Title: Analysis of chemotherapeutic response in ovarian cancers using publically available high-throughput data.

Authors:

Jesus Gonzalez Bosquet, M.D., Ph.D. 1, Douglas C. Marchion, Ph.D. 2, HyeSook Chon, M.D. 3, Johnathan M. Lancaster, M.D., Ph.D. 3, Stephen Chanock, M.D. 4.

1 Division of Gynecologic Oncology, Department of Obstetrics & Gynecology, University of Iowa Hospitals & Clinics, Iowa City, IA.

2 Department of Women’s Oncology, Experimental Therapeutics Program, Department of Oncologic Sciences, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL

3 Department of Women’s Oncology, Gynecologic Oncology, Department of Oncologic Sciences, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL

4 Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD.

The authors disclose no potential conflicts of interest

Corresponding author:

Jesus Gonzalez Bosquet, MD, PhD

3511 Dorset Street NW,

Rochester, MN, 55901, USA.

Tel: (501) 529 5611

Email: jesus.gonzalezbosquet@gmail.com

Running title: Molecular analysis of chemo-response in ovarian cancer
Keywords: Ovarian cancer, chemotherapy response, data integration, gene signature

Word Count: 5,110
Number of Figures: 4
Number of Tables: 2
Number of Supplementary Figures: 6
Number of Supplementary Tables: 9
Abstract

A third of patients with epithelial ovarian cancer (OVCA) will not respond to standard treatment. The determination of a robust signature that predicts chemo-response could lead to the identification of molecular markers for response as well as possible clinical implementation in the future to identify patients at risk of failing therapy. This pilot study was designed to identify biological processes affecting candidate pathways associated with chemo-response and to create a robust gene signature for follow-up studies. After identifying common pathways associated with chemo-response in serous OVCA in 3 independent gene expression experiments, we assessed the biological processes associated with them using The Cancer Genome Atlas (TCGA) dataset for serous OVCA. We identified differential copy number alterations (CNA), mutations, DNA methylation, and miRNA expression between patients that responded to standard treatment and those who did not or recurred prematurely. We correlated these significant parameters to gene expression to create a signature of 422 genes associated with chemo-response. A consensus clustering of this signature identified 2 differentiated clusters with unique molecular patterns: cluster #1 was significant for cellular signaling and immune response (mainly cell–mediated); and cluster #2 was significant for pathways involving DNA damage repair and replication, cell cycle and apoptosis. Validation through consensus clustering was performed in 5 independent OVCA gene expression experiments. Genes were located in the same cluster with consistent agreement in all 5 studies (Kappa coefficient ≥ 0.6 in 4). Integrating high throughput biological data has created a robust molecular signature that predicts chemo-response in OVCA.
Introduction

Epithelial ovarian cancer (OVCA) has the highest mortality rate of all gynecologic cancers, mainly because more than 70% of patients present with advanced stage and will have disseminated intra-peritoneal disease at diagnosis. But also because between 20-30% will not respond to the initial treatment consisting of a combination of cytoreductive surgery and a platinum-based chemotherapy. Even in some patients with a complete initial response to chemotherapy, the disease will recur and eventually develop resistance to multiple drugs. Several mechanisms have been described to contribute to chemo-response including drug efflux, increased cellular glutathione levels, increased DNA repair, and drug tolerance, but the exact mechanisms are not fully defined, and there are no valid clinical biomarkers or molecular signatures that effectively predict response to chemotherapy. Understanding the underlying processes could lead to the identification of prognostic signatures, which in turn, could be used to stratify patients who are likely to develop resistance to standard chemotherapy and thus could benefit from alternative strategies.

In previous studies we have identified a series of molecular signaling pathways associated with OVCA response to chemotherapy in vitro as well as in clinical settings. We selected pathways associated with chemo-response to design a pilot study aimed to identify biological processes that influence expression. For this purpose we used publicly available data from The Cancer Genome Atlas (TCGA) ovarian cancer datasets, gene expression from ovarian cancer samples available at the Gene Expression Omnibus (GEO) repository, and OVCA data from our laboratory previously published. All biological data from these diverse sources also included information concerning response to chemotherapy. The objectives of this study were to identify biological processes affecting gene expression of candidate pathways and create a...
molecular signature associated with chemo-response. Also, we aimed to validate this molecular signature across independent gene expression microarray experiments.

Material and Methods

Sources of data

Only serous OVCA specimens were used for these comparisons. The study included the following sets:

1. OVCA cultured cell lines A2008, A2780CP, A2780S, C13, IGROV1, OV2008, OVCAR5, and T8. Cells were subjected to sequential treatment with increasing doses of cisplatin. Both cisplatin resistance and genome-wide expression changes were measured serially at baseline and after 3 and 6 cisplatin-treatment/expansion cycles. Gene expression using Affymetrix Human U133 Plus 2.0 arrays (Affymetrix, Santa Clara, CA) was uploaded at the Gene Expression Omnibus (GEO), accession number GSE23553.

