Analysis of Chemotherapeutic Response in Ovarian Cancers Using Publicly Available High-Throughput Data

Jesus Gonzalez Bosquet1, Douglas C. Marchion2, HyeSook Chon3, Johnathan M. Lancaster3, and Stephen Chanock4

Abstract

A third of patients with epithelial ovarian cancer (OVCA) will not respond to standard treatment. The determination of a robust signature that predicts chemoresponse 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 biologic processes affecting candidate pathways associated with chemoresponse and to create a robust gene signature for follow-up studies. After identifying common pathways associated with chemoresponse in serous OVCA in three independent gene-expression experiments, we assessed the biologic 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 with gene expression to create a signature of 422 genes associated with chemoresponse. A consensus clustering of this signature identified two 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 five independent OVCA gene-expression experiments. Genes were located in the same cluster with consistent agreement in all five studies (κ coefficient ≥ 0.6 in 4). Integrating high-throughput biologic data have created a robust molecular signature that predicts chemoresponse in OVCA. Cancer Res; 74(14); 3902–12. ©2014 AACR.

Introduction

Epithelial ovarian cancer (OVCA) has the highest mortality rate of all gynecologic cancers (1), mainly because more than 70% of patients present with advanced stage and will have disseminated intraperitoneal disease at diagnosis (2). But also because between 20% and 30% will not respond to the initial treatment consisting of a combination of cytoreductive surgery and a platinum-based chemotherapy (3). Even in some patients with a complete initial response to chemotherapy, the disease will recur and eventually develop resistance to multiple drugs (4). Several mechanisms have been described to contribute to chemoresponse, including drug efflux, increased cellular glutathione levels, increased DNA repair, and drug tolerance, but the exact mechanisms are not fully defined (5–7), and there are no valid clinical biomarkers or molecular signatures that effectively predict response to chemotherapy (8, 9). 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 (9).

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 (10–13). We selected pathways associated with chemoresponse to design a pilot study aimed to identify biologic 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 biologic data from these diverse sources also included information about response to chemotherapy. The objectives of this study were to identify biologic processes affecting gene expression of candidate pathways and create a molecular signature associated with chemoresponse. Also, we aimed to validate this molecular signature across independent gene-expression microarray experiments.
Materials 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 (12). 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 three and six cisplatin treatment/expansion cycles. Gene expression using Affymetrix Human U133 Plus 2.0 arrays (Affymetrix) was uploaded at the GEO, accession number GSE23553.

2. GEO clinical datasets: 127 serous OVCA with chemoresponse information from the datasets GSE23554 (12) and GSE3149 (14) with Affymetrix Human Genome U133 Plus 2.0 and U133A arrays; 240 serous OVCA with chemoresponse information from the dataset GSE9891 (15) with Affymetrix Human Genome U133 Plus 2.0 arrays; 50 serous OVCA with chemoresponse information from the dataset GSE28739 (16) with Agilent-012097 Human 1A arrays (V2; Agilent Technologies); 110 serous OVCA with chemoresponse information from the dataset GSE17260 (17) with Agilent-014850 Whole Human Genome 44 K arrays; 185 serous OVCA from the dataset GSE28712 (18) with Affymetrix human U133A arrays.

3. TCGA (www.cancergenome.nih.gov): More than 20 different cancer types are included in this initiative, including more than 560 serous OVCAs, the most common histologic 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 were downloaded, normalized, formatted, and organized for the integration and analysis with other biologic 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.

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; refs. 3, 19, 20).

Data analysis

Copy-number alteration. Samples from Agilent Human Genome CGH Microarray 24A (Agilent Technologies) 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 chromosome (21). The copy number at a particular genomic location was computed on the basis of the segmentation mean log ratio data. We found regions with frequent CNA among all samples by performing genomic identification of significant targets in cancer (GISTIC) analysis (22). The significance of CNA at a particular genomic location is determined on the basis of false discovery rate (FDR), as previously described (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 (23). Somatic mutation information resulting from Illumina Genome Analyzer DNA Sequencing GAIIx platform (Illumina Inc.) 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, nonsense, deletion, 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 whether mutated genes were associated or influenced expression of those pathways. To correlate mutated genes with gene expression, we used the Spearman 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 (24). Gene expression and correlation with CNA. Raw gene-expression data were 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 available.

We found regions with frequent CNA among all samples by performing genomic identification of significant targets in cancer (GISTIC) analysis (22). The significance of CNA at a particular genomic location is determined on the basis of false discovery rate (FDR), as previously described (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 (23). Somatic mutation information resulting from Illumina Genome Analyzer DNA Sequencing GAIIx platform (Illumina Inc.) 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, nonsense, deletion, 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 whether mutated genes were associated or influenced expression of those pathways. To correlate mutated genes with gene expression, we used the Spearman 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 (24).

Gene expression and correlation with CNA. Raw gene-expression data were 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 chemoresponse. 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 on the basis of 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 the Spearman rank correlation test, as the expression between genes is not completely independent one from another. Statistical significance was assessed with qFDR and P value, corrected for multiple analyses (24).

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-validation misclassification rate is computed, usually in the form of the leave-one-out cross-validation method. For TCGA gene-expression analysis, 40 samples in each group (CR vs. IR) had 90% power of detect the differentially expressed gene, with a type I error of 0.001. A similar experiment, using the same software (BRB-ArrayTools), same statistics, same outcomes definitions (CR vs. 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# GSE8981, GSE28739, and GSE23554.

**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.): 572 cancer, 2 normal. There were 453 unique DNA-methylation arrays from serous OVCA with clinical information about chemoresponse. Differential DNA methylation of gene promoters was computed on the basis of \( \beta \) values. \( \beta \) values for each sample and locus were calculated as \([M/(M+U)]; \text{ref. 23}\). Differences of gene \( \beta \) 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 nonlinear relationships between DNA methylation and gene expression, along with \( P \) values. To control the FDR, we used the \( q \) value for statistical significance (qFDR; ref. 24) and Bonferroni correction for multiple comparisons.

**miRNA expression analysis and its correlation with gene expression.** Raw miRNA expression data were 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.): 585 cancer, 10 normal. There were 455 unique miRNA expression arrays from serous OVCA with clinical information about chemoresponse (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 nonlinear relationships between miRNA expression and gene expression, along with \( P \) values. To control the FDR, we used the \( q \) value for statistical significance (qFDR; ref. 24) and Bonferroni correction for multiple comparisons.

**Transcription factor–binding sites and their association with gene expression.** To identify transcription factor (TF) and their binding sites within the CNA-correlated pathway (CCP) gene subset, we used publicly available search tools, The Transcription Factor Database (TRANSFAC; ref. 25). TRANSFAC is a knowledge-base containing published data on eukaryotic TFs, their experimentally proven binding sites, and regulated genes (26) that uses 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 was 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.

**Nonnegative matrix factorization consensus clustering of the final model.** Nonnegative matrix factorization (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 biologic 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 on the basis of the most stable \( k \)-factor decomposition and visual inspection of gene-by-gene correlation matrices (for details of the method; refs. 15, 23).

**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 uses Excel (Microsoft) as front end, and with tools developed in the R statistical system. BRB ArrayTools were developed by Dr. Richard Simon (Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, MD) 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 overrepresented and significant pathways among the selected list of genes, we used MetaCore (GeneGo, Inc.), 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 chemoresponse.**

Our analysis focused on gene-expression data from different sources to identify genes and pathways involved in OVCA chemoresistance. Sources included OVCA cultured cell lines that underwent progressive higher doses of chemotherapy and tested with a chemosensitivity analysis measured by IC\textsubscript{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 GEO repository 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. Chemoresponse, as defined in methods, was a significant independent survival factor in TCGA dataset survival analysis (Cox proportional HR), even after control for age, stage, and optimal surgery ($P < 10^{-15}$; Supplementary Fig. S1). We determined genes that were differentially expressed between patients with CR and patients with IR in the three clinical databases; furthermore, we evaluated genes differentially expressed in the OVCA cell lines based on in vitro chemosensitivity 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, PDGFD, OMD, PKD2, SNRPA, COL8A1, GCNT1, CDK5RAP3, PRPF40A, RAB35, MAPK14, PARN, NCRNA00184, ERCC5, C13orf33, LHFP, KIAA1033, GJB1, SVEP1, TPM1, and IMPACT (Supplementary Fig. S2). 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 chemoresponse in OVCA were used as candidates pathways: the O-glycan biosynthesis pathway ($P = 7 \times 10^{-3}$), described in our previous in vitro studies of chemosensitivity (10), the transport RAB1A regulation pathway ($P = 5 \times 10^{-3}$) that controls vesicle traffic within the cell (27) and members of the RAS oncogene family (11, 28), MAPK14 and PDGFD (29) part of the MAPK signaling pathway ($P = 10^{-4}$), involved in the initiation of a G2 delay after UV radiation (30) and also identified in in vitro studies (11). Only these candidate pathways, previously associated with chemoresponse, 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 biologic data from TCGA patients included in our study are summarized in Table 1. The TRANSFAC database (25, 31, 32) was used to predict TF-binding sites in our analysis.

Table 1. Clinical and biologic data from TCGA patients included in the study

<table>
<thead>
<tr>
<th></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 (average)</td>
<td>58.7</td>
<td>59.6</td>
</tr>
<tr>
<td>Grades</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>Stages</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>
<tr>
<td>Biologic information (number of samples)</td>
<td></td>
<td></td>
</tr>
<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>

NOTE: 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 biologic information: gene mutation, CNA, DNA methylation, and miRNA expression. Abbreviation: N.S., nonsignificant.
CNA analysis and correlation with gene expression

Of the 452 patients with clinical information about chemoresistance (294 CR and 158 IR) and gene-expression data from specimens collected at diagnosis, 435 underwent CGH to determine CNA. Supplementary Figure S3 and Supplementary Table S1 summarize whole genome significant somatic gains and losses determined by the GISTIC analysis (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 (23). In a genome-wide assessment, there were four significant regional alterations with gains, 3q22.1-q29, 8p11.21-q24.3, 12p13.33-p11.21, and 20q11.21-q13.33, and eight with losses, 4q13.3-q35.2, 6p15.1-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 chemoresistance 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 rank correlation (Supplementary Table S2). Positive statistical correlation was defined as CNA gain with increased gene expression, or CNA loss with decreased gene expression, with a P value of < 10^{-4} to account for multiple comparisons (23). These 2,364 genes with CNA and correlated gene expression, including 68 genes from the three candidate pathways, may influence or be associated with chemoresistance through our candidate pathways, and will be used for the remainder of our analyses and referred as the CCP gene subset.

Differential gene expression at the univariate significance level P < 0.001 was used to create a 69 gene-expression

<table>
<thead>
<tr>
<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><strong>Copy-number gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Focal CNA gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>1</td>
<td>q</td>
<td>157287924</td>
<td>3811143</td>
<td>3.11</td>
<td>1q21.1-q23.3</td>
<td>89</td>
</tr>
<tr>
<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>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>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>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>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>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>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>20</td>
<td>p</td>
<td>12719</td>
<td>4947364</td>
<td>18.26</td>
<td>20q11.21-q13.33</td>
<td>93</td>
</tr>
<tr>
<td><strong>Regional CNA gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>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>
<tr>
<td><strong>Copy-number losses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Focal CNA losses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>5</td>
<td>q</td>
<td>51632554</td>
<td>47760099</td>
<td>35.84</td>
<td>5q11.2-q21.1</td>
<td>225</td>
</tr>
<tr>
<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>15</td>
<td>q</td>
<td>18362555</td>
<td>26596065</td>
<td>32.04</td>
<td>15q11.1-q12.1</td>
<td>311</td>
</tr>
<tr>
<td><strong>Regional CNA losses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>13</td>
<td>q</td>
<td>18194544</td>
<td>9528796</td>
<td>97.89</td>
<td>13q11-q34</td>
<td>382</td>
</tr>
<tr>
<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>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>18</td>
<td>q</td>
<td>33659682</td>
<td>42456164</td>
<td>70.88</td>
<td>18q12.2-q23</td>
<td>151</td>
</tr>
<tr>
<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>

NOTE: Only those genomic regions with gene copy-number gain/loss that include genes of the candidate pathways were included.
signature predictive of CR versus IR in the CCP subset of genes (Supplementary Fig. S4A and Supplementary Table S3). Validation of the gene signature was performed in previously published gene-expression studies, including one from our own institution (12, 15, 16). For this validation, we used the same study design used in the testing set (TCGA), a retrospective design of microarray gene expression with additional information about chemoresponse, the same software used in the testing set (BRB ArrayTools), same statistics (t test at significance level \(P < 0.001\)), same outcomes definitions (CR vs. 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 of \(<0.001\) for GEO\# GSE9891 (15), a \(P\) value of 0.01 for GSE28739 (16), and a \(P\) value of 0.02 for GSE23554 (Supplementary Fig. S5; ref. 12).

**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 (23). Mutations of 6,716 genes were available for 192 patient samples with information about chemoresponse at the TCGA data site. We performed a logistic regression analysis to determine which mutations were associated with chemoresponse (CR vs. IR). \(KRT72\), \(MYD51\), \(ODZ1\), \(SMARCA4\), and \(TP53\) were statistically associated with the outcome. \(SMARCA4\) had mutations in 0.7% of CR samples and in 7.3% of IR samples (\(P = 0.04\)); and \(TP53\) had mutations in 88.3% of CR samples and in 72.7% of IR samples (\(P = 0.01\)). Spearman rank correlation, with correction for multiple comparisons, was used to assess the correlation between significant mutated genes and gene expression of the CCP gene subset (\(P\) value and qFDR \(< 10^{-6}\); Supplementary Fig. S4B). 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 (36). As we did with mutations, an analysis of DNA methylation status was performed in 440 patients with high-throughput data and information about chemoresponse available. Sixty-nine genes were differentially methylated between the CR and IR groups (Supplementary Fig. S4C). A list of these genes could be reviewed at Supplementary Table S4. 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 \(< 10^{-6}\)). 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 chemoresponse. A heatmap of the 38 miRNA differentially expressed between the CR and IR groups is represented in Supplementary Fig. S4D. A list of these miRNA could be reviewed at Supplementary Table S5. 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 \(< 10^{-6}\)). Differentially expressed miRNAs and their correlated genes were also included to the model or signature.

**TF-binding sites and their association with gene expression**

Because gene expression is regulated at the transcriptional level, we examined web-based tools that use 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 chemoresponse (31). Of the 2,364 genes 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 Fig. S6A and Supplementary Table S6). Only these 59 genes harboring TF-binding sites that presented both CNAs and differentially gene expression were also introduced in the final model.

**External validation of the final model with NMF consensus clustering**

The combined data from all significant and correlated genes included 422 unique different genes introduced in the final model (Supplementary Table S7). A cluster analysis of these final models was performed with the NMF consensus clustering, a type of unsupervised learning algorithm that has been shown to identify molecular patterns when applied to gene-expression data (15). This analysis yielded two clusters with differentiated gene patterns (Fig. 1A). Attempts to create models with more than two clusters resulted in lower cophenetic correlation coefficients (Fig. 1B), and less harmonic consensus matrices (Fig. 1C).

Validation of this molecular pattern with two defined clusters observed in the final signature model of TCGA data was performed in five 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 (12, 14, 15, 17, 18). Comparison of these six sets of two clusters (TCGA and five validation sets) showed an unprecedented external validation of molecular signatures associated with chemoresponse in OVCA (Fig. 2).

We also determined whether 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 $k$ coefficient was considered "good or substantial" for three of them (0.61–0.80; ref. 12, 14, 17), "almost perfect" for another one (0.81–1; ref. 15), and "moderate" in only one of them (0.41–0.60; Fig. 2; Supplementary Fig. S7; ref. 18).

Pathway enrichment analysis
To identify which pathways and biologic processes were overrepresented in both gene clusters identified within the final gene model, further analysis was conducted with

Figure 1. Consensus matrices of the final signature model. A, consensus matrix of the final model, including 422 genes, suggesting an optimal and more robust result, with limited overlap between clusters, for clustering with $k$-factor = 2. B, cophenetic correlation coefficients representation, with an optimal result for $k = 2$. C, other consensus matrices for $k = 3$ to 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 including 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 6 were less harmonious (not shown). All cophenetic correlation coefficients had optimal results for $k = 2$ (not shown). The level of agreement ($k$ coefficient) for placement of an individual gene in the same cluster than in TCGA analysis is also shown for each dataset. A, GSE9891 database (15). B, GSE3149 database (14). C, GSE26712 database (18). D, GSE23554 database (12). E, GSE17260 database (17).
these pathways, like Table S9). Two thirds of all cell cycle and DNA repair genes

ATF6B

presented significant representation of cellular signaling and immune response (mainly cell-mediated) pathways, but also several types of metabolic pathways (Fig. 4; Supplementary Table S8). 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 (Fig. 4; Supplementary Table S9). Two thirds of all cell cycle and DNA repair genes presented elevated expression in CR tumors when compared with IR tumors, probably driven by mutated key elements of these pathways, like TP53, and increased expression of TFs, like ATF6B, CRTC1, E2F1, SIN3B, and NFIX.

Discussion

Patients suffering from platinum-refractory or -resistant ovarian cancer have a median overall survival of around 12 to 13 months (3). 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 (38). The objective of our study was to identify biologic processes associated with chemoresponse 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 chemoresponse may help design new strategies as molecular-targeted therapy is gaining traction (39). Our study determined chemoresponse as the most significant independent clinical factor for survival ($P < 10^{-15}$) in the TCGA database, agreeing with daily clinical practice and published clinical trials (20).

Despite increasing knowledge about mechanisms of chemoresponse in tumor cells there are no valid clinical biomarkers or molecular signatures that could effectively predict response to chemotherapy (8, 9). Initial analysis of gene profiling aiming to identify functional processes associated with chemoresponse in OVCA had showed little overlap between studies looking for expression signatures or pathways associated with response to therapy (40). Previously, several groups have used integration of "omics" data in OVCA to predict other clinical outcomes (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 chemoresponse in OVCA. This model, with two clusters involving previous known mechanisms of chemoresponse (9), was the most robust in the TCGA database; but what is even more important and unprecedented in our study is the external validation of the TCGA model with other five independent gene-expression studies of OVCA (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 microarray 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 five independent gene-expression sets, in which the agreement was moderate (Fig. 2). We think that adding elements to the model that did not result exclusively from the differential gene expression of microarrays (CNA, mutations, methylation, miRNA, and 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 (8, 9).

Most genes included in our molecular signature for chemoresponse are drawn from cellular functions previously associated with response to chemotherapy (9). These biologic 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 (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 S7) may result in diverse effects, including cell proliferation, invasion, angiogenesis, apoptosis evasion, and response to chemotherapy (44, 45). The MAPK signaling pathway is also connected to the Ras pathway (which includes PAK4), that also regulates cell morphology, cytoskeletal organization, and cell proliferation.
and migration; PAK4 can also function as an antiapoptotic protein (46). PAK proteins are critical effectors that link Rho GTPases to cytoskeleton reorganization and nuclear signaling. Both PAK4 and RHOT1 are included in cluster 2. The Ras gene family (which 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 Ras onco-gene action (47). MAPKs may also have a role in early gene expression by modifying the chromatin environment of target genes (48), an action regulated through phosphorylation of various substrates, including TFs and chromatin constituents. NCAPG, a component of the condensin complex that is required for both interphase and mitotic condensation, is present in cluster 2 of the chemoresponse model. Animal models with condensin mutations, DNA damage induced by UV radiation is not repaired, and cells arrested by hydroxyurea do not recover (49). In our gene signature, it is notable that a set of genes map to the DNA repair pathway (mainly through homologous recombination), like RAD52, and elements of the PARP family, like PARP12. PARPs inhibitors have been proven to be efficacious in the treatment of OVCA in carriers of BRCA1
or BRCA2 mutations (50). Also notable are the results of bevacizumab, a humanized antibody against VEGF, in the adjuvant treatment of OVCA (51). PDGFβR, 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 (52).

The 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 histologic type (serous) and outcome of interest (chemoresponse) to improve homogeneity. We believe that those are major factors influencing in the important external validation of the molecular signature in five independent gene-expression experiments, which is unprecedented in microarray analysis (40). Other major factor influencing a significant validation is the integration of diverse biologic 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 (12, 15). These datasets had the closest clinical information, including chemoresponse, and study design to TCGA. To avoid data over fitting in the validation process, though, we added three independent gene-expression experiments not used before, despite presenting minor differences in study design and platform content (14, 17, 18).

A major limitation of this study is its retrospective nature. Although the outcome of interest (chemoresponse) was collected, we had to reformat it to fit our definition of CR and IR. 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 more than 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 (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 whether the signature is valid in other solid tumors treated with comparable platinum-based combinations.

In summary, integration of diverse biologic data into gene expression may strengthen gene signature models for prediction of OVCA chemoresponse. Robust validation over five independent publicly available gene-expression experiments supports these findings. Nonetheless, this model has to be validated in vitro or 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.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: J. Gonzalez Bosquet, H. Chon, J.M. Lancaster, S. Chanock
Development of methodology: J. Gonzalez Bosquet, J.M. Lancaster
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Gonzalez Bosquet, S. Chanock
Analysis and interpretation of data (e.g., statistical analysis, biosubstatistics, computational analysis): J. Gonzalez Bosquet, S. Chanock
Writing, review, and/or revision of the manuscript: J. Gonzalez Bosquet, J.M. Lancaster
Study supervision: J. Gonzalez Bosquet, S. Chanock

Acknowledgments
The authors thank “TCGA Research Network” for generating, curating, and providing high quality biologic 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 and Research Institute, University of South Florida (Tampa, FL); and by the "St. Louis Ovarian Cancer Awareness Research Grant" from the 2012–2013 Foundation for Women’s Cancer grants. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 23, 2014; revised March 28, 2014; accepted April 21, 2014; published OnlineFirst May 21, 2014.

References


## Analysis of Chemotherapeutic Response in Ovarian Cancers Using Publicly Available High-Throughput Data

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

*Cancer Res* 2014;74:3902-3912. 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

### Cited articles
This article cites 52 articles, 20 of which you can access for free at:
http://cancerres.aacrjournals.org/content/74/14/3902.full#ref-list-1

### 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, contact the AACR Publications Department at permissions@aacr.org.