doi:10.1038/nature10166

Contents

Supplementary Methods S1. Biospecimen Collection and Clinical Data

A. Specimen Acquisition

Sample inclusion criteria

Biospecimens were collected from newly diagnosed patients with ovarian serous adenocarcinoma who were undergoing surgical resection and had received no prior treatment for their disease, including chemotherapy or radiotherapy. All cases had to be of serous histology but were collected regardless of surgical stage or histologic grade. Cases were staged according to the 1988 FIGO staging system. Each frozen tumor specimen had to have a companion normal tissue specimen, which could be adjacent normal tissue, peripheral lymphocytes, or previously extracted germline DNA. Each tumor specimen was approximately 1 cm^3 in size and weighed between 100mg and 200mg, in general. Each specimen was embedded in optimal cutting temperature (OCT) medium and histologic sections were obtained from top and bottom portions for review. Each case was reviewed by a board-certified pathologist to confirm that the frozen section was histologically consistent with ovarian serous adenocarcinoma. The top and bottom sections had to contain an average of 70% tumor cell nuclei with less than 20% necrosis. Specimens were shipped overnight from one of 15 tissue source sites using a cryoport that maintained an average temperature of less then -180°C. The tissue source sites contributing biospecimens included Memorial Sloan-Kettering Cancer Center, Washington University in St. Louis, University of Pittsburgh, Mayo Clinic, Duke University, Gynecologic Oncology Group, Cedars-Sinai Medical Center, University of California San Francisco, Harvard Medical School, MD Anderson Cancer Center, British Columbia Cancer Agency, Fox Chase Cancer Center, Imperial College London, International Genomics Consortium, and Roswell Park Cancer Institute.

Sample processing

DNA and RNA fractions were isolated from the tissue using an AllPrep DNA/RNA mini kit (Qiagen). Frozen tissue was homogenized with a Covaris adaptive focused acoustics tissue disruptor. DNA was selectively recovered from the lysate by chromatography on a spin column and the column was then washed. DNA was eluted in 0.1X TE buffer and then precipitated with 1/10 volume of 3M sodium acetate (pH 5.5) and 2.5 volumes of absolute ethanol. TRIzol was added to the flow-through from the DNA capture column, which contained RNA, and the solution heated at 65*°C* for 5 minutes. Chloroform was added and the phases were separated via centrifugation. To isolate microRNA, 10% of this total RNA fraction was mixed with 1/10 volume 3M sodium acetate (pH 5.5) and 2.5 volumes and absolute ethanol. Ethanol was added to the remaining 90% of the aqueous phase to provide appropriate binding conditions for RNA. The sample was then applied to an RNeasy spin column, treated with DNase I to remove residual contaminating DNA, then washed and eluted in 0.1X TE buffer.

Quality Control of Molecular Analytes

Matched normal patient DNA was extracted and purified from the blood or tissue using a QIAamp DNA Blood Midi Kit/QIAamp Mini Kit from QIAGEN. DNA and RNA from these purifications were quantitated by measuring optical density at 260, 280 and 320 nm wavelengths. The purity was assessed by the A260 and A280 absorbance ratio. All DNA samples were further qualified by agarose gel electrophoresis to confirm molecular weight distributions. To estimate the quality of the RNA, we used the RNA 6000 Nano assay on the Agilent

Bioanalyzer, which provided two estimates of the integrity of the 28S and 18S ribosomal RNA: RIN (RNA Integrity Number) and the 28S/18S ratio. Acceptable values were 28S/18S ratio ≥ 1 or RIN ≥7.

To date, 1020 ovarian cases have been received by the BCR and 564 (55%) have passed quality control. The biospecimens included in this report come from 518 ovarian samples included in batches 9, 11-15, 17-19, 21, 22, and 24 (Figure S1.1). For the present analyses, grade 1 and FIGO stage I tumors were excluded as they may represent a disease biologically and distinct from high-grade advanced stage ovarian carcinoma. In total, 22 cases were excluded due to: grade 1 (5), stage I (15), wrong diagnostic site (1), previous treatment (1).

B. Clinical data annotation

Clinical data collection

Clinical data were obtained from TSSs through data collection forms. Data forms were entered electronically at the BCR and XML files were generated. The XML files were parsed into flat text files at University of North Carolina and posted at the DCC. Clinical data can be accessed and downloaded from the TCGA Data Portal at http://tcga.cancer.gov. Demographics, histopathologic information, treatment details including chemotherapy drugs, doses and routes of administration, and outcome parameters were collected.

Clinical data definitions

The definitions of most clinical variables were implicit and select variables were defined as follows: TUMORRESIDUALDISEASE was defined as the size of residual disease at the conclusion of the primary surgical procedure. This field was used to define surgical cytoreduction as optimal or suboptimal. Optimal was defined as no residual disease greater than 1cm and included the variable categories of no macroscopic disease (*i.e.* microscopic residual disease) and 1 to 10mm. Suboptimal was defined as residual disease greater than 1cm and included the variable categories of 11 to 20mm and greater than 20mm.
PRIMARYTHERAPYOUTCOMESUCCESS was defined as the response to treatment determined after primary surgery and subsequent adjuvant chemotherapy.
PERSONNEOPLASMCANCERSTATUS was defined as the last known status of disease. For the purpose of these analyses, the date of surgery was used as a surrogate for the date of initial diagnosis, since treatment planning and intervention for these cases undergoing initial surgical resection began at that time point. Overall survival was defined as the interval from the date of initial surgical resection to the date of last known contact or death. Progression free survival was defined as the interval from the date of initial surgical resection to the date of progression, date of recurrence, or date of last known contact if the patient was alive and has not recurred. For the purpose of these analyses, any patient who had died without a date of progression or recurrence was excluded from analyses of progression free survival.

Chemotherapy treatment details were reviewed to identify drugs prescribed for adjuvant therapy. The date of last primary platinum treatment was also determined from the available chemotherapy details and included adjuvant therapy and consolidation treatment when given consecutively following adjuvant therapy. The platinum free interval was defined as the interval from the date of last primary platinum treatment to the date of progression, date of recurrence, or date of last known contact if the patient is alive and has not recurred. Platinum status was defined as resistant if the platinum free interval was less than six months and the patient had progressed or recurred. Platinum status was defined as sensitive if the platinum free interval is six months or greater, there was no evidence of progression or recurrence, and the follow-up interval was at least six months from the date of last primary platinum treatment. Patients who have not progressed or recurred and been followed for less than six months from the date of last primary platinum treatment were excluded from analyses regarding platinum status.

Clinical data analysis

Standard statistical tests were used to analyze the clinical data including, but not limited to, X^2 test, Fischer's exact text, Student's t test, log-rank text, and Cox proportional hazard analysis, as appropriate. Descriptive statistics were also included. All statistical tests were two-sided and statistical significance was considered when $P < 0.05$. Analyses of clinical data were primarily performed using SPSS v.18 (SPSS, Chicago, IL). Clinical data were available for 488 patients included in this report. Key clinical data variables are provided in Table S1.2.

C. <u>Additional clinical data results</u>

Demographics and histopathology

As indicated in the main text and shown in Table S1.1, the characteristic of the TCGA ovarian cases reflect the general population of women with advanced ovarian cancer. The average age at diagnosis was 60.2 years, all cases were of serous histology, 73% of the cases were FIGO stage IIIC, 16% were FIGO stage IV, 88% were grade 3, and 73% had optimal surgical cytoreduction.

Overall and progression-free survival

Univariate correlations between select clinical variables and progression-free survival (PFS) or overall survival (OS) are shown in Table S1.3. Age at diagnosis was associated with overall survival. Platinum status was associated with both PFS and OS. Stage III and optimal surgical cytoreduction both demonstrate a trend toward improved OS. The median PFS and OS was 16.7 and 43.4 months for FIGO stage III patients and 14.0 and 32.9 months for FIGO stage IV patients, respectively (*P* = 0.12 for PFS and *P* = 0.07 for OS). The median PFS and OS was 16.7 and 44.2 months for optimally debulked patients and 15.4 and 36.2 months for suboptimally debulked patients, respectively (*P* = 0.34 for PFS and *P* = 0.06 for OS). The median OS was 57.9 months for platinum sensitive patients and 33.2 months for platinum resistant patients (*P* = 3.5e-19).

In a multivariate analysis shown in Table S1.4, including all clinical variables from Table S1.3, age at diagnosis and platinum status were independently associated with OS (HR 1.02, 95% CI: 1.00-1.03; HR 3.69, 95% CI: 2.60-5.21). Stage and platinum status were independently associated with PFS (HR 0.80, 95% CI: 0.66-0.96; HR 25.6, 95% CI: 15.9-41.7).

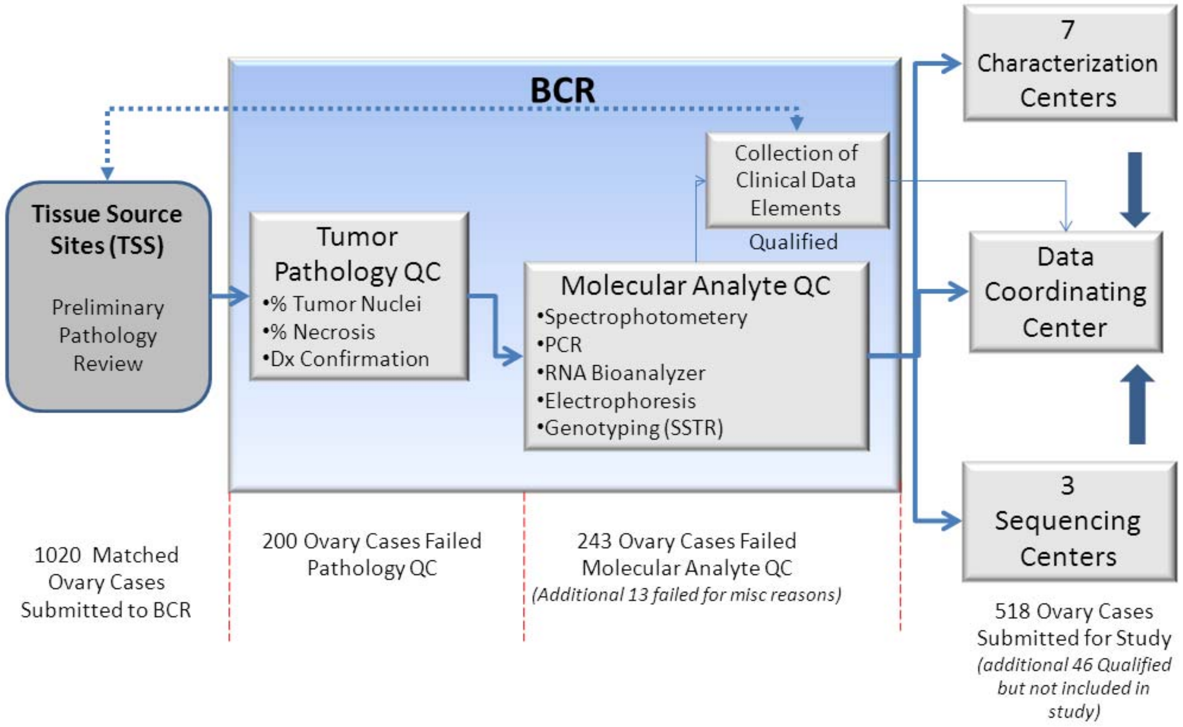
Surgical outcome is associated with OS, PFS, and platinum status

Many recent studies have demonstrated that patients left with microscopic residual disease after surgical cytoreduction have an improved outcome when compared with other optimally or suboptimally debulked patients[1-4]. We therefore further examined the PFS and OS of the TCGA ovarian patients in relation size of residual disease after surgical cytoreduction. Size of residual disease was microscopic in 90 (21%) cases, between 1 and 10mm in 223 (52%) cases, between 11 and 20mm in 30 (7%) cases, and more than 20mm in 89 (21%) cases. TCGA ovarian patients left with microscopic residual disease had improved PFS and OS when compared to patients left with optimal, non-microscopic disease or suboptimal disease (Figure S1.2 and Table S1.5). The median PFS was 21.8 months for patients with microscopic residual disease and 15.0 months for patients with more than microscopic residual disease (*P* = 0.001). The median OS was 57.4 months for patients with microscopic residual disease and 38.1 months for patients with more than microscopic residual disease (*P* = 3e-4). Microscopic residual disease was found to be independently associated with OS in a multivariate analysis.

An association between surgical cytoreduction and platinum sensitivity has also been previously reported[5]. Considering the improved PFS and OS identified in patients with microscopic residual disease, we explored the relationship between surgical cytoreduction and platinum sensitivity in the TCGA ovarian cases. We found no association between platinum status and surgical cytoreduction when defined traditionally as optimal or suboptimal. However, when considering microscopic residual disease separately from other optimally or suboptimally debulked patients, there was an association between surgical outcome and platinum status. Patients with microscopic residual disease were more likely to be platinum sensitive than patients with more than microscopic residual disease (*P* = 0.02, df=4; *P* = 0.003, df=2, Odds ratio = 3.1, 95%CI: 1.44-6.68; Table S1.6). These data suggest that surgical cytoreduction may have a direct impact on platinum status. Logistic regression analyses indicate that microscopic residual disease is independently associated with platinum status (*P* = 0.005).

Figures

Figure S1.1. Biospecimen processing and quality control. This figure summarized the flow of biospecimens from the Tissue Source Sites (TSS) through the Biospecimen Core Resource (BCR) and into the molecular analysis pipeline.


BCR
Tissue Source Sites (TSS)
Preliminary Pathology Review
Tumor Pathology QC
•% Tumor Nuclei
•% Necrosis
•Dx Confirmation
Collection of Clinical Data Elements
Qualified
Molecular Analyte QC
•Spectrophotometery
•PCR
•RNA Bioanalyzer
•Electrophoresis
•Genotyping (SSTR)
7 Characterization Centers
Data Coordinating Center
3 Sequencing Centers
1020 Matched Ovary Cases Submitted to BCR
200 Ovary Cases Failed Pathology QC
243 Ovary Cases Failed Molecular Analyte QC
(Additional 13 failed for misc reasons)
518 Ovary Cases Submitted for Study
(additional 46 Qualified but not included in study)

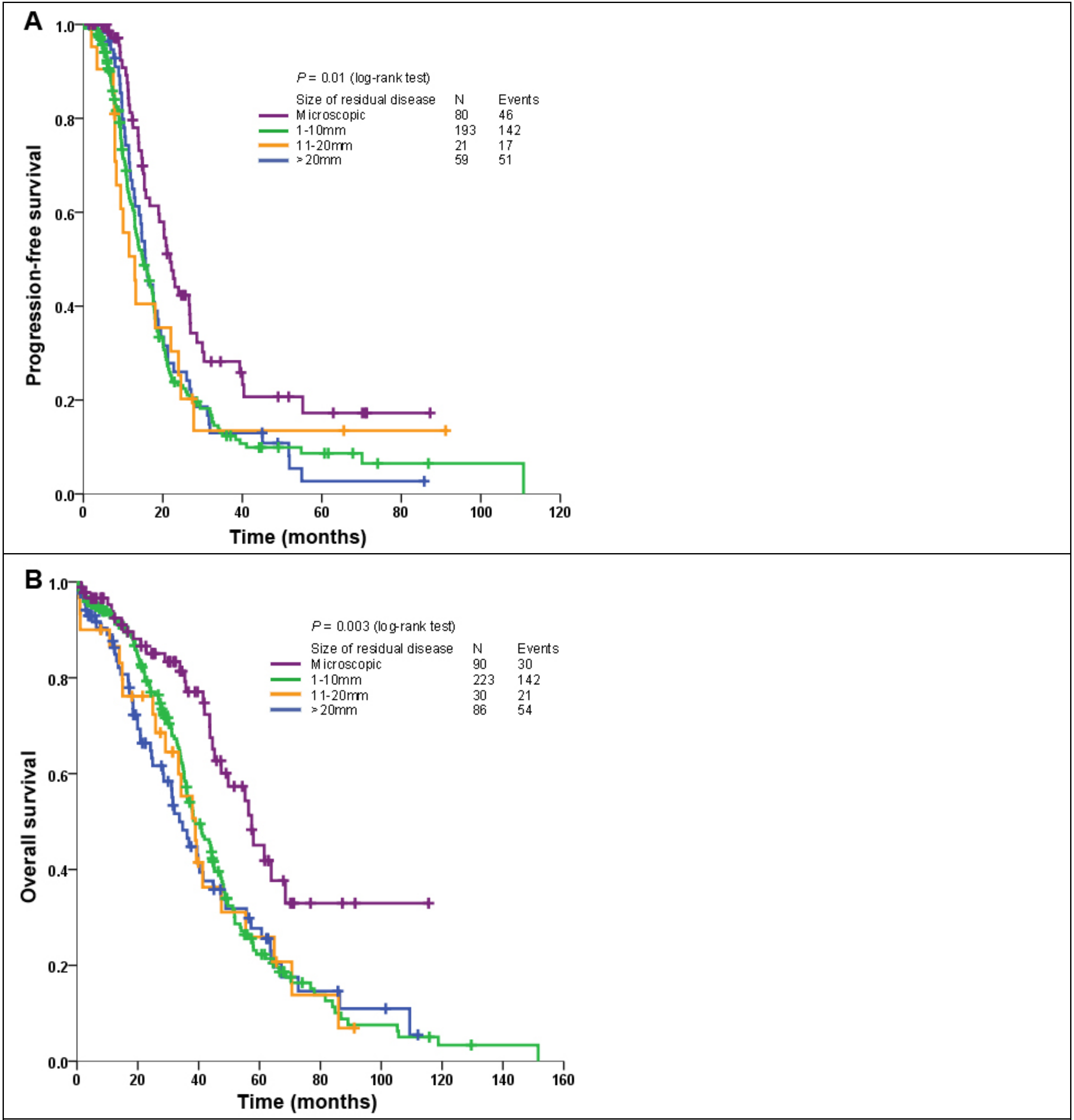


Figure S1.2. Progression-free and overall survival as a function of residual disease. Progression-free (A) and overall (B) survival is improved in patients left with microscopic residual disease after initial surgical cytoreduction. Other optimal patients left with more than microscopic residual disease do not have improved outcome when compared with suboptimal patients.

Table S1.1. Clinical-pathologic characteristics of TCGA ovarian cases*

Cohort	Training	Validation	Total
Number of patients	229	259	488
Age			
Mean, years (STD)	60.4 (11.5)	60.0 (11.4)	60.2 (11.4)
range	35-87	27-85	27-87
Tumor stage#			
II	4 (2%)	20 (8%)	24 (5%)
III	180 (79%)	201 (79%)	381 (79%)
IV	44 (19%)	35 (14%)	79 (16%)
Tumor grade^			
2	12 (6%)	45 (18%)	57 (12%)
3	212 (95%)	208 (82%)	419 (88%)
Number of patients	232	135	367
Histology			
Serous	229 (100%)	258 (100%)	487 (100%)
Surgical outcome			
Optimal (≤ 1cm)	154 (76%)	159 (69%)	313 (73%)
Suboptimal (>1 cm)	48 (24%)	71 (31%)	119 (28%)
Platinum status			
Sensitive	92 (65%)	105 (72%)	197 (69%)
Resistant	50 (35%)	40 (28%)	90 (31%)
Recurrent disease			
No	70 (31%)	67 (26%)	137 (28%)
Yes	159 (69%)	190 (74%)	349 (72%)
Vital status			
Alive	104 (45%)	111 (44%)	215 (45%)
Dead	125 (55%)	143 (56%)	268 (56%)
Median PFS, months (±SE)	14.9 (1.1)	17.9 (1.1)	16.8 (0.8)
Median OS, months (±SE)	44.4 (2.7)	41.5 (2.4)	43.6 (2.2)
Adjuvant chemotherapy regimen			
Single agent platinum	10 (5%)	2 (1%)	12 (3%)
Platinum/Taxane doublet	169 (79%)	156 (75%)	325 (77%)
Other platinum doublet	2 (1%)	11(5%)	13 (3%)
Platinum/Taxane triplet	32 (15%)	40 (19%)	72 (17%)

* Numbers do not sum to the total due to unavailable values. 30 samples have been excluded as described in supplement.

#P<0.01, X^2 test; ^P<0.001, X^2 test

PFS, progression-free survival; OS, overall survival; STD, standard deviation; SE, standard error

Table S1.3. Univariate analysis of overall and progression-free survival for TCGA ovarian cases

	Progression-free survival			Overall survival		
	HR	95% CI	*P*	HR	95% CI	*P*
Age (years)	1.00	0.99-1.01	0.99	**1.02**	**1.01-1.03**	**0.002**
Grade, 3 vs 2	1.33	0.95-1.86	0.10	1.35	0.94-1.94	0.11
Stage, III vs IV	0.88	0.75-1.04	0.13	0.87	0.74-1.01	0.07
TCGA cohort, training vs validation	1.05	0.94-1.19	0.38	0.99	0.88-1.12	0.91
Platinum status, resistant vs sensitive	**24.28**	**15.9-37.1**	**2.3e-49**	**3.94**	**2.86-5.43**	**6.0e-17**
Surgical outcome, optimal vs suboptimal	0.87	0.66-1.15	0.34	0.77	0.59-1.02	0.06

HR, hazard ratio

Table S1.4. Multivariate analysis of overall and progression-free survival for TCGA ovarian cases. All variables from Table S1.3 were included in the model. Statistically significant independent associations are emphasized in bold.

	Progression-free survival			Overall survival		
	HR	95% CI	*P*	HR	95% CI	*P*
Age (years)	1.00	0.99-1.01	0.88	**1.02**	**1.00-1.03**	**0.01**
Grade, 3 vs 2	2.42	0.78-1.81	0.41	1.12	0.69-1.81	0.65
Stage, III vs IV	**0.80**	**0.66-0.96**	**0.02**	0.96	0.77-1.20	0.74
TCGA cohort, training vs validation	1.03	0.90-1.19	0.66	0.99	0.84-1.18	0.92
Platinum status, resistant vs sensitive	**25.64**	**15.9-41.7**	**4.9e-41**	**3.69**	**2.60-5.21**	**2.0e-13**
Surgical outcome, optimal vs suboptimal	0.88	0.64-1.20	0.42	0.93	0.64-1.36	0.72

HR, hazard ratio

Table S1.5. Progression-free and overall survival as a function of size of residual disease after surgical cytoreduction.

	Progression-free survival[#]		Overall survival[#]	
Size of residual	Median	95% CI	Median	95% CI
Microscopic	21.8	19.0-24.7	57.4	45.8-69.0
1-10mm	15.0	12.5-17.6	39.0	33.4-44.6
11-20mm	13.0	8.4-17.5	39.0	31.7-46.2
>20mm	15.6	15.6-17.6	33.7	27.1-40.2

[#] Data is shown in months.

Table S1.6. Platinum status as a function of size of residual disease after surgical cytoreduction.[#]

Size of residual	Sensitive	Resistant
Microscopic	47 (84)	9 (16)
1-10mm	83 (61)	53 (39)
11-20mm	10 (59)	7 (41)
>20mm	35 (69)	16 (31)
Microscopic	47 (84)	9 (16)
More than microscopic*	128 (63)	76 (37)

[#] Data expressed as n (%).

*Includes 1-10mm, 11-20mm, and >20mm

References:

1. Armstrong DK, Bundy B, Wenzel L, et al. Intraperitoneal cisplatin and paclitaxel in ovarian cancer. N Engl J Med. Jan 5 2006;354(1):34-43. PMID: 16394300
2. Bookman MA, Brady MF, McGuire WP, et al. Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: a Phase III Trial of the Gynecologic Cancer Intergroup. J Clin Oncol. Mar 20 2009;27(9):1419-1425. PMID: 19224846
3. Eisenhauer EL, Abu-Rustum NR, Sonoda Y, et al. The effect of maximal surgical cytoreduction on sensitivity to platinum-taxane chemotherapy and subsequent survival in patients with advanced ovarian cancer. Gynecol Oncol. 2008 Feb;108(2):276-81. PMID: 18063020.
4. Winter WE 3rd, Maxwell GL, Tian C, et al. Prognostic factors for stage III epithelial ovarian cancer: a Gynecologic Oncology Group Study. J Clin Oncol. 2007 Aug 20;25(24):3621-7. PMID: 17704411.
5. Chi DS, Eisenhauer EL, Lang J, et al. What is the optimal goal of primary cytoreductive surgery for bulky stage IIIC epithelial ovarian carcinoma (EOC)? Gynecol Oncol. 2006 Nov;103(2):559-64. PMID: 16714056.

1. Chi, D.S. et al. What is the optimal goal of primary cytoreductive surgery for bulky stage IIIC epithelial ovarian carcinoma (EOC)? *Gynecol Oncol* **103**, 559-64 (2006).
2. Armstrong, D.K. et al. Intraperitoneal cisplatin and paclitaxel in ovarian cancer. *N Engl J Med* **354**, 34-43 (2006).
3. Bookman, M.A. et al. Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: a Phase III Trial of the Gynecologic Cancer Intergroup. *J Clin Oncol* **27**, 1419-25 (2009).
4. Winter, W.E., 3rd et al. Prognostic factors for stage III epithelial ovarian cancer: a Gynecologic Oncology Group Study. *J Clin Oncol* **25**, 3621-7 (2007).
5. Eisenhauer, E.L. et al. The effect of maximal surgical cytoreduction on sensitivity to platinum-taxane chemotherapy and subsequent survival in patients with advanced ovarian cancer. *Gynecol Oncol* **108**, 276-81 (2008).

Supplementary Methods S2: Exome Sequencing

The Genome Center at Washington University

Library Construction

Whole genome amplified (WGA) DNA samples (3µg) were constructed into Illumina libraries according to the manufacturer's protocol (Illumina Inc, San Diego, CA) with the following modifications: 1) DNA was fragmented using Covaris S2 DNA Sonicator (Covaris, Inc. Woburn, MA). Fragment sizes ranged between 100 and 500bp. 2) Illumina adapter-ligated DNA was amplified in a single 50µl PCR for five cycles. 3) Solid Phase Reversible Immobilization (SPRI) bead cleanup was used to purify the PCR and select for 300-500bp fragments.

Exome Capture and Sequencing

Sequencing libraries were hybridized with a customized version of the Agilent SureSelect All Exome v2.0 kit, which targets ~33 Mbp of coding sequence from ~18,500 genes, according to the manufacturer's protocol (1). Illumina library quantification was completed using the KAPA SYBR FAST qPCR Kit (KAPA Biosystems, Woburn, MA). The qPCR result was used to determine the quantity of library necessary to produce 180,000 clusters on a single lane of the Illumina GAIIx. Three lanes of 2x100bp paired-end sequence were generated per capture library.

BRAC1, *BRCA2*, and *TP53* were also sequenced using ABI 3730.

Alignment, De-duplication, and BAM File Generation

Illumina reads were mapped to the Ensembl release 45 version of Human NCBI Build 36 using BWA (2) v0.5.7 with soft trimming (-q 5). For each sample, individual lane alignments in BAM format were merged together using SAMtools (3) r544. Duplicates were marked in the merged BAM files by the MarkDuplicates class of Picard (4) v1.17. Reads with mapping quality of zero, or that were marked as duplicates by Picard, were excluded from further analysis.

Sample Identity Verification

To verify the identity of each BAM file, we compared SAMtools (3) filtered SNP calls with high-density SNP array data (Affymetrix) from the Cancer Genome Atlas research consortium using a customized Perl script. We required 8x coverage for a SNP genotyped as heterozygous in the array data, or 4x for a SNP genotyped as homozygous, to perform the comparison. On average across 176 samples, genotypes were compared at ~40,000 SNP positions with >98.5% concordance, suggesting that no samples were switched or significantly contaminated.

Somatic Mutation Calling

We have developed an automated pipeline for comprehensive identification of somatic mutations in exome data. Our approach combines the predictions of multiple algorithms:

1.) VarScan 2 (5). A heuristic somatic mutation caller that calls consensus genotypes, compares supporting read counts, and assesses the significance between tumor and normal using a Fisher's Exact Test. We applied the following thresholds for somatic mutations: coverage >= 3x, phred base quality >= 15, tumor variant frequency >= 15%, normal variant frequency <= 4%, FET p-value < 0.01.

2.) SomaticSniper. Our previously published (6) somatic mutation caller for whole genome resequencing data. We required that somatic mutations have average mapping quality >= 40, somatic score >= 40, and a tumor consensus genotype that matched the filtered SNP consensus for tumor from SAMtools (3).
3.) GATK (4) IndelGenotyper v2.0. A heuristic indel caller that compares tumor/normal data and classifies each variant as germline or somatic. We specified a window size of 300.

SNVs from VarScan and SomaticSniper were merged into a single non-redundant file. To remove false positives from paralogous alignments, local mis-alignments, sequencing error, and other factors, we filtered SNVs to remove any with strand bias, read position bias, or multiple high-quality mismatches in supporting reads. Indels from all three algorithms were merged into a single non-redundant file and filtered to remove small events around homopolymers, which are likely false positives.

Annotation and Tiering

We annotated the filtered high confidence somatic mutations using gene structure and UCSC annotation information, assigning each mutation to one of four tiers as previously described (6). Briefly, Tier 1 mutations alter coding sequence (nonsynonymous, synonymous, splice site, or noncoding RNA); Tier 2 mutations affect conserved or regulatory sequences; Tier 3 mutations occur in non-repetitive regions of the human genome, and Tier 4 mutations occur in repetitive non-coding regions.

Mutation Validation

All Tier 1 variants reviewed as somatic or ambiguous underwent PCR primer design and amplification using DNA from the tumor sample and matched DNA control. Amplifications were performed independently, then pooled together into tumor and normal PCR libraries. Gel fractionation was used to remove small fragments from each PCR pool prior to 454 library construction and sequencing using Titanium protocols. Read sequences and quality scores were extracted from 454 data files using *sffinfo* (Roche) then aligned to the Ensembl release 45 version of Human NCBI Build 36 using SSAHA2 (7) with the SAM output option. Reads with multiple top-scoring alignments were excluded from further analysis. Alignments were imported to BAM format using SAMtools (3). The validation status was determined by comparing tumor-normal read counts for each allele using VarScan 2 (5). To be validated, sites were required to have at least 30 reads with base quality >= 15 (Phred score) in both normal and tumor pools. To be validated as Somatic, a variant must have a somatic p-value (a read-count weighted measure of the significance of the allele frequency difference between normal and tumor) of less than 0.01, as calculated by VarScan 2 using Fisher’s Exact Test. A small fraction of sites were validated using ABI 3730.

Broad Institute

Library Construction and Exome Capture

We follow the procedure described by Gnirke et al.(1) adapted for production-scale exome capture library construction. Exome targets were generated based on CCDS genes, representing

188,260 exons from ~18,500 genes. DNA oligonucleotides were PCR amplified, then transcribed in vitro in the presence of biotinylated UTP to generate single•stranded RNA "bait." Genomic DNA from primary tumor and patient•matched blood normal was sheared, ligated to Illumina sequencing adapters, and selected for lengths between 200 to 350 bp. This "pond" of DNA was hybridized with an excess of bait in solution. The "catch" was pulled down by magnetic beads coated with streptavidin, then eluted.

Resulting exome sequencing libraries from the process described above were sequenced on three lanes of an Illumina GA•II sequencer, using 76 bp paired-end reads.

Illumina Sequencing

Libraries were quantified using a SYBR Green qPCR protocol with specific probes for the ends of the adapters. The qPCR assay measures the quantity of fragments properly adapter•ligated that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to the manufacturer's protocol (Illumina) using V2 Chemistry and V2 Flowcells (1.4mm channel width). SYBR Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were paired•end sequenced on Genome Analyzer II's, using V3 Sequencing•by•Synthesis kits and analyzed with the standard Illumina GAPipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis. The Illumina pipeline generates data files that contain the reads and qualities.

Sequence Data Processing Pipeline

The sequencing data•processing pipeline, called "Picard" (http://picard.sourceforge.net/; Fennel T. et al., unpublished), developed by the Sequencing Platform at the Broad Institute, starts with the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces, at the end of the pipeline, a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads with well-calibrated qualities together with their alignments to the genome (only for reads that were successfully aligned).

Several of the tools used in these pipelines were developed jointly by the Broad's Sequencing Platform, Medical and Population Genetics Program and the Cancer Program (additional details regarding parts of the pipeline focused on germline events, also used for medical and population genetics, will be described elsewhere; DePristo et al., submitted).

Picard consists specifically of four steps (briefly described below): (1) recalibration of base qualities, (2) alignment to the genome, (3) aggregation of lane and library data, and (4) marking of duplicate reads.

(1) Base-quality recalibration

Each base is associated with a Phred-like quality Q score(8)representing the probability that the base call is erroneous. The Q score represents –10*log10(Probability of error), rounded to an integer value. In order to make sure that Q30 bases indeed have a 1 in a 1000 chance of being wrong we used a GATK tool (http://www.broadinstitute.org/gatk) that empirically recalibrates the qualities based on the original Q score (generated by the Illumina software), the read•cycle, the lane, the tile, the base in question and the preceding base. The original quality scores are also kept in the BAM file in the read-level OQ tag.

(2) Alignment to the genome
Alignment is performed using MAQ3 [http://maq.sourceforge.net/] to the NCBI Human Reference Genome Build 36.3. The reads in the BAM file are sorted according to their chromosomal position. Unaligned reads are also stored in the BAM file such that all reads that passed the Illumina quality filter (PF reads) are kept in the BAM.
(3) Aggregation of lane- and library-level data
Multiple lanes and libraries are aggregated into a single BAM per sample. Lane•level BAM files are combined to library•level BAM files and these are then combined to sample-level BAM files. The BAM files contain read groups that represent the library and lane information. Information regarding the read groups appears in the BAM header (see the BAM file specifications in http://samtools.sourceforge.net/SAM1.pdf).
(4) Marking of duplicated reads
Molecular duplicates are flagged using the MarkDuplicates algorithm from Picard (http://picard.sourceforge.net/). The method identifies pairs of reads in which both ends map to the exact same genomic position as being multiple reads of the same DNA molecule and hence marks all but the first as duplicates.
The BAM files that are produced by the Picard pipeline are then delivered to dbGaP.

Local Realignment of Indels

This pre-processing step is performed before actual variant (SNV and short indel) calling. The rationale behind the local realignment is that while initial read mappings are usually and mostly correct at coarse-grained level (*e.g.* the overall position of the read on the reference is correct), finer details of some alignment can only be determined in the presence of the additional evidence from other reads at the locus. This problem manifests itself predominantly where the actual sequence contains indels. For instance, single read aligner normally would not place an indel near the end of the read. In other cases, a sequencing error in the read may cause an incorrect gap opening, *etc*. In most cases, at the locus with true indel event some alignments will contain the indel, and some others will not or will have it misplaced after the initial alignment. In the realignment step, we consider only the reads initially mapped into a small interval around putative event (hence making the procedure local and computationally tractable), and explicitly align them all to either reference or alternative consensus model(s). As a result, we are able to refine alignments for a number of reads.

Detection of Single Nucleotide Variations

Single nucleotide mutation detection for both whole genome and capture data was performed using a highly sensitive and specific method called *muTector* (Cibulskis K. et al, in preparation). In brief, muTector consists of three steps:

(i) Preprocessing the aligned reads in the tumor and normal sequencing data. In this step we ignore reads with too many mismatches or very low quality scores since these represent noisy reads that introduce more noise than signal.

(ii) A statistical analysis that identifies sites that are likely to carry somatic mutations with high confidence. The statistical analysis predicts a somatic mutation by using two Bayesian classifiers – the first aims to detect whether the tumor is non-reference at a given site and, for those sites that are found as non-reference, the second classifier makes sure the normal does not carry the variant allele. In practice the classification is performed by calculating a LOD score (log odds) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events. For the tumors we calculate

$$\mathrm{LOD}_T = \log_{10}\left(\frac{P(\text{observed data in tumor}\,|\,\text{site is mutated})}{P(\text{observed data in tumor}\,|\,\text{site is reference})}\right)$$

, and for the normal

$$\mathrm{LOD}_N = \log_{10}\left(\frac{P(\text{observed data in normal}\,|\,\text{site is reference})}{P(\text{observed data in normal}\,|\,\text{site is mutated})}\right).$$

Thresholds were chosen for each statistic such that our false positive rate is sufficiently low.

(iii)Post-processing of candidate somatic mutations to eliminate artifacts of next-generation sequencing, short read alignment and hybrid capture. For example, sequence context can cause hallucinated alternate alleles but often only in a single direction. Therefore, we test that the alternate alleles supporting the mutations are observed in both directions.

As muTector attempts to call mutations it also generates a coverage file in a wiggle file format(9), which indicates for every base whether it is sufficiently covered in the tumor and normal to be sensitive enough to call mutations. We currently use cutoffs of at least 14 reads in the tumor and at least 8 in the normal (these cutoffs are applied after removing noisy reads in the preprocessing step).

Detection of Small Insertion and Deletions

We have found the local realignment step to be very important for indel calling. Indeed, our results indicate that after the initially missing evidence for an indel is recovered through the realignment procedure, good specificity and sensitivity can be achieved using simple cutoff and filter-based approach. Our current indel calling procedure is implemented in two steps. First, high sensitivity calls are made based on count thresholds (minimum coverage, minimum fraction of indel-supporting reads at the locus). Second, these high sensitivity calls are filtered based on local alignment statistics around the putative event (average number of additional mismatches per indel allele-supporting read, average mismatch rate and base quality in a small NQS window around the indel). All calls are made in tumor samples and classified as somatic or germline based on the presence of any evidence (not necessarily strong enough to make an independent call) for the same event in matching normal sample.

Mutation Validation

Validation of somatic variants, both single nucleotide and short insertions and deletions, was performed using Sequenom Mass Spectrometry. This genotyping technology utilizes AssayDesigner v.3.1 software to design PCR and extension primers for low and high multiplex SNP and IN/DEL assays. Oligos were synthesized and mass-spec QCed at Integrated DNA Technologies, Inc. To minimize reagent and labor cost, individual genotyping reactions are multiplexed. SNPs are amplified in multiplex PCR reactions consisting of a maximum of twenty-four loci each. The volume of the PCR reaction is kept exceedingly small (6 µl) and only 10 ng of DNA per 24-36 multiplex SNP pool is consumed.

Following amplification, the Single Base Extension reaction is performed on the Shrimp Alkaline Phosphatase treated PCR product using iPLEX enzyme™ and mass-modified terminators™ (Sequenom iPLEX-GOLD reagents kit, San Diego). A small volume (~7 nl) of reaction is then loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl of matrix (3-hydroxypicolinic acid). SpectroCHIPs are analyzed in automated mode by a MassArray MALDI-TOF Compact system with a solid phase laser mass spectrometer (Bruker Daltonics Inc., 2005). The resulting spectra are called by real-time SpectroCaller algorithm and analyzed by SpectroTyper v.4.0 software, which combines base calling with the clustering algorithm.

Human Genome Center at Baylor College of Medicine

Library Construction

Whole genome amplified (WGA) DNA samples (5ug) were constructed into SOLiD pre-capture libraries according to a modified version of the manufacturer's protocol (Applied Biosystems, Inc.). Briefly, DNA was sheared into fragments approximately 120 bp in size with the Covaris S2 or E210 system as per manufacturer instructions (Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair (NEBNext End-Repair Module; Cat. No. E6050L) and A-tailing (NEBNext dA-Tailing Module; Cat. No. E6053L), followed by purification using a QIAquick PCR purification kit (Cat. No. 28106). Resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit (Cat. No. M2200L). Solid Phase Reversible Immobilization (SPRI) bead cleanup (Beckman Coulter Genomics, Inc.; Cat. No. A29152) was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR LM-PCR was performed using Platinum PCR Supermix HIFi (Invitrogen; Cat. No.12532-016) and 6 cycles of amplification. Following bead purification, PCR products were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500 (Cat. No. 5067-1506). Primer sequences and a complete library construction protocol are available on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

Exome Capture and DNA Sequencing

Precapture libraries libraries (2 ug) were hybridized in solution with either NimbleGen SeqCap EZ Exome Probes (~26 Mbs of coding sequence from ~17,000 genes), or a custom designed solution probe Vcrome1, (~43.9 Mbs of coding sequence from ~23,000), according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture LM-PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500 (Cat. No. 5067-1506). Capture efficiency was evaluated by performing a qPCR-based SYBER Green assay (Applied Biosystems; Cat. No. 4368708) with built-in controls (RUNX2, PRKG1, SMG1, and NLK). Capture library enrichment was estimated at 7 to 9-fold over background. Captured libraries were further processed for sequencing, with approximately 6-12 Gbs of sequence generated per capture library on either SOLiD V3 or V4 instruments (Applied Biosystems, Inc). A complete capture protocol can be found on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

Alignment, De-duplication, and BAM File Generation

SOLiD reads were mapped to the Ensembl release 45 version of the Human NCBI Build 36 using BFAST v0.6.4 using standard parameters. For each sample, individual runs in BAM format were merged together and duplicates were marked in the merged BAM files using Picard v1.22. Duplicate reads were excluded from further analysis.

Sample Identity Verification

To verify the identity of each BAM file, we compared our sequencing genotypes with high-density SNP array data (Affymetrix) from The Cancer Genome Atlas research consortium using a customized concordance analysis pipeline. The pipeline applies two genotype calling methods, e-GenoTyping, which screens raw reads for expected alleles from each read subset. After SNP calling, which uses the BAM file as input, filters duplicate reads and low mapping quality reads to produces a list of SNPs/INDELs in the format of SAMtools pileup. Our concordance metric incorporates allele frequency to reward matches or penalize mismatches between rare alleles. The metric also rewards exact matches or mismatches, and penalizes one allele matches, which are statically more common among unrelated samples. About 40,000 sites from the SNP array fall in the capture design region and contribute to the concordance metric. Samples are judged to be concordant and uncontaminated when they score significantly higher against their own SNP array data than any other patient, and the average of the scores against other samples is significantly lower. A match, by e-GenoTyping, is 95 +/- 3%, whereas unrelated samples score approximately 70%. For SNP call analysis, a match is 98 +/- 2%, and whereas unrelated samples score approximately 70%

Somatic Mutation Calling

The aligned reads from whole exome sequencing were prefiltered to remove reads with 3 or more non-reference bases, including insertions or deletions. This removed approximately 1-3% of reads from the BAM file. Base substitution and indels in SOLiD reads observed by *pileup* (SamTools 0.1.7) were collected and filtered to remove variant bases with SNP quality <100. For any given variant position in the tumor the variant allele frequency must be 15%. In addition, at least one read harboring the variant must have mapping quality=255 (i.e., uniquely mapped), and one variant must be Phred quality 40. Variants were discarded if they were observed only at the ends of reads, in position 38-50, or if they exhibited strand bias. Variants were annotated as somatic mutation if they were not observed in the normal. Putative somatic variants observed less than 5 times, or in which the coverage in the normal sequence was less than 9 were set aside. Greater than 80% of the target bases have sufficient coverage in both tumor and normal exome to make somatic mutation calls.

Annotation

Variants were annotated using gene structures from the NCBI RefSeq transcript set. Coding base substitutions were classified as missense, nonsense, splice site, or silent. Insertions and deletions were classified as in frame or frame shifting and submitted to the TCGA Data Coordinating Center.

Mutation Validation

All somatic and LOH variants were validated using a sequencing chemistry different from the discovery chemistry (SOLiD). PCR primers are designed to amplify the mutation target in both the tumor and the normal and the amplification products were sequenced using the AB 3730 Sanger, or 454 pyrosequencing methods. PCR reactions are cleaned using Exo Sap IT, (VWR, Inc.) for Sanger sequencing and by Solid Phase Reversible Immobilization (SPRI) beads (AMPure XP, Beckman Coulter Genomics) for 454 sequencing. For the Sanger-based validation, forward and reverse reads are generated and variants are called using SNP Detector v 3.0 software. Amplicons in which SNPdetector did not call the mutation were visually examined for evidence of the mutant allele. For the 454 sequencing based validation PCR products from tumor and normal samples are made into separate pools of 1000 amplicons. 454 Titanium sequencing libraries are generated and each pool is sequenced in a separate 454 run. Reads that map to their cognate amplicon in the genome reference sequence are realigned to the amplicon reference with crossmatch and the variant coordinate is examined for the presence of the somatic mutation. There must be at least 50 matching reads in tumor and normal to make a variant call although typically there are 500-2000 reads per amplicon. A variant is validated if it is observed in the tumor in at least 5% of the reads and not be observed at all in the normal. Variants in which the mutant allele frequency was 40% or greater in the original SOLiD sequencing data exhibited an 80% validation rate in Sanger sequencing. Allele frequencies less than 40% validate with decreasing efficiency by Sanger sequence, and must be validated by 454.

Downstream Analyses of Mutations

Mutation Annotation

Translational annotation of all somatic mutations is based on the combination of all human transcripts obtained from Ensembl Release 54_36p and the concurrent release of Entrez Gene (NCBI/Genbank) from May 2009 (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ASN_BINARY/Mammalia/Homo_sapiens.ags.gz). The reference alleles and positions were derived from the sequence and coordinates of NCBI Build36. All transcripts from both databases were annotated and a single representative transcript was selected for each somatic mutation based on the significance of the predicted functional effect of each mutation, ordered from most significant to least significant as follows: nonsense, frameshift, splice site, in frame, missense, no stop (nonstop/readthrough), silent, and RNA. Splice site mutations were restricted to substitutions, deletions, or insertions overlapping the 2bp intronic sequence defined as the splice donor or splice acceptor. RNA mutations were restricted solely to transcripts without an annotated open reading frame. Mutations affecting 3'UTR, 5'UTR, intronic sequence, and intergenic sequence were discarded for the purposes of downstream analysis. Mutations are encoded in MAF format and are available in the external file **Table_S2.1.MAF**. It should be noted that the overall list of somatic mutations is enriched for non-silent mutations versus silent mutations due to several factors, including the selection of a single representative transcript for each gene when multiple isoforms exist and a focus on reporting the annotation for the single transcript having the most most significant deleterious effect and largest open-reading-frame - all of which results in the under-representation of transcripts with smaller ORFs and omission of silent mutations that occur in alternate reading frames when multiple exons with different reading frames span a single mutation site.

Pfam Annotation and Analysis

Pfam(10, 11) protein domain annotation for all Ensembl and Entrez Gene transcripts was obtained by importing results from the Ensembl database or running InterProScan(12) for all transcripts where no domain annotation could be retrieved from the database. All non-silent mutations within identical Pfam domains were consolidated across gene families and statistically relevant clusters were selected for further analysis.

Background Mutation Rate Calculation

The overall background mutation rate was determined by dividing the total number of mutations by the total number of covered bases. This yielded an estimate of the BMR that was conservative (i.e. high), due to the fact that it includes all driver events as well as all passengers. We refined our estimate of the BMR by excluding the following highly mutated (and likely driver-containing) genes: *TP53*, *BRCA1*, *BRCA2*, *NF1*, and *RB1*. This lowered the final BMR estimates by approximately 2 percent. To account for the fact that certain base contexts and mutation types are known to have increased mutation rates, e.g. C residues in CpG dinucleotides, we calculated context-specific background mutation rates for each categories (**Table S2.2**).

Table S2.2a: Method 1

Class Mutati	on Rate
AT Transitions	3.86E-07
AT Transversions	5.38E-07
CG Transitions	6.29E-07
CG Transversions	1.33E-06
CpG Transitions	4.72E-06
CpG Transversions	1.47E-06
Indels (frame-shift and in-frame indels)	8.92E-08
Overall BMR	1.74E-06

Table S2.2b Method 2

Class Mutati	on Rate
AT Mutations	8.54E-07
CG Transversions	1.20E-06
other CG Transitions	5.45E-07
CpG Transitions	4.31E-06
Indels + Null	2.20E-07
Overall BMR	1.70E-06

Identification of Significantly Mutated Genes

The multitude of variables associated with genes and their somatic mutations makes assessing mutational significance a challenging problem. Specifically, there are various methods by which one could take account of the *collective* effects of these variables. Here, we use two approaches that both consider the diverse mutation rates for transitions and transversions under various sequence contexts, but that differ in the number and content of the mutation categories (**Supplementary Table S2.2**), as well as how the various category contributions are combined to obtain a final P-value. The detailed implementations are outlined below.