2. GEO clinical datasets: 127 serous OVCA with chemo-response information from the dataset GSE23554 and dataset GSE31494 with Affymetrix Human Genome U133 Plus 2.0 and U133A arrays; 240 serous OVCA with chemo-response information from the dataset GSE98915 with Affymetrix Human Genome U133 Plus 2.0 arrays; 50 serous OVCA with chemo-response information from the dataset GSE287396 with Agilent-012097 Human 1A arrays (V2) (Agilent Technologies, Santa Clara, CA); 110 serous OVCA with chemo-response information from the dataset GSE172607 with Agilent-014850 Whole Human Genome 4x44K arrays; 185 serous OVCA from the dataset GSE267128 with Affymetrix human U133A arrays.
3. TCGA (www.cancergenome.nih.gov): Over 20 different cancer types are included in this initiative, including more than 560 serous epithelial ovarian cancers, the most common histological subtype of OVCA. TCGA comprehensive genomic information includes copy number variation, SNPs, miRNA expression, gene expression (mRNA), and DNA methylation as well as clinical and outcome information. Data from TCGA was downloaded, normalized, formatted and organized for the integration and analysis with other biological datasets in accordance with the precepts of the TCGA data sharing agreements.

All data collection and processing, including the consenting process, were performed after approval by a local institutional review board and in accord with the TCGA Human Subjects Protection and Data Access Policies, adopted by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI).

**Clinical Outcomes**

Complete response (CR) was defined as complete disappearance of all disease up to 6 months after treatment. In patients with incomplete response (IR) the disease either not responded or progressed during treatment (refractory) or recurred within 6 months of treatment completion (resistant)³,¹⁹,²⁰.

**Data analysis**

**Copy number alteration (CNA):** Samples from Agilent Human Genome CGH Microarray 244A (Agilent Technologies, Santa Clara, CA) were processed and DNA sequences were aligned to NCBI Build 36 of the human genome. Circular Binary Segmentation was used to identify regions with altered copy number in each chromosome²¹. The copy number at a
particular genomic location was computed based on the segmentation mean log ratio data. We found regions with frequent CNA among all samples by performing genomic identification of significant targets in cancer (or GISTIC) analysis\textsuperscript{22}. The significance of CNA at a particular genomic location is determined based on false discovery rate (FDR), as previously described\textsuperscript{23}. To determine the performance of our strategy, we initially proceed with the analysis of the whole sample set, and then repeated the methodology with only samples from patients that demonstrated CR (n=294) and IR (n=158). CNA in both scenarios was consistent.

**Mutation analysis:** Somatic mutation detection, calling, annotation and validation have been extensively discussed previously\textsuperscript{23}. Somatic mutation information resulting from Illumina Genome Analyzer DNA Sequencing GAIIx platform (Illumina Inc., San Diego, CA) was downloaded and formatted for analysis. Mutation information was downloaded as Level 3, or validated somatic mutations. Somatic mutation information was available from 137 samples from patients with CR and 55 with IR. For those patients, there were 6,716 unique genes presenting some type of validated somatic mutation. These included: frame shift insertions and deletions, in-frame insertions or deletions, missense, nonsense and nonstop mutations, silence, splice site and translation start site mutations. All independent significant mutated genes were correlated with gene expression of the candidate pathways genes to determine if mutated genes were associated or influenced expression of those pathways. To correlate mutated genes with gene expression we used Spearman's rank correlation test, as both variables are not completely independent one from another. We assessed statistical significance of the correlation by computing q-value for FDR (qFDR) and p-value, corrected for multiple analyses\textsuperscript{24}. 
**Gene expression and correlation with CNA:** Raw gene expression data was downloaded from the TCGA Data Portal (Level 1), extracted, loaded, and normalized with the analytical software, BRB-ArrayTools. In total 594 microarrays samples analyzed with Affymetrix HT Human Genome U133 Array: 584 coming from cancer tissue and 10 coming from ovarian normal samples. DNA sequences were aligned to NCBI Build 36 of the human genome. There were 452 arrays with clinical information about chemo-response. During the Circular Binary Segmentation analysis of the CNA, a gene-centric table is created, which contains a value for each gene covered in the genomic array. This value is assigned based on the segmentation mean log ratios. The gene-centric table is required for the correlation analysis between copy number and gene expression. Positive correlation between gene expression and CNA (increased CNA/increased gene expression, and decreased CNA/decreased gene expression) was performed using Spearman's rank correlation test, as the expression between genes is not completely independent one from another. Statistical significance was assessed with qFDR and p-value, and corrected for multiple analyses.

To construct a gene signature profile that would classify patients between complete responders and incomplete responders, we used the Class Prediction Tool of BRB-ArrayTools. Genes differentially expressed between both classes at significance level $p<0.001$ were included in the predictor and evaluated with several methods. To assess how accurately the groups are predicted by this multivariate class predictor, a cross-validated misclassification rate is computed, usually in the form of Leave-one-out cross-validation method. For TCGA gene expression analysis, 40 samples in each group (CR versus IR) had 90% power of detect differentially expressed gene, with a type 1 error of 0.001. A similar experiment, using the same software (BRB ArrayTools), same statistics, same outcomes definitions (CR versus IR), and list...
of genes from the gene signature identified in the testing set (TCGA), was designed to validate the results of the signature profile in independent available databases: GEO\# GSE9891\textsuperscript{15}, GSE28739\textsuperscript{16}, and GSE23554\textsuperscript{12}.

**Methylation analysis and correlation with gene expression:** DNA methylation data with beta-values, methylated (M) and unmethylated (U) intensities were downloaded from the TCGA Data Portal (Level 2), extracted, loaded, and normalized. In total 574 arrays samples of Illumina Infinium Human DNA Methylation 27 (Illumina Inc., San Diego, CA): 572 cancer, 2 normal. There were 453 unique DNA-methylation arrays from serous OVCA with clinical information about chemo-response. Differential DNA methylation of gene promoters was computed based on beta-values. Beta-values for each sample and locus were calculated as $(M/(M+U))$\textsuperscript{23}. Differences of gene’s beta-values between the classes (CR vs. IR) at the univariate significance level of $p<0.001$ were considered significant. Rank-based Spearman correlation was used to allow for non-linear relationships between DNA methylation and gene expression, along with p-values. To control the false discovery rate (FDR) we used the q-value for statistical significance (qFDR)\textsuperscript{24} and Bonferroni correction for multiple comparisons.

**miRNA expression analysis and its correlation with gene expression:** Raw miRNA expression data was downloaded from the TCGA Data Portal (Level 1), extracted, loaded, and normalized with the analytical software, BRB-ArrayTools. In total 595 microarrays samples of Agilent Human miRNA Microarray Rel12.0 (Agilent Technologies Inc., Santa Clara, CA): 585 cancer, 10 normal. There were 455 unique miRNA expression arrays from serous OVCA with clinical information about chemo-response\textsuperscript{23}. Differences of miRNA expression between the
classes (CR vs. IR) at the univariate significance level of p<0.05 were considered significant, as there were 619 unique miRNA tested. Rank-based Spearman correlation was used to allow for non-linear relationships between miRNA expression and gene expression, along with p-values. To control the false discovery rate (FDR) we used the q-value for statistical significance (qFDR)\(^24\) and Bonferroni correction for multiple comparisons.

**Transcription factor (TF) binding sites and their association with gene expression:**

To identify TF and their binding sites within the CNA-Correlated-Pathway (CCP) gene subset, we used publicly available search tools, The Transcription Factor Database (TRANSFAC®)\(^25\). TRANSFAC is a knowledge-base containing published data on eukaryotic transcription factors, their experimentally-proven binding sites, and regulated genes\(^26\) that utilizes a range of tools and algorithms to search DNA sequences for predicted TF binding sites through high-throughput promoter analysis. Differential gene expression between CR and IR were performed on those genes within the CCP set found to have TF binding sites by TRANSFAC database. Differences of gene expression between the classes (CR vs. IR) at the univariate significance level of p<0.01 were considered significant, as \(\approx 1,700\) genes were introduced in the analysis.

**Non-negative matrix factorization (NMF) consensus clustering of final model:** NMF is an unsupervised learning algorithm that has been shown to identify molecular patterns when applied to gene expression data. NMF detects context dependent patterns of gene expression in complex biological systems\(^15,23\). This method computes multiple k-factor factorization decompositions of the expression matrix and evaluates the stability of the solutions using a cophenetic coefficient. The final subclasses of genes were defined based on the most stable
k-factor decomposition and visual inspection of gene by gene correlation matrices (for details of the method\textsuperscript{15,21}).

**Software**

The majority of analyses were performed using **R statistical package** for statistical computing and graphics (www.r-project.org) as background, using Bioconductor packages as open source software for bioinformatics (bioconductor.org). Analysis of comparative genomic hybridization (CGH) to assess CNA and analysis of gene expression were performed using Biometric Research Branch (BRB) ArrayTools, an integrated package for the visualization and statistical analysis that utilizes Excel (Microsoft, Redmond, WA) as front end, and with tools developed in the R statistical system. BRB-ArrayTools were developed by Dr. Richard Simon and the BRB-ArrayTools development team.

**MultiExperiment Viewer** was used to implement the NMF consensus clustering and is part of the TM4 suite of tools (http://www.tm4.org/) developed in Java, an open-source, and freely available collection of tools of use to a wide range of laboratories conducting microarray experiments.

**Pathway enrichment analysis:** To identify over-represented and significant pathways among the selected list of genes we used MetaCore™ (GeneGo, Inc., Carlsbad, CA), an integrated and curated “knowledge-based” platform for pathway analysis. The p-value of significant associated pathways represents the probability that a particular gene of an experiment is placed into a pathway by chance, considering the numbers of genes in the experiment, and total genes across
all pathways.

Results:

Selection of candidate pathways associated with chemo-response

Our analysis focused on gene expression data from different sources to identify genes and pathways involved in OVCA chemo-resistance. Sources included OVCA cultured cell lines that underwent progressive higher doses of chemotherapy and tested with a chemo-sensitivity analysis measured by IC$_{50}$. Then gene expression of 48 samples was compared before and after the treatment$^{12}$. Also, we included clinical samples gathered from our own institution, 127 samples, available at the Gene Expression Omnibus repository (GEO) with accession number GSE23554$^{10,12,13}$, OVCA samples from 240 patients from the GEO database GSE9891$^{15}$, and data from TCGA, with 465 OVCA samples$^{23}$. Only serous OVCA specimens with information of response to chemotherapy were used for comparison. Chemo-response, as defined in methods, was a significant independent survival factor in TCGA dataset survival analysis (Cox proportional hazard ratio), even after control for age, stage and optimal surgery (p-value <10$^{-15}$, Supplementary Figure 1). We determined genes that were differentially expressed between patients with CR and patients with IR in the 3 clinical databases; furthermore, we evaluated genes differentially expressed in the OVCA cell lines based on in vitro chemo-sensitivity analysis. A series of genes were common between these comparison analyses: TIMP3, OLFML3, C10orf26, COPZ2, PDGFD, OMD, PKD2, SNRPA, COL8A1, GCNT1, CDK5RAP3, PRPF40A, RAB35, MAPK14, PARN, NCRNA00184, ERCC5, C13orf33, LHFP, KIAA1033, GJB1, SVEP1, TPM1, and IMPACT (Supplementary Figure 2). We introduced these common differentially
expressed genes in the pathway enrichment analysis by GeneGO (MetaCore™) and only those significant pathways that were previously described as associated with chemo response in OVCA were used as candidates pathways: the O-glycan biosynthesis pathway (p-value=7x10^{-3}), described in our previous in vitro studies of chemo-sensitivity\textsuperscript{10}, the Transport RAB1A regulation pathway (p-value=5x10^{-3}) that controls vesicle traffic within the cell\textsuperscript{27} and members of the RAS oncogene family\textsuperscript{11,28}, MAPK14 and PDGFD\textsuperscript{29} part of the MAPK signaling pathway (p-value=10^{-4}), involved in the initiation of a G2 delay after ultraviolet radiation\textsuperscript{30} and also identified in in vitro studies\textsuperscript{11}. Only these candidate pathways, previously associated with chemo-response, were used in the rest of the study.

**Identification of elements influencing expression of candidate pathways genes**

TCGA datasets for genomic copy number alterations (CNA), mutation and methylation analysis, and miRNA expression were used to identify the elements that could potentially influence the expression of our candidate pathways. Main clinical and biological data from TCGA patients included in our study are summarized in Table 1. The Transcription Factor (TRANSFAC) database\textsuperscript{25,31,32} was used to predict TF binding sites in our analysis.

**CNA analysis and correlation with gene expression:**

Of the 452 patients with clinical information about chemo-response (294 CR and 158 IR) and gene expression data from specimens collected at diagnosis, 435 underwent comparative genomic hybridization to determine CNA. Supplementary Figure 3 and Supplementary Table 1 summarize whole genome significant somatic gains and losses determined by the genomic identification of significant targets in cancer (GISTIC) analysis\textsuperscript{22}. CNA could be divided into
regional alterations, when gain or loss of genetic material affects more than 50% of the chromosomal arm and focal alterations when gains/losses are smaller. In a genome-wide assessment, there were 4 significant regional alterations with gains, 3q22.1-q29, 8p11.21-q24.3, 12p13.33-p11.21, and 20q11.21-q13.33, and 8 with losses, 4q13.3-q35.2, 6q15-q27, 8p23.3-p12, 13q11-q34, 16q12.2-q24.3, 17p13.3-q21.2), 18q12.2-q23, and 22q11.22-q13.33. Apart from the candidate pathways genes, these altered regions contained other genes that may be important or associated with unknown mechanisms of chemo-response in OVCA. Details of regional and focal CNA somatic gains and losses affecting genes of the candidate pathways are summarized in Table 2.

Of the genes included in these candidate regions with CNA (6,622 genes) only 2,364 genes showed statistically significant positive correlation with gene expression, determined by Spearman's rank correlation (Supplementary Table 2). Positive statistical correlation was defined as CNA gain with increased gene expression, or CNA loss with decreased gene expression, with a p-value <10^{-4} to account for multiple comparisons. These 2,364 genes with CNA and correlated gene expression, including 68 genes from the 3 candidate pathways, may influence or be associated with chemo-response through our candidate pathways, and will be used for the remainder of our analyses and referred as the CNA-Correlated-Pathway (CCP) gene subset.

Differential gene expression at the univariate significance level p<0.001 was used to create a 69 gene expression signature predictive of CR versus IR in the CCP subset of genes (Supplementary Figure 4 panel A, Supplementary Table 3). Validation of the gene signature was performed in previously published gene expression studies, including one from our own institution. For this validation we used the same study design used in the testing set (TCGA), a retrospective design of micro-array gene expression with additional information.
about chemo-response, the same software used in the testing set (BRB ArrayTools), same statistics (t-test at significance level $p<0.001$), same outcomes definitions (CR versus IR), and the same list of genes from the gene signature in the testing set. The gene signature profile was validated in all independent databases with a $p$-value $<0.001$ for GEO# GSE9891$^{15}$, $p$-value of 0.01 for GSE28739$^{16}$, and $p$-value of 0.02 for GSE23554$^{12}$ (Supplementary Figure 5).

**Somatic mutations and their correlation with gene expression:**

Somatic mutations in genes have been shown to influence gene expression patterns, of both individual genes and distinct pathways$^{33-35}$. Only validated mutations were used for these series of analyses$^{33}$. Mutations of 6,716 genes were available for 192 patient’s samples with information about chemo-response at TCGA data site. We performed a logistic regression analysis to determine which mutations were associated with chemo-response (CR versus IR). *KRT72, MYO5C, ODZ1, SMARCA4,* and *TP53* were statistically associated with the outcome. *SMARCA4* had mutations in 0.7% of CR samples and in 7.3% of IR samples ($p$-value=0.04); and *TP53* had mutations in 88.3% of CR samples and in 72.7% of IR samples ($p$-value=0.01). Spearman's rank correlation, with correction for multiple comparisons, was used to assess the correlation between significant mutated genes and gene expression of CCP gene subset ($p$-value and qFDR $\leq 10^{-4}$) (Supplementary Figure 4 panel B). Significant mutated genes as well as their 22 correlated genes with significant change in gene expression were included in the model and added to the significant gene signature identified from the CCP gene subset.

**Methylation analysis and its correlation with gene expression:**

Epigenetic gene regulation also may affect expression of candidate pathways by
inactivating gene function\textsuperscript{36}. As we did with mutations, an analysis of DNA methylation status was performed in 440 patients with high-throughput data and information about chemo-response available. 69 genes were differentially methylated between the CR and IR groups (Supplementary Figure 4 panel C). A list of these genes could be reviewed at Supplementary Table 4. In order to explore possible interactions between differentially methylated genes and their possible influence in the expression of our candidate pathways we performed a correlation, with correction for multiple comparisons, between these differentially methylated genes and the expression of the CCP gene subset (p-value and qFDR \(\leq 10^{-6}\)). Differentially methylated genes and their correlated expressed genes were also added to significant gene signature and mutations found previously to improve the molecular model.

**miRNA expression analysis and its correlation with gene expression:**

Gene expression is also regulated by miRNAs. As previously, differentially miRNA expression was performed in 455 patients with available high-throughput data and information about chemo-response. A heat map of the 38 miRNA differentially expressed between the CR and IR groups is represented in Supplementary Figure 4, panel D. A list of these miRNA could be reviewed at Supplementary Table 5. Possible interactions between differentially expressed miRNA and their possible influence in the expression of our candidate pathways were explored with a corrected correlation between miRNA expression and gene expression of the CCP subset (p-value and qFDR \(\leq 10^{-6}\)). Differentially expressed miRNAs and their correlated genes were also included to the model or signature.

**Transcription factor (TF) binding sites and their association with gene expression:**
Since gene expression is regulated at the transcriptional level, we examined web-based tools that utilize algorithms to search DNA sequences for predicted TF binding sites through high-throughput promoter analysis. TRANSFAC® was used to predict TF binding sites in genes of the candidate pathways associated with OVCA chemo-response\textsuperscript{31}. Of the 2,364 gene of the CCP subset, 1,772 genes were identified to have TF binding sites in their promoter area; 59 of these genes had a differential gene expression between CR and IR (Supplementary Figure 6, panel A, Supplementary Table 6). Only these 59 genes harboring TF binding sites that presented both copy number alterations and differentially gene expression were also introduced in the final model.

**External validation of final model with non-negative matrix factorization (NMF) consensus clustering:**

The combined data from all significant and correlated genes included 422 unique different genes introduced in the final model (Supplementary Table 7). A cluster analysis of these final model was performed with the non-negative matrix factorization consensus clustering, a type of unsupervised learning algorithm that has been shown to identify molecular patterns when applied to gene expression data\textsuperscript{15}. This analysis yielded two clusters with differentiated gene patterns (Figure 1, Panel A). Attempts to create models with more than 2 clusters resulted in lower cophenetic correlation coefficients (Figure 1, Panel B), and less harmonic consensus matrices (Figure 1, Panel C).

Validation of this molecular pattern with 2 defined clusters observed in the final signature model of TCGA data were performed in 5 publicly available independent OVCA gene expression datasets with the same analytical tool, NMF consensus clustering. All these
independent gene experiments also showed two differentiated clusters with the highest cophenetic correlation coefficients and the most harmonious consensus matrices\textsuperscript{12,14,15,17,18}. Comparison of these 6 sets of two clusters (TCGA and 5 validation sets) showed an unprecedented external validation of molecular signatures associated with chemo-response in OVCA (Figure 2).

We also determined if the individual genes were placed by the NMF consensus clustering within the same cluster in all databases used for validation. The level of agreement measured with kappa coefficient was considered ‘good or substantial’ for 3 of them (0.61-0.80)\textsuperscript{12,14,17}, ‘almost perfect’ for another one (0.81-1)\textsuperscript{15} and ‘moderate’ in only one of them (0.41-0.60)\textsuperscript{18}(Figure 2, and Supplementary Figure 7).

**Pathway enrichment analysis:**

To identify which pathways and biological processes were overrepresented in both gene clusters identified within the final gene model, further analysis was conducted with MetaCore\textsuperscript{TM} and clusterProfiler\textsuperscript{37}, from the R statistical package, which mines the KEGG database (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg). The correlation matrix of each cluster denotes overrepresented predominant pathways (Figure 3). Cluster #1 showed a significant representation of cellular signaling and immune response (mainly cell–mediated) pathways, but also several types of metabolic pathways (Figure 4, and Supplementary Table 8). Three fourths (75\%) of all signaling and metabolic pathways genes in cluster #1 were overexpressed in the IR tumors with respect to the CR samples, with the general perception that IR tumors were engaged in higher metabolic rates through external and internal stimuli. Cluster #2 was significant for pathways involving DNA damage repair and replication as well as cell
cycle and apoptosis, all of them with strong influence by mutated TP53. Cluster #2 also presented significant pathways related to cancer and cytoskeleton configuration and structure (Figure 4, and Supplementary Table 9). Two thirds of all cell cycle and DNA repair genes presented elevated expression in CR tumors when compared to IR tumors, probably driven by mutated key elements of these pathways, like TP53, and increased expression of transcription factors, like ATF6B, CRTCl, E2F1, SIN3B and NFIX.

Discussion:

Patients suffering from platinum-refractory or resistant ovarian cancer have a median overall survival around 12-13 months. Furthermore, these patients become resistant to multiple drugs early on the course of treatment of their disease and thus, it is challenging to establish efficacious treatment strategies. The objective of our study was to identify biological processes associated with chemo-response so we may, in the future, identify patients at risk for standard chemotherapy resistance, thus eligible for novel strategies. Also, knowledge of mechanisms involved in chemo-response may help design new strategies as molecular-targeted therapy is gaining traction. Our study determined chemo-response as the most significant independent clinical factor for survival (p-value <10^{-15}) in the TCGA database, agreeing with daily clinical practice and published clinical trials.

Despite increasing knowledge about mechanisms of chemo-response in tumor cells there are no valid clinical biomarkers or molecular signatures that could effectively predict response to chemotherapy. Initial analysis of gene profiling aiming to identify functional processes associated with chemo-response in OVCA had showed little overlap between studies looking for expression signatures or pathways associated with response to therapy. Previously, several
groups have used integration of ‘omics data in OVCA to predict other clinical outcomes\textsuperscript{41-43}. By integrating the comprehensive characterization of TCGA data, namely, CNA, gene mutations, DNA methylation, miRNA expression, and TF binding sites location, into an analytical framework for gene expression, we have created a robust molecular signature that predicts chemo-response in OVCA. This model, with 2 clusters involving previous known mechanisms of chemo-response\textsuperscript{9}, was the most robust in TCGA database; but what is even more important and unprecedented in our study is the external validation of the TCGA model with other 5 independent gene expression studies of OVCA\textsuperscript{12,14,15,17,18}. These findings demonstrate consistency of this signature across diverse studies and platforms that we believe is due to the selection of micro-array experiments with the same tumor type of OVCA (serous), statistical design of analyses, and adequately powered. Furthermore, both clusters in the model included the same individual genes with substantial agreement in all but one of the 5 independent gene expression sets, where the agreement was moderate (Figure 2). We think that adding elements to the model that did not result exclusively from the differential gene expression of micro-arrays (CNA, mutations, methylation, miRNA, TF binding sites) added stability to the molecular signature and provided enough range to overcome validation difficulties observed by gene expression experiments alone due to tumor heterogeneity\textsuperscript{8,9}.

Most genes included in our molecular signature for chemo-response are drawn from cellular functions previously associated with response to chemotherapy\textsuperscript{9}. These biological processes include cell signaling pathways, immune response pathways, and several types of metabolic pathways that are involved in DNA damage repair and replication, cell cycle and apoptosis, all of them also have been associated with cancer transformation and proliferation\textsuperscript{44}. In ovarian cancers, signaling transduction cascades from the pathways $PI3K/AKT/mTOR$ and
Ras/Raf/MEK/MAPK/ERK (with representation in our molecular model, Supplementary Table 7) may result in diverse effects, including cell proliferation, invasion, angiogenesis, apoptosis evasion, and response to chemotherapy\textsuperscript{44,45}. The MAPK signaling pathway is also connected to the Ras pathway (which includes \textit{PAK4}), that also regulates cell morphology, cytoskeletal organization, and cell proliferation and migration; \textit{PAK4} can also function as an anti-apoptotic protein\textsuperscript{46}. PAK proteins are critical effectors that link Rho GTPases to cytoskeleton reorganization and nuclear signaling. Both \textit{PAK4} and \textit{RHOT1} are included in cluster #2. The Ras gene family (which \textit{RASA1} is part) encodes membrane-associated, guanine nucleotide-binding proteins that are involved also in the control of cellular proliferation and differentiation, and have a weak intrinsic GTPase activity, effectors of \textit{Ras} oncogene action\textsuperscript{47}. MAP kinases may also have a role in early gene expression by modifying the chromatin environment of target genes\textsuperscript{48}, an action regulated through phosphorylation of various substrates, including transcription factors and chromatin constituents. \textit{NCAPG}, a component of the condensin complex that is required for both interphase and mitotic condensation, is present in cluster #2 of the chemo-response model. Animal models with condensin mutations, DNA damage induced by ultraviolet radiation is not repaired and cells arrested by hydroxyurea do not recover\textsuperscript{49}. In our gene signature, it is notable that a set of genes map to DNA repair pathway (mainly through homologous recombination), like \textit{RAD52}, and elements of the PARP family, like \textit{PARP12}. PARPs inhibitors have been proven to be efficacious in the treatment of OVCA in carriers of \textit{BRCA1} or \textit{BRCA2} mutations\textsuperscript{50}. Also notable are the results of bevacizumab, a humanized antibody against VEGF, in the adjuvant treatment of OVCA\textsuperscript{51}. \textit{PDGFB}, a component of the VEGF signaling pathway, is present in cluster #1, with other numerous components of cell signaling pathways. With all these interconnections between signaling pathways, DNA damage repair, and cell cycle, alternative
strategies to standard therapies may have to involve a combination of cross-specific drugs to avoid by-pass of the blocked path\textsuperscript{52}.

A strength of our study is that our discovery set has been based on a large genomic dataset, which has high benchmarks for quality control and processing (TCGA). The large sample size confers adequate power to detect important patterns while at the same time, it overcomes possible bias introduced by outliers; moreover, it permits better selection of the histological type (serous) and outcome of interest (chemo-response) to improve homogeneity. We believe that those are major factors influencing in the important external validation of the molecular signature in 5 independent gene expression experiments, which is unprecedented in micro-array analysis\textsuperscript{40}. Other major factor influencing a significant validation is the integration of diverse biological data, other than gene expression, in the final model. For the validation process we used two OVCA datasets that were used initially to identify candidate pathways\textsuperscript{12,15}. These datasets had the closest clinical information, including chemo-response, and study design to TCGA. To avoid data over fitting in the validation process, though, we added 3 independent gene expression experiments not used before, despite presenting minor differences in study design and platform content\textsuperscript{14,17,18}.

