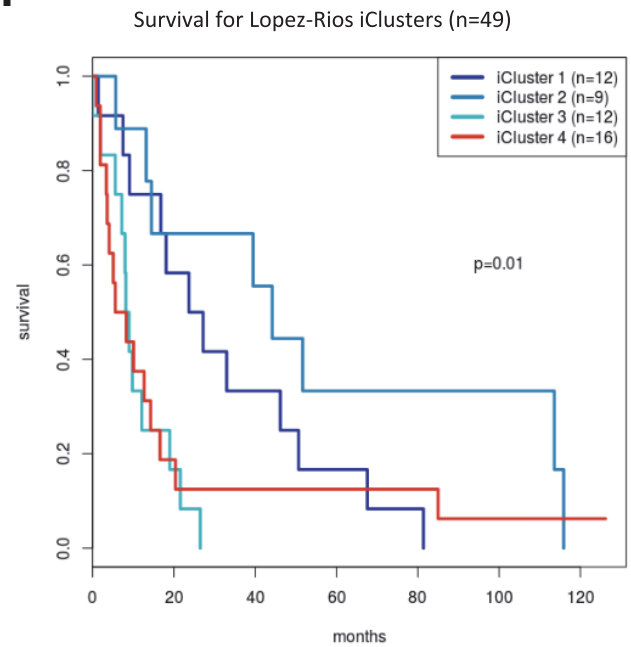



**E**

Histology breakdown for Lopez-Rios iClusters (n=52)

| | iCluster 1 n=12 | iCluster 2 n=9 | iCluster 3 n=12 | iCluster 4 n=18 | Total n=52 | P-value |
|---|---|---|---|---|---|---|
| Histology | | | | | | 0.007 |
| Biphasic | 1 | 0 | 4 | 5 | 10 | |
| Epithelioid | 11 | 10 | 8 | 8 | 37 | |
| Sarcomatoid | 0 | 0 | 0 | 5 | 5 | |

**F**




**G**

Cox Regression with iCluster, Age and Histology (n=49)

| Characteristic | HR (95% CI) | P |
|---|---|---|
| iCluster group (ref iCluster 1) | | |
| iCluster2 | 0.74 (0.28, 1.95) | |
| iCluster3 | 2.38 (0.99, 5.69) | 0.10 |
| iCluster4 | 1.53 (0.67, 3.49) | |
| Age (continuous) | 1.01 (0.98, 1.05) | 0.39 |
| Histology (ref Epithelioid) | | |
| Non-Epithelioid | 2.52 (1.14, 5.59) | 0.03 |

**H**

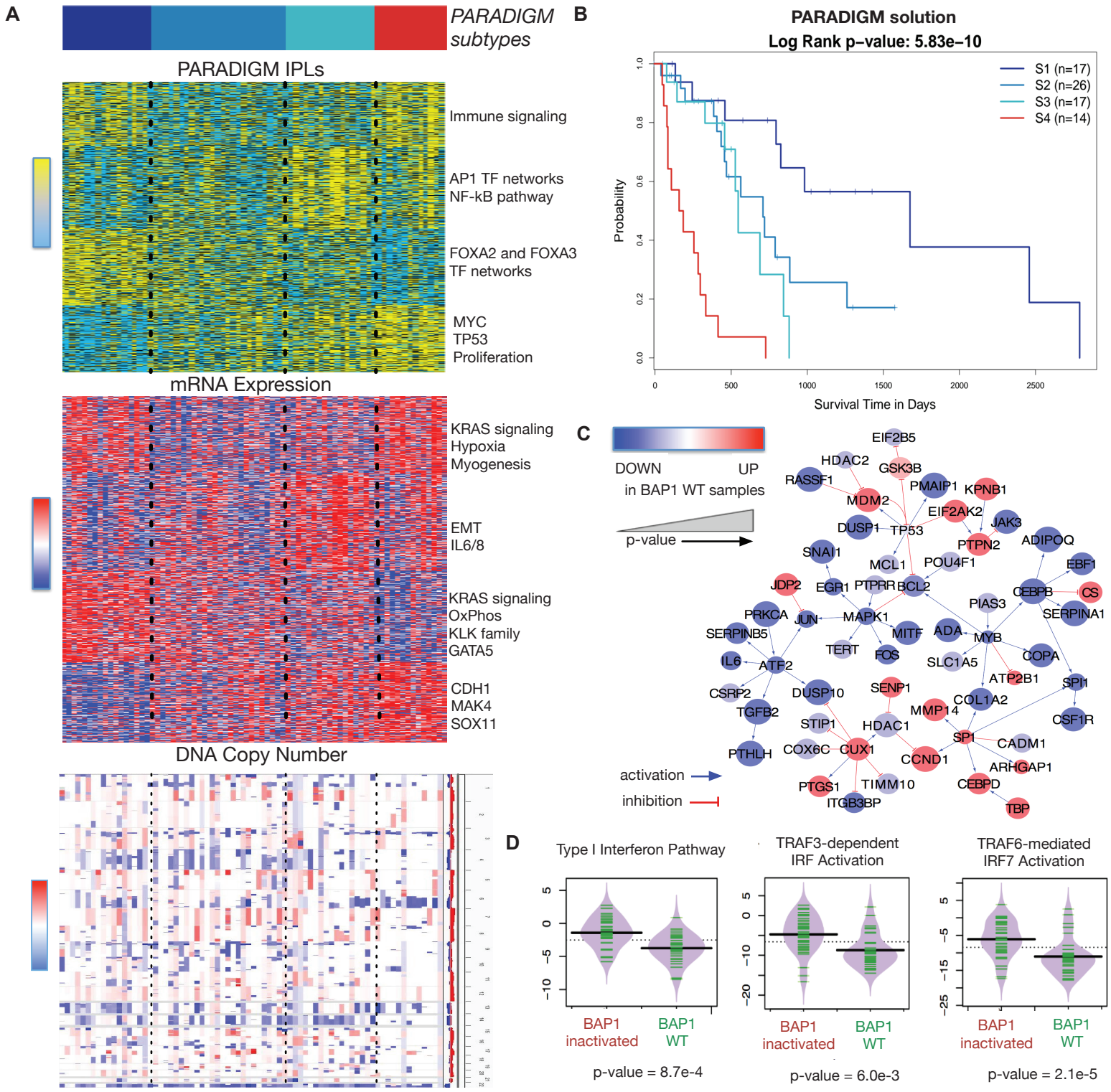
**Cox Regression with iCluster, Histology and Age (n=74)**

| Characteristic | HR (95% CI) | P |
|---|---|---|
| iCluster group (ref iCluster 1) | | |
| iCluster 4 | 5.71 (2.49-13.10) | |
| iCluster 3 | 1.49 (0.63-3.52) | **4.6E-04** |
| iCluster 2 | 2.50 (1.14-5.51) | |
| Histology (ref Epithelioid) | | |
| Non-Epithelioid | 2.15 (1.17-3.94) | **0.02** |
| Age (continuous) | 0.98 (0.95-1.02) | 0.37 |

**Cox Regression with Paradigm Clusters, Histology and Age (n=74)**

| Characteristic | HR (95% CI) | P |
|---|---|---|
| Paradigm group (ref cluster 1) | | |
| Paradigm Cluster 4 | 23.63 (7.57-73.80) | |
| Paradigm Cluster 3 | 3.58 (1.48-8.68) | **6.4E-07** |
| Paradigm Cluster 2 | 2.76 (1.23-6.17) | |
| Histology (ref Epithelioid) | | |
| Non-Epithelioid | 1.60 (0.78-03.26) | 0.20 |
| Age (continuous) | 0.99 (0.96-1.03) | 0.70 |

Supplementary Figure S6




Supplementary Figure 7