The first approach, MUSIC (MutationSigificanceInCancer, Dees *et al.*, in preparation), is inspired by the analysis of mutation patterns in a variety of cancer types (13, 14). Transitions generally occur at a higher rate than transversions. Substitution rates are also influenced by flanking sequences, an obvious example being that cytosines in CpG dinucleotides have a significantly higher mutation rate than cytosines in other sequence contexts. Moreover, the rate of indel events is roughly an order of magnitude lower than the rate of substitutions. These observations suggest scoring mutations according to their prevalence. Here, independent tests are first performed for the individual observations within the different sequence mutation categories (eg. A/T transitions, A/T transversions, C/G transitions, C/G transversions, CpG transitions, CpG transversions, and indels). Then, methods like Fisher's test, likelihood test, and convolution, can be used on the category-specific binomials to obtain an overall P-value. Specifically, Fisher's approach combines P-values from individual categories into one and a final result is calculated based on binomial distribution given the estimated background mutation rate. The likelihood ratio (LR) test calculates a P-value based on an LR between two hypotheses, the null hypothesis (i.e. true mutation rate = BMR) and alternative hypothesis (i.e. true mutation rate = maximum likelihood estimate), and uses a Chi-square distribution of LR. The convolution test is based on a semi-exact binned distribution/histogram of the product of point probabilities from individual categories. The false discovery rate for multiple-gene testing is controlled in all 3 methods using the standard Benjamini and Hochberg False Discovery Rate (FDR) procedure. Significant genes identified by this method are in **Table S2.3a**. It must be noted that the convolution test found TTN as significant. In contrast, Fisher's test and likelihood test both placed TTN as non-significant. We noticed that TTN has a high fraction of nonsilent mutations (67/83 or 80.7%) One possible explanation for the higher than expected number of nonsilent mutations in this gene is that an excessive number of rare or poorly characterized non-functional exons were targeted and sequenced when consolidated coding region targets were identified using a combination of all transcript isoforms present in Genbank, Ensembl, and UCSC. Dozens of transcript isoforms exist for TTN in these databases, however, no single common isoform exists in the CCDS database and many individual isoforms contain unique exons.

Table S2.3a. Genes significantly mutated by Method 1.

Rank G	ene	Mutations	AT Transitions	AT Transversions	CG Transitions	CG Transversions	CpG Transitions	CpG Transversions	Indels F	isher p-value	Likelihood Ratio p-value	Convolution p-value	Fisher FDR	Likelihood Ratio FDR	Convolution FDR
1	TP53	302	46	26	56 4	5	61 1	4	54	0.00E+00	0.00E+00	0.00E+00	0.0000	0.0000	0.0000
2	BRCA1	11	0	1	1 2		1 0		6	2.62E-05	6.83E-07	7.62E-08	0.2648	0.0069	0.0008
3	CSMD3	19	2	7	2 5		1 2		0	1.40E-04	9.66E-05	1.05E-06	0.9450	0.0886	0.0071
4	NF1	13	3	0	1 3		1 1		4	4.47E-04	1.40E-04	3.05E-06	1.0000	0.0914	0.0154
5	CDK12	9	2	1	2 0		0 1		3	8.05E-04	4.21E-05	5.77E-06	1.0000	0.0784	0.0233
6	FAT3	19	3	6	2 4		4 0		0	6.79E-04	1.24E-04	8.06E-06	1.0000	0.0914	0.0236
7	TTN**	67	9	15	7	19	13	2	2	1.00E-02	2.61E-02	8.20E-06	1.0000 0	.2257 0	.0236
8	GABRA6	6	1	0	3 1		1 0		0	7.56E-03	2.10E-04	4.68E-05	1.0000	0.0914	0.1181
9	BRCA2	10	1	1	0 2		1 0		5	4.78E-03	2.40E-04	6.36E-05	1.0000	0.0914	0.1426

Genes with convolution FDR <0.15 were included in Table S2.3a. We consider genes with convolution FDR <0.15 and also with Fisher's test FDR and/or Likelihood FDR <0.15 as significant. **TTN is not significant based on this criteria.

1. The second algorithm, MutSig (Lawrence et al., manuscript in preparation), is based in part on methods we have published elsewhere(15, 16). In brief, we tabulate the number of mutations and the number of covered bases for each gene. The counts are broken down by mutation context category: transitions at CpG dinucleotides, transitions at other C:G basepairs, transversions at C:G basepairs, mutations at A:T basepairs, one for indel and "null" mutations, which included indels, nonsense mutations, splice-site mutations, and non-stop (read-through) mutations. For each gene, we calculate the probability of seeing the observed constellation of mutations, i.e. the product P_1 x P_2 x … x P_m, or a more extreme one, given the background mutation rates calculated across the dataset.

(This is done by convoluting a set of binomial distributions, as described previously(2). This final P-value is then adjusted for multiple hypotheses according to the conventional Benjamini-Hochberg procedure for controlling False Discovery Rate (FDR). Mutations identified by this method are in **Table S2.3b**.

Table S2.3b Significant by method 2.

Rank Gen	e	Coverage	Mutations	CpG Transitions	Other CG Transitions	CG Transversions	AT Mutations	Indels and Null P-valu	e	FDR
1	TP53	393540	302	50	33 39		64 11	6	<1.00e-11	<1.89e-07
2 BR	CA1	1833605	11	0	0	1	0	10	3.20E-09	0.00003
3 NF	1	2607776	13	1	0	1	3	8	1.40E-06	0.0088
4 F	AT3	3637009	19	4	2	3	9	1	8.35E-06	0.039
5 G	ABRA6	439842	6	1	3	1	1	0	0.000022	0.08
6 RB	1	828029	6	0	0	1	0	5	0.000029	0.08
7 CS	MD3	3608502	19	1	2	7	8	1	0.00003	0.08
8 BR	CA2	2831480	10	1	0	0	2	7	0.000037	0.087
9 CD	K12	1524427	9	0	0	1	3	5	0.00006	0.13

Recurrent Mutation Identification and Proximity Analysis

Positional clustering of somatic mutations was exploited as a potential signal for functional elements. All somatic mutations having protein-altering translational effects were annotated based on the best representative transcript from Entrez Gene or Ensembl and the position of amino acid residues corresponding to affected nucleotides was identified. Mutations within a single gene were compared to identify those instances where mutations occurred within close proximity and clustered mutations were assigned to 10 bins representing a separation of 0-9 amino acid residues. For those deletions spanning sequence coding for multiple amino acid residues, the closest amino acid to the cluster was selected. Silent mutations and mutations affecting non-protein-coding RNA sequences were not included in proximity analysis.

Comparison of Mutations with COSMIC and OMIM databases

Detected and annotated mutations were compared to the COSMIC (version 48) and OMIM (Downloaded on 08/27/2010) databases. For each mutation, all possible matching transcripts were used to determine possible amino acid changes with respect to both residue and position within the associated protein. This change was then checked against the two databases. For COSMIC, if the genomic coordinate or amino acid position is identical to any record in the database, the mutation was declared a match. For OMIM, only amino acid position was used for comparison. In cases where a mutation affected a splice site, we checked to see if its genomic coordinates were present in COSMIC. OMIM does not include genomic coordinates. Mutations are recorded in the separate file **Table S2.4.xls**.

Hand curation of *TP53*

Given the high rate of *TP53* we examined the *TP53* gene by hand. 25 additional mutations were discovered and 20 retained after subsequent 3730 validation attempts. The results are shown in **Table S2.5**.

Table S2.5: Additional TP53 mutations discovered by hand curation

Patient Cen	ter	Classif ication	Chr	Start	End	Ref	Tum1 Tu	m2 T	ype	Transcript	Protein change
TCGA-10-0927 b	cm	DEL	17	7520294	7520306	TTGCTTGGGA CGG	- -	F	rame_Shift_Del	NM_001126113.1	p.P36fs
TCGA-13-0717	bcm	DEL	17	7518270	7518275	TGCCGC	- -	I	n_Frame_Del	NM_001126117.1	p.G112_M114> V
TCGA-23-1120	bcm	SNP	17	7520083	7520083	C	A A	M	issense	NM_001126113.1	p.R110L
TCGA-25-1634	bcm	DEL	17	7520201	7520202	GA	- -	F	rame_Shift_Del	NM_001126113.1	p.A70fs
TCGA-36-1574	bcm	SNP	17	7517845	7517845	C	T T	M	issense	NM_001126117.1	p.R141H
TCGA-36-1576	bcm	DEL	17	7514743	7514743	G	- -	F	rame_Shift_Del	NM_001126115.1	p.R205fs
TCGA-04-1342	broad I	NS	17	7517800 75	17801	-	T	T	Frame_Shift_Ins	NM_001126117.1	p.N156fs
TCGA-04-1356	broad S	NP	17	7518915 7	518915	T	C	C	Missense	NM_001126117.1	p.Y88C
TCGA-13-1510	broad S	NP	17	7518937 75	18937	G	A	A	Nonsense	NM_001126117.1	p.R81*
TCGA-23-1027	broad S	NP	17	7518982 7	518982	C	A	A	Nonsense	NM_001126117.1	p.E66*
TCGA-23-2079	broad S	NP	17	7520032 75	20032	C	A	A	Splice_Region	NM_001126117.1	p.T125_splice
TCGA-24-1431	broad S	NP	17	7519000 75	19000	G	A	A	Nonsense	NM_001126117.1	p.Q60*
TCGA-24-2035	broad D	EL	17	7517625 7	517626	GA	-	-	Frame_Shift_Del	NM_001126117.1	p.S183fs
TCGA-24-2280	broad S	NP	17	7519000 75	19000	G	A	A	Nonsense	NM_001126117.1	p.Q60*
TCGA-24-2298	broad S	NP	17	7518335 75	18335	T	C	C	Splice_site	NM_001126117.1	p.V93_splice
TCGA-25-2398	broad I	NS	17	7519193 75	19194	-	CCGGGCGGGG GTGTGGAATC AGTG	CCGGGCGGGG GTGTGGAATC AGTG	In_Frame_Ins N	M_001126117.1	p.G22>GTDSTP PPG
TCGA-59-2354	broad S	NP	17	7518915 7	518915	T	C	C	Missense	NM_001126117.1	p.Y88C

TCGA-61-2012	broad S	NP	17	7518915 7	518915	T	C	C	Missense	NM_001126117.1	p.Y88C
TCGA-10-0930	wustl	SNP	17	7520032	7520032	C	G G	S	plice_Region	NM_001126117.1	p.T125_splice
TCGA-13-0723 w	ustl	DEL	17	7519195	7519207	CGGGCGGGGG TGT	- -	F	rame_Shift_Del	NM_001126117.1	p.T18fs
TCGA-13-0897	wustl	INS	17	7517815	7517816	-	A A	F	rame_Shift_Ins	NM_001126117.1	p.R151fs
TCGA-13-1506 w	ustl	INS	17	7520215	7520216	-	ATTCTGGG	ATTCTGGG	Frame_Shift_Ins	NM_001126113.1	p.M66fs
TCGA-13-1506	wustl	SNP	17	7520228	7520228	C	T T M		issense	NM_001126113.1	p.E62K
TCGA-24-1417	wustl	SNP	17	7518264	7518264	G	A A	M	issense	NM_001126117.1	p.R116W
TCGA-24-1549 w	ustl	DEL	17	7518315	7518346	TGGTACAGTC AGAGCCAACC TAGGAGATAA CA	G G	S	plice_Site_Del	NM_001126117.1	p.V93_splice

Analysis of mRNA expression levels

We examined the mRNA expression level for the target genes. Most appeared to show at least some expression in the majority of samples. Two genes, *GABRA6* and *FAT3* did not (**Figure S2.1**).

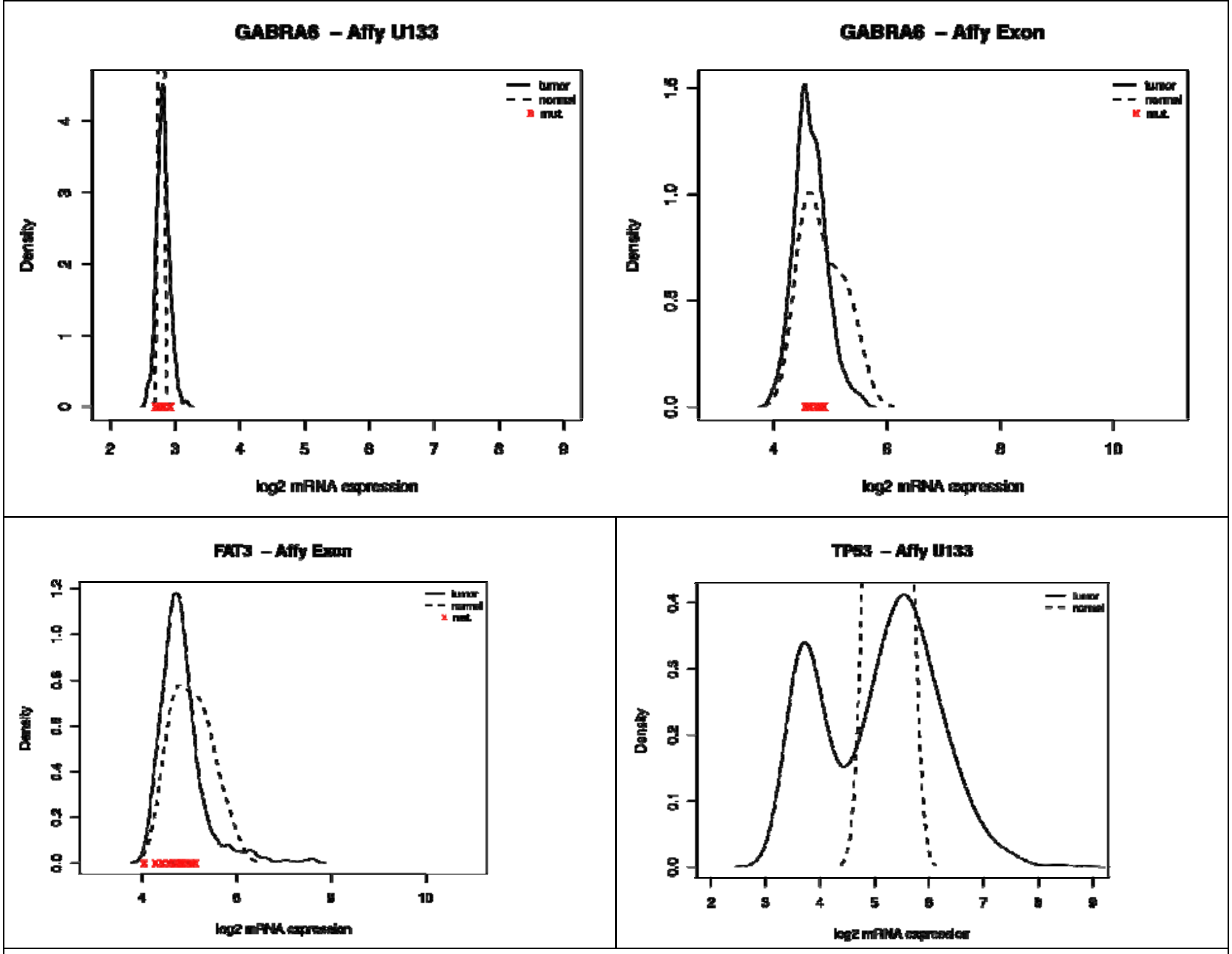


Figure S2.1. mRNA expression levels for three significantly mutated genes. *GABRA6* (top two panels) show very low (likely absent) expression in all tumors and in nearly all normal samples on U133A (left) and Exon array (right). *FAT3* shows very low expression in tumors and low expression in normals (lower left). *TP53* is well expressed in many samples and very low in others (see **Supplement S8**).

References

1. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182-189.
2. Getz, G., Hofling, H., Mesirov, J.P., Golub, T.R., Meyerson, M., Tibshirani, R., and Lander, E.S. 2007. Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* 317:1500.
3. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.
4. McKenna, A.H., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*.
5. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*.
6. Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 361:1058-1066.
7. Ning, Z., Cox, A.J., and Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res* 11:1725-1729.
8. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175-185.
9. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., et al. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 38:D613-619.
10. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. 2004. The Pfam protein families database. *Nucleic Acids Res* 32:D138-141.
11. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. The Pfam protein families database. *Nucleic Acids Res* 38:D211-222.
12. Mulder, N., and Apweiler, R. 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 396:59-70.
13. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* 446:153-158.
14. Rubin, A.F., and Green, P. 2009. Mutation patterns in cancer genomes. *Proc Natl Acad Sci U S A* 106:21766-21770.
15. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455:1069-1075.
16. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061-1068.

Supplementary Materials: Functional Mutations

Hannah Carter, Josue Samayoa, Dewey Kim, Rachel Karchin

We applied CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) [1, 2] to identify and prioritize somatic missense mutations most likely to generate functional changes that enhance tumor cell proliferation (drivers). These prioritizations are independent of mutation frequency, thus CHASM can potentially detect driver missense mutations and driver genes that would not otherwise be detected by methods dependent on mutation recurrence.

Excel worksheet with scores and associated annotations is available at
http://karchinlab.org/S3_Functional_Mutations_Supp.xls

Results

We identified 122 validated somatic missense mutations (in 113 genes) that are strong candidates for future studies, in spite of the fact that they did not occur in genes identified as significantly mutated (by frequency) in this TCGA study. These mutations occur in genes associated with several pathways potentially important for oncogenesis, including MAPK signaling, NFKB signaling, apoptosis, angiogenesis and inflammatory response pathways.

Of the 147 validated missense mutations that occurred in significantly mutated genes, 80 were scored as significant driver mutations by CHASM (FDR < 0.25). Seventy-six of these were in TP53, one in NF1 and three in CDK12 (CRKRS). Fifty-one TP53 mutations and one NF1 mutation were in the CHASM training set and 20 TP53 mutations occurred at positions that were in the training set but had different amino acid substitutions (Methods).

Methods

CHASM uses a Random Forest [3, 4] trained on a positive class of driver missense mutations and a negative class of passenger missense mutations. The positive class consists of 3299 missense mutations previously identified as playing a functional role in oncogenic transformation from the COSMIC database [5] and the negative class is synthetically generated with a computer algorithm that attempts to mimic patterns of random point mutations in a specific cancer type (*e.g.* serous ovarian carcinoma). The algorithm samples from eight multinomial distributions derived from background substitution rates (by di-nucleotide context) in the sequenced TCGA ovarian samples (S3 Table 1).

To avoid overfitting, the training set is split into two partitions, one of which is used for feature selection and the other to actually train the random forest classifier.

Each missense mutation is represented by 62 predictive features, including measures of evolutionary conservation, amino acid physiochemical properties, predicted protein structure, and annotations curated from the literature, retrieved from the UniProtKB database [6]. While an information theoretic analysis indicates that all of these features contribute to classification performance [1], the most important features in this study are (S3 Figure 1):

- Location in an enzymatic domain involved in post-translational modification;
- Compatibility with observed amino acid residues in an alignment of protein orthologs;
- Frequency of SNPs in the exon in which the mutation occurs;
- Average PhastCons [7] nucleotide-level conservation in the exon in which the mutation occurs;
- Fraction of sequence directly neighboring the mutated site composed of basic amino acids (K,R).

- Negative entropy in the column of amino acids that align to the mutated position in a protein superfamily multiple sequence alignment.

Discussion

A potential strength of random forest classifiers is that they are able to utilize interactions among predictive features. We explored visualization of higher-order feature interactions, using principal component analysis (PCA) [8], a transformation in which high-dimensional data is projected into a lower-dimensional co-ordinate system. In the transformed space, each dimension represents a weighted linear combination of the original data, and the first few dimensions span most of the dynamic range in the data. By visualizing these dimensions (the "principal components"), it is often possible to see statistical patterns and hidden structure in a data set.

Using PCA, we explored whether the driver and passenger mutations in the CHASM training set could be discriminated from each other without supervised learning, solely on the basis of differences in their feature distributions. We generated a matrix in which the rows represented the training set mutations plus the 122 ovarian somatic mutations and the columns represented the 20 most important features. PCA analysis yielded a projection of this data onto a three dimensional coordinate system (S3 Figure 2). In the transformed space, the passenger mutations tend to form a cluster (blue), while the drivers tend to radiate out from this cluster (red). Thus, there appears to be some separation between the two classes in a space that represents interacting features, even when the location of a mutation in the space does not depend on its class membership. Interestingly, most of the 122 predicted ovarian driver mutations (green) also radiate out from the cluster of passengers.

We also investigated whether the 122 mutations could be associated with regions of high mutation density ("hot spots"). One approach to identifying hot spots focuses on individual genes. However, we observed that the 122 mutations frequently occurred in the same few kinase families, including mitogen-activated protein kinases (MAPKs), cyclin-dependent kinases (CDKs), and serine-threonine kinases (STKs) and that these mutations tended to occur in hot spots in both primary protein sequence and tertiary protein structure. Kinase family mutation hot spots have been previously associated with disease-related mutations [9, 10].

Serine-threonine kinases (STKs). Four STKs contain mutations that were classified by CHASM as drivers (FDR<0.25) (STK10 L85P, STK38 K354N, STK33 T140M, STK38L L359I). We aligned these four sequences with CLUSTALW [11] and identified putative hot spots (S3 Figure 3). Because amino acid residues may be close together in tertiary protein structure, but not in primary amino acid sequence, we also mapped these mutations onto an X-ray crystal structure of STK10 (PDB ID 2j7t) [12] (S3 Figure 4). On the tertiary structure, the mutations appear to form two distinct hot spots, one of which is in close proximity to the kinase active site.

Mitogen-activated and cyclin-dependent kinases (MAPKs and CDKs). Six MAPKs and 7 CDKs contain mutations classified by CHASM as drivers (FDR<0.25). We identified putative hotspots in the alignments of these MAPK and CDK proteins (data not shown). However, these proteins are closely related (pairwise BLAST E-values << 0.001) and we explored whether we could identify more hotspots by combining them in a single alignment. We also added the kinases GSK3A and RAGE (pairwise BLAST E-values << 0.001) to the alignment (S3 Figure 5). This alignment reveals a hotspot (GSK3A I122M, CDC7 D62G) proximal to the kinase glycine-rich loop (which plays a critical role in ordering of the ATP binding pocket [13]) and several other putative hotspots. Notably, mutations are seen at the equivalent aspartic acid in two CDKs (D73 in CDK1 and D90 in CDK19). This kind of event has previously been associated with important somatic mutations in cancer, for example mutation at R132 in IDH1 and its equivalent R172 in IDH2 were both shown to be drivers in glioblastoma multiforme [14]. We mapped the mutations onto an X-ray crystal structure of MAP3K7 (PDB ID 2eva) [15]. With respect to tertiary structure, there appear to be three main hot spots: (CDC7 D62G, GSK3A I122M); (MAP3K7 G45V, MAP2K6 V163G, CDK17 R312G, CDK17 K343N, CDK15 V210M, MAP3K13 V317W, CDK12 Y901C, MK08 L198F, M4K3 D196N); and (CDK8 E165Q, RAGE Q257P, CDK12 L996F). Notably, the second hot spot clusters tightly around the kinase active site (ATP binding pocket) and includes a mutation (CDK17 R312G) within the highly conserved HRD motif. Furthermore, this analysis reveals clustering in three dimensional space for residues distant in primary sequence. Such clustering would not be expected by chance and suggests key functional roles for these regions.

While the functional relevance of some of these hot spots is unclear, these results provide additional support for the hypothesis that regions of high mutation density exist in kinases and that identifying these hotspots may be useful in discovering driver mutations in cancer.

References

1. Carter, H., et al., *Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.* Cancer Res, 2009. **69**(16): p. 6660-7.
2. Carter, H., et al., *Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM).* Cancer Biol Ther, 2010. **10**(6).
3. Amit, Y. and D. Geman, *Shape quantization and recognition with randomized trees.* Neural computation, 1997. **9**(7): p. 1545-1588.
4. Breiman, L., *Random forest.* Machine Learning, 2001. **45**: p. 5-32.
5. Forbes, S., et al., *Cosmic 2005.* Br J Cancer, 2006. **94**(2): p. 318-22.
6. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information*. 2006.
7. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
8. Wood, F., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometr. Intel. Lab. Syst, 1987. **2**: p. 37-52.
9. Yue, P., et al., *Inferring the functional effects of mutation through clusters of mutations in homologous proteins.* Hum Mutat, 2010. **31**(3): p. 264-71.
10. Dixit, A., et al., *Sequence and structure signatures of cancer mutation hotspots in protein kinases.* PLoS One, 2009. **4**(10): p. e7485.
11. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.
12. Pike, A.C., et al., *Activation segment dimerization: a mechanism for kinase autophosphorylation of non-consensus sites.* EMBO J, 2008. **27**(4): p. 704-14.
13. Kornev, A.P., et al., *Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17783-8.
14. Yan, H., et al., *IDH1 and IDH2 Mutations in Gliomas.* The New England Journal of Medicine, 2009. **360**(8): p. 765.
15. Brown, K., et al., *Structural basis for the interaction of TAK1 kinase with its activating protein TAB1.* J Mol Biol, 2005. **354**(5): p. 1013-20.
16. van der Maaten, L.J.P., E.O. Postma, and H.J. van den Herik, *Dimensionality Reduction: A Comparative Review*. 2007.

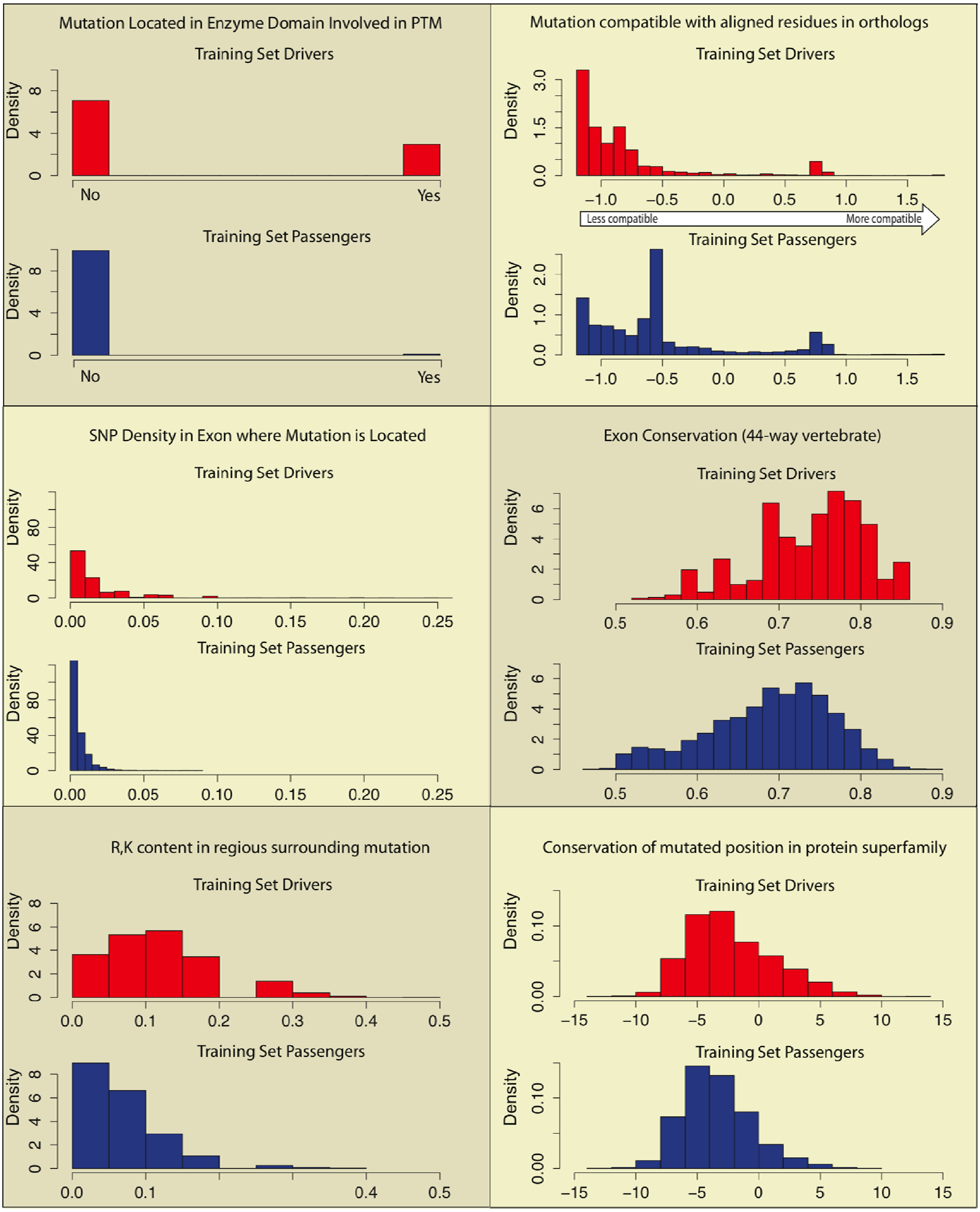
S3 Tables

S3 Table 1. Eight multinomial distributions represent the (normalized) rates of each type of nucleotide substitution in serous ovarian carcinoma, depending on di-nucleotide context. Frequencies provided by Mike Lawrence of the Broad Institute.

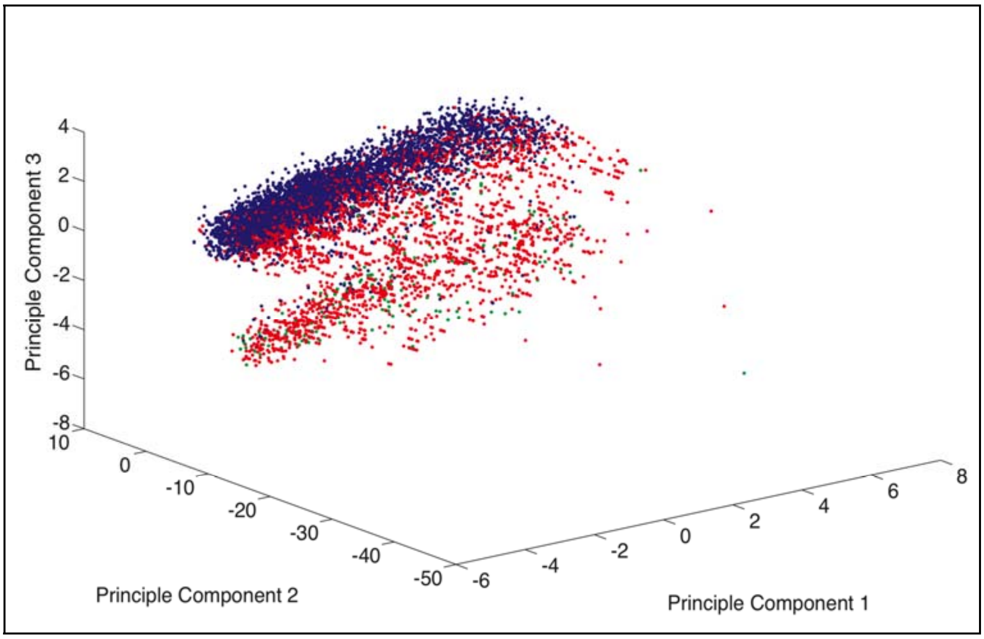
	C in CpG	G in CpG	C in TpC	G in GpA	A	C	G	T
A	0.13	0.77	0.27	0.40	0.00	0.39	0.36	0.28
C	0.00	0.09	0.00	0.37	0.28	0.00	0.23	0.48
G	0.09	0.00	0.30	0.00	0.42	0.22	0.00	0.24
T	0.78	0.14	0.43	0.23	0.30	0.39	0.41	0.00

S3 Figures

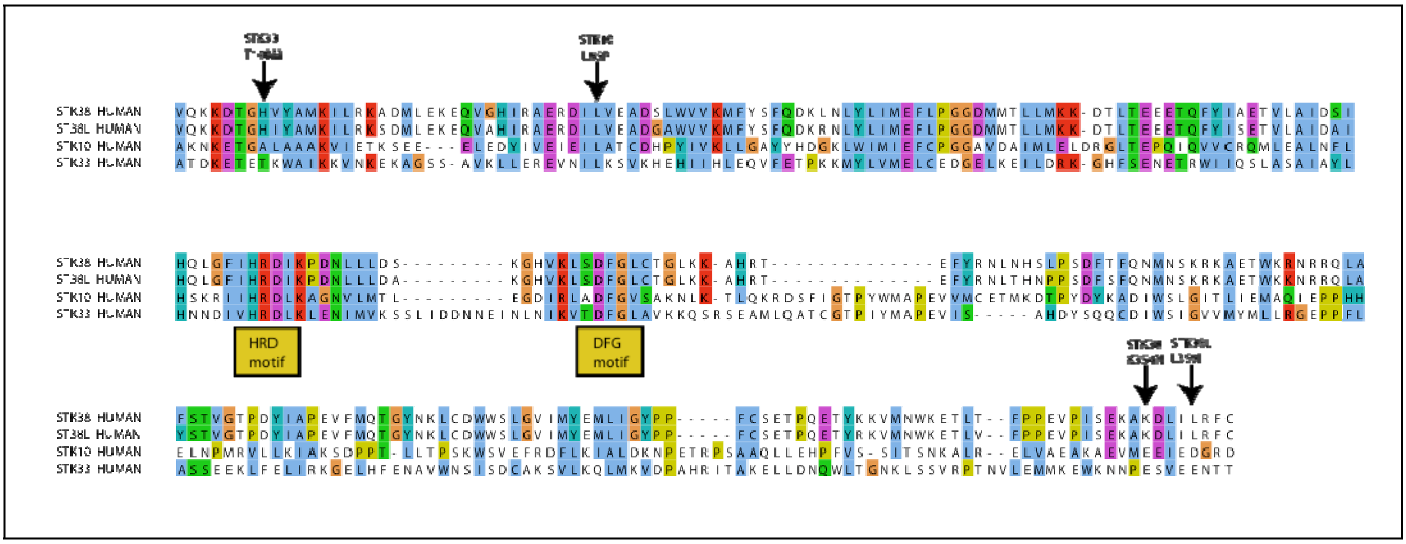
S3 Figure 1. Histogram analysis of the most important mutation features identified in this study. Feature value distributions differ between the driver and passenger missense mutations in the CHASM training set



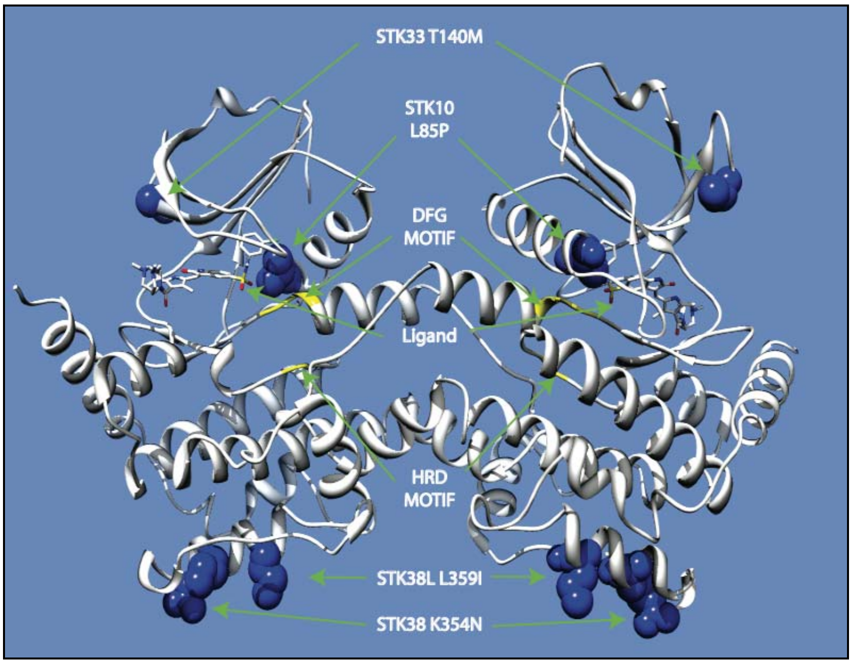
S3 Figure 2. Principal components analysis reveals substructure in the CHASM training set driver mutations (red), passenger mutations (blue) and 122 predicted ovarian driver mutations (green). In the transformed space, the predicted drivers are distributed more like the training set drivers than the training set passengers. Plot done with Matlab Toolbox for Dimensionality Reduction [16].



S3 Figure 3. Serine-threonine kinases (STKs) with missense mutations in this study that were scored as significant or weakly significant by CHASM. Sequences were aligned with CLUSTALW and mutations were mapped onto alignment columns. Two mutational hot spots [STK33 T140M, STK10 L85P] and [STK38 K354N, STK38L L359I] cluster in three dimensions (see also, S3 Figure4). Also shown are the highly conserved kinase "HRD" and "DFG" motifs, which are involved in metal ion coordination and ATP binding [13].

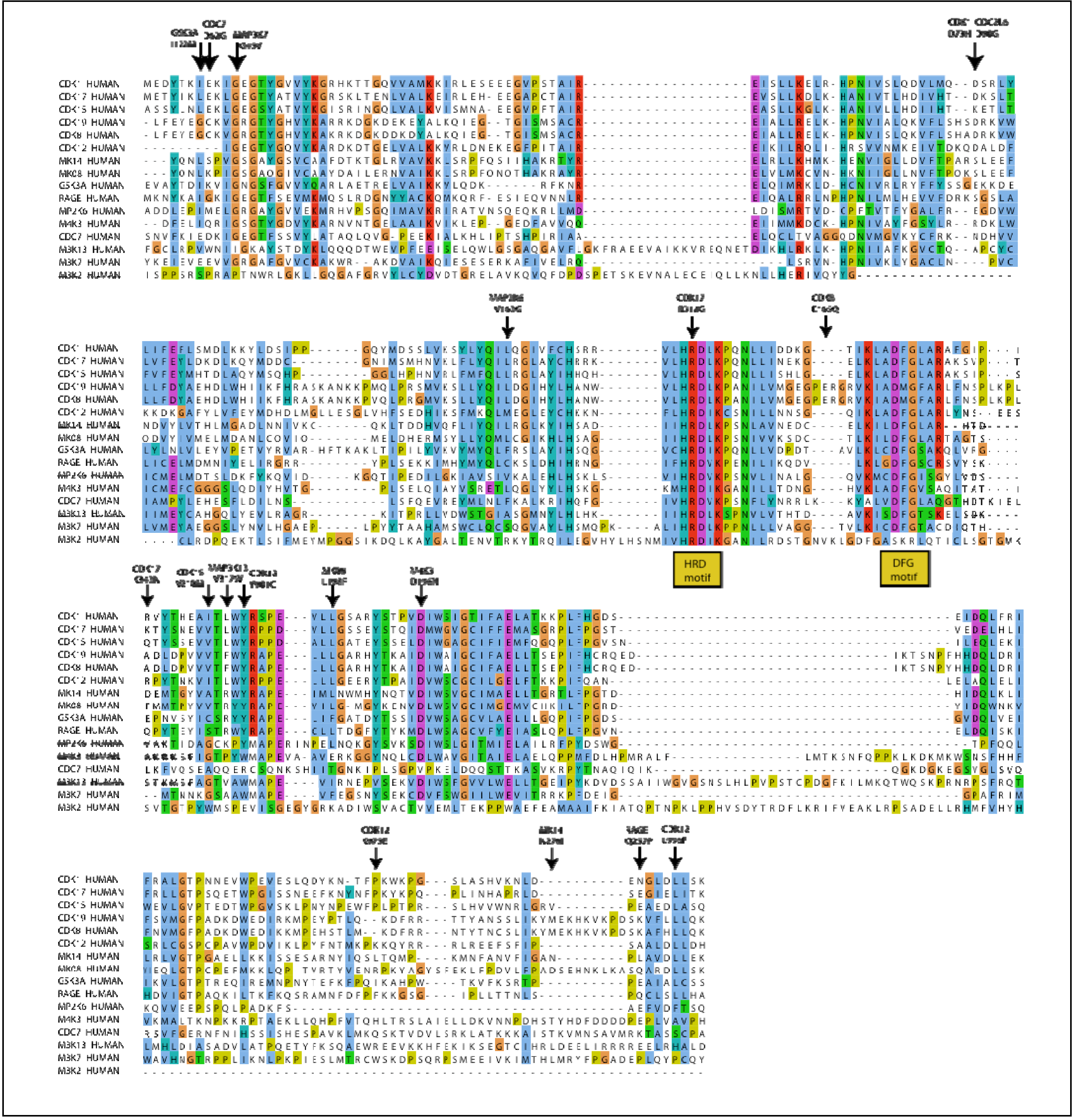

STK33 T140M
STK10 L85P
STK38 HUMAN VQKKDTGHVYAMKILRKADMLEKEQVGHIRAERDILVEADSLWVVKMFYSFQDKLNLYLIMEFLPGGDMMTLLMKK-DTLTEEETQFYIAETVLAIDSI
STK38L HUMAN VQKKDTGHIYAMKILRKSDMLEKEQVAHIRAERDILVEADGAWVVKMFYSFQDKRNLYLIMEFLPGGDMMTLLMKK-DTLTEEETQFYISETVLAIDAI
STK10 HUMAN AKNKETGALAAAKVIETKSEE---ELEDYIVEIEILATCDHPYIVKLLGAYYHDGKLWIMIEFCPGGAVDAIMLELDRGLTEPQIQVVCRQMLEALNFL
STK33 HUMAN ATDKETETKWAIKKVNKEKAGSS-AVKLLEREVNILKSVKHEHIIHLEQVFETPKKMYLVMELCEDGELKEILDRK-GHFSENETRWIIQSLASAIAYL
STK38 HUMAN HQLGFIHRDIKPDNLLLDS----------KGHVKLSDFGLCTGLKK-AHRT----------------EFYRNLNHSLPSDFTFQNMNSKRKAETWKRNRRQLA
STK38L HUMAN HQLGFIHRDIKPDNLLLDA----------KGHVKLSDFGLCTGLKK-AHRT----------------EFYRNLTHNPPSDFSFQNMNSKRKAETWKKNRRQLA
STK10 HUMAN HSKRIIHRDLKAGNVLMTL----------EGDIRLADFGVSAKNLK-TLQKRDSFIGTPYWMAPEVVMCETMKDTPYDYKADIWSLGITLIEMAQIEPPHH
STK33 HUMAN HNNDIVHRDLKLENIMVKSSLIDDNNEINLNIKVTDFGLAVKKQSRSEAMLQATCGTPIYMAPEVIS-----AHDYSQQCDIWSIGVVMYMLLRGEPPFL
HRD motif
DFG motif
STK38 K354N
STK38L L359I
STK38 HUMAN FSTVGTPDYIAPEVFMQTGYNKLCDWWSLGVIMYEMLIGYPP-----FCSETPQETYKKVMNWKETLT--FPPEVPISEKAKDLILRFC
STK38L HUMAN YSTVGTPDYIAPEVFMQTGYNKLCDWWSLGVIMYEMLIGYPP-----FCSETPQETYRKVMNWKETLV--FPPEVPISEKAKDLILRFC
STK10 HUMAN ELNPMRVLLKIAKSDPPT-LLTPSKWSVEFRDFLKIALDKNPETRPSAAQLLEHPFVS-SITSNKALR--ELVAEAKAEVMEEIEDGRD
STK33 HUMAN ASSEEKLFELIRKGELHFENAVWNSISDCAKSVLKQLMKVDPAHRITAKELLDNQWLTGNKLSSVRPTNVLEMMKEWKNNPESVEENTT

S3 Figure 4. STK mutations mapped onto the X-ray crystal structure of STK10 (PDB ID 2j7t) [12]. The location of the mutations in the protein's tertiary structure shows that there are two mutation hot spots (protein is shown in dimerized form). The hotspot containing [STK33 T140M and STK10 L85P] is in close proximity to the kinase active site.

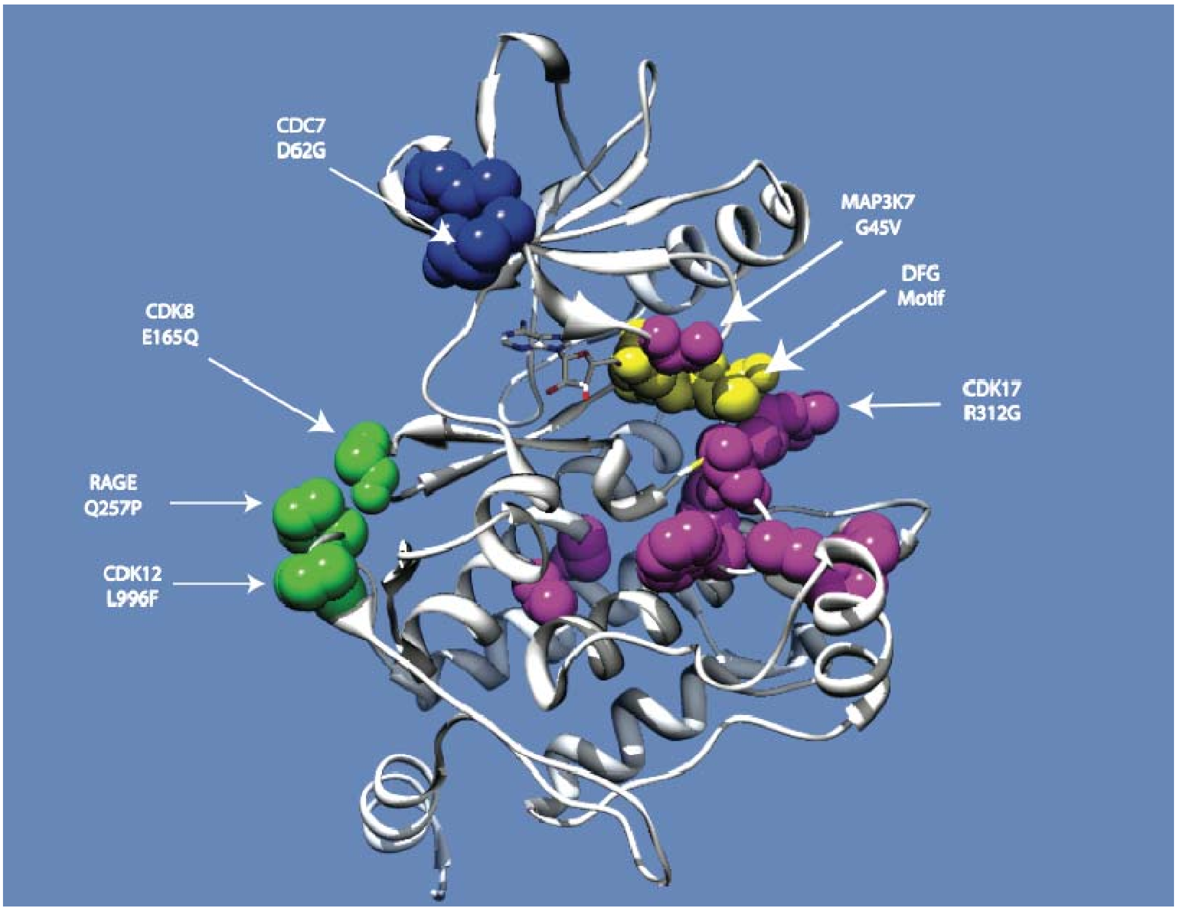


doi:10.1038/nature10166

S3 Figure 5. MAPKs, CDKs, GSK3A and RAGE contain mutations scored as significant by CHASM. Sequences were aligned with CLUSTALW and mutations were mapped onto alignment columns. Several putative hot spots were identified. The CDK1 D73H and CDK19 D90G mutations map to the same alignment column and are most likely functionally equivalent.



S3 Figure 6. MAPK, CDK, GSK3A and RAGE mutations predicted as significant by CHASM mapped onto an X-ray crystal structure of MAP3K7 (PDB ID 2eva). The location of the mutations suggests three hot spots: (CDC7 D62G, GSK3A I122M) shown in blue; (MAP3K7 G45V, MAP2K6 V163G, CDK17 R312G, CDK17 K343N, CDK15 V210M, MAP3K13 V317W, CDK12 Y901C, MK08 L198F, M4K3 D196N) shown in magenta; and (CDK8 E165Q, RAGE Q257P, CDK12 L996F) shown in green. The first and second clusters are located in close proximity to the glycine-rich loop and kinase active site (ATP binding pocket) respectively. Furthermore, the second cluster, magenta, contains a residue in the highly conserved HRD motif (CDK17 R312G). Notice that mutations CDK8 E165Q and MAPK3 G45V cluster in three dimensional space with residues far away in primary sequence (S3 Figure 5).



Supplemental Methods S4:

Functional mutations in ovarian cancer

Method

To predict the functional impact of protein missense mutations, we used a new computational approach [1], which is based on the assessment of evolutionary conservation of amino-acid residues in a protein family multiple sequence alignment. Given the genomic coordinates of a missense mutation, the reference base and the substituted nucleotide, a fully automated computational protocol (available at http://mutationassessor.org) searches for sequence homologs, builds a multiple sequence alignment, clusters sequences into subfamilies and scores a mutation by the whole-family conservation and the conservation within a particular subfamily [2]. Mutations that affect conserved residues are more likely to be functional. The scoring method was validated and calibrated by separation of a large set of disease-associated variants (~20K) from common (benign) polymorphisms (~35K) with an accuracy of ~80% [1].

The automated computational protocol integrates various types of information related to the functional impact of a mutation: it determines the position of a mutated residue in the 3D structure of a mutated protein and its sequence homologs; it determines binding sites using available 3D complexes; it returns functional information on a protein region affected by a mutation; it also reports known cancer mutations and functional variants affecting a mutated position and various other cancer-related annotations.

When assessing the impact of ovarian cancer mutations, we also took into account the gene expression level of a mutated protein. Mutations in genes that are not expressed are not expected to have a functional impact. Specifically, we computed a percentile of genes in a sample that are expressed at a lower level than the mutated gene.

All functional information is summarized in 22 annotation columns added to the supplemental mutations **Table_S2.1.MAF** (external MAF file). See below for detailed descriptions of the annotation columns.

Results

For mutations that change an amino acid (missense), we predicted the likely functional impact on protein function using a combination of evolutionary information from protein-family sequence alignments and residue placement in known or homology-deduced three-dimensional protein and complex structures. We compared the distribution of predicted functional ovarian cancer mutations to disease-associated, polymorphic and COSMIC mutations, and we show that ovarian cancer mutations are highly enriched in functional mutations compared to polymorphic variants (**Figure S4.1**).

Including the 2070 truncating mutations (deletions, insertions, nonsense, splicesites), in-frame deletions and insertions, and taking into account only the 2939 predicted functional missense mutations in genes that are well expressed (the highest 70% of expressed genes in a sample), we predict that 5009 (26%) of the total 19359 somatic mutations affect protein function (**Figure S4.2**).

It is plausible that only a fraction of these functional mutations contribute causally to oncogenesis or disease progression, but this fraction is generally unknown. Selecting missense mutations predicted to affect protein function with a high score in genes known to be functionally involved in at least one type of cancer (source: MSKCC CancerGenes [3] & COSMIC [4]) results in 387 (2%) candidate oncogenic mutations, 183 of which are mutations in *TP53* (**Figure S4.2**).

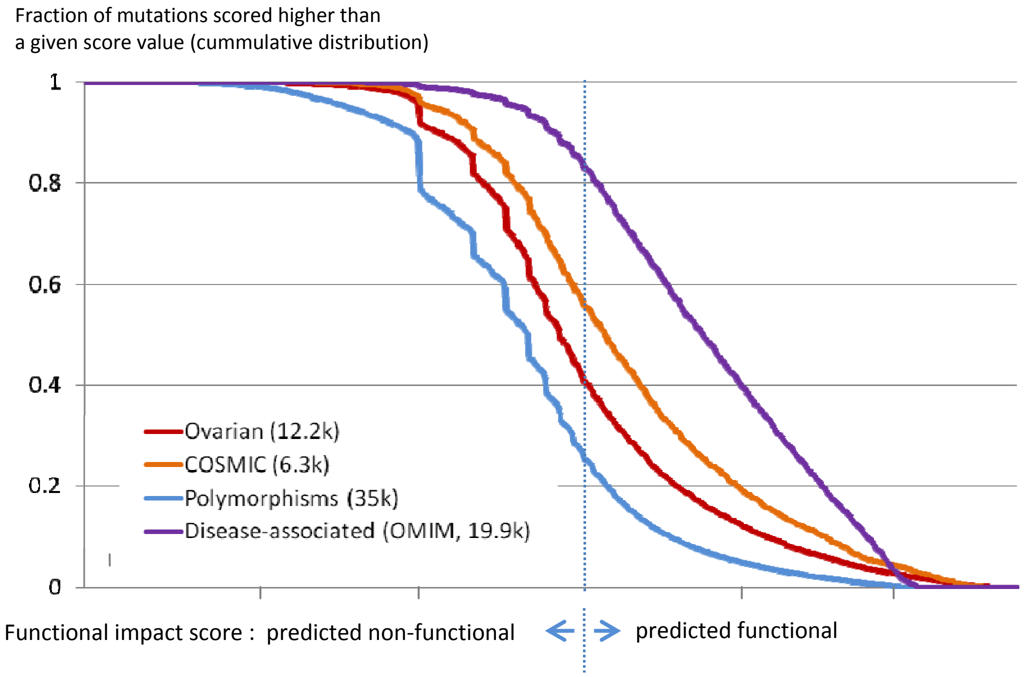


Figure S4.1. Distribution of ovarian cancer mutations by computed functional impact score. A higher score indicates an increased likelihood that a given mutation is functional and possibly causative for disease. The functional impact scores of mutations in the ovarian cancer data set are compared to mutations listed in the Human Polymorphisms and Disease Mutations Index (HUMSAVAR, release 2010_08), and in the Catalogue of Somatic Mutations in Cancer (COSMIC, v45) [4]. The optimal separation (~80%) between two variant classes – disease-associated and polymorphic - is achieved at the score threshold of ~1 [1]. There is an enrichment of ovarian cancer and COSMIC mutations at higher scores compared to polymorphic variants, which are presumed to be non-functional. About 56% of the 6.3K Cosmic mutations scored higher than the threshold value. Of the 12943 ovarian cancer missense mutations, the functional impact was assessed for 12170 mutations, and among them 4897 (~38% of the total missense mutations) scored higher than the threshold value. Plausibly, only a subset of these mutations is oncogenic.

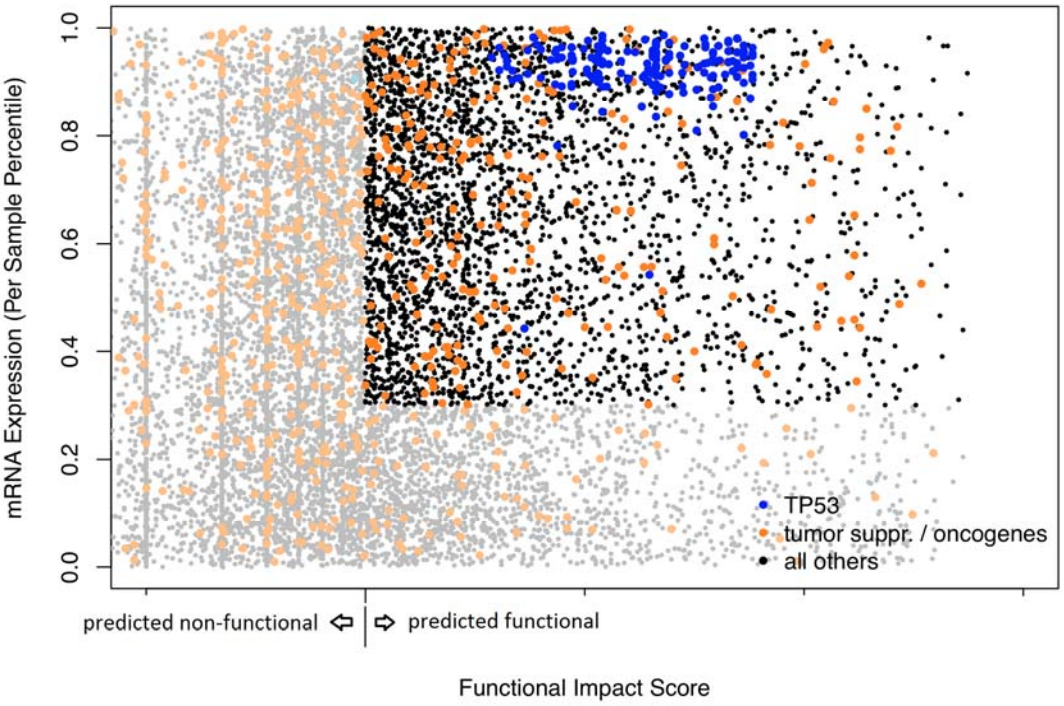


Figure S4.2. A quarter of the ovarian cancer mutations are expressed and predicted to be functional. A significant fraction of genes affected by likely functional mutations have low expression levels and are unlikely to have much functional impact. It total, 2939 mutations are selected that exceed a threshold value of predicted significant functional impact (1.0) and have moderate to high mRNA levels (>0.3). All but two of the observed missense mutations in TP53 (185 unique mutations) are above these thresholds (blue) and therefore are likely to negatively affect the function of the TP53 tumor repressor. The predicted high functional impact of TP53 mutations serves as a positive control of the accuracy of these predictions, if one assumes that there is independent knowledge about the functional role of TP53 alterations in ovarian cancer.

Description of the annotation columns produced by automated protocol for assessment of functional impact of mutations (http://mutationassessment.org) added to the supplemental mutations table:

- **MA:variant** - genomic position, reference nucleotide, observed nucleotide (mutations validated as "wildtype" are ignored)
- **MA:GE.rank** - mRNA expression percentile of the mutated gene in the mutated sample (1 = gene with the highest expression in the sample, 0 = gene with the lowest expression, based on Affymetrix U133 data)
- **MA:CNA** - GISTIC [5] DNA copy number gene status for the gene in the mutated sample (-2 = homozygous deletion, -1 = hemizygous deletion, 0 = neutral, 1 = gain, 2 = high level amplification)
- **MA:OV.variant.samples** - total number of samples with the same exact mutation
- **MA:OV.gene.samples** - total number of samples with a mutation in the affected gene
- **MA:mapping.issue** - explanation if a mutation could not be analyzed (no Uniprot ID is the most common reason)
- **MA:FImpact** - predicted functional impact category
- **MA:FI.score** - predicted functional impact score [1]
- **MA:Func.region** - 1 = mutated position is within one of the following regions annotated by Uniprot: CARBOHYD, CA_BIND, CROSSLNK, DISULFID, DNA_BIND, METAL, MOD_RES, MOTIF, NON_STD, NP_BIND, SITE, ZN_FING
- **MA:bindsite.protein** - the mutated residue has at least one heavy atom at a distance of less than 4.5A to another protein (including protein dimers)
- **MA:bindsite.DNA/RNA** - the mutated residue has at least one heavy atom at a distance of less than 4.5A to a DNA or RNA molecule
- **MA:bindsite.sm.mol** - the mutated residue has at least one heavy atom at a distance of less than 4.5A to a small molecule. The following small molecules are ignored: PO4, PI, SO4,S UL, CL, BR, NO3, SCN, NH4, K, NA ,LI, MG, DOD, NAG, MAN, GOL, SO4, CL, CO3, FS4 (Polyphen)
- **MA:CancerGenes** - all annotations by CancerGenes
- **MA:TS** - 1 = gene is annotated as a "tumor suppressor" by CancerGenes
- **MA:OG** - 1 = gene is annotated as an "oncogene" by CancerGenes
- **MA:COSMIC.mutations** - list of all Cosmic mutations in the affected position
- **MA:COSMIC.cancers** - list of cancer types in which Cosmic mutations in the affected position of this gene were detected
- **MA:Uniprot.regions** - all affected Uniprot regions (Feature // Description; /// separates multiple regions)
- **MA:TS.interacts** - 1 = protein interacts with another tumor suppressor (interactions from PIANA, mainly HPRD, TS annotation from CancerGenes)
- **MA:OG.interacts** - 1 = protein interacts with another oncogene (interactions from PIANA, mainly HPRD, OG annotation from CancerGenes)
- **MA:Pfam.domain** - Pfam domain affected by the mutation
- **MA:link.var** - hyperlink to a summary analysis table (at http://mutationassessor.org)
- **MA:link.PDB** - hyperlink to a PDB Jmol page with the mutated residue highlighted
- **MA:link.MSA** - hyperlink to a multiple sequence alignment (MSA) with the mutation column highlighted

References

1. Reva, B.A., Antipin, Y.A. and Sander, C. (2010) Functional Impact of Protein Cancer Mutations: Assessment based on Evolutionary Information, in preparation (in submission to *Nucl. Acids Res.*)

2. Reva, BA, Antipin, YA, and Sander, C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol*.; *8*(11)*:* R232.

3. Higgins, M.E., Claremont, M., Major, J.E., Sander, C. and Lash, A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res*, **35**, D721-726.

4. Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A. and Stratton, M.R. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*, *Chapter 10, Unit 10 11*.

5. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* 104(50):20007-12. Epub 2007 Dec 6.

Supplementary Methods S5: Copy Number Analysis

Segmented copy number profiles for ovarian carcinoma and matched control DNAs (489 Agilent 1M arrays for tumor and for normal, or 978 total – see Table 1) 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 (for details, see Beroukhim *et al.*[1] and Mermel et al, submitted). As has been previously reported for multiple cancer types[1], across all samples the copy number events can be clearly divided into at least two classes on the basis of their observed frequency: focal copy number events much smaller than a chromosome arm, and broad copy number events that span a chromosome arm or entire chromosome (see **Figure S3.1**). As broad and focal copy number changes appear to have markedly different rates of occurrence, and may have distinct biological consequences, we analyzed them separately. A length threshold of 50% of a chromosome arm was used to distinguish between broad and focal events (see **Figures S3.2** and **S3.3**). To remove false positive segments resulting from hyper-segmentation, we further filtered segments using an amplitude threshold at a copy-difference of 0.1 (data not shown).

For broad copy number changes, the frequency with which chromosomal arms are measured to undergo gain or loss is negatively correlated with the size of that arm (**Figure S3.4**). To determine which arms were significantly enriched/depleted among copy gains and losses after accounting for this trend, we compared the expected frequency of gain and loss for each arm, determined by linear regression, with the actual frequency observed over the entire dataset. Since samples with gain of a chromosome arm cannot have loss of the same arm, we computed the frequency of gains and loss among the undeleted and unamplified samples, respectively. By decoupling the gains and losses in this way, the frequency metric follows a binomial distribution; z-scores for each arm were calculated using the normal approximation to the binomial (**Table S5.1**), and the resulting one-sided p-values were corrected for multiple hypotheses testing using the Benjamini-Hochberg FDR method. The frequency of samples that have segments whose length is at least 50% of a chromosome arm, displaying gains (relative copy number >2.1) and losses (relative copy number <1.1) is shown for each chromosome arm (**Table S5.1**).

Focal copy number changes in the 489 ovarian carcinoma DNA samples were analyzed using the GISTIC methodology[2] with modifications as described in further detail in (Mermel et al, submitted). Briefly, each marker was scored according to the mean amplitude and frequency of focal amplification across the dataset, and significance values were computed by comparing to the distribution of scores obtained by random permutation of the markers across the genome. Significant peak regions of amplification (or deletion) are identified using an iterative peel-off procedure that distributes the score associated with amplified (or deleted) segments among all peaks that overlap them (weighted according to each peak's score) until no new region crosses the significance threshold of q-value ≤0 .25 on each chromosome. Finally, by taking into account the auto-correlation within the GISTIC score profiles, we compute for each peak region a confidence interval that is predicted to contain the true driver gene or genes with at least 99% probability (see Supplementary methods of Beroukhim et al. [1], Mermel et al, submitted). The output of focal GISTIC, defining the key peaks of amplification and deletion in the 489 ovarian carcinoma DNA samples, appears in **Table S5.2**.

Regions are defined as possessing deep deletions, shallow deletions, neutral copy number, low gain, and high gain in each sample using sample-specific thresholds as previously described[3]. In brief, high gains are segments with copy number that exceed the maximum median chromosomal arm copy number for that sample by at least 0.1; low gains are segments with copy numbers from 2.1 to the high gain threshold; neutral segments have copy numbers between 1.9 and 2.1; shallow losses have copy numbers between 1.9 and the deep deletion threshold; and deep deletions have copy numbers that are below the minimum median chromosomal arm copy number for that sample by at least 0.1. Frequencies of all these events are tallied across the 489 ovarian carcinoma samples (**Table S5.2**).

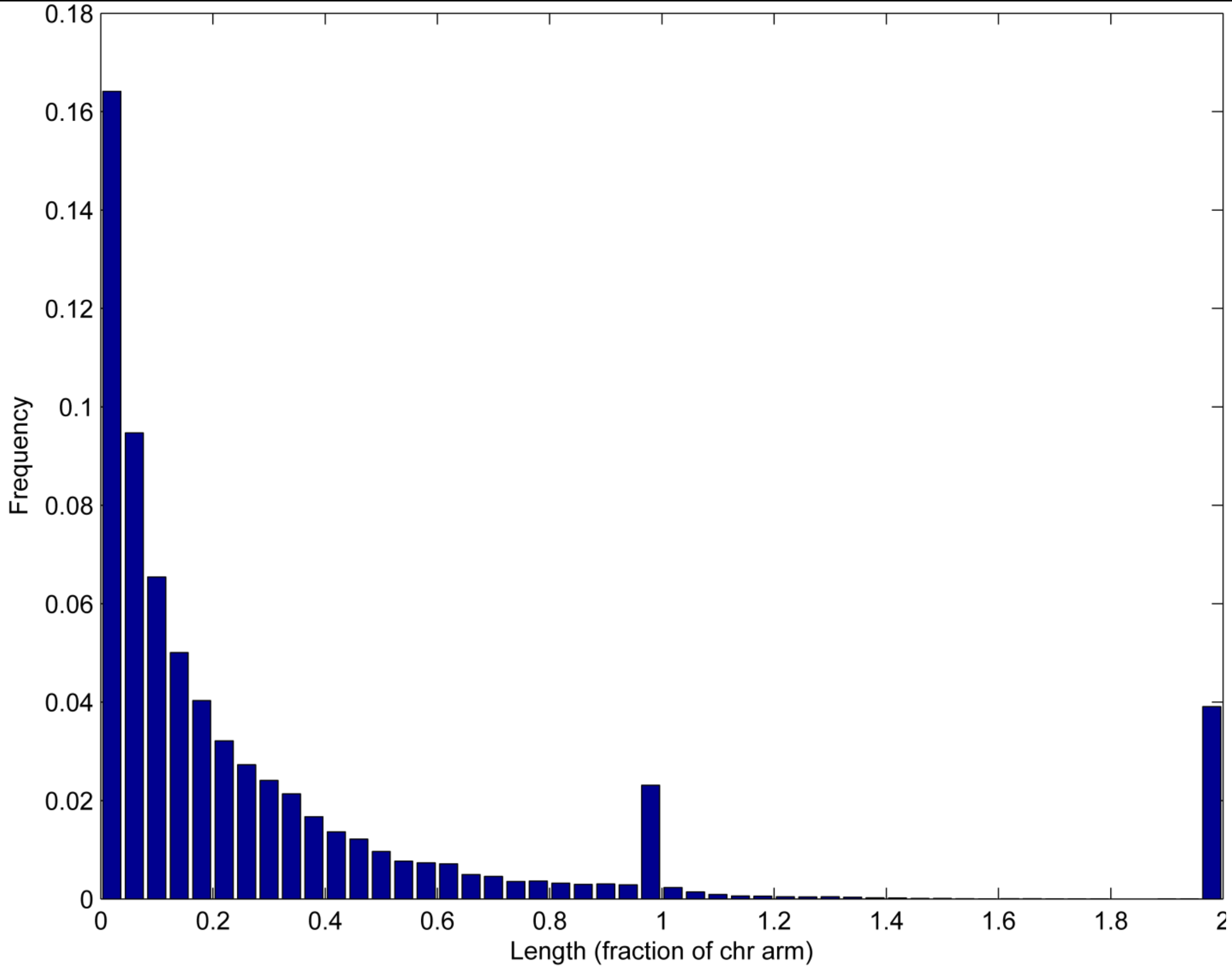
A subset of genes located in the 63 recurrent focal amplification regions were selected based on an Ingenuity database search and visual scanning genes of interest. We searched the subset of genes against information obtained from DrugBank (http://www.drugbank.ca)[4,5], including drug and associated target information on 1589 genes and 2010 drug entries. Since we are searching for inhibitors target the frequently amplified and actively expressed genes, we retained only inhibitors from the resulting list. The list was further manually curated using http://clinicaltrials.gov and the literature. The final list of 22 genes and their corresponding therapeutic compounds is in **Table S5.3**.

References

1. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905.
2. Beroukhim, R. et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007-12 (2007).
3. TCGA Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
4. Wishart, D.S., et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6.
5. Wishart, D.S., et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.

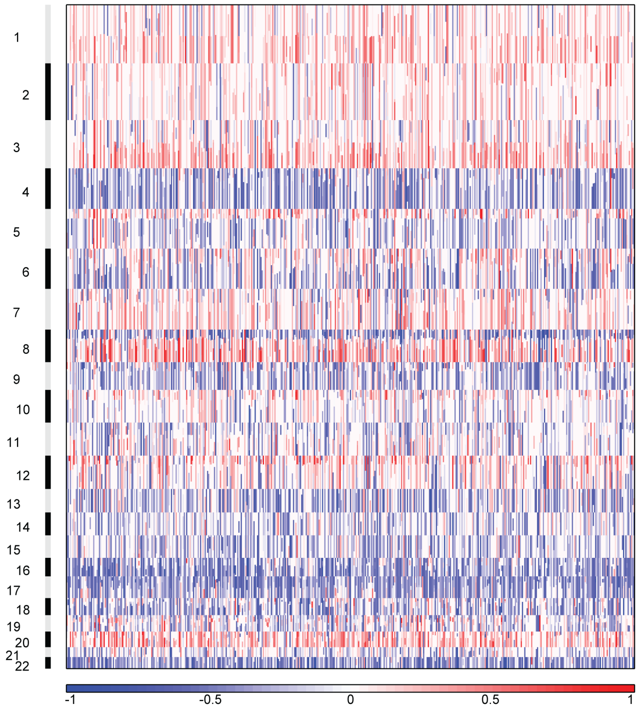
Supplementary Figures

Figure S5.1 Histogram distribution of copy number change frequency scaled by chromosome arm length. Note peaks at 1 and 2 indicating arm or chromosome gains.



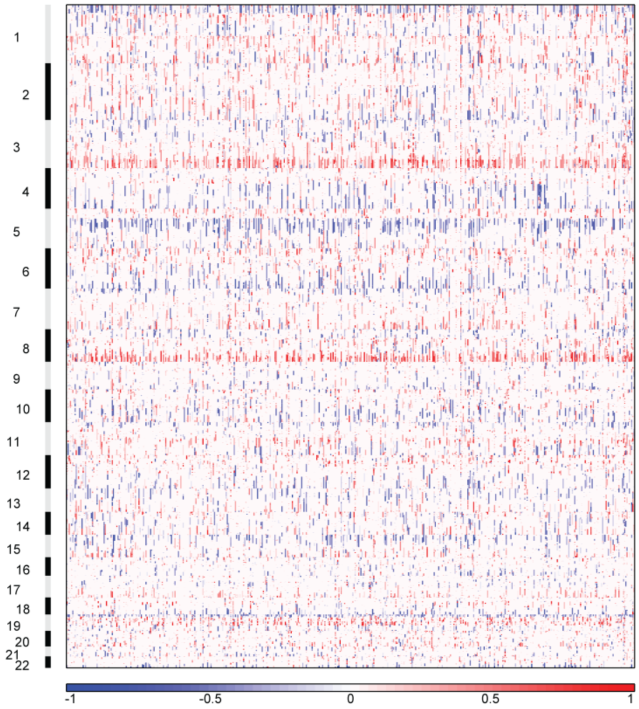
Supplementary Figure S5.1

Figure S5.2. Arm events found in ovarian serous carcinoma.



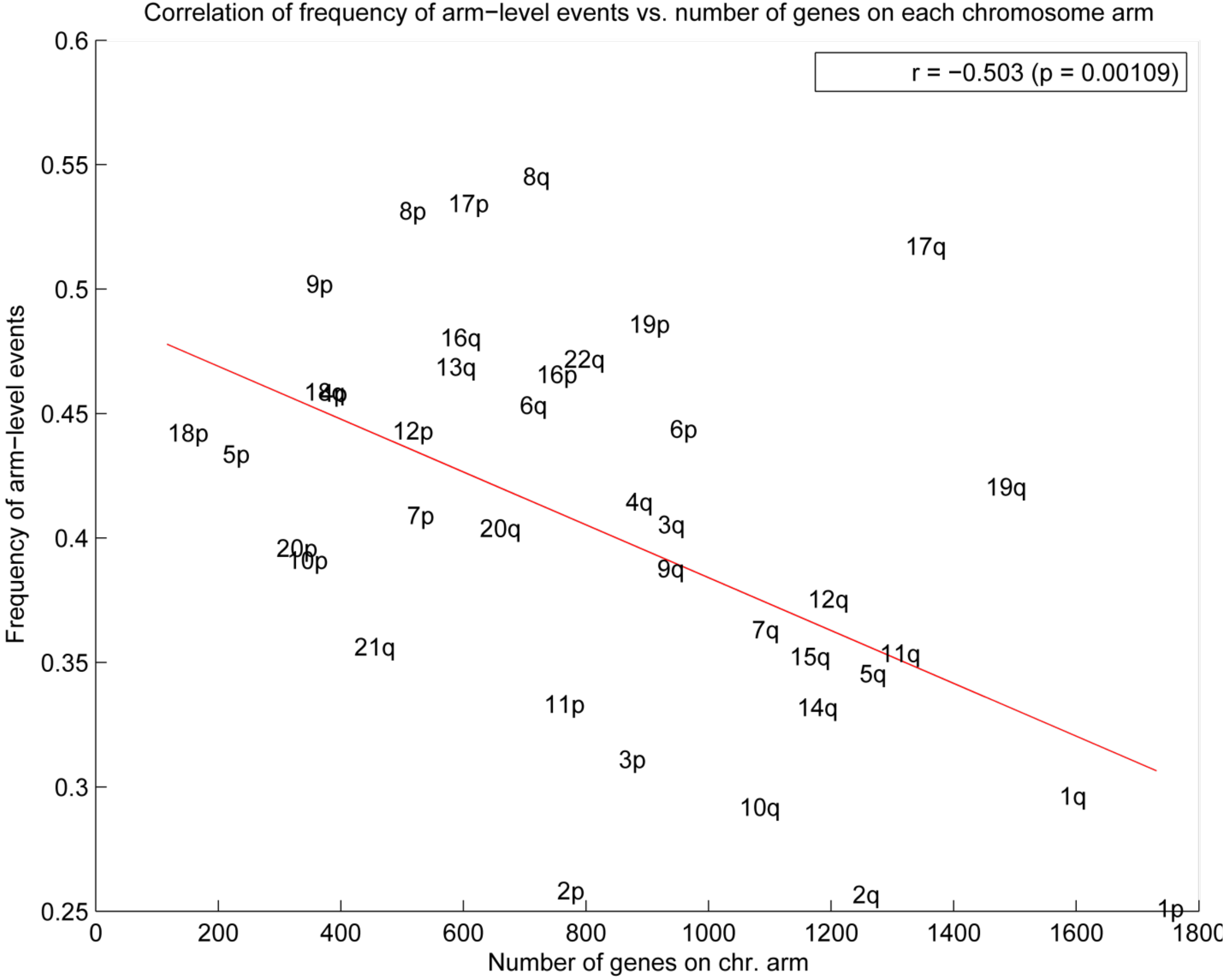
Supplementary Figure S5.2

Figure S5.3. Focal copy number changes plotted alone.



Supplementary Figure S5.3

Figure S5.4. Correlation between gene number and loss frequency.



Supplementary Figure S5.4

Supplementary Methods S6. mRNA and miRNA Expression Profiling

A. Affymetrix Exon 1.0

Sample verification and RNA QC

Total RNA samples were received from Biospecimen Core Resource (BCR). Samples were normalized to approximately 100ng/ul concentration to perform the sample QC. Total RNA concentration, quality and protein contamination were determined by Nanodrop measurements. RNA integrity number (RIN) and 28s/18s ratio were determined by the Bioanalyzer (Agilent, Santa Clara, CA). To evaluate the possible DNA contamination in RNA, quantitative RT-PCR was performed using iScript one Step RT-PCR Kit SYBR Green assay, and delta CT values were computed against controls to check if the samples exceeded the genomic DNA contamination of 10ng/ul. All the quality values computed at LBNL CGCC were used to compare to the quality data provided by BCR. For each microarray experiment, with each batch of ovarian samples, we included three universal RNA samples as controls for the experiment. We used Universal Human Reference RNA (Stratagene) Cat# 740000 (Stratagene, La Jolla, CA.), Human Universal Reference Total RNA Cat# 636538 (BD Clontech, Palo Alto, CA), and Ovarian total RNA Cat# R1234035-50 (Biochain Institute, Hayward, CA).

Whole transcript sense target labeling assay

2 µg of total RNA was subjected to ribosomal RNA removal procedure using Ribominus kit by Invitrogen Corporation (Carlsbad, CA). Double-stranded cDNA was synthesized from rRNA depleted RNA with random hexamers tagged with a T7 promoter sequence (T7-(N)6 primer). The double-stranded cDNA was then used as a template for T7 RNA polymerase producing cRNA. A second cycle of cDNA synthesis was performed using random hexamers to reverse transcribe the cRNA from the first cycle to produce single-stranded DNA (using dATP, dTTP, dGTP, and dUTP) in the sense orientation. cDNA was fragmented using DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE1). The fragmented DNA was then labeled with terminal deoxynucleotidyl transferase that conjugates biotinylated nucleotides. 5.5 µg of this biotin–labeled DNA was hybridized overnight with Affymetrix Human Exon1.0 ST microarrays and washed and scanned on Affymetrix GeneChip® Scanner 3000 7G scanner with an autoloader, according to the instructions from Affymetrix GeneChip Whole-TranscriptSense Target–Labeling Assay manual. Each scanned CEL image of the array was checked for any significant artifacts.

Data Processing

RMA was applied in combination with affymetrix.aroma to all CEL files. This generated gene centric expression values, using a CDF file based on remapping of probes to the human genome 36.1 resulting in expression values for 18,632 genes.

B. Agilent 244K Whole Genome Expression Array

mRNA labeling

One to 2 ug of total RNA sample and Stratagene Universal Human Reference were amplified and labeled using Agilent's Low RNA Input Linear Amplification Kit. The total yield of amplified RNA (aRNA) and Cy dye incorporation was measured by NanoDrop.

Array Hybridization and Imaging

Sample and reference (7-10 ug of each) were co-hybridized to a Custom Agilent 244K Gene Expression Microarray. Arrays were scanned on an Agilent Scanner and probe information was obtained with Agilent's Feature Extraction Software. Each scanned image is viewed for visible artifacts, and if multiple artifacts are present, the array is rejected. Agilent Feature Extraction software creates a QC report for each array that includes: (1) Net Signal Statistics: Signal range distributions for the red and green channels are presented and compared. Samples with large differences between the red and green channel for net signal are flagged as samples/arrays to be watched. (2) Distribution of Outliers: Samples with the % feature non-uniformity >1% are flagged as samples/arrays to be watched. (3) MA plots: Log of the Processed Signal is plotted versus Log of the Ratio (R/G) for each gene to help identify biases in intensity or dye. (4) Reproducibility of SpikeIns (an internal hybridization control): reproducibility of Agilent SpikeIns are measured by % coefficient of variation (<15) and SpikeIn linearity with R2 values close to 1. If any array fails three of the QC criteria it is rejected.

Data Processing

Data was lowess normalized and the ratio of the Cy5 channel (sample) and Cy3 channel (reference) was log2 transformed to create gene expression values for 18,624 genes.

C. Affymetrix HT-HG-U133A

Sample Labeling

One µg of total RNA was converted to complementary RNA (cRNA) target using the Genechip® HT One-Cycle cDNA synthesis Kit (Affymetrix 900687) and the GeneChip® HT IVT Labeling Kit (Affymetrix 900688). Total RNA was first reverse transcribed using a T7-Oligo(dT) Promoter primer in the first strand cDNA synthesis reaction. Following RNAse H-mediated second strand cDNA synthesis, the double stranded cDNA was purified and served as a template for an *in vitro* transcription (IVT) reaction. The IVT reaction was carried out in the presence of T7 RNA polymerase and a biotinylated nucleotide analog / ribonucleotide mix for cRNA amplification and biotin labeling. The biotinlyated cRNA targets were then cleaned up and fragmented.

Array Hybridization

Samples were analyzed using Affymetrix HT-HG-U133A peg arrays (Affymetrix 900751). The hybridization and subsequent washing and staining were performed on the Affymetrix GeneChip® Array Station (GCAS) automation platform.

Data Processing

All samples included in the current study met broadly accepted quality control standards, including percentage present, GAPDH 3'/5' ratio and NUSE IQR. RMA[1] was applied in combination with affymetrix.aroma (http://www.aroma-project.org/). In order to generate gene centric expression values, using a CDF file based on remapping of probes to the human genome 36.1, similarly as described before[2]. This resulted in expression values for 12,042 genes.

D. Creation of a unified Expression Data set

Data from each platform were normalized and summarized separately, as described above, resulting in gene expression estimates for each sample and gene on each platform. Relative gene expression values were calculated per platform by subtracting the mean expression value across patients from the gene estimate and dividing by the standard deviation across patients. Factor analysis was applied to integrate three expression values (one for each platform), on genes present at all three platforms (n = 11,864). The factor analysis provided estimates of relative gene expression scaled to have the same underlying variation among patients for all genes. We rescaled the unified gene expression of each gene by estimates of the standard deviation across patients. To obtain a single estimate of standard deviation per gene, we took the median absolute deviation (MAD) for each platform and then averaged these estimates, restricting to those platforms with high correlation to the unified gene estimates. This gave a single estimate of variation per gene that was then used to rescale the unified gene estimates.

E. Subtype discovery and validation

Gene filtering

Two filters were applied to eliminate unreliably measured genes and to limit the clustering to relevant genes, similar to the filters used in the TCGA GBM data set[2]. The first filter removed genes that had poor unified gene measurements by keeping only genes in which at least two of the three platforms' original measurements had correlation with the unified gene estimate of at least 0.7, resulting in 8,596 genes. The second filter selected 1,500 genes with the highest variability across patients, using the MAD.

Identification of expression subclasses using Non-negative Matrix Factorization clustering

Subclasses of a data set consisting of unified expression data of 489 samples and 1,500 genes were computed by reducing the dimensionality of the expression data from thousands genes to a few metagenes by applying a consensus non-negative matrix factorization (NMF) clustering method[3]. This method computes multiple k-factor factorization decompositions of the expression matrix and evaluates the stability of the solutions using a cophenetic coefficient. Consensus matrices and sample correlation matrices are shown for k=2 to k=6 (**Figure S6.1**). The final subclasses were defined based on the most stable k-factor decomposition and visual inspection of sample by sample correlation matrices, in both TCGA and Tothill data set (see below). Clustering with k=4 gave the most consistent result in both sets. The silhouette width was computed to filter out expression profiles that were included in a subclass, but that were not a robust representative of the subclass. This resulted in the removal of 51 of 135 samples of the Differentiated subclass; 12 of 107 samples of the Immunoreactive subclass; 0 of 109 samples of the Mesenchymal subclass; and 13 of 138 samples of the Proliferative subclass. Differentially expressed marker genes were determined for each subclass by comparing the subclass versus the other three subclasses, using Significance Analysis of Microarrays[4] (SAM). NMF clustering and SAM were both applied as implemented in the R statistical computing environment.

Identification of expression subclasses of an independent public data set

To confirm the presence of four expression subtypes in ovarian serous carcinoma, publicly available expression profiles were preprocessed and clustered in an identical fashion to the

TCGA data. Affymetrix HG-U133 plus 2 CEL-files of 245 stage II-IV ovarian serous cancer patient samples were downloaded from the Gene Expression Omnibus (accession GSE9899[5]). Gene expression values were calculated for 17,256 genes using a gene centric CDF[6], RMA and quantile normalization. Mean row subtraction of log transformed data was applied, and the 1,500 most variable genes were selecting using a MAD filter. NMF clustering was applied on a data set of 245 samples and 1,500 genes as described above. Consensus matrices and sample correlation matrices are shown for k=2 to k=6 (**Figure S6.2**). Differentially expressed marker genes were determined for each subclass by comparing the subclass versus the other three subclasses, using SAM.

Comparison of expression subclasses in the TCGA and Tothill data sets

SAM calculated an F-score for each gene, resulting in a vector of 11,864 F-scores for each subclass. A four by four correlation matrix was computed, with the input vectors being the F-score vectors from the four TCGA expression subtypes, and the F-score vectors from the four Tothill expression subtypes (**Figure S6.3**).

Correlation with Copy Number and Gene Mutations

Segmented copy number profiles were available for 489 ovarian carcinoma samples and matched normal controls (TCGA Research Network, 2011). Chromosomal regions were categorized using sample-specific thresholds into one of the three following levels: 1) homozygous deletions, 2) neutral copy number or 3) focal gains. The frequencies of the three levels of copy number for each of the 113 regions reported to be significantly altered in this manuscript were determined per subtype. Only tumor samples for which the expression profile exhibited a positive silhouette width to the centroid of their expression subtype were included in the analysis (*n*=413).

13,707 somatic mutations present in 271 core samples were tested for statistically higher than random significance using MutSig (**Table S2.3a**, Lawrence et al., manuscript in preparation). Frequencies of mutations for ten genes identified as significantly mutated by MutSig were assessed per subtype.

Association of copy number alterations and gene mutations with subtype was determined by comparing each subtype versus the remaining three subtypes using a chi square test. The Family-wise Error rate of p-values between 113 copy number alterations and subtypes was controlled by using the Hochberg method implemented in p.adjust (R Development Core Team, 2008). The Family-wise Error rate of p-values between 10 gene mutations and subtypes was controlled by using the Hochberg method implemented in p.adjust (R Development Core Team, 2008). Results are shown in **Figure S6.4**.

F. Identification of a signature predictive of survival in ovarian cancer

Gene expression values within each dataset used were normalized to standard deviations from the median, and overall survival was capped at 60 months. Using a training dataset of gene expression profiles from 215 stage II-IV ovarian tumors from the TCGA (batches 9, 11-15), a prognostic gene signature for overall survival was defined (genes with univariate Cox $P < 0.01$), comprised of 108 genes correlated with poor (worse) prognosis and 85 genes correlated with good (better) prognosis. Cox analysis to define the signature was carried out in September of 2009, using the training dataset and all patient information available at the time. The signature was tested in a validation set consisting of 255 samples available through TCGA (batches 17-19, 21, 22 and 24) in June of 2010, using the most up-to-date patient information. Four tumors from the Bonome *et al.*[7] data set were excluded when validating the TCGA signature since they had also been included in the TCGA cohort. The prognostic t-score was defined as the two-sided t-statistic comparing, within each tumor profile, the average of the poor prognosis genes with the average of the good prognosis genes (*e.g.* the t-score for a given tumor being high when both the "poor prognosis" genes in the signature were high and the "good prognosis" genes were low).

G. miRNA profiling and subtype identification

miRNA Labeling

100-400ng of total RNA was labeled by ligation to cyanine 3-pCp molecules using the Agilent miRNA Micorarray labeling protocol (Agilent Technologies, Santa Clara, CA) using T4 ligase (NEB, Ipswich, MA).

Array Hybridization

Labeled miRNAs were hybridized to Agilent 8 x 15K Human miRNA-specific microarrays overnight. Arrays were scanned on an Agilent Technologies Scanner and probe information was obtained with Agilent's Feature Extraction Software. Each scanned image is viewed for visible artifacts. Agilent's Feature Extraction output reports four main microRNA-specific quality check criteria, including: (1) Additive Error Estimate, measure of the background. Samples with additive error between 5-12 counts/pixel are flagged as watched, samples with additive error greater than 12 counts/pixel are flagged as failed. (2) Percentage of Feature Population Outliers: samples with populations outlier between 7-10% are flagged as watched; samples with population outlier greater than 10% are flagged as failed. (3) Median Percent Coefficient of Variation (%CV) for replicate probes: measure of reproducibility. Samples with %CV between 8-15% are flagged as watched; samples with %CV greater than 15% are flagged as failed. (4) 75th Percentile of Total Gene Signal. Samples with 75th percentile total gene signal less than 35 are flagged as watched. Any sample that failed any of the first three criteria is repeated. All samples used in this study passed quality control.

Data Processing

Data was quantile normalized on the probe level. Signals from probes measuring the same microRNA are summed up to generate gene-centric total gene signal, followed by log2 transformation. Distance Weighted Discrimination (DWD) method is applied to data for batch-correction.

Summary of miRNA consensus clustering

We used the Firehose pipeline at the Broad Institute to calculate miRNA clusters based on a consensus non-negative matrix factorization (NMF) clustering method[8,9]. Using the mean row subtraction of expression data, we filtered the data to 150 most variable mirs. Consensus NMF clustering[8,9] of 487 samples and 150 mirs identified 3 subtypes (**Figure S6.5**), with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width[10,11] calculation for selecting the robust clusters(**Figure S6.6)**. Samples with asigned clusters are available at List of samples with 3 subtypes and silhouette width and List of samples belonging to each cluster in different k clusters. Samples most representative of the clusters, hereby called "core samples" were identified based on positive silhouette width[10], indicating higher similarity to their own class than to any other class member (**Figure S6.6**).

Supplementary Figures

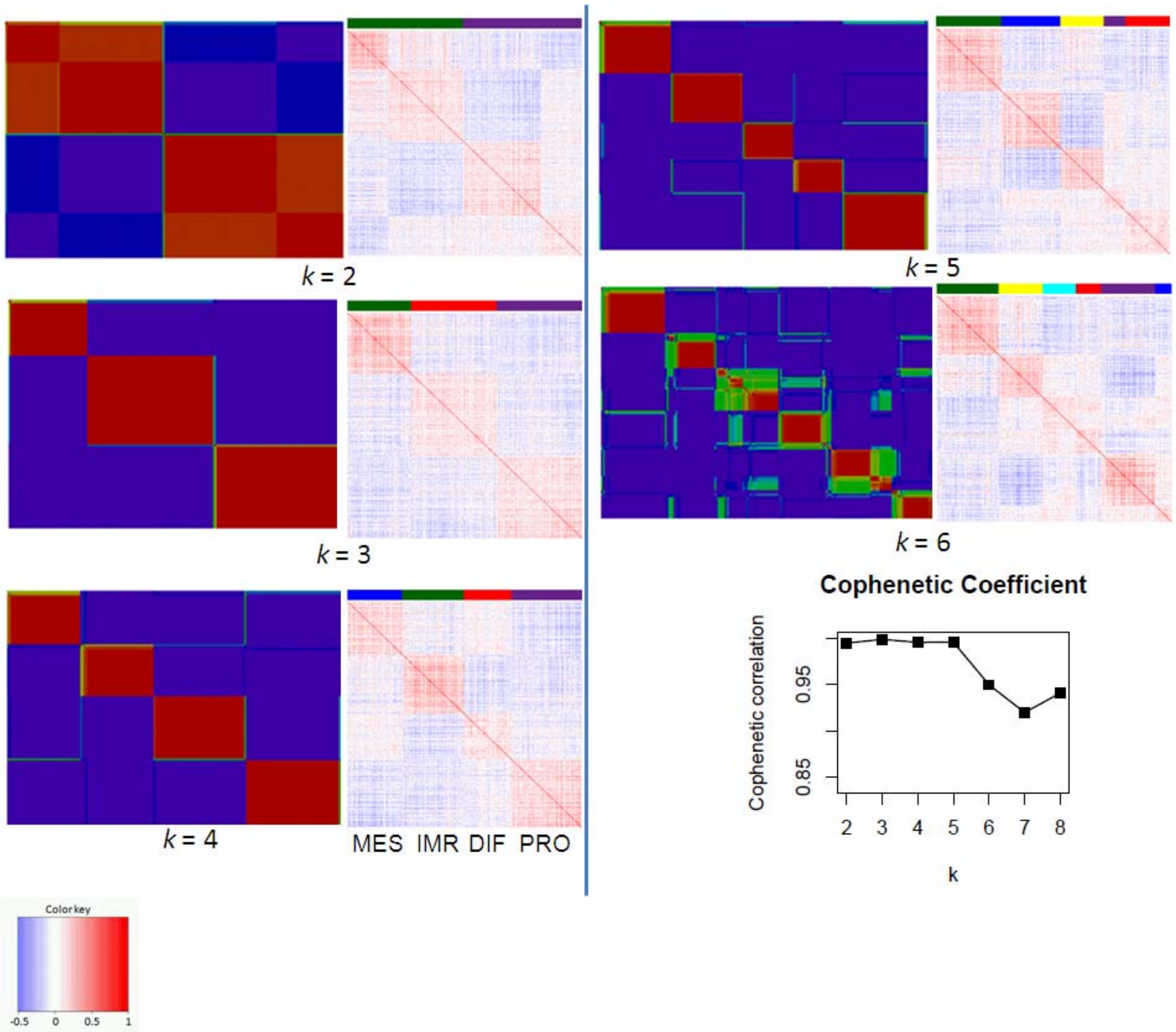


Figure S6.1. CNMF clustering of 1,500 variably expressed genes and 489 TCGA Ovarian serous carcinoma samples of stage II-III-IV. Consensus matrices (left panel) and correlation matrices (right panel) are shown for clustering with *k*=2 to *k*=6. The cophenetic coefficient shows a consistently high value between *k*=2 and *k*=5. Clustering with *k*=4 shows four robust clusters with limited overlap between clusters.

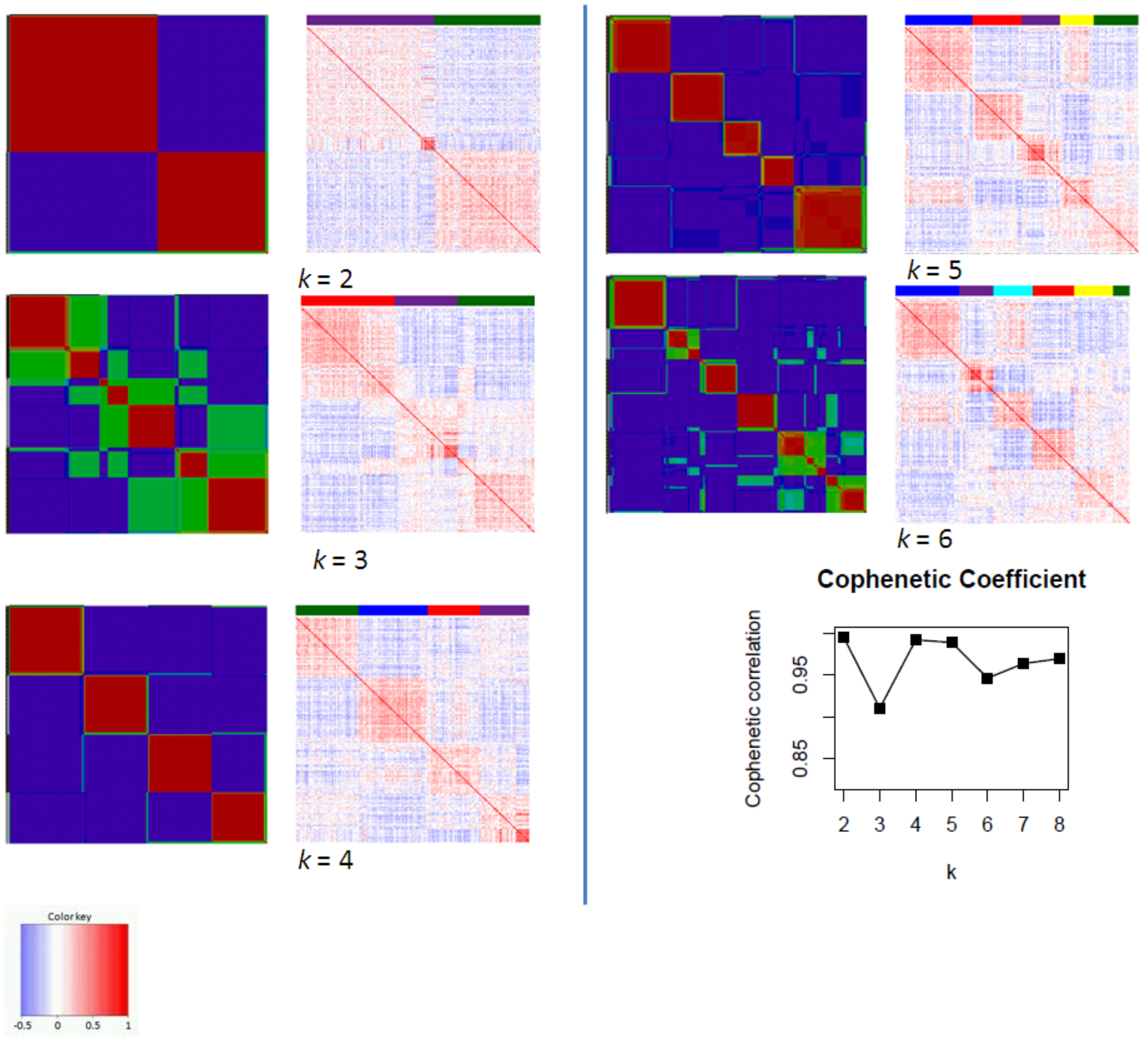


Figure S6.2. CNMF clustering of 1,500 variably expressed genes and 245 Ovarian stage II-III-IV serous carcinoma samples (Tothill et al, PMID 18698038). Endometroid and lower grade samples present in the original study were excluded. Consensus matrices (left panel) and correlation matrices (right panel) are shown for clustering with *k*=2 to *k*=6. The cophenetic coefficient suggests an optimal result for *k*=2, *k*=4, *k*=5. Clustering with *k*=4 shows four robust clusters with limited overlap between clusters. Low malignancy samples display particularly strong correlation as indicated by the red block in the right lower corner.

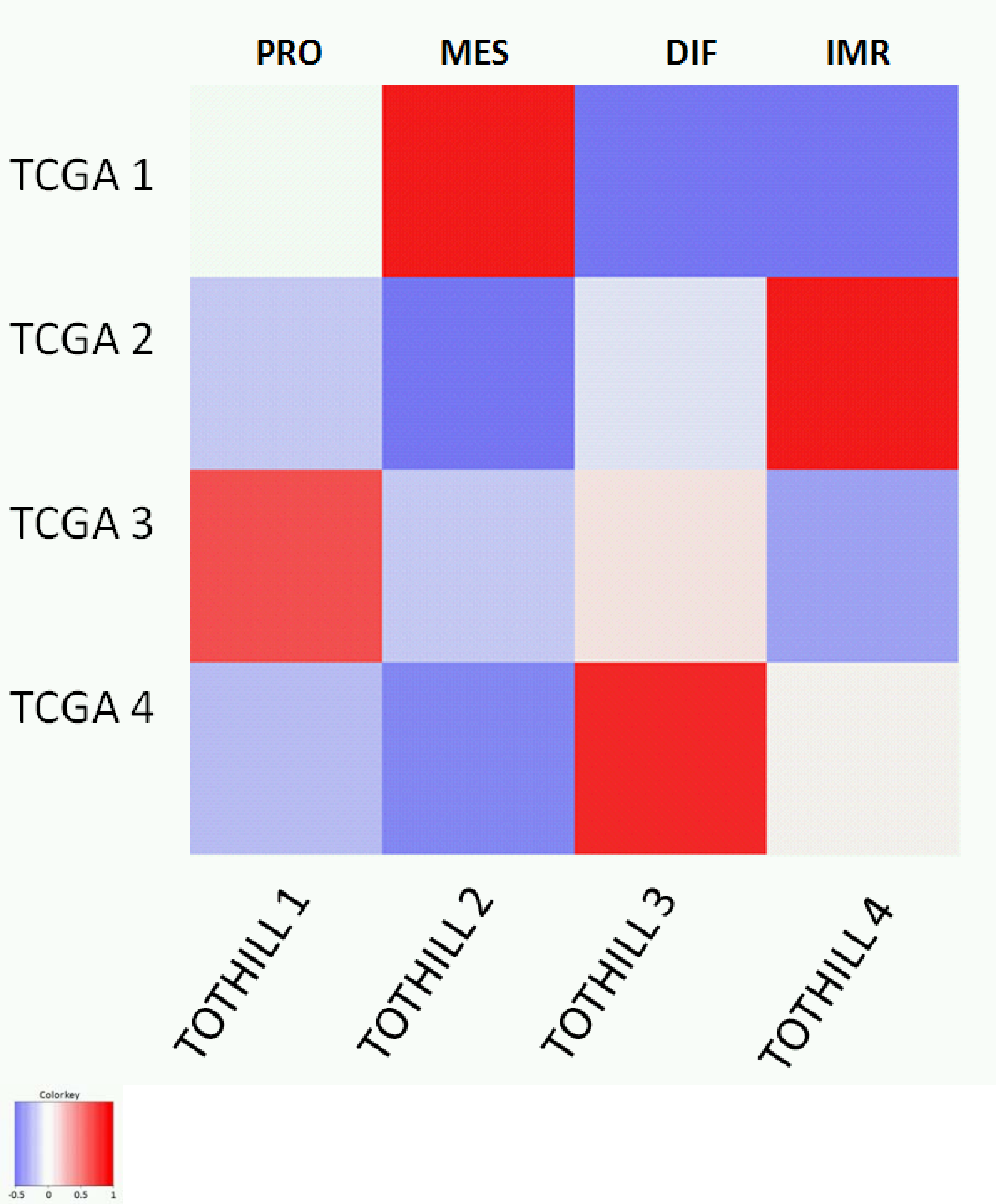


Figure S6.3. Correlation of F-test score vectors between TCGA and Tothill clusters. F-test scores were determined by comparing samples of cluster *K* to the remaining clusters within the data set.

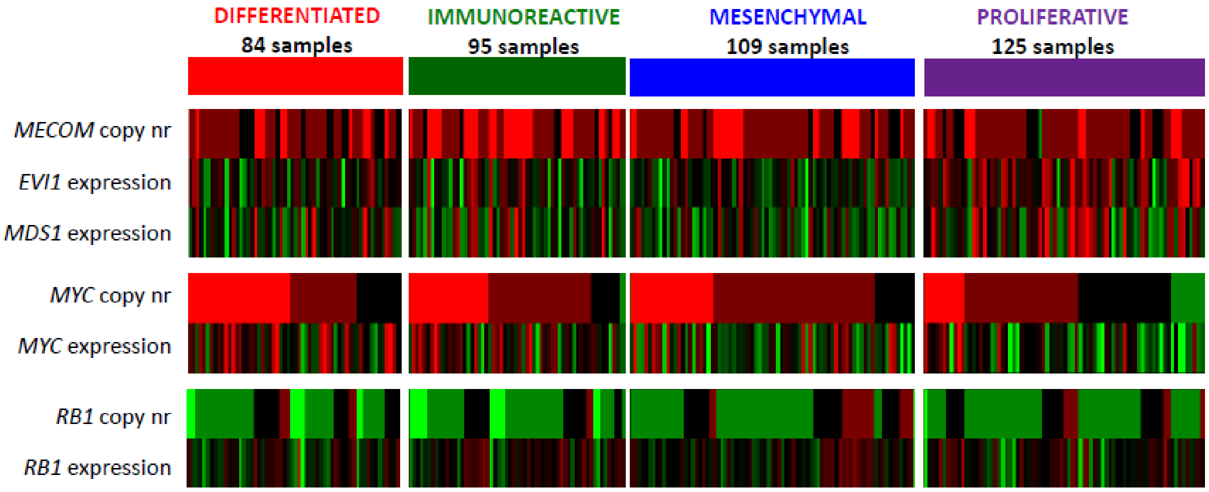


Figure S6.4. Genomic abnormalities significantly associated with subtype. Copy number associations were calculated using copy number profiles from 413 tumors (Supplemental Table S6.2a). Mutation data was included from 271 tumors (Supplemental Table S6.2b). Data shown is from 271 tumors for which copy number, mutation and expression data was available. Significantly subtype associated abnormalities present in at least 5% of tumor samples are shown. A complete list of abnormalities and their association to subtype is shown Supplemental Table S6.2.

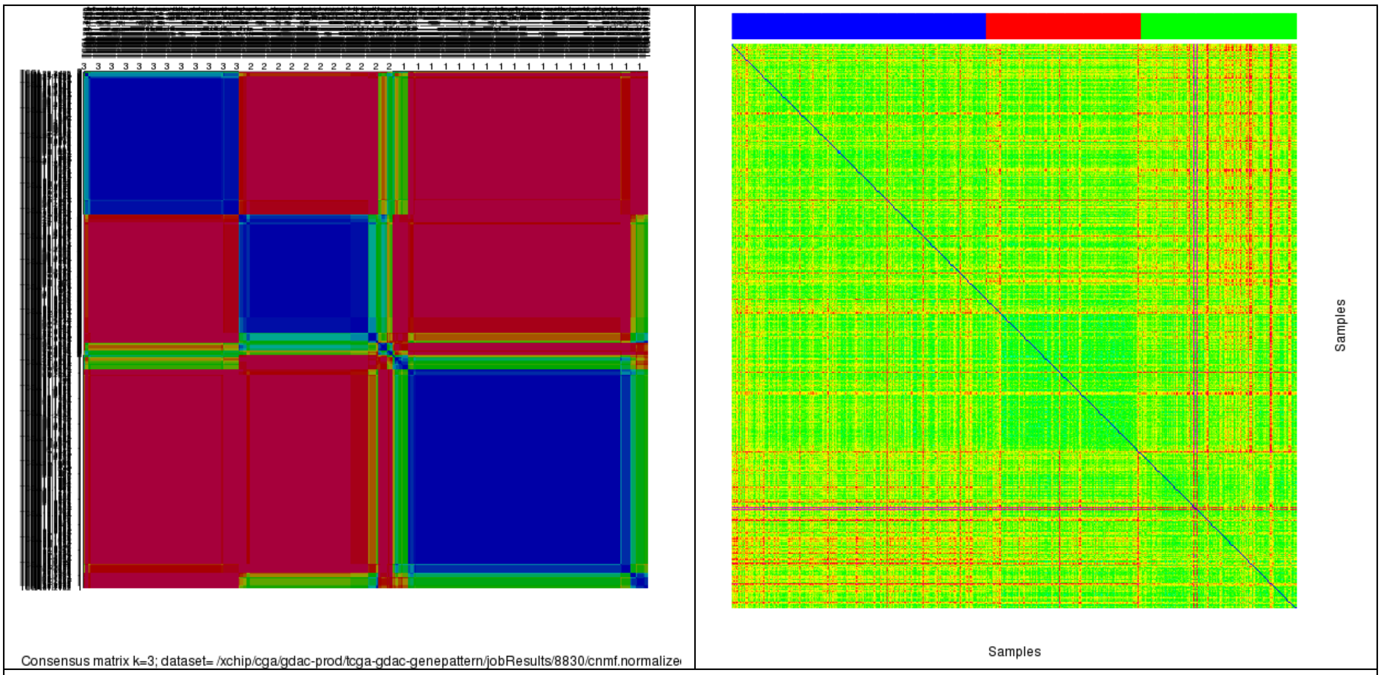


Figure S6.5 Consensus NMF clustering of 150 variably expressed genes and 487 samples. The consensus matrix (left panel) and correlation matrix (right panel) after clustering show 3 robust clusters with limited overlap between clusters.

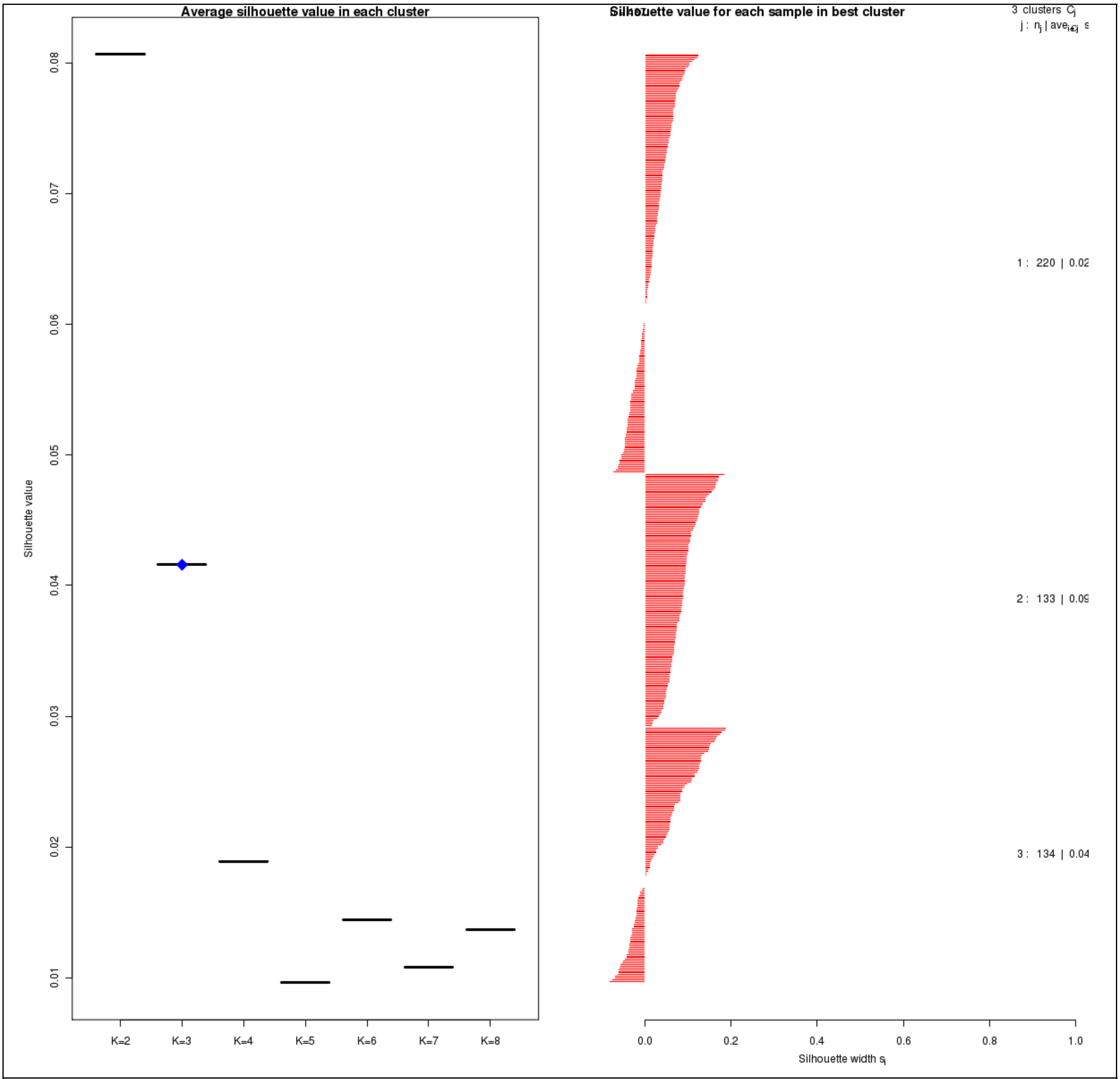
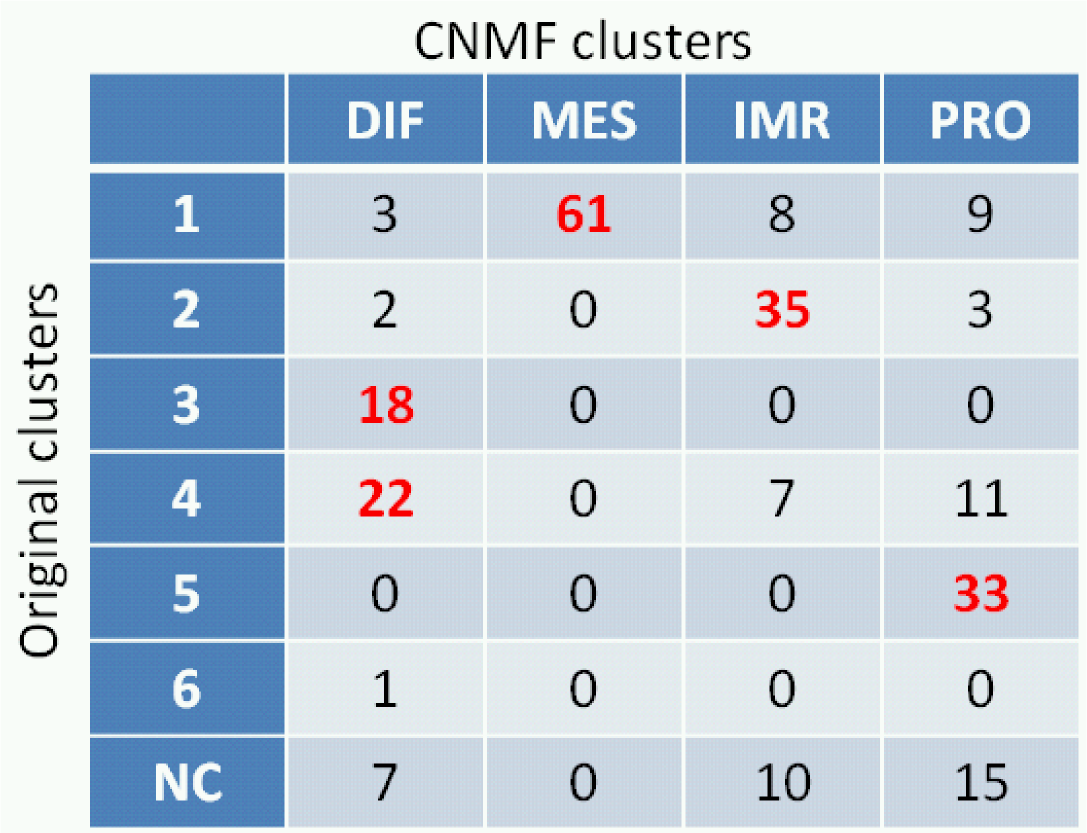


Figure S6.6 The robust cluster was pointed out by blue symbol(left panel) and the silhouette width of each sample in robust cluster was shown on right panel. Silhouette width is defined as the ratio of average distance of each sample to samples in the same cluster to the smallest distance to samples not in the same cluster. It was calculated and the average silhouette width for all samples within one cluster was shown above according to different clusters(left panel). If silhouette width is close to 1, it means that sample is well clustered. If silhouette width is close to -1, it means that sample is misclassified.

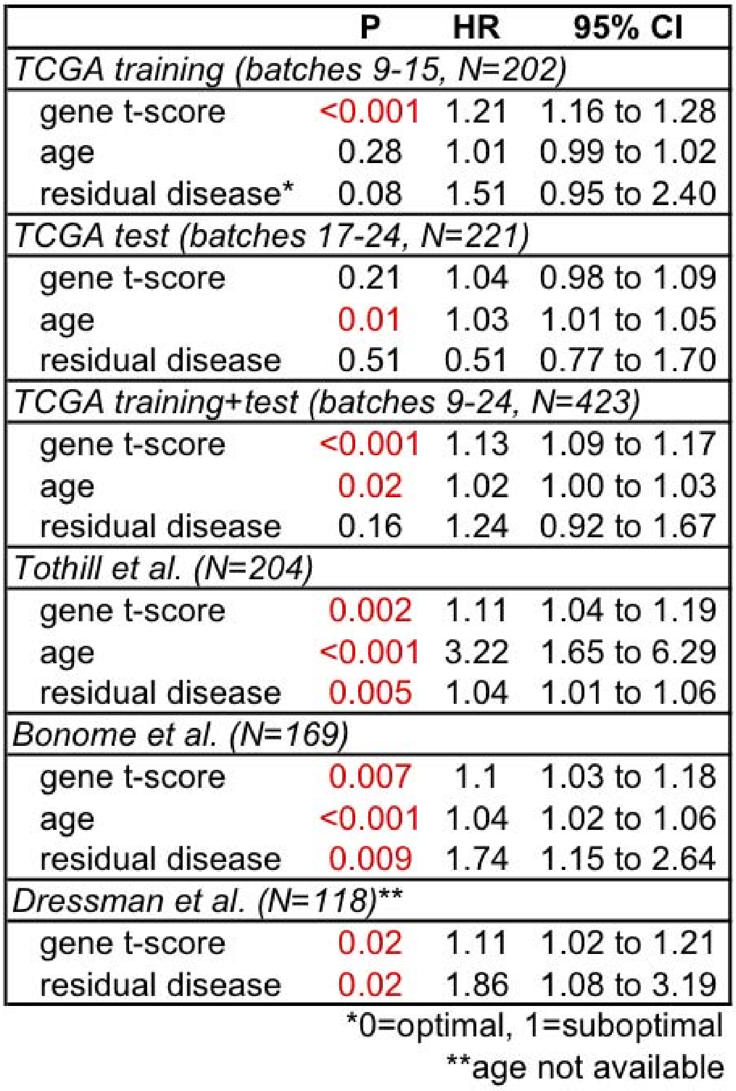


CNMF clusters

Original clusters

	DIF	MES	IMR	PRO
1	3	61	8	9
2	2	0	35	3
3	18	0	0	0
4	22	0	7	11
5	0	0	0	33
6	1	0	0	0
NC	7	0	10	15

Table S6.3. Overlap in cluster membership of 245 Ovarian stage II-III-IV serous carcinoma samples between CNMF clustering and hierarchical clustering as described (Tothill et al, PMID 18698038). Note the overlap between the high grade clusters 1-2-4-5 and the CNMF clusters.



	P	HR	95% CI
TCGA training (batches 9-15, N=202)			
gene t-score	<0.001	1.21	1.16 to 1.28
age	0.28	1.01	0.99 to 1.02
residual disease*	0.08	1.51	0.95 to 2.40
TCGA test (batches 17-24, N=221)			
gene t-score	0.21	1.04	0.98 to 1.09
age	0.01	1.03	1.01 to 1.05
residual disease	0.51	0.51	0.77 to 1.70
TCGA training+test (batches 9-24, N=423)			
gene t-score	<0.001	1.13	1.09 to 1.17
age	0.02	1.02	1.00 to 1.03
residual disease	0.16	1.24	0.92 to 1.67
Tothill et al. (N=204)			
gene t-score	0.002	1.11	1.04 to 1.19
age	<0.001	3.22	1.65 to 6.29
residual disease	0.005	1.04	1.01 to 1.06
Bonome et al. (N=169)			
gene t-score	0.007	1.1	1.03 to 1.18
age	<0.001	1.04	1.02 to 1.06
residual disease	0.009	1.74	1.15 to 2.64
*Dressman et al. (N=118)***			
gene t-score	0.02	1.11	1.02 to 1.21
residual disease	0.02	1.86	1.08 to 3.19

*0=optimal, 1=suboptimal

**age not available

Table S6.4. Multivariate Cox model, in which worse patient outcome was evaluated among tumors in both TCGA and Tothill *et al*. datasets, in relation to the 192-gene t-score (from **Figure 2**) as well as to clinical variables age and surgical outcome (i.e. presence of residual disease).

References

1. Irizarry, R.A. et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).
2. Verhaak, R.G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110.
3. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-9 (2004).
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**, 5116-21 (2001).
5. Tothill, R.W. et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* **14**, 5198-208 (2008).
6. Liu, H. et al. AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. *Bioinformatics* **23**, 2385-90 (2007).
7. Bonome, T. et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res* **68**, 5478-86 (2008).

8. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101, 4164-9 (2004).
9. Broad Genepattern: NMFConsensus http://genepattern.broadinstitute.org/gp/pages/index.jsf .
10. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53-65 (1987).
11. R silhouette package: http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/silhouette.html .

M7. DNA Methylation Supplemental Text

Cancer-associated epigenetic silencing of genes

Epigenetic silencing is increasingly recognized as an important alternative to deletion or mutation in the inactivation of gene function in cancer (PMID: 9988266, 12042769). The identification of genes that are epigenetically silenced in a cancer-specific manner adds to our understanding of the full complement of molecular alterations that contribute to oncogenesis, and might shed light on the early detection, prevention, treatment and prognosis of the disease.

DNA methylation, repressive histone modifications, and other marks can work in concert to achieve an aberrant epigenetically silenced state at susceptible gene promoters. For the purposes of this study, the assessment of epigenetic marks is confined to the quantitative measurement of DNA methylation levels at a limited number of CpG dinucleotides (an average of two per promoter) at 14,475 genes, using the Illumina Infinium HumanMethylation27 BeadChip assay (see Supplemental Methods).

The goal of this section is to identify genes with evidence for cancer-specific promoter hypermethylation with an associated decrease in gene expression. The general properties of such genes are: low levels of promoter methylation in control tissues thought to represent potential cells of origin, high levels of promoter methylation and an associated lower expression in at least some of the tumors, and a good inverse correlation between promoter methylation levels and gene expression. The relationship between DNA methylation and gene expression is complex and highly variable among the 12,233 genes in our data set for which we have both DNA methylation and gene expression measurements. Therefore, the appropriate selection criteria may vary depending on the nature of this relationship. In designing a strategy to identify epigenetically silenced genes, we considered the following three issues.

First, the impact of CpG methylation on transcriptional potential depends on the density of the methylated CpGs, and their location relative to the transcription start site and functional promoter elements. The constraints of the HumanMethylation27 BeadChip design, with an average of two CpGs per promoter, do not allow for a comprehensive assessment of each gene, and the locations of the measured CpGs may be uninformative for some genes. Therefore, we anticipate that we will be unable to identify some silenced genes. Some genes may have alternative promoters for which the methylation status is not assessed. Genes lacking promoter methylation in some tumors or in normal tissues may not be expressed due to a lack of appropriate transcription factors. These factors all contribute to a complex relationship between our DNA methylation measurements and observed gene expression levels. We used a rank-based Spearman correlation to allow for nonlinear relationships between DNA methylation and gene expression.

Second, the identification of cancer-associated DNA methylation alterations requires a comparison of tumor DNA methylation data to a control tissue, ideally representing (or at least enriched for) the cell-of-origin of serous ovarian cancer. The normal ovary surface epithelium and fallopian tube epithelium have both been proposed as originating tissues for ovarian cancer. Our study includes eight full-thickness fallopian tube samples, but no ovarian surface epithelial samples. The presence of other cell types in the full-thickness samples may introduce DNA methylation profiles that differ from the normal epithelial comparator tissue. A technical concern is that all of the fallopian samples were run in a single analysis batch (batch 9), and the clinical parameters of the tumor samples within this batch are not representative of those in the entire study. We focused on identifying genes with large DNA methylation and gene expression differences between fallopian tube and tumors, to address concerns regarding the small sample size of the comparison control group, the confounding of biology and batch-associated measurement biases, which cannot be fully removed using conventional normalization approaches, and variations in stromal contamination among the tumors.

Third, epigenetic silencing of different genes is likely to occur at varying frequencies in the tumor data set. To capture genes with a low frequency of epigenetic silencing, we focused on tumors with high levels of DNA methylation at that locus, by comparing the 90th percentile of tumors to the mean of the fallopian tube samples.

We describe the strategy for identifying epigenetically silenced genes in detail in the methods section. In brief, we apply four separate filtering criteria: 1) low mean DNA methylation in fallopian tube samples, 2) large difference in DNA methylation between the 90th percentile tumor and mean fallopian tube methylation, 3) large difference in mean gene expression between the fallopian tubes, and the 10% of tumor samples with the highest DNA methylation for that gene, and 4) strong inverse correlation between DNA methylation and gene expression. For each filter, we established a relaxed threshold and a stringent threshold. To set minimal criteria for each filter, while capturing genes with different silencing patterns and frequencies, we required candidate epigenetically silenced genes to pass all four relaxed thresholds, and at least three out of four more stringent thresholds.

The list of genes resulting from our analysis should be considered a preliminary list of epigenetic silencing candidates. This list will have likely missed some important, functionally relevant silenced genes, while including others inappropriately. These candidate genes for which we have observed correlative evidence for epigenetic silencing, will require experimental validation by promoter methylation cassette analysis or DNA methyltransferase inhibitor treatment of cell lines. A complete list of the 168 genes with evidence for epigenetic silencing is provided in **Table S7.1**.

BRCA1 is one of the 168 genes identified with this method. A scatterplot of DNA methylation versus gene expression is shown in **Figure S7.1A**. *BRCA1* silencing via promoter hypermethylation has been reported previously in breast and ovarian cancer (PMID: 10749912), and recent studies have reported *BRCA1* hypermethylation in varying percentages of ovarian cancer patients, but mostly within 10-20% (PMID: 11034065; PMID: 10749912; PMID: 18208621). With the procedure described in the supplemental methods below, we identified 56 out of 489 samples (11.5%) with *BRCA1* inactivation via promoter hypermethylation in the current high-grade, high stage serous ovarian cancer cohort (indicated by blue dots in **Figure S7.1A**). We validated the *BRCA1* promoter hypermethylation with the MethyLight technology (PMID: 10344733; 10734209). MethyLight is a real-time PCR based method for DNA methylation quantification. The MethyLight PMR value (See Methods) for *BRCA1* showed strong correlation (Pearson Correlation Coeffient = 0.78-0.90) with the beta values measured by the four *BRCA1* probes (**Figure S7.2**). A receiver operating characteristic (ROC) curve (**Figure S7.3**) showed that the PMR values correlate very well with the *BRCA1* epigenetic silencing calls, with an AUC of 0.99. Using PMR>10 as the cutoff for promoter hypermethylation, as described by Weisenberger et al (PMID: 16804544), MethyLight confirms the promoter hypermethylation in 55 of the 56 samples (98.2%) previously identified on the Infinium platform, and the absence of such methylation in 436 of the 441 samples (98.9%), including 433 tumors and 8 normal fallopian tube samples previously identified to be negative of *BRCA1* epigenetic silencing. Overall, the two methods showed concordance on 491 of the 497 samples (98.8%), and confirmed the observed *BRCA1* epigenetic silencing.

Notably, *BRCA1* epigenetic silencing is mutually exclusive with all BRCA1/2 mutations (the pathway section of the main text). A previous population based study showed that *BRCA1* epigenetic silencing was only seen in ovarian cancer patients without a family history associated with a breast/ovarian cancer syndrome (PMID: 11034065), suggesting that *BRCA1* promoter hypermethylation is unlikely to be inherited, but rather an acquired somatic change that leads to *BRCA1* inactivation in sporadic ovarian cancers. *BRCA1* hypermethylated cases are considerably younger and occur more frequently than the *BRCA1* somatic mutation cases. This suggests that epigenetic silencing of *BRCA1* might be a more efficient somatic mechanism of inactivation for this gene than mutation.

RAB25 (**Figure 7.1B**) is ranked highest among the 168 genes, based on the DNA methylation-expression correlation. Previously, *RAB25* was reported to have a >1.3-fold copy number increase in about half of advanced serous epithelial ovarian cancers and marked mRNA up-regulation in most of ovarian cancers, compared to normal ovarian epithelium, and the copy number and expression levels of *RAB25* were associated with disease-free survival or overall survival in ovarian and breast cancers (PMID: 15502842). RNA interference targeting *RAB25* has been showed to slow down cell proliferation and inhibit tumor growth in in vivo and in vitro ovarian cancer models (PMID: 17393986). Other papers also highlight the role of *RAB25* in cancer development. Somewhat contrary to these reports on ovarian and breast cancers, a recent paper (PMID: 20197623) indicated that loss of *RAB25* promotes intestinal neoplasia, and is associated with human colorectal adenocarcinomas. Our study did not observe significant amplification of this region. On the contrary, our results indicate that *RAB25* down-regulation actually occurs in a subset of ovarian tumors (**Figure S7.1B**). This result, in line with the reported *RAB25* loss in intestinal neoplasia, suggests that loss of *RAB25* might play a role in ovarian tumorigenesis.

Among the 168 genes, *AMT* (**Figure S7.1C**), *SPARCL1* (**Figure S7.1D**) and *CCL21* (**Figure S7.1E**), are also noteworthy because they show promoter hypermethylation in the vast majority of tumors. *SPARCL1*, a member of the *SPARC* family and anti-adhesive extracellular matrix protein, was originally shown to be down-regulated in many epithelium-derived cancers (PMID: 9485012; PMID: 9443398). A gene closely related to *SPARCL1*, *SPARC*, has been shown to have tumor-suppressor activity in human ovarian epithelial cells (PMID: 8649850), and one driver mutation of *SPARC* was observed in our study. Loss of *SPARCL1* expression has been shown to be associated with increased proliferation and cell cycle progression (PMID: 10735494), highlighting its role in tumorigenesis. *CCL21* has been shown to be a chemoattractant for T cells and dendritic cells (PMID: 9927506). Anti-tumor properties of this gene have been attributed to its role as a chemo-attractant (PMID: 12740040) and as an angiostatic modulator (PMID: 10925282).

Discovery of DNA methylation subgroups

Using the resampling-based consensus clustering method as previously described (S. Monti, et al. Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, Machine Learning Journal, 52(1-2):91-118, 2003.), we identified four DNA methylation clusters (**Figure S7.3**). However, there is no clear evidence for the existence of a well-defined CpG Island Methylator Phenotype (CIMP), as has been identified for colorectal carcinoma (PMID: 10411935) and glioblastoma (PMID: 20399149), characterized by concerted hypermethylation at CpG islands. There is a moderate, but statistically significant overlap between DNA methylation clusters and gene expression subtypes ($p<2.2*10^{-16}$, χ^2 test. **Table S7.3**), Adjusted Rand Index = 0.07).

Patients belonging to the four clusters differ significantly in age at diagnosis (One-way ANOVA, p=$5*10^{-7}$) (**Figure S7.4B**). The mean ages of the patients in the four clusters are 59.1, 65.8, 57.1, and 62.1 for MC1, MC2, MC3, and MC4, respectively. Tukey HSD test revealed that the real differences lie between clusters MC1 and MC2 (adjusted *p*=0.0005), between MC2 and MC3 (adjusted *p*=0.000002), and between MC3 and MC4 (adjusted *p*=0.0009).

Patients in the four DNA methylation clusters also differ significantly in overall survival with data censored at five years (Median survival time: Cluster MC1 – 48.9 months, Cluster MC2 – 35.8 months; Cluster MC3 – 40.9 months; Cluster MC4 – 43.6 months; Logrank test, *p*=0.04.) (**Figure S7.4A**). After adjusting for age using the Cox regression model, Cluster MC1 has the best survival and Cluster MC3 has a significantly worse survival compared to Cluster MC1 (Hazard Ratio = 1.43, *p*=0.04). Cluster MC2 has marginally significantly worse age-adjusted survival (Hazard Ratio = 1.42, *p*=0.09) than MC1.

The four DNA methylation clusters differ significantly in their frequencies of *BRCA* inactivation events (p=$4.9*10^{-6,}$ Fisher's exact test), which include *BRCA1/2* mutation and *BRCA1* epigenetic silencing. Altogether, Cluster MC1 and MC3 have the highest frequencies of *BRCA* inactivation (46.6% and 44.5%, respectively), while Cluster MC2 has the lowest such frequency (13.2%, **Table S7.2**). This trend holds true when we stratify the *BRCA* inactivation events to epigenetic silencing (p=$1.0*10^{-5}$, Fisher's exact test) and *BRCA1/2* mutation events (*p*=0.04, Fisher's exact test).

Although the four DNA methylation clusters differ in their DNA methylation profiles and their biology, the overall average silhouette width (P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. 1987. Journal of Computational and Applied Mathematics. 20. 53-65.) is poor (0.02), indicating weakly defined clusters with substantial within-group heterogeneity. Indeed, other clustering methods yield varying subgroupings of the tumors based on their DNA methylation profiles. Therefore, the DNA methylation cluster memberships reported here should be considered preliminary, with alternative groupings possible. Nevertheless, the significant overlap with expression clusters, and other biological differences between the DNA methylation clusters suggest some validity to these subdivisions.

Batch Effect In TCGA DNA Methylation Data

Prevalent batch effects were observed in the DNA methylation dataset (affecting 78.8% of the features, F test), consisting of 489 tumor samples, as also discussed by Leek et al (PMID: 20838408). Principal component analysis (PCA) shows that the second PC is significantly associated with batch (adjusted R-square = 0.59).

A natural solution would be to employ normalization tools to remove the technical variation associated with batch. However, DNA methylation data expressed as beta values do not fit the assumptions for existing normalization methods developed for gene expression, such as quantile normalization and loess normalization(PMID: 20125086). Most of these methods assume that most genes are not expressed and/or that most genes are not differentially expressed. However, neither assumptions hold for DNA methylation data, where global hypomethylation observed in many cancers may affect a majority of probes. Moreover, samples can differ substantially in their total methylcytosine content. Quantile normalization would erase such differences. Meanwhile, for beta-distributed data like DNA methylation beta values, the variance is associated with the mean (heteroscedasticity). Therefore, we cannot apply linear model-based methods without transforming the data properly (logit or probit). This heteroscedasticity also hinders us from a simple application of empirical Bayes normalization methods like ComBat (PMID: 16632515) since the variances of different probes cannot be modeled as the same. In other words, there is a lack of a reasonable prior.

What is more, there are substantial biological differences across batches. For example, five-year survival rate (Logrank test, p=0.04, for differences across batches) differ from 0% (Batch 9) to 58% (Batch 11). The known or unknown confounding biological differences between batches may result in removing biology in the normalization process. While putting in a design matrix to adjust for the known biological factors is feasible, the unknown confounding biological difference is a more serious problem. While well-established methods to remove unknown factors such as surrogate variable analysis (PMID: 17907809) are outstanding methods in dealing with unknown factors, will actually be undesirably harsh for ***unsupervised*** analyses aimed at finding molecular subtypes. We, however, advise the use of this method in ***supervised*** analyses with TCGA data. This confounding biology reemphasizes the importance of randomized experimental design in high-throughput studies, as has been emphasized by Verdugo et al and Leek et al (PMID: 19617374, 17907809)

Although most of the probes are susceptible to the batch effect, the size of the batch variation is small. The mean absolute beta values difference across pairwise comparison of different batches is 0.030 (median difference 0.026) for the probes that varied significantly across batches. Given the complications discussed above, we chose not to normalize the data in a way that might remove biological differences or introduce artifacts, but rather focus our analyses on probes with large biological variation and limited technological variation. For the clustering analysis, we removed probes with relatively large batch variation using technical replicates (weighted average of deviation from equality of >0.05, as described in the method section). By combining this filtering approach with a selection of the most variant probes, we were able to obtain a dataset in which the large biological variation predominates over the weak technical variation. None of the top ten PCs were associated with batch in the reduced data set used for clustering. Cohen's f^2 (PMID: 19565683) for the 858 probes range from 0.004 to 0.183, with a mean of 0.029 (1st quartile: 0.018, 3rd quartile: 0.035). 857 probes out of 858 have an effect size of smaller than medium (rule of thumb 0.15; equivalent to a multiple regression R of 0.13). On average, only about 2.8% of the variance for each probe is explained by batch in the dataset used for clustering. We can also see from the bottom side bar in **Figure S7.4** that analytical batch is not driving the clustering. For the epigenetically silenced genes, we used large effect size as the threshold, rather than statistical significance. Here, we also observed that batch effects do not have a deterministic role in the epigenetically silenced genes identified. We would advise those who use the TCGA DNA methylation data to be aware of the technical variation in the data, and to consider developing normalization methods appropriate for beta-distributed DNA methylation data.

Supplemental Methods for DNA Methylation

DNA Methylation Assays

We performed the Illumina Infinium DNA methylation assay on 519 TCGA ovarian samples and eight fallopian tube samples from Batches 9,11-15 and 17-19, 21-22, and batch 24. The Illumina Infinium HumanMethylation27 arrays interrogate 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing in the NCBI Database (Genome Build 36). Bisulfite conversion was performed on 1 µg genomic DNA from each patient using the Zymo EZ96 kit (Zymo Research, Orange, CA) as recommended by the manufacturer. We evaluated DNA quantity and completeness of bisulfite-conversion using MethyLight control quality control (QC) reactions as previously described (PMID: 18987824). All TCGA samples passed these QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified and enzymatically fragmented. The bisulfite-converted, WGA-DNA samples were then purified and hybridized to the BeadChip arrays, in which bisulfite-converted DNA molecules anneal to locus-specific DNA oligomers that are bound to individual bead types. Each CpG locus can hybridize to methylated (CpG) or unmethylated (TpG) oligo bead types. DNA methylation-specific primer annealing is followed by single-base extension using labeled nucleotides. Both unmethylated and methylated bead types for a specific CpG locus incorporate the same labeled nucleotide, as determined by the base immediately preceding the cytosine being interrogated by the assay, and subsequently will be detected in a single channel. Each beadchip, containing 12 subarrays, was then fluorescently stained after extension, scanned, and the intensities of the methylated (M) and unmethylated (U) bead types for each CpG locus across all samples are measured. Mean non-background corrected M and U signal intensities for each locus were extracted from Illumina BeadStudio (or GenomeStudio) software. The beta value DNA methylation scores for each sample and locus were calculated as (M/(M+U)).

Detection p-values were calculated by comparing the set of analytical probe replicates for each locus to the set of 16 negative control probes. The negative controls are modeled by normal distribution. The detection p-value for the probe with intensity I_{probe} , is calculated as: $1-Z((|I_{probe}-\mu_{neg} |)/\sigma_{neg})$. In this formula, μ_{neg} and σ_{neg} are the average and the standard deviation of the signals from the negative controls, and Z is the one-sided tail probability of the standard normal distribution. For each probe, the detection p-values are calculated separately for methylated and unmethylated probe signal intensities, and the smaller detection p-value was taken as the final detection p-value for the probe. Data points with detection p-values > 0.05 were deemed not significantly different from background, and were masked as "NA".

TCGA Data Packages

The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga.cancer.gov/dataportal).

LEVEL 1: Level 1 data contain the non-background corrected signal intensities of the methylated (M) and unmethylated (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 methylated and unmethylated bead types as well as the standard error of methylated and unmethylated signal intensities are also provided for each sample and probe. Similar values are also provided for the negative control probes. It is important to note that the identity of the dye is representative of the nucleotide adjacent to the CpG dinucleotide. The methylation discrimination is derived from separate measurements from the two different types of beads present for each locus. For some loci, both 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 as level 1 data to the TCGA Data Coordination Center (DCC) as “Methylated Signal Intensity” and “Unmethylated Signal Intensity”. The information of which dye is used for each locus is supplied in the manifest deposited with the DCC.

LEVEL 2: Level 2 data files contain the beta value calculations for each probe and sample. Data points with detection p-values > 0.05 were deemed not significantly different from background, and were masked as NA.

□LEVEL 3: Level 3 data contain beta value calculations, gene IDs and genomic coordinates for each probe on the array. In addition, data for probes that contain known single nucleotide polymorphisms (SNPs) after comparison to the dbSNP database (Build 128) and data for probes that contain repetitive element DNA sequences in more than 10 bp of each 50 bp probe sequence are masked with an “NA” descriptor.

The data packages used for the following analyses are listed below. Please note that with continuing updates of genomic databases, data archive revisions become available at the TCGA data portal. The following data archives were used for the analyses described in this manuscript:

Batches 9 and 11: jhu-usc.edu_OV.HumanMethylation27.Level_3.1.3.0
(Batches 9 and 11 included all fallopian tube samples)
Batch 12: jhu-usc.edu_OV.HumanMethylation27.Level_3.2.3.0
Batch 13: jhu-usc.edu_OV.HumanMethylation27.Level_3.3.3.0
Batch 14: jhu-usc.edu_OV.HumanMethylation27.Level_3.4.2.0
Batch 15: jhu-usc.edu_OV.HumanMethylation27.Level_3.5.1.0
Batch 17: jhu-usc.edu_OV.HumanMethylation27.Level_3.6.1.0
Batch 18: jhu-usc.edu_OV.HumanMethylation27.Level_3.7.1.0
Batch 19: jhu-usc.edu_OV.HumanMethylation27.Level_3.8.1.0
Batch 21: jhu-usc.edu_OV.HumanMethylation27.Level_2.9.0.0
Batch 22: jhu-usc.edu_OV.HumanMethylation27.Level_2.10.0.0
Batch 24: jhu-usc.edu_OV.HumanMethylation27.Level_2.11.0.0

During the course of data production, the platform manifest was updated to reflect current HUGO gene symbols, and to mask probes containing recently identified SNPs or repeats. This manifest update started with Batch 21. To ensure consistent gene symbol usage across all batches within this study, we used Level 2 data for batches 21, 22, and 24, and generated Level 3 data for these batches using identical procedures as for the earlier batches (masking the same probes, and reconciling gene symbols).

Cancer-associated epigenetic silencing

We used Level 3 DNA methylation data on 23,679 DNA methylation probes, and the median-integrated gene expression data set on 18,868 genes. The median based integrated expression data set was assembled using row-centered Level 3 data generated on the LBL-HuEx, UNC-Agilent and Broad-U133A platforms. This data set included every gene and every samples that has been profiled on one of these platform. If a gene was only assayed on one platform (n=1,116), this measurement was used. If the gene was assayed on two platforms (n=5,890), the average of the two measurements was used; if the gene was assayed on all platforms (gene on all three platforms n=11,864) the median measurement was used. This data set contains 541 samples (including 8 fallopian tube samples) and 18,868 genes. A set of 21,273 Infinium probes that interrogate 12,233 genes have matched gene expression data. We determined the Spearman correlation between DNA methylation and gene expression for 497 samples (including 489 tumor samples and eight fallopian tube samples. We used the non-parametric Spearman method, as bivariate normality could not be assumed (DNA methylation data are not normally distributed). Spearman's rank correlation coefficient (ρ) on the gene expression and DNA methylation was computed for each probe, along with a p value testing against the null hypothesis that ρ truly equals zero. The Benjamini-Hochberg procedure was used to control the false discovery rate. Given the small sample size for the fallopian tubes, we excluded 49 probes that failed on any of the fallopian tube samples, with 23,630 DNA methylation probes remaining, of which 21,229 covering 12,206 genes, could be matched to expression.

We describe below the strategy used to identify candidate epigenetically silenced genes. We apply four separate filtering criteria. The thresholds for each filter were selected based on inspection of scatterplots of DNA methylation versus gene expression for the genes passing the relevant filter criterion in the current data set. The four filters are:

1) The sample mean beta value of eight normal fallopian tubes with a relaxed threshold of 0.5 and a stringent threshold of 0.4.

2) The difference in DNA methylation between the 90th percentile tumor and mean fallopian tube methylation, with a relaxed threshold of 0.1 and a stringent threshold of 0.3.

3) The fold difference in mean gene expression between the fallopian tubes, and the 10% of tumor samples with the highest DNA methylation for that gene, with a relaxed threshold of 2-fold and a stringent threshold of 3-fold.

4) Spearman’s correlation coefficient between DNA methylation and gene expression calculated jointly across 489 tumor and 8 fallopian samples, with a relaxed ρ threshold of –0.2 and a stringent threshold of –0.3.

We required candidate epigenetically silenced genes to pass all four relaxed thresholds, and at least three out of four more stringent thresholds. If there were multiple CpGs for the same gene promoter, the CpG with the highest absolute Spearman’s Rho was retained for that gene. A complete list of the 168 genes is shown in **Table S7.1**, ranked by descending absolute Spearman’s Rho.

Definition of BRCA1 epigenetically silenced cases

We analyzed the relationship between DNA methylation and gene expression for nine different probes located in or near the BRCA1 promoter region, and found statistically significant inverse correlations for four of the nine probes (cg19531713, cg19088651, cg08993267, cg04658354). The target CpG sites of those probes are located in the CpG island that contains the transcription start site of BRCA1. The Spearman ρ of these correlations ranges from –0.28 to –0.37 (Benjamini-Hochberg adjusted $P<0.0001$). We did not see good inverse correlations for the other five probes located in other two CpG islands further away from the transcription start site.

For each of the aforementioned four probes, we used K-means clustering (assuming K=2) on the two-dimensional space of DNA methylation and expression data to separate the epigenetically silenced group and non-epigenetically silenced group of samples. Expression data were scaled to have the same range as DNA methylation data for this clustering. We then combined the calls from the four probes. Since data was lacking for some probes in some samples, we relied on the fraction of the four probes calling a particular sample epigenetically silenced for BRCA1, rather than on a fixed number of probes. Samples with >50% consensus on belonging to the epigenetically silenced group across the four probes were classified as samples with silencing of BRCA1 by promoter hypermethylation.

Validation of BRCA1 epigenetic silencing with MethyLight

The Infinium DNA Methylation data for the *BRCA1* promoter was valiadated for all 489 TCGA ovarian serous adenocarcinoma samples using MethyLight technology (PMID: 10344733; 10734209; 16326863; 16804544). The BRCA1-M1 MethyLight assay (HB-045) utilized primer and probe oligomers described previously (PMID: 15159323). MethyLight data are reported as a ratio between the value derived from the real-time PCR standard curve plotted as log (quantity) versus threshold C(t) value for the BRCA1 methylation reaction and likewise for a methylation-independent control reaction based on interspersed ALU repeats (PMID:16326863; 16804544). This calculation was performed for both the sample and an M.SssI-treated genomic DNA sample, which was used as a methylated reference. We calculated the percent of methylated reference (PMR) for each sample as: 100 X (BRCA1-M1 / ALU)sample / (BRCA1-M1 / ALU)M.SssI-Reference.

Unsupervised clustering analysis for DNA methylation subtype discovery

We performed unsupervised clustering analysis on 489 high-grade, high-stage TCGA samples based on DNA methylation data. We first removed 3,899 probes containing a single nucleotide polymorphism (SNP) within five base pairs of the target CpG site and those containing repeat element sequences of ≥10 base pairs.

Most normalization methods developed for gene expression arrays are not suitable for Infinium DNA methylation data for the following reasons: 1) The Cy3 and Cy5 dyes are not tied to the methylation status of the probes, 2) The majority of loci cannot be assumed to be unmethylated, 3) The total signal or methylation levels of different samples cannot be assumed to be equal, and 4) The measurements of the methylated and unmethylated probes are not independent. Moreover, we observed significant differences in biology and clinical parameters between analysis batches of samples. Therefore, rather than attempt to dissociate biology from technical batch effects through normalization, we chose to rely on robust probes for the unsupervised clustering, by eliminating probes that introduce technical noise. We compared technical replicates of the same sample (TCGA-07-0227) that were measured across ten different batches, to identify probes that are prone to variation across batches. The underlying assumption is that technical replicates should yield identical beta values for each probe. We determined the deviation from this assumption for each probe by calculating the distance of point (x,y) to y=x, in which x is the beta value of that probe for the same sample in one batch, and y the beta value for the other batch. Mathematically, this distance (D) can be calculated as $D=(y-x)/\sqrt{2}$) for each probe. We then calculated the rank-weighted mean (with penalty for probe failures) of the distances (D) calculated from 45 pair-wise comparisons for each of the 23,679 probes. To exclude as much technical noise as possible, we removed 10,589 probes with a weighed mean D greater than 0.05. Of the remaining 12,990 probes, we selected the most variant 858 probes for the clustering analysis. The variant probes were selected as the union of top 5% of probes with the largest standard deviation and top 5% with the largest adjusted standard deviation (σ') normalized for the Bernoulli distribution standard deviation for the associated mean ($\sigma'=\sigma/(\sqrt{(\mu(1-\mu))})$), since the maximal standard deviation of a beta distribution is influenced by its mean, and equals the standard deviation of a Bernoulli distribution.

Ovarian DNA methylation subtypes were discovered using consensus clustering (GenePattern, v 3.2.3. PMID: 16642009). The optimal number of clusters was determined with 1,000 resampling iterations (seed value: 12,345) using K-means clustering algorithm for K=2,3,4,…,10, with Euclidean distance as the distance measure.

We developed signatures for the clusters by selecting the top 50 probes for each cluster compared to the other clusters. The union of the probes (192 unique probes) was then clustered on the 489 samples to generate **Figure 3**. The R software (version 2.11.1) (http://www.r-project.org) was used for all data analyses.

Batch Effect Investigation

A fast singular value decomposition (SVD) done with the *corpcor* R package was used to extract the principal components (PCs). Each of the top ten PCs was tested for association with batch using linear regression. A univariate F test was used to test for the association of each probe and analytical batch. The Benjamini-Hochberg method was used to adjust for multiple comparisons and control the false-discovery rate. Cohen's f^2 was used to assess the effect size of the batch effect.

Other statistical analyses

Pearson's χ^2 test was used to assess the unequal distributions of categorical outcomes (e.g., differences in the frequencies of *BRCA1* inactivation events across DNA methylation clusters), with DNA methylation clusters. For covariates with fewer than five observations in any cell in the R*C contingency table, Fisher's exact test was used instead. A Logrank test was used to test against the null hypothesis that there was no difference between the Kaplan-Meier survival curves. Proportional hazards regression (Cox Regression) was used for parametric analysis to estimate hazard ratios associated with unit changes in any continuous variable, or the comparison of survival after adjusting for other variables, or test for an interaction term between two variables. The differences were considered significant if the two-sided p values are <0.05. All statistical tests were performed in R (http://www.r-project.org).

Table S7.1. A list of the 168 candidate epigenetically silenced genes. Genes are ranked by descending absolute Spearman's Rho. For genes with multiple CpG probes, we have listed the probe with the strongest inverse correlation.

Official Gene Symbol	ProbeID	Sample Mean Beta Values of Fallopian Tube	Beta Value Difference	Log2 (Fold Change)	Spearman's Rho
RAB25	cg19580810	0.33	0.50	4.57	-0.89
LYPLAL1	cg02665570	0.28	0.46	2.77	-0.86
ZNF597	cg24333473	0.13	0.60	2.11	-0.77
VTCN1	cg22424746	0.43	0.30	4.51	-0.76
VSIG2	cg02082342	0.37	0.43	2.58	-0.75
BANK1	cg25023994	0.33	0.41	1.96	-0.74
C3	cg17612991	0.35	0.43	5.07	-0.74
DNALI1	cg21488617	0.22	0.38	3.11	-0.72
LDHD	cg03991512	0.15	0.55	1.31	-0.70
UBB	cg06537829	0.16	0.48	3.29	-0.69
LTC4S	cg11394785	0.08	0.66	1.98	-0.68
SULT1C4	cg17966192	0.19	0.45	1.09	-0.67
NAP1L5	cg12759554	0.49	0.37	1.93	-0.67
CFTR	cg25509184	0.17	0.71	2.26	-0.65
CMBL	cg11882252	0.10	0.70	2.41	-0.64
TMEM173	cg16983159	0.10	0.40	2.90	-0.63
EYA4	cg21296676	0.35	0.47	2.46	-0.62
S100A16	cg23851011	0.45	0.35	1.68	-0.62
LRG1	cg24926276	0.29	0.26	3.62	-0.62
ALDOC	cg06367117	0.13	0.57	1.25	-0.61
CDO1	cg07644368	0.40	0.44	2.08	-0.61
ZNF671	cg19246110	0.11	0.75	1.45	-0.59
ZNF502	cg21672276	0.11	0.21	2.18	-0.58
TSPYL5	cg15747595	0.22	0.47	3.80	-0.58
MT1E	cg20083730	0.06	0.31	2.42	-0.57
TRIM4	cg01626227	0.06	0.32	1.92	-0.56
CFI	cg12243271	0.28	0.23	3.16	-0.56
DNAJB13	cg19692710	0.36	0.31	1.87	-0.56
AQP9	cg11098259	0.23	0.39	2.98	-0.56
THY1	cg12508624	0.12	0.64	1.21	-0.55
CDH16	cg14221831	0.43	0.36	1.66	-0.55
KLK11	cg09702010	0.20	0.38	4.00	-0.55
CLIP3	cg06432655	0.23	0.44	1.49	-0.55
NUPR1	cg05590982	0.18	0.44	2.68	-0.54
EHF	cg18414381	0.27	0.14	2.79	-0.54
UQCRH	cg21576698	0.42	0.49	3.30	-0.54
LCN12	cg19534945	0.40	0.46	2.29	-0.54
TMEM71	cg20955688	0.16	0.53	1.97	-0.54
TMEM140	cg06456031	0.20	0.33	1.32	-0.54
SCG5	cg15787039	0.14	0.36	2.43	-0.54

KLK10	cg06130787	0.15	0.21	1.81	-0.54
WT1	cg16463460	0.18	0.25	4.79	-0.53
SCARA3	cg26847866	0.15	0.22	2.51	-0.53
APOL6	cg19853703	0.07	0.14	1.92	-0.52
SLAIN1	cg08504583	0.38	0.45	2.25	-0.52
PCDHB5	cg03349953	0.17	0.51	2.03	-0.51
HSPB2	cg13210534	0.12	0.33	1.65	-0.51
CHI3L2	cg26366091	0.22	0.39	1.09	-0.51
SLC15A2	cg10523671	0.36	0.53	2.43	-0.50
RBP1	cg13099330	0.09	0.58	3.53	-0.50
CDH6	cg10919204	0.34	0.38	1.60	-0.50
LRRC34	cg24777454	0.10	0.50	3.13	-0.50
SPDEF	cg07705908	0.34	0.21	1.66	-0.50
PDLIM4	cg01305625	0.35	0.23	1.61	-0.49
RERG	cg19205533	0.31	0.38	4.03	-0.49
EFS	cg07197059	0.22	0.47	1.06	-0.49
HSPA1A	cg05920090	0.04	0.18	3.93	-0.49
HOXB8	cg15539420	0.33	0.38	1.13	-0.49
MFAP4	cg09606564	0.19	0.45	3.35	-0.48
AMT	cg25021247	0.38	0.35	1.90	-0.48
CRAT	cg26805528	0.17	0.55	1.64	-0.48
PCDHB2	cg02260587	0.16	0.45	1.80	-0.48
CRISPLD1	cg01410472	0.05	0.31	2.88	-0.48
HOXB2	cg09313705	0.31	0.32	3.24	-0.48
FXYD1	cg27461196	0.36	0.31	2.99	-0.47
APOBEC3G	cg26022401	0.09	0.24	3.43	-0.47
HP	cg06172871	0.39	0.30	3.00	-0.46
ZNF300	cg19014419	0.12	0.24	1.97	-0.46
TRIM22	cg12461141	0.32	0.39	3.31	-0.46
RIPK3	cg20822579	0.21	0.36	1.54	-0.45
HLA-DMA	cg14833385	0.12	0.11	3.47	-0.45
DDR2	cg22740835	0.15	0.69	2.42	-0.45
STAT5A	cg03001305	0.22	0.33	1.24	-0.45
WDR69	cg14329157	0.26	0.45	5.30	-0.45
PTGDS	cg11546621	0.41	0.37	2.62	-0.45
DDO	cg20011134	0.34	0.46	1.34	-0.44
GYPC	cg13901526	0.22	0.54	2.27	-0.44
PAM	cg20131596	0.20	0.28	2.20	-0.44
CRYAB	cg15227610	0.33	0.40	2.90	-0.44
FADS2	cg06781209	0.17	0.47	1.20	-0.44
FCGRT	cg15528736	0.31	0.47	1.92	-0.43
ARSE	cg11964613	0.35	0.48	1.32	-0.43
RNASE1	cg05958352	0.41	0.35	1.62	-0.43
AGT	cg19125606	0.43	0.44	1.95	-0.43
TBX2	cg12163132	0.19	0.12	2.36	-0.43
PKIA	cg04689061	0.17	0.37	1.22	-0.42
THNSL2	cg07952391	0.08	0.27	2.05	-0.41

FOXJ1	cg24164563	0.08	0.19	3.66	-0.41
CPNE8	cg23495733	0.13	0.72	1.68	-0.41
CYBRD1	cg10731149	0.12	0.14	1.64	-0.41
IL20RA	cg22487322	0.43	0.35	3.08	-0.41
SPATA18	cg09022993	0.10	0.51	4.31	-0.40
PLSCR4	cg24315815	0.27	0.22	2.99	-0.40
C1S	cg05538432	0.35	0.28	3.59	-0.40
VNN2	cg10044101	0.45	0.34	1.79	-0.40
TMEM101	cg12259256	0.09	0.59	2.92	-0.39
FOLR1	cg03699566	0.30	0.23	3.56	-0.39
VAMP5	cg11108890	0.30	0.54	1.36	-0.39
GSTM2	cg16670497	0.04	0.11	1.69	-0.39
PART1	cg09712066	0.37	0.12	3.50	-0.38
PNOC	cg03642518	0.24	0.26	3.26	-0.38
SEMA3E	cg18464137	0.14	0.41	1.10	-0.38
SERPINA3	cg06190732	0.31	0.44	4.76	-0.38
TRIM59	cg10273210	0.09	0.32	1.18	-0.38
LIMS3	cg18879041	0.20	0.13	3.81	-0.38
CYP4B1	cg23440155	0.16	0.16	4.08	-0.38
SPAG6	cg06908778	0.20	0.62	4.94	-0.38
OVGP1	cg09558502	0.13	0.10	6.66	-0.38
SERPINB1	cg06148264	0.16	0.13	3.02	-0.38
GIMAP2	cg25918245	0.24	0.24	1.96	-0.37
CLEC11A	cg13152535	0.26	0.32	1.12	-0.37
IQGAP2	cg02387679	0.14	0.49	1.08	-0.37
WIT1	cg19718882	0.10	0.13	2.10	-0.37
KIAA1324	cg16797831	0.13	0.40	3.78	-0.37
ANGPTL1	cg07044282	0.39	0.20	2.64	-0.36
MCAM	cg21096399	0.29	0.34	1.00	-0.36
CRIP1	cg02000005	0.18	0.37	2.65	-0.36
SLC47A2	cg24743310	0.36	0.16	3.66	-0.36
GNB4	cg17483510	0.10	0.36	1.22	-0.36
GAS2L2	cg24922045	0.30	0.17	2.26	-0.35
ZMYND12	cg06346081	0.16	0.14	2.36	-0.35
ALDH3B1	cg07730301	0.16	0.13	1.86	-0.35
SLC44A4	cg07363637	0.38	0.16	4.05	-0.35
NUAK1	cg23555120	0.48	0.36	1.63	-0.35
HOXB5	cg01405107	0.13	0.48	1.89	-0.35
LY75	cg23995753	0.23	0.32	2.68	-0.35
CXCR7	cg03626672	0.17	0.16	1.92	-0.35
PLAT	cg12091331	0.14	0.19	3.65	-0.34
CTSO	cg11754095	0.10	0.11	2.60	-0.34
ZNF655	cg13636404	0.05	0.12	1.99	-0.34
CAMK2N1	cg08398233	0.17	0.29	1.94	-0.34
BRCA1	cg04658354	0.06	0.46	1.17	-0.34
ANXA6	cg21623671	0.10	0.10	1.61	-0.34
GIPC2	cg09107315	0.32	0.44	1.38	-0.33

IL1R2	cg20340242	0.35	0.35	1.63	-0.33
IGF1	cg01305421	0.43	0.35	2.23	-0.33
KCTD14	cg17272843	0.08	0.21	2.67	-0.33
STEAP2	cg27626102	0.13	0.23	1.92	-0.33
NPDC1	cg26581729	0.38	0.42	3.19	-0.32
FBLN2	cg00201234	0.30	0.40	1.14	-0.32
H1F0	cg07141002	0.25	0.16	1.64	-0.32
GCNT3	cg06817269	0.37	0.13	2.50	-0.32
NDN	cg12532169	0.48	0.30	3.90	-0.32
TRIM2	cg12793610	0.23	0.20	2.04	-0.31
CPXM2	cg09619146	0.23	0.43	2.02	-0.31
HNF1B	cg12788467	0.19	0.62	2.31	-0.31
CTSS	cg08578023	0.15	0.21	2.39	-0.31
NME5	cg25507001	0.11	0.11	3.48	-0.31
SLC16A5	cg09300114	0.23	0.14	1.77	-0.31
PEG3	cg18668753	0.41	0.34	2.19	-0.30
BLNK	cg16779976	0.25	0.20	2.22	-0.30
RARRES2	cg17279839	0.12	0.50	3.14	-0.30
SPARCL1	cg19466563	0.16	0.62	3.47	-0.28
CBX7	cg23124451	0.33	0.50	2.17	-0.27
CCDC65	cg02620769	0.04	0.39	3.98	-0.27
APH1B	cg17207590	0.26	0.38	1.98	-0.27
TSC22D3	cg00404599	0.38	0.38	1.67	-0.26
CCL21	cg27443224	0.35	0.45	3.38	-0.25
MRGPRF	cg22933847	0.34	0.33	1.95	-0.25
HSPA2	cg16319578	0.16	0.47	2.77	-0.24
PENK	cg24645221	0.07	0.47	1.85	-0.24
LONRF2	cg12232463	0.36	0.47	3.29	-0.23
SERPING1	cg09061733	0.18	0.32	2.19	-0.22
ALDH1A3	cg21359747	0.23	0.70	1.92	-0.22
MGP	cg00431549	0.30	0.31	2.99	-0.22
CYYR1	cg10238818	0.07	0.48	2.45	-0.21
TCTEX1D1	cg24110050	0.25	0.56	3.80	-0.21
AIF1	cg21440587	0.17	0.46	1.76	-0.21

Table S7.2. Distribution of *BRCA* inactivation events across promoter methylation clusters.

DNA Methylation Cluster	MC1	MC2	MC3	MC4	TOTAL	P value	Total Sample set
All Samples	131	64	156	138	489	-	-
Sequenced Samples	80	38	115	83	316	-	-
Samples With Known *BRCA* Status*	88	38	128	84	338	-	-
All *BRCA* Inactivtion Events	41	5	57	17	120	4.9E-06	338
All *BRCA* Inactivtion Events (%)	46.6%	13.2%	44.5%	20.2%	35.5%	-	338
***BRCA1* Epigenetic Silencing**	19	1	30	6	56	1.0E-05	489
BRCA* Mutation*	22	4	27	11	64	0.04	338
All Germline Mutations	18	2	20	6	46	0.03	338
***BRCA1* Germline Mutation**	11	1	11	4	27	0.19	338
***BRCA2* Germline Mutation**	7	1	10	2	20	0.17	338
All Somatic Mutations	4	2	8	5	19	0.97	338
***BRCA1* Somatic Mutation**	1	2	7	0	10	0.05	338
***BRCA2* Somatic Mutation**	3	0	1	5	9	0.09	338

* 316 sequenced cases + 22 samples with epigenetic silencing but no sequencing data; due to the known mutual exclusivitiy observed, we assume no mutation events in those 22 samples.

** Two cases in Cluster MC3 have both more than one mutations: TCGA-20-1684 has BRCA2 germline mutations and BRCA1 somatic mutation; TCGA-13-1501 has BRCA2 germline mutations and BRCA1 germline mutation; Otherwise all inactivation events are mutually exclusive.

Table S7.3. Overlap between gene expression and DNA methylation subtypes.

	Differentiated Gene Expression Subtype	Immunoreactive Gene Expression Subtype	Mesenchymal Gene Expression Subtype	Proliferative Gene Expression Subtype	Total
DNA Methylation Subtype 1 (MC1)	55	31	24	21	**131**
DNA Methylation Subtype 2 (MC2)	3	2	10	49	**64**
DNA Methylation Subtype 3 (MC3)	41	32	60	23	**156**
DNA Methylation Subtype 4 (MC4)	36	42	15	45	**138**
Total	**135**	**107**	**109**	**138**	**489**

Figure S7.1A. Scatterplots showing BRCA1 gene expression versus promoter methylation. The color and size of the dots represent tissue type (red/large – fallopian tube samples, n=8; other colors/small – ovarian tumors, n=489. Specifically, blue dots represent tumors with *BRCA1* epigenetically silencing; green dots represent tumors with *BRCA1* germline mutation; purple dots represent tumors with *BRCA1* somatic mutation. Unsequenced tumors were shown with hollow dots). Plotted in the y-axis is the relative mRNA expression level of *BRCA1* as log ratios reported in the median-integrated expression data set, and in the x-axis is the DNA methylation beta value.

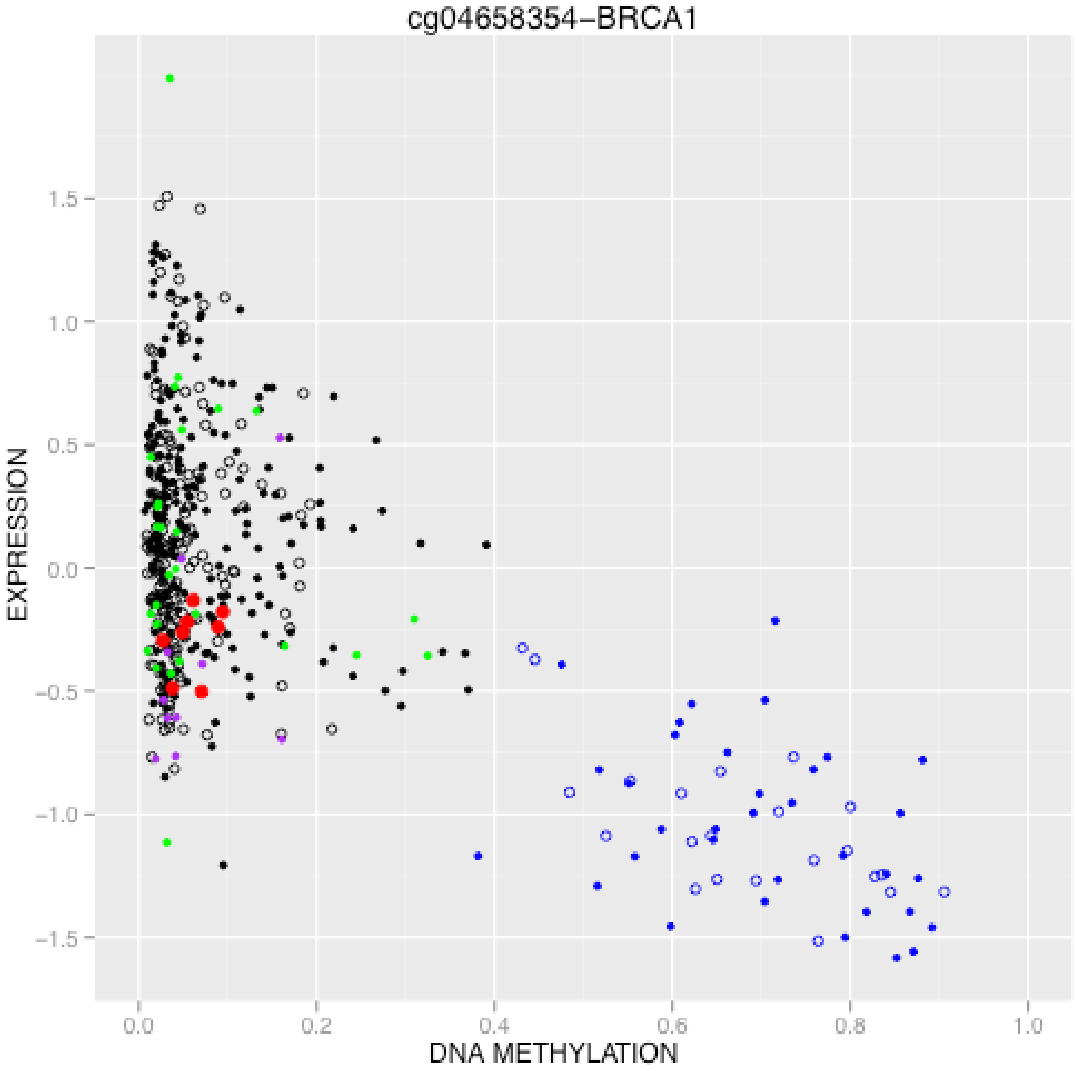


Figure S7.1B. Scatterplots showing RAB25 gene expression versus promoter methylation. The color and size of the dots represent tissue type (red/large – fallopian tube samples, n=8; black/small – ovarian tumors, n=489).

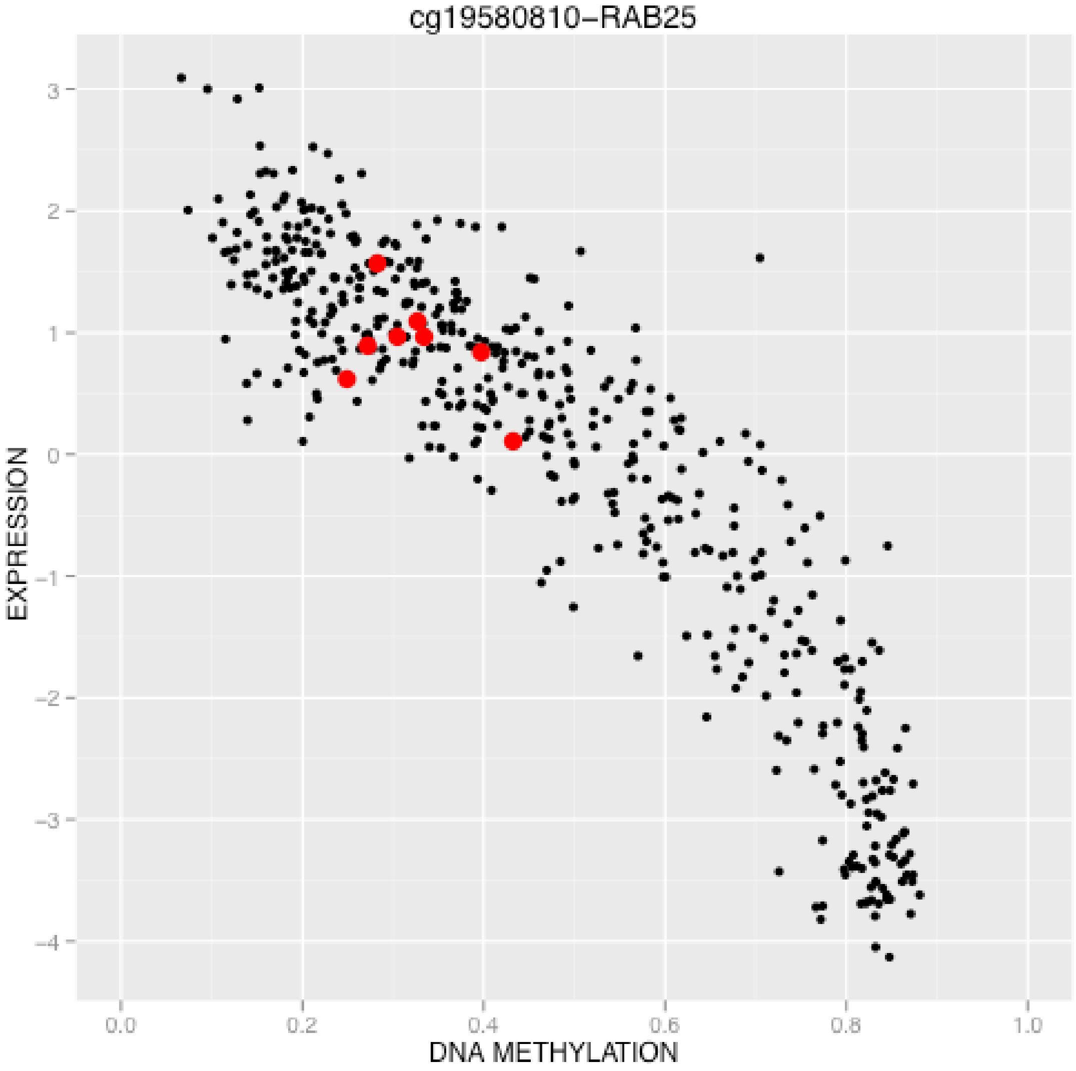


Figure S7.1C. Scatterplots showing AMT gene expression versus promoter methylation. The color and size of the dots represent tissue type (red/large – fallopian tube samples, n=8; black/small – ovarian tumors, n=489).

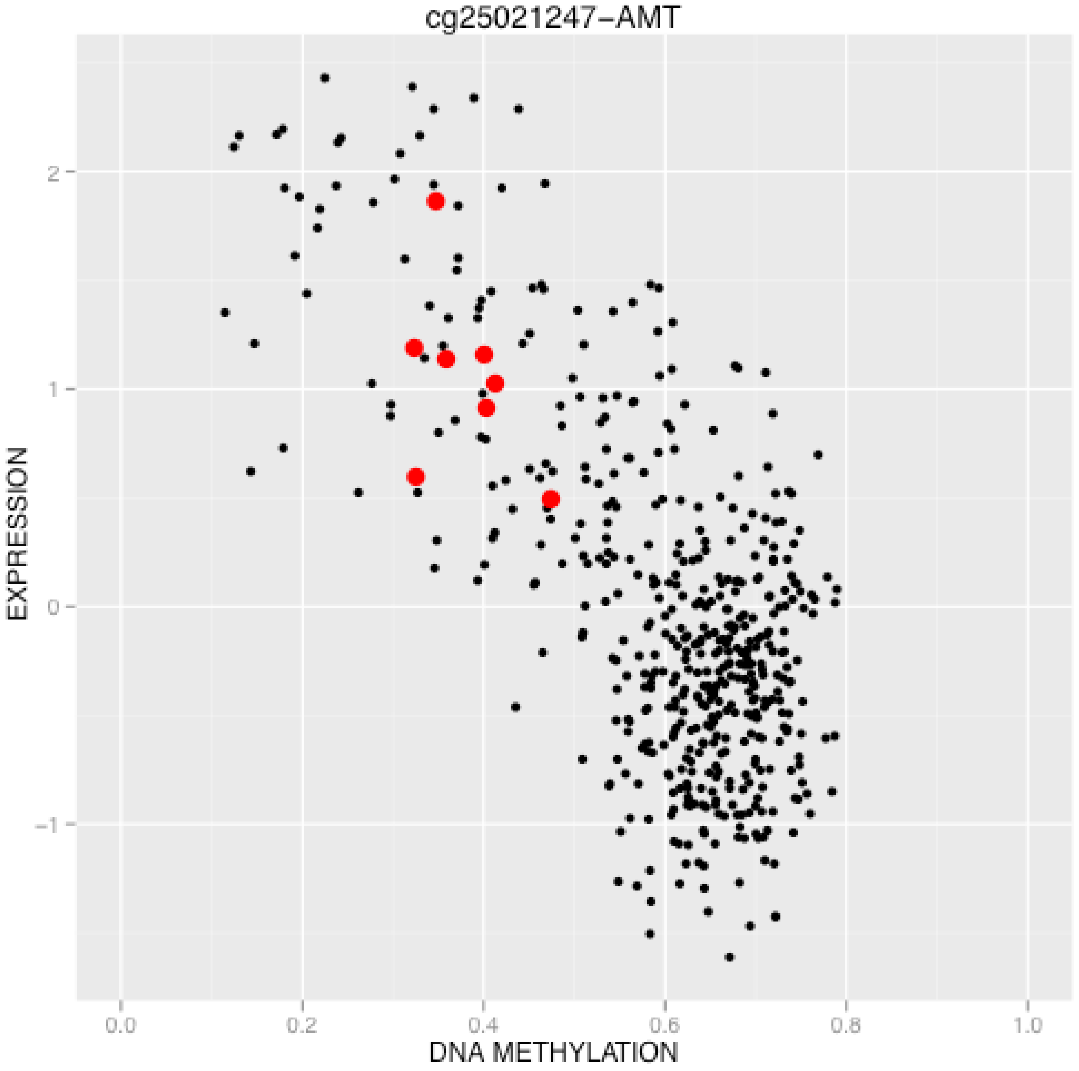


Figure S7.1D. Scatterplots showing *SPARCL1* gene expression versus promoter methylation. The color and size of the dots represent tissue type (red/large – fallopian tube samples, n=8; black/small – ovarian tumors, n=489).

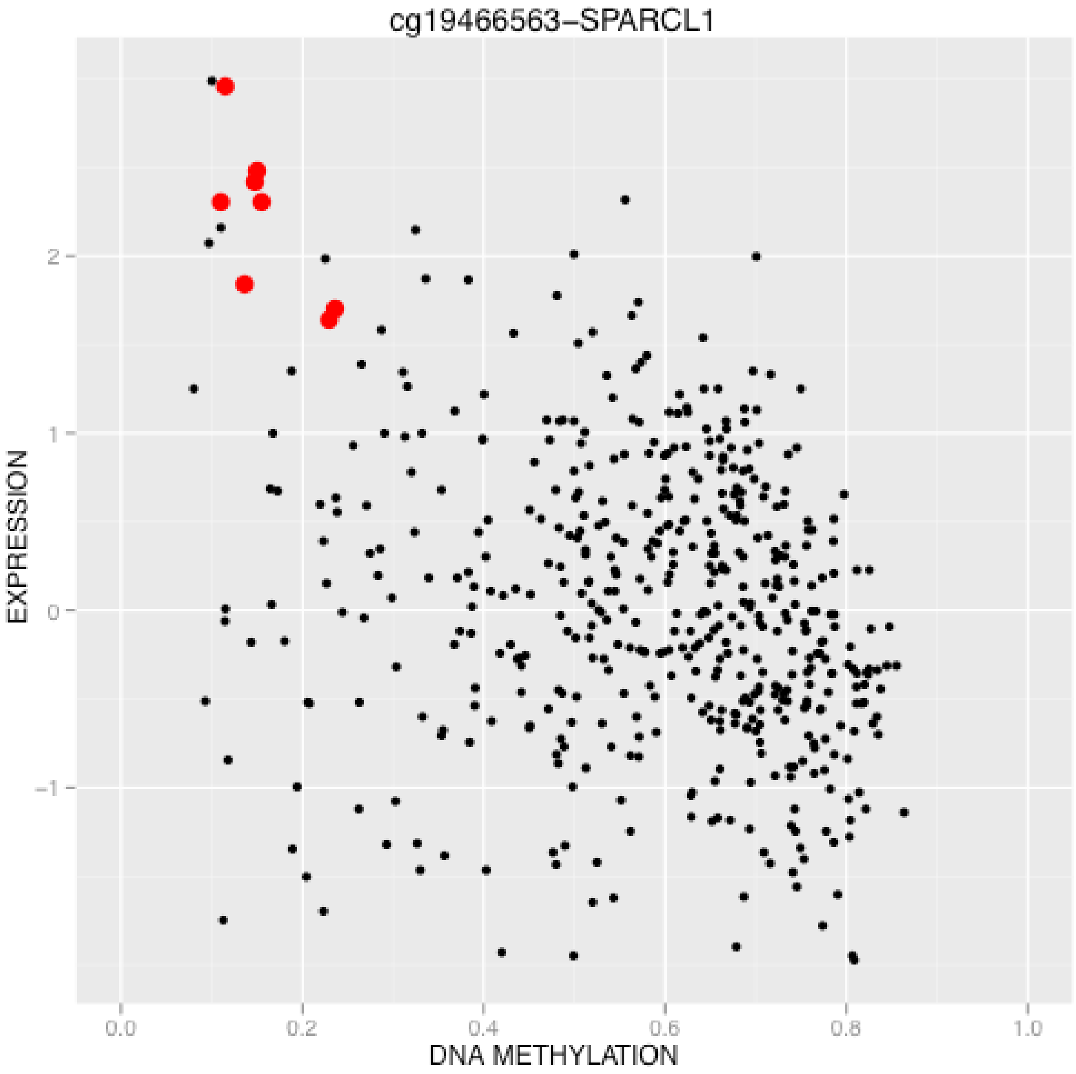


Figure S7.1E. Scatterplots showing *CCL21* gene expression versus promoter methylation. The color and size of the dots represent tissue type (red/large – fallopian tube samples, n=8; black/small – ovarian tumors, n=489).

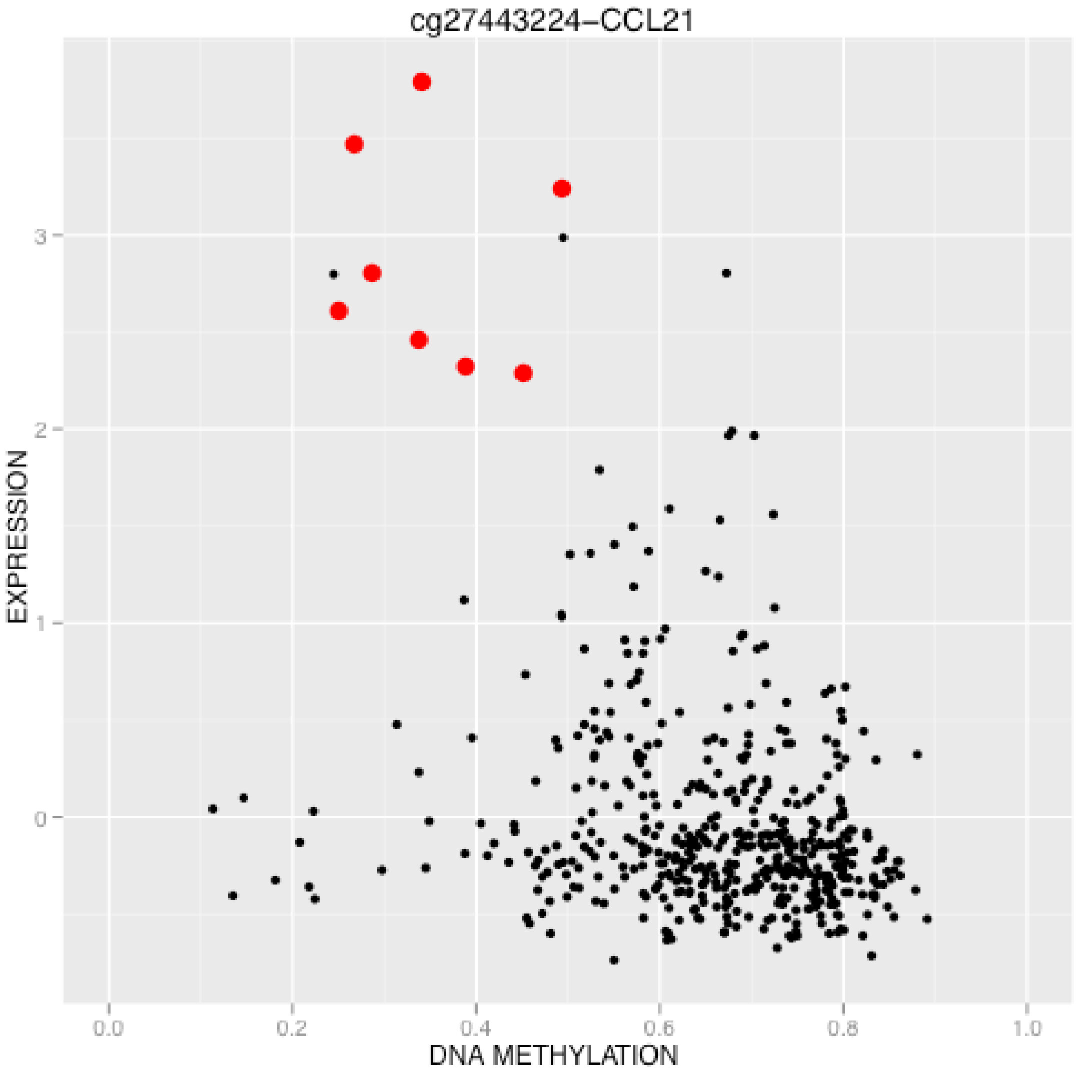


Figure S7.2. Scatter plots showing pairwise comparison of the Infinium beta values and MethyLight PMR values for the 489 ovarian serous adenocarcinomas. Left four columns (upper four rows) are the four *BRCA1* probes used for making the epigenetical silencing calls. The fifth column(row) shows the PMR values given by MethyLight. The lower left panels show the pairwise comparison for each of the five measurements. Each dot represents a sample. The red line indicates a Loess regression fit (alpha=1.2). The numbers at the upper right panels show the Pearson's Correlation Coefficient of the two measurements at each intersection. The Infinium probes and MethyLight probe are arranged by genomic location. All five are located in the same CpG Island that flanks *BRCA1* transcription start site by the Takai Jones definition.

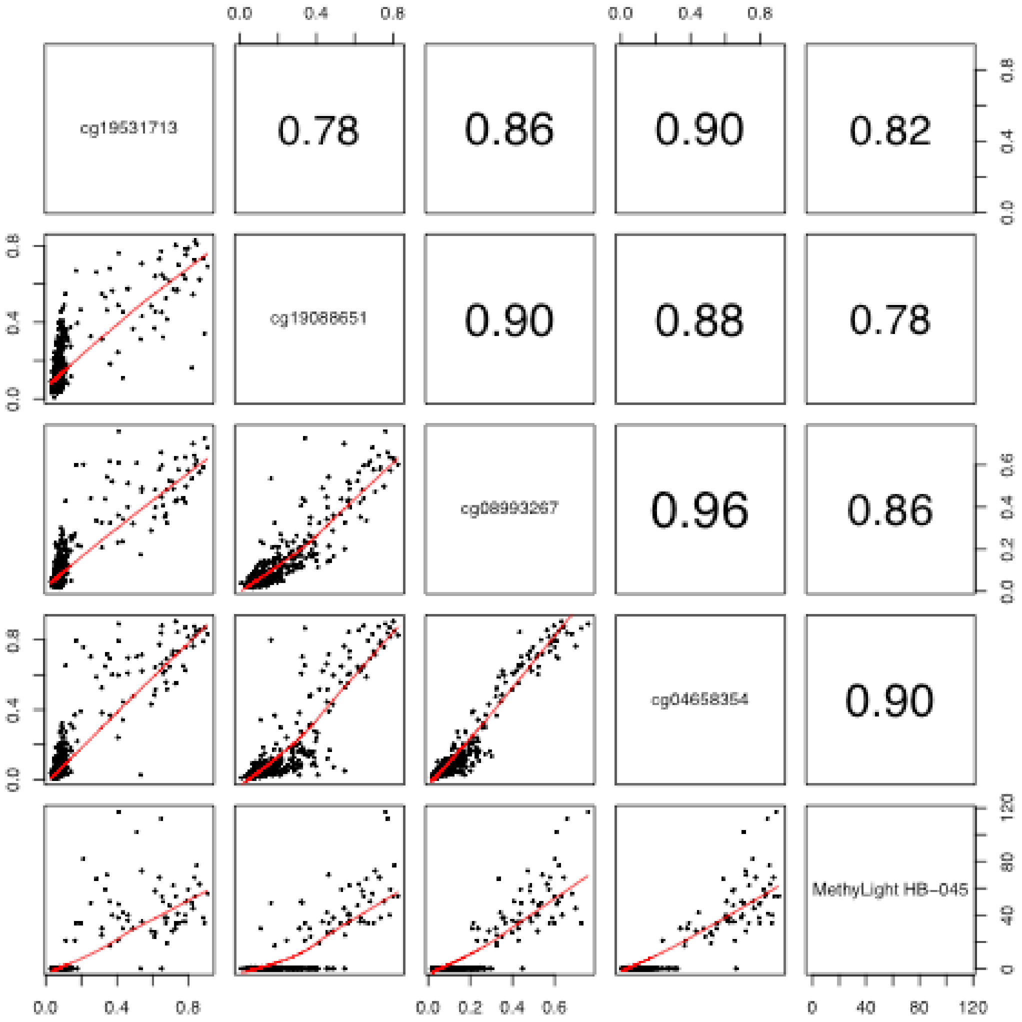

cg19531713
0.78
0.86
0.90
0.82
cg19088651
0.90
0.88
0.78
cg08993267
0.96
0.86
cg04658354
0.90
MethyLight HB−045
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0
40
80
120

Figure S7.3. ROC curve of *BRCA1* methylation validation by MethyLight. The ability of MethyLight measurement of *BRCA1* methylation to discriminate BRCA1 epigenetic silenced cases from non-silenced cases, as determined by the Illumina Infinium DNA methylation and gene expression measurements (see methods) is depicted as an ROC curve. (AUC=0.99).

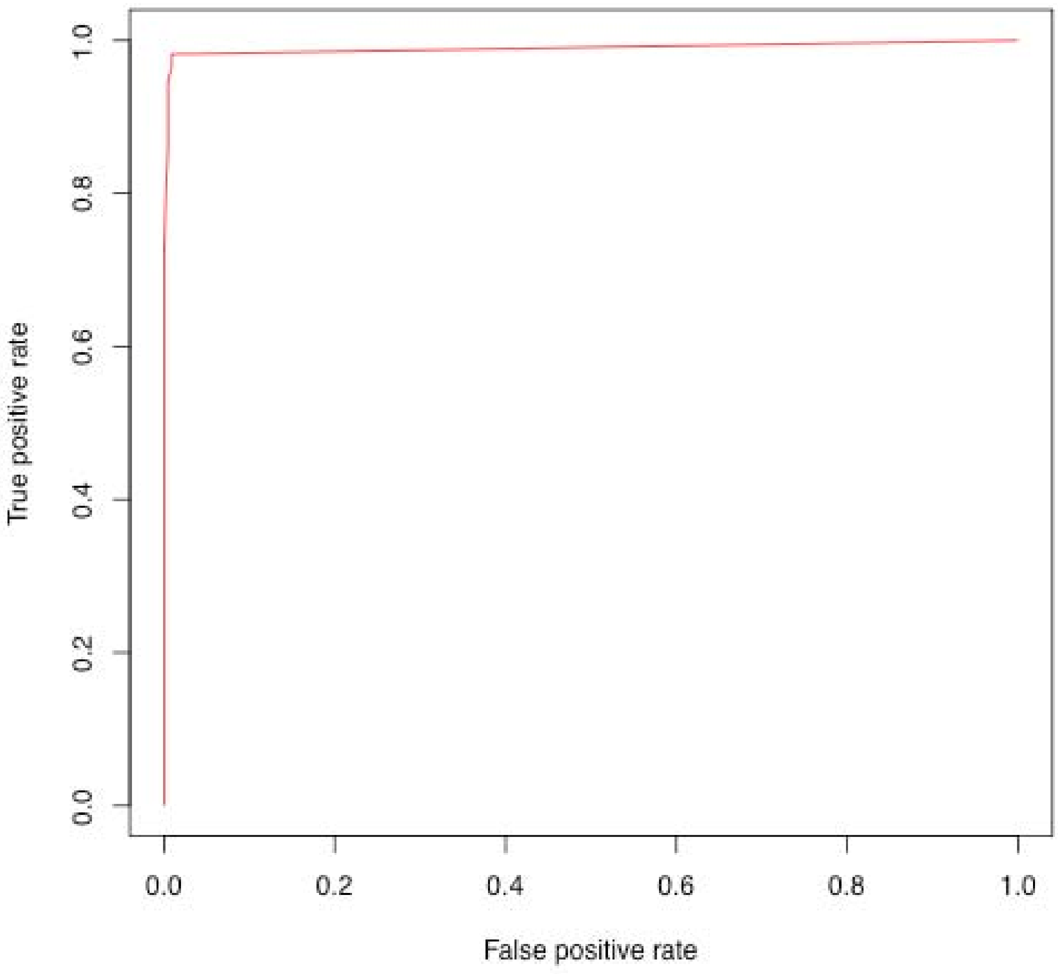


Figure S7.4. Consensus clustering was performed on 489 serous ovarian tumor samples with 858 Infinium probes, selected as described in Supplemental Methods. DNA cluster membership was determined by 1,000 resampling iterations of consensus clustering using the K-means algorithm. Hierarchical clustering of the 192 most discriminant probes is shown in the heatmap, with eight fallopian tube samples shown on the left. DNA methylation levels (beta value) are shown with a color spectrum as indicated in the color key panel, with blue indicating no methylation (beta value=0), to red, indicating full methylation (beta value=1). White indicates missing value. DNA methylation cluster memberships of the tumors are indicated by the color bar: *blue*, Cluster MC1 (n=131); *green*, Cluster MC2 (n=64); *red*, Cluster MC3 (n=156), *purple*, Cluster MC4 (n=138). Other color bars indicate various molecular features as indicated in the color key. There is no association between the DNA methylation clusters and analytical batch (bottom bar, p=0.85, χ^2 test)

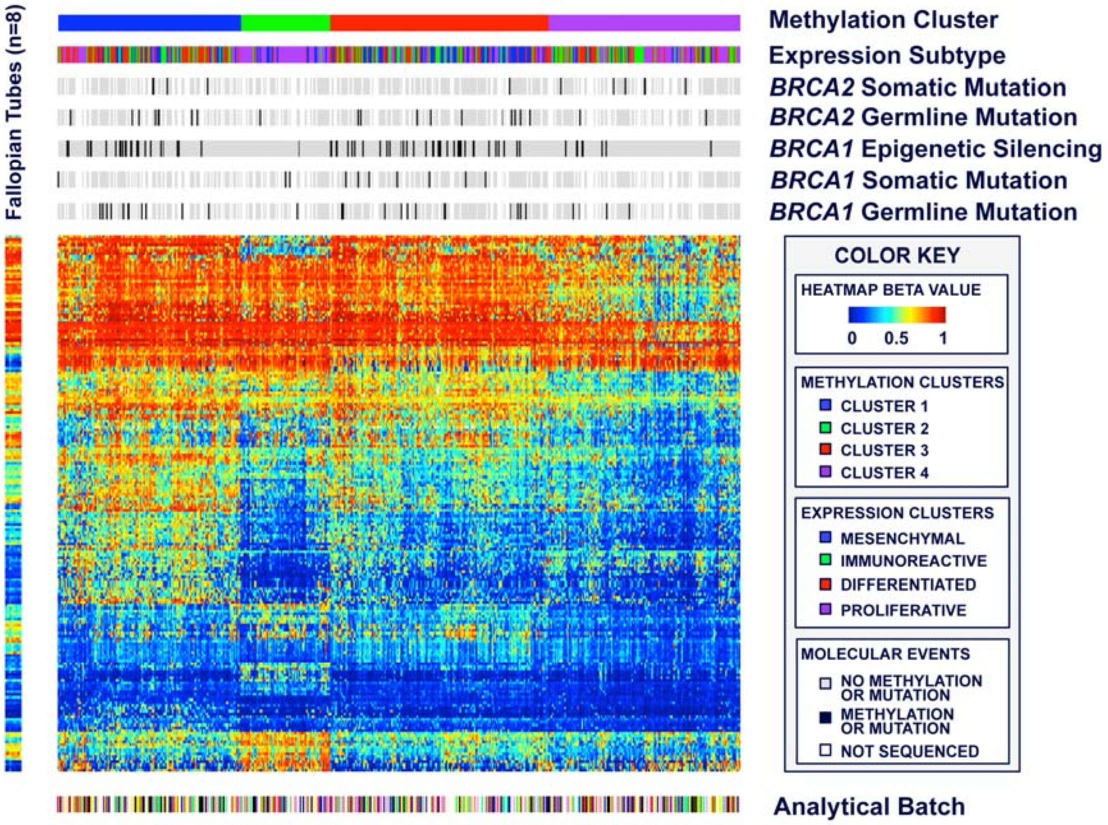
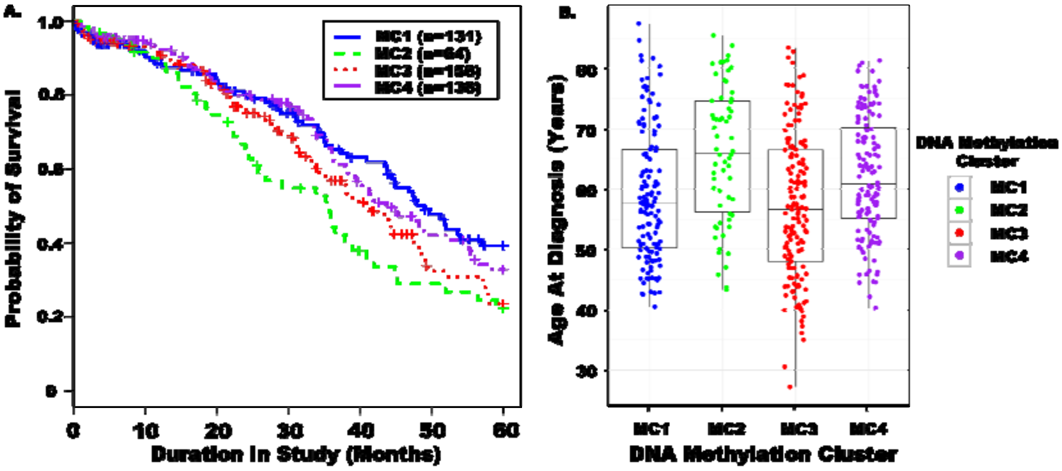


Figure S7.5. Clinical relevance of the DNA methylation clusters. A. Kaplan-Meier curves showing the differential survival of the four DNA methylation clusters with five-year censored survival data. Samples are colored according to their cluster membership as described in **Figure S7.2**. The four clusters differ in overall survival (Median survival time: Median survival time: Cluster MC1 – 48.9 months, Cluster MC2 – 35.8 months; Cluster MC3 – 40.9 months; Cluster MC4 – 43.6 months; Logrank test, p=0.04.) B. The distributions of age at diagnosis for patients in the three DNA methylation clusters are shown in the box-plots, and patients in the three DNA methylation clusters differ in age at diagnosis (One-way ANOVA, p=5*10-7). Tukey HSD test revealed that patients in cluster MC2 are an average of 6.7 years older than the patients in cluster MC1 (95% CI: 2.33-11.15; mean age: 65.8 v.s. 59.1 years; adjusted p=0.0005) and 8.7 years (95% CI: 4.4 -13.0 years; mean age: 65.8 v.s. 57.1 years; adjusted p=0.000002) older than cluster MC3 (and 57.1 years, adjusted p=0.0005 and 0.000002 respectively), and Cluster MC4 patients are 5.0 years older than patients belonging to cluster MC3 (95% CI: 1.6 – 8.4 years; mean age: 62.1 v.s. 57.1 years, adjusted p=0.0009).



Supplemental Methods 8:

8.1 Introduction

We analyzed several pathways that are generally altered in different cancer types, specifically the RAS/PI3K, RB, and p53 signaling pathways, as well as the homologous recombination (HR) pathway, which has germline as well as somatic alterations in ovarian cancer. For all pathway analyses, we used the set of cases (N=316) with complete data (mRNA expression, DNA copy-number, methylation, and protein mutations).

Figure S8.1 outlines the assessment approach used to determine whether a particular gene was altered or not altered in a particular sample. Our approach was based on first examining each gene across all samples, and binning each gene into one of four categories:

- Category 1: Gene is altered by mutations.
- Category 2: Gene is primarily altered by copy number alterations, and mRNA expression levels correlate with copy number changes.
- Category 3: Gene is epigenetically silenced.
- Category 4: Gene has evidence of a bimodal expression pattern, unrelated to copy number status.

As outlined in Figure S8.1, we then used different alteration criteria for each of the four categories. For example, for Category 2 genes, we classified each gene as a likely oncogene or tumor suppressor, and a gene was called altered in a specific sample if the gene was altered by a high level copy-number amplification or homozygous deletion (as defined by GISTIC, see

Supplemental Methods 5). Category 3 epigenetically silenced genes were defined by k-means clustering; for example, for *BRCA1*, we used k-means clustering on the two-dimensional space of DNA methylation and expression data to separate the epigenetically silenced group and the non-epigenetically silenced group of samples. Finally, for category 4 genes, alteration status was defined by relative expression compared to the expression distribution in tumor samples diploid in the particular gene, ≥ one standard deviation. In all categories, a gene was called altered if the gene contained a non-synonymous, somatic (or in the case of BRCA1/2, a germline) mutation in a protein-coding region.

A pathway was considered altered in a given sample, if at least one gene in the pathway was altered.

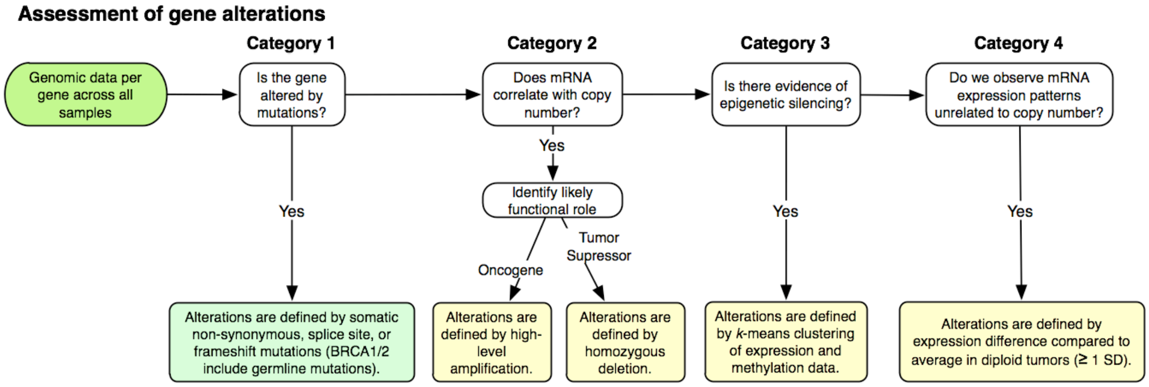

Assessment of gene alterations
Category 1
Category 2
Category 3
Category 4
Genomic data per gene across all samples
Is the gene altered by mutations?
Does mRNA correlate with copy number?
Is there evidence of epigenetic silencing?
Do we observe mRNA expression patterns unrelated to copy number?
Yes
Identify likely functional role
Oncogene
Tumor Supressor
Yes
Yes
Yes
Alterations are defined by somatic non-synonymous, splice site, or frameshift mutations (BRCA1/2 include germline mutations).
Alterations are defined by high-level amplification.
Alterations are defined by homozygous deletion.
Alterations are defined by *k*-means clustering of expression and methylation data.
Alterations are defined by expression difference compared to average in diploid tumors (≥ 1 SD).

Figure S8.1. Assessment of gene alterations used in pathway analysis.

8.2. Cancer Pathways

TP53 pathway

For the TP53 protein, we observe a mutation rate of 87%. With the depth of coverage of TP53 with the hybrid capture and next generation sequencing approaches, it is possible and even likely that a subset of mutations in TP53 were missed raising the possibility that TP53 mutations are essentially universal. Samples with truncating TP53 mutations, i.e. nonsense, splice, and frame shift mutations (approximately one third of cases) have markedly lower TP53 expression than those with missense mutations or in-frame deletions (Figure S8.2), possibly caused by nonsense-mediated decay (NMD) of mRNA (17 samples with low expression are candidates for missed truncating mutations). Amplifications of *MDM2* and *MDM4* are uncommon, occurring in 4% and 3% of cases, respectively.

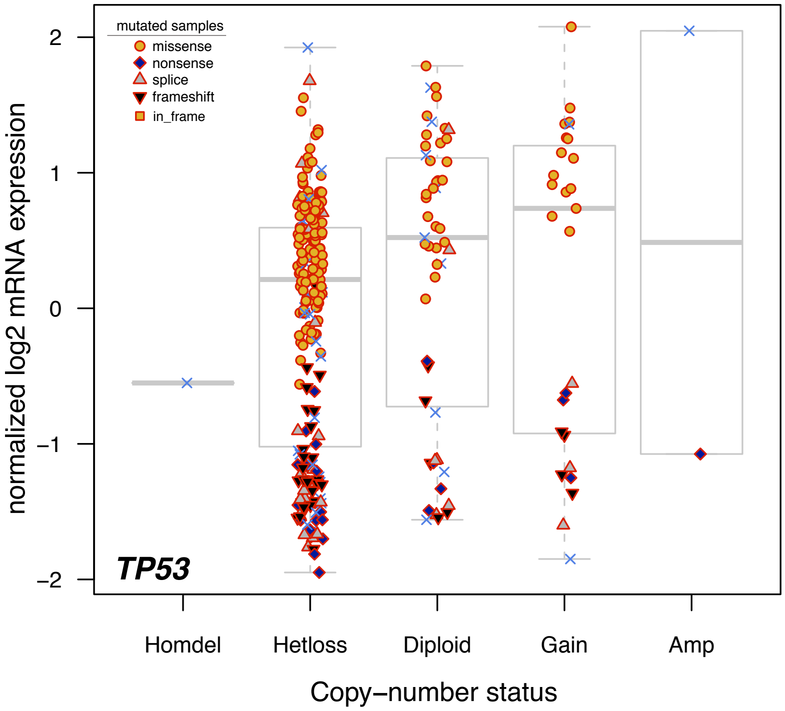


Figure S8.2: Truncating mutations of TP53 lead to markedly lower transcript levels, independent of copy-number status.

RB pathway

Amplification of *CCNE1* is one of the most common focal copy number change events in serous ovarian cancer, occurring at a frequency of 20%. *RB1*, immediately downstream of *CCNE1*, is deleted in 25 samples and mutated in an additional nine samples (10.8% of cases combined). As is the case with *PTEN* and *NF1* (see below), some of the *RB1* deletions are intragenic, i.e., do not affect the entire gene, and cases with intragenic deletions have low mRNA expression at the exon level but not the whole gene level (data not shown).

CDKN2A, a negative regulator of cyclins and cyclin-dependent kinases, is frequently altered in various types of cancer, typically by deletion or epigenetic regulation. In this data set, we observe a striking bimodal expression pattern, with approximately one third of the cases with very low or no expression (Figure S8.3). There is no evidence for *CDKN2A* promoter methylation in the samples with low expression. Low *CDKN2A* mRNA expression is mutually exclusive with *CCNE1* amplification and *RB1* deletion/mutation events ($P = 4.726\text{e-}11$, two-sided Fisher's Exact Test, Figure S8.4).

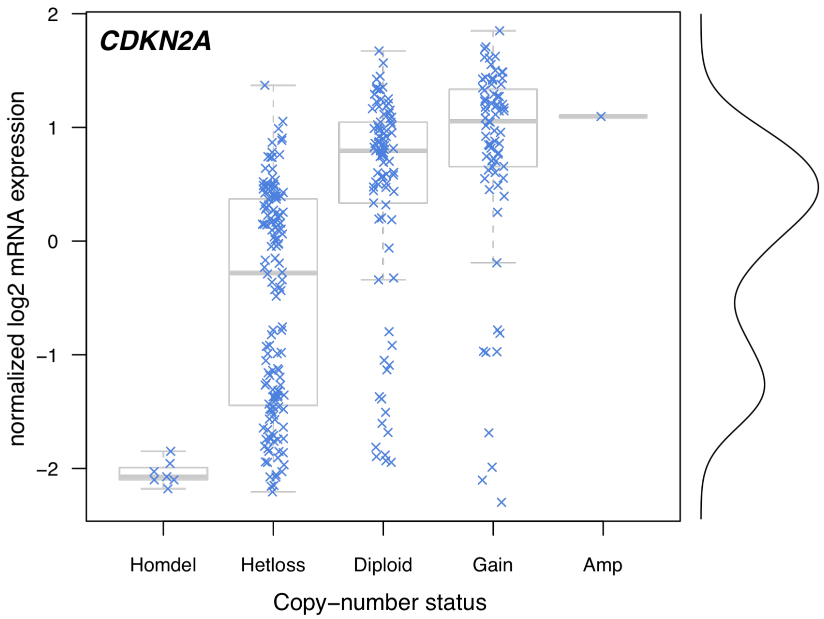


Figure S8.3: Bimodal expression pattern of *CDKN2A*.

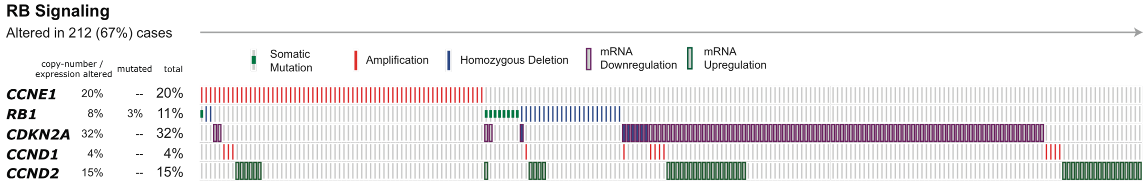


Figure S8.4: Alteration pattern in the RB signaling pathway. Each column represents an individual case; each row represents a gene. Only cases with RB signaling alterations (N=212) are shown. The percent altered is relative to N=316.

RAS/PI-3-Kinase-signaling

Various key members of the RAS/PI3K pathway are frequently altered by several different mechanisms in ovarian cancer[1]. The most commonly altered genes in the pathway are *PTEN* (homozygous deletion or mutation), *PIK3CA* (amplification or mutation), *KRAS* (amplification or mutation), *NF1* (homozygous deletion or mutation), as well as *AKT1* and *AKT2* (amplification) (Figure S8.5). Known activating mutations are observed in *PIK3CA* (two cases, E545A and H1047R), *KRAS* (two cases, both G12V), and *BRAF* (one case, N581S).

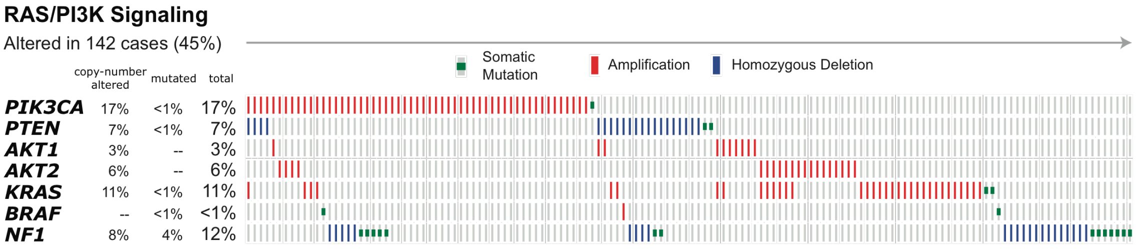


Figure S8.5: Alteration pattern in the RAS/PI3K signaling pathway. Each column represents an individual case; each row represents a gene. Only cases with RAS/PI-3-K signaling alterations (N=142) are shown. The percent altered is relative to N=316.

A fraction of the homozygous deletions of *PTEN* and *NF1* are intragenic, i.e. they only affect part of the gene. In these cases, we usually observe lower expression of the deleted exons than of the rest of the gene (Figure S8.6, A-C).

We also observed uncommon but focal amplification of *ERBB2* (4 cases, 1.3%) and *ERBB3* (12 cases, 3.8%) (Figure S8.6, D-E). While *ERBB2* expression is markedly increased with amplification, expression increase of *ERBB3* is only modest.

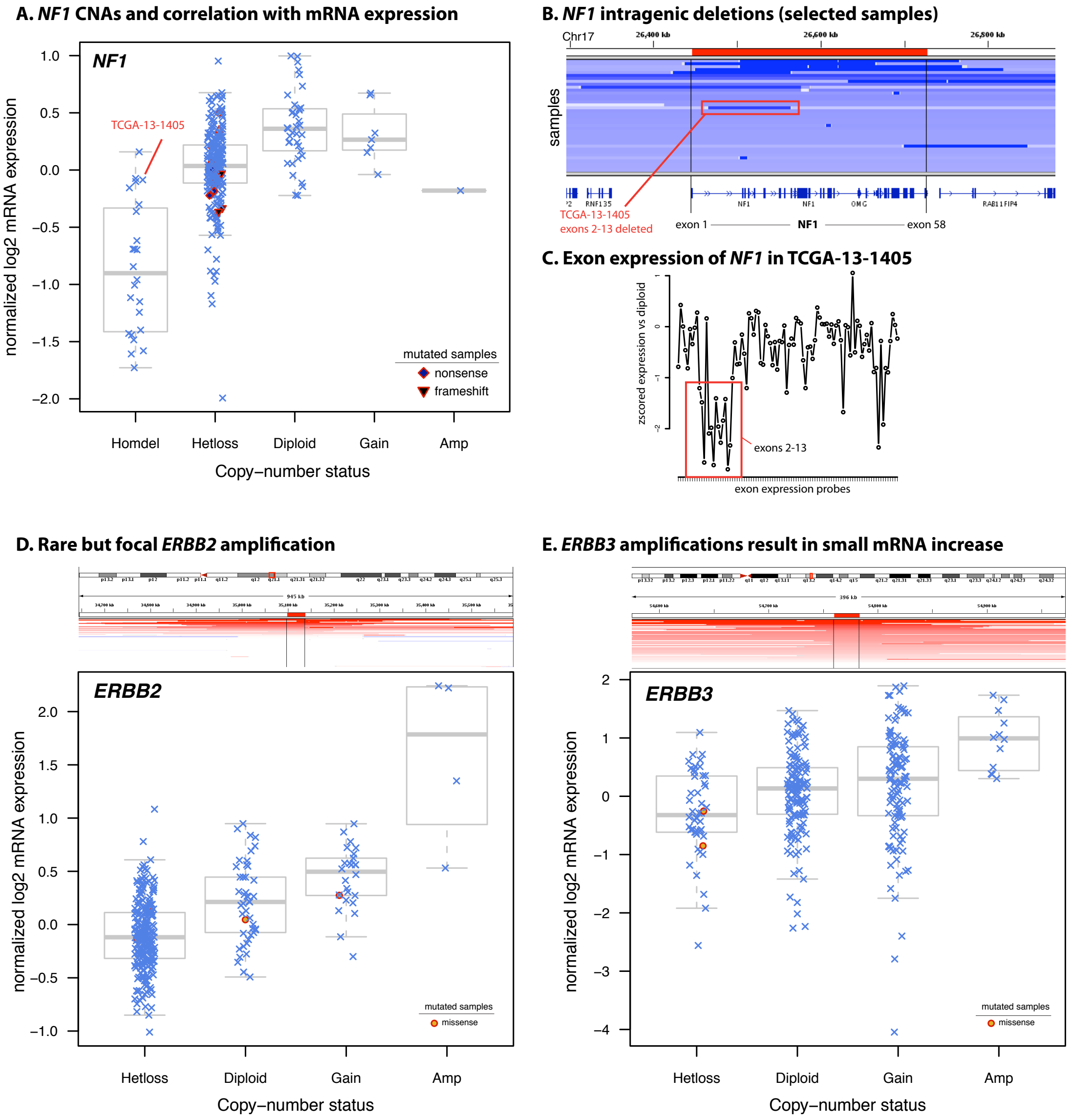


Figure S8.6: ***NF1* Deletion and *ERBB2*/*ERBB3* Amplification. A)** Correlation between *NF1* copy-number state and mRNA expression. Some samples with homozygous deletion of *NF1* do not have low mRNA expression, usually because they are only partially deleted, with possible full loss of function. **B)** Intragenic deletions of *NF1* are frequent, sometimes only affecting one exon. **C)** Sample TCGA-13-1405 has a deletion of exons 2-13 of *NF1*, and these exons show the lowest expression values across the gene. **D)** The few samples with focal, high-level amplification of *ERBB2* result in markedly increased mRNA expression. **E)** *ERBB3* expression is only modestly increased by gene amplification.

8.3 Homologous Recombination (HR)

Introduction

Approximately 10-15% of ovarian cancers appear to be hereditary, and the majority of these cases are due to germline mutations in *BRCA1* or *BRCA2*[1]. A subset of sporadic ovarian tumors appear to share distinctive DNA-repair defects with *BRCA1*/*BRCA2* germline mutation carriers, a phenomenon broadly described as "BRCAness"[2,3,4]. DNA-repair defects can be caused by germline or somatic alterations to the homologous recombination (HR) DNA repair pathway, including somatic mutation of *BRCA1*/*BRCA1* and epigenetic silencing of *BRCA1*, alterations to the core set of Fanconi Anemia genes, and additional genetic alterations to other key members of the HR pathway. For example, somatic mutations in *BRCA1* and *BRCA2* have previously been observed in sporadic ovarian cancer, but these events were considered relatively rare in ovarian cancer -- early studies have reported somatic mutation rates of 7-9% in *BRCA1* and 4% in *BRCA2*[5,6,7]. Additionally, *BRCA1* silencing via promoter hypermethylation has been reported in ovarian cancer[8,9], and recent studies have observed *BRCA1* hypermethylation in 18% of ovarian patients[10]. Other recent studies have identified EMSY amplification[11,12] and *FANCF* hypermethylation[13] as two additional means of inactivating the BRCA pathway in a broader spectrum of sporadic ovarian cancers.

Identifying ovarian cancer cases with defects in BRCA or the homologous recombination (HR) pathway is of increased clinical relevance due to the advent of new PARP inhibitors,[14,15] with potentially synthetic lethal effect when applied to cells with pre-existing defects in HR DNA repair. *In vitro* experiments have demonstrated that PARP inhibitors uniquely affect the survival of tumors cells with defects in HR, while leaving normal cells intact, and that *BRCA1* and *BRCA2* deficient cells are up to 1000 times more sensitive to the current set of PARP inhibitors[16,17]. Multiple PARP inhibitor drugs are currently in clinical trials in breast and ovarian cancer[14], and early Phase 1 and 2 trials in *BRCA1*/*BRCA2* mutation carriers appear promising[18,19]. High-throughput screening has also identified PARP sensitivity in cells deficient in other HR pathway members, including *RAD51*, *RAD54*, *DSS1*, *RPA1*, *NBS1*, *ATR*, *ATM*, *CHK1*, *CHK2*, *FANCD2*, *FANCA*, and *FANCC*[20]. *PTEN* deficiency has also been recently identified to cause homologous recombination defects in human tumor cells, and to sensitize tumor cells to PARP inhibitors[21]. Many investigators have therefore hypothesized that PARP inhibitors may be effective against a much larger group of tumors, beyond just BRCA1/BRCA2 mutation carriers[3,14,15].

A key challenge is to determine the extent of BRCA defects in sporadic ovarian cancers, develop biomarkers for these defects and for the response to, e.g., PARP inhibitor therapy, and apply this knowledge to identify patients likely to benefit from PARP inhibition therapy.

Analysis of alterations in HR DNA repair processes

For the analysis of the homologous recombination (HR) and BRCA pathways, four levels of analysis were performed:

- First, a detailed analysis of BRCA1/2 mutations and epigenetic silencing of *BRCA1*.

- Second, a detailed analysis of well-annotated genes known to be involved in the canonical HR pathway. This includes, for example, the set of Fanconi Anemia genes, *C11orf30* (EMSY), *RAD51*, the DNA damage sensing genes *ATM* and *ATR* and *PTEN*.

- Third, a global, but less detailed assessment of approximately 40 other HR-related genes. Additional genes were derived from an extended literature and pathway search, and Gene Ontology annotation.

- Fourth, to investigate potential cross-talk with other genes and pathways, we compared the complete set of BRCA inactivation events to all recurrently altered copy number peaks, as defined by GISTIC, looking for trends in mutual exclusivity and co-occurrence.

BRCA Alterations

BRCA Mutations

BRCA1 is mutated in 37 of 316 cases (11.7%): Twenty-seven (8.5%) cases have germline mutations and 10 (3.2%) have somatic mutations (Table S8.1, Figure S8.7A). Thirteen of the observed *BRCA1* germline mutations correspond to the well-known 'founder' mutations 185/187delAG and 5382/5385insC, both of which have been extensively studied in Ashkenazi Jewish populations[22,23,24,25]. *BRCA2* is mutated in 29 of 316 cases (9.2%): Twenty (6.3%) cases have germline mutation and 9 cases (2.9%) have somatic mutations (Table S8.1, Figure S8.7B). Five of the observed *BRCA2* germline mutations correspond to the well-known 6174delT founder mutation[24,26]. Thirty of the 37 (81%) *BRCA1* mutations are accompanied by heterozygous loss of *BRCA1*, indicating that both alleles are inactivated, as predicted by Knudson's two-hit hypothesis for a tumor suppressor gene (Figure S8.8A). Twenty-one of the 29 (72.4%) *BRCA2* mutations are accompanied by heterozygous loss (Figure S8.8B). Eighty-eight percent of germline *BRCA1* mutations matched to existing records in the Breast Cancer Information Core (BIC) Database (http://research.nhgri.nih.gov/projects/bic/), compared to 40% for somatic mutations; similarly, 58% of germline *BRCA2* mutations matched to existing BIC records, compared to 30% for somatic mutations.

In total, *BRCA1* or *BRCA2* are mutated in 64/316 cases (20.3%, Table S8.3). This corresponds to a germline mutation rate of 14.6% and a somatic mutation rate of 6.0%. The observed mutation rates are within range of previous reports. For example, a 2010 study involving 235 women with ovarian cancer found germline and somatic mutation rates of approximately 11.5% and 7% respectively[4], and a 2005 U.S. based survey involving a total of 232 women found *BRCA1/2* germline mutations in 13.8% of all cases, and 14.8% of serous cases[27].

With the exception of two cases, *BRCA1* and *BRCA2* mutations are mutually exclusive, but the mutual exclusivity is not statistically significant (N=316 P = 0.5518, two-sided Fisher's exact test).

Table S8.1: *BRCA1* Mutations

BRCA1 **Germline Mutations, (sorted by nucleotide position)**

Case ID	Mutation Type	Mutation	Chromosome Location	NT Position†	Note	# of Records in BIC Database††	Copy Number Status
TCGA-10-0931	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Heterozygous Loss
TCGA-13-1408	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Heterozygous Loss
TCGA-23-1027	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Diploid
TCGA-23-1118	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Heterozygous Loss
TCGA-23-2078	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Heterozygous Loss
TCGA-23-2079	Frame Shift Deletion	p.E23fs	17:38529571-38529572	187	185/187DelAG Founder Mutation [23,24].	1980	Diploid
TCGA-13-0887	Frame Shift Deletion	p.C24fs	17:38529570-38529571	188	185/187DelAG Founder Mutation [23,24].	1980	Heterozygous Loss
TCGA-13-1494	Split Site SNP	e3-1	17:38512077-38512077	N/A		N/A	Heterozygous Loss
TCGA-13-0893	Frame Shift Insertion	p.R504fs	17:38499565-38499566	1627	1627Ins ATAAATTAAA	0	Heterozygous Loss
TCGA-13-0903	Frame Shift Deletion	p.R504fs	17:38499564-38499564	1629	DelC	2	Heterozygous Loss
TCGA-61-2109	Frame Shift Deletion	p.K654fs	17:38499113-38499113	2080	DelA	31	Heterozygous Loss
TCGA-04-1356	Frame Shift Deletion	p.N723fs	17:38498908-38498908	2285	DelC	0	Heterozygous Loss
TCGA-59-2348	Nonsense Mutation	p.E797*	17:38498685-38498685	2508	2508 G to T (Glu to Stop)	3	Heterozygous Loss
TCGA-13-1512	Frame Shift Deletion	p.D825fs	17:38498599-38498599	2594	DelC	55	Heterozygous Loss
TCGA-09-1669	Frame Shift Deletion	p.E1346fs	17:38497039-38497039	4154	DelA	50	Heterozygous Loss
TCGA-25-2392	Frame Shift Deletion	p.E1346fs	17:38497039-38497039	4154	DelA	50	Diploid
TCGA-24-2298	Frame Shift Insertion	p.Q1395fs	17:38496488-38496489	4302	4302InsTC.	1	Diploid
TCGA-24-1470	Frame Shift Deletion	p.T1677fs	17:38473195-38473198	5146	DelTAAC	1	Heterozygous Loss
TCGA-57-1582	Frame Shift Deletion	p.R1726fs	17:38468889-38468892	5296	DelGAAA	39	Gain
TCGA-09-2051	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Heterozygous Loss
TCGA-13-0883	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Heterozygous Loss
TCGA-23-1122	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Amplification
TCGA-23-2077	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Heterozygous Loss
TCGA-23-2081	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Heterozygous Loss
TCGA-25-2401	Frame Shift Insertion	p.Q1756fs	17:38462605-38462606	5385	5382/5385 insC Founder Mutation [24,25].	1063	Heterozygous Loss
TCGA-09-2045	Frame Shift Deletion	p.Q1779fs	17:38454735-38454735	5454	DelC	5	Heterozygous Loss
TCGA-61-2008	Nonsense Mutation	p.W1815*	17:38453208-38453208	5564	5564 G to A	0	Heterozygous Loss

BRCA1 **Somatic Mutations, (sorted by nucleotide position)**

Case ID	Mutation Type	Mutation	Chromosome Location	NT Position†	Note††	# of Records in BIC Database††	Copy Number Status
TCGA-13-0804	Missense Mutation	p.C47W	17:38512070-38512070	260	260 C to G	0	Heterozygous Loss
TCGA-25-1625	Nonsense Mutation	p.E116*	17:38509760-38509760	465	465 G to T	0	Heterozygous Loss
TCGA-29-2427	Nonsense Mutation	p.L431*	17:38499782-38499782	1411	1411 T to G (Leu to Stop).	1	Heterozygous Loss
TCGA-25-1630	Frame Shift Deletion	p.A521fs	17:38499517-38499517	1676	1676DelG	0	Heterozygous Loss
TCGA-23-1026	Frame Shift Deletion	p.G813fs	17:38498636-38498636	2557	2557DelG	0	Heterozygous Loss
TCGA-25-1632	Frame Shift Insertion	p.S1216fs	17:38497425-38497426	3767	3767Ins AGAACTTA. Three 3767 InsA records recorded in BIC Database.	3	Heterozygous Loss
TCGA-13-1489	Frame Shift Insertion	p.N1265fs	17:38497279-38497280	3913	3913InsAA.	0	Heterozygous Loss
TCGA-04-1357	Nonsense Mutation	p.Q1538*	17:38479937-38479937	4731	4731 C to T	3	Diploid
TCGA-24-2035	Frame Shift Deletion	p.G1710fs	17:38469440-38469440	5248	5248DelG	0	Heterozygous Loss
TCGA-13-0730	Nonsense Mutation	p.R1835*	17:38451310-38451310	5622	5622 C to T (Arg to Stop)	63	Heterozygous Loss

† Nucleotide positions are reported in reference to *BRCA1* GenBank record U14680, as per The Breast Cancer Information Core Database (http://research.nhgri.nih.gov/projects/bic/).

†† Mutations were matched by nucleotide position and compared to existing mutation records in the Breast Cancer Information Core (BIC) Database (http://research.nhgri.nih.gov/projects/bic/) on August 30, 2010.

Table S8.2: *BRCA2* Mutations

BRCA2 **Germline Mutations, (sorted by nucleotide position)**

Case ID	Mutation Type	Mutation	Chromosome Location	NT Position†	Note††	# of Records in BIC Database††	Copy Number Status
TCGA-24-0975	Splice Site SNP	e6+2	13:31798752-31798752	N/A		N/A	Heterozygous Loss
TCGA-24-2288	Frame Shift Deletion	p.V220fs	13:31801605-31801606	885	del TG	0	Heterozygous Loss
TCGA-13-0900	Frame Shift Deletion	p.N257fs	13:31803141-31803145	995	delCAAAT	1	Heterozygous Loss
TCGA-04-1367	Nonsense Mutation	p.E294*	13:31804495-31804495	1108	1108 G to T	0	Heterozygous Loss
TCGA-25-2404	Frame Shift Deletion	p.K343fs	13:31804640-31804640	1253	1253 DelA	0	Heterozygous Loss
TCGA-24-1463	Frame Shift Insertion	p.I605fs	13:31805420-31805421	2033	2033 InsA	0	Diploid
TCGA-24-1417	Frame Shift Deletion	p.N1706fs	13:31811604-31811607	5340	delAATA	0	Heterozygous Loss
TCGA-24-2024	Frame Shift Deletion	p.Y1710fs	13:31811620-31811623	5356	delTATG	0	Heterozygous Loss
TCGA-04-1336	Frame Shift Deletion	p.T1738fs	13:31811703-31811706	5439	delTACT	0	Heterozygous Loss
TCGA-13-0913	Frame Shift Deletion	p.E1857fs	13:31812061-31812065	5797	delGAAAC	0	Heterozygous Loss
TCGA-13-0886	Frame Shift Deletion	p.S1982fs	13:31812438-31812438	6174	6174delT Founder Mutation [24,26].	1087	Heterozygous Loss

TCGA-13-1498	Frame Shift Deletion	p.S1982fs	13:31812438-31812438	6174	6174delT Founder Mutation [24,26].	1087	Diploid
TCGA-13-1499	Frame Shift Deletion	p.S1982fs	13:31812438-31812438	6174	6174delT Founder Mutation [24,26].	1087	Heterozygous Loss
TCGA-24-2280	Frame Shift Deletion	p.S1982fs	13:31812438-31812438	6174	6174delT Founder Mutation [24,26].	1087	Heterozygous Loss
TCGA-59-2351	Frame Shift Deletion	p.S1982fs	13:31812438-31812438	6174	6174delT Founder Mutation [24,26].	1087	Heterozygous Loss
TCGA-13-0726	Nonsense Mutation	p.R2394*	13:31827170-31827170	7408	7408 A to T	5	Heterozygous Loss
TCGA-24-2293	Nonsense Mutation	p.R2520*	13:31828687-31828687	7786	7786 C to C	44	Diploid
TCGA-24-1562	Nonsense Mutation	p.K3326*	13:31870626-31870626	10204	10204 A to T	293	Diploid
TCGA-13-1512	Nonsense Mutation	p.K3326*	13:31870626-31870626	10204	10204 A to T.	293	Diploid
TCGA-23-1026	Nonsense Mutation	p.K3326*	13:31870626-31870626	10204	10204 A to T	293	Diploid

***BRCA2* Somatic Mutations, (sorted by nucleotide position)**

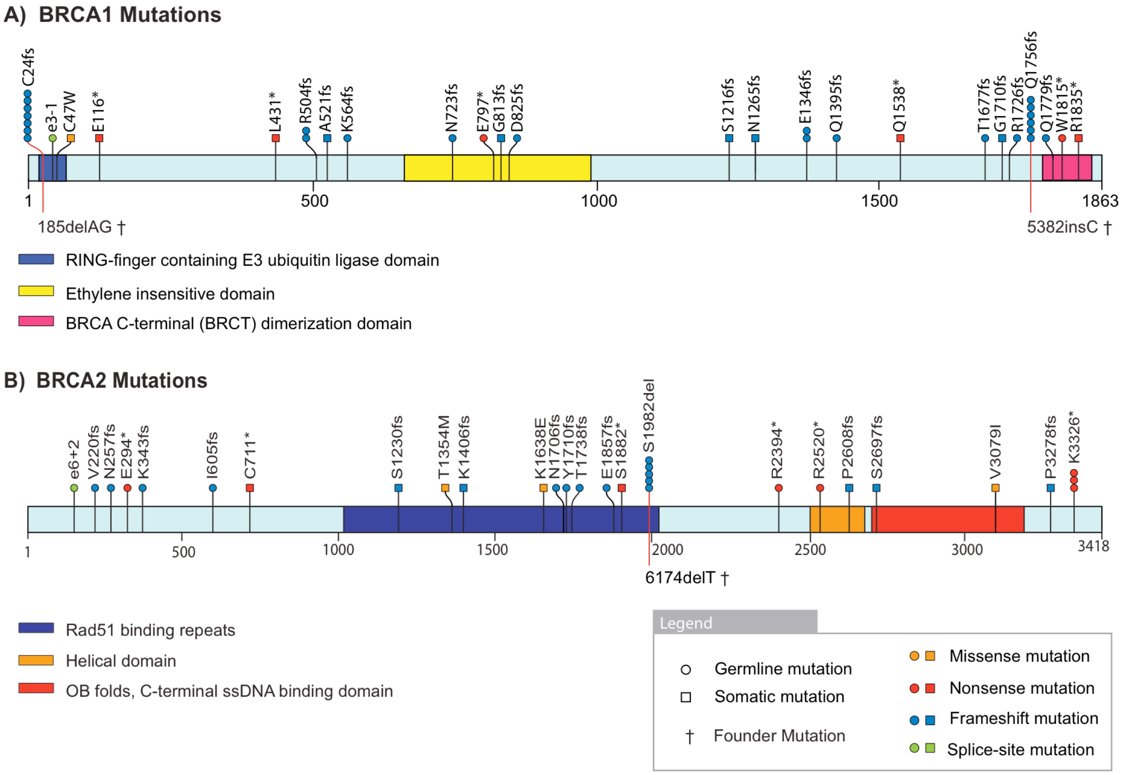
Case ID	Mutation Type	Mutation	Chromosome Location	Nucleotide Position†	Note	# of Records in BIC Database††	Copy Number Status
TCGA-04-1331	Nonsense Mutation	p.C711*	13:31808625-31808625	NT Position: 2361	2361 C to A	0	Heterozygous Loss
TCGA-13-0890	Frame Shift Deletion	p.S1230fs	13:31810178-31810178	NT Position: 3914	3914DelT	0	Heterozygous Loss
TCGA-23-1030	Missense Mutation	p.T1354M	13:31810553-31810553	NT Position: 4289	4289 C to T	11	Diploid
TCGA-13-0885	Frame Shift Deletion	p.K1406fs	13:31810708-31810711	NT Position: 4444	delAAAG	0	Heterozygous Loss
(2 mutations)	Frame Shift Deletion	p.E1407fs	13:31810710-31810713	NT Position: 4446	delAGAA	1	Heterozygous Loss
TCGA-24-1103	Missense Mutation	p.K1638E	13:31811404-31811404	NT Position: 5140	5140 A to G	0	Heterozygous Loss
TCGA-09-2050	Nonsense Mutation	p.S1882*	13:31812137-31812137	NT Position: 5873	5873 C to A.	28	Heterozygous Loss
TCGA-24-1555	Frame Shift Deletion	p.P2608fs	13:31834675-31834675	NT Position: 8049	8049DelT.	0	Heterozygous Loss
TCGA-13-1481	Frame Shift Deletion	p.S2697fs	13:31835426-31835441	NT Position: 8316	8315DelTGAGCGCAAATATATC.	0	Diploid
TCGA-23-1120	Frame Shift Deletion	p.P3278fs	13:31870481-31870481	NT Position: 10059	10059DelG.	0	Heterozygous Loss

† Nucleotide positions are reported in reference to *BRCA2* GenBank record U43746, as per The Breast Cancer Information Core Database (http://research.nhgri.nih.gov/projects/bic/).

†† Mutations were matched by nucleotide position and compared to existing mutation records in the Breast Cancer Information Core (BIC) Database (http://research.nhgri.nih.gov/projects/bic/) on August 30, 2010.

Table S8.3: BRCA Mutation Rates

Gene	Germline Mutation Rate	Somatic Mutation Rate	Total Mutation Rate
BRCA1	8.54%	3.16%	11.71%
BRCA2	6.33%	2.85%	9.18%
Both Genes	14.56%	6.01%	20.25%


A) BRCA1 Mutations

B) BRCA2 Mutations

Figure S8.7: Summary of BRCA Mutations. All BRCA1/2 germline and somatic mutations are displayed along the protein domain structure. A) BRCA1 Mutations. Thirteen cases, all germline, have well-known BRCA1 founder mutations at 185/187delAG and 5382/5385insC. B) BRCA2 mutations. Five cases, all germline, have known BRCA2 founder mutations at 6174delT.

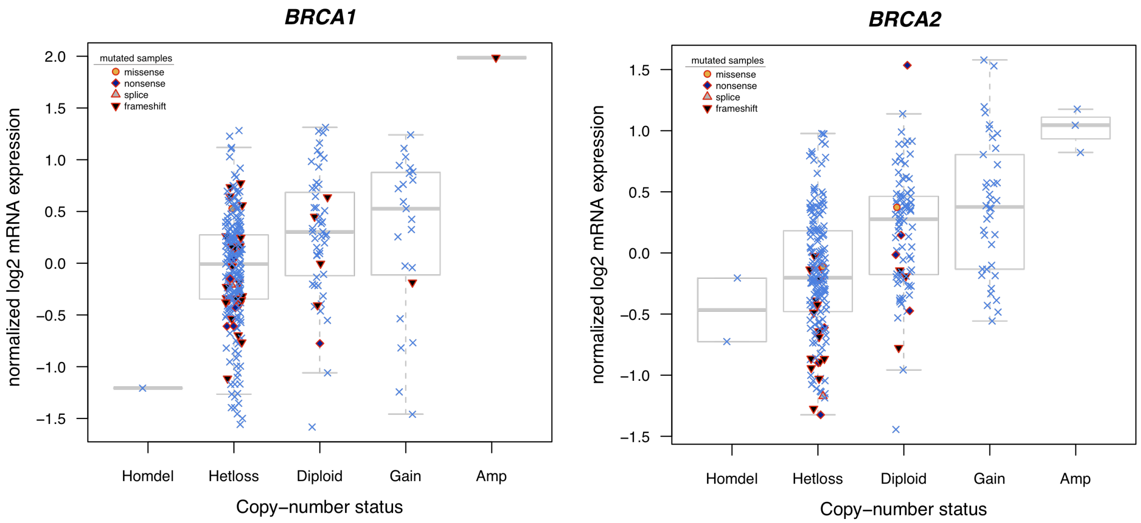


Figure S8.8: Heterozygous loss associated with BRCA1/2. A) Thirty of the 37 (81%) of BRCA1 mutations are accompanied by heterozygous loss; **B)** Twenty-one of the 29 (72.4%) of the BRCA2 mutations are accompanied by heterozygous loss.

Epigenetic Silencing of BRCA1

BRCA1 silencing via promoter hypermethylation has been reported previously in ovarian and breast cancer[8,9], and recent studies have reported *BRCA1* hypermethylation in 18% of ovarian patients[10].

As described in Supplemental Methods 7, we analyzed the relationship between DNA methylation and gene expression for nine different probes located in or near the *BRCA1* promoter region, and found statistically significant inverse correlations for four of the nine probes (cg19531713, cg19088651, cg08993267, cg04658354). The target CpG sites of those probes are located in the CpG island that contains the transcription start site of *BRCA1*. For each of the aforementioned four probes, we used *k*-means clustering on the two-dimensional space of DNA methylation and expression data to separate the epigenetically silenced group and the non-epigenetically silenced group of samples. Expression data were scaled to have the same range as DNA methylation data for the purpose of clustering. We then combined the calls from the four probes. Since data was lacking for some probes in some samples, we relied on the fraction of the four probes calling a particular sample in the hypermethylated group, rather than on a fixed number of probes. Samples with >50% consensus on belonging to the hypermethylated group across the four probes were classified as samples with silencing of *BRCA1* by promoter hypermethylation.

Using this method, we identified 34 of 316 cases (10.8%) with epigenetic silencing of *BRCA1*. Notably, epigenetic silencing of *BRCA1* is mutually exclusive of *BRCA1/2* mutations *(P* = 4.45 e-04, two-sided Fisher's exact test). This mutual exclusivity provides evidence of strong selective pressure to inactivate BRCA via either mutation or epigenetic silencing.

Analysis of the Core HR Pathway

Amplification of EMSY

Previous studies have identified amplification and overexpression of EMSY (*C11orf30*) as an alternative means by which tumors selectively inactivate the BRCA pathway. EMSY was discovered in a yeast two-hybrid screen with BRCA2, and the EMSY protein binds specifically to the transactivation domain in BRCA2[12]. An excess of EMSY can result in an inhibition of BRCA2 transcriptional activity, and overexpression of EMSY may eliminate selective pressure in sporadic breast and ovarian cancer to inactivate BRCA2[28]. The EMSY protein is also known to be co-located with *BRCA2* at chromosomal sites of DNA damage and to interact with proteins involved in the regulation of chromatin[29].

Previous studies have identified amplification of EMSY in 13% of sporadic primary breast cancer and 17% of high-grade sporadic ovarian cancer[2,11]. Ovarian tumors with EMSY amplification have been associated with significantly worse outcome[30]. However, in a multivariate analysis that included histological subtype, grade, stage, age and EMSY amplification as the covariates, only stage and age were significant prognostic predictors[30]. EMSY is located at 11q13, a region known to be amplified in multiple cancers, including breast, ovarian, head and neck, lung, and bladder cancer[12]. The amplicon is gene dense, and the region likely contains a cassette of genes rather than a single oncogene -- for example, in ovarian cancer, the amplicon tends to include several genes including EMSY, *LRRC32* (GARP), and *PAK1*[12].

For the unified case list (N=316), we identified 19 cases with EMSY amplification (GISTIC) and 6 cases with EMSY mutation (Figure S8.9). By this analysis, there is evidence for EMSY alteration in 7.9% of cases. However, we do not observe co-occurrence or mutual exclusivity between BRCA inactivation events (mutations plus methylation) and EMSY amplification and mutation ($P = 0.8248$, two-sided Fisher's Exact Test).

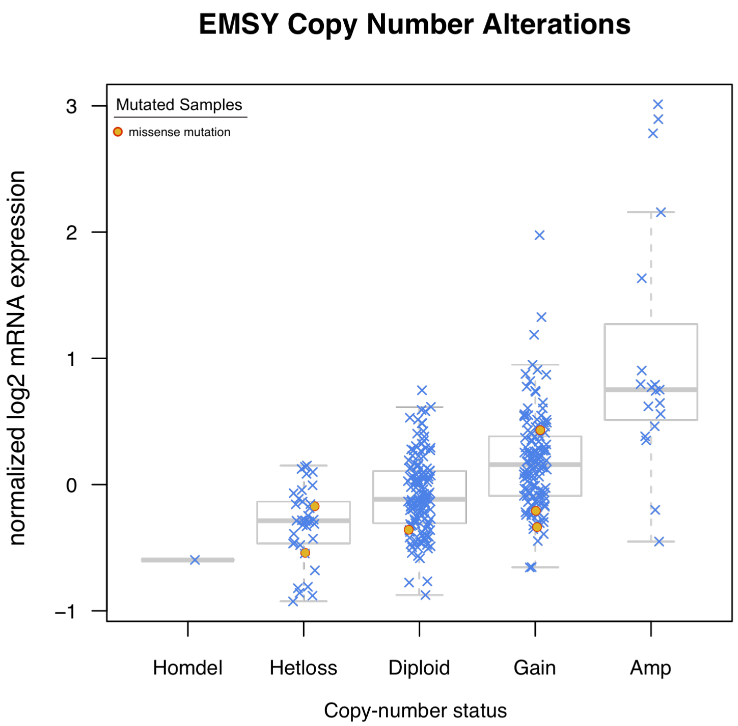


Figure S8.9: EMSY/C11orf30 Copy Number Alterations. Normalized log2 mRNA expression v. GISTIC copy number status for EMSY.

Absence of FANCF Hypermethylation

A number of recent studies have identified hypermethylation of FANCF as an alternative means of altering the BRCA pathway in sporadic cancers, including ovarian cancer [2]. For example, a 2008 study observed hypermethylation of FANCF in 13.2% of 53 ovarian tumors samples[13]. However, in the TCGA data, we observe no clear evidence of FANCF silencing by hypermethylation (Figure S8.10).

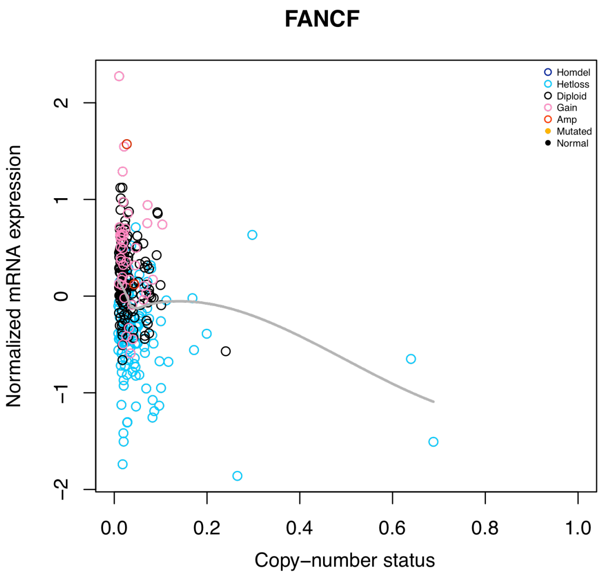


Figure S8.10: DNA methylation beta values v. normalized log2 mRNA expression levels for FANCF. We observe no clear evidence of hypermethylation of FANCF.

Homozygous Deletions of PTEN

PTEN deficiency has been identified to cause homologous recombination defects in human tumor cells, and to sensitize tumor cells to PARP inhibitors[21]. However, the exact role of PTEN in homologous recombination and DNA repair remains controversial and an area of active research[31]. DNA copy-number analysis identifies a focal deletion region at 10q23.31 (q-value: 5.41E-11), which includes only *PTEN*. This corresponds to 21 cases (6.7%) of *PTEN* homozygous deletion, each of which is associated with down-regulation at the mRNA level (Figure S8.11). We also observe two somatic mutations in *PTEN*. However, we do not observe co-occurrence or mutual exclusivity between BRCA inactivation events (mutations plus methylation) and PTEN homozygous deletion and mutation (*P* = 0.3607, two-sided Fisher's Exact Test).

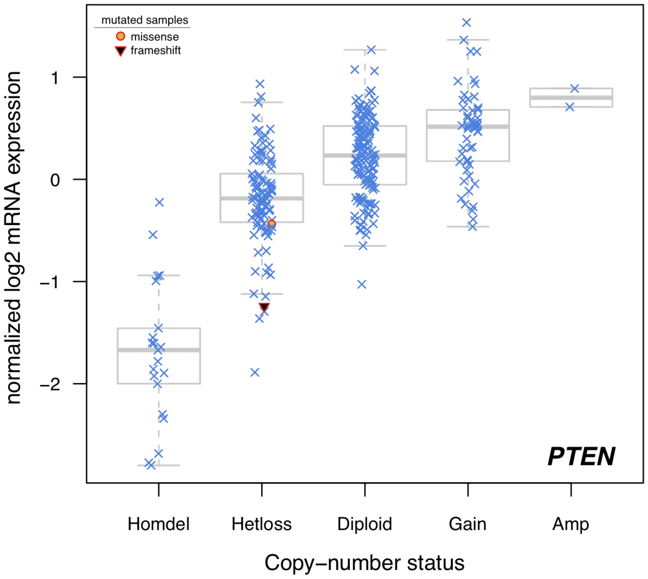


Figure S8.11: PTEN Copy Number Alterations. PTEN is homozygously deleted in 21 cases (6.65%), and homozygous deletions are associated with down-regulation at the mRNA level. N=316 cases.

Fanconi Anemia and Other Core HR Genes

Table S8.4 provides mutation and copy number alteration rates for other well-annotated genes known to be involved in homologous recombination (HR), derived from literature curation[32,33,34,35]. A fingerprint of the complete set of HR genes is provided in Figure S8.12. Due to the low mutation rates observed in the Fanconi Anemia genes, we do not observe co-occurrence or mutual exclusivity between BRCA inactivation events (mutations plus methylation) and Fanconi Anemia mutations (*P* = 0.7834, two-sided Fisher's Exact Test).

Table S8.4: Analysis of other Core Members of the HR Pathway

Fanconi Anemia Genes, Total Mutation Rate: 5.06%

Gene Symbol	Entrez Gene ID	*In Vitro* Sensitivity to PARPi*	Number of Samples Mutated (N=316)	% of Samples Mutated (N=316)	Copy Number Alterations†
C19orf40	91442		0	0.00%	7.91%
FANCA	2175	Yes	3	0.95%	2.85%
FANCB	2187		0	0.00%	0.00%
FANCC	2176	Yes	2	0.63%	1.58%
FANCD2	2177	Yes	1	0.32%	0.95%
FANCE	2178		1	0.32%	2.53%
FANCF	2188		0	0.00%	0.63%
FANCG	2189		1	0.32%	0.00%
FANCI	55215		2	0.63%	1.58%
FANCL	55120		2	0.63%	1.58%
FANCM	57697		1	0.32%	0.95%
PALB2	79728		4	1.27%	0.63%

Core HR RAD Genes, Total Mutation Rate: 1.58%

Gene Symbol	Entrez Gene ID	*In Vitro* Sensitivity to PARPi*	Number of Samples Mutated (N=316)	% of Samples Mutated (N=316)	Copy Number Alterations†
RAD50	10111		2	0.63%	1.27%
RAD51	5888	Yes	1	0.32%	1.27%
RAD51C	5889		0	0.00%	0.63%
RAD51L1	5890		0	0.00%	2.22%
RAD51L3	5892		0	0.00%	0.95%
RAD52	5893		0	0.00%	7.28%
RAD54B	25788		0	0.00%	4.11%
RAD54L	8438		2	0.63%	5.38%

DNA damage response genes involved in HR, Total Mutation Rate: 2.22%

Gene Symbol	Entrez Gene ID	*In Vitro* Sensitivity to PARPi*	Number of Samples Mutated (N=316)	% of Samples Mutated (N=316)	Copy Number Alterations†
ATM	472	Yes	4	1.27%	1.27%
ATR	545	Yes	2	0.63%	3.80%
CHEK1	1111	Yes	0	0.00%	3.48%
CHEK2	11200	Yes	1	0.32%	1.90%

**In Vitro* Sensitivity to PARPi based on: [20].

† Copy number rates include amplifications and homozygous deletions as determined by GISTIC copy-number analysis.

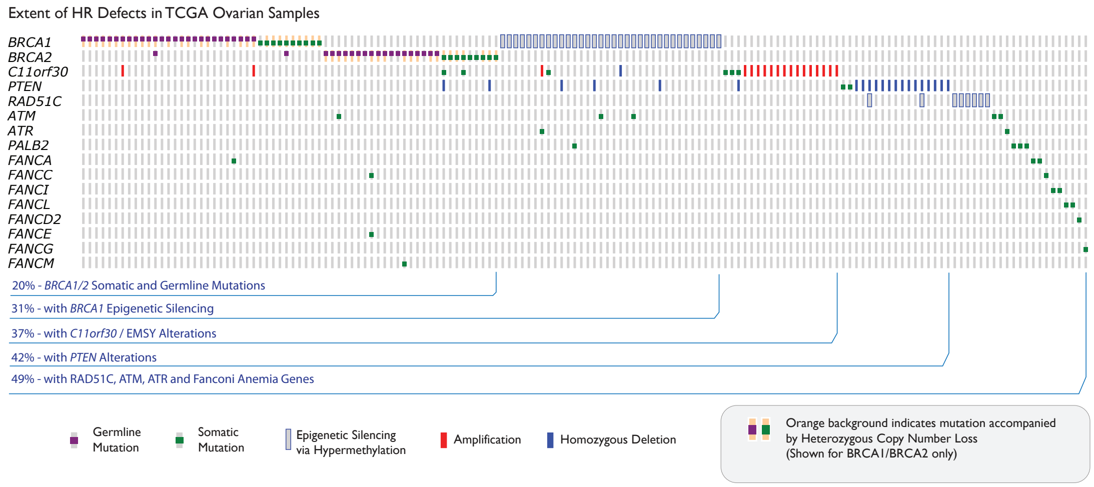


Figure S8.12: Genomic Fingerprint of HR Pathway Alterations. Each column represents an individual case; each row represents a gene. Only cases with HR defects (N=154) are shown. Copy number alterations are only shown for EMSY and *PTEN*. While it is not yet clear if all of these HR defects result in a sufficient decrease in homologous recombination to result in sensitization to PARP inhibitors, our findings indicate that HR defects occur in a substantial fraction of sporadic ovarian tumors. We therefore suggest comprehensive profiling of these molecular alterations in ongoing and future clinical trials of PARP inhibitors.

Extended HR Analysis

To extend the analysis beyond well-annotated genes involved in HR, a more global, but less detailed analysis was performed on 42 other potentially relevant genes. These additional genes were derived from an extended literature and pathway search, and Gene Ontology annotation. More specifically, the list was derived from the ATM/BRCA pathway from BioCarta, the Homologous Recombination Repair pathway from Reactome[36,37], and Gene Ontology GO:0000724: double-strand break repair via homologous recombination. The complete list of genes analyzed, along with mutation rates and GISTIC copy number analysis is provided in Table S8.5. Within the larger gene set, we observe only very low mutations rates. For example, the Bloom syndrome gene (*BLM*) participates in genome maintenance, is essential for BRCA1 function[38] and is mutated in four cases. Additionally, several genes including *BCL2L1*, *OBFC2B* and *RBBP8* appear within relatively narrow recurrent regions of amplification, as defined by GISTIC copy number analysis.

Table S8.5: Analysis of Other Potential HR Genes

Gene Symbol	Entrez Gene ID	# of Samples Mutated (N=316)	% of Samples Mutated (N=316)	Within GISTIC Peak (Amp/Del; Total Number of Genes within Peak appear in brackets)
BBC3	27113	0	0.00%	Deletion (323)
BCL2	596	0	0.00%	
BCL2L1	598	0	0.00%	Amplification (2)
BLM	641	4	1.27%	Amplification (62)
BTBD12	84464	2	0.63%	
DMC1	11144	0	0.00%	
EME1	146956	0	0.00%	
EME2	197342	0	0.00%	
ERCC4	2072	1	0.32%	
GEN1	348654	2	0.63%	
GIYD1	548593	0	0.00%	
H2AFX	3014	0	0.00%	Deletion (269)
HUS1	3364	1	0.32%	
LIG1	3978	0	0.00%	Deletion (323)
MDC1	9656	2	0.63%	
MDM2	4193	0	0.00%	
MRE11A	4361	0	0.00%	
MUS81	80198	1	0.32%	
NBN	4683	0	0.00%	
OBFC2A	64859	0	0.00%	
OBFC2B	79035	1	0.32%	Amplification (18)
PCNA	5111	0	0.00%	
PMAIP1	5366	0	0.00%	
POLD1	5424	1	0.32%	
POLD2	5425	2	0.63%	
POLD3	10714	1	0.32%	
POLD4	57804	0	0.00%	
RAD1	5810	1	0.32%	Amplification (80)
RAD17	5884	0	0.00%	Deletion (51)
RAD9A	5883	0	0.00%	
RBBP8	5932	1	0.32%	Amplification (11)
RPA1	6117	2	0.63%	
RPA2	6118	1	0.32%	Deletion (188)
RPA3	6119	0	0.00%	Deletion (84)
RTEL1	51750	0	0.00%	Amplification (39)
SHFM1	7979	0	0.00%	
TEX15	56154	4	1.27%	
TP53BP1	7158	4	1.27%	
TREX1	11277	0	0.00%	
UBE2N	7334	0	0.00%	Deletion (375)
XRCC2	7516	0	0.00%	Amplification (92)
XRCC3	7517	1	0.32%	

Survival Analysis of Cases with HR Defects

Previous studies have observed better outcome in BRCA-positive patients, including longer tumor-free intervals between relapses, and improved overall survival[39]. Previous studies have also observed shorter overall survival for patients with *BRCA1* hypermethylation[9].

In the TCGA ovarian data, we observe mutual exclusivity between *BRCA1* epigenetic silencing and *BRCA1/2* mutations (see above), and we therefore focused our survival analysis on comparing three patients groups: *BRCA1* epigenetically silenced, *BRCA1/2* mutated, and BRCA Wildtype (WT). Within the complete data set (N=316), we observe differences in age between the three groups (P = 0.01576, Kruskal Wallis Test). Post-hoc pairwise comparisons show differences between BRCA mutated and BRCA WT (57.74 years versus 61.84 years, Bonferroni adjusted P = 0.061, Wilcoxon signed-rank test). Univariate survival analysis of BRCA status shows divergent outcome for the two types of events, with BRCA mutated cases exhibiting better overall survival (OS) than BRCA wild-type (median OS 66.5 *versus* 41.9 months, P = 3.08 e-04, log-rank test, Figure S8.13), and *BRCA1* epigenetically silenced cases exhibiting similar survival to BRCA1/2 WT (median OS 41.5 *versus* 41.9 months, P = 0.69, log-rank test, Figure S8.13). In a multivariate survival analysis of BRCA mutated versus BRCA WT cases, mutation status and age were significant prognostic predictors (BRCA mutation status, P = 0.00375, Age, P = 0.02742). We therefore observe evidence of selective pressure to alter BRCA genes via distinct genetic mechanisms, but statistically significant differences in outcomes for patients. Sequencing additional samples will allow further exploration in the distinct outcome patterns seen in *BRCA1 versus BRCA2* and germline *versus* somatic events.

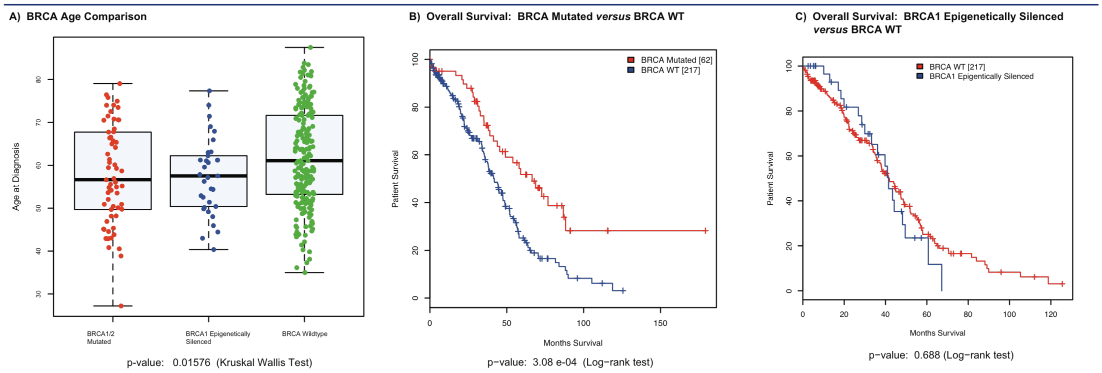


Figure S8.13: BRCA survival analysis. A) BRCA age comparison for the three BRCA categories analyzed. **B)** Kaplan-Meier curve comparing the survival of patients with BRCA mutation *versus* BRCA wild-type (WT). **C)** Kaplan-Meier curve comparing the survival of patients with BRCA1 epigenetic silencing *versus* BRCA wild-type (WT).

Effect of BRCA inactivation on genome stability

We investigated the effect of *BRCA1*/*BRCA2* mutations and *BRCA1* silencing on the overall level of DNA copy-number alterations. We computed the fraction of the genome that is not diploid for each case, and found that BRCA-altered cases to not exhibit increased levels of copy-number alterations (Figure S8.14). The result is similar when using the number of breakpoints in the DNA copy-number profiles (data not shown).

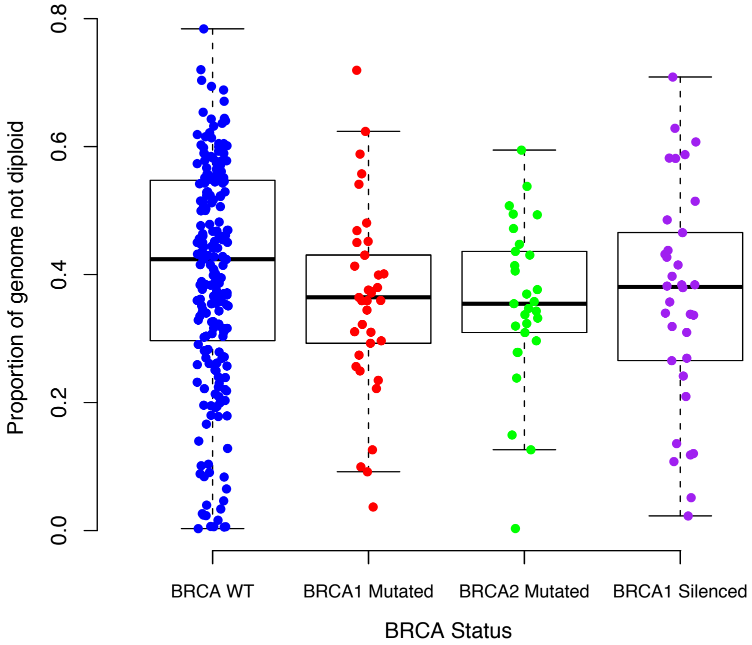


Figure S8.14: Cases with BRCA-alterations do not exhibit increased genomic instability.

Correlation of BRCA inactivation with recurrently altered copy number peaks in other genomic regions

To investigate potential cross-talk with other genes and pathways, we looked for potential correlations of BRCA inactivation events (mutation plus methylation, 98 samples, see above) with significantly altered copy number events as reported by GISTIC analysis (63 peaks of amplification and 50 peaks of deletion).

For each GISTIC peak, we defined the set of samples that is affected by the DNA copy-number alteration. We only considered samples as altered if at least half of the genes in the region are affected by homozygous deletion or high-level amplification. Each peak-associated set of samples was then tested for enrichment and depletion in BRCA inactivation by a two-tailed Fisher's exact test. Significant correlations were selected after Benjamini-Hochberg correction for false discovery (FDR < 5%) (Table S8.6).

We found a significant enrichment of BRCA inactivation for *MYC* amplified cases (49.0% of BRCA altered cases have *MYC* amplification *versus* 24.3% of BRCA wild type cases, FDR-adjusted *P* = 0.002, Table S8.6). *CCNE1* amplified cases show significant depletion of BRCA alteration (8.2% of BRCA altered cases have *CCNE1* amplification *versus* 25.7% of BRCA wild type cases, FDR adjusted *P* = 0.009). Unlike CCNE1, cases with alterations in *RB1* and *CDKN2A* (the other two main genes in the RB pathway, see 8.2. Cancer Pathways above), had overlap with BRCA alterations (*P* = 0.18 and *P* = 0.6, respectively, two-sided Fisher's Exact Test).

The observed tendency towards mutual exclusivity between BRCA inactivation and CCNE1 amplification prompted us to reevaluate the previously reported poor survival associated with *CCNE1*-amplification[40,41]. In evaluating the full case set, we observe worse outcome for CCNE1 amplified cases, in line with previous studies (P = 0.0718, Log Rank Test, Figure S8.15A). However, if we remove all BRCA inactivated cases, and examine survival differences in CCNE1 amplified cases within BRCA WT cases only, significant worse outcome is no longer detectable (*P* = 0.24, log-rank test, Figure S8.15B), suggesting that the previously reported survival difference can be explained by the better survival of BRCA-mutated cases.

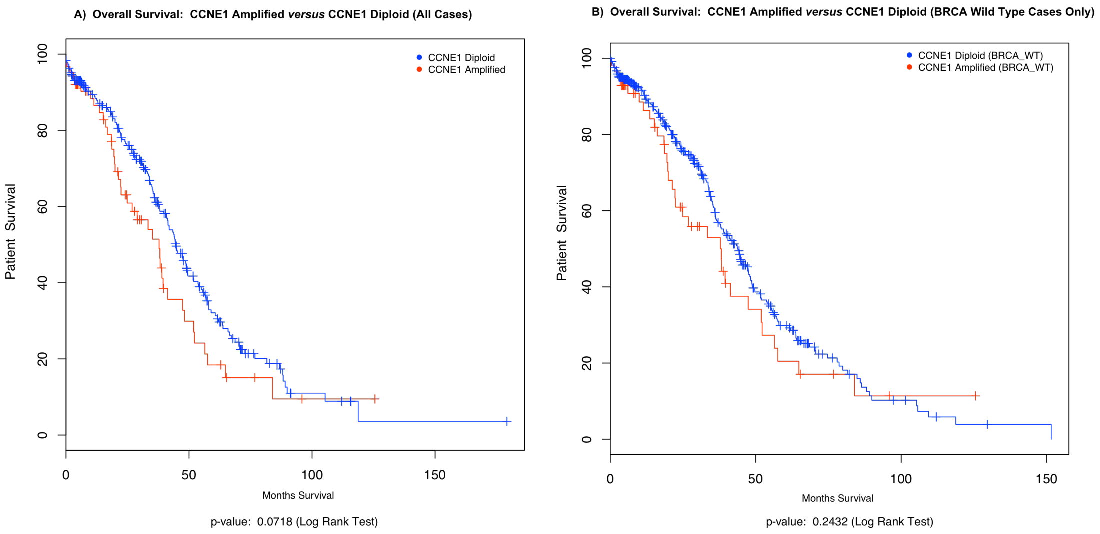


Figure S8.15: Overall Survival for CCNE1 amplified cases. Survival of CCNE1 amplified cases is compared to CCNE1 wild type cases: Among all cases **(A)**, and among BRCA wild type cases only **(B)**.

Table S8.6: Correlation between BRCA alterations and DNA copy-number events: Each peak is identified by its corresponding cytoband, and the regions are marked as either amplified (AMP) or deleted (DEL). The number of co-occurring cases with BRCA altered and BRCA wild type cases are in columns "BRCA Altered" and "BRCA WT" respectively. Fisher's p-values are reported (only regions with p<0.05 are in the table) with the corresponding FDR-corrected values. The red box highlights regions with significant enrichment/depletion after FDR correction.

GISTIC Region	Alteration	BRCA Altered	BRCA WT	Fisher's exact test	FDR	Relation	Genes in the regions
8q24.21	AMP	48	53	2.51E-05	0.00203	Co-oc.	MYC PVT1
19q12	AMP	8	56	0.00023	0.00963	Mut.Ex	CCNE1
8q24.3	AMP	37	42	0.00069	0.01877	Co-oc.	ZNF7 ZNF623 SHARPIN VPS28 PUF60 COMMD5 HSF1 GRINA DGAT1 GPAA1 EXOSC4 PYCRL CYC1 FAM83H GPR172A TSTA3 LRRC14 ADCK5 ZNF34 BOP1 ZC3H3 RPL8 PPP1R16A ZNF251 EEF1D CPSF1 MAF1 TIGD5 KIAA1688 ZNF707 PLEC1 NRBP2 ZNF696 FBXL6 SCRIB SLC39A4 MFSD3 OPLAH TOP1MT KIFC2 RECQL4 NFKBIL2 NAPRT1 RHPN1 C8ORFK29 ZFP41 MAPK15 PARP10 KIAA1875 GPT MGC70857 GLI4 ZNF517 SCXB FOXH1 SPATC1 MAFA SCRT1 LY6H CYHR1 C8orf30A C8orf51 GSDMD EPPK1 BREA2 C8orf31 GPIHBP1 LRRC24 C8orf73 MIR661 HEATR7A MIR937 MIR939 SCXA LOC100130274
19p13.13	AMP	3	32	0.00163	0.03290	Mut.Ex	CCDC130 TRMT1 STX10 CC2D1A PRKACA ZSWIM4 IER2 ASF1B NFIX RFX1 IL27RA CACNA1A NANOS3 RLN3 PODNL1 LYL1 C19orf53 C19orf57 MRI1 SAMD1 DCAF15 NACC1 LOC113230 PALM3 MIR181C MIR23A MIR24-2 MIR27A MIR181D
1q21.2	AMP	1	21	0.00349	0.05653	Mut.Ex	SETDB1 ARNT TARS2 VPS72 GOLPH3L PRUNE PIP5K1A LYSMD1 ENSA SCNM1 LASS2 CDC42SE1 MCL1 FAM63A SEMA6C HORMAD1 BNIPL MLLT11 TMOD4 ANXA9 CTSS ADAMTSL4 GABPB2 TNFAIP8L2 CTSK ECM1 RPRD2 C1orf56
19p12	AMP	1	17	0.01623	0.219	Mut.Ex	ZNF431 ZNF430 ZNF100 ZNF429 ZNF708 ZNF85 ZNF714 ZNF43 ZNF493 ZNF738 LOC641367
19p13.2	AMP	3	24	0.01731	0.20036	Mut.Ex	KEAP1 TYK2 EIF3G MRPL4 CDC37 KRI1 FDX1L QTRT1 DNM2 PPAN ATG4D ILF3 DNMT1 SLC44A2 AP1M2 ICAM3 CDKN2D RAVER1 PDE4A ICAM5 ICAM1 ICAM4 P2RY11 ANGPTL6 RDH8 COL5A3 S1PR2 S1PR5 C19orf66 LOC147727 C3P1 SNORD105 PPAN-P2RY11 MIR638 SNORD105B ZGLP1
4q13.3	AMP	0	11	0.02022	0.20474	Mut.Ex	COX18 ANKRD17 MTHFD2L BTC AREG ADAMTS3 RASSF6 EREG IL8 CXCL2 CXCL3 AFP AFM CXCL5 NPFFR2 ALB EPGN CXCL1 PF4 PF4V1 SLC4A4 GC CXCL6 PPBPL2 PPBP PPBPL1
18q11.2	AMP	0	12	0.02096	0.18864	Mut.Ex	TAF4B KCTD1
18q12.1	AMP	0	12	0.02096	0.18864	Mut.Ex	KIAA1012 RNF138 DSG2 FAM59A B4GALT6 DSC2 RNF125 DSG1 MEP1B DSC3 DSC1 DSG4 DSG3 TTR MCART2
3q29	AMP	20	23	0.02162	0.15922	Co-oc.	NCBP2 LSG1 WDR53 PAK2 OPA1 DLG1 LRCH3 RNF168 PPP1R2 FYTTD1 KIAA0226 LOC152217 SENP5 PCYT1A RPL35A PIGX ATP13A3 LMLN SDHALP2 BDH1 TMEM44 HRASLS TNK2 IQCG MUC20 TFRC PIGZ FAM43A MFI2 FGF12 MUC4 LRRC33 HES1 APOD ATP13A5 ZDHHC19 LRRC15 TM4SF19 LOC348840 ATP13A4 GP5 CPN2 ACAP2 UBXN7 MGC2889 C3orf34 C3orf59 C3orf21 OSTalpha FBXO45 LOC220729 TCTEX1D2 C3orf43 SDHALP1 MIR570 FAM157A MIR922 LOC100128023 LOC100131551
14q11.2	AMP	1	15	0.02692	0.18169	Mut.Ex	METT11D1 ZNF219 NDRG2 FLJ10357 SLC39A2 TPPP2 RNASE13 RNASE7 RNASE8 RNASE3 RNASE2 C14orf176
6q27	DEL	4	1	0.03363	0.20953	Co-oc.	FAM120B TBP PDCD2 PSMB1 FGFR1OP PHF10 SFT2D1 MLLT4 WDR27 BRP44L QKI PARK2 RNASET2 TCTE3 RPS6KA2 PACRG CCR6 DLL1 TTLL2 KIF25 UNC93A PDE10A DACT2 LOC441177 TCP10 FRMD1 PRR18 SMOC2 T GPR31 THBS2 C6orf123 C6orf70 C6orf208 C6orf176 LOC154449 C6orf118 LOC285796 C6orf120 TCP10L2 C6orf122 C6orf124 HGC6.3
8q24.12	AMP	30	44	0.04580	0.26502	Co-oc.	DEPDC6 COL14A1

References

1 Bast, J., Robert C, Hennessy, B. & Mills, G. B. The biology of ovarian cancer: new opportunities for translation. Nat Rev Cancer 9, 415-428, doi:10.1038/nrc2644 (2009).

2 Turner, N., Tutt, A. & Ashworth, A. Hallmarks of ``BRCAness'' in sporadic cancers. Nat Rev Cancer 4, 814-819, doi:10.1038/nrc1457 (2004).

3 Bast, R. C., Jr. & Mills, G. B. Personalizing therapy for ovarian cancer: BRCAness and beyond. J Clin Oncol 28, 3545-3548.

4 Hennessy, B. T. et al. Somatic mutations in BRCA1 and BRCA2 could expand the number of patients that benefit from poly (ADP ribose) polymerase inhibitors in ovarian cancer. J Clin Oncol 28, 3570-3576.

5 Merajver, S. D. et al. Somatic mutations in the BRCA1 gene in sporadic ovarian tumours. Nat Genet 9, 439-443, doi:10.1038/ng0495-439 (1995).

6 Berchuck, A. et al. Frequency of germline and somatic BRCA1 mutations in ovarian cancer. Clin Cancer Res 4, 2433-2437 (1998).

7 Foster, K. A. et al. Somatic and germline mutations of the BRCA2 gene in sporadic ovarian cancer. Cancer Res 56, 3622-3625 (1996).

8 Esteller, M. et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. Journal of the National Cancer Institute 92, 564-569 (2000).

9 Chiang, J. W., Karlan, B. Y., Cass, L. & Baldwin, R. L. BRCA1 promoter methylation predicts adverse ovarian cancer prognosis. Gynecol Oncol 101, 403-410, doi:S0090-8258(05)00974-1 [pii] 10.1016/j.ygyno.2005.10.034 (2006).

10 Press, J. Z. et al. Ovarian carcinomas with genetic and epigenetic BRCA1 loss have distinct molecular abnormalities. BMC Cancer 8, 17, doi:1471-2407-8-17 [pii]10.1186/1471-2407-8-17 (2008).

11 Hughes-Davies, L. et al. EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115, 523-535 (2003).

12 Brown, L. A. et al. Amplification of EMSY, a novel oncogene on 11q13, in high grade ovarian surface epithelial carcinomas. Gynecol Oncol 100, 264-270, doi:10.1016/j.ygyno.2005.08.026 (2006).

13 Lim, S. L. et al. Promoter hypermethylation of FANCF and outcome in advanced ovarian cancer. Br J Cancer 98, 1452-1456, doi:10.1038/sj.bjc.6604325 (2008).

14 Drew, Y. & Calvert, H. The potential of PARP inhibitors in genetic breast and ovarian cancers. Ann N Y Acad Sci 1138, 136-145, doi:10.1196/annals.1414.020 (2008).

15 Iglehart, J. D. & Silver, D. P. Synthetic lethality--a new direction in cancer-drug development. N Engl J Med 361, 189-191, doi:10.1056/NEJMe0903044 (2009).

16 Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434, 917-921, doi:10.1038/nature03445 (2005).

17 Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature 434, 913-917, doi:10.1038/nature03443 (2005).

18 Fong, P. C. et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med 361, 123-134, doi:10.1056/NEJMoa0900212 (2009).

19 Audeh, M. W. et al. Phase II trial of the oral PARP inhibitor olaparib (AZD2281) in BRCA-deficient advanced ovarian cancer. J Clin Oncol (Meeting Abstracts) 27, 5500- (2009).

20 McCabe, N. et al. Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. Cancer Res 66, 8109-8115, doi:10.1158/0008-5472.CAN-06-0140 (2006).

21 Mendes-Pereira, A. M. et al. Synthetic lethal targeting of PTEN mutant cells with PARP inhibitors. EMBO Mol Med 1, 315-322, doi:10.1002/emmm.200900041 (2009).

22 Ferla, R. et al. Founder mutations in BRCA1 and BRCA2 genes. Ann Oncol 18 Suppl 6, vi93-98, doi:18/suppl_6/vi93 [pii] 10.1093/annonc/mdm234 (2007).

23 Struewing, J. P. et al. The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. Nat Genet 11, 198-200, doi:10.1038/ng1095-198 (1995).

24 Roa, B. B., Boyd, A. A., Volcik, K. & Richards, C. S. Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nat Genet 14, 185-187, doi:10.1038/ng1096-185 (1996).

25 Backe, J. et al. Frequency of BRCA1 mutation 5382insC in German breast cancer patients. Gynecol Oncol 72, 402-406, doi:S0090-8258(98)95270-2 [pii] 10.1006/gyno.1998.5270 (1999).

26 Neuhausen, S. et al. Recurrent BRCA2 6174delT mutations in Ashkenazi Jewish women affected by breast cancer. Nat Genet 13, 126-128, doi:10.1038/ng0596-126 (1996).

27 Pal, T. et al. BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. Cancer 104, 2807-2816, doi:10.1002/cncr.21536 (2005).

28 Livingston, D. M. EMSY, a BRCA-2 partner in crime. Nat Med 10, 127-128, doi:10.1038/nm0204-127 (2004).

29 Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 10, 59-64, doi:10.1038/nrc2771 (2010).

30 Brown, L. A. et al. Amplification of 11q13 in ovarian carcinoma. Genes Chromosomes Cancer 47, 481-489, doi:10.1002/gcc.20549 (2008).

31 Gupta, A. et al. Cell cycle checkpoint defects contribute to genomic instability in PTEN deficient cells independent of DNA DSB repair. Cell cycle (Georgetown, Tex 8, 2198-2210 (2009).

32 Wood, R. D., Mitchell, M. & Lindahl, T. Human DNA repair genes, 2005. Mutat Res 577, 275-283, doi:10.1016/j.mrfmmm.2005.03.007 (2005).

33 Venkitaraman, A. R. Tracing the network connecting BRCA and Fanconi anaemia proteins. Nat Rev Cancer 4, 266-276, doi:10.1038/nrc1321 nrc1321 [pii] (2004).

34 Venkitaraman, A. R. A growing network of cancer-susceptibility genes. N Engl J Med 348, 1917-1919, doi:10.1056/NEJMcibr023150 (2003).

35 Lord, C. J., McDonald, S., Swift, S., Turner, N. C. & Ashworth, A. A high-throughput RNA interference screen for DNA repair determinants of PARP inhibitor sensitivity. DNA Repair (Amst) 7, 2010-2019, doi:10.1016/j.dnarep.2008.08.014 (2008).

36 Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33, D428-432, doi:10.1093/nar/gki072 (2005).

37 Matthews, L. et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37, D619-622, doi:10.1093/nar/gkn863 (2009).

38 Davalos, A. R. & Campisi, J. Bloom syndrome cells undergo p53-dependent apoptosis and delayed assembly of BRCA1 and NBS1 repair complexes at stalled replication forks. J Cell Biol 162, 1197-1209, doi:10.1083/jcb.200304016 jcb.200304016 [pii] (2003).

39 Tan, D. S. et al. "BRCAness" syndrome in ovarian cancer: a case-control study describing the clinical features and outcome of patients with epithelial ovarian cancer associated with BRCA1 and BRCA2 mutations. J Clin Oncol 26, 5530-5536 (2008).

40 Nakayama, N. et al. Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer. Cancer 116, 2621-2634.

41 Etemadmoghadam, D. et al. Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. Clin Cancer Res 15, 1417-1427 (2009).

Supplementary Methods M9: HotNet

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

We analyzed the combined mutation and copy number data for the 316 samples. For each sequenced gene, we defined the gene as *altered* in a sample if the gene had a somatic mutation, or if the gene was present in a focal aberration according to GISTIC analysis (*i.e.* the gene was annotated as -2 or 2). We excluded alterations in *TP53* because this gene is mutated in the majority of samples and we aimed to identify other pathways not associated with p53. We also excluded alterations in TTN, since it is mutated at a rate higher than expected by the background mutation rate, but analysis of these mutations indicates that they are likely artifacts. Moreover, we removed CNAs for which the sign of the aberration was not the same in at least 90% of altered samples.

The resulting alteration data on 316 samples was input to HotNet (Figure S9.1). We used the interaction network derived from the Human Protein Reference Database (HPRD) [2]. For the HotNet statistical test, we generated random datasets in the following manner. We simulated mutations using the estimated background mutation rate (1.7368×10^{-6}). We simulated CNAs using the observed distribution of CNA lengths and permuting their positions. The latter minimizes potential artifacts resulting from functionally related genes that are both neighbors on the interaction network and close enough on the genome that they are affected by the same CNA. To further reduce such artifacts, we applied two additional heuristics. First, we removed candidate subnetworks returned by HotNet that contain 3 or more genes in the same focal CNA in more than 1% of the samples. Second, for subnetworks with 2 genes g_1, g_2 in the same focal CNA in more than 1% of the samples, we removed the genes that were not found in the subnetwork when alterations in either g_1 or g_2 are removed. Using this approach HotNet identified 32 candidate subnetworks containing at least 7 genes ($p = 0.03$) with a corresponding FDR = 0.68 for the list of subnetworks. The FDR is a conservative estimate of the ratio of false positives among all subnetworks reported by HotNet, and implies that (at least) a third of the subnetworks reported by HotNet are significant. A total of 27 subnetworks remained after CNA filtering (Table S9.1 and Figure S9.2).

To gain additional support for individual subnetworks and to focus attention on subnetworks with known biological function, we computed the overlap between the genes in candidate subnetworks and: i) known pathways from the KEGG database [3]; ii) protein complexes from PINdb [4]. Of those 27 subnetworks returned by HotNet, 12 had statistically significant ($p \leq 0.05$) overlap with at least one KEGG pathway or PINdb protein complex (Table S9.2). Among the most significant subnetworks were the core promoter recognition complex TFIID, the Notch signalling pathway, the cohesin complex, and RNA polymerase II. In particular, the Notch signalling pathway was altered in 22% of samples (Figure 3B, main text). Notch signalling has been implicated in many cancers [5] and the NOTCH3-JAG1 interaction has been shown to be important for proliferation of ovarian cancer cells [6].

For each of the 12 subnetworks reported in Table S9.2, we tested the association between the mutation status of each subnetwork and the expression subtypes of the samples, using a χ^2 test. For each subnetwork, we built a contingency table in which the column variable is the expression subtype of a sample (Differentiated, Immunoreactive, Proliferative, or Mesenchymal) and the row variable is the mutation status (altered or unaltered) of the subnetwork in the sample. The *p*-value for this test is reported for each subnetwork in Table S9.2. Six subnetworks showed a statistically significant association between the mutation status of the subnetwork and the expression subtype (FDR<0.05 after correction for the 12 hypothesis tests). Although each of these six subnetworks has a dominant expression subtype, there was insufficient power to associate only one expression subtype to the subnetwork.

Tables

Table S9.1: The 27 subnetworks identified by HotNet. (see XLS SupplementMethodsM9Table2.xls)

Table S9.2: Significantly altered subnetworks identified with HotNet

Genes	*KEGG pathway enrichment (p-value, corrected)*	*PINdb complexes : top enrichment (p-value, corrected)*	*Altered Samples (%)*	*Association to Expression Subtype p-value*
MAML2 MAML3 MAML1 RBPJ NOTCH3 JAG2 JAG1	Notch signaling pathway (<10^{-13})		70 (22%)	0.18
HSF1 TAF7 TAF5 TAF4 TAF2 TAF1 TAF6 TAF9 TAF12 CPSF1	Basal transcription factors(<10^{-13})	TAF4b-TFIID TAF4b-TFIID; 4b-IID; 4b/4-IID (<10^{-13}) TAF4-TFIID TAF4-TFIID; 4-IID;4/4-IIB (<10^{-13}) TFTC SAGA-like; hSAGA; TBP-free TAFII-containing (<10^{-13}) TFIID hTFIID; transcription initiation factor (<10^{-13})	133 (42%)	0.008
GJA1 IGF1 POLR2L SMYD3 CTGF POLR2G POLR2K POLR2C POLR2B S100A4 FBLN1 POLR2I POLR2H NOV	RNA polymerase (2x10^{-13}) Pyrimidine metabolism (6x10^{-10}) Purine metabolism (8x10^{-9}) Huntington's disease (3x10^{-8})	PolII(G) RNAPII; Gdown1-containing Pol II (<10^{-13}) TAP-tagged RNAPII RNAPII; RNA polymerase II (<10^{-13}) RNA polymerase II DNA-directed RNA polymerase II; RNAP II; RNAPII; RNA pol II; RNA polII (<10^{-13}) Integrator DSS1-associated; RNAPII-associated (10^{-2})	150 (47%)	0.73
LRRC8D NEK1 GADD45A SKIV2L2 MPZL1 GTF2IRD1 EXOSC7 UPF3B GADD45GIP1 AKR1A1 UPF1 SKIV2L UPF2 EXOSC2 EXOSC1 SMPD4 GADD45B EXOSC4	RNA degradation (2x10^{-8})		146 (46%)	0.11
MAP2 LMNB1 PLEC1 ITGB4 ACTC1 ITGA6 SPTA1 MSN SPTAN1 ACTG1 COL17A1	Hypertrophic cardiomyopathy (HCM) (3x10^{-4}) Dilated cardiomyopathy (3x10^{-4}) Regulation of actin cytoskeleton (10^{-2}) Arrhythmogenic right ventricular cardiomyopathy (ARVC) (10^{-2})		131 (41%)	0.01
SMC1A NUMA1 STAG2 STAG1 CHRAC1 POLE3 RAD21 SMARCA5	Cell cycle (4x10^{-4})	SNF2h/cohesion SNF2h; ISWI-contaning (2x10^{-8}) HuCHRAC CHRAC; hCHRAC; ISWI; human ISWI; human ISWI-containing (10^{-6}) cohesin-1 14S cohesin; SA1-cohesin; SA1-containing (10^{-6}) cohesin-2 14S cohesin; SA2-cohesin; SA2-containing (10^{-6})	109 (34%)	0.009
LY6E VAV1 FCGR2B EPHA2 CD247 ZAP70 LAT SLA FGFR1	T cell receptor signaling pathway(4x10^{-4}) Natural killer cell mediated cytotoxicity (9x10^{-4}) Fc gamma R-mediated phagocytosis (10^{-2})		121 (38%)	0.02
KARS CD48 VARS RPS6KA1 EEF1D GARS KTN1 CTBP1 CTBP2 C10orf4 HEXDC	Aminoacyl-tRNA biosynthesis (5x10^{-4})	CtBP CtBP co-repressor; CtBP corepressor; CtBP1-containing (4x10^{-2})	103 (33%)	0.0003
RASD2 IRS4 PIK3CA NRAS MRAS APPL1 PIK3R3 GABRB1 APLP2	Neurotrophin signaling pathway (8x10^{-4}) Insulin signaling pathway (10^{-3}) Non-small cell lung cancer (3x10^{-3}) Type II diabetes mellitus (3x10^{-3})		84 (27%)	0.54
LRP2 HSPA5 P4HB HSP90B1 ASGR1 CTSL1 APCS TG CANX	Antigen processing and presentation (4x10^{-3})		118 (37%)	0.01
NRCAM MACF1 NRXN2 CNTNAP2 GOLGA4 CNTNAP4 PLXND1	Cell adhesion molecules (CAMs) (10^{-2})		69 (22%)	0.49

AP2M1 AQP4 ADRA1B EHD2 AP1M2 KCNJ11 MED4 LY9 TGOLN2 LDLR GAK LAMP1 RRP12 LDLRAP1 TBC1D5	Endocytosis ($2x10^{-2}$)		113 (36%)	0.27

Figures

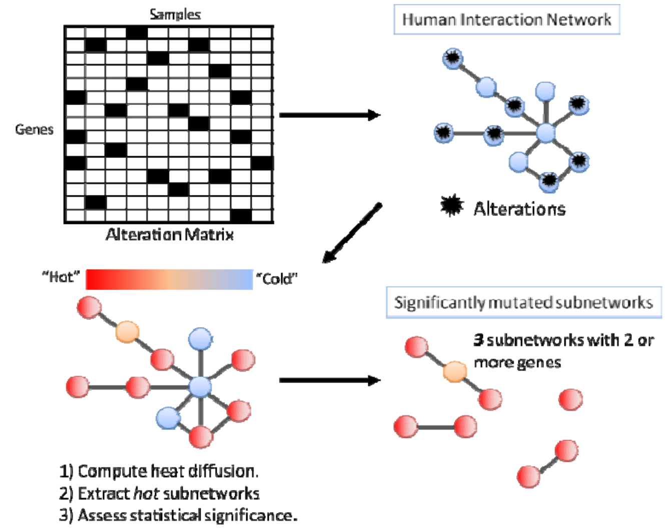

Samples
Genes
Alteration Matrix
Human Interaction Network
Alterations
"Hot"
"Cold"
Significantly mutated subnetworks
3 subnetworks with 2 or more genes
1) Compute heat diffusion.
2) Extract *hot* subnetworks
3) Assess statistical significance.

Figure S9.1: Overview of HotNet

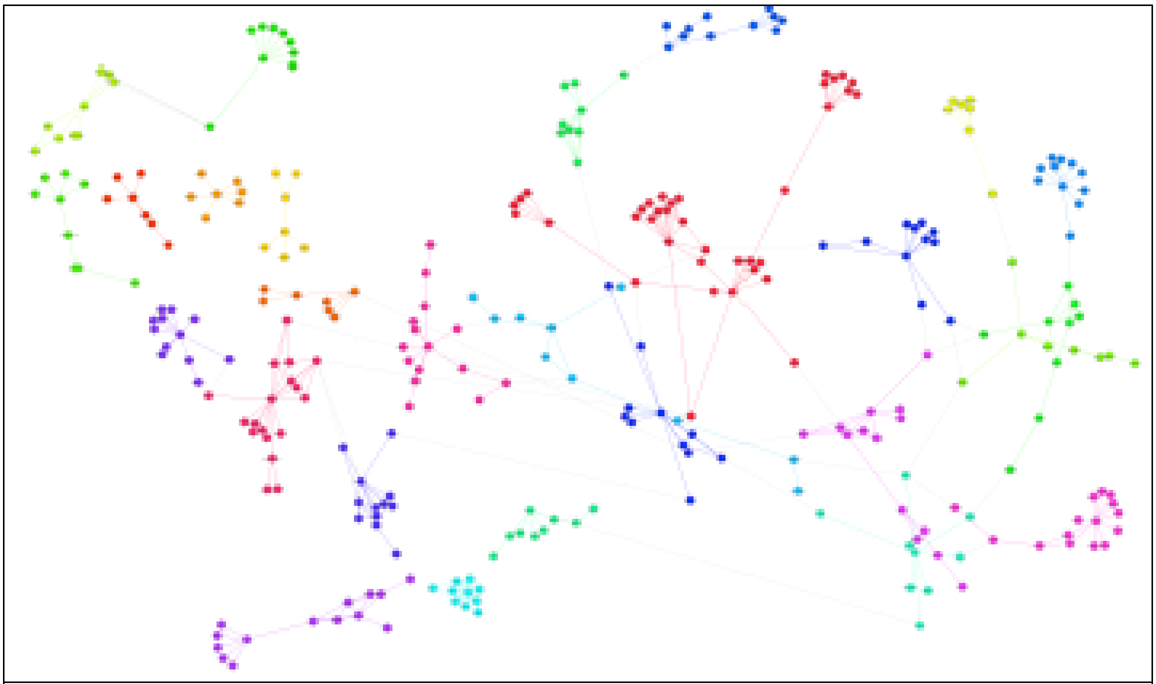


Figure S9.2: The 27 subnetworks identified by HotNet. Nodes correspond to proteins, and are colored using a different color for each subnetwork. Edges correspond to interactions in HPRD [2]: colored edges are interactions between proteins in the same subnetwork, while gray edges are interactions between proteins in different subnetworks.

References

1. Vandin, F., E. Upfal, and B. Raphael, *Algorithms for Detecting Significantly Mutated Pathways in Cancer*, in *Research in Computational Molecular Biology*, B. Berger, Editor. 2010, Springer Berlin / Heidelberg. p. 506-521.
2. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
3. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
4. Luc, P.V. and P. Tempst, *PINdb: a database of nuclear protein complexes from human and yeast.* Bioinformatics, 2004. **20**(9): p. 1413-5.
5. Bray, S.J., *Notch signalling: a simple pathway becomes complex.* Nat Rev Mol Cell Biol, 2006. **7**(9): p. 678-89.
6. Choi, J.H., et al., *Jagged-1 and Notch3 juxtacrine loop regulates ovarian tumor growth and adhesion.* Cancer Res, 2008. **68**(14): p. 5716-23.

Supplemental Methods S10. Integrated Pathway Analysis

Supplemental Results

Frequently altered pathways in ovarian serous carcinomas

To identify significantly altered pathways through an integrated analysis of both copy number and gene expression, we applied the recently developed pathway activity inference method PARADIGM (PMID: 20529912). The computational model incorporates copy number changes, gene expression data, and pathway structures to produce an *integrated pathway activity* (IPA) for every gene, complex, and genetic process present in the pathway database. We use the term “entity” to refer to any molecule in a pathway be it a gene, complex, or small molecule. The IPA of an entity refers only to the final activity. For a gene, the IPA only refers to the inferred activity of the active state of the protein, which is inferred from copy number, gene expression, and the signaling of other genes in the pathway. We applied PARADIGM to the ovarian samples and found alterations in many different genes and processes present in pathways contained in the National Cancer Institutes’ Pathway Interaction Database (NCI-PID). We assessed the significance of the inferred alterations using 1000 random simulations in which pathways with the same structure were used but arbitrary genes were assigned at different points in the pathway. In other words, one random simulation for a given pathway kept the set of interactions fixed so that an arbitrary set of genes were connected together with the pathway’s interactions. The significance of all samples’ IPAs was assessed against the same null distribution to obtain a significance level for each entity in each sample. IPAs and the percentage of samples in which they are significant are listed in **Table S10.1** and IPAs with a standard deviation of at least 0.1 are displayed as a heatmap in **Figure S10.1**.

Table S10.2 shows the pathways altered by at least three standard deviations with respect to permuted samples found by PARADIGM. The FOXM1 transcription factor network was altered in the largest number of samples among all pathways tested – 67% of entities with altered activities when averaged across samples. In comparison, pathways with the next highest level of altered activities in the ovarian cohort included PLK1 signaling events (27%), Aurora B signaling (24%), and Thromboxane A2 receptor signaling (20%). Thus, among the pathways in NCI-PID, the FOXM1 network harbors significantly more altered activities than other pathways with respect to the ovarian samples.

The FOXM1 transcription factor network was found to be differentially altered in the tumor samples compared to the normal controls in the highest proportion of the patient samples (**Figure S10.2**, **Table S10.1**). FOXM1 is a multifunctional transcription factor with three known dominant splice forms, each regulating distinct subsets of genes with a variety of roles in cell proliferation and DNA repair. The FOXM1c isoform directly regulates several targets with known roles in cell proliferation including AUKB, PLK1, CDC25, and BIRC5 (PMID:15671063). On the other hand, the FOXM1b isoform regulates a completely different subset of genes that include the DNA repair genes BRCA2 and XRCC1 (PMID:17101782). CHEK2, which is under indirect control of ATM, directly regulates FOXM1s expression level.

We asked whether the IPAs of the FOXM1 transcription factor itself were more highly altered than the IPAs of other transcription factors. We compared the FOXM1 level of activity to all of the other 203 transcription factors in the NCI-PID. Even compared to other transcription factors in the NCI set, the FOXM1 transcription factor had significantly higher levels of activity (p<0.0001; K-S test) suggesting further that it may be an important signature (**Figure S10.3**).

Because FOXM1 is also expressed in many different normal tissues of epithelial origin, we asked whether the signature identified by PARADIGM was due to an epithelial signature that would be considered normal in other tissues. To answer this, we downloaded an independent dataset from GEO (GSE10971) (PMID:18593983) in which fallopian tube epithelium and ovarian tumor tissue were microdissected and gene expression was assayed. We found that the levels of FOXM1 were significantly higher in the tumor samples compared to the normals, suggesting FOXM1 regulation is indeed elevated in cancerous tissue beyond what is seen in normal epithelial tissue (**Figure S10.4**).

Because the entire cohort for the TCGA ovarian contained samples derived from high-grade serous tumors, we asked whether the FOXM1 signature was specific to high-grade serous. We obtained the log expression of FOXM1 and several of its targets from the dataset of Etemadmoghadam et al. (2009) (PMID:19193619) in which both low- and high-grade serous tumors had been transcriptionally profiled. This independent data confirmed that FOXM1 and several of its targets are significantly up-regulated in serous ovarian relative to low-grade ovarian cancers (**Figure S10.5**). To determine if the 25 genes in the FOXM1 transcription factor network contained a significant proportion of genes with higher expression in high-grade disease, we performed a Student's t-test using the data from Etemadmoghadam. 723 genes in the genome (5.4%) were found to be significantly up-regulated in high- versus low-grade cancer at the 0.05 significance level (corrected for multiple testing using the Benjamini-Hochberg method). The FOXM1 network was found to have 13 of its genes (52%) differentially regulated, which is a significant proportion based on the hypergeometric test ($P < 3.8*10^{-12}$). Thus, high expression of the FOXM1 network genes does appear to be specifically associated with high-grade disease when compared to the expression of typical genes in the genome.

FOXM1's role in many different cancers including breast and lung has been well documented but its role in ovarian cancer has not been investigated. FOXM1 is a multifunctional transcription factor with 3 known splice forms, each regulating distinct subsets of genes with a variety of roles in cell proliferation and DNA repair. An excerpt of FOXM1's interaction network relevant to this analysis is shown in the main text as **Figure 3D**. The FOXM1a isoform directly regulates several targets with known roles in cell proliferation including AUKB, PLK1, CDC25, and BIRC5. In contrast, the FOXM1b isoform regulates a completely different subset of genes that include the DNA repair genes BRCA2 and XRCC1. CHEK2, which is under indirect control of ATM, directly regulates FOXM1's expression level. In addition to increased expression of FOXM1 in most of the ovarian patients, a small subset also have increased copy number amplifications detected by CBS (19% with copy number increases in the top 5% quantile of all genes in the genome measured). Thus the alternative splicing regulation of FOXM1 may be involved in the control switch between DNA repair and cell proliferation. However, there is insufficient data at this point to support this claim since the exon structure distinguishing the isoforms and positions of the Exon array probes make it difficult to distinguish individual isoform activities. Future high-throughput sequencing of the mRNA of these samples may help determine the differential levels of the FOXM1 isoforms. The observation that PARADIGM detected the highest level of altered activity centered on this transcription factor suggests that FOXM1 resides at a critical regulatory point in the cell.

SUPPLEMENTAL METHODS

Data Sets and Pathway Interactions

Both copy number and expression data were incorporated into PARADIGM inference. Since a set of eight normal tissue controls was available for analysis in the expression data, each patient's gene-value was normalized by subtracting the gene's median level observed in the normal fallopian control. Copy number data was normalized to reflect the difference in copy number between a gene's level detected in tumor versus a blood normal. For input to PARADIGM, expression data was taken from the same integrated dataset used for subtype analysis and the copy number was taken from the segmented calls of MSKCC Agilent 1M copy number data.

A collection of pathways was obtained from NCI-PID on September 15, 2009 containing 131 pathways, 11,563 interactions, and 7,204 entities. An entity is molecule, complex, small molecule, or abstract concept represented as "nodes" in PARADIGM's graphical model. The abstract concepts correspond to general cellular processes (such as "apoptosis" or "absorption of light,") and families of genes that share functional activity such as the RAS family of signal transducers. We collected interactions including protein-protein interactions, transcriptional regulatory interactions, protein modifications such as phosphorylation and ubiquitinylation interactions.

Inference of integrated molecular activities in pathway context.

We used PARADIGM, which assigns an integrated pathway activity (IPA) reflecting the copy number, gene expression, and pathway context of each entity.

The significance of IPAs was assessed using permutations of gene- and patient-specific cross-sections of data. Data for 1000 "null" patients was created by randomly selecting a gene-expression and copy number pair of values for each gene in the genome. To assess the significance of the PARADIGM IPAs, we constructed a null distribution by assigning random genes to pathways while preserving the pathway structure.

Identification of FOXM1 Pathway

While all of the genes in the FOXM1 network were used to assess the statistical significance during the random simulations, in order to allow visualization of the FOXM1 pathway, entities directly connected to FOXM1 with significantly altered IPAs according to **Figure S10.2** were chosen for inclusion in **Figure 3D** of the main text. Among these, genes with roles in DNA repair and cell cycle control found to have literature support for interactions with FOXM1 were displayed. BRCC complex members, not found in the original NCI-PID pathway, were included in the plot along with BRCA2, which is a target of FOXM1 according to NCI-PID. Upstream DNA repair targets were identified by finding upstream regulators of CHEK2 in other NCI pathways (e.g. an indirect link from ATM was found in the PLK3 signaling pathway).

Clustering

The use of inferred activities, which represent a change in probability of activity and not activity directly, it enables entities of various types to be clustered together into one heatmap. To globally visualize the results of PARADIGM inference, Eisen Cluster 3.0 was used to perform feature filtering and clustering.   A standard deviation filtering of 0.1 resulted in 1598 out of 7204 pathway entities remaining, and average linkage, uncentered correlation hierarchical cluster was performed on both the entities and samples.

Supplemental Figures

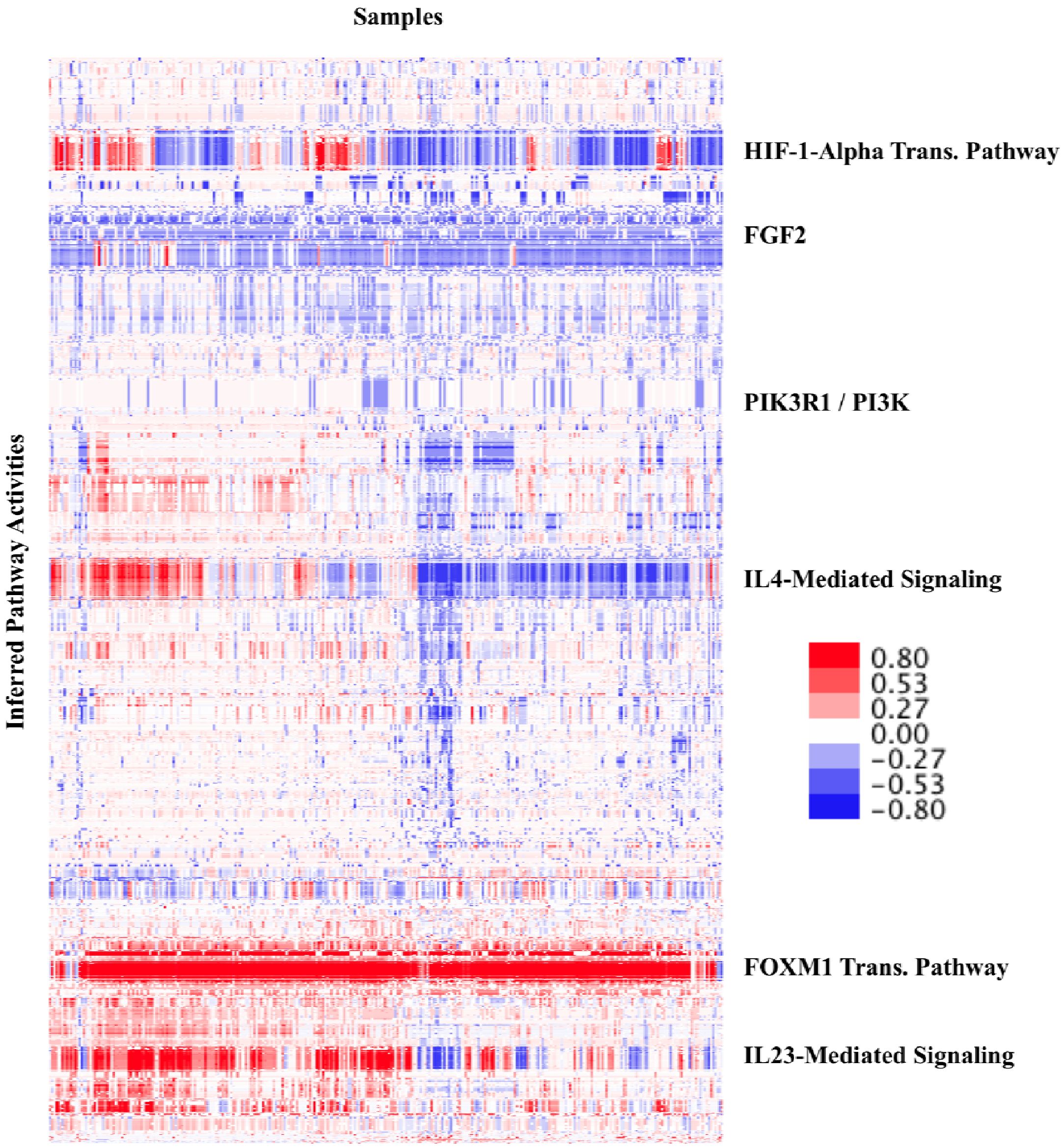


Figure S10.1. Heatmap of Inferred Pathway Activities (IPAs). IPAs representing 1598 inferences of molecular entities (rows) inferred to be activated (red) or inactivated (blue) are plotted for each of 316 patient tumor samples (columns). IPAs were hierarchically clustered by pathway entity and tumor sample, and labels on the right show sections of the heatmap enriched with entities of individual pathways.. The colorbar legend is in log base 10.

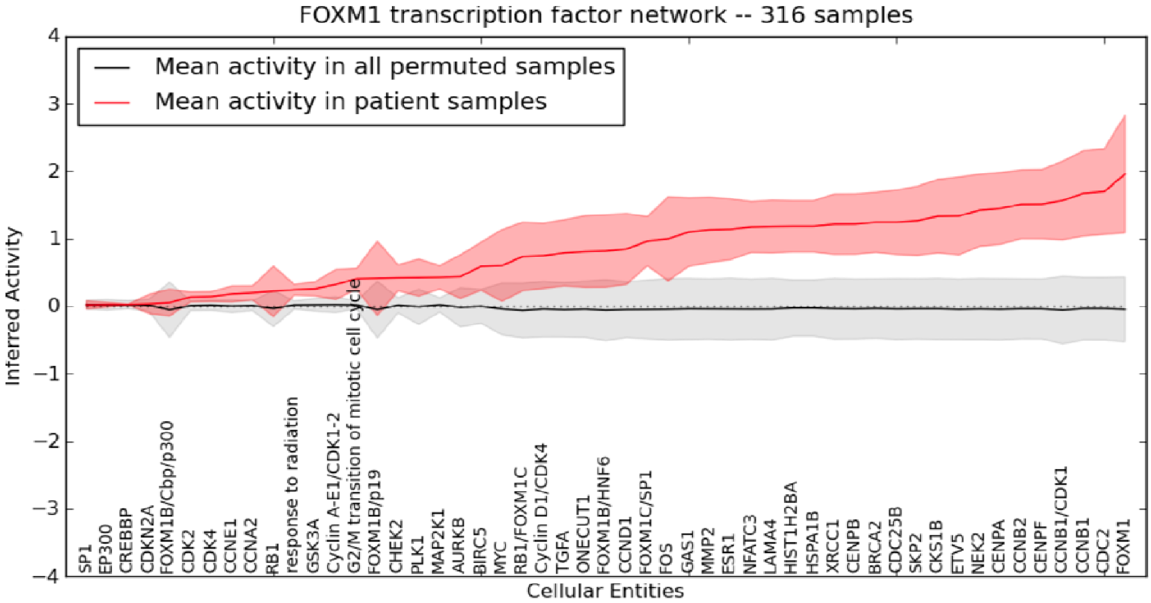


Figure S10.2. Summary of FOXM1 integrated pathway activities (IPAs) across all samples. The arithmetic mean of IPAs across tumor samples for each entity in the FOXM1 transcription factor network is shown in red, with heavier red shading indicating two standard deviations. Gray line and shading indicates the mean and two standard deviations for IPAs derived from the 1000 ”null” samples.

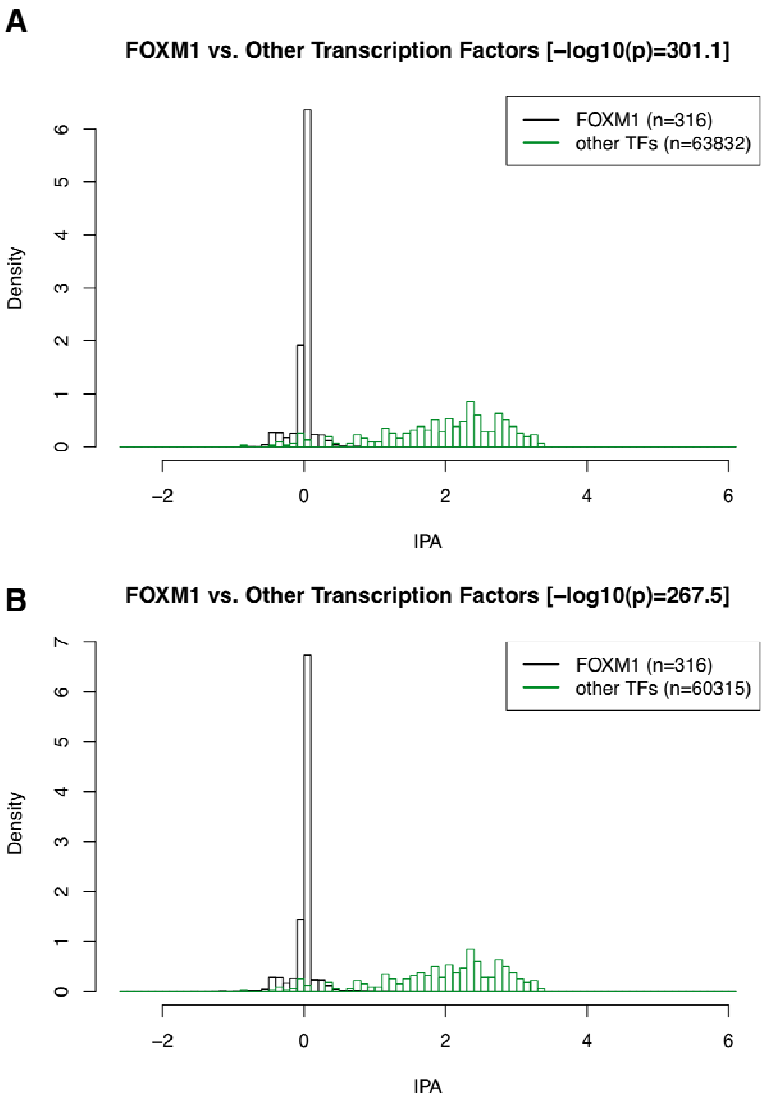
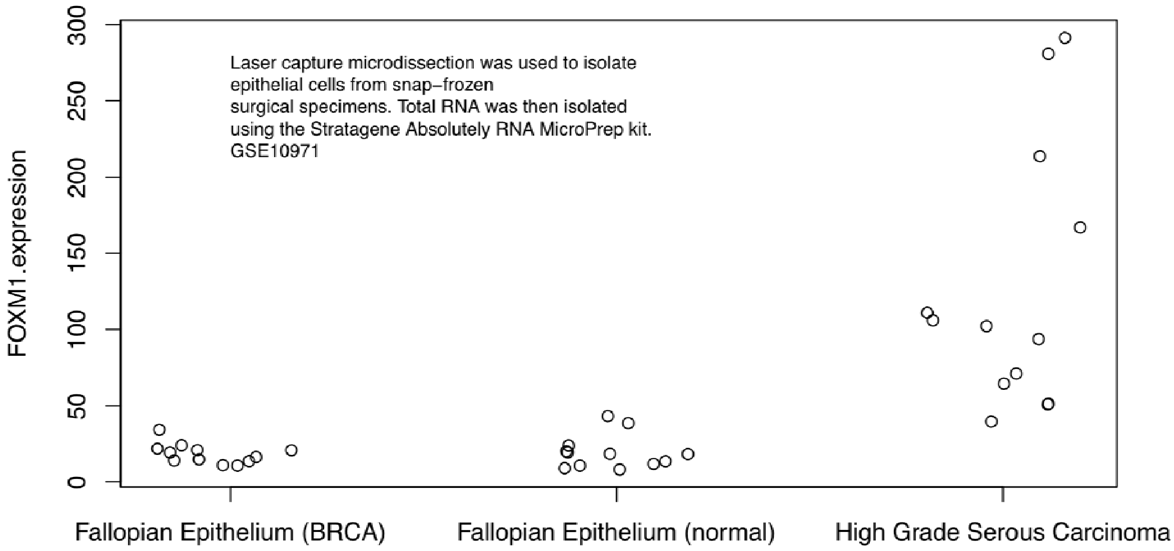


Figure S10.3. Comparison of IPAs of FOXM1 to those of other tested transcription factors (TFs) in NCI Pathway Interaction Database. **A.** Histogram of IPAs with non-active (zero-valued) IPAs removed. FOXM1 targets are significantly more activated than other NCI TFs ($P < 10^{-267}$; Kolmogorov-Smirnov (KS) test). B. Histogram of all IPAs including non-active IPAs. Using all IPAs, FOXM1's activity relative to other TFs is interpreted with somewhat higher significance ($P < 10^{-301}$; KS test).



Figures S10.4. FOXM1 is not expressed in fallopian epithelium compared to serous ovarian carcinoma. FOXM1's expression levels in fallopian tube was compared to its levels in serous ovarian carcinoma using the data from Tone et al (PMID: 18593983). FOXM1's expression is much lower in fallopian tube, including in samples carrying BRCA 1/2 mutations, indicating that FOXM1's elevated expression observed in the TCGA serous ovarian cancers is not simply due to an epithelial signature.

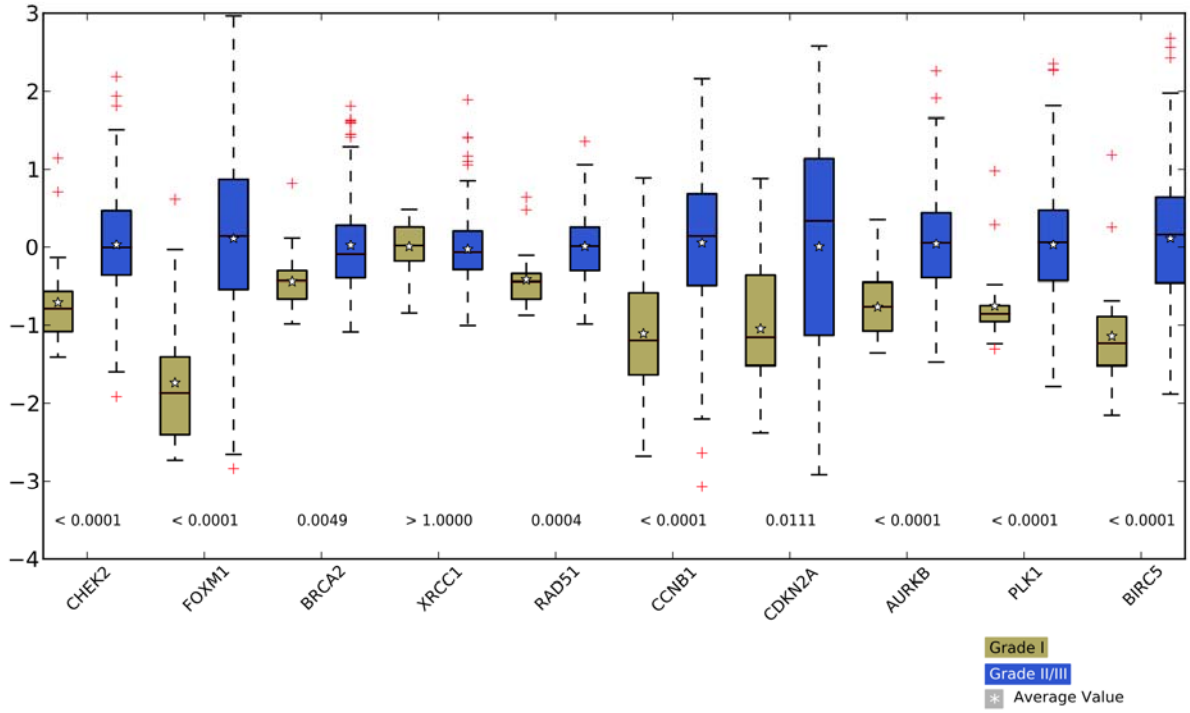


Figure S10.5. Expression of FOXM1 transcription factor network genes in high grade versus low grade carcinoma. Expression levels for FOXM1 and nine selected FOXM1 targets (based on NCI-PID) were plotted for both low-grade (I; tan boxes; 26 samples) and high-grade (II/III; blue boxes; 296 samples) ovarian carcinomas. Seven out of the nine targets were showed to have significantly high expression of FOXM1 in the high-grade carcinomas (Student’s t-test; p-values noted under boxplots). CDKN2A may also be differentially expressed but had a borderline t-statistic (P = 0.01). XRCC1 was detected as differentially expressed.

Supplement S11:
Calculations to identify batch effects in the ovarian mRNA expression data

Nianxiang Zhang, Rehan Akbani, Keith A. Baggerly, John N. Weinstein
University of Texas, M. D. Anderson Cancer Center Genome Data Analysis Center – 11/8/10

Supplemental Methods:

We used our correlation of correlations parameter[1] to check for differences among sample batches.

$$U_{ij} = \frac{\sum_{d=1}^{D} X_{di}X_{dj} - \frac{1}{D}\sum_{d=1}^{D} X_{di}\sum_{d=1}^{D} X_{dj}}{\sqrt{\sum_{d=1}^{D} X_{di}^{2} - \frac{1}{D}\left(\sum_{d=1}^{D} X_{di}\right)^{2}}\sqrt{\sum_{d=1}^{D} X_{dj}^{2} - \frac{1}{D}\left(\sum_{d=1}^{D} X_{dj}\right)^{2}}},$$

X_{di} and X_{dj} are the mRNA expression levels of genes i and j in sample d;
D is the number of samples; U_{ij} is the correlation of genes *i* and *j* in batch A; the expression for V_{ij}, the correlation of genes *i* and *j* in batch B, is analogous. Then,

$$r_c = \frac{\sum_{i<j} U_{ij}V_{ij} - \frac{2}{n(n-1)}\sum_{i<j} U_{ij}\sum_{i<j} V_{ij}}{\sqrt{\sum_{i<j} U_{ij}^{2} - \frac{2}{n(n-1)}\left(\sum_{i<j} U_{ij}\right)^{2}}\sqrt{\sum_{i<j} V_{ij}^{2} - \frac{2}{n(n-1)}\left(\sum_{i<j} V_{ij}\right)^{2}}},$$

where n denotes the number of genes, and r_c is the correlation of correlations.

First, we filtered out noisy genes from informative genes by selecting only those with standard deviation greater than 0.5 and median absolute deviation greater than 0.5. Using 2000 randomly sampled informative genes, we calculated all pairwise Pearson correlations between genes within a batch, obtaining a vector for each batch. The Pearson correlation between two such vectors is the r_c for that pair of batches. Our null hypothesis was that no batch effect is present. We assessed statistical significance using a permutation test. To do so, we randomly permuted the sample-to-batch mapping (thereby removing any systematic structure), calculated a “null” set of r_c values using 200 randomly sampled genes (to save computation time), and repeated the procedure 10,000 times to assemble null cumulative distribution functions (CDFs). Significance levels of the observed r_c values were then estimated from the empirical CDFs with one-sided tests. (Preliminary testing had shown that the 2000-gene and 200-gene random samples described above were easily sufficient in size to represent the entire gene sets accurately.)

The p-values shown in **Figure S11.1** serve as indicators of the severity of batch effects in mRNA expression data. Low p-values indicate significant batch effects, and values below 0.05

are shown in red. The p-values test the corresponding correlation of correlations (CR) metric, r_c, defined above. Given that gene expression measurements are nonlinear and batch effects go beyond mere scaling, the CR metric is useful in that it characterizes batch differences by capturing changes in gene-gene interaction networks within batches. It provides an alternative approach to assessment of batch effects, complementing the more routine principal component (PCA) and clustering analyses (see **Figures S11.2-9**). Since it provides statistical tests and p-values for batch effects, it is more objective than visualization of hierarchically defined clusters or PCA figures.

We observed modest batch effects for the Affymetrix U133A mRNA data (9 significant pairs out of 78 batch pairs), but more severe batch effects for Agilent G4502A and Affymetrix Human Exon array data (23 and 22 significant pairs, respectively). However, batch effects were largely absent (3 significant pairs) in the unified data used for mRNA analyses in the main text of the article.

Supplementary figure:

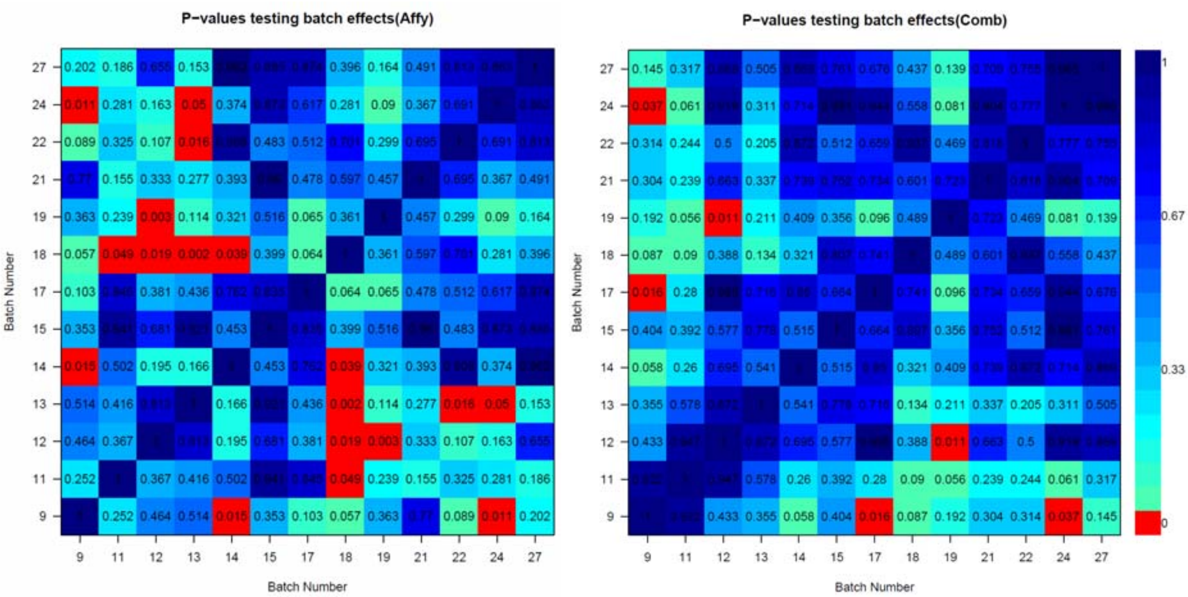


Figure S11.1 Assessment of batch effects in the TCGA ovarian cancer gene expression data. P-values were obtained from permutation tests of pairwise correlation of correlations (r_c) levels between batches for Affymetrix U133A level 3 data (left) and unified gene expression data from the three TCGA platforms (right). Red indicates a significant difference between batches (p<0.05, without multiple-comparisons correction).

To complement those results, we performed PCA and unsupervised hierarchical clustering on the unified expression data (**Figures S11.2-5**)[2]. For **Figure S11.4**, we used the average linkage algorithm and Euclidean distance metric (all in the log frame, without subtracting the mean across samples or across genes). Genes were filtered by removing those with standard deviation ≤ 0.5 across the samples. Results were similar when we used a 1-Pearson correlation metric with the average linkage rule (**Figure S11.5**). We annotated the samples by their batch number (top colored bar) and center of origin (bottom colored bar) to ascertain whether there were any significant center-specific effects or batch effects. The samples did not cluster to an appreciable degree by center or by batch in either PCA or clustering analysis. The results suggest that observed patterns or signatures in the data are not artifacts of the methods used by each individual centre or technical differences among batches. However, the importance of batch effects or bias obviously depends on the nature and strength of a prediction or pattern relationship being derived from the data. R-scripts for checking the calculations here can be obtained on request (nzhang@mdanderson.org).

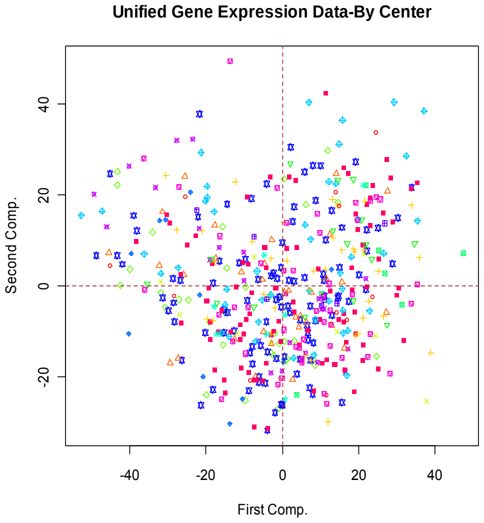


Figure S11.2: PCA plot of unified gene expression data, annotated by center.

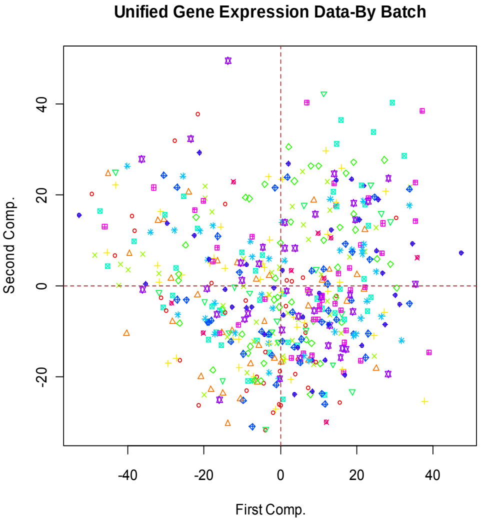
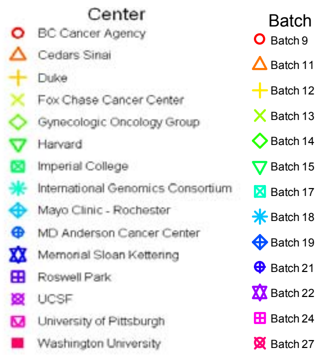


Figure S11.3: PCA plot of unified gene expression data, annotated by batch number.



Legend for Figures S11.2 & S11.3.

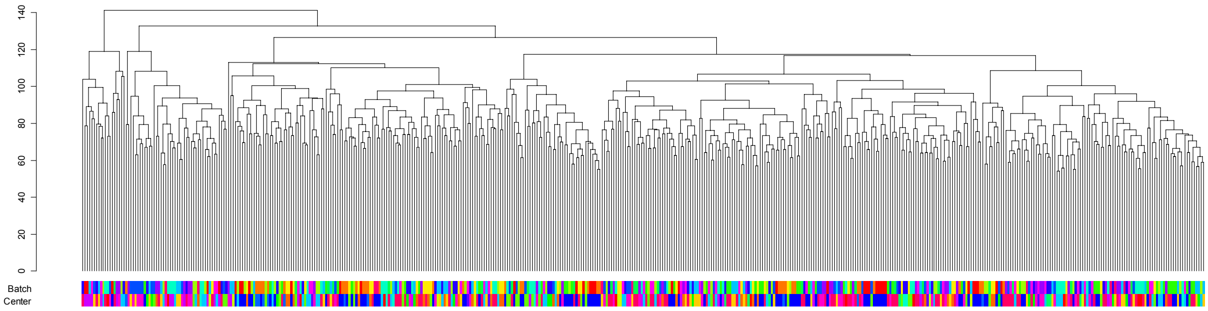


Fig S11.4: Unsupervised heirarchical clustering (Euclidean distance metric, average linkage alrogithm) of unified mRNA expression data. The top colored bar indicates batch number; the bottom colored bar indicates center.

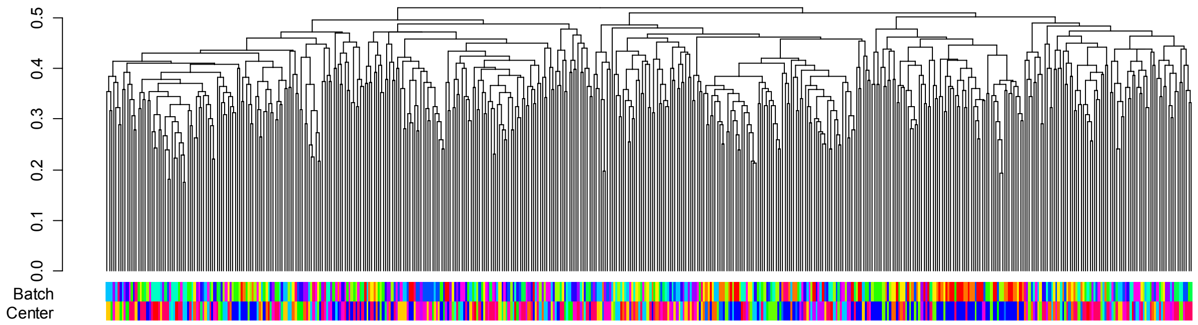


Fig S11.5: Unsupervised heirarchical clustering (1-Pearson correlation metric, average linkage algorithm) of unified mRNA expression data. The top colored bar indicates batch number; the bottom colored bar indicates center.

Center	Batch
🟥 BC Cancer Agency	🟥 Batch 9
🟧 Cedars Sinai	🟧 Batch 11
🟨 Duke	🟨 Batch 12
🟩 Fox Chase Cancer Center	🟩 Batch 13
🟩 Gynecologic Oncology Group	🟩 Batch 14
🟩 Harvard	🟩 Batch 15
🟩 Imperial College	🟦 Batch 17
🟦 International Genomics Consortium	🟦 Batch 18
🟦 Mayo Clinic - Rochester	🟦 Batch 19
🟦 MD Anderson Cancer Center	🟦 Batch 21
🟦 Memorial Sloan Kettering	🟪 Batch 22
🟪 Roswell Park	🟪 Batch 24
🟪 UCSF	🟥 Batch 27
🟪 University of Pittsburgh	
🟥 Washington University	

Legend for Figures S11.4-9.

We applied Johnson et al.'s empirical Bayes algorithm[3] to the unified gene expression data and performed hierarchical clustering on the data after adjusting by batch number (**Figure S11.6-7**) and by center (**Figure S11.8-9**) for both Euclidean distance and 1-Pearson correlation metrics. When we compared **Figures S11.4-5** (before adjustment) with **Figures S11.6-7** and **S11.8-9** (after batch and center adjustment, respectively), we saw only modest improvements in uniformity. Furthermore, given that some batches have known clinical differences from other batches, it is unclear to what extent the differences corrected for were biological or technical in origin. Therefore, we decided in favor of using unadjusted batches for the mRNA analyses.

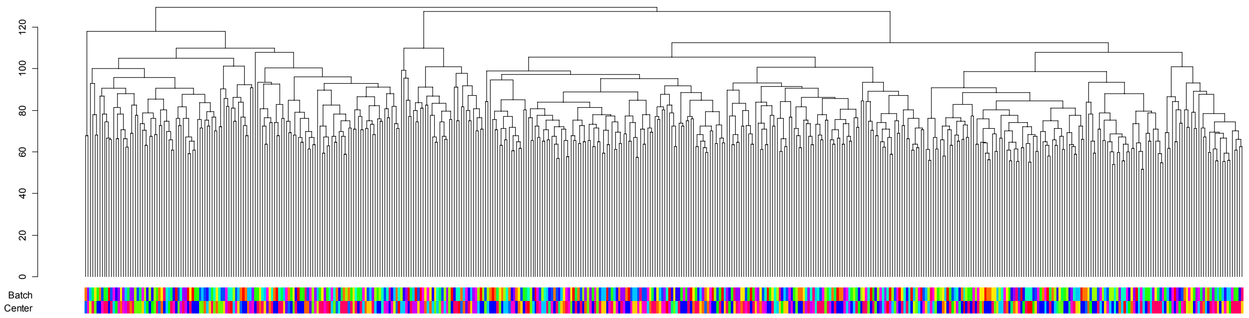


Fig S11.6: Unsupervised hierarchical clustering (Euclidean distance metric, average linkage algorithm) of unified mRNA expression data after adjusting by batch number using the empirical Bayes algorithm of Johnson, et al. The top colored bar indicates batch number; the bottom colored bar indicates center (see legend above)

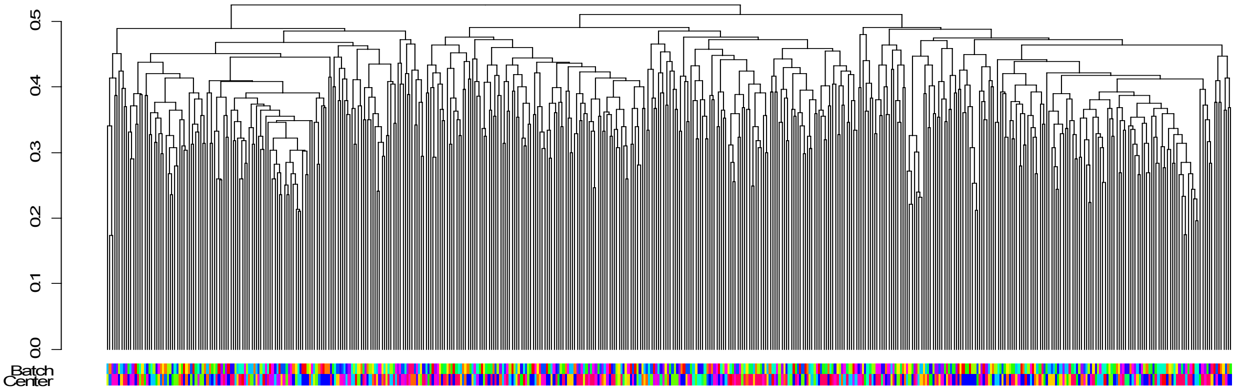


Fig S11.7: Unsupervised hierarchical clustering (1-Pearson correlation metric, average linkage) of unified mRNA expression data after adjusting by batch number as in Figure S11.6. The top colored bar indicates batch number; the bottom colored bar indicates center (see legend above).

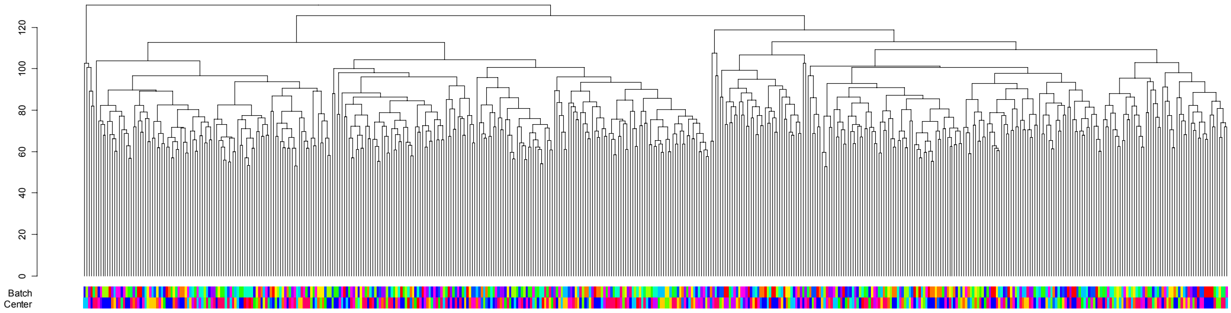


Fig S11.8: Unsupervised hierarchical clustering (Euclidean distance metric, average linkage algorithm) of unified mRNA expression data after adjusting by center of origin as in Figure S11.6. The top colored bar indicates batch number; the bottom colored bar indicates center (see legend above)

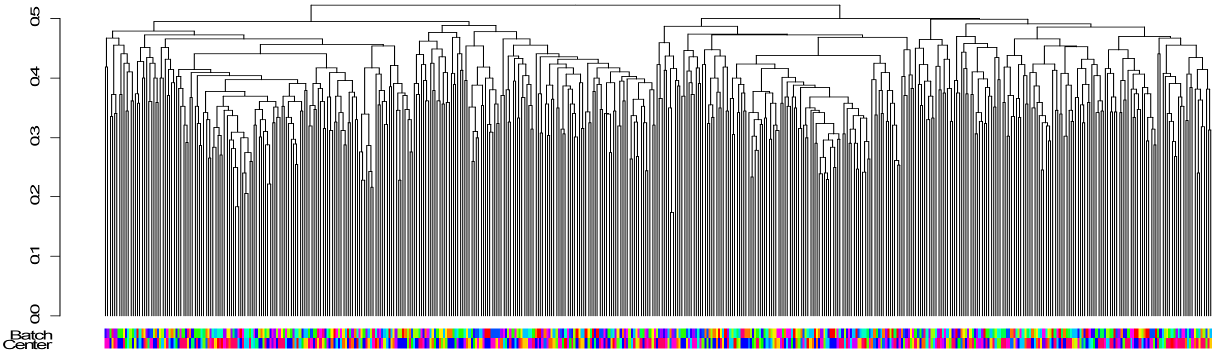


Fig S11.9: Unsupervised hierarchical clustering (1-Pearson correlation metric, average linkage algorithm)of unified mRNA expression data after adjusting by center of origin as in Figure S11.6. The top colored bar indicates batch number; the bottom colored bar indicates center (see legend above).

References:

1. Johnson, W. E., Li, C., & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).

2. Scherf, U, *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* **24**, 236-244 (2000).

3. Weinstein, J. N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343-349 (1997).