scientific reports

Check for updates

OPEN

Creation and validation of models to predict response to primary treatment in serous ovarian cancer

Jesus Gonzalez Bosquet[1,2]✉, Eric J. Devor[3], Andreea M. Newtson[1], Brian J. Smith[2,4], David P. Bender[1,2], Michael J. Goodheart[1,2], Megan E. McDonald[1], Terry A. Braun[2,5], Kristina W. Thiel[3] & Kimberly K. Leslie[2,3]

Nearly a third of patients with high-grade serous ovarian cancer (HGSC) do not respond to initial therapy and have an overall poor prognosis. However, there are no validated tools that accurately predict which patients will not respond. Our objective is to create and validate accurate models of prediction for treatment response in HGSC. This is a retrospective case–control study that integrates comprehensive clinical and genomic data from 88 patients with HGSC from a single institution. Responders were those patients with a progression-free survival of at least 6 months after treatment. Only patients with complete clinical information and frozen specimen at surgery were included. Gene, miRNA, exon, and long non-coding RNA (lncRNA) expression, gene copy number, genomic variation, and fusion-gene determination were extracted from RNA-sequencing data. DNA methylation analysis was performed. Initial selection of informative variables was performed with univariate ANOVA with cross-validation. Significant variables ($p < 0.05$) were included in multivariate lasso regression prediction models. Initial models included only one variable. Variables were then combined to create complex models. Model performance was measured with area under the curve (AUC). Validation of all models was performed using TCGA HGSC database. By integrating clinical and genomic variables, we achieved prediction performances of over 95% in AUC. Most performances in the validation set did not differ from the training set. Models with DNA methylation or lncRNA underperformed in the validation set. Integrating comprehensive clinical and genomic data from patients with HGSC results in accurate and robust prediction models of treatment response.

Abbreviations

HGSC	High-grade serous ovarian cancer
RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
lncRNA	Long non-coding RNA
miRNA	Micro RNA
ANOVA	Analysis of variance
AUC	Area under the curve
TCGA	The cancer genome atlas
CI	Confidence interval
PFS	Progression-free survival
OS	Overall survival
IRB	Institutional review board
WHTR	Women's health tissue repository
gDNAs	Genomic DNAs
SNV	Single nucleotide variation
CNV	Copy number variation

[1]Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA. [2]Holden Comprehensive Cancer Center, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA. [3]Department of Obstetrics and Gynecology, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA. [4]Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA 52242, USA. [5]Coordinated Laboratory for Computational Genomics, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA. ✉email: jesus-gonzalezbosquet@uiowa.edu

GCE	Gene copy estimation
NCBI	National Center for Biotechnology Information
PCA	Principal component analysis
PC	Principal components
UI	University of Iowa
ROC	Receiver operating characteristic
FIGO	International federation of obstetrics and gynecology
GEO	Gene expression omnibus
FDR	False discovery rate

Despite notable advances in the treatment of ovarian cancer, it continues to be one of the leading causes of cancer death among women in the United States[1]. The most common type of ovarian cancer is high-grade serous cancer (HGSC). HGSC typically presents as advanced disease, and standard treatment consists of combined primary cytoreductive surgery and platinum-based chemotherapy[2]. Platinum is considered the most effective drug for HGSC[2]. Patients that respond to initial therapy and have progression-free survival (PFS) of at least 6 months are termed "platinum-sensitive" or "responders" and have a median survival of well over four years[3]. In patients that have no residual disease after the initial surgery and respond to chemotherapy, median survivals can reach over 10 years[3]. However, in nearly a third of patients, HGSC progresses during initial chemotherapy (termed "platinum-refractory") or recurs < 6 months after finishing treatment (termed "platinum-resistant")[2,4–6]. The majority of these patients with suboptimal response to initial treatment (termed "non-responders" herein) will die from their disease within two years[4,7,8] and are typically treated in the second-line setting with alternative therapies that do not contain platinum[2].

In recent years, significant efforts have been dedicated to test new targeted drugs in clinical trials to increase PFS of the patients that already respond to primary chemotherapy, with celebrated successes[9–12]. However, few resources have been dedicated to identify those patients that are at risk of failing initial treatment before its administration, and there is no validated test that can predict robustly and accurately this outcome[13,14]. By contrast, in breast cancer gene signatures have been identified that can accurately predict recurrence[15] and chemotherapeutic response[16,17]. These signatures have been validated in independent clinical studies[17–20]. For example, one of these signatures, OncotypeDx, used 600 cases to create an association model and validated the model in an additional 400 cases[15,16]. The majority of previous attempts to define predictors of treatment response in HGSC have been limited by a small number of patients, mixture of histological types and stages, and lack of validation in independent datasets[13,14]. One of the more successful efforts used serum markers, including kallikreins and CA 125[13,14]. The performance of these prediction models ranges from 75–85% (measured as the area under the receiver operator curve (AUC)). Adding clinical characteristics to serum markers increases the performance of the model to an AUC of 90%[14]. Others have integrated the Cancer Genome Atlas (TCGA) genomic data to predict overall survival (OS) and PFS, with performances that ranged from AUCs of 81 to 87%[21]. However, none of these models have been validated in independent datasets, nor have they been validated prospectively.

Using publicly available multi-dimensional datasets with clinical data, like TCGA, we previously built prediction models that distinguish between different outcomes in ovarian and endometrial cancer; we validated these models in independent datasets[22–27]. However, these models had some limitations due to suboptimal clinical data: many patients were lost to follow-up, and others had little information about clinical variables that influenced treatment response, such as stage of disease or number of cycles of chemotherapy. Also, some datasets did not have complete molecular information because expression analyses were performed on different platforms, with different probes. This last limitation severely impacted the performance of the validation studies[22].

Herein we tested the hypothesis that integrating comprehensive clinical and genomic data from patients with HGSC will ultimately result in a more accurate and robust prediction models of response to treatment. The primary objective of our study was to create models of prediction to standard therapy in patients with HGSC. The secondary objective was to validate these models in an independent dataset. Also, we intend to extract maximum genomic information from RNA-sequencing (RNA-seq) so resulting models would be feasible and affordable for any laboratory.

Methods

This is a retrospective case–control study that used clinical and genomic information to create models to predict initial response to standard therapy for HGSC patients. The prediction was made using only data that could be obtained before the administration of the initial chemotherapy. Also, as we mentioned, we intend to extract maximum genomic information from RNA-sequencing (RNA-seq) experiments.

Outcomes definition. HGSC patients were classified as responders or non-responders. Responders were those with a progression-free survival of at least 6 months after the first platinum-based treatment. Non-responders were those who did not respond (platinum-resistant) or progressed during treatment (platinum-refractory).

Patient inclusion criteria. Ovarian cancer patients with high grade serous histology and complete clinical and pathological data were included. Patients that had less than 6 months of follow-up after completing initial chemotherapy, unknown disease status after 6 months of completion of initial chemotherapy, or incomplete data about the chemotherapy delivered were excluded. Also, patients without DNA and RNA of sufficient quality (see below) for RNA-seq or DNA methylation analysis were excluded from the study. Based on the definition of response to treatment, there were a total of 50 patients classified as responders and 38 as non-responders included in the study (Fig. 1). All patients received combination platinum-based chemotherapy initially. How-

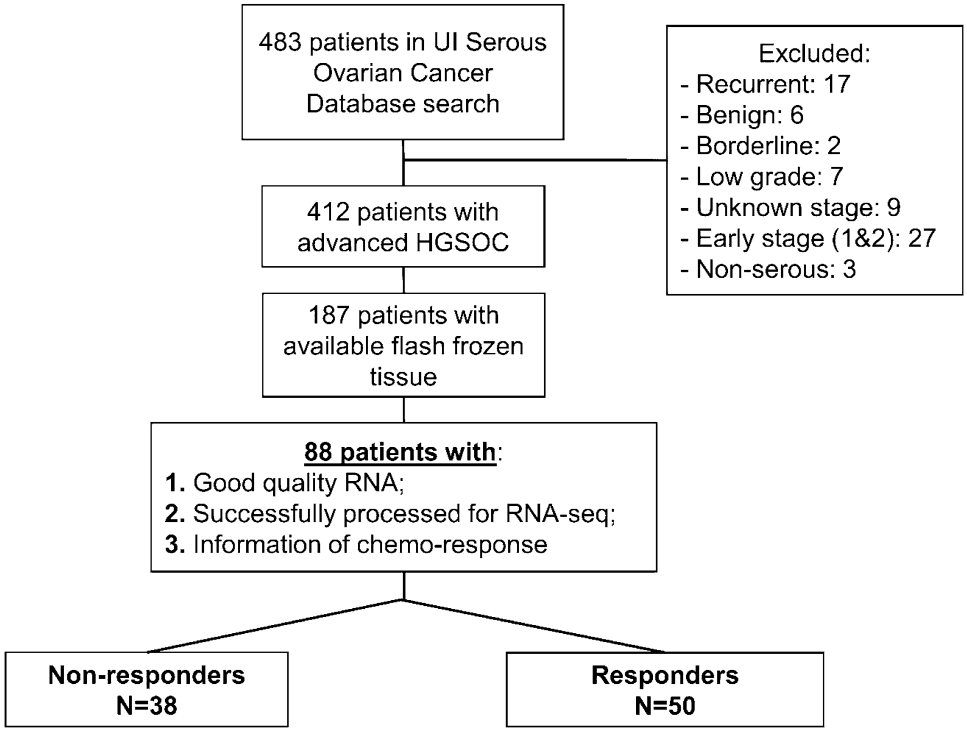

Figure 1. Selection criteria for patients in clinical prediction analysis.

ever, in 2 patients the regimen was changed before finishing because of disease progression in one case, and stable disease in the other. 66% of patient in our analysis received Taxol as initial treatment.

The institutional review board (IRB) of the University of Iowa (UI) approved the current study including human subjects/materials on April 25, 2018 (IRB Number 201804817: '*Prediction Models in Ovarian Cancer*'). The UI Department of Obstetrics and Gynecology maintains a Women's Health Tissue Repository (WHTR) containing more than 60,000 biological samples, including more than 2500 primary gynecologic tumors[28]. All tissues in the WHTR are collected under informed consent of patients in accordance with University of Iowa IRB guidelines (IRB Number 200910784 and IRB Number 200209010). Tumor samples were collected, reviewed by a board-certified pathologist and flash frozen. HGSC diagnosis was confirm in paraffin. Specimens had less than 30% of necrosis.

Clinical data. Clinical and pathological data were collected from the electronic medical record. Clinical variables previously observed to be associated with treatment response were included in the data collection[27]. Only baseline clinical and pathological characteristics that can be obtained before starting initial chemotherapy were collected. Table 1 shows the main clinical variables collected for the study. Differences between clinical variables between responders and non-responders were assessed with logistic regression. P-values < 0.05 were considered statistically significant. All clinical variables initially used in the analysis are described in Supplementary Methods. Statistical analysis and graphics were performed with R statistical package and computer environment[29].

Biological data. *RNA purification and sequencing.* Of the 187 patients identified in the original HGSC panel, 88 primary tumor tissues with sufficient RNA yield and quality were available for analysis; 50 were responders and 38 non-responders (Fig. 1). Most tumors were collected from the ovaries, 63%; 30% were extracted from the omentum, 3% from a pelvic mass and the rest, 4%, from an abdominal mass. The were no differences between both groups ($p = 0.2$), responders and non-responders, in this distribution. At the time of diagnosis, these HGSC were considered of ovarian origin. Now we assume they would be tubal. Only 3 of them had no ovaries (previously removed) and were considered as primary peritoneal.

Total cellular RNA was purified from primary tumor tissue using the mirVana (Thermo Fisher, Waltham, USA) RNA purification kit following the manufacturers' instructions. Yield and quality of purified cellular RNA was assessed using a Trinean DropSense 16 spectrophotometer and an Agilent Model 2100 bioanalyzer. Samples with an RNA integrity number (RIN)[30] greater than or equal to 7.0 were selected for RNA sequencing.

RNA processing and sequencing has been described elsewhere[31]. Briefly, equal mass total RNA (500 ng) was quantified by Qubit measurement (Thermo Fisher, Waltham, USA). Each qualifying tumor was fragmented, converted to cDNA and ligated to bar-coded sequencing adaptors using Illumina TriSeq stranded total RNA library preparation (Illumina, San Diego, CA, USA). Molar concentrations of the indexed libraries were confirmed on the Agilent Model 2100 bioanalyzer and libraries were then combined into equimolar pools for sequencing. The concentration of the pools was confirmed using the Illumina Library Quantification Kit (KAPA Biosystems,

			Responders	Non-responders	
			N = 50	N = 38	p-value
Age (median, range)			56 (25–81)	64 (33–83)	0.009
Charlson Comorbidity Index*	1–3		9	6	0.039
	4–6		35	21	
	>6		1	6	
	Unknown		5	5	
FIGO stage	3		39	25	0.069
	4		7	12	
	Unknown		4	1	
Disease in upper abdomen (other than omentum) by imaging	Yes	Large bowel (N = 4)	28	29	0.051
		Porta—hepatis (N = 4)			
		Mesenteric mets (N = 4)			
		Other (N = 22)			
	No		22	9	
Disease in the chest by imaging	Yes	Chest (N = 4)	6	0	0.991
		Pleural effusion (N = 5)			
	No		44	38	
Grade	2		8	11	0.146
	3		35	23	
	Unknown		7	4	
Residual disease after surgery	Microscopic		12	3	0.053
	Macroscopic		37	35	
	Unknown		1	0	
	Optimal (< 1 cm)		37	20	0.039
	Suboptimal (≥ 1 cm)		13	18	
Removal of pelvic LN	Yes		9	4	0.333
	No		41	34	
Removal of para-aortic LN	Yes		5	3	0.734
	No		45	35	
Surgical complexity score#	Low		22	23	0.990
	Intermediate		28	12	
	High		0	3	
Neoadjuvant chemotherapy	Yes		2	10	0.008
	No		47	28	
	Unknown		1	0	
Number of cycles delivered	<6		2	8	0.344
	≥6		48	30	
Dose dense chemotherapy+	Yes		1	1	0.844
	No		49	37	

Table 1. Patient characteristics and association with treatment response. *Charlson Comorbidity Index is a measure of the prognostic burden of all associated morbidities to predict mortality, and is the most validated measure of the prognostic impact of multiple chronic illnesses (www.charlsoncomorbidity.com). #Surgical complexity score: score to predict surgical morbidity and 90-day mortality after primary debulking surgery for HGSC[67]. +Dose dense chemotherapy: increases the dose intensity of the regimen. In serous ovarian cancer, dose dense therapy consists in increasing IV administration of paclitaxel from every 3 weeks to weekly.

Wilmington, MA, USA). Sequencing was then carried out on the Illumina HiSeq 4000 genome sequencing platform using 150 bp paired-end SBS chemistry. All library preparation and sequencing were performed in the Genome Facility of the University of Iowa Institute of Human Genetics (IIHG). Quality control (QC) of both DNA methylation arrays and RNA-seq experiments were performed to minimalize technical biases (see details in Supplementary Methods).

DNA methylation assay. Genomic DNAs (gDNAs) were purified from frozen tumor tissues using the DNeasy Blood and Tissue Kit according to manufacturer's (QIAGEN) recommendations. Yield and purity were assessed on a NanoDrop Model 2000 spectrophotometer and used a 260 nm/280 nm absorbance ratio of ~ 1.8 with minimal to no degradation as shown through horizontal agarose gel eletrophoresis. For more details please see the original publication of DNA methylation assessment in HGSC[32]. Bisulfite converted gDNAs from HGSC

tumors were submitted to the Genomics Core Facility of the IIHG for processing on methylationEPIC arrays. The Illumina Infinium MethylationEPIC BeadChip Kit (Illumina, San Diego, CA, USA) allows quantification of more than 850,000 methylation sites across the human genome. Bisulfite-converted samples were denatured and neutralized before they were isothermally amplified overnight. The amplified product was fragmented enzymatically. After isopropanol precipitation, fragmented DNA was resuspended and placed onto Illumina methylationEPIC BeadChip and hybridized. There are two different bead types for each CpG locus, representing methylated or unmethylated DNA. The BeadChip was washed to remove unhybridized DNA, followed by extension and staining. The arrays were scanned with the Illumina iScan and methylation intensity measured. Analysis was performed using the Minfi R statistical package[33].

Pre-processing of biological data. RNA-seq reads were mapped and aligned to the human reference genome (version hg38) using STAR, a paired-end enabled algorithm[34]. BAM files were produced after alignment. We used featureCount to measure gene expression from BAM files[35]. After the gene counts were generated, we used DESeq2 package to import, normalize and prepare data for analysis[36]. We independently used gene expression and micro RNA (miRNA) expression for the association analysis. Exon specific expression needed different mapping references for alignment, therefore ENSEMBL was used to annotate single exons during the mapping process. Then, single exon features were extracted from these newly created BAM files with the DEXSeq package[37]. BAM files for each sample were also used for genomic or single nucleotide variation (SNV) discovery and base-calling against the human genome reference utilizing SAMtools and BCFtools for sorting and indexing[38]. After filtering for duplicates, known non-synonymous single-nucleotide variants, and synonymous variants, results were annotated with ANNOVAR and formatted to display the number of variants per gene and sample[39]. We included only non-synonymous variants. To estimate gene copy we used SAMtools and CopywriteR using BAM files as input[40]. CopywriteR is a suite of tools that uses off-target sequenced data to detect CNV and, initially, was conceived to be used with DNA sequencing products. However, due to the particularities of the method, that uses off-target (not exonic) reads uniformly distributed along the genome, which also are available even in low-coverage sequencing, we used this method to create variables that would be proxies for gene copy in the prediction model (gene copy estimation, or GCE). CopywriteR software uses the segmentation algorithm CBS (circular binary segmentation) to create segmentation files that contain log2-transformed, normalized ratios of read counts, that can be used to do further prediction analyses. Long non-coding RNA (lncRNA) were determined using BAM files as input[41]. Fusion-genes were determined from *fastq* files processed with the STAR-Fusion suite[42]. Supplementary Fig. S3 depicts the pipeline and analytics used for file pre-processing before modelling.

Statistical analysis. *Variable selection for prediction modeling.* In the prediction model, we only used those variables that could be assessed at baseline, prior to initiation of treatment. RNA features were used only to create predictions models of response to treatment, not for other comparisons. Most RNA features were used as continuous variables. Only presence and absence of SNV and fusion-genes were used as dichotomous variables: present or not. Our approach was to (1) reduce the number of variables using a univariate selection of prediction variables with cross-validation; (2) utilize those significant variables from the univariate selection process in a multivariate model to predict response risk. Rather than introducing all variables directly in the prediction model, this approach was chosen because it would likely lead to a model with less complexity (i.e., fewer variables) and can be more easily validated retrospectively and prospectively. To reduce the number of variables, initially, we introduced only features that were different for both groups in a univariate analysis with ANOVA (p-value < 0.05). Then, cross-validation with 10 replicates for each fold (tenfold) was applied to select those variables that were more informative for prediction of response, as implemented by the *caret* R package[43]. Features that were selected by this univariate analysis were then used for multivariate *lasso* regression modeling. Unless resampling is included in this initial feature selection step, cross-validation of the subsequent models could be biased[44]. Thus, variable selection for all classes of clinical and biological data (gene, miRNA, exon and lncRNA expressions, GCE, fusion-gene, genomic variation, and DNA methylation analysis) was performed with cross-validation to decrease the possibility of overfitting the final model[43]. As result of this selection process, poorly annotated features, present in one or few samples, were eliminated early in the analysis.

Prediction model construction. Selected clinical and types of molecular variables from the k-fold cross-validation process were analyzed individually and in combination to determine their prediction potential for treatment response. The lasso method, as implemented in the glmnet R package[45], was used to develop a regression model to predict responders versus non-responders. We selected lasso because it is a multivariate regression method that allows simultaneous selection and estimation of the effects of variables, while accounting and adjusting for confounding factors. In our experience, lasso consistently lowers number of samples and computes AUC without reporting any errors, as compared to other prediction methods[22]. We evaluated the performance of our model using the AUC and its 95% confidence interval (CI). AUC was estimated with 1,000 replicates of tenfold cross-validation to avoid over-fitting of the model (internal validation)[46]. Bias-corrected and accelerated bootstrap CIs were computed for resulting AUCs. A value of 0.5 indicates a lack of model predictive performance, and 1.0 indicates perfect predictive performance, or the best model.

Model validation. For external validation of response prediction models created with UI data, we used a publicly available TCGA HGSC dataset[47]. We included only patients with follow-up of at least 6 months after completing initial chemotherapy, with known disease status after 6 months of completion of initial chemotherapy, and data about type of chemotherapy delivered[25]. As before, only baseline (pre-treatment) *clinical data* were used for validation (see Supplementary Methods). Not all clinical data used in the UI cohort was available in TCGA.

This is a limitation of using TCGA data for model building[27]. Based on these inclusion criteria, there were a total of 189 patients classified as responders and 149 as non-responders in TCGA for validation of prediction models (Supplementary Table S2).

To make validation possible, only TCGA patients with RNA-seq HGSC tumors were included in the analysis. BAM files, resulting from RNA-seq alignment to hg38 human genome version were downloaded (DbGaP Access #16,003—NCBI) and pre-processed as for the UI dataset (see *Pre-processing of biological data*). Fusion-genes need to be determined from *fastq/fq* files, so we converted BAM files into *fq* files with BCFtools[38]. After sorting and indexing, *fq* files were processed with the STAR-Fusion suite to obtain fusion-genes[42].

The *validation analysis* applies the UI-built model to the TCGA data to predict or discriminate between responder or non-responder classes. For validation of clinical data, we constructed new UI models with clinical variables available in TCGA. The same was done for other types of data with missing variables in TCGA dataset, DNA methylation and fusion-genes. For validation, we used the best UI-built models of treatment response. Next, we used the R package *pROC* to determine thresholds, or cut-offs, for the UI-built model applied to the TCGA data (see details in Supplementary Methods)[48]. Threshold values that yielded sensitivities of > 90% were ranked from highest to lowest sensitivity, negative predictive value and AUC. Among the ranked results, the top-ranked set of tuning parameters was used to fit a final score of the model to the entire set of patients and define the classification rule. A sensitivity threshold of over 90% will identify most of the patients at risk of failing treatment. Our goal is to capture the highest proportion of non-responders for clinical use of the model, while also aiming for acceptable specificity. Similar thresholds have been used to assess tests for malignancy, recurrence or failure of ovarian cancer treatment[49–53]. We coupled high sensitivity with high accuracy measured by AUC: 0.8–0.9 is considered 'a very good' diagnostic accuracy, 0.9–1 is considered 'excellent.'

In previous studies we observed that TCGA patient population has different genetic admixture than UI patient population[54]. That difference may influence the performance of validation analyses. To account and adjust for those genetic differences we extracted genotypes from VCF files obtained after RNA-seq. Then we employed two different strategies for the adjustment: 1) we performed a principal component analysis (PCA) to differentiate genotypes from UI and TCGA datasets and used the first 3 principal components (PC) for adjustment; 2) we performed a lasso regression analysis (PCA) to obtain the genotypes that differentiated UI from TCGA, and used them for adjustment (for details see Supplementary Methods).

Survival analysis. To assess the association of survival with response, survival analysis was performed using Cox proportional hazard ratios.**

Ethical approval and consent to participate. Tumor samples were obtained under informed consent after approval by the University of Iowa Institutional Review Board: IRB# 201,804,817 (approved 5/9/2018) and 200,209,010 (approved 9/19/2005). The institutional review board (IRB) of the University of Iowa (UI) approved the current study including human subjects/materials on April 25, 2018 (IRB Number 201804817: '*Prediction Models in Ovarian Cancer*'). All data collection and processing, including the consenting process, were performed after approval by the University of Iowa IRB.

Consent for publication. All authors have reviewed and approved the manuscript for submission.

Results

In the UI database, 43% of patients were non-responders, and in TCGA HGSC dataset 44% of patients were non-responders, chi-square p-value = 0.88 (Table 1 and Supplementary Table S1, respectively). Non-responder UI patients had higher Charlson comorbidity index score, more residual disease after surgery, and received more frequently neoadjuvant chemotherapy before surgery (Table 1). Non-responder TCGA patients had more residual disease after surgery (Supplementary Table S2). Median survival was 39.3 months (95% CI: 31, 58.2) for UI responders and 57.7 months (95% CI: 44.3, 82) for TCGA responders. Median survival was 12.5 months (95% CI: 8, 19.1) for UI non-responders and 22.7 months (95% CI: 15.9, 26.3) for TCGA non-responders. Based on these 95% CIs, there were no differences in survival for responders and non-responders in the UI and TCGA datasets.

Variable selection for prediction modeling. After the univariate analysis of all clinical and genomic data with ANOVA as described in Methods, we identified those variables that were more informative for the outcome of interest: treatment response (Fig. 2). The number clinical, gene, miRNA, exon and lncRNA expressions, GCE, fusion-gene presence, SNV, and DNA methylation variables selected after the univariate and multivariate analysis, and included in the prediction analyses are detailed in Table 2. Notably, in the genomic variation analysis, we found BRCA2 variants in 24% (12 out of 50) of responders and in 26% (10 out of 38) of non-responders (p = 0.52); and BRCA1 variants in 36% (18 out of 50) of responders and 50% (19 out of 38) in non-responders (p = 0.06). These differences were non-significant; therefore BRCA1&2 were not selected for the prediction analysis.

Prediction model construction. Prediction models of response were built initially with one type of selected data: clinical, gene, exon, miRNA, and lncRNA expression, SNV, GCE, fusion-gene presence, or DNA methylation (Table 2). See Supplementary Table S3 for more details about the variables after lasso prediction. Next, we built models integrating 2 and 3 types of data. The performance of all models was evaluated using the AUC and its 95% CI. By integrating clinical and genomic variables, we achieved prediction performances of over 95%. Figure 3 represents all prediction models with 1, 2, or 3 variables with AUC over 90% (N = 59). Adding 4 or

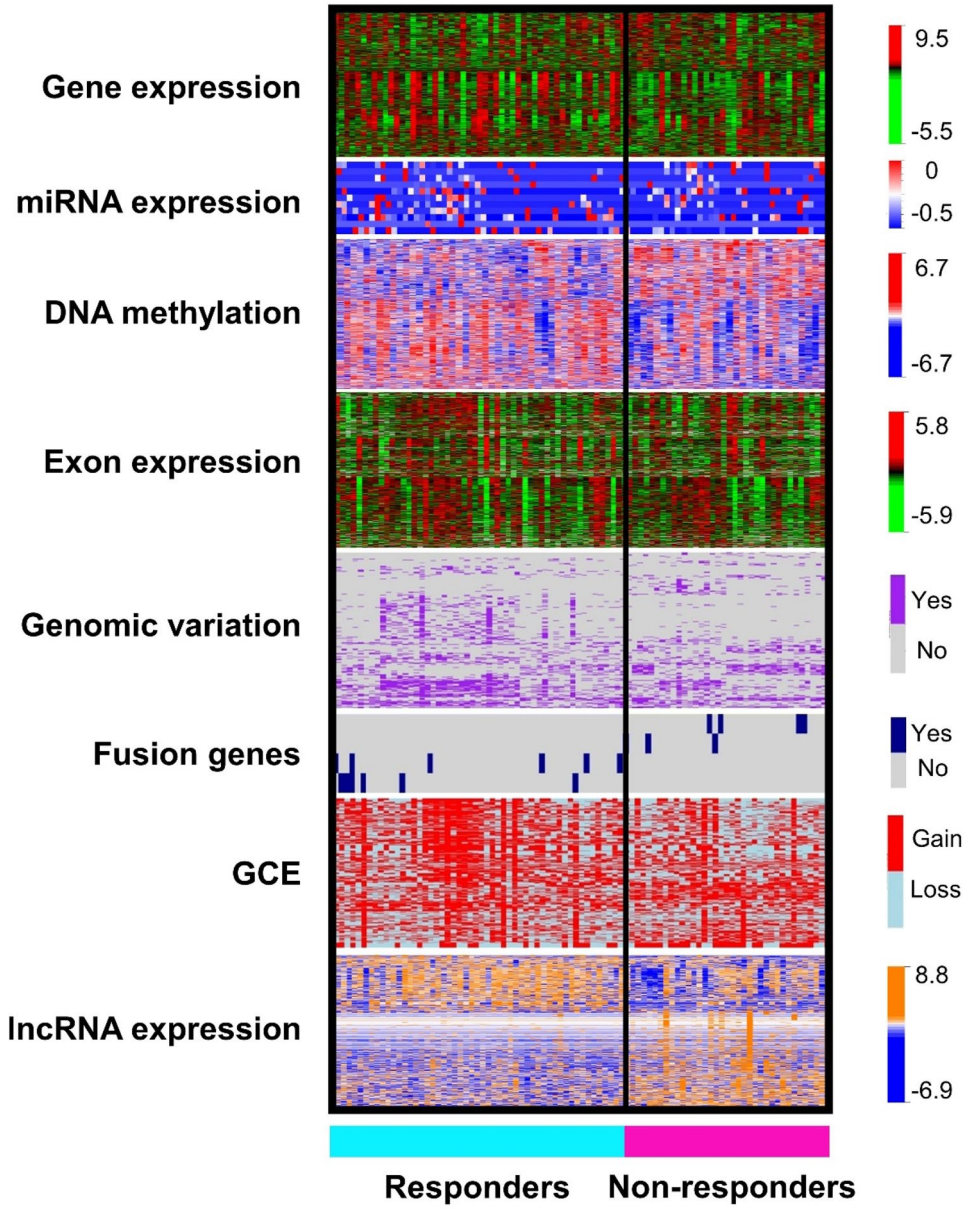

Figure 2. Heatmap of selected variables after univariate ANOVA analysis. Representation of the significant variables after univariate analysis ($p < 0.05$) for different types of genomic data: gene, miRNA, exon, and long non-coding RNA (lncRNA) expression, DNA methylation, genomic variation, fusion-gene presence, and gene copy estimation (GCE). At the right side of each heatmap there are color-coded range of values for all genomic variables. Heatmaps were generated with R package *Heatplus*[68] (R version 3.6.3. http://www.r-project.org).

more types of data increased model complexity without a significant improvement in performance. For details about all prediction models review Supplementary Figure S7.

Model validation. One of the limitations of using TCGA for validation of prediction models was that not all clinical data used in the UI cohort were available in TCGA dataset. For validation analysis, we took all models built using 1, 2, or 3 variables in the UI dataset and inserted TCGA data to assess how well the UI-build models discriminate between responders and non-responders in the TCGA dataset (107 different models)[47]. We selected a sensitivity threshold over 90% in order to identify most of the patients at risk of failing treatment (see the rationale in methods). Notably, validation of models containing DNA methylation and lncRNA data underperformed (Fig. 4A). If we eliminated models with DNA methylation and lncRNA, 80% (51 out of 64) of UI-built models had an AUC 95% CI in the TCGA validation set that overlapped with the UI training set interval (Fig. 4B).

Type of data	Initial number of variables	Variables after selection: univariable ANOVA analysis with k-fold cross-validation	Variables after multivariable prediction model with lasso
Clinical	45	–	7
Gene expression: mRNA	23,528	2214	62
miRNA expression	1914	12	11
Gene copy estimation: GCE	28,917	1098	83
Genomic variation	13,840	327	54
DNA methylation	66,042	4961	35
Long non-coding RNA	16,325	773	69
Fusion genes	597	147	104
Individual exon expression	63,677	4387	61

Table 2. Variable selection and variables after prediction model construction with type of data. To reduce the number of variables, we used univariate analysis of all data with ANOVA to select the variables that were more informative for prediction of response, with a p-value < 0.05 (3rd column). Features that were statistically significant in this univariate analysis were then used for multivariate *lasso* regression modeling. In the last column are the number of variables resulting after performing that prediction model with only one variable. Variables in this last column were used to build prediction models integrating 2 or 3 types of data.

To adjust for different population genetic backgrounds between UI and TCGA cohorts, we used PCA and lasso regression analysis with genotypes that differentiated UI from TCGA. The preliminary studies with PCA did not increase validation performances (see Supplementary Methods), so we next carried out the adjustment with genotypes that differentiated UI and TCGA datasets. When adjusting for genetic variation, validation of the UI-built models had an AUC 95% CI in the TCGA validation that overlapped 97% (62 out of 64 models) with the UI training set interval (Fig. 4C). We did not observe any improvement after adjusting for genotypes in models containing DNA methylation and lncRNA data (Fig. 4D). These results indicate that differences in performance between UI-built and TCGA validation models that do not contain DNA methylation and/or lncRNA data may be related to different genetic background of both datasets.

Discussion

Prediction of treatment response in HGSC patients before treatment initiation is a difficult task. Not many groups have attempted these types of predictions and, of the few reported models, none have been introduced successfully into clinical practice. The reasons are varied but can be classified into two groups: (1) prediction model performances did not warrant their introduction in routine practice[2,13]; or (2) models could not be validated in independent datasets and thus could not be generally applied[14,21]. We present a comprehensive study of treatment response prediction in patients with advanced stage HGSC. Most importantly, these prediction models were validated in an independent dataset with similar in-depth genomic assessment, TCGA.

Prediction models were built by integrating different types of clinical and molecular data. Simpler models were built after selecting those variables more informative for the outcome, treatment response, and using cross-validation to minimize over-fitting. As we increased model complexity (with 2–3 types of data), performances on the training set reached levels over 95% in terms of AUC. These performances are promising and could provide robust clinical decision support to discriminate responders versus non-responders. The most predictive models included diverse types of data, but notably, those around 100% AUC included epigenetic regulators of gene expression, either by DNA methylation or miRNA expression. Also, other gene modulators like lncRNA expression and individual exon expression, a readout for splice variant expression[55], were involved in these high performing models. Finally, clinical information was also an important component of the best prediction models. Other components of high performing models, like fusion genes, have not been characterized yet in HGSC outcome prediction or prognosis, but have been associated with acquired resistance[56].

Epigenetic gene regulation is one of the mechanisms that regulate treatment response. For example, whole genome DNA methylation analysis found that epigenetic regulation of potentially clinically relevant genes predicts response to platinum[57–59]. Also, lncRNAs have been associated with epigenetic regulation of HGSC[60], and specific lncRNAs have been associated with chemo-resistance[61]. Therefore, it is not surprising that some of the best prediction models are composed of diverse lncRNAs. The role of miRNAs in response to treatment in HGSC in vitro is well documented[62], and some miRNAs also have been associated with chemo-response modulation[63]. The presence of epigenetic regulators and modifiers of treatment response in high performance prediction models of response to therapy seems to support these regulatory mechanisms. We must be cautious about extrapolating functional conclusions from prediction models. Prediction analysis is a form of statistical learning that uses data obtained in the past to predict outcomes, or behavior, of other individuals in the future. Prediction analysis is based on association and does not infer causation[64].

Prediction models of response created with UI data were validated in TCGA HGSC data. Notably, models containing DNA methylation and/or lncRNA data did not perform well in validation analyses. Moreover, when we adjusted for different genetic backgrounds between UI and TCGA samples, validation performances of models *not* containing DNA methylation or lncRNA improved, with 97% of overlap in the 95% CI of AUCs. Conversely, validation models containing DNA methylation and/or lncRNA did not improve despite adjusting for genetic background. We speculate that genetic background differences between the UI-sampled population and TCGA

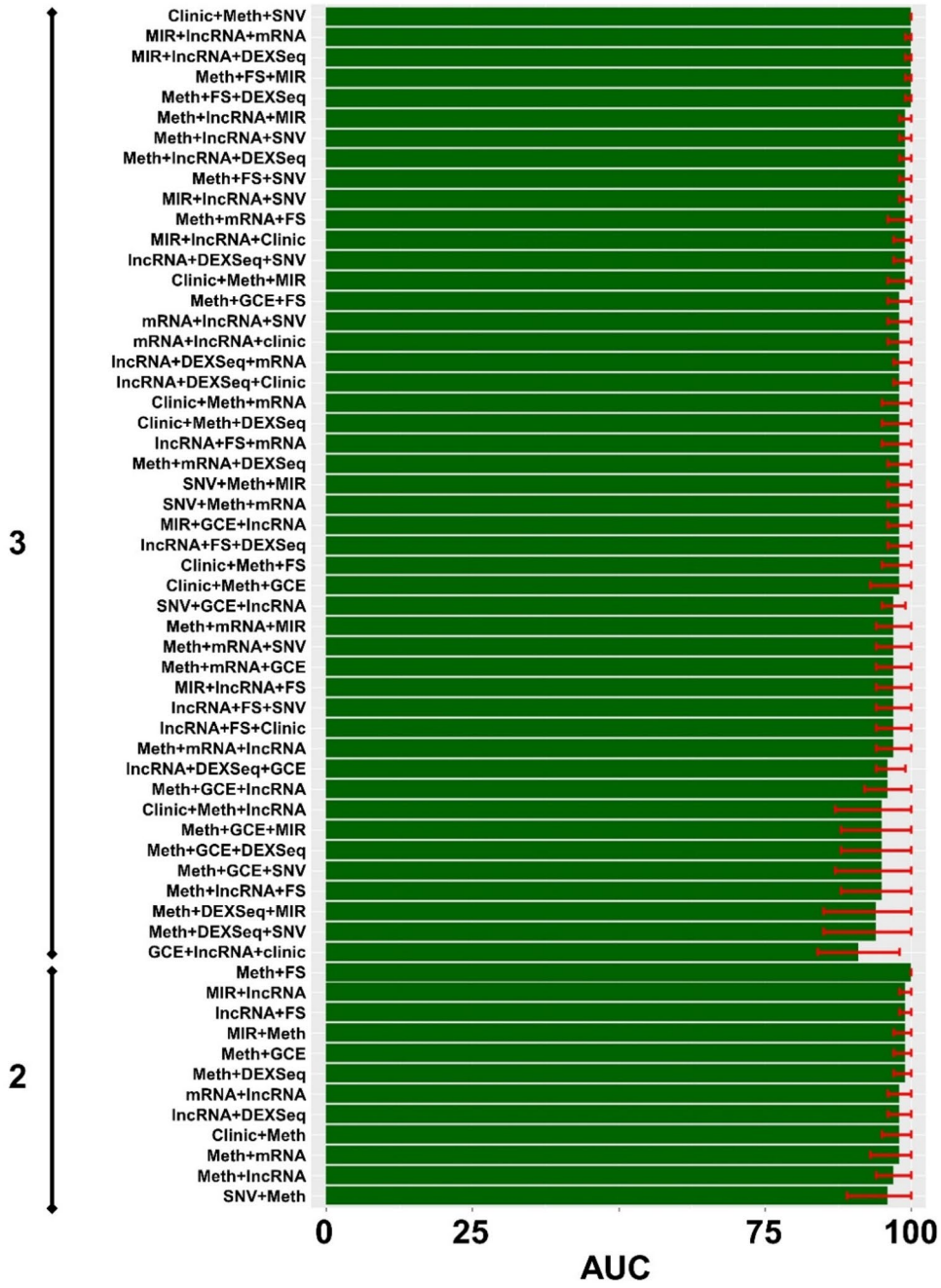

Figure 3. High performing prediction models of response. On the left is the number of types of data: **2**: combination of 2 types of data; **3**: combination of 3 types of data. Different performances are displayed in ascending order. The x axis is AUC as a percentage (0–100%). Although we tested 107 models with 1, 2 and 3 variables, we only represented those models with performances over 90% measured in AUC, N = 59 (to see all of them see Supplementary Figure S7). FS: Fusion genes; Meth: DNA methylation; SNV: single nucleotide variation; GCE: gene copy estimation; DEXSeq: exon expression; lncRNA: long non-coding RNA; MIR: micro RNA, mRNA: gene expression. Graphics were generated with R package *ggplot* (R version 3.6.3. http://www.r-project.org)[69].

account for some variation in the validation of these prediction models of treatment response, except for those models including DNA methylation and/or lncRNA. The poor performance of validation TCGA prediction models including DNA methylation and/or lncRNA data was likely due to other reasons. DNA methylation analysis for TCGA HGSC data was performed using Illumina Infinium HumanMethylation27K BeadChip arrays[47]. Methylation in UI was performed with an EPIC BeadChip 850 k arrays (both arrays from Illumina Inc.). The 27 K methylation array interrogates mainly CpG islands in gene promoter regions, while the 850 K array also explores DNA methylation outside the promoter areas with whole genome coverage[32]. To determine lncRNA

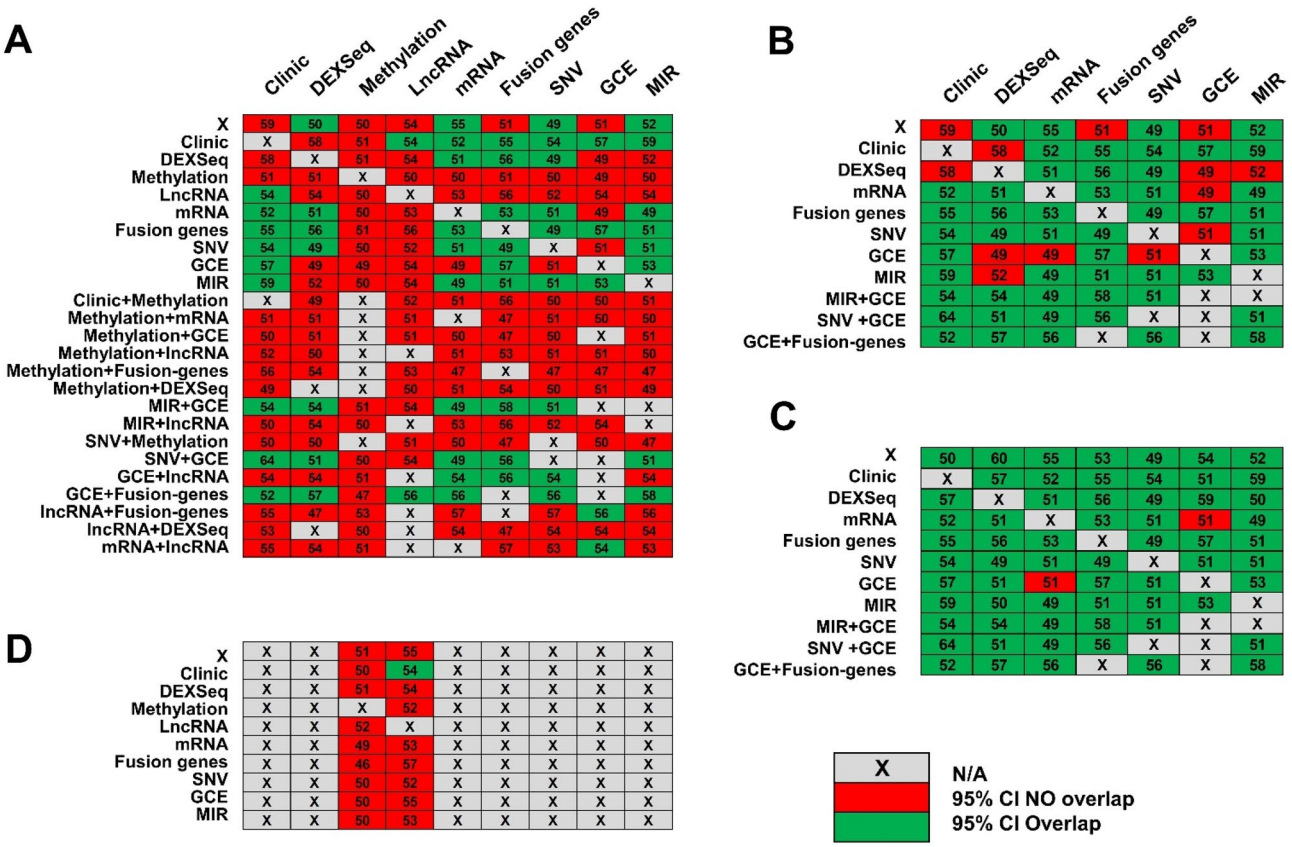

Figure 4. Validation of UI prediction models of response to treatment in TCGA datasets. The columns represent different types of clinical and molecular data: DEXSeq (individual exon expression), Methylation (DNA methylation), LncRNA (long non-coding RNA expression), mRNA (gene expression), Fusion genes (presence of fusion genes), SNV (Single Nucleotide Variation), GCE (gene copy estimation), and MIR (microRNA expression). The rows also represent different types of data, either individually or in combination. (**A**) Validation of all prediction models of response in TCGA: Models containing DNA methylation and lncRNA data underperformed: AUC 95% CIs of TCGA-validation models did not overlap with AUC 95% CIs of UI-built training models (red cells in the graphic). Green cells represent those AUC 95% CIs of TCGA-validation models that overlap with UI-built AUC 95% CIs. (**B**) When we removed models containing DNA methylation and lncRNA data, 80% of CIs from UI-built training models overlapped with CIs from TCGA-validation models. (**C**) When prediction models without DNA methylation and lncRNA data were adjusted with genotyping data (as detailed in in Supplementary Methods), 97% of CIs from UI-built models overlap ped with CIs of TCGA-validation models. (**D**) TCGA-validation performance did not improve when adjusting for different genotypes between UI and TCGA. This was tested in models with 1 and 2 types of data (18 models).

expression in UI HGSC data, we used data from RNA sequencing that was carried out on the Illumina HiSeq 4000 genome sequencing platform using 150 bp paired-end sequencing by synthesis (SBS) chemistry[32]. LncRNA expression was extracted from TCGA HGSC sequenced with the Illumina HiSeq 2000 genome sequencing platform that uses 75 bp paired-end SBS chemistry. Differences between 150 and 75 bp sequencing products may have contributed to differences in background noise and total lncRNA counts. Indeed, differences between the platforms may contribute to the decrease in performance validation with lncRNA data. Other technical differences between both databases, like libraries preparations between both sets, may influence critically overall prediction model performance.

A strength of this study is that we used diverse databases of genomic and clinical variables to build prediction models of response. We postulated that a complete database containing all variables involved in malignant cell functions would make prediction models more accurate[22,27,31,65]. Therefore, we extracted as much information from the HGSC specimens as possible to improve our models. Likewise, with clinical data we extracted as much baseline clinical information that could affect the primary outcome of interest. These variables may have been known previously to affect the outcome, or not. Public databases not designed specifically for prediction assessment, like TCGA, may lack of some characteristics that result in important discrepancies of model performance. In the present study, we were able to adjust for these discrepancies in some of the models, except for those containing DNA methylation and/or lncRNA data. Another strength is the outcome definition. Progression-free survival of at least 6 months after the first platinum-based treatment, or responders, is an standard definition of response to chemotherapy[2,5,6]. Indeed, patients that do not respond to initial standard treatment, platinum-resistant, or progressed during treatment, platinum-refractory, are considered a different population

when chemotherapy or clinical trials are considered[4,8]. Treatment response was reviewed in all UI patients, and any patient that did not meet the inclusion criteria was excluded from the analysis. Also, to be included in validation analyses all TCGA patients had to meet treatment response definition criteria. Finally, validation of all models of prediction in a public, well known, independent database (TCGA) also strengthens our study. One of the inherent limitations of the study comes from its design: to build prediction models, initially, the outcome must be known, so the initial step of the analysis and validation is retrospective. The advantage of extracting data from patients of a single institution is the uniformity of diagnosis, outcome definition, specimen collection and processing, treatment philosophy and surveillance. This resulted in a homogenous population with quality biological and clinical data. The selection process may have some disadvantages, though, and the selected samples may lack diversity and it may be limited by the number of HGSC eligible patients. We are the largest of only two Gynecological Oncologic practices that serves around 80–90% of women with gynecological cancer in the state of Iowa (USA). Thus, the samples we have studied represents the female racial composition of the State of Iowa: 95.5% white, 1.1% black, 3.4 other (Latina, Asian, Pacific)[54]. We adjusted these differences in genetic variation with TCGA during validation (see Results). In future studies, it may be especially important that every center that treats women with HGSC knows exactly the racial and genetic composition of the population they treat, so they can correct or adjust for these differences. We acknowledge the limited sample size used to construct prediction models, but the necessity of having an accurate outcome definition and homogeneous population is even more critical. Previous prediction studies are plagued with patients with heterogenous clinical characteristics and outcome definitions, and with different histological types of ovarian cancers that made generalizability even more difficult[2,13,14,21].

Before these models can be applied clinical, they must be validated prospectively. Despite internal and external validation, models of prediction still may have biases due to overfitting. In the prospective model those biases could be detected and corrected before clinical application. Then, for models of prediction to be applied, there is a process that must be followed. After a biopsy is taken (either CT-guided or during surgery) and HGSC histologic type has been confirmed, we will determine all components of the best prediction model validated prospectively. As long as CT-guided biopsies have enough tumor cellularity (over 2/3), and not too much necrosis (< 30%), it would be enough and comparable to the initial model. Sequencing is rapidly evolving with single cell RNA-seq technology[66]. Each of these components will be transformed as they were formatted in the initial analysis (i.e., log transformed, coding values, etc.) and the values would be applied to the weight of each variable in the selected model. The addition of all values will give us a final score. That score, and where it is located with respect to a chosen threshold, will assess the risk of the patient to fail initial therapy. We could design customized assays, PCR-based, for the genomic features of the model that would reduce costs and complexity and would improve the turnaround time so the results could be used even before surgery. With this information, clinicians could have a good sense of which patients would respond to treatment, and they would be better informed to plan initial treatment. For example, if a patient has a higher score for response, the surgeon would take that into consideration to balance the effort in cytoreduction with possible complications. Knowing that the patient may not respond as well to initial treatment may increase the surgical effort to minimize residual disease. Also, if a patient has a lower score for response, we may want to involve them in clinical trials that add targeted treatment to the initial chemotherapy backbone to improve outcomes.

Conclusions

Based on our results and previous reporting from other prediction studies[22,27,31,65] we can conclude that our hypothesis holds true: integrating comprehensive clinical and genomic data from patients with HGSC results in accurate and robust prediction models of treatment response. We have described high performance prediction models of response for initial treatment in HGSC. Based on these performances, some of these prediction models could be useful to provide clinical decision support that will differentiate responders to non-responders. Furthermore, these models were validated in an independent, trusted, well known database.

Data availability

Data for the prediction model has been submitted to the GEO at NCBI website: https://www.ncbi.nlm.nih.gov/geo/. Datasets with methylation data can be browsed by their accession number: GSE133556. Datasets with RNA-seq can be browsed by their accession number: GSE156699. The validation part of this study was performed in silico, with de-identified publicly available data. All data from TCGA is available at their website: https://portal.gdc.cancer.gov/. Software utilized by this study is also publicly available at Bioconductor website: http://bioconductor.org/.

Received: 21 August 2020; Accepted: 24 February 2021
Published online: 16 March 2021

References

1. Torre, L. A. *et al.* Ovarian cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 284–296 (2018).
2. Cannistra, S. A. Cancer of the ovary. *N. Engl. J. Med.* **351**, 2519–2529 (2004).
3. Walker, J. L. *et al.* Randomized trial of intravenous versus intraperitoneal chemotherapy plus bevacizumab in advanced ovarian carcinoma: an NRG oncology/gynecologic oncology group study. *J. Clin. Oncol.* **37**, 1380–1390 (2019).
4. Friedlander, M. L. *et al.* Clinical trials of palliative chemotherapy in platinum-resistant or -refractory ovarian cancer: time to think differently?. *J. Clin. Oncol.* **31**, 2362 (2013).
5. Therasse, P., Arbuck, S. G., Eisenhauer, E. A. *et al.* New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J. Natl. Cancer Inst.* **92**, 205–216 (2000).

6. Friedlander, M. *et al.* Symptom control in patients with recurrent ovarian cancer: measuring the benefit of palliative chemotherapy in women with platinum refractory/resistant ovarian cancer. *Int. J. Gynecol. Cancer* **19**(Suppl 2), S44–S48 (2009).
7. American Cancer Society. *Cancer Facts & Figures 2014* (American Cancer Society, 2014).
8. Davis, A., Tinker, A. V. & Friedlander, M. "Platinum resistant" ovarian cancer: What is it, who to treat and how to measure benefit?. *Gynecol. Oncol.* **133**, 624–631 (2014).
9. Perren, T. J. *et al.* A phase 3 trial of bevacizumab in ovarian cancer. *N. Engl. J. Med.* **365**, 2484–2496 (2011).
10. Burger, R. A. *et al.* Incorporation of bevacizumab in the primary treatment of ovarian cancer. *N. Engl. J. Med.* **365**, 2473–2483 (2011).
11. Pujade-Lauraine, E. *et al.* Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* **18**, 1274–1284 (2017).
12. Pujade-Lauraine, E. *et al.* Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: The AURELIA open-label randomized phase III trial. *J. Clin. Oncol.* **32**, 1302–1308 (2014).
13. Oikonomopoulou, K. *et al.* Prediction of ovarian cancer prognosis and response to chemotherapy by a serum-based multiparametric biomarker panel. *Br. J. Cancer* **99**, 1103–1113 (2008).
14. Zheng, Y. *et al.* A multiparametric panel for ovarian cancer diagnosis, prognosis, and response to chemotherapy. *Clin. Cancer Res.* **13**, 6984–6992 (2007).
15. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
16. Wan, Y. W., Qian, Y., Rathnagiriswaran, S., Castranova, V. & Guo, N. L. A breast cancer prognostic signature predicts clinical outcomes in multiple tumor types. *Oncol. Rep.* **24**, 489–494 (2010).
17. Rathnagiriswaran, S. *et al.* A population-based gene signature is predictive of breast cancer survival and chemoresponse. *Int. J. Oncol.* **36**, 607–616 (2010).
18. Nielsen, T. *et al.* Analytical validation of the PAM50-based prosigna breast cancer prognostic gene signature assay and ncounter analysis system using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer* **14**, 177 (2014).
19. Van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
20. van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(530), 536 (2002).
21. Fu, A., Chang, H. R. & Zhang, Z. F. Integrated multiomic predictors for ovarian cancer survival. *Carcinogenesis* **39**, 860–868 (2018).
22. Gonzalez Bosquet, J. *et al.* Prediction of chemo-response in serous ovarian cancer. *Mol. Cancer* **15**, 66 (2016).
23. Dai, D. *et al.* Stratification of endