doi:10.1038/nature12113

Integrated Genomic Characterization of Endometrial Carcinoma
Supplementary Materials

Supplementary Methods S1: Biospecimen collection and clinical data:
- Table S1.1 Tumor stage, histology and grade for 370 patients, n (%)*.
- Table S1.2 Specimen and assay summary
- Table S1.3 Microsatellite markers
- Data File S1.1 Key Clinical Data

Supplementary Methods S2: Copy number analysis:
- Figure S2.1 GISTIC 2.0 analysis for each copy number cluster
- Data File S2.1 GISTIC amplification and deletion peak annotations

Supplementary Methods S3: DNA sequencing - exome and genome:
- Table S3.1 Genes involved in recurrent translocations
- Figure S3.1 Non-silent mutation matrix for 11 SMGs
- Figure S3.2 Recurrent translocations involving the BCL family
- Figure S3.3 A complex rearrangement involving FNDC3B and MYC genes
- Date File S3.1 SMG lists in 248 endometrial cases
- Date File S3.2 SMG differences between mutation spectra cohorts

Supplementary Methods S4: RNA sequencing:
- Table S4.1 Putative fusion candidates
- Figure S4.1 Silhouette widths
- Figure S4.2 Most significantly enriched pathways in different subtypes
- Figure S4.3 Hormone receptor RNA and protein expression
- Figure S4.4 Immune cell infiltrates across subtypes

Supplementary Methods S5: Reverse Phase Protein Arrays
- Table S5.1 Correlations of RPPA clusters
- Figure S5.1 Supervised hierarchical clustering
- Figure S5.2 Unsupervised hierarchical clustering
- Data File S5.1 RPPA Antibody List

Supplementary Methods S6: miRNA sequencing:
- Figure S6.1 NMF consensus clustering
- Figure S6.2 Discriminatory miRNA to discriminate tumor groups
- Figure S6.3 Relationships of platform, purity and batch

Supplementary Methods S7: DNA methylation:
- Figure S7.1 Unsupervised clustering

Supplementary Methods S8: Integrative clustering
- Figure S8.1 Integrative clustering

Supplementary Methods S9: Super Clusters
- Figure S9.1 Super Clusters

Supplementary Methods S10: Batch effect analysis

- Figure S10.1 Hierarchical clustering for mRNA Sequencing
- Figure S10.2 Principal components analysis (PCA) for mRNA according to batch
- Figure S10.3 PCA for mRNA according to tissue source site (TSS)
- Figure S10.4 Hierarchical clustering for miRNA Sequencing
- Figure S10.5 PCA for miRNA according to batch
- Figure S10.6 PCA for miRNA according to TSS
- Figure S10.7 Hierarchical clustering for DNA methylation
- Figure S10.8 PCA for DNA methylation according to batch
- Figure S10.9 PCA for DNA methylation according to TSS
- Figure S10.10 Hierarchical clustering for SNP data
- Figure S10.11 PCA for SNP data according to batch
- Figure S10.12 PCA for SNP data according to TSS

Supplementary Methods S11: PARADIGM analyses

- Figure S11.1 Top 1000 varying pathway features within PARADIGM clusters
- Figure S11.2 DIPSC analysis between endometrial subtypes
- Figure S11.3 Differentially activated pathways between endometrial subtypes
- Figure S11.4 PARADIGM-SHIFT analysis of p53 mutations
- Figure S11.5 Machine learning classifiers between endometrial, breast and ovarian cancers
- Figure S11.6 Comparison of endometrial, breast and ovarian
- Figure S11.7 Machine learning classifiers between hormonal endometrial and luminal breast cancers
- Figure S11.8 Differentially activated pathways in endometrial subtypes
- Figure S11.9 Comparison of hormonal endometrial with luminal breast cancers

Supplementary Methods S12: Cross-tumor comparison

- Figure S12.1 Dendrogram showing relatedness of SCNAs across tumor types
- Figure S12.2 Supervised analysis of transcriptomic datasets
- Figure S12.3 Methylation profiles across tumor types

Additional supporting material can be found at the TCGA Data Portal page for this study at: https://tcga-data.nci.nih.gov/docs/publications/ucec_2013/

Supplementary Methods S1: Biospecimen collection and clinical data

Sample inclusion criteria and pathology review

Biospecimens were collected at diagnosis from patients with endometrioid adenocarcinomas and serous carcinomas 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 200 Grade 1/2, 200 Grade 3, and 100 serous cancer subtypes. 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 which 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. Normal endometrium from 11 patients without a history of cancer was included in this study. Each tumor specimen was shipped overnight from one of 17 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. Each H&E stained section of frozen OCT-embedded tumor processed centrally by TCGA was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically generally consistent with the diagnosis and the adjacent normal specimen (when provided) contained 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 *mir*Vana 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 either the QiaAmp blood midi kit (Qiagen).

Each specimen was quantified by measuring Abs_{260} 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 was sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only those specimens yielding a minimum of 6.9 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, 837 endometrioid adenocarcinomas and serous carcinomas cases were received by the BCR and 65% passed pathology and molecular quality control. The

biospecimens included in this report come from 373 endometrioid adenocarcinomas and serous carcinomas cases included in batches 49, 59, 73, 75, 81, 92, 94, 104, 110, 118, 121, 125, 137, 143, 156.

Microsatellite Instability Testing

Microsatellite instability (MSI) status of endometrioid adenocarcinomas and serous carcinomas 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.[1] 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.

Clinical data analyses

Clinical data included in this report were downloaded from the TCGA Data Portal on May 13, 2012. 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 using various staging systems. If the FIGO 2009 staging system for endometrial cancer was not specified by the TSS, the FIGO stage was recalculated into the 2009 staging system directly from submitted pathology reports into major substage divisions (stage I, II, III, and IV). Body mass index (BMI) was calculated using the following formula: BMI = weight (kg) / [height (m) * height (m)]. All serous and mixed cases were designated as grade 3 based on customary practices. 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 S1.1 and Supplementary data file S1.1. The 373 patients had a median age of 63 years (range, 31-90 years). Of the 88 patients who have received adjuvant chemotherapy, 86 (98%) received a platinum-containing regimen, most commonly a platinum/taxane doublet (90%).

Table S1.1 Tumor stage, histology, grade, and adjuvant treatment for 370 patients, n (%)*.

Stage	Endometrioid Grade 1	Endometrioid Grade 2	Endometrioid Grade 3	Mixed Grade 3[#]	Serous Grade 3	Total
I	78 (89)	83 (79)	70 (63)	6 (46)	17 (32)	254 (69)
II	3 (3)	9 (9)	6 (5)	2 (15)	5 (9)	25 (7)
III	7 (8)	12 (11)	26 (23)	4 (31)	25 (47)	74 (20)
IV	0	1 (1)	9 (8)	1 (8)	6 (11)	17 (5)
Adjuvant Therapy						
RT	12 (14)	28 (27)	22 (20)	1 (8)	7 (13)	70 (19)
Chemo	2 (2)	6 (6)	14 (13)	3 (23)	13 (25)	38 (10)
ChemoRT	2 (2)	9 (9)	18 (16)	4 (31)	17 (32)	50 (14)
Unknown	70 (80)	61 (58)	57 (51)	5 (39)	16 (30)	209 (57)
Total	88 (100)	105 (100)	111 (100)	13 (100)	53 (100)	370 (100)

*3 patients have missing data

[#]Mixed serous and endometrioid

RT = radiation therapy, Chemo = cytotoxic chemotherapy, ChemoRT = cytotoxic chemotherapy and radiation therapy, Unknown = data not provided

Table S1.2 Specimen and assay summary.

Assay	Number of endometrial patient specimens
Exome sequencing	248 pairs
Whole genome sequencing	107 pairs
RNA sequencing	333
miRNA sequencing	367
DNA methylation (Infinium HM450)	256
DNA methylation (Infinium HM27)	117
DNA copy number (Affymetrix SNP6.0)	363 pairs
Reverse phase protein arrays	293

Table S1.3 Microsatellite markers.

MARKER	LOCUS	GENOME DATABASE/ GENEBANK ID	PRODUCT SIZE (basepairs)
BAT25	4q11-12 KIT gene, intron 16 (T25 repeat)	GDB: 9834508 U63834	148
BAT26	2p22-21 hMSH2 gene, exon 5 (A26 repeat)	GDB: 9834505 U41210	116
BAT40	Chromosome 11, Intron 2 (within 3-β-HSD gene)	GenBank: M38180	94-112
TGFBRII	3p22	MIM: 190182 Unigen HS: 82028 Locus ID: 7048	60-80
D2S123	2p16.3 (CA repeat)	GDB: 187953 Z16551	114~174
D5S346	5q22.2 (CA repeat)	GDB: 181171 M73547	96~122
D17S250	17q12 (CA repeat)	GDB: 177030 X54562	146~165

Data File S1.3. Key Clinical Data.

datafile.S1.1.KeyClinicalData.xls

Section References

1. Umar, A. *et al*. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) & microsatellite instability. *J Natl Cancer Inst* **96**: 261-268 (2004).

Supplementary Methods S2: Copy number analysis

SNP Based Copy Number Analysis

DNA from each tumor or germline-derived sample was hybridized to the Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute.[1] From raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus.[2] For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor.[3] 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 undergo segmentation using Circular Binary Segmentation.[4] 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.[4] Analysis of broad copy number alterations was then conducted as previously described.[2] Significant focal copy number alterations were identified from segmented data using GISTIC 2.0.[5] Hierarchical clustering of copy number data was performed using R on thresholded relative copy number data in significantly reoccurring amplifications or deletions regions identified by GISTIC 2.0 analysis.[5]

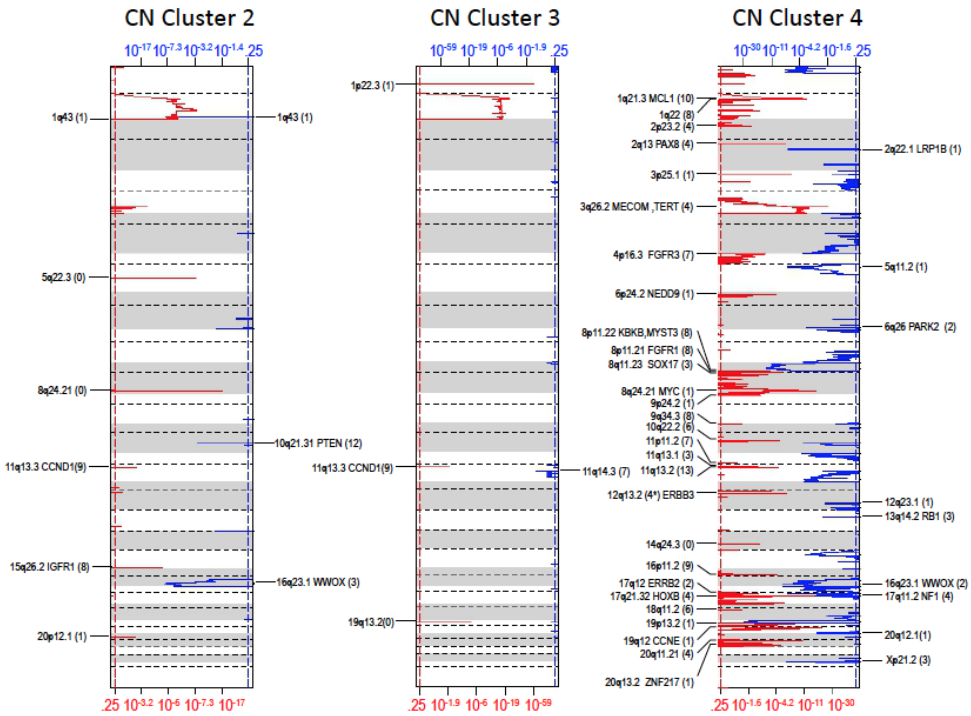


Figure S2.1. GISTIC 2.0 analysis of significantly reoccurring somatic copy number alterations from tumors in each copy number cluster. Significantly reoccurring focally amplified (red) and deleted (blue) regions are plotted along the genome by false-discovery rates. Annotations include well-localized regions with 13 or fewer genes and a false discovery Q < 0.15. Known cancer genes or genes identified by genome-wide loss-of-function screens are shown next to peaks. The number of genes included in each region is given in brackets. In the peak marked by *, the region identified by GISTIC was expanded to include an adjacent oncogene. No significantly reoccurring amplified or deleted regions were identified in analysis of copy number cluster 1. Copy number clusters 2 and 3 had focal and/or broad SCNAs, distinguished primarily by more frequent 1q amplification in cluster 3 than cluster 2. Additional recurring SCNAs in clusters 2 and 3 included focal 11q13.3 amplification, which encompassed CCND1. Focal changes in cluster 4 contained most serous and 'serous-like' tumors which were similar to those in serous tumors alone.

Data File S2.1: GISTIC amplification and deletion peak annotations.

Data file listing individual genes within GISTIC amplification and deletion peaks with chromosomal location, peak boundaries, and false discovery rates.

datafile.S2.1.UcecGisticPeaks.xls

Section References

1. McCarroll, S.A. *et al*. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**:1166-1174 (2008).
2. Korn, J.M. *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**:1253-1260 (2008).
3. The Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**:609-615 (2011).
4. Olshen, A.B. *et al*. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**:557-572 (2004).
5. Mermel, C.H. *et al*. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**:R41 (2011).

Supplementary Methods S3: DNA sequencing – exome and genome

Supplementary Methods for Exome Sequencing

The exomes of 248 tumor and normal pairs were targeted using Agilent SureSelect v2 or Nimblegen SeqCap v2, and sequenced on Illumina GAIIx or HiSeq 2000 platforms. An average of 86.5% of targeted base pairs received a read-depth of at least 20x. Somatic single nucleotide variants (SNVs) were called using Samtools,[1] VarScan 2,[2] and SomaticSniper,[3] while small indels were called using GATK,[4] VarScan 2,[2] and Pindel.[5] After combining these SNVs and indels across callers, they were filtered for potential false-positives based on criteria described by Koboldt et al.,[2] and a few additional stringent filters, including one that removes short indels near homopolymers. Most short indels near homopolymers belonged to cases with microsatellite instability (MSI), but they could not be confidently distinguished from commonly seen sequencing artifacts at homopolymers. The final list of somatic mutations in endometrial cancer was studied using the MuSiC suite of tools.[6]

Somatic variants across 222 non-ultramutated cases were targeted for resequencing with new cDNA libraries. In the final curated list of somatic variants, 98.7% of SNVs and 81.7% of indels had at least 1 supporting read in the tumor, and <3% supporting reads in the normal. Variants with <8% supporting reads in the tumor and <3% in the normal, were not considered validated.

Germline variants were obtained from the results of Samtools, VarScan 2, GATK, and Pindel. They were then shortlisted by filtering out variants that fit these criteria:

1. Annotated to transcripts that are unvalidated, provisional, of have reported errors
2. Annotated to known problematic genes like olfactory receptors
3. Annotated to non-coding RNA genes, or other non-coding loci
4. Near the 3' end of a transcript, within 5% its length, unless in a protein domain
5. Seen in more than 2% of the cases in the cohort (248 cases)
6. Minor allele frequency of more than 1% in either 1000 genomes or NHLBI

Somatic mutations identified from whole exome sequencing

A total of 184,824 somatic mutations comprising 181,930 point mutations and 2,931 indels (ranging from 1 to 82 bps) were identified in the targeted exons and splice junctions. Roughly 80% of these events (146,814 point mutations and 438 indels) were from 26 tumors, characterized by the recurrent hotspot mutations P286R and V411L in *POLE*, a catalytic subunit of DNA polymerase epsilon that is involved in nuclear DNA replication and repair (Figure 2). These ultramutated tumors have a distinctive mutation spectrum, exemplified by elevated frequency of C→A transversions, as compared to lung cancer among smokers.[7] The 35,116 point mutations across the remaining 222 non-ultramutated tumors included 23,051 missense, 9,359 silent, 1,678 nonsense, 31 read-through, 601 splice-site mutations, and 396 in non-coding RNA genes. The 2,493 remaining indels included 1,459 frame-shift, 829 in-frame, 103 splice-site, and 102 in non-coding RNA genes. Of these 222 tumors, 175 were histologically classified as endometrioid, 43 as serous, and 4 as mixed. The 175 endometrioid tumors were comprised of 68 instances of grade 1, 68 of grade 2, and 39 of grade 3. Using 5 Bethesda markers for microsatellite instability (MSI), 109 endometrioid tumors were classified as microsatellite stable

(MSS), 3 as MSI-Low (MSI-L), and 63 as MSI-High (MSI-H). Based on these classifications, various subcohorts of the 222 non-ultramutated tumors were processed using the MuSiC suite of tools[6] to identify significantly mutated genes (Data File S3.2), mutation hotspots, mutual exclusivity or co-occurrence of mutations in genes, and correlations to clinical data.

Mutational significance in endometrioid histology

Across the 175 non-ultramutated endometrioid tumors, 37 significantly mutated genes (SMGs) were identified, with a convolution test false discovery rate (FDR) of 2% or less (Data File S3.2). In addition to previously implicated[8,9] genes *PTEN* (77.7%), *PIK3CA* (53.1%), *PIK3R1* (37.1%), *CTNNB1* (36.6%), *ARID1A* (35.4%), *KRAS* (24.6%), *CTCF* (20.6%), *RPL22* (12.6%), *TP53* (11.4%), *FGFR2* (10.9%), *ARID5B* (10.9%), *ATR* (6.9%), and *CCND1* (5.7%), several additional SMGs were identified, including *MLL4* (9.1%), *BCOR* (8.0%), *SPOP* (5.7%), *SIN3A* (5.7%), *MKI67* (5.7%), *FBXW7* (5.1%), *FOXA2* (5.1%), and *NRAS* (2.9%). Consistent with previous results,[10] our analysis showed that grade 3 endometrioid tumors had a higher frequency of *TP53* mutations (30.8%) than in grade 2 (11.8%) or grade 1 (0%), while *CTNNB1* mutations were more frequent in grade 1 (47.1%) and grade 2 (36.8%) than in grade 3 (17.9%). Further, 61 of the 136 tumors with altered *PTEN* had multiple non-silent *PTEN* mutations in the same tumor (52 tumors with 2 each, 9 with 3 each). Comparison of tumors with high mutation rates due to microsatellite instability (MSI) and lower mutation rates, revealed that MSI endometrioid tumors had a higher frequency of *MLL4* mutations (22.2%) than in microsatellite stable (MSS) tumors (1.8%). *ARID1A* had a similar mutation frequency in MSI-H (34.9%) and MSS (35.8%) endometrioid tumors, but *ARID5B* mutations were more frequent in MSI-H (20.6%) than in MSS (5.5%) endometrioid tumors. *RPL22* mutations were all frame-shift indels near homopolymers (Lys15), and significantly more frequent in MSI-H (34.9%) than in MSS (0%). Similarly, *ATR* mutations were mostly frame-shift indels, and more frequent in MSI-H (15.9%) than in MSS (1.8%).

Several of the novel SMGs identified have been implicated in cancers of other tissue types. For example, *FBXW7* (F-box/WD repeat-containing protein 7) and *MKI67* (Proliferation-related Ki-67 antigen) are frequently mutated in colorectal cancers.[11] *MLL4* (myeloid/lymphoid or mixed-lineage leukemia 4) shares a functional domain with epigenetic regulator *MLL2,* which is frequently mutated in non-Hodgkin lymphomas.[12] Recurrent *SPOP* mutations have also been reported in prostate cancer at, or near the F133 residue.[13] Interestingly, we identified 4 novel recurrent *SPOP* mutation sites (E50, M117, R121, and D140) across the 248 endometrial tumors, with E50K mutations recurrent in 3 tumors. 8 of 14 non-silent *BCOR* mutations identified across the 175 endometrioid tumors were N1459S, a highly recurrent site in our data set that has not been reported in COSMIC. *BCOR* mutations have been identified in various cancer types, including acute myeloid leukemia with normal karyotype.[14]

Transcriptional repressor *CTCF* and transcriptional activator *FOXA1* are frequently mutated in ductal breast cancer and occur mutually exclusively.[15] *CTCF* is a negative regulator of *FOXA1*, a key determinant of estrogen receptor function and endocrine response.[16] Although *FOXA1* mutations were absent across non-ultramutated endometrial tumors, *FOXA2* mutations were observed in 9 of the 175 endometrioid tumors, 4 of which co-occurred with *CTCF* mutations. *FOXA1* and *FOXA2* have distinct transcriptional circuitry,[17] and play a role in determining the

gender specificity of liver cancer.[18] They were also shown to oppositely regulate genes like *DIO1*,[19] a thyroid hormone activator, which was mutated in 3 of the 175 endometrioid tumors. It has also been shown that *FOXA1*-chromatin interaction is crucial in activating a cell-type-specific enhancer downstream of the cyclin D1 oncogene (*CCND1*), the latter being the primary recruitment site of estrogen receptor alpha (*ESR1*) in estrogen-responsive breast cancer cells.[20] Mutations were also seen near the phosphorylation site (T286) of Cyclin D1 (*CCND1*) in 10 of the 175 endometrioid tumors, as was previously reported.[21] And a similar hotspot of mutations in *ESR1* (Y537N, Y537S, Y537C, and D538G) was seen in 4 of the 175 endometrioid tumors, mutually exclusively of the 9 tumors with *FOXA2* mutations.

Mutational significance in serous histology

In the 43 serous tumors, 14 SMGs were identified with an FDR of 10% or less (Table S3.1), including previously implicated genes *TP53* (90.7%), *PIK3CA* (41.9%), *FBXW7* (30.2%), and *PPP2R1A* (27.9%). Additional SMGs included *CHD4* (16.3%), *CSMD3* (11.6%), *COL11A1* (11.6%), *PRPF18* (7%), *SPOP* (7%), and *CDH19* (7%). Non-silent mutations were also seen in *FGFR2* (7%), *ARID1A* (7%), *FOXA2* (4.6%), and *USP36* (4.6%), though they were not identified as *significantly* mutated possibly due to small sample size. While the majority of non-silent *FGFR2* mutations were seen in the 175 non-ultramutated endometrioid tumors (20 tumors), three serous tumors also harbored non-silent *FGFR2* mutations, including the recurrent N550K in one case.

Correlations to histological classifications and mutation spectra

The non-silent mutation statuses of 66 SMGs were correlated against qualitative clinical data types using Fisher's test, and against quantitative data types using the Wilcoxon rank-sum test. These 66 SMGs were selected from the union of SMGs in the 175 non-ultramutated endometrioid (58 SMGs) and 43 ultra-mutated serous (14 SMGs) samples (Data File S3.2). A loose FDR threshold of 15% was used in both cohorts (<15% in at least 2 of the 3 tests). In correlations to histology, endometrioid tumors were differentiated from serous tumors by frequent mutations in *PTEN* (p=5.5E-24), *CTNNB1* (p=1.8E-08), *ARID1A* (*P* = 9.5E-05), *CTCF* (*P* = 0.0001), and *KRAS* (*P* = 0.0003), while serous tumors were identifiable by mutations in *TP53* (*P* = 6.4E-23) and *PPP2R1A* (*P* = 4.8E-04). The 66 SMGs indicated above were run through MuSiC's clinical-correlation tool against the samples in the Figure 2 mutation spectra cohorts to identify 48 genes (Data File S3.1) that most significantly differentiated (FDR<1%) the cohorts. These 48 genes were then manually shortened to the genes in Figure 2d, based on mutation frequency and clustering across the cohorts.

Mutations in *PTEN* and *PIK3CA* compared to other tumor lineages

PIK3CA mutations are present in greater than 10% of breast, colorectal, head and neck, and lung squamous carcinomas. In endometrioid tumors, 33% of the *PIK3CA* mutations resided in exon 2, more than twice as many as were found in these other tumor types, including uterine serous carcinomas. Also, 22% of *PTEN* mutations occurred at R130, which is more than four times the rate seen in glioblastoma or in solid tumors with frequent *PTEN* mutations. Thus the location of recurrent mutations in *PIK3CA* and *PTEN* in endometrial carcinoma was different than in most other tumor lineages.

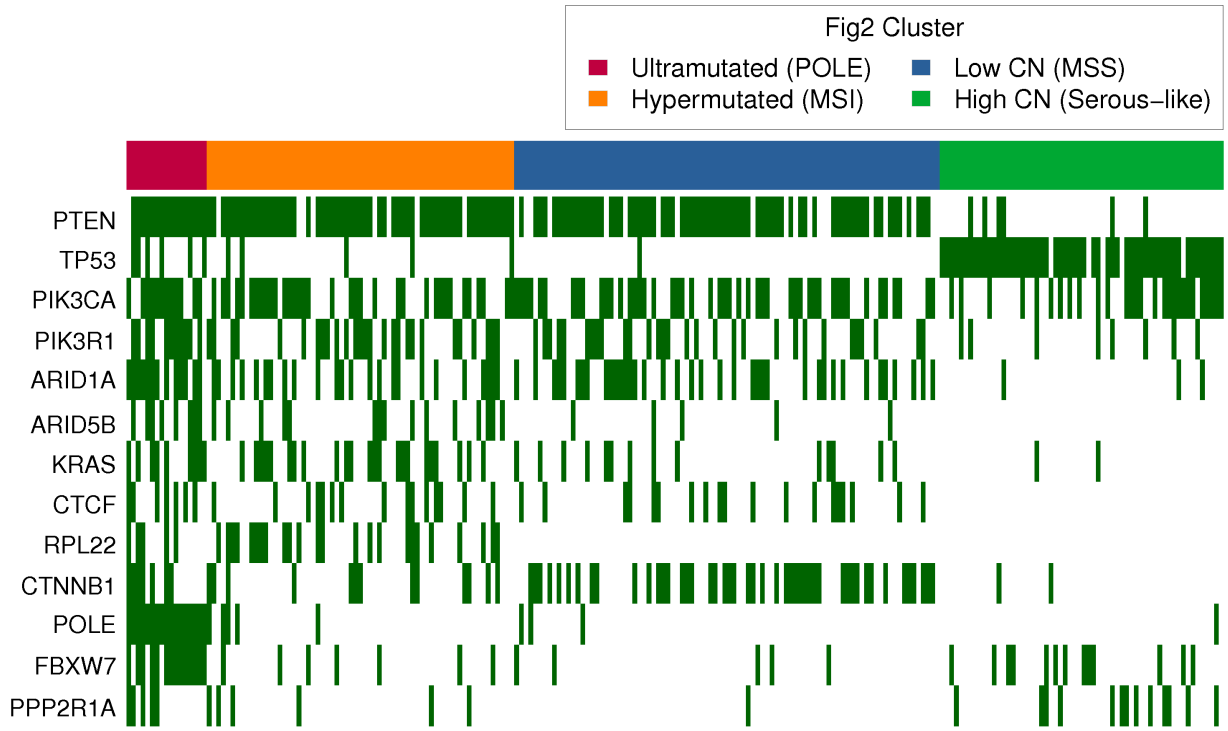


Figure S3.1 Non-silent mutation matrix for 13 SMGs with significantly different frequencies across the mutation spectra cohorts. A green bar indicates that at least 1 non-silent SNV or indel was identified in the tumor. *PTEN* and *TP53* distinguished the CN-high serous-like tumors from the remaining tumors. *PIK3CA* and *PIK3R1* mutations were mutually exclusive and *PIK3R1* was mutated significantly less in the CN-high group. *ARID1A* and *ARID5B* mutations also appear to be mutually exclusive and both have low mutation frequency in CN-high group. *KRAS* and *CTCF* mutations are rarely seen in the CN-high group. *CTNNB1* mutations appear more frequently in MSS tumors with lower mutation rates, than in endometrioid tumors with MSI. *RPL22* mutations were almost exclusive to cases with MSI. *FBXW7* and *PPP2R1A* mutations appear mutually exclusive, and are more common in CN-high serous-like tumors. *ARID1A* had a similar non-silent mutation frequency in MSI (36.9%) and MSS endometrioid tumors (42.2%), but a low frequency in high SCNA serous tumors (5%).

Date File S3.1 List of SMGs with significant frequency differences between mutation spectra cohorts.

The non-silent mutation status of 66 SMGs was collected across the 232 cases in Figure 2. The mutation frequency of each gene in the 4 groups, was tested against the expected mutation frequency by permutation, using Fisher's test. Multiple testing correction was applied using a false-discovery rate (FDR) threshold of <1%, leaving behind 48 SMGs that are significantly differently mutated across the 4 mutation spectra groups.

datafile.S3.1.SmgListMutationCohorts.xls

Date File S3.2 SMG lists in 15 subcohorts of 248 endometrial cases

These are results of 3 tests for significantly mutated genes (SMGs) from the MuSiC suite of tools, performed across 15 subcohorts of 248 tumor-normal pairs with exomes sequenced. Columns A to F contain gene names, the number of non-silent mutations, the altered cases, and the frequency of altered cases across that cohort. Column G contains the number of bases across the exons of that gene, which have sufficient coverage for variant detection (summed across samples in the cohort). Column H reports the number of non-silent mutations per million base pairs with sufficient coverage. The remaining columns report p-values and false discovery rates (FDR) for each gene tested by the three tests as described in the MuSiC manuscript[6] - Fisher's Combined P-value (FCPT), Likelihood Ratio (LRT), and Convolution (CT). Genes are included if the FDR cutoffs are <15% in at least two tests as implemented in MuSiC.[6]

datafile.S3.2.SmgListTopGenes.xls

Supplementary Methods for genome sequencing:

WGS (low-pass) Based Analysis of Structural Variations.
From 700 to 500 ng of each sample gDNA were sheared using Covaris E220 to about 250 bp fragments, than converted to a pair-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturers' protocols. All libraries were sequenced by HiSeq2000 using one sample – one lane, pair-end 2x51bp setup. Tumor and its matching normal were usually loaded to the same flowcell. Average sequence coverage was found to be 6.07, read quality 38.6, 94% reads mapped. Raw data were converted to FASTQ format then were fed to BWA alignment software to generate .bam files.

Identification of copy number variants. To characterize somatic copy number alterations in the tumor genome, we applied a new algorithm called BIC-seq to low-coverage whole-genome sequencing data. First, we counted the uniquely aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both fitness and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio difference smaller than 0.1 (log2 scale) between tumor and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

Translocation discovery with BreakDancer and MEERKAT. Structural Variation detection is performed with the program BreakDancer on a .bam file constructed from HiSeq sequencing of each tumor pair. The first step requires a configuration file of each bam file for each tumor pair with the bam2cfg.pl perl module of the program. After the configuration file, the perl module BreakDancerMax.pl is run on the configuration file in order to call structural variants in the tumor and control files. Each tumor structural variant file is filtered with its matched normal to remove any false positives. Structural variations are also detected by Meerkat which require at least two discordant read pairs supporting one event and at least one read covering the breakpoint junction. Each variant detected from tumor genome is filtered with all normal genomes to remove germline events. The structural variants are filtered out if both breakpoints fall into simple repeats or satellite repeats.

We detected 1,166 candidate structural variant (inter-, intra-, del-, inv-) events (average=11/tumor). Among the translocation events that involved at least one gene, 358 had one of the breakpoints in an intergenic region, whereas the remaining 551 juxtaposed coding regions of two genes in putative fusion events of which 423 were predicted to code for in-frame events (Table S3.1).

Validation of translocation hits. To understand the translocations at the structural level, we PCR amplified the junction fragments using primers from regions of the two chromosomes close to the region of putative breakpoints and the DNA from this product was subjected to sequencing using the Sanger method on a capillary electrophoresis unit. We attempted to validate the translocations using two different approaches. MEERKAT determines translocations on the basis of discordant reads as well as reads that span the translocation junction (split reads). We also attempted to validate several translocations by attempting to PCR amplify the junctions of the translocation and sequencing the products. Based on these two approaches we validated 29/50 (58%) of translocations. Therefore, it is possible that the false discovery rate could be as high as 42%.

Table S3.1 Genes involved in recurrent translocations in endometrial cancer from 106 tumor / normal pairs.

Genes	Type	Samples	Detected by DNA	Detected by RNA
ARHGAP	Interchromosomal , Deletion, Intrachromosomal	7	All	
BCL	Intrachromsomal, Deletion, Interchromosomal	5	All	2
AKAP	Interchromosomal , Deletion, Intrachromosomal	5	All	
EIF2C2	Intrachromsomal	3	All	
CRHR1-MAPT	Intrachromsomal	2	All	
GNG5-RPF1	Inversion	2	All	
CACNA2D2	Inversion, Deletion	2	All	
ASXL2	Inversion, Interchromosomal	2	All	
CSNK1D/1E/1G2	Interchromosomal , Tandem Duplication	4	2	4
CRHR1-MAPT	Interchromosomal	2	All	
NCOA3-EYA2	Tandem Duplication	1		1
GADD45GIP-CSTF1	Interchromosomal	1	1	1
*CDK12**	Inversion or Translocation	3	1	2
ERBB2-TSPAN11	Interchromosomal	1		1
PIK3CA-KCNMB3	Inversion	1		1
*SRP68**	Deletion or Translocation	2	1	1
EYA2-SLC2A10	Tandem Duplication	1	1	1

** Fusions with different partners*

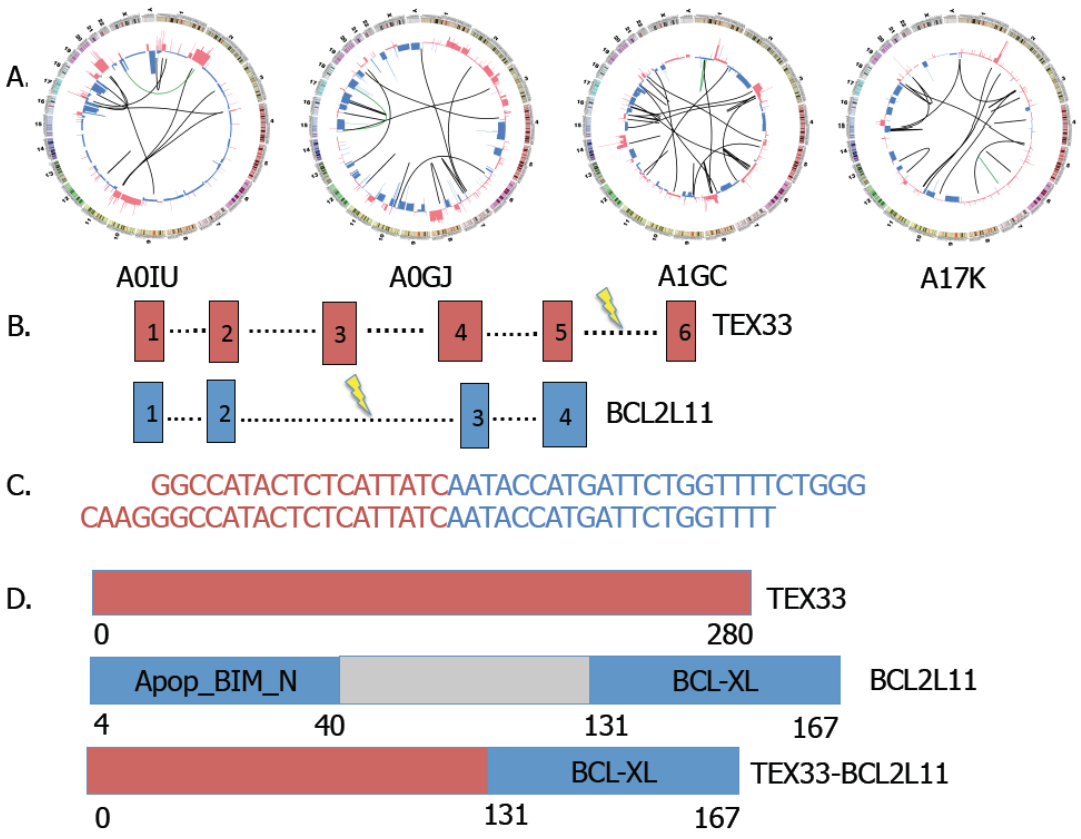


Figure S3.2 Recurrent translocations involving members of the BCL family of genes.
A. Circular diagrams showing CNV and translocations in four tumors. Translocations involving BCL genes are indicated in blue. B. Translocation involving TEX33 and BCL2L11 genes. Exon-intron structures of the two genes and the sites of breakpoints are shown. C. Sequence at the translocation breakpoints. A PCR fragment spanning the translocation breakpoint was isolated and sequenced. A second sequence spanning the breakpoint was identified by our MEERKAT software. Nucleotides in red correspond to TEX33 gene and nucleotides in blue correspond to BCL2L11 gene. D. Predicted structure of the fusion protein resulting from the translocation. The fusion results in the loss of the BIM domain of BCL2L11 that is required for intrinsic mitochondrial mediated apoptosis. Loss of the BIM domain may cause reduction in apoptosis.

Section References

1. Li, H. *et al*; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-2079 (2009).
2. Koboldt, D.C. *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**:568-576 (2012).
3. Larson, D.E. *et al*. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**:311-317 (2012).
4. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**:1297-1303 (2010).
5. Ye, K., Schulz, M.H., 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* **25**:2865-2871 (2009).
6. Dees, N.D. *et al*. MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**:1589-1598 (2012).
7. Ding, L. *et al*. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**:1069-1075 (2008).
8. Liang, H. *et al*. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Res* **22**:2120-2129 (2012).
9. Cheung, L.W. *et al*. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. *Cancer Discov* **1**:170-185 (2011).
10. McConechy, M.K. *et al*. Use of mutation profiles to refine the classification of endometrial carcinomas. *J Pathol* **228**:20-30 (2012).
11. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061-1068 (2008).
12. Morin, R.D. et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**:298-303 (2011).
13. Barbieri, C.E. *et al*. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44**:685-689 (2012).
14. Grossmann, V. *et al*. Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118**:6153-6163 (2011).
15. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**:61-70 (2012).
16. Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D., Carroll, J.S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**:27-33 (2011).
17. Bochkis, I.M. Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2. *PLoS Genet* **8**:e1002770 (2012).

18. Li, Z., Tuteja, G., Schug, J., Kaestner, K.H. Foxa1 and Foxa2 are essential for sexual dimorphism in liver cancer. *Cell* **148**:72-83 (2012).
19. Kanamoto, N. *et al*. Forkhead box A1 (FOXA1) and A2 (FOXA2) oppositely regulate human type 1 iodothyronine deiodinase gene in liver. *Endocrinology* **153**:492-500 (2012).
20. Eeckhoute, J., Carroll, J.S., Geistlinger, T.R., Torres-Arzayus, M.I., Brown, M. A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev* **20**:2513-2526 (2006).
21. Moreno-Bueno, G. *et al*. Cyclin D1 gene (CCND1) mutations in endometrial cancer. *Oncogene* **22**:6115-6118 (2003).

Supplementary Methods S4: RNA sequencing.

Identification of Gene Expression-Based Subtypes

The gene expression profiling of 333 endometrial tumors was filtered to eliminate unreliably measured genes and to limit the clustering to relevant genes.[1,2] Genes that are not well-characterized on the basis of HGNC's description or have small expression values in at least one-fourth of the samples[3] were removed. We then filtered out genes with small signal-to-noise ratio (SNR) where SNR was calculated per gene by subtracting from the gene estimate the mean expression value across patients and then dividing it by its standard deviation across patients.[1] Next, we calculated the gene expression profile variances across the samples that were subsequently used to rank the genes in a descending order. The final filter excluded genes with smaller variability and selected the top 15 percent of genes with the highest values of variances. Implementation of these filters resulted in 1368 genes with reliably measured and highly variable expression. The expression data were then median centered and log transformed.

Next, we applied k-means unsupervised clustering with a randomized selection of the initial cluster centroids from the samples as our basis for consensus clustering, to detect robust clusters. This clustering approach uses a two-phase iterative algorithm that assigns samples to clusters so that the sum of distances from each sample to its cluster centroid, over all clusters, is a minimum. The distance metric was one minus the Pearson's correlation coefficient and each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation. The procedure was repeated over 1000 times, each with a new set of initial cluster centroid positions to avoid a local minimum. Silhouette width values were calculated accordingly for all samples. Silhouette width is defined as the ratio of each sample's average distance to samples in the same cluster to the smallest distance to samples not in the same cluster.

Average silhouette width and percentage of samples with larger silhouette width of greater than 0.2 were calculated for different number of clustering, k (Figure S4.4). Except the two-cluster assignment that is essentially driven by the histopathological classification, clustering with k = 3 gave the highest average silhouette value and percentage of number of patients with larger silhouette values, and was thus subject to further investigation.

We applied significance analysis of microarray (SAM)[4] to identify marker genes that are associated with the transcriptome subtypes. Each class was compared to the other two classes combined, and each class was compared to the other individual classes in a pairwise manner. We provided both rank order and test statistic for all of these analyses.[1] Genes exhibiting positive expression difference and statistical significance in all these analyses were selected as gene signatures associated with the subtypes. A combined *P* value for each gene was calculated and used for ranking the genes in the gene signatures. The subtype of TCGA samples and the normalized gene expression profiling were visualized in the heatmap using the 450 genes (Figure S4.1, 150 top ranked genes per class selected from the gene signatures).[5] The identified gene signatures were then subject to pathway analysis (Ingenuity Pathway Analysis,

version 12710793) and the statistical significance of pathway enrichment was determined by Fisher's exact test (Figure S4.5). The results from the pathway analysis were used to term the gene expression clusters. In addition, both hormone receptors (*ESR1* and *PGR*) were significantly higher in the hormonal subtype at both RNA and protein levels (Figure S4.6). A similar approach was applied to the RPPA profiling in order to identify the differentially expression cancer- related proteins and phosphor-proteins that are associated with the gene expression subtypes (Supplementary Methods S5).

Unsupervised Clustering

Unsupervised k-means clustering of 333 endometrial tumors using 1,368 mRNAs that had most variable expression identified three robust clusters, termed 'mitotic', 'hormonal', and 'immunoreactive' based on pathway analysis (Fig. S4.1). The mitotic subtype (n=126) was characterized by TP53 mutation and included most of the serous/mixed histology tumors (57 of 62) and endometrioid grade 3 tumors (57 of 102). High expression of ESR1, PGR, and their downstream targets in the hormonal subtype (n = 111) revealed unique biology in this group of patients, who may be more responsive to hormonal therapy (Fig. S4.3). Intriguingly, immune response genes characterized the immunoreactive subtype (n = 96); however, this subtype showed the same level of infiltrating immune cells as the other subtypes (Fig. S4.7). It is possible that tumor cells contributed to the immunoreactive gene expression; alternatively, immune cells in the immunoreactive subtype might be activated, contributing to the unique tumor environment. Both hormonal and immunoreactive subtypes were primarily composed of endometrioid grade 1 or 2 tumors and PTEN mutated cases.

Lymphocyte Contents Across the Gene Subtypes

To determine whether or not the cluster assignment was biased from the immune cell infiltrate, we examined the percentage of lymphocytes, macrophages and neutrophils from both the top and bottom sides of non-malignant tissue. The one-way ANOVA test showed that there was no significant difference in the percentage of individuals or their summation (Figure S4.7), suggesting that the immunoreactive subtype was due to the underlying molecular features instead of contamination from the inflammatory cells in the tumor tissue.

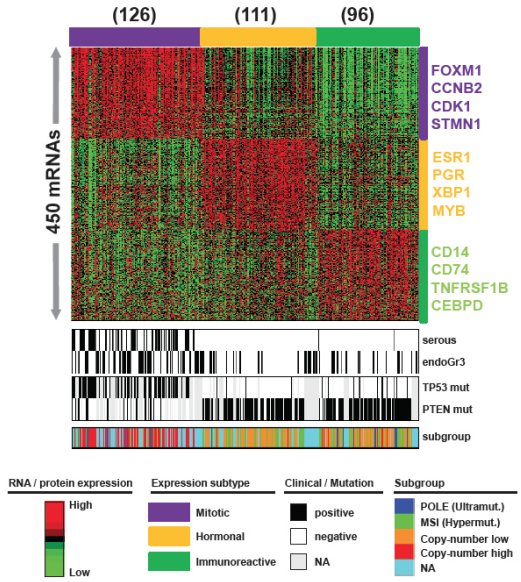


Figure S4.1 Gene expression subtypes in endometrial carcinomas.
Three gene expression subtypes were identified via unsupervised k-means clustering of TCGA endometrial tumors and significantly correlated with clinical (histology and endometrioid grade) and molecular (*TP53* / *PTEN* mutations) features, and mutation clusters.

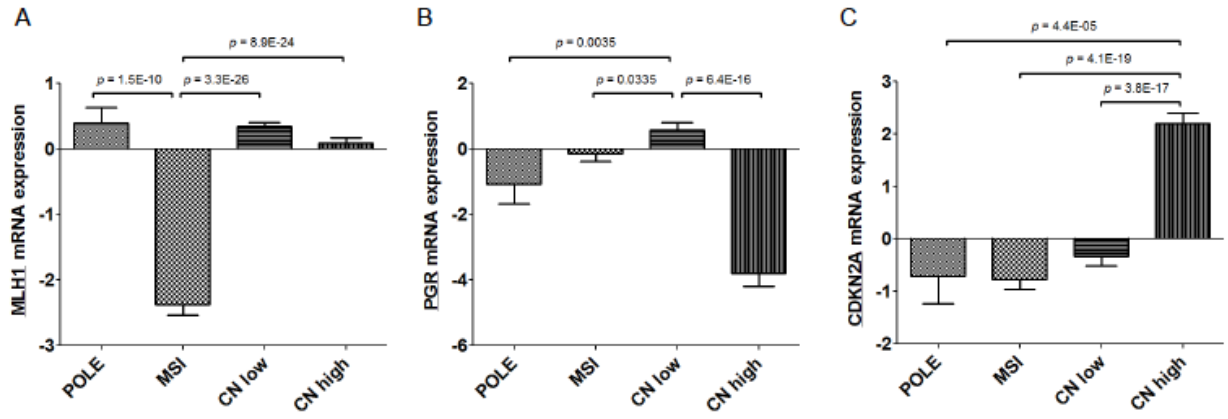


Figure S4.2 Gene expression across integrated subtypes. (A) *MLH1* mRNA expression is significantly lower in the MSI cluster. (B) *PGR* mRNA expression is significantly higher in the CN low cluster. (C) *CDKN2A* mRNA expression is significantly higher in the CN high cluster.

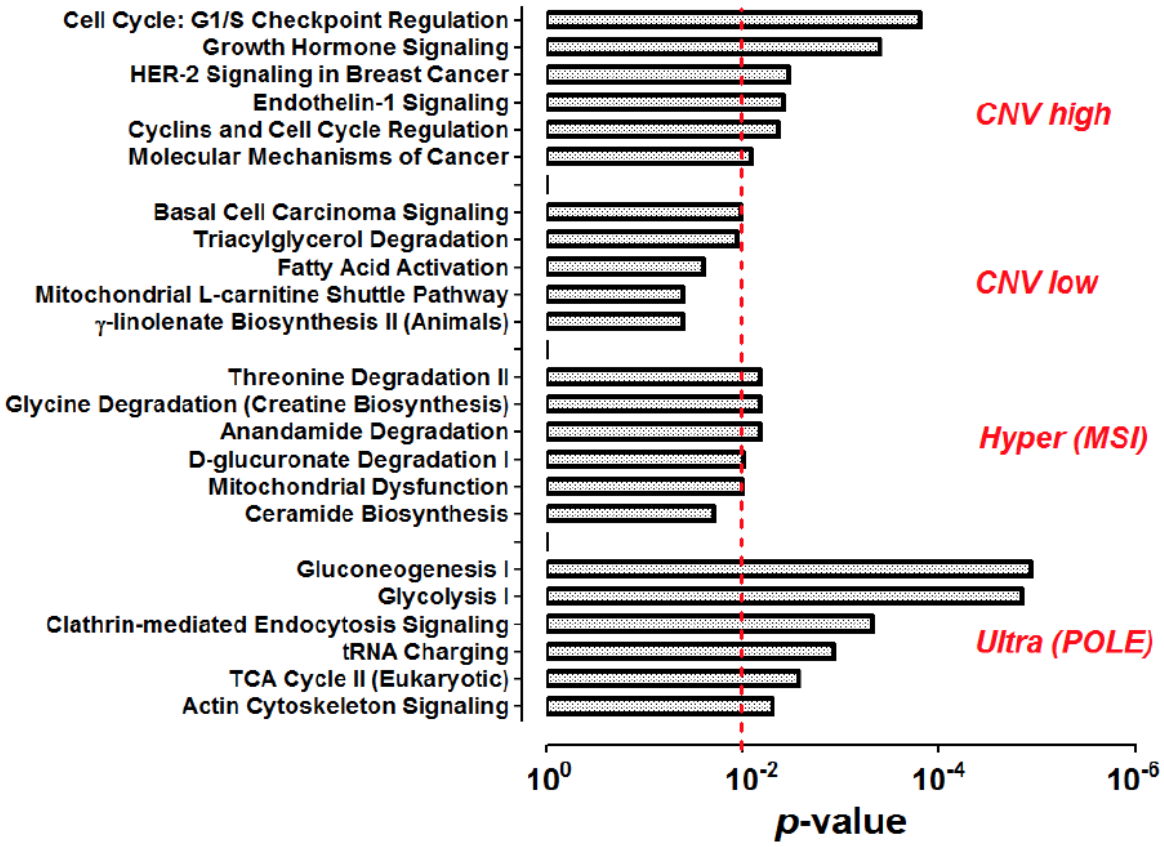
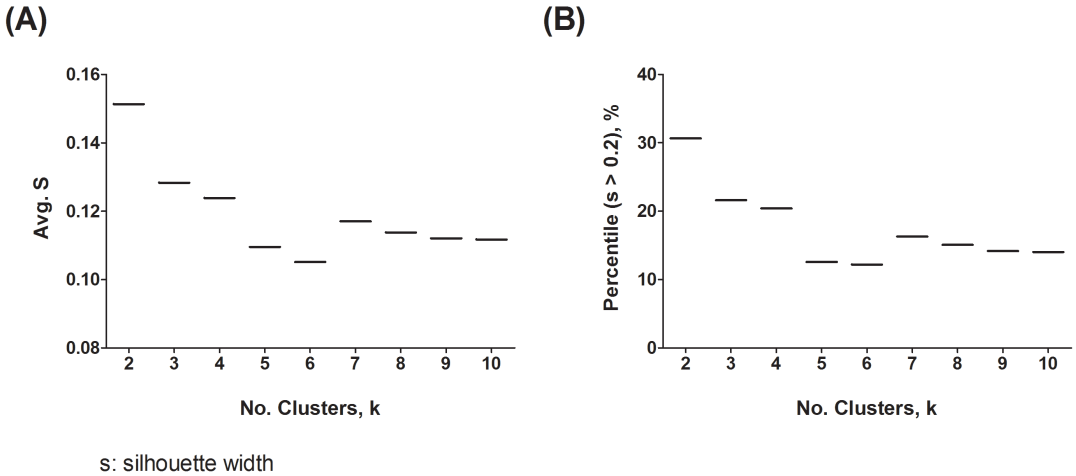


Figure S4.3 Significantly enriched pathways across integrated subtypes.



s: silhouette width

Figure S4.4 (A). Average of silhouette width values in different number of clustering, k. (B). Percentile of samples with silhouette width values of greater than 0.2 in different number of clustering, k.

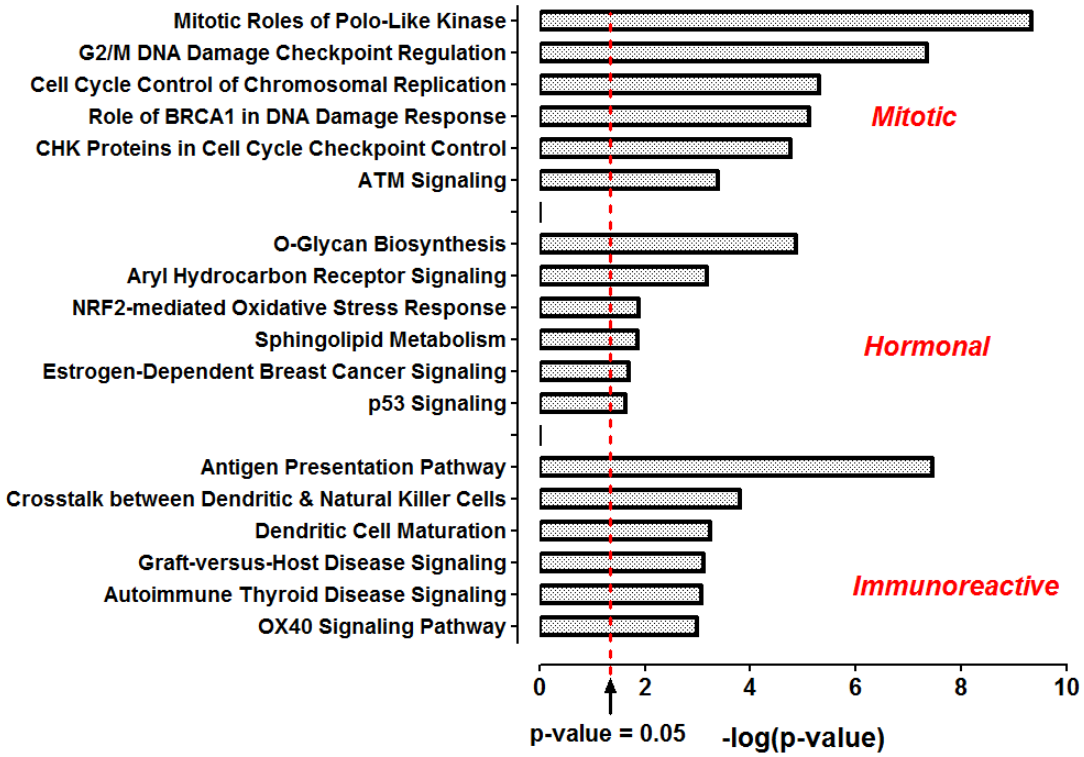


Figure S4.5 The most significantly enriched pathways in different gene expression subtypes.

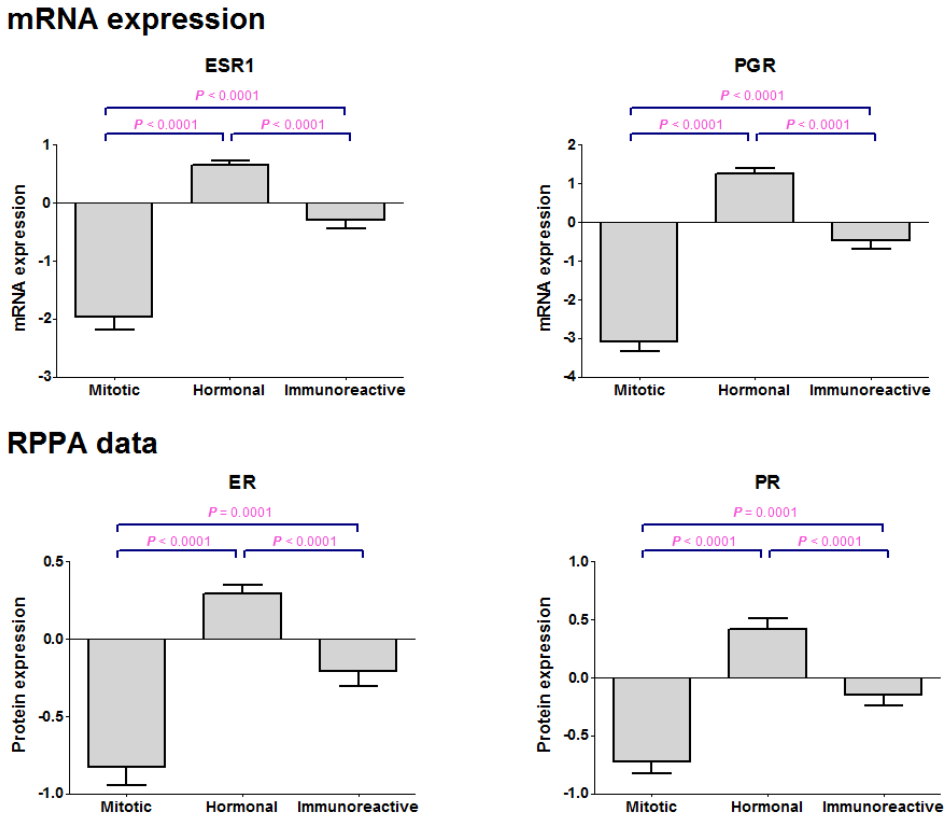


Figure S4.6 Both hormone receptors (*ESR1* and *PGR*) were significantly higher in the hormonal subtype at both RNA and protein levels.

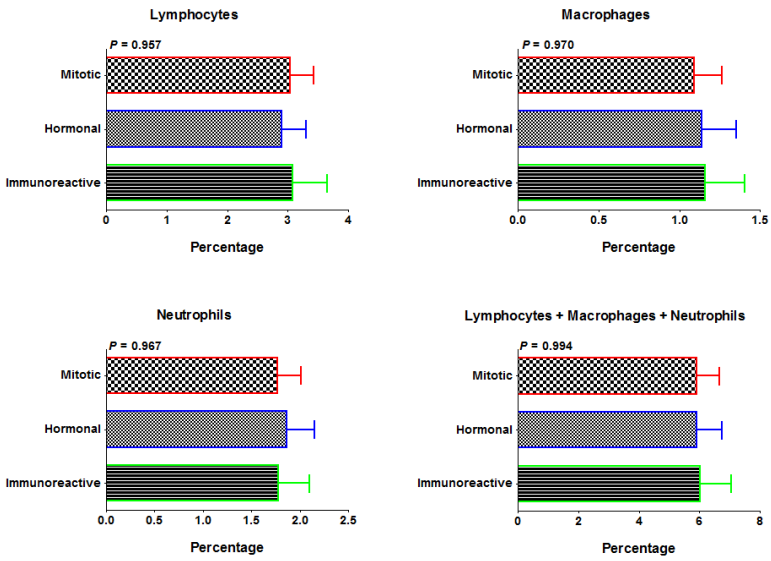


Figure S4.7 Percentage of lymphocytes, macrophages, neutrophils and their sum across the three gene expression subtypes.

Identification of fusion genes from RNA sequencing data - Methods

To identify fusion genes based on whole transcriptome sequencing data, full 76 bp Illumina reads were first aligned against Ensembl 68 transcript sequences. Reads that did not align against the transcriptome were then aligned against the GRCh37 human reference genome. In both steps, alignments were performed using Bowtie version 2.0.0-beta7, with the mismatch threshold parameter set to *–score-min L,0,-0.2* to allow for a maximum of two nucleotide mismatches. Reads that did not align to the transcriptome or genome were split to produce 25 bp anchors from both ends of the read. The anchors were aligned against the GRCh37 reference genome with zero mismatches allowed. Anchor pairs where one or both anchors did not align to the genome were discarded, as were anchors that aligned within 100 kb of one another or within a single gene. To determine the exact location of the fusion junction corresponding to each remaining anchor pair, the anchors were extended to full 76 bp reads, with the breakpoint positioned so that the number of nucleotide mismatches between the breakpoint flanks and the 76 bp read was minimized. Microhomologies at the fusion junction often render it impossible to exactly locate the breakpoint, so a range of breakpoint locations was accepted if a microhomology was found. If a region of microhomology overlapped an exon boundary, the RNA level breakpoint was fixed to the exon boundary.

Fusion genes involving immunoglobulin or HLA loci were discarded, because these hypervariable regions often produce many false positive fusions. Fusion genes involving ribosomal RNA genes and related proteins were also discarded, as highly expressed genes such as these often display a high number of PCR chimaeras, technical artifacts where pieces of two transcripts are merged together during PCR amplification.

Putative fusion genes were also discarded if they involved more than 2 nucleotide mismatches in all junction overlapping reads, or if all the reads together did not cover at least 40 bp on both sides of the junction. Fusion gene candidates were also discarded if the flanks of the two breakpoints were too homologous. Finally, a fusion gene candidate was discarded if the candidate was also found in one of the three available normal tissue samples.

Identification of fusion genes from RNA sequencing data - Results

We identified NCOA3-EYA2 fusions in two endometrial cancer patients. Patient TCGA-AP-A053 harbored a fusion that fused NCOA3 exon 1 to EYA2 exon 2. Patient TCGA-D1-A179 harbored a fusion that fused NCOA3 exon 21 to EYA2 exon 8. The two genes are both located in chromosome 20, with EYA2 situated directly upstream of NCOA3. The fusion gene probably arises due to a 500 kb tandem duplication in this locus. NCOA3 is a nuclear receptor coactivator that is frequently amplified and overexpressed in breast and ovarian cancers. It is therefore possible that NCOA3-EYA2 fusions merely represent a passenger event that arises as a side effect of NCOA3 amplification.

A single GADD45GIP1-CSTF1 fusion was detected in patient TCGA-D1-A17K. In the fusion, GADD45GIP1 exon 1 is fused to CSTF1 exon 6. The fusion is caused by interchromosomal rearrangement and does not produce a functional fusion protein as the fusion fuses the CDS of GADD45GIP1 to the 3’ UTR of CSTF1. GADD45GIP1 is a nuclear-localized protein that regulates

the cell cycle by inhibiting G1 to S phase progression. CSTF1 is a cleavage stimulation factor. We hypothesize that the fusion causes GADD45GIP1 loss of function, leading to cell cycle deregulation.

We observed a complex CDK12 rearrangement in patient TCGA-AP-A053. This patient harbored the fusions PSMD3-CDK12 and CDK12-EIF4A3. In the former fusion, PSMD3 exon 7 is fused to CDK12 exon 6, resulting in the formation of a non-frameshifted fusion protein. The location of the two genes suggests that the fusion is probably caused by a 500kb tandem duplication on 17q. The genes CDK12 and PSMD3 flank the ERBB2 locus on chromosome 17, suggesting that the fusion is a passenger event that arose as a side effect of ERBB2 amplification.

Table S4.1: Putative fusion candidates identified from RNA sequencing data

Fusion	Junction	Frequency	Mechanism	Predicted biological impact
NCOA3-EYA2	chr20:+:46130763 -> chr20:+:45618640; chr20:+:46280020 -> chr20:+:45717878	2 / 322	Tandem duplication	Variant #1 produces full length EYA2. Variant #2 produces an in-frame chimeric protein. Fusion is possibly a side effect of NCOA3 amplification.
ARL3-ACO1	chr10:-:104445573 -> chr9:+:32440463	1 / 322	Interchromosomal	In-frame chimeric protein.
GDA-EPS8L2	chr9:+:74764598 -> chr11:+:720062	1 / 322	Interchromosomal	In-frame chimeric protein.
SCNN1A-LPAR5	chr12:-:6463604 -> chr12:-:6730630	1 / 322	Tandem duplication	SCNN1A coding region fused with LPAR5 5'-UTR.
SSU72-ASCC1	chr1:-:1509858 -> chr10:-:73887925	1 / 322	Interchromosomal	Frameshifted chimeric protein.
FMNL2-GPBAR1	chr2:+:153437563 -> chr2:+:219127403	1 / 322	Deletion	FMNL2 coding region fused with GPBAR1 5'-UTR.
ZNF3-ACTL6B	chr7:-:99677159 -> chr7:-:100247758	1 / 322	Tandem duplication	5'-UTR of ZNF3 fused with the ACTL6B coding region.
MTMR3-GJB1	chr22:+:30762237 -> chrX:+:70443542	1 / 322	Interchromosomal	CCDC157-GJB1 is also possible (same junction sequence).
NFAT5-SNTB2	chr16:+:69600277 -> chr16:+:69279505	1 / 322	Tandem duplication	In-frame chimeric protein.
EYA2-SLC2A10	chr20:+:45523626 -> chr20:+:45353680	1 / 322	Tandem duplication	5' UTR to CDS fusion -> truncated SLC2A10?
KDM4A-PNKD	chr1:+:44160565 -> chr2:+:219182678	1 / 322	Interchromosomal	PNKD breakpoint is intronic.
NPLOC4-SIRT7	chr17:-:79589192 -> chr17:-:79872406	1 / 322	Tandem duplication	Frameshifted chimeric protein.
KIAA0100-MYO18A	chr17:-:26945809 -> chr17:-:27449271	1 / 322	Tandem duplication	In-frame chimeric protein.
MICALL1-IFT122	chr22:+:38329119 -> chr3:+:129236312	1 / 322	Interchromosomal	In-frame chimeric protein.
PUM1-NKAIN1	chr1:-:31532051 -> chr1:-:31661034	1 / 322	Tandem duplication	In-frame chimeric protein.
UNK-MYO15B	chr17:+:73781065 -> chr17:+:73620851	1 / 322	Tandem duplication	Frameshifted chimeric protein.
CSNK1D-POLR2E	chr17:-:80223562 -> chr19:-:1090144	1 / 322	Interchromosomal	Frameshifted chimeric protein.
DOT1L-CSNK1G2	chr19:+:2180755 ->	1 / 322	Tandem duplication	DOT1L coding region fused

	chr19:+:1969507			with the CSNK1G2 5′-UTR.
GALNT14-BRE	chr2:-:31360824 -> chr2:+:28117417	1 / 322	Inversion	GALNT14 coding region fused with the BRE 5′-UTR.
SLC25A13-SHFM1	chr7:-:95926210 -> chr7:-:96324203	1 / 322	Tandem duplication	Frameshifted chimeric protein.
TANC1-SLC4A1AP	chr2:+:159922483 -> chr2:+:27907904	1 / 322	Intrachromosomal	Frameshifted chimeric protein.
GADD45GIP-CSTF1	chr19:-:13067677 -> chr20:+:54978524	1 / 322	Interchromosomal	Frameshifted chimeric protein. GADD45GIP1 loss of function -> cell cycle deregulation.
PSMD3-CDK12	chr17:+:38151321 -> chr17:+:37657503	1 / 322	Tandem duplication	In-frame chimeric protein. Fusion is possibly a side effect of ERBB2 amplification.
CDK12-EIF4A3	chr17:+:37657692 -> chr17:-:78116886	1 / 322	Inversion	Involves an unannotated exon.
NSUN5-BCL7B	chr7:-:72722428 -> chr7:-:72957974	1 / 322	Tandem duplication	In-frame chimeric protein. Fusion is possibly a side effect of FZD9 amplification?
BCL9-RASAL2	chr1:+:147087648 -> chr1:+:178399568	1 / 322	Deletion	BCL9 coding region fused with a RASAL2 intron.
ERBB2-TSPAN11	chr17:+:37868300 -> chr12:+:31110098	1 / 322	Interchromosomal	Involves an unannotated exon.
PIK3CA-KCNMB3	chr3:+:178886391 -> chr3:-:178968722	1 / 322	Inversion	Involves an unannotated exon.
SRP68-MIEN1	chr17:-:74063298 -> chr17:-:37886544	1 / 322	Deletion	In-frame chimeric protein.
DDI2-EFHD2	chr1:+:15953293 -> chr1:+:15752367	1 / 322	Tandem duplication	Frameshifted chimeric protein. DDI2 loss of function may lead to loss of DNA damage response.
LRIG1-SLC25A26	chr3:-:66550614 -> chr3:+:66419902	1 / 322	Inversion	Frameshifted chimeric protein.
DNAJC1-SFTPD	chr10:-:22217969 -> chr10:-:81706418	1 / 322	Intrachromosomal	DNAJC1 coding region fused with SFTPD 5′-UTR.
CSNK1E-SPATA21	chr22:-:38694791 -> chr1:+:16717870	1 / 322	Interchromosomal	In-frame chimeric protein.
MTMR3-HEXIM2	chr22:+:30279348 -> chr17:+:43246382	1 / 322	Interchromosomal	MTMR3 5′-UTR fused with the HEXIM2 coding region.
CSNK1E-DYNLRB1	chr22:-:38794443 -> chr20:+:33114073	1 / 322	Interchromosomal	CSNK1E 5′-UTR fused with the DYNLRB1 coding region.
ZNF704-GFER	chr8:-:81733609 -> chr16:+:2035867	1 / 322	Interchromosomal	In-frame chimeric protein.
DLG4-SHBG	chr17:-:7106222 -> chr17:+:7536522	1 / 322	Inversion	In-frame chimeric protein.
QRICH1-SHISA5	chr3:-:49094295 -> chr3:-:48520665	1 / 322	Deletion	Frameshifted chimeric protein.

CPNE2-NUP93	chr16:+:57171194 -> chr16:+:56857619	1 / 322	Tandem duplication	In-frame chimeric protein.
POLR2B-FIP1L1	chr4:+:57845170 -> chr4:+:54308820	1 / 322	Tandem duplication	In-frame chimeric protein.
AACS-MAL	chr12:+:125561157 -> chr2:+:95713704	1 / 322	Interchromosomal	Frameshifted chimeric protein.
BICD1-METTL20	chr12:+:32459056 -> chr12:+:31814775	1 / 322	Tandem duplication	BICD1 coding region fused with the METTL20 5'-UTR.
CTTN-ANO1	chr11:+:70275305 -> chr11:+:69970461	1 / 322	Tandem duplication	In-frame chimeric protein.
TRHDE-LGR5	chr12:+:72680696 -> chr12:+:71918186	1 / 322	Tandem duplication	Frameshifted chimeric protein.
CELSR1-KCNJ4	chr22:-:46787083 -> chr22:-:38824176	1 / 322	Deletion	CELSR1 coding region fused with the KCNJ4 5'-UTR.
ITCH-RALY	chr20:+:32951155 -> chr20:+:32619328	1 / 322	Tandem duplication	ITCH 5'-UTR fused with the RALY 5'-UTR.
KIF26A-CKB	chr14:+:104618798 -> chr14:-:103988842	1 / 322	Inversion	KIF26A coding region fused with the CKB 5'-UTR.
ODF2-SLC27A4	chr9:+:131236020 -> chr9:+:131122613	1 / 322	Tandem duplication	Frameshifted chimeric protein.

Section references

1. Verhaak, R.G., Hoadley, K.A., Purdom, E., Layes, D.N. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**:98-110 (2010).
2. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**:609-615 (2011).
3. Salvesen, H.B. *et al*. Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *PNAS* **106**:4834-4839 (2009).
4. Tusher, V.G., Tibshirani, R., Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**:5116-5121 (2001).
5. TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**:330-337 (2012).

Supplementary Methods S5: Reverse phase protein arrays

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.[1-5] Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually 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 170 validated primary antibodies (Data File S5.1) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Slides were scanned in CanoScan 9000F. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI,[3,5] available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model.[1] During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[5] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[3,5,6] using median centering across antibodies (level 3 data). In total, 170 antibodies and 316 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.[7] These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described.[7]

Two RPPA arrays were quantitated and processed (including normalization and load controlling) as described previously, using MicroVigene (VigeneTech, Inc., Carlisle, MA) and the R package SuperCurve (version-1.3), available at http://bioinformatics.mdanderson.org/OOMPA.[1,3] Raw data (level 1), SuperCurve nonparametric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

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. Among 316 samples with RPPA data, 302 have available clinical data; and nine samples were removed due to the concern of data quality. Further analyses were performed on the remaining 293 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.[8] processed clinically similar ALL samples in two batches and observed differences in their protein data distributions. 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 two batches of endometrial tumor protein expression data. To remove those technical effects, we median centered the samples on each slide. Then, we divided the slide by its standard deviation. The procedure adjusted the location and scale of each slide, so that its median became zero and standard deviation became one. Multiple slides from different batches could then be compared against each other. Of course, that meant that we couldn't directly compare the expression levels of one protein with another, but RPPA has already that limitation built in. Our normalization procedure significantly reduced technical effects, thereby allowing us to merge the datasets from different batches.

Hierarchical clustering

We used bootstrap to resample (N=3000) the proteins to estimate the number of sample clusters. Pearson correlation was used as distance matrix and Ward was used as a linkage algorithm in the unsupervised hierarchical clustering analysis. This method clustered samples and counted how frequently two samples are in the same cluster. The bootstrap resampling analysis identified five robust sample clusters. The five clusters and their protein expression patterns can be viewed through the next generation clustered heat map (NG-CHM) pipeline developed at the University of Texas MD Anderson Cancer Center.

Unsupervised clustering

Unsupervised hierarchical clustering analysis revealed five robust protein clusters (Fig. S5.1). These five protein subtypes were significantly correlated with histology (P=2.2×10^{-16}) and grade (P=1.57×10^{-7}) as well as with the subtypes/clusters defined by other genomic data including copy number variation (P=1.67×10^{-11}), DNA methylation (P=2.42×10^{-10}), *MHL1* hypermethylation (P=5.0×10^{-7}), microRNA (P=2.54×10^{-9}), and mRNA (P=1.4×10^{-13}). RPPA cluster 1 (signaling on) was associated with activation of signaling pathways, by expression of hormone receptors, and by an enrichment of *PIK3R1* and *KRAS* mutations. RPPA cluster 2 (serous) mainly consisted of serous or serous-like samples with a strong concordance with mRNA cluster 1 and an enrichment of *TP53* mutations. RPPA cluster 3 (signaling off) was associated with low levels of signaling pathway activity. RPPA cluster 4 (RAS/reactive) had selective activity of the

RAS/MAPK pathway without PI3K pathway activation as well as evidence for a reactive stroma. RPPA cluster 5 (reactive) had high levels of collagen VI, caveolin 1, and VEGFR compatible with activated stroma, and a depletion of *PIK3R1* and *KRAS* mutations.

Supervised clustering

We measured the expression of 170 cancer-related proteins and phospho-proteins using RPPA in 293 qualified tumor samples. Supervised analysis of the RPPA profiling data identified 36 of 170 cancer-related proteins and phospho-proteins (e.g., TP53, CCNB1, CDK1, ER, PR, AR, p-STAT3) significantly associated with and supportive of the transcriptome clusters (Fig. S5.2). In particular, elevated phosphorylated STAT3 (STAT3-pY705) was observed in the immunoreactive subtype, consistent with a key role of *STAT3* transcriptional activity in regulating immune response. The integrated CN high group had elevated CCNE1, CCNB1, and CDK1 consistent with an increased proliferative rate as well as elevated p53 and phospho-CHK2 suggestive of elevated levels of DNA damage. The CN low group had elevated SYK, which is associated with lymphocytic infiltration.

Correlation and survival analysis

Chi square test was used to evaluate the correlations between RPPA clusters and histology, grade, stage, or the clusters determined by other genomic data. Log-rank test and Kaplan-Meier survival curves were used to compare overall survival (OS) or progression-free survival (PFS) between different clusters of patients. A significance level of 0.05 was used (Table S5.1).

Table S5.1. Chi square test estimates correlation between RPPA clusters and clinical variables, clusters defined by other platforms, and gene mutations.

RPPA Clusters vs	***P***
Clinical Variables	
Histology	2.2e-16
Grade	1.57e-07
Stage	0.098
Clusters Defined by Other Platforms	
CNA-K4	1.67e-11
Methylation	2.42e-10
MHL1 Hypermethylation	5.01e-07
Micro RNA	2.54e-09
mRNA	1.4e-13
Mutation	0.17
Gene Mutation	
PIK3CA	0.57
PIK3R1	0.014
PTEN	2.35e-10
KRAS	0.014
ARID1A	0.13
TP53	3.7e-11

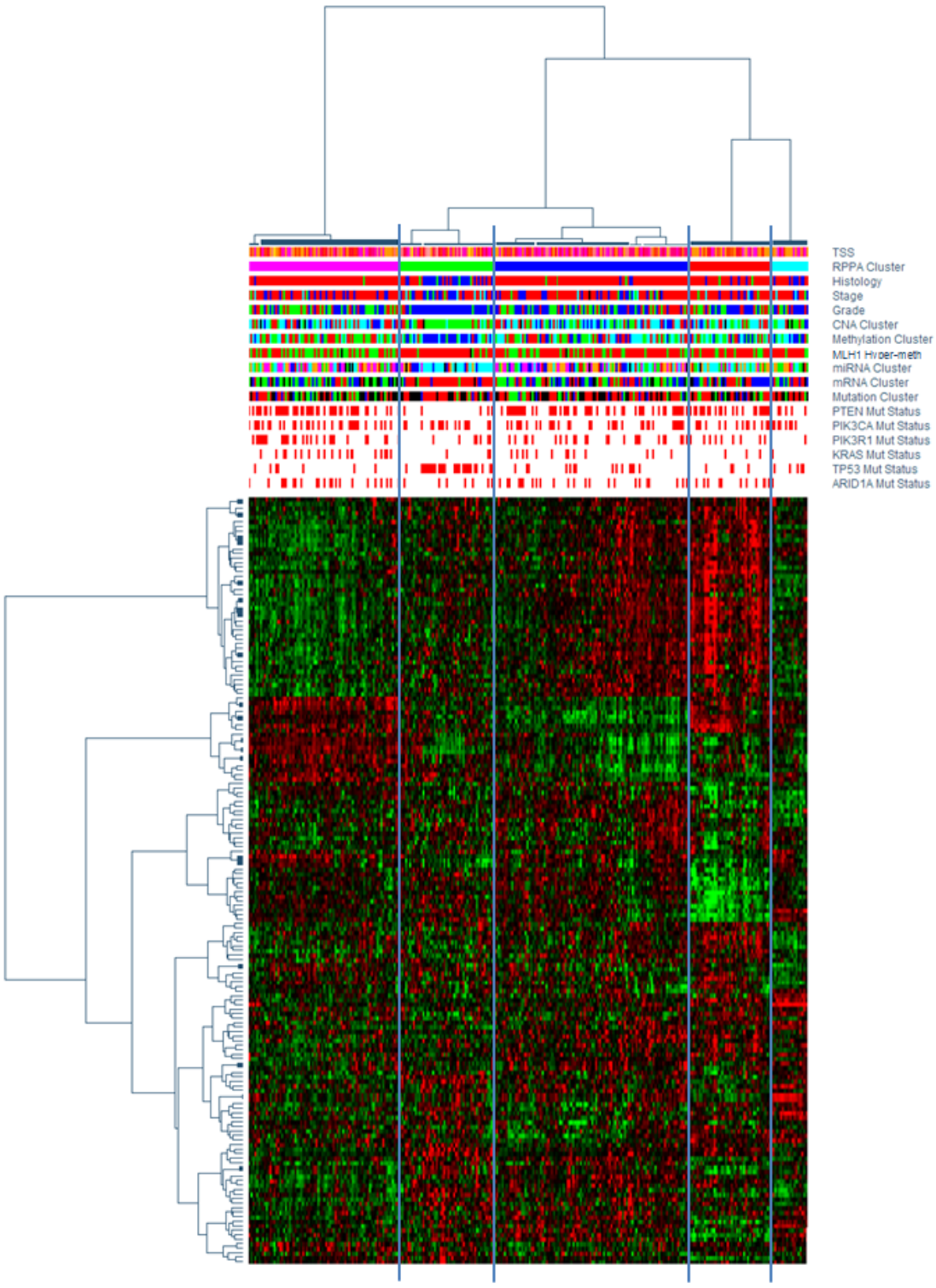


Figure S5.1 Unsupervised hierarchical clustering of 293 samples and 170 antibodies, showing 5 RPPA clusters. The green cluster (shown in second row from top) corresponds to the serous and serous-like samples (shown in third row from top). The heat map can be dynamically explored at:
http://bioinformatics.mdanderson.org/main/TCGA/Supplements/NGCHM-UCEC

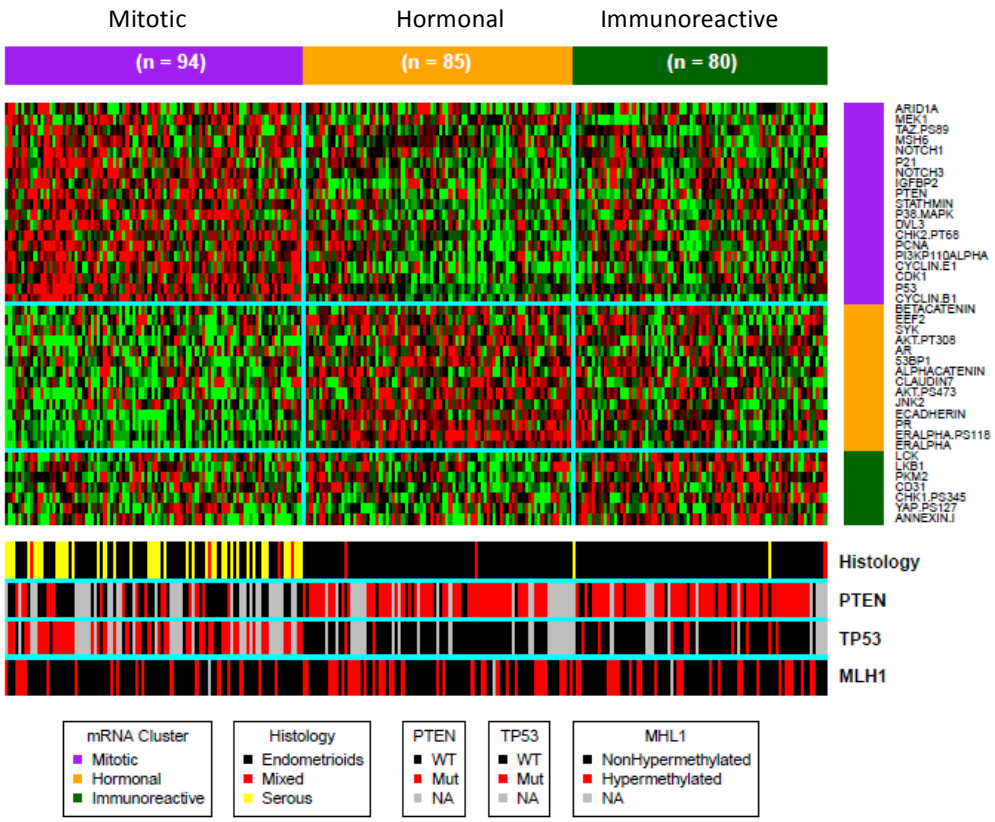


Figure S5.2 Supervised hierarchical clustering of RPPA data for the mRNA clusters shown in Fig. S4.1.

Data File S5.1 RPPA Antibody List

datafile.S5.1.RPPAAntibodyList.xls

Section References

1. Tibes, R. *et al*. 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 (2006).
2. Liang, J. *et al*. The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. *Nat Cell Biol* **9**:218-224 (2007).
3. Hu, J. *et al*. Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**:1986-1994 (2007).
4. Hennessy, B.T. *et al*. Pharmacodynamic markers of perifosine efficacy. *Clin Cancer Res* **13**:7421-7431 (2007)
5. Coombes, K. *et al*. SuperCurve: SuperCurve Package. R package version 1.4.1 (2011).
6. Gonzalez-Angulo, A. *et al*. Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clin Proteomics* **8**:11 (2011).
7. Hennessy, B. *et al*. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin Proteomics* **6**:129-151 (2010).
8. Neeley, E.S., Kornblau, S.M., Coombes, K.R., Baggerly, K.A. Variable slope normalization of reverse phase protein arrays. *Bioinformatics* **25**:1384-1389 (2009).

Supplementary Methods S6: miRNA sequencing

To identify compact lists of genes and miRNA mature or star strands that were differentially abundant between unsupervised groups of tumor samples, we calculated minimal sets of strands that allowed a classifier to discriminate samples in a group from all other samples.[1] Such a set was defined by a minimum in a profile of the out-of-bag (OOB) error as a function of the number of most-important variables, using an accuracy importance metric (Fig. S5.2). We note that discriminatory calculations were done between groups of tumor samples, rather than between tumor and normal samples.

Unsupervised consensus clustering[17] of miRNA-seq abundance profiles for 367 tumor samples suggested six sample groups (Supplementary Fig. 6.1). The consensus membership heatmap, per-group silhouette width profiles, and average silhouette widths of at least 0.84 suggested that the groups were distinct. Differences between groups were significantly associated with hypermethylated *MLH1*, histology, grade, and stage, but not overall or PFS. All groups except the more heterogeneous group 4 had few discriminatory miRNAs (Supplementary Fig. 6.2).

Unsupervised groups that were more homogeneous in the consensus membership heatmap and the silhouette width profile tended to have smaller sets of discriminatory miRNAs (e.g. groups 2, 5 and 6 vs. groups 3 and 4 in Fig. S5.2). A microRNA that was assigned a high classifier importance had an abundance distribution in a group that was distinct relative to all other samples (e.g. group 3 in Fig. S5.2). Regardless of the importance assigned by the classifier, discriminators (mature and star strands) that are more abundant are likely to be more influential in disease processes,[2] and we show abundance distributions as box-whisker plots only for the most highly-ranked and abundant discriminators.

Group 1's 34 samples were discriminated by the mature and star strands of miR-10b, and -503, -584, -34a and -361. The abundance of miR-10b's mature strand was comparable to adjacent normals for samples in this group, but was far lower in all other tumor groups. While the downregulated star strand (miR-10b*) has been reported as associated with cell cycle inhibition in breast tumors,[3] and this strand was downregulated in all tumor groups except group 1, its abundance was low in all samples (<60 RPM), and its importance in endometrial disease processes is uncertain.

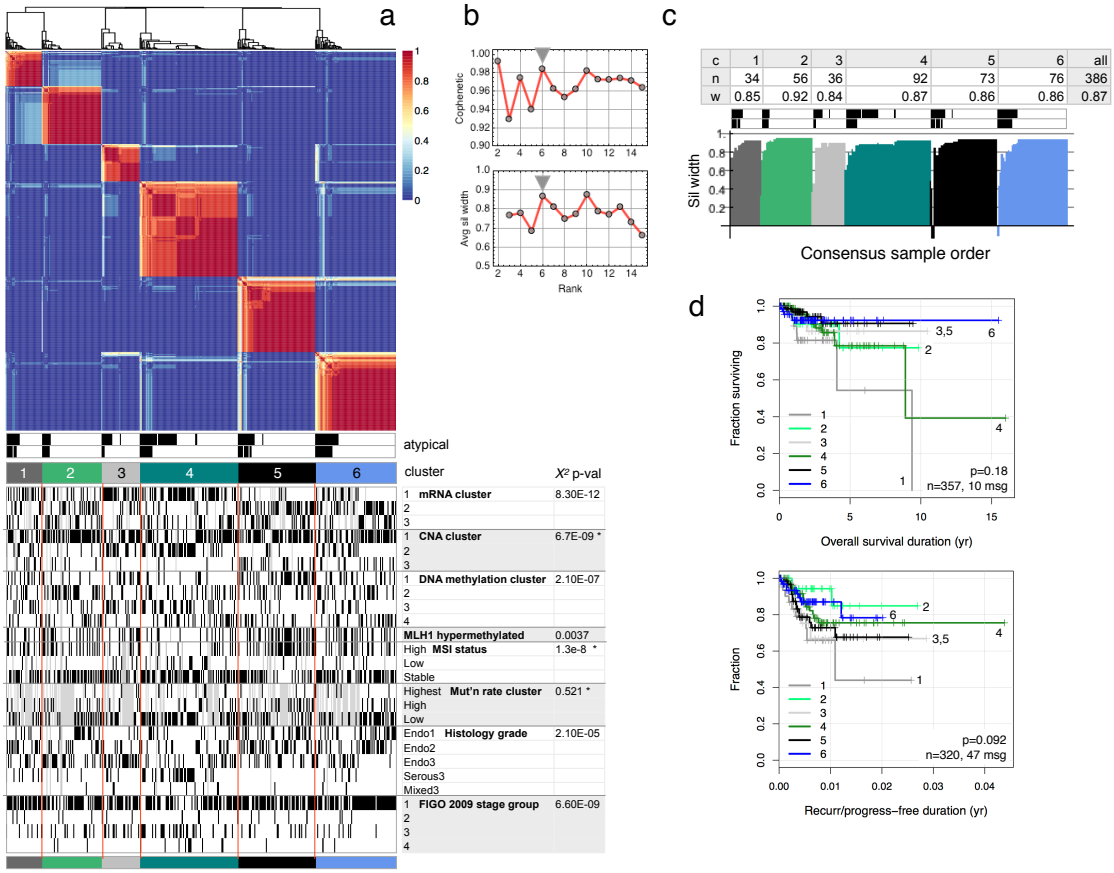
Cluster 2 had lower purity values than most other groups (Fig. S5.3c) and a relatively homogeneous consensus heatmap. The 56 samples in this cluster were discriminated by the mature strands of miR-143 and -1. Of these, only miR-143 was relatively abundant; its abundance was lower in all tumor groups than in tissue normals.

Cluster 3's 36 samples had high purity values but relatively inhomogeneous consensus membership values. It was discriminated by miR-9, -183 and -182, abundances of which were high relative to other tumor samples and to normals. The star strand of mir-9 was also discriminatory, but its abundance in this group was low (median ~100 RPM).

Cluster 4 consisted of 92 samples and contained most of the serous grade 3 tumors. Its purity was relatively low (Fig. S5.3c). Consistent with it containing a relatively large fraction of atypical samples, the cluster had a long list of 18 discriminators. The most important discriminators were the mature and star strands of miR-21; the star strand was relatively abundant in this cluster (~3000 RPM).

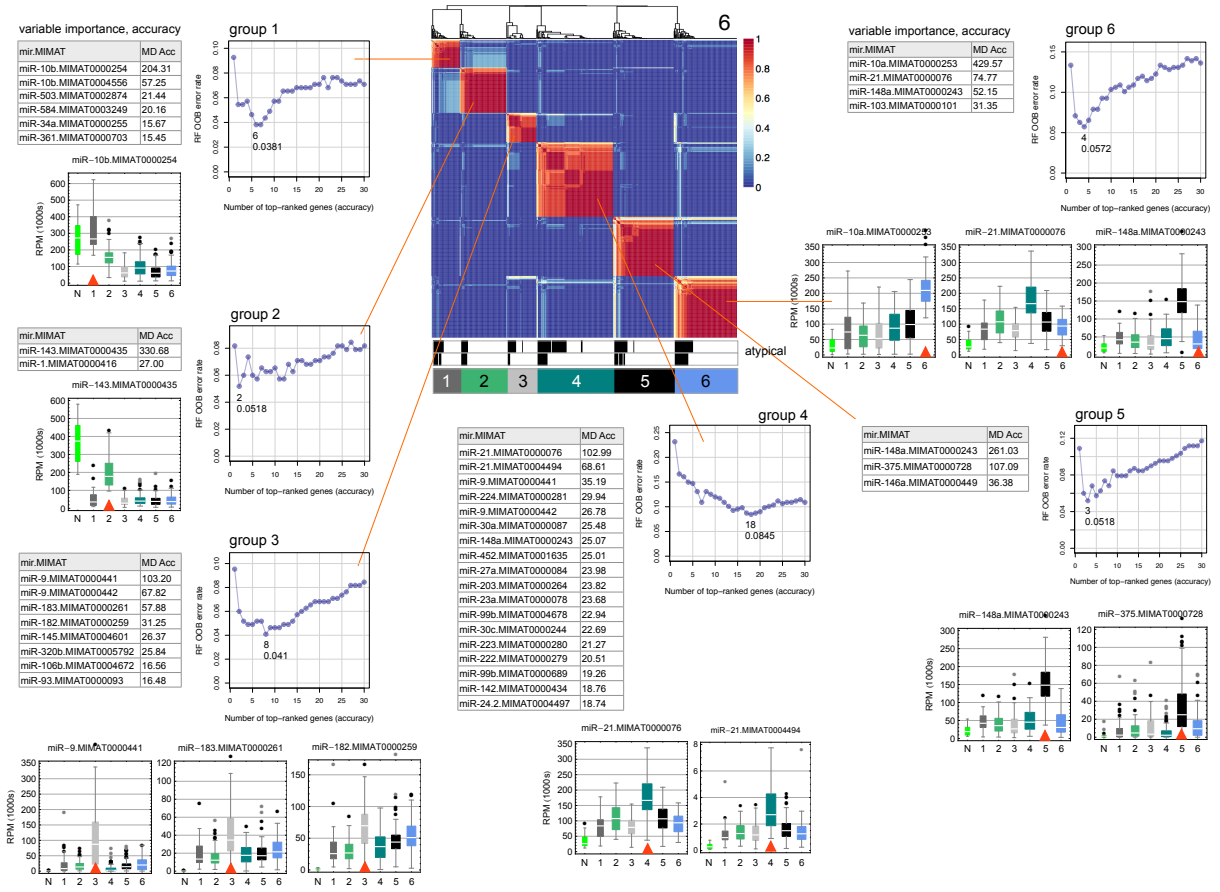
Cluster 5's 73 samples contained the highest proportion of hypermethylated MLH1 and high MSI. It was discriminated largely by the mature strands of miR-148a and -375, both of which were more abundant in this group than in other tumor groups or in normal tissue. Validated targets of miR-148a include CDNK1B, DNMT1 and DNMT3B.[4] Mir-375's pre-miRNA abundance was anticorrelated to RPPA data for cyclin E1 (CCNE1, r=-0.413, P=5x10^{-8}, explorer.cancerregulome.org).

Cluster 6's 76 samples were discriminated by miR-10a, -21, -148a and -103. The abundances of the first three were relatively high, with that of miR-10a the highest and most distinct. The abundance of the mir-10a pre-miRNA was anticorrelated to RPPA data for CDKN1A (r=-0.38, P = 6x10^{-8}, explorer.cancerregulome.org).

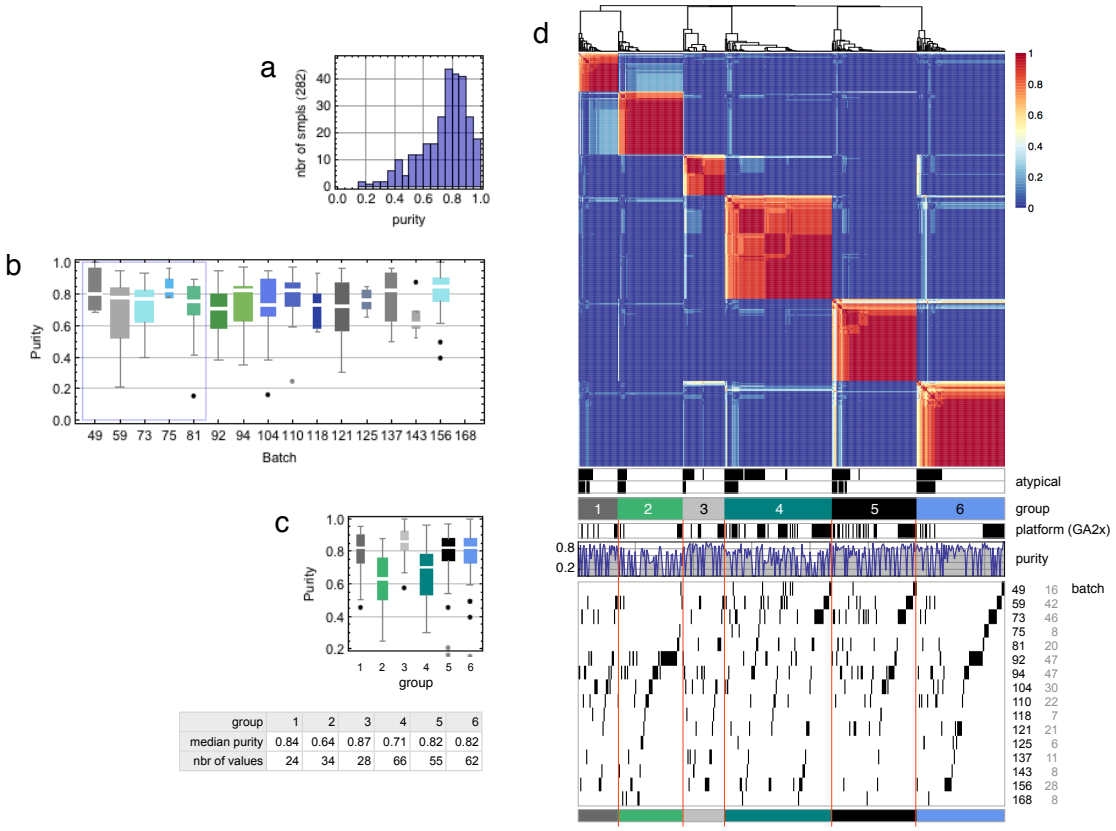


c	1	2	3	4	5	6	all
n	34	56	36	92	73	76	386
w	0.85	0.92	0.84	0.87	0.86	0.86	0.87

Supplemental Figure 6.1. NMF consensus clustering identified six sample groups. The NMF input was the normalized abundance (RPM) matrix for the 304 most-variant mature and star strands, for 367 tumor samples. **a)** Consensus membership heatmap for six clusters. Horizontal bands below the heatmap show (top to bottom) atypical members of each cluster (black) based on f=0.95 and f=0.90 per-cluster silhouette width thresholds, then covariates with Chi-square association *P* values. **b)** Profiles of cophenetic correlation coefficient and average silhouette width for solutions with 2 to 15 clusters, with the preferred six-cluster solution indicated by gray triangles. **c)** Summary group metrics (number of samples, average silhouette width), and silhouette width profile with samples in consensus heatmap order. **d)** Kaplan Meier curves for overall survival and recurrence/progression.



Supplemental Figure 6.2. Discriminatory miRNA mature or star strands used by a classifier to discriminate samples in a tumor group from all other tumor samples. Tables show the miRNA strands corresponding to the first minimum in a profile of the out-of-bag (OOB) error as a function of the number of most-important variables, using an accuracy metric.[2] Box-whisker plots show per-group RPM abundances for the most highly-ranked and abundant miRNAs, and include 19 tissue normals. MicroRNA names are a base name with a miRBase v16 MIMAT ID that is specific to either a 5p or a 3p strand.



group	1	2	3	4	5	6
median purity	0.84	0.64	0.87	0.71	0.82	0.82
nbr of values	24	34	28	66	55	62

Supplemental Figure 6.3. Relationship of unsupervised clusters to sequencing platform, tumor purity and BCR batch number. **a,b)** Distribution of tumor purity from SNP6 data: **a)** for 282 samples, and **b)** by BCR batch. The blue box shows samples sequenced on GAIIx systems. **c)** Distribution and summary table for purity as a function of unsupervised groups. **d)** NMF heatmap showing consensus membership values, with sequencing platform, purity profile and BCR batches indicated.

Section References

1. Biagioni, F. *et al*. miR-10b*, a master inhibitor of the cell cycle, is down-regulated in human breast tumours. *EMBO Mol Med* **4**:1214-1229 (2012).
2. Mehrian-Shai, R. *et al*. Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. *Proc Natl Acad Sci U S A* **104**:5563-5568 (2007).
3. Mullokandov, G. *et al*. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat Methods* **9**:840-846 (2012).
4. Xiao, F. *et al*. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**:D105-110 (2009).

Supplementary Methods S7: DNA methylation

Array-based DNA methylation assay

We used two Illumina Infinium DNA methylation platforms, HumanMethylation27 (HM27) BeadChip and HumanMethylation450 (HM450) BeadChip (Illumina, San Diego, CA) to obtain gene promoter and gene body DNA methylation profiles of 373 TCGA endometrial cancer samples and 27 adjacent non-tumor endometrial tissue samples. The Infinium HM27 array targets 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36). The Infinium HM450 array targets 482,421 CpG sites and covers 99% of RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. It covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them. The assay probe sequences and information on each interrogated CpG site on both Infinium DNA methylation 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.[1] All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified (WGA) and enzymatically fragmented prior to hybridization to the arrays. BeadArrays were scanned using the Illumina iScan technology, and the IDAT files (Level 1 data) were used to extract the intensities (Level 2 data) and calculate the beta value (Level 3 data) for each probe and sample with the R-based methylumi package.

The level of DNA methylation at each CpG locus is summarized as 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 2 and 3 in HM27 and Level 3 in HM450 data packages, as detailed below.

TCGA data packages

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

HM27: *Level 1* - Level 1 data packages contain the non-background corrected signal intensities of the M and U probes and the mean negative control cy5 (red) and cy3 (green) signal intensities. A detection *P* value for each data point, the number of replicate beads for M and U

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

HM450: *Level 1* - Level 1 data contain raw IDAT files. IDAT files are the direct output from the scanning program. *Level 2* - Level 2 data contain background corrected signal intensities of the M and U probes. *Level 3* - Level 3 data files contain β-value calculations and masked data points with "NA" from the probes that are annotated as having a SNP within 10 base pairs of the interrogated locus (HM27 carryover or recently discovered). The genomic characteristics for each probe are available for download via Illumina (www.illumina.com).

Unsupervised clustering analysis of DNA methylation data

The shared probe set between HM27 and HM450 platforms (N=25,978) were used for this analysis. We removed probes that contained any masked data due to detection *P* value, repeats and SNPs and non-uniquely mapped probes (n=22,071 remaining). We observed batch and platform specific effects. To alleviate systematic platform-specific effects (dye bias, background level, etc) we fit a LOESS regression model between the two platforms using M values, stratified by the number of CpGs in the probe (CpG=1,2,3,4,5,6+), and normalized the HM450 data against the HM27 data. M value is the log 2 ratio of Methylated (M) intensity and Unmethylated (U) intensity and better satisfies the linearity assumption. In order to further filter out probes with high technical variances, we applied a two-way nested ANOVA for platform and batch effects with batch nested in platform (M value ~ Platform + Platform/Batch) and removed probes with above-median F value for either platform or batch. We then selected probes with standard deviation of >1.8 (n=785 probes) based on M values for unsupervised clustering. Beta values were used for clustering with a mixture model based method, RPMM (recursively partitioned mixture model for Beta and Gaussian Mixtures) well suited for beta-distributed DNA methylation measurements.[2] We performed RPMM clustering on the above-mentioned 785 probes for the 373 tumor samples with beta mixture model. A fanny algorithm (a nonparametric clustering algorithm) was used for initialization and level-weighted version of Bayesian information criterion (BIC) as a split criterion for an existing cluster as implemented in the RPMM package. The clustering result was visualized with a modified version of

heatmap.plus, with samples within each cluster group seriated by hierarchical clustering. The statistical analysis was done in R.

Unsupervised clustering of DNA methylation data from the 373 endometrial tumor samples revealed four unique DNA methylation subtypes (MC1-4), typified by a heavily methylated subtype (MC1) reminiscent of the CpG island methylator phenotype (CIMP) phenotype described in colon and glioblastoma,[18-20] and by a serous-like cluster (MC3) composed primarily of serous tumors (Supplementary Fig. 7.1). The CIMP phenotype was associated with the MSI phenotype, attributable to promoter hypermethylation of *MLH1*. This association was similar to the colorectal CIMP but not the glioblastoma CIMP, suggesting a potential shared mechanism for epithelial CIMP tumors. MC2 tumors exhibited relatively high cancer-specific DNA methylation levels, only lower than MC1. DNA hypermethylation observed in MC4 was the lowest among the non-serous-like tumors. MC3, the serous-like cluster, had minimal DNA methylation changes compared to normal endometrium.

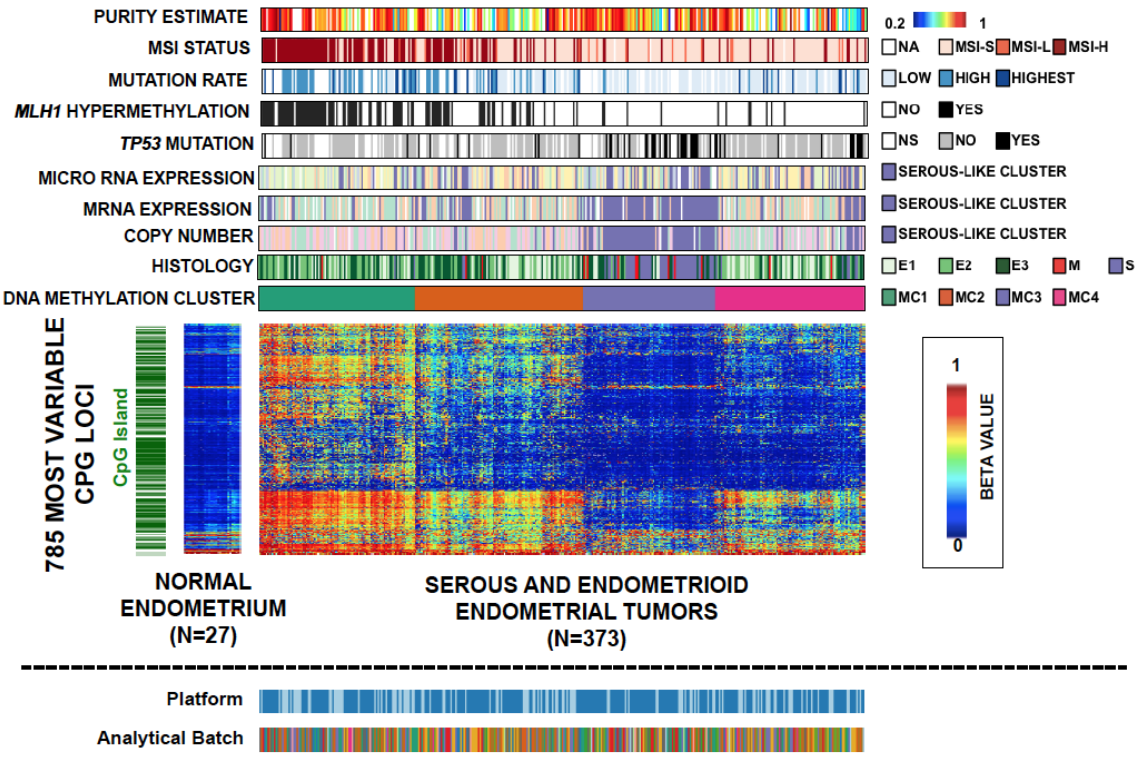


Figure S7.1. Unsupervised clustering of the DNA methylation data reveals four subtypes. A spectrum of blue to red in the heatmap indicates low to high DNA methylation (0% to 100%). Four DNA methylation subtypes among the 373 tumors (column) are visualized for 785 CpG loci (row) used for the clustering. Column-side color bars indicate different features of each sample, the bottom of which shows grouping of the samples as determined by RPMM (recursively partitioned mixture model for Beta and Gaussian Mixtures). Within each cluster the samples are seriated by hierarchical clustering. 27 normal endometrium samples are also plotted for the same loci for comparison. Platform (light blue – HumanMethylation27; dark blue – HumanMethylation450) and analytical batch for each sample are plotted on the bottom to make sure that the clustering results are not driven by technical variations.

Section References

1. Campan, M., Weisenberger, D.J., Trinh, B., Laird, P.W. MethyLight. *Methods Mol Biol* **507**:325-337 (2009).
2. Houseman, E.A. *et al*. Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* 9:365 (2008).

Supplementary Methods S8: Integrative clustering using iCluster

Integrative clustering of somatic mutation, DNA copy number, DNA methylation, and mRNA expression data was performed using the iCluster framework originally described in Shen et al..[1] The problem is formulated as a joint multivariate regression of multiple data types with respect to a set of common latent variables that represent the underlying tumor subtypes. A penalized likelihood approach was used with lasso[2] penalty terms for balancing the fitness and the complexity of the model. We applied an extended algorithm that generalizes the original method to encompass both discrete and continuous data types using the generalized linear model framework (Shen et al, manuscript submitted). In brief, for the binary mutation data matrix, we assume each entry is a realization of a Bernoulli random variable x_{ij} associated with the *j*th gene in the *i*th sample, with marginal mutation probability π_{ij} and the logit function as its canonical link to equate to the latent variables. For data types that are on a continuous scale (copy number, methylation, and mRNA expression data), a Gaussian distribution with identity link function was used. Data processing procedures for iCluster analysis is performed as described in the TCGA squamous cell lung cancer study.[3]

Model selection

The number of clusters (K) is unknown and needs to be estimated. We compute a deviance ratio metric which can be interpreted as the percentage of variation explained by the current model, and K is chosen to maximize the deviance ratio. To determine the optimal combination of the lasso penalty parameter values, a very large search space needs to be covered. We used an efficient sampling method that utilizes the uniform design (UD).[4] A theoretical advantage of the uniform design over an exhaustive grid search is the uniform space filling property that avoides wasteful computation at close-by points. For each sampled “experimental” point, we fit an iCluster model with the sampled parameter setting. The best parameter setting was chosen that minimizes the Bayesian information criterion (BIC).

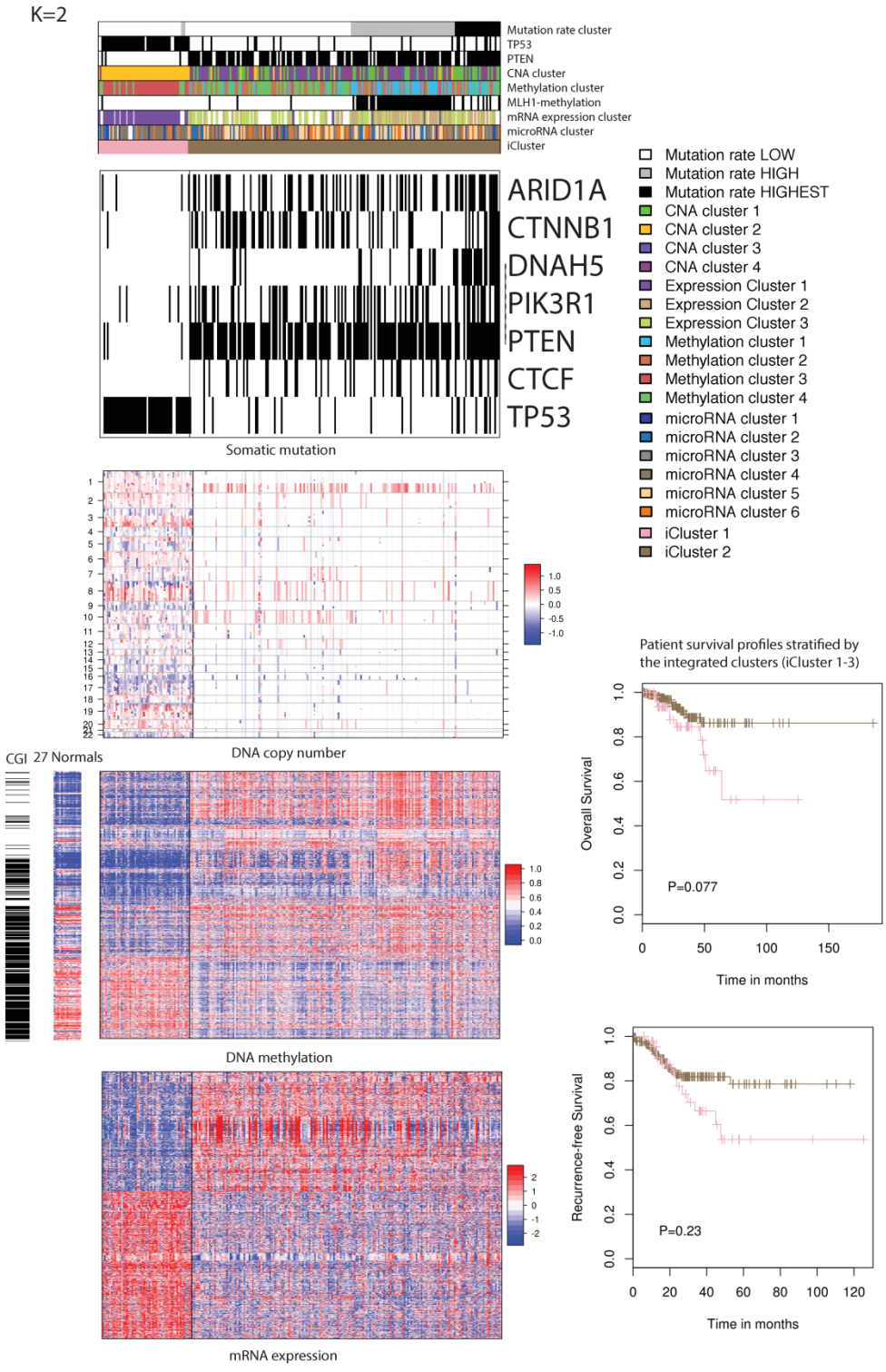

K=2
Mutation rate cluster
TP53
PTEN
CNA cluster
Methylation cluster
MLH1-methylation
mRNA expression cluster
microRNA cluster
iCluster
ARID1A
CTNNB1
DNAH5
PIK3R1
PTEN
CTCF
TP53
Somatic mutation
Mutation rate LOW
Mutation rate HIGH
Mutation rate HIGHEST
CNA cluster 1
CNA cluster 2
CNA cluster 3
CNA cluster 4
Expression Cluster 1
Expression Cluster 2
Expression Cluster 3
Methylation cluster 1
Methylation cluster 2
Methylation cluster 3
Methylation cluster 4
microRNA cluster 1
microRNA cluster 2
microRNA cluster 3
microRNA cluster 4
microRNA cluster 5
microRNA cluster 6
iCluster 1
iCluster 2
DNA copy number
CGI
27 Normals
DNA methylation
mRNA expression
Patient survival profiles stratified by the integrated clusters (iCluster 1-3)
Overall Survival
P=0.077
Time in months
Recurrence-free Survival
P=0.23
Time in months

Figure S8.1. Integrative Clustering. iCluster reveals two distinct molecular subgroups. Heatmap displays coordinated patterns of alteration in somatic mutation, DNA copy number, DNA methylation, and mRNA expression.

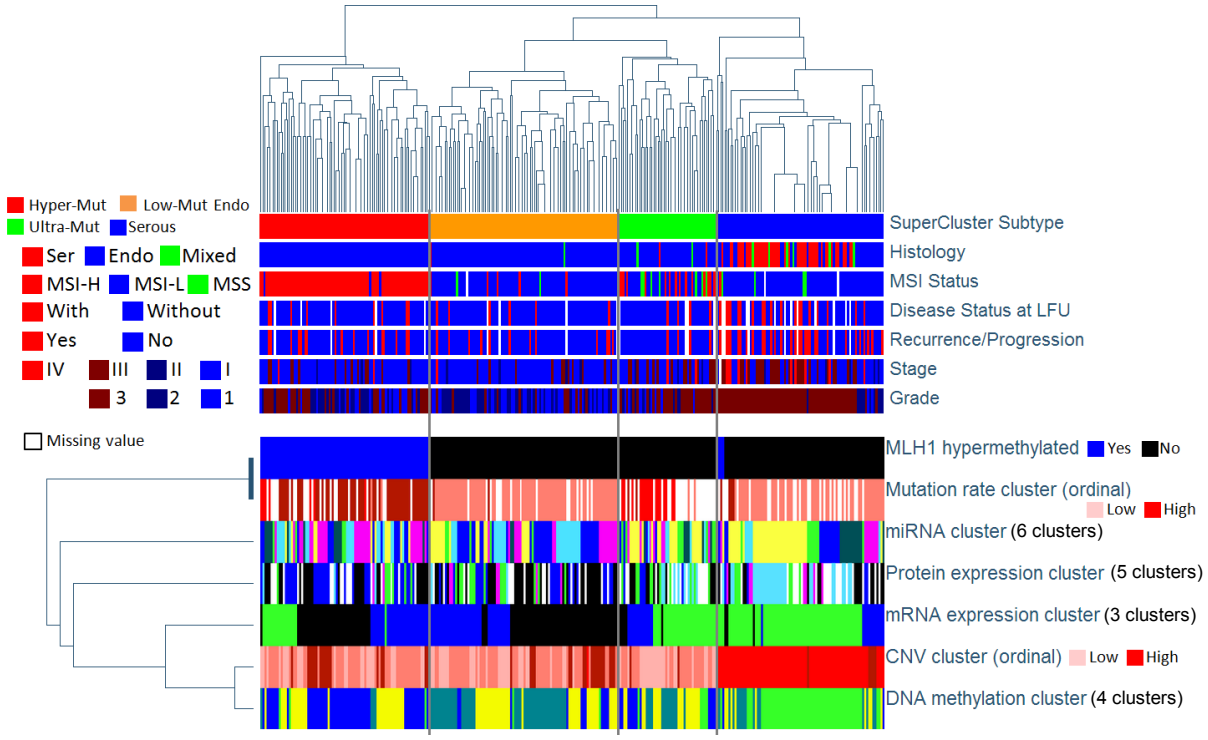
Section references

1. Shen, R., Olshen A.B., 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* **25**:2906-2912 (2009).
2. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**:267-288 (1996).
3. TCGA network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**:519-525 (2012).
4. Fang, K. T., & Wang, Y. Number-theoretic methods in statistics. 1st ed. Monographs on statistics and applied probability. London ; New York: Chapman & Hall. xii, 340 (1994).

Supplementary Methods S9: Super Clusters

We developed a new clustering algorithm, called SuperCluster, to derive overall subtypes for the samples based on their cluster memberships of different data types (mRNA, protein, CNV, etc.). The algorithm adjusted the contribution from each data type so that their weights were equal. Mutation and CNV clusters were treated as ordinal variables, whereas the others were treated as nominal. The results are shown in Fig. S11.1 where four super clusters can be seen. The hyper-mutator super cluster (red) is characterized by hyper mutator samples that have MLH1 silenced, high overall DNA methylation, and MSI-high status. The low mutator endometrioid super cluster (orange) is characterized by low mutation rates and few CNVs. The ultra-mutator super cluster (green) has ultra-mutator samples that don't have MLH1 silenced and have a mixture of MSI low and high samples. The serous super cluster (blue) has low mutation rates, very high CNVs, and enrichment of specific mRNA, miRNA, protein and DNA methylation subtypes. It has most of the serous samples and tends to be high grade and stage. It also has a high rate of recurrence and poor disease related outcome.

(a)



(b)

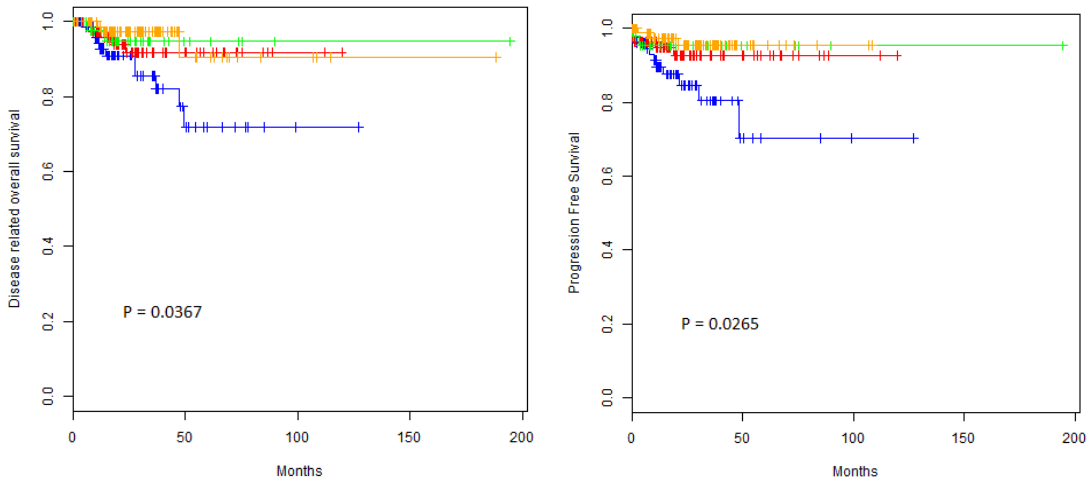


Figure S9.1. (a) Top panel: Overall super clusters derived from clusters on individual data types. The columns contain samples. The rows contain the cluster memberships of different data types. The annotation bars on top of the heat map show clinical associations of the super clusters (not used to derive the clusters). The heat map can be explored dynamically at: http://bioinformatics.mdanderson.org/main/TCGA/Supplements/NGCHM-UCEC
(b) Bottom panel: Disease related overall survival (left) and progression free survival (right) showing the serous super cluster has poor outcome.

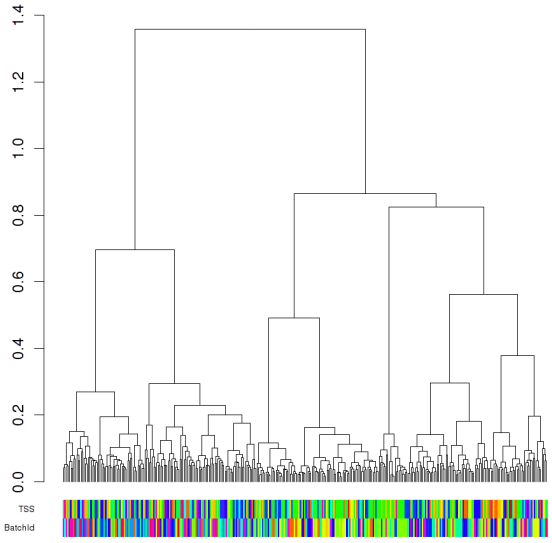
Supplementary Methods S10: Batch effects analysis

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the TCGA Uterine Corpus Endometrioid carcinoma (UCEC) data sets. Four different data sets were analyzed: mRNA seq (Illumina GA RNA Seq), miRNA seq (RNA-seq Illumina HighSeq), DNA methylation (Infinium HM27K and Infinium HM450K microarray), and SNP (GW SNP 6). All the data sets were at level 3, since that is the level at which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS). Two different algorithms were used; hierarchical clustering and PCA. For hierarchical clustering, we used the average linkage algorithm with 1 minus Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or TSS.

For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results of the analysis are show in Figs. 1-12. The results can also be analyzed dynamically online at http://bioinformatics.mdanderson.org/tcgabatcheffects/

mRNA Seq – Illumina GA RNA-Seq

Figures S10.1-S10.3 show clustering and PCA plots for the mRNA seq data (Illumina GA RNA-Seq platform). The plots show that the batches and tissue source sites are well mixed, indicating that batch effects are negligible.



Legends

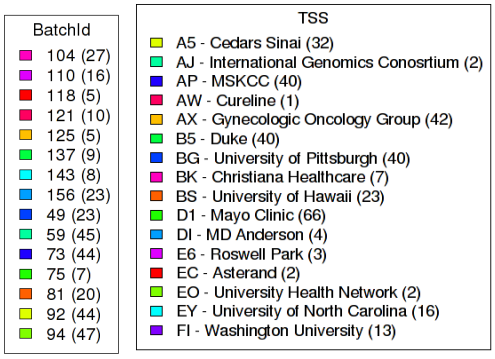


Figure S10.1. Hierarchical clustering plot for mRNA Sequencing (Illumina GA RNA-Seq).

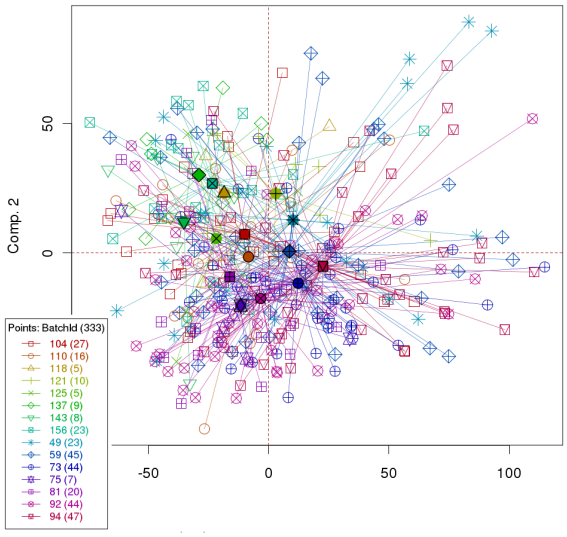


Figure S10.2. PCA: First two principal components for mRNA Sequencing (Illumina GA RNA-Seq), with samples connected by centroids according to batch ID.

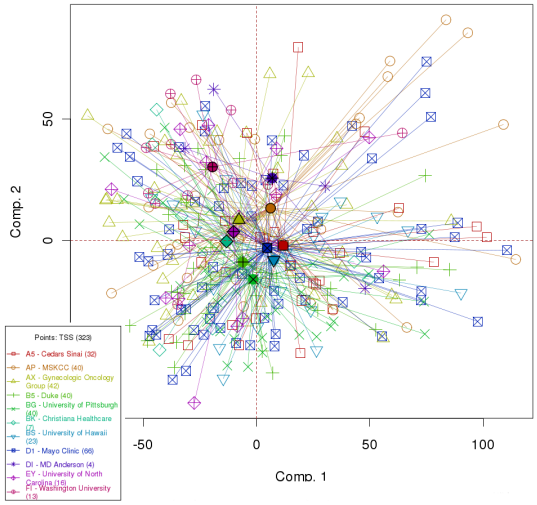
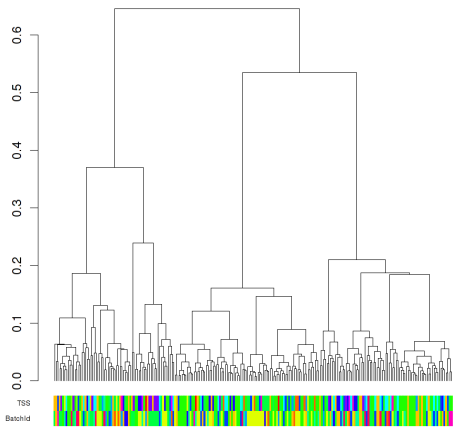


Figure S10.3. PCA: First two principal components for mRNA Sequencing (Illumina GA), with samples connected by centroids according to TSS.

miRNA expression sequencing (RNA-seq Illumina HighSeq)

Figures S10.4-S10.6 show the clustering and PCA plots for the miRNA expression sequencing platform (RNA-seq Illumina HighSeq). The results once again show that batch effects are negligible by both, batch ID and TSS.



Legends

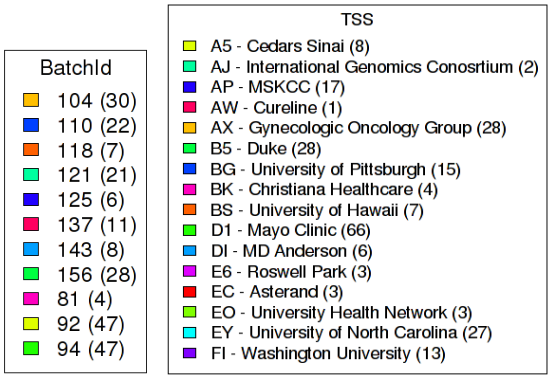


Figure S10.4. Hierarchical clustering for miRNA- seq (RNA-seq Illumina HighSeq)

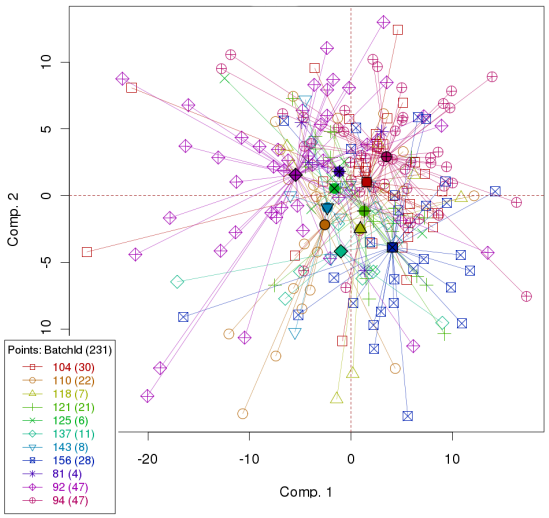


Figure S10.5. PCA: First two principal components for miRNA seq (RNA-seq Illumina HighSeq), with samples connected by centroids according to batch ID.

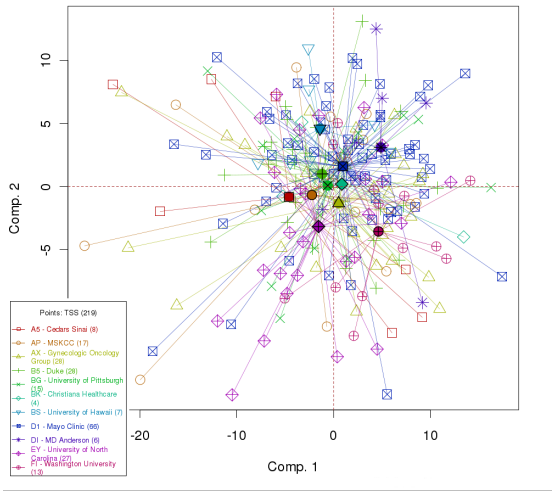
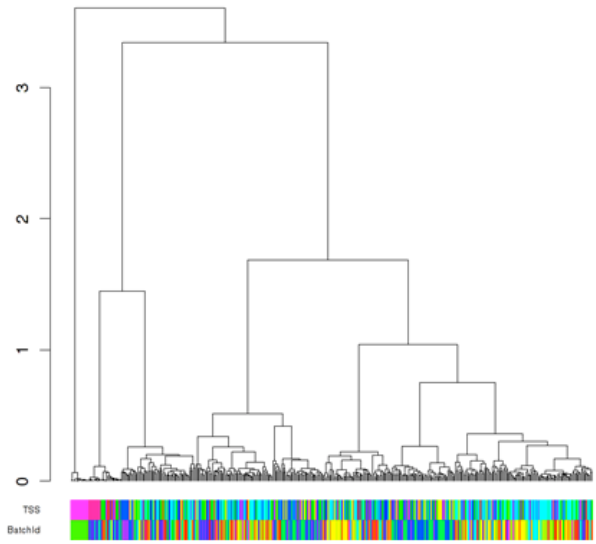


Figure S10.6. PCA: First two principal components for miRNA seq (RNA-seq Illumina HighSeq), with samples connected by centroids according to TSS.

DNA methylation (Infinium HM27K and 450K microarray)

Figures S10.7-S10.9 show clustering and PCA plots for the DNA methylation platforms (Infinium HM27K and 450K microarray). The results show that about 14 samples stand out whose batch IDs or tissue source sites were not available at the time the analysis were done. However, it was unlikely that all 14 came from the same batch or TSS, so we didn't think that batch effects correction was warranted. Besides those samples, the batches were well mixed and no major batch effects were seen.



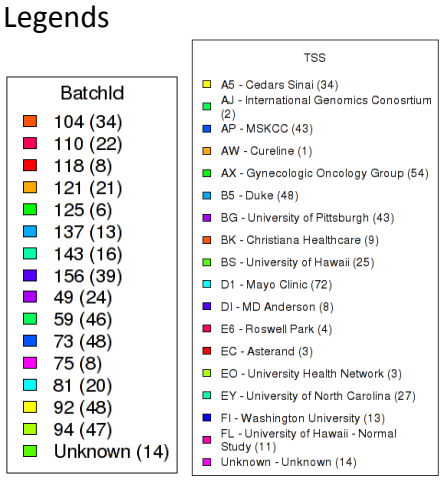


Figure S10.7. Hierarchical clustering for DNA methylation data (HM27K and 450K)

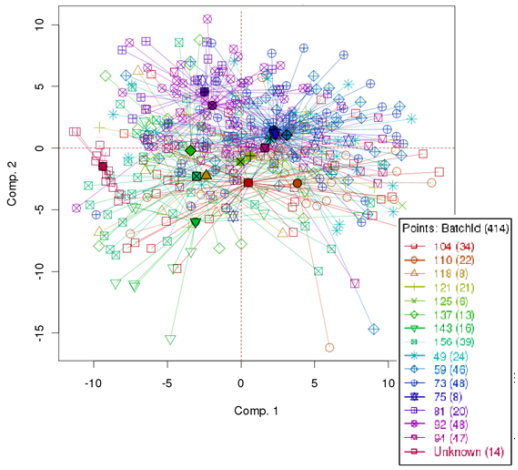


Figure S10.8. PCA: First two principal components for DNA methylation data with samples connected by centroids according to batch ID.

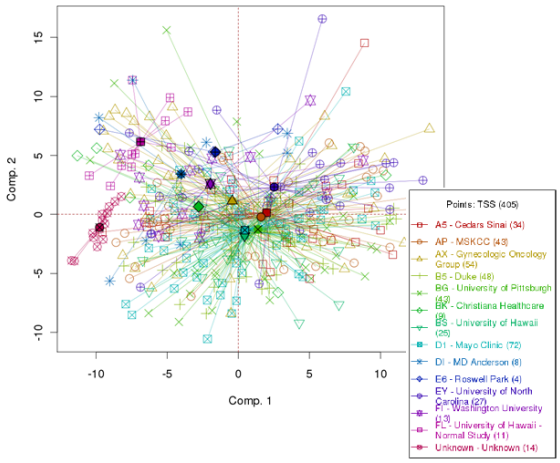
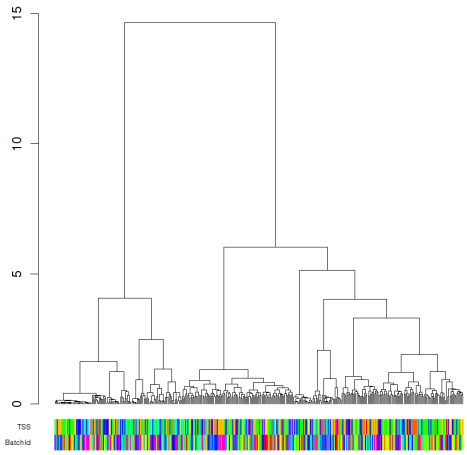


Figure S10.9. PCA: First two principal components for DNA methylation data with samples connected by centroids according to TSS.

SNPs (GW SNP 6)

Figures S10.10-S10.12 show clustering and PCA plots for the SNP platform. At level 3, the TCGA SNP data resembles copy number data when we use chromosomal segment counts (rather than actual SNPs). We mapped the chromosomal segments to genes and then used them to construct the plots shown in the figures. The copy number data has a lower limit when it comes to deletions, but copy number gain can potentially have high values, which is why the points seem skewed in the PCA plots. However, we can see from the plots that the batches and TSSs are well mixed, with negligible batch effects.



Legends

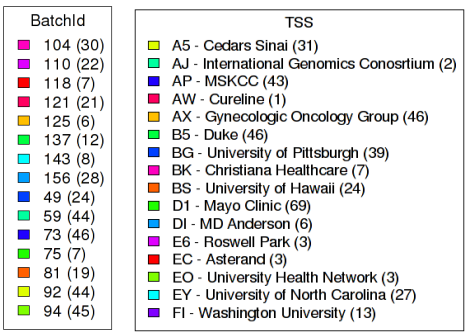


Figure S10.10. Hierarchical clustering plot for SNP data.

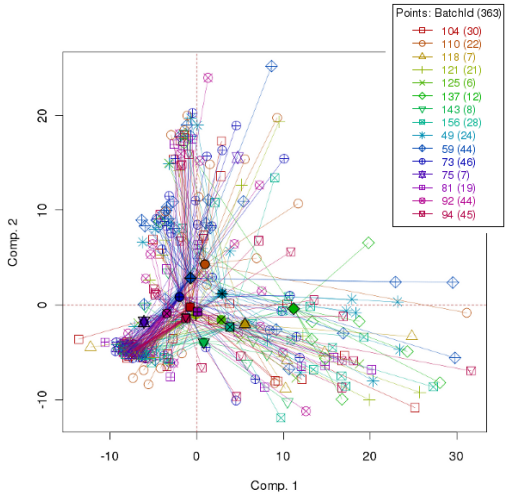


Figure S10.11. PCA for SNPs, with samples connected by centroids according to batch ID.

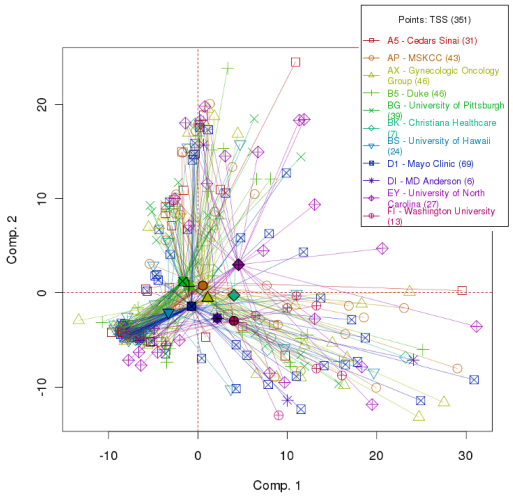


Figure S10.12. PCA for SNPs, with samples connected by centroids according to TSS.

Conclusions

We tested batch effects in four different platform types; mRNA-seq, miRNA-seq, DNA methylation, and SNP. None of the platforms showed any major batch effects by either batch ID or tissue source site.

Supplementary Methods S11: Pathways and integrated analyses

PARADIGM integrated pathway analysis of copy number and expression data

Integration of copy number, mRNA expression and pathway interaction data was performed on the 324 samples using the PARADIGM software.[1] Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. Expression and gene copy number data was obtained from the Endometrial Data Snapshot page (http://23.23.224.116/ucec_tcga/). The mRNA data was converted to relative mRNA expression levels by taking the log2 ratio of each gene in each sample to the gene's median computed over 5 normal controls. Data was rank transformed and discretized prior to PARADIGM analysis.

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov and the Reactome database from http://reactome.org. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and henceforth referred to collectively as pathway concepts. The resulting pathway structure contained a total of 17151 concepts, representing 7111 proteins, 7813 complexes, 1574 families, 52 RNAs, 15 miRNAs and 586 processes.

The PARADIGM algorithm infers an integrated pathway level (IPL) for each gene that reflects a gene's activity in a tumor sample relative to the normal controls. Including only pathway concepts with relative activities distinguishable from normal (0.05 absolute activity) in at least one patient sample, non-zero activity in at least 10% of the samples and showing variation between samples (variance > 0.05) yielded over 10,000 concepts. To identify patient subtypes implicated from shared patterns of pathway inference, we ran Consensus Clustering using the median-centered IPLs implemented with the ConsensusClusterPlus package in R [http://www.R-project.org] with 80% subsampling over 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric (Figure S11.1).

Consensus clustering of ~10K varying IPLs yielded 5 PARADIGM clusters with distinct pathway activation patterns and significant associations with subtypes obtained from other platforms. Cluster 1, showing the lowest FOXA1/ER and MYC signaling, appears associated with the immunoreactive expression subtype. Cluster 3, with relatively high MYC but low FOXA1/ER signaling, is comprised almost entirely of High CN and proliferative cases (44/49). The remaining large cluster, Cluster 5, shows high MYC, HIF1 and FOXA1 signaling, and is relatively enriched in the hormonal mRNA subtype.

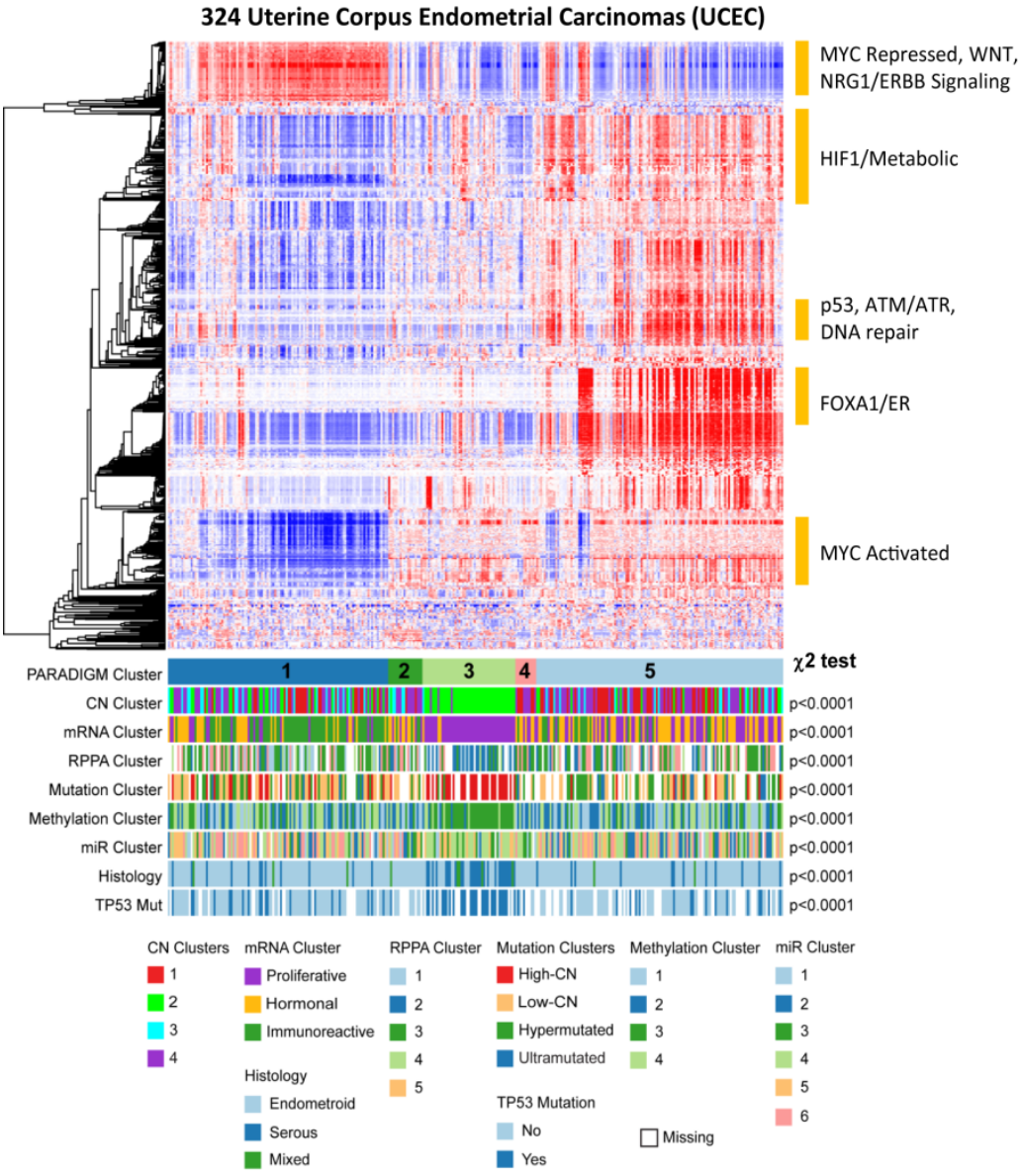


Figure S11.1. Heatmap display of top 1000 varying pathway features within PARADIGM consensus clusters. Samples were arranged in order of their consensus cluster membership. The copy number, mRNA, RPPA, mutation, methylation, and microRNA cluster membership assignments, histology and TP53 mutation status for each sample are displayed below. For each variable, the *P* value from the χ2 test of associations with consensus clusters was displayed. Selected pathways showing distinct activation patterns among the consensus clusters were labeled (orange bar).

Histologically serous endometrial cancers share common pathway activation with the CN-High subtype and p53 mutation.

We employed the Differential Pathway Signature Correlation (DIPSC) algorithm to assess the correlation between phenotypes such as mutations, subtype, and histology. The DIPSC method provides us with a measure of the relationship of one phenotype to another by comparing a phenotype signature – a vector of values associated with PARADIGM IPLs. We exclude all non-gene IPLs (such as processes) and further filter the data by a more stringent variance filter (standard deviation > 0.05). For each phenotype, we dichotomize the samples into a phenotypic set, for example serous histology, and its complement, and derive a phenotype signature by Two-Set SAM analysis.[2] The DIPSC method accounts for sample overlap by performing a bootstrap analysis that randomly assigns samples into independent cohorts. Pairwise Pearson R correlation is performed between the cohorts. The bootstrap process is repeated 1000 times to create an estimate of the mean and standard deviation statistic and determine a *P* value, which must be lower than 0.01 for inclusion in this study. The final signature vector is the mean. The final correlation of correlation figure is then assembled using Cluster 3.0 and visualized with Java TreeView (Figure S11.2A).

DIPSC analysis subsets significantly mutated genes and tumor phenotypes into three major groups based on the correlation between phenotypic signatures. These pathway signature correlation groups correspond to distinct mutation subtypes (Group 0: CN-High, Group 1: Non, Group 2: Hyper and Ultra mutated subtypes). Of note, the phenotypic signature of histologically serous endometrial cancers clusters with that of the CN-High subtype and p53 mutation; and this group (Group 0) appears very distinct from other endometrial cases (Group 1 and 2). It is also interesting to note that the correlation profile of the CN-High phenotypic signature suggests it associates primarily, but strongly, with only a small number of mutations (p53: Rp = 0.7, FBXW7: Rp = 0.425), which is in contrast to the other mutation subtypes, showing a consistent (but lower) correlation with a larger number of mutations (Figure S11.2B). We speculate here that the pathway correlations between p53 and FBXW7 mutations within the CN-High subtype may be of significant functional consequence, as murine models has implicated FBXW7 as fail-safe mechanism against tumorigenesis in a p53 deficient background with potential links to genomic instability.[3-5] However, further studies are needed to evaluate if FBXW7 in conjunction with p53 mutations play a role in the etiology of the CN-High endometrial cancers.

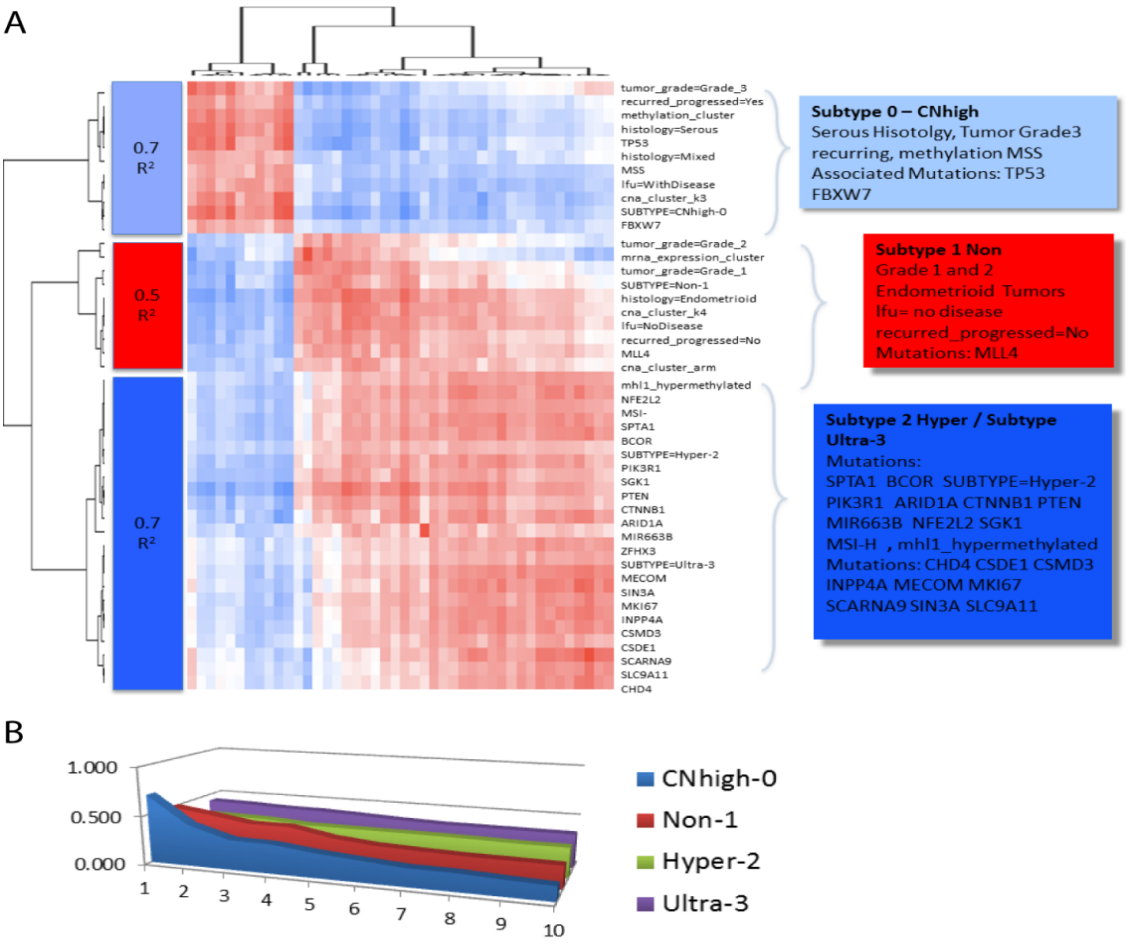


Figure S11.2. DIPSC analysis demonstrates serous endometrial, CN-High subtype, and p53 mutation shares similar phenotypic signatures. (A) Correlation of Differential Pathway Signatures. Phenotypes of uterine endometrial cancer cluster into three groups that align with mutation subtypes as described, but the hypermutator and ultramutator subtypes have stronger within group correlations than between group correlations. (B) Top 10 Differential pathway correlations of mutations with subtype. The graph shows the correlation profile of each subtype's differential pathway signature (blue: CN-High, red: Non, green: hypermutated, purple: ultra-mutated) with its top 10 correlated mutations.

Pathway-based biomarkers of CN-High versus Other subtypes.

IPLs differentially activated between the CN-High and the other subtypes (Ultramutator, Hypermutated, Non) were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg (BH) FDR correction. Only features deemed significant (FDR corrected $P < 0.05$) by both tests were selected. Pathways enriched among differentially activated IPLs were assessed using the EASE score with BH FDR correction; and sub-networks were constructed to identify regulatory hubs based on interconnectivity and visualized using Cytoscape (Figure S11.3).

~4.2K IPLs were found to be significantly differentially activated between the CN-High vs. other subtypes. Pathway enrichment and subnetwork analysis independently implicated p53 and XBP1 signaling as major hubs showing differential activation in the CN-High subtype relative to other UCEC cases. Lower activation of p53 and FOXA1/ER/XBP1 signaling are observed among the CN-High cases. These observations are in line with the high *TP53* mutation frequencies (55/60) and low fraction of hormonal expression cluster members (1/60) within the CN-High subtype.

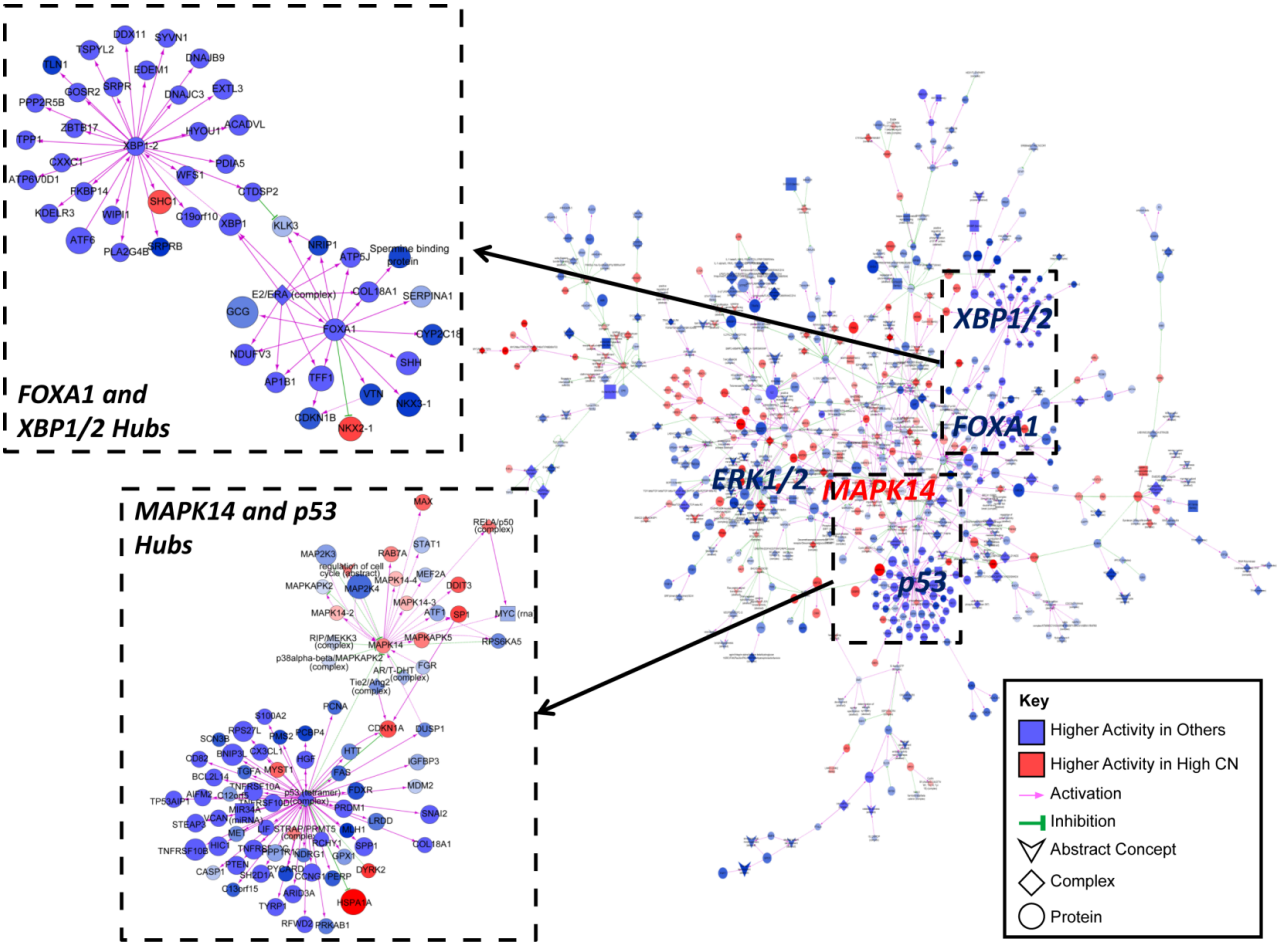


Figure S11.3. Differentially activated pathway features between CN-High and others UCEC subtypes. Largest interconnected regulatory subnetwork of differentially activated IPLs is displayed, with network hubs showing interconnectivity > 15 edges labeled. A zoomed in view of the p53, MAPK14 and XBP1/2 and FOXA1 hubs are also shown. Color intensity reflects activity differences between subtype (red: higher in CN-High, blue: higher in Others). Purple arrows denote activation. Green tees represent inhibition. Node shapes reflects pathway concept type (inverted v: abstract concept, diamond: complex, circle: protein). Node size is scaled to the significance of differential activation.

***TP53* truncating, but not missense, mutations are implicated as loss-of-function mutations by evaluating the discrepancy between up- and down-stream pathway signals.**

As p53 appears to be the key regulatory hub down-regulated in CN-High endometrial cancers, we employed the PARADIGM-SHIFT algorithm[6] to compare the pathway impact of truncating and missense mutations. Samples were defined to have truncating mutations if they were annotated with insertions, deletions, nonsense, or splice site mutations in *TP53*. Alternatively, samples are defined to have missense mutations if they were annotated with missense mutations in *TP53*. Under these definitions, there were 180 samples with available copy number and expression data to run PARADIGM-SHIFT analysis on truncating versus non mutant and 216 samples to run missense versus non mutant. *TP53* mutation neighborhoods were selected in a supervised fashion by selecting features based on a rank ratio of the features determined by a linear SVM. PARADIGM-SHIFT (P-Shift) scores for p53 (reflecting the discrepancy between activity as inferred by up-/down- stream pathway signals) was computed as the difference in activity between two runs of PARADIGM - one in which only upstream regulators are connected (R-run) and one where only downstream targets are connected (T-run). We then assessed the accuracy of the models by using the absolute P-Shift score as a classifier to predict *TP53* mutation status with 5-fold cross validation. The average AUC over the 5-folds for predicting truncating mutations (against non-mutants) is 0.57. In contrast, the average AUC for the prediction of missense mutations (vs. non-mutants) is only 0.47, suggesting that PARADIGM-SHIFT may not be effective at distinguishing missense mutants from non-mutants.

Comparing the distribution of P-Shift scores between truncating mutants and non-mutants shows an enrichment of negative P-Shift scores in the truncating mutant samples indicative of a loss-of-function (LOF) mutation. The significance of this LOF call was determined by running a background model in which the selected network topology is fixed, but the data is permuted. Under this background model, the LOF call was found to have a z-score of -1.7. This is in contrast to when the P-Shift score distributions between missense mutants and non-mutants are compared, where no significant enrichment is observed under a similarly generated background model (Figure S11.4A-B). Altogether, these findings suggests that the signaling consequences of truncating and missense *TP53* mutations may not be equivalent; and that only truncating mutations are implicated as LOF based on the discrepancy of up- vs. down- stream activity signals. The entire sample set was used for training to determine the functional impact of truncating mutations of *TP53* on the network (Figure S11.4C). Interestingly, the pattern of activity of upstream regulators NGFR and SORT1 mirrors the profile of P-Shift score, where samples with high NGFR and SORT1 activities also have negative P-Shift scores. This highlights NGFR and SORT1 as major contributors to the discrepancy between up/down-stream signals in *TP53* truncation mutants, and implicates these features as potentially important upstream regulators of p53 signaling in endometrial cancers.

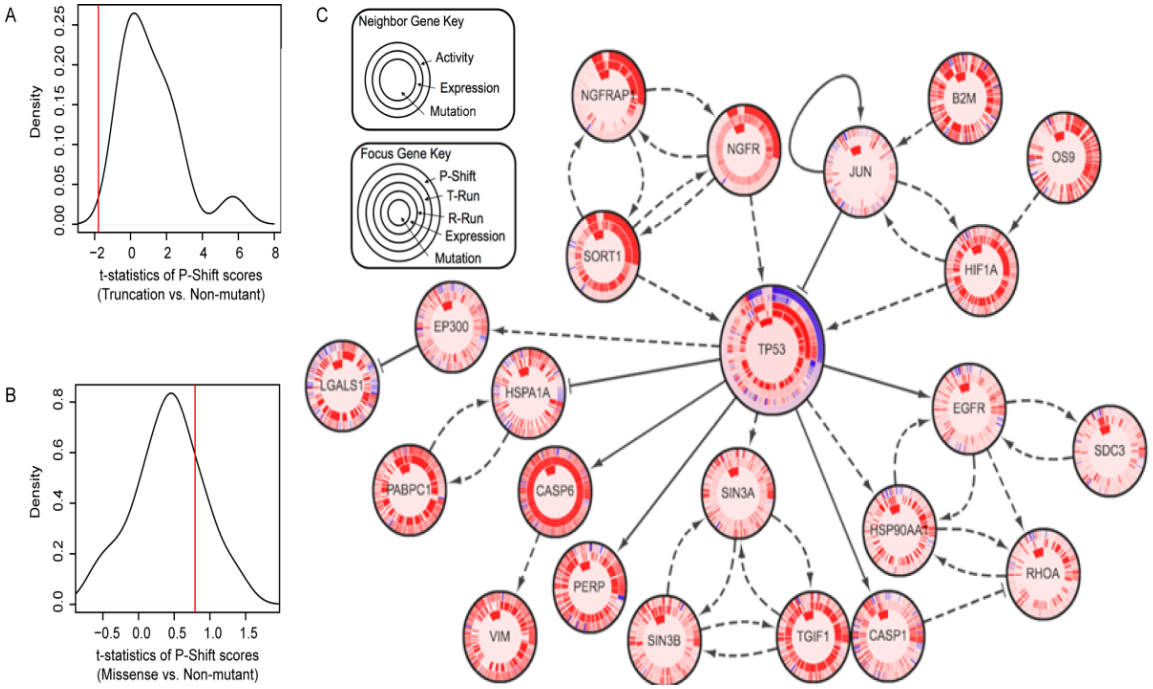


Figure S11.4. PARADIGM-SHIFT analysis of *TP53* truncation vs. missense mutations. (A-B) Distribution of t-statistics of the difference in P-Shift scores between non-mutants and (A) truncation mutants, (B) missense mutant under the permuted background models. Red line shows t-statistic based on actual data. (C) Circlemap display of mutation neighborhood selected for *TP53* truncating mutations. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation. Samples were sorted first by the *TP53* mutation status (inner ring), then by PARADIGM-SHIFT score.

Machine learning classifiers (based on expression data) classify serous CN-High endometrial cancers as basal-like and vice versa.

We first asked whether (mRNA expression) data supports the hypothesis that CN-High endometrial cancers share common molecular signatures with the TCGA basal breast and serous ovarian samples. To address this, we asked whether machine-learning classifiers, trained to recognize basal from luminal samples also classify ovarian and CN-High UCEC samples as basal. Models are trained on 80% of the breast cancer dataset using many different algorithms; and the linear model with the highest accuracy in a 5X5 fold cross-validation is selected. The test set comprise of the remaining 20% of breast cancer samples, and the serous ovarian and UCEC samples. To estimate the distribution of permuted background scores, we generated 1000 randomly selected and permutated samples from the test set and scored them using the predictive model. Each prediction is then assigned a z-score based on this permuted background score distribution; and samples with scores inside the margin of the permuted background are deemed ambiguously classified and ignored (Figure S11.5A-B). Similarly, we also constructed a reciprocal classifier, trained to recognize CN-High from other UCEC cases from 80% of UCEC cases, and applied it to a test set of the remaining 20% of UCEC, breast and ovarian samples (Figure S11.5C-D). Performance of the classifiers at distinguishing basal from luminal breast samples and CN-High from other UCEC samples within the test set is assessed by ROC analysis.

When we applied the basal-luminal breast cancer classifier, similar to the serous ovarian samples, most of the CN-High endometrial cancers are predicted to be basal-like. In fact, the basal-luminal classifier, although trained on breast cancer data, was able to very accurately distinguish CN-High UCEC samples from other subtypes (AUC = 0.93) when applied to the endometrial samples. Similarly, a reciprocal classifier, trained to recognize CN-High from other endometrial subtypes, classifies most of the basal-like breast and serous ovarian cancers to be CN-High like; and is able to distinguish basal from luminal breast cancers with great accuracy (AUC=0.97). Altogether, these findings suggest CN-High, basal-like breast and ovarian cancers share a common molecular signature, which distinguishes them from other endometrial and luminal breast cancers.

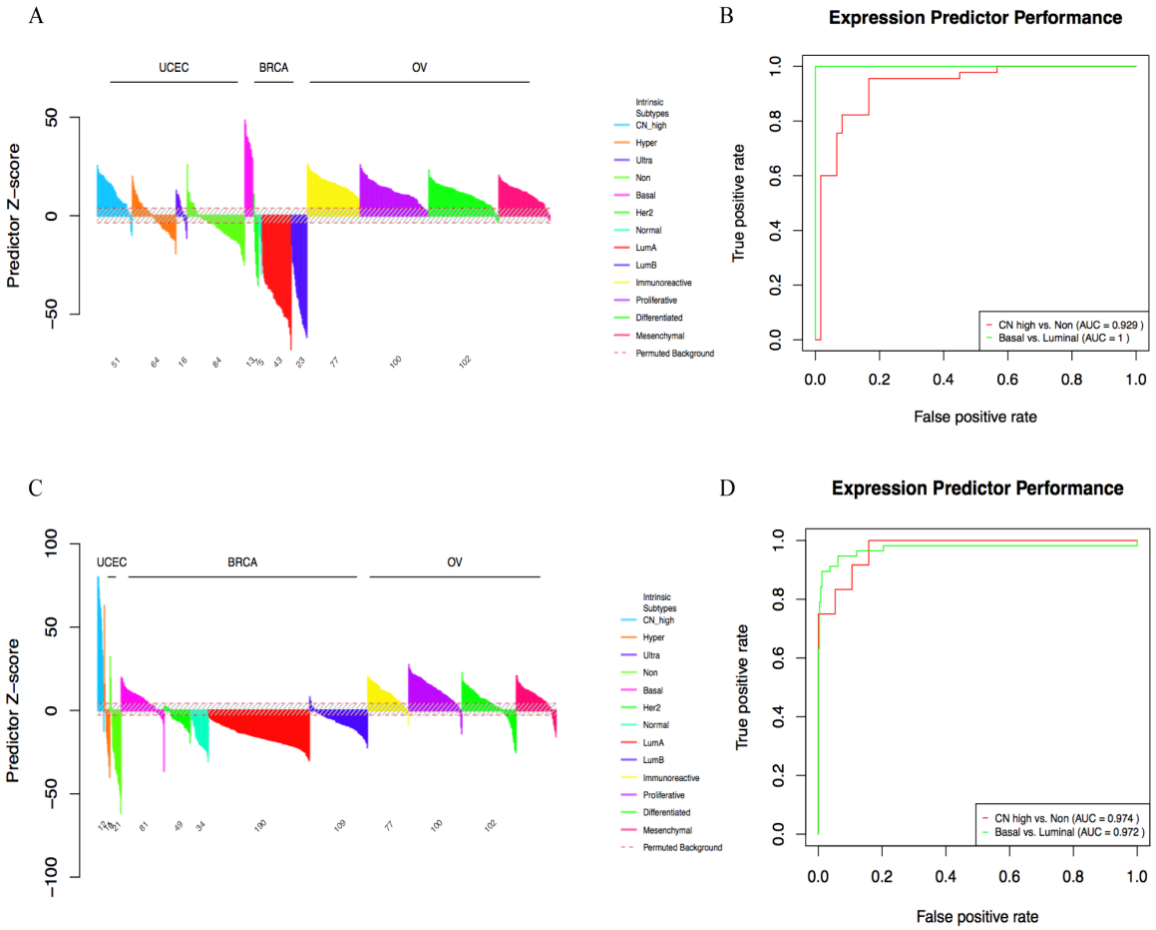


Figure S11.5. Machine learning classifiers demonstrate similarities between basal breast, serous ovarian and CN-High endometrial cancers. (A) Plot of basal-luminal prediction scores by tumor origin and subtype. (B) ROC curves for prediction of CN-High and basal subtypes within the UCEC and breast test samples respectively using the basal-luminal predictor. (C) Plot of CN-High-Others prediction scores by tumor origin and subtype. (D) ROC curves for prediction of CN-High and basal subtype with the UCEC and breast test samples respectively using the CN-High-Others predictor

Genes with shared expression patterns in basal breast, serous ovarian and CN-High endometrial cancers are significantly interconnected by known pathway interactions.
Integrated pathway levels generated from PARADIGM for the 377 TCGA ovarian and 81 basal breast cancer samples were obtained; and the average IPL across samples was computed. Among the ~14K features present in the ovarian and breast dataset, 3436 were mapped to IPLs showing significant differential activation between CN-High vs. others UCEC subtypes identified as described above. Restricting to these IPLs, a linear fit of average ovarian/basal breast activity onto the CN-High vs. Others differential score was performed (Figure S11.6A). A 'CN-High' score was computed as the orthogonal projection of the average ovarian activity onto the linear fit. Features with 'CN-High' scores at least one standard deviations from the mean were defined as significant; and regulatory sub-networks within the SuperPathway structure linking these features were identified and displayed using Cytoscape (Figure S11.6B).

The PARADIGM inference differentials for CN-High versus other endometrial cancers are significantly concordant with overall inferred activity in the TCGA basal breast and serous ovarian cohorts (r=0.47). 961 features were selected as having significant basal breast/ovarian to CN-High associations using the 'CN-High' score. Subnetwork analysis suggests higher activity of the MAPK14 and MAX hubs and lower activity of the p53 and XBP1-2 hubs are common pathways features shared between basal breast, ovarian and CN-High endometrial cancers. Pathway enrichment analysis independently confirms XBP1, p53 and MYC signaling as significantly enriched among features showing significant basal/ovarian to CN-High associations.

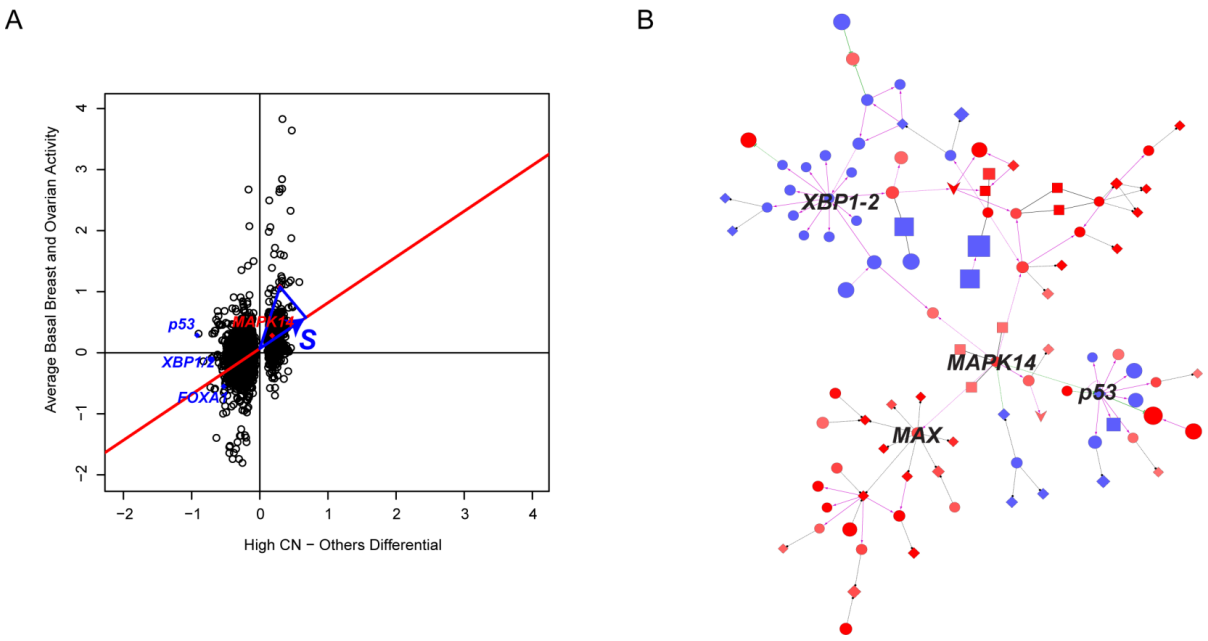


Figure S11.6. Comparison of the High CN Subtype with basal breast and ovarian cancers. (A) Scatterplot of average basal breast and ovarian activity vs. High CN-Others differential scores. Regression line of ovarian activity on the High CN-Others differential was fit (red line) and the orthogonal projection of a given point (blue arrow) onto the linear fit was determined to calculate the 'CN-High' score (s). Highlighted in red is the MAPK14 Hub significantly activated in the High CN subtype; and highlighted in blue are specific points representing important regulatory hubs (p53, XBP1-2, FOXA1) with lower activity in High CN UCEC. (B) PARADIGM analysis reveals common networks in basal breast, ovarian and High CN endometrial cancers. Basal breast/ovarian to High CN subtype associations were assessed using a 'CN-High' score and only significant features were retained. The largest interconnected sub-networks linking significant features are shown as a Cytoscape plot: positive values (red) indicate higher activity in CN-High endometrial cancers and negative values (blue) indicate lower activity. Node shapes correspond to complexes (diamonds), proteins (circles) microRNAs (squares), and cellular processes (inverted v-shapes). Network hubs (greater than 5 connections) are highlighted in boxes and labeled.

Machine learning classifiers (based on expression data) classify hormonal cluster endometrial cancers as luminal-like.

To assess whether the hormonal expression cluster endometrial samples are similar to luminal breast cancers, we applied the top basal vs. luminal classifier (as described above), and computed a *P* value (Fisher's Exact test) for the disproportion of samples predicted as luminal in the hormonal expression cluster (Figure S11.7A). We also asked whether the hormonal endometrial cluster is more similar the luminal A or luminal B subtype using the same methodology by developing a luminal A vs. B classifier (Figure S11.7B).

The basal-luminal classifier predicts the majority of the hormonal endometrial samples as luminal-like, suggesting the hormonal endometrial samples may share molecular signatures with luminal breast cancers. However, when we applied the luminal A vs. B classifier, we did not find any significant association between the hormonal endometrial samples with either of the two luminal subtypes (P = 0.336), suggesting that the hormonal endometrial samples is more similar to a mixed population of luminal breast cancers, rather than a specific luminal breast subtype.

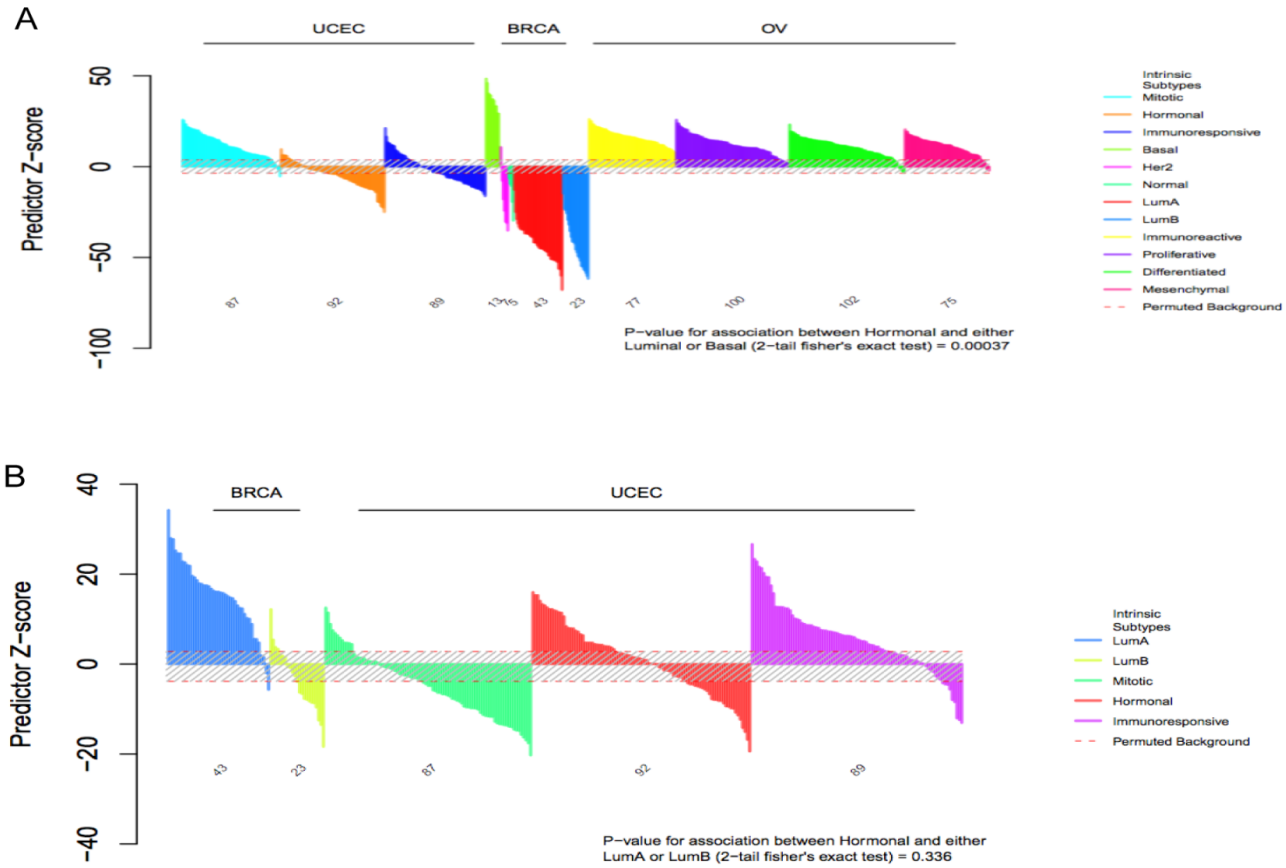


Figure S11.7. Machine learning classifiers demonstrate similarities between luminal breast and hormonal endometrial cancers. (A) Plot of basal-luminal prediction scores by tumor origin and subtype in UCEC, 20% breast and ovarian cancer data. Color code for each tumor subtype is shown on the right. The *P* value of association between hormonal endometrial and luminal/basal breast cancers is shown below. (B) Plot of luminal A vs. B prediction scores by tumor origin and subtype in UCEC, 20% luminal breast cancer data. Color code for each tumor subtype is shown on the right. The *P* value of association between hormonal endometrial and luminal A/B breast cancers is shown below.

Pathway-based biomarkers of hormonal versus other endometrial subtypes.
IPLs differentially activated between the hormonal and the other subtypes (Proliferative and Immunoreactive) were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg (BH) FDR correction. Only features deemed significant (FDR corrected $P < 0.05$) by both tests were selected. Pathways enriched among differentially activated IPLs were assessed using the EASE score with BH FDR correction; and sub-networks were constructed to identify regulatory hubs based on interconnectivity and visualized using Cytoscape (Figure S11.8).

~2.9K IPLs were found to be significantly differentially activated between the hormonal vs. other endometrial subtypes. Pathway enrichment and subnetwork analysis independently implicated XBP1, FOXA1, and MYB signaling as major hubs showing differential activation in the hormonal relative to other UCEC cases.

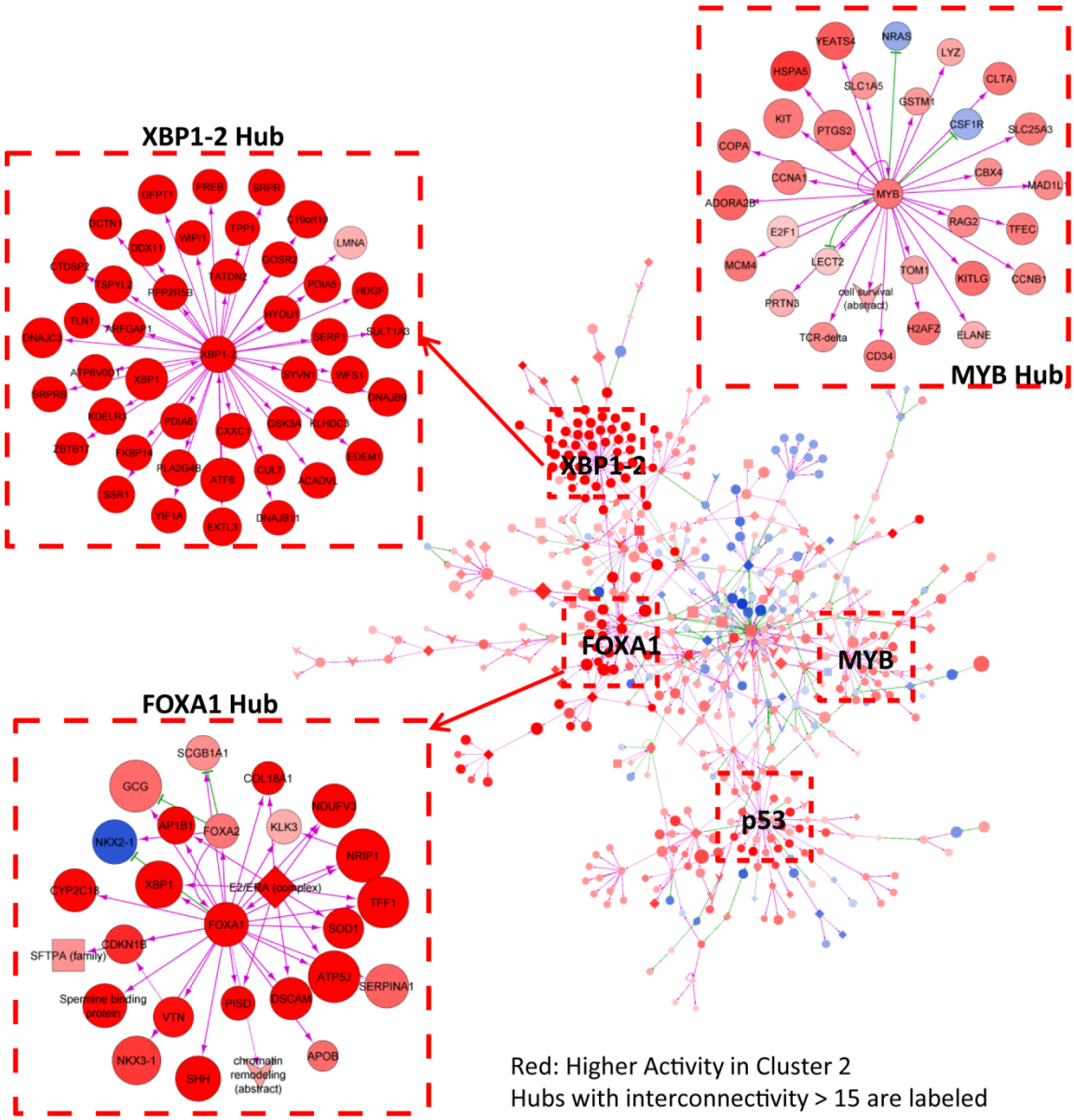


Figure S11.8. Differentially activated pathway features between Hormonal and others UCEC subtypes. Largest interconnected regulatory subnetwork of differentially activated IPLs is displayed, with network hubs showing interconnectivity > 15 edges labeled. A zoomed in view of the FOXA1, MYB and XBP1/2 hubs are also shown. Color intensity reflects activity differences between subtype (red: higher in hormonal, blue: higher in others). Purple arrows denote activation. Green tees represent inhibition. Node shapes reflects pathway concept type (inverted v: abstract concept, diamond: complex, circle: protein). Node size is scaled to the significance of differential activation.

Hormonal endometrial shows activation of FOXA1/ER signaling pathways similar to luminal breast cancers, but exhibits higher HIF1A/ART, MYC/Max and FOXM1 signaling.

Extending the methodology used to identify commonly activated pathways between High-CN and basal breast/ovarian cancers, we computed the PARADIGM inference differential between the hormonal and other endometrial cancers and the differential between luminal and basal breast cancers for all IPLs. We performed a linear fit of the luminal-basal differential on the hormonal-others differential, and defined a 'hormonal endometrial-like' score of the orthogonal projection of the luminal-basal differential onto the linear fit. Features with scores at least two standard deviations away from the mean is deemed significant; and regulatory sub-networks within the SuperPathway structure linking these features were identified and displayed using Cytoscape (Figure S11.9A). In addition, we performed the reciprocal analysis, and defined a 'luminal breast-like' score as the orthogonal projection of the hormonal-others differential onto the luminal-basal breast cancer differential. Significant features were similarly identified and displayed (Figure S11.9B); and the resulting networks were compared for common and distinct pathway hubs of differential activation.

Although our machine learning classifiers suggests that hormonal endometrial cancers are luminal breast cancer-like, the PARADIGM inference differential between the luminal-basal breast and hormonal-others endometrial cancers are very weakly (albeit significantly) correlated ($r = 0.04$). 297 features were deemed to show significant 'hormonal-like' scores; and pathway enrichment and subnetwork analysis independently identifies FOXA1/ER, XBP1 and p53 as major regulatory hubs among these 'hormonal endometrial-like' features. In comparison, in the reciprocal analysis, FOXA1/ER, HIF1A, MYC/Max and FOXM1 were implicated as major regulatory hubs among 350 features showing significant 'luminal breast-like' scores. Altogether, these findings suggests that while hormonal UCEC and luminal breast cancers shares common activation of the FOXA1/ER signaling pathways, the lower activity of the HIF1A, MYC/Max and FOXM1 hubs, which differentiated luminal from basal breast cancers may not be a distinguishing feature of the hormonal endometrial subtype. Also of note, although XBP1-2 is a feature showing significant 'hormonal endometrial-like' and 'luminal breast-like' scores, it appears to be a larger regulatory signaling hub within the hormonal endometrial-like features. Given the differential responses of breast and endometrial cancers to estrogen stimulation, and that XBP1 is an estrogen-regulated target, further investigation of signaling differences downstream of XBP1 between these tumor types may be warranted.

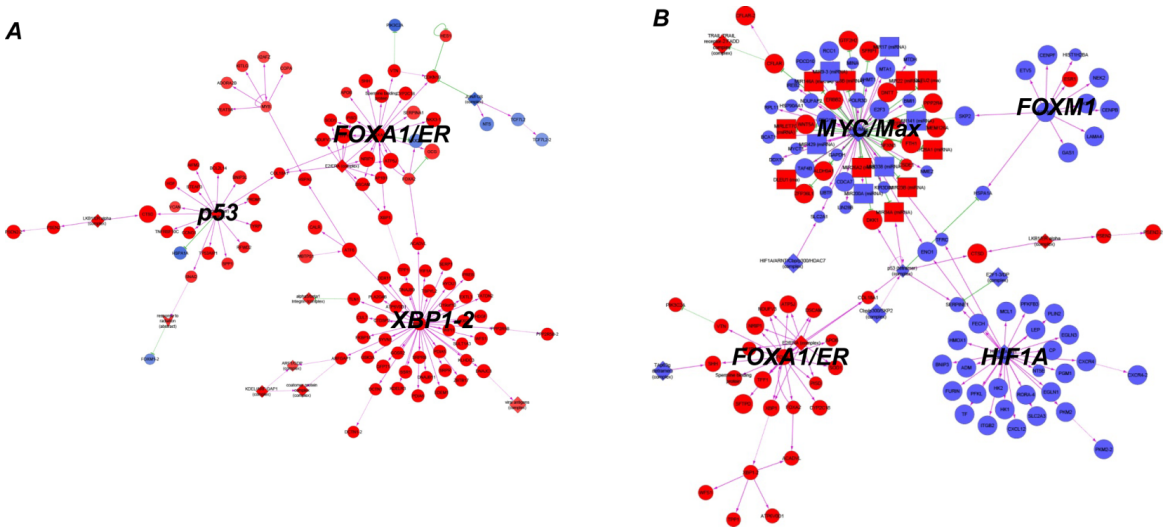


Figure S11.9. Comparison of the hormonal endometrial cancers with luminal breast cancers. PARADIGM analysis reveals common, as well as distinct, regulatory networks differentiating hormonal-other endometrial and luminal-basal breast cancers. (A) Luminal breast/ hormonal endometrial subtype associations were assessed using a 'hormonal-endometrial-like' score defined by projecting the luminal-basal differential onto the hormonal-others differential. The largest interconnected sub-networks linking significant features are show as a Cytoscape plot: positive values (red) indicate higher activity in hormonal endometrial cancers and negative values (blue) indicate lower activity. (B) Hormonal endometrial/luminal breast associations were assessed using a 'luminal-breast-like' score defined by projecting the hormonal-others differential onto the luminal-basal differential. The largest interconnected sub-networks linking significant features are show as a Cytoscape plot: positive values (red) indicate higher activity in luminal breast cancers and negative values (blue) indicate lower activity. Node shapes correspond to complexes (diamonds), proteins (circles) microRNAs (squares), and cellular processes (inverted v-shapes). Network hubs (greater than 5 connections) are highlighted in boxes and labeled.

Section references

1. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237-245, doi:10.1093/bioinformatics/btq182 (2010).
2. Tusher, V.G., Tibshirani, R., Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**:5116-5121 (2001).
3. Mao, J.H. *et al*. Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature* **432**:775-779 (2004).
4. Matsuoka, S., *et al*. Fbxw7 acts as a critical fail-safe against premature loss of hematopoietic stem cells and development of T-ALL. *Genes Dev* **22**:986-991 (2008).
5. Grim, J.E. *et al*. Fbw7 and p53 cooperatively suppress advanced and chromosomally unstable intestinal cancer. *Mol Cell Biol* **32**:2160-2167 (2012).
6. Ng, S. *et al*. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**:i640-i646 (2012).

Supplementary Methods S12: Cross tumor comparison

Shared regions of amplification across uterine serous, basal-like breast and high-grade serous ovarian (HGSOC) cancers include 1q21.3 (*MCL1*), 3q26.2 (*MECOM*), 4p16.3 (*FGF3*), 8p21.21 (*MYC*), 12q13.2 (*ERBB3*) and 19q13.2 (*CCNE1*). There are a few significant differences in amplifications between ovarian serous and tumors in the endometrial serous-like cluster (Figure 5a). However, unlike ovarian tumors, *ERBB2* is amplified in many of the endometrial serous-like tumors (26% of uterine serous compared with ~2% in ovarian and breast). Furthermore, the majority of these cases have concurrent *PIK3CA* mutations; an indication of resistance to *ERBB2* targeted therapy in multiple studies.[1] Also notable is that unlike ovarian tumors, uterine and breast tumors lack amplification of *KRAS* yet none have frequent *KRAS* mutations. Whole arm chromosome deletions of 4, 5 and 9 and arm level deletions in 8p that commonly occur in uterine serous tumors are also present in many ovarian and breast tumors.

The MC3 DNA methylation subtype largely overlaps with the tumors with frequent copy number changes, indicating that those tumors are copy-number instead of DNA-methylation driven (Figure S12.3). However, the MC3 tumors are not entirely identical to their ovarian counterparts. For example, extensive *BRCA1* and *BRCA2* inactivation through promoter hypermethylation was reported for both HGSOC and basal-like breast but only one out of 81 MC3 samples has *BRCA1* methylation. Also of note, the endometrial hormonal transcriptomic subtype was highly correlated with the luminal A breast cancer subtype that is characterized by a good outcome.

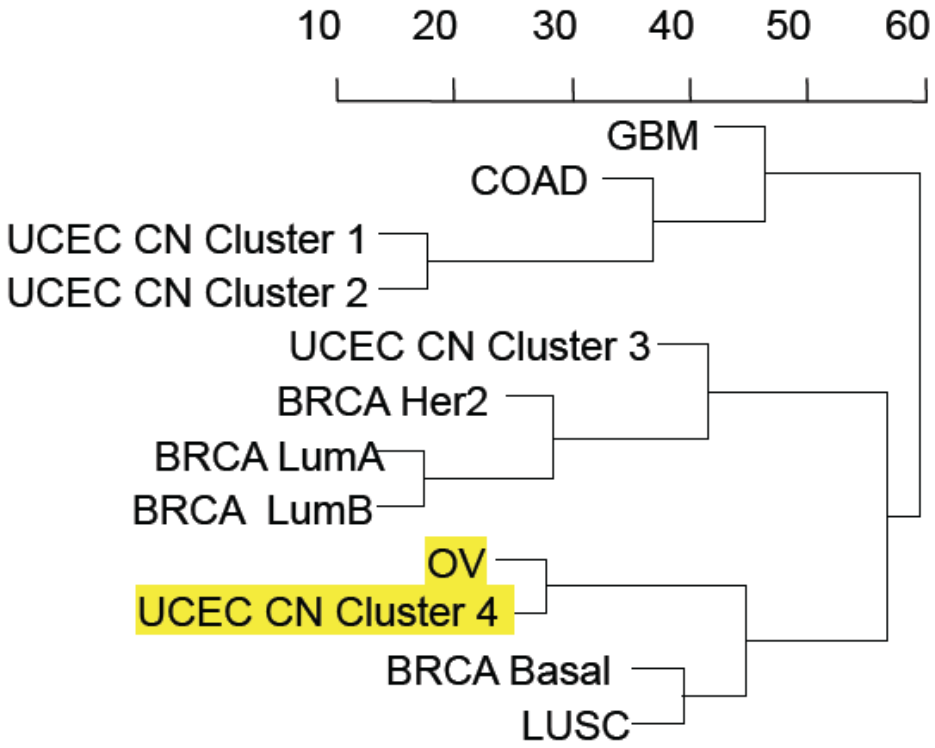


Figure S12.1 Cross tumor dendrogram from unsupervised hierarchical clustering showing relatedness of serous ovarian cancer, serous-like endometrial cancer from copy number cluster 4, and basal-like breast cancer. Clustering is based on the average gene level copy number change of each tumor type or subtype.

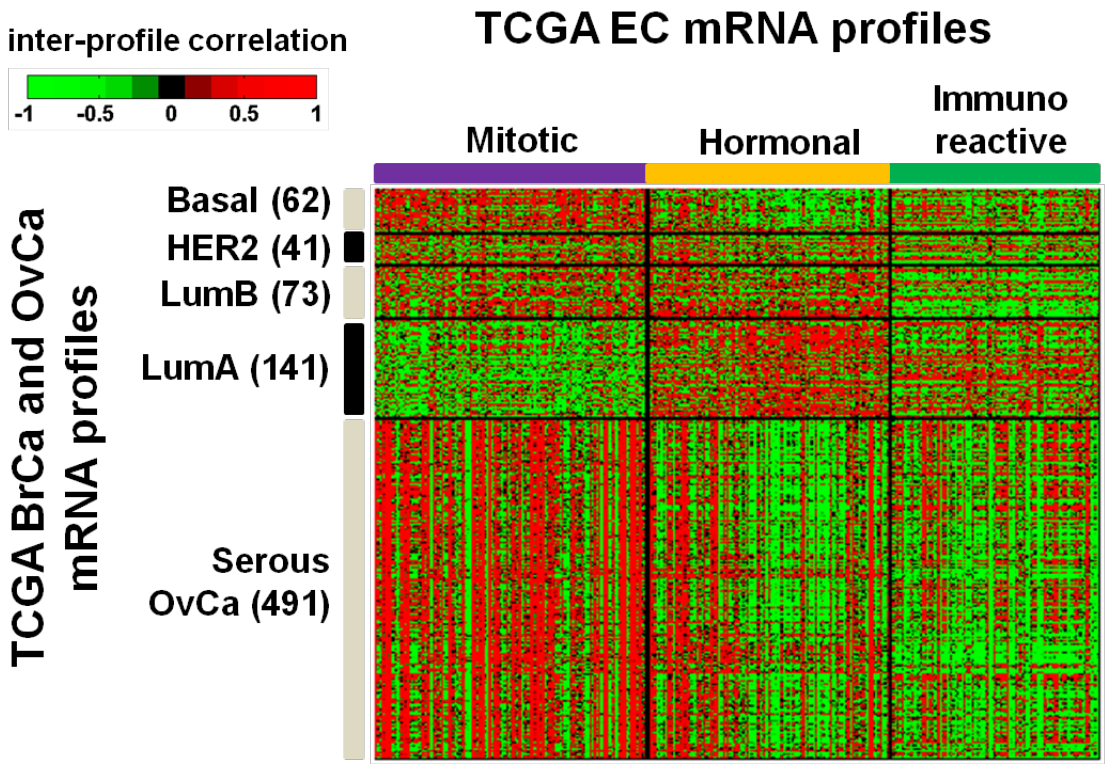


Figure S12.2 Supervised analysis of transcriptomic datasets for endometrial cancer, TCGA breast cancer and TCGA ovarian cancer.

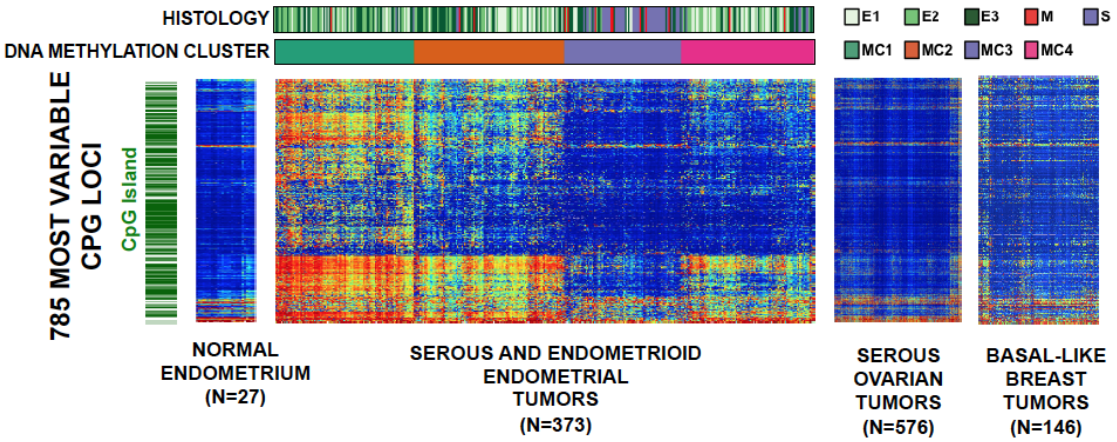


Figure S12.3 Methylation profiles across endometrial cancer, TCGA basal-like breast cancer and TCGA ovarian cancer.

Section references

1. Berns, K. et al. A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer Cell* **12**:395-402 (2007).