A major limitation of this study is its retrospective nature. Although the outcome of interest (chemo-response) was collected, we had to reformat it to fit our definition of complete and incomplete response. Only patients with information about the outcome variable were finally introduced in the analysis. There was also extensive information about the specifics of patient treatment after diagnosis, including types of drugs received. Nearly all patients received standard treatment, with over 99% of them getting a platinum-based chemotherapy (Table 1). Initial treatment with platinum may select for some of molecular patterns observed in our signature,
like cell cycle or DNA repair pathways, due to the DNA adducts induced by platinum\textsuperscript{53}. Consequently, our molecular model would be most applicable to OVCA that would receive initial platinum-based treatment, and less suited for other scenarios with different initial therapeutic strategies. Similarly, it would be interesting to investigate if the signature is valid in other solid tumors treated with comparable platinum-based combinations.

In summary, integration of diverse biological data into gene expression may strengthen gene signature models for prediction of OVCA chemo-response. Robust validation over 5 independent publicly available gene expression experiments supports these findings. Nonetheless, this model has to be validated \textit{in vitro} or \textit{vivo} models before can be tested clinically. It would be very interesting to apply this molecular signature model to other solid tumors that usually receive platinum-based therapy.
Acknowledgements:

We would like to thank ‘TCGA Research Network’ for generating, curating and providing high quality biological and clinical data.
Grant Support:

The research was supported in part by the U.S. Army Medical Research and Materiel Command, in support of the proposal, ‘National Functional Genomics Center’ (award number: W81XWH-08-2-0101) at the H. Lee Moffitt Cancer Center & Research Institute, University of South Florida, Tampa, Florida; and by the ‘St. Louis Ovarian Cancer Awareness Research Grant’ from the 2012-2013 Foundation for Women’s Cancer grants.
References:


**Table 1:** Clinical variables were associated to gene expression analysis primarily through the ID number of each patient. The same ID number was used to correlate gene expression data with other biological information: gene mutation, copy number alteration (CNA), DNA methylation and miRNA expression. CR: Complete responders. IR: Incomplete responders.

**Table 2:** Only those genomic regions with gene copy number gain/loss that include genes of the candidate pathways were included. CNA: Copy number alterations.
**Figure 1:** Consensus matrices of the final signature model.

**Panel A.** Consensus matrix of the final model including 422 genes, suggesting and optimal and more robust result, with limited overlap between clusters, for clustering with k-factor = 2.

**Panel B.** Cophenetic correlation coefficients representation, with an optimal result for k=2.

**Panel C.** Other consensus matrices for k=3 to k=6 showing less harmonious models.

**Figure 2:** Validation of the NMF consensus clustering in independent publicly available databases. In all of them we used 422 genes included the final model. All of them also showed robust clustering with k = 2, and very limited overlap between clusters. Other consensus matrices for k=3 to k=6 were less harmonious (not shown). All cophenetic correlation coefficients had optimal results for k=2 (not shown). The level of agreement (kappa coefficient) for placement of an individual gene in the same cluster than in TCGA analysis are also shown for each dataset.

**Panel A.** GSE9891 database\(^{15}\).

**Panel B.** GSE3149 database\(^{14}\).

**Panel C.** GSE26712 database\(^{18}\).

**Panel D.** GSE23554 database\(^{12}\).

**Panel E.** GSE17260 database\(^{17}\).

**Figure 3:** Gene expression correlation matrix of the 422 and association with pathway analysis.

Pathway analysis of both clusters showed an overrepresentation of cellular signaling and immune response pathways in cluster #1 (blue), and DNA repair/replication within the context of cell cycle pathways in cluster #2 (yellow).
Figure 4: Gene expression profile of genes included in cluster #1 and #2 and their association with significant pathways identified with enrichment pathway analyses. Cluster 1: significantly associated with cell signaling (blue), immune response (yellow) and a variety of metabolic pathways (purple). Cluster 2: significantly associated with DNA repair/replication and cell cycle (green), pathways in cancer (red) and a cell adhesion/cytoskeleton pathways (maroon).
Table 1: Clinical and biological data from TCGA patients included in the study

<table>
<thead>
<tr>
<th>Clinical information of patients with gene expression</th>
<th>CR</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>294</td>
<td>158</td>
</tr>
<tr>
<td>Age (Avg.)</td>
<td>58.7</td>
<td>59.6</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Grade 2</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>Grade 3</td>
<td>247</td>
<td>135</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Stage II</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Stage III</td>
<td>226</td>
<td>123</td>
</tr>
<tr>
<td>Stage IV</td>
<td>39</td>
<td>29</td>
</tr>
<tr>
<td>Surgical outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal (&lt;1 cm residual)</td>
<td>185</td>
<td>89</td>
</tr>
<tr>
<td>Suboptimal (&gt;1 cm residual)</td>
<td>51</td>
<td>57</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contained Platinum</td>
<td>282</td>
<td>147</td>
</tr>
<tr>
<td>Not contained Platinum</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Contained Taxane</td>
<td>159</td>
<td>77</td>
</tr>
<tr>
<td>All</td>
<td>283</td>
<td>150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological information (number of samples)</th>
<th>CR</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGH</td>
<td>283</td>
<td>152</td>
</tr>
<tr>
<td>Mutation analysis</td>
<td>137</td>
<td>55</td>
</tr>
<tr>
<td>Gene expression arrays</td>
<td>294</td>
<td>158</td>
</tr>
<tr>
<td>DNA methylation arrays</td>
<td>287</td>
<td>153</td>
</tr>
<tr>
<td>miRNA expression arrays</td>
<td>295</td>
<td>160</td>
</tr>
</tbody>
</table>
Table 2: CNA affecting candidate pathway genes

<table>
<thead>
<tr>
<th>Focal CNA gains</th>
<th>chromosome</th>
<th>arm</th>
<th>start</th>
<th>size</th>
<th>size/arm (%)</th>
<th>cytoband</th>
<th>number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>q</td>
<td>143809389</td>
<td>13468627</td>
<td>10.98</td>
<td>1q21.1-q23.1</td>
<td>338</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>q</td>
<td>157287924</td>
<td>3811143</td>
<td>3.11</td>
<td>1q23.1-q23.3</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>q</td>
<td>154605621</td>
<td>1322</td>
<td>0.00</td>
<td>2q24.1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>p</td>
<td>31563829</td>
<td>1041471</td>
<td>1.72</td>
<td>6p21.33-p21.32</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>p</td>
<td>9797279</td>
<td>7685593</td>
<td>12.70</td>
<td>6p24.3-p22.3</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>q</td>
<td>128194352</td>
<td>30626911</td>
<td>30.66</td>
<td>7q32.1-q36.3</td>
<td>265</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>q</td>
<td>76074403</td>
<td>2562103</td>
<td>3.16</td>
<td>11q13.5-q14.1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>p</td>
<td>9457512</td>
<td>10236513</td>
<td>35.92</td>
<td>19p13.2-p13.11</td>
<td>323</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>q</td>
<td>32545047</td>
<td>13265170</td>
<td>37.37</td>
<td>19q12-q13.2</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>p</td>
<td>12719</td>
<td>4947364</td>
<td>18.26</td>
<td>20p13</td>
<td>93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regional CNA gains</th>
<th>chromosome</th>
<th>arm</th>
<th>start</th>
<th>size</th>
<th>size/arm (%)</th>
<th>cytoband</th>
<th>number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>q</td>
<td>133198618</td>
<td>66130974</td>
<td>61.06</td>
<td>3q22.1-q29</td>
<td>386</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>p</td>
<td>41244139</td>
<td>105020707</td>
<td>232.35</td>
<td>8p11.21-q24.3</td>
<td>492</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>p</td>
<td>33854</td>
<td>32885882</td>
<td>92.90</td>
<td>12p13.33-p11.21</td>
<td>314</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>q</td>
<td>29297270</td>
<td>33122264</td>
<td>94.91</td>
<td>20q11.21-q13.33</td>
<td>411</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Focal CNA losses</th>
<th>chromosome</th>
<th>arm</th>
<th>start</th>
<th>size</th>
<th>size/arm (%)</th>
<th>cytoband</th>
<th>number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>q</td>
<td>154605621</td>
<td>1322</td>
<td>0.00</td>
<td>2q24.1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>q</td>
<td>51632554</td>
<td>47776009</td>
<td>35.84</td>
<td>5q11.2-q21.1</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>p</td>
<td>738240</td>
<td>4983</td>
<td>0.01</td>
<td>12p13.33</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>q</td>
<td>18362555</td>
<td>26590605</td>
<td>32.04</td>
<td>15q11.1-q21.1</td>
<td>311</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regional CNA losses</th>
<th>chromosome</th>
<th>arm</th>
<th>start</th>
<th>size</th>
<th>size/arm (%)</th>
<th>cytoband</th>
<th>number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>q</td>
<td>72720228</td>
<td>118433385</td>
<td>84.47</td>
<td>4q13.3-q35.2</td>
<td>479</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>q</td>
<td>90029071</td>
<td>80724035</td>
<td>73.05</td>
<td>6q15-q27</td>
<td>388</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>p</td>
<td>151472</td>
<td>37851200</td>
<td>83.74</td>
<td>8p23.3-p12</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>q</td>
<td>18194544</td>
<td>95928796</td>
<td>97.89</td>
<td>13q11-q34</td>
<td>382</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>q</td>
<td>54405667</td>
<td>34284904</td>
<td>67.49</td>
<td>16q12.2-q24.3</td>
<td>354</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>p</td>
<td>1909371</td>
<td>34801218</td>
<td>156.76</td>
<td>17p13.3-q21.2</td>
<td>614</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>q</td>
<td>33659862</td>
<td>42456164</td>
<td>70.88</td>
<td>18q12.2-q23</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>q</td>
<td>21579649</td>
<td>27991419</td>
<td>73.28</td>
<td>22q11.22-q13.33</td>
<td>402</td>
</tr>
</tbody>
</table>
A. GSE9891 (Tothill) Kappa: 0.86

B. GSE3149 (Bild) Kappa: 0.75

C. GSE26712 (Bonome) Kappa: 0.45

D. GSE23554 (Marchion) Kappa: 0.66

E. GSE17260 (Yoshihara) Kappa: 0.60
Fig. 3

Cellular signaling
Immune response

Cell cycle
DNA damage repair/replication
Fig. 4

Cluster 1

Cluster 2

- Cell signaling pathway
- Immune response pathway
- Metabolism pathway
- DNA repair/replication and Cell cycle
- Pathways in cancer
- Cell adhesion/Regulation of cytoskeleton
Analysis of chemotherapeutic response in ovarian cancers using publically available high-throughput data

Jesus Gonzalez Bosquet, Douglas C Marchion, HyeSook Chon, et al.

Cancer Res  Published OnlineFirst May 21, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-14-0186

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2014/07/21/0008-5472.CAN-14-0186.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://cancerres.aacrjournals.org/content/early/2014/05/21/0008-5472.CAN-14-0186. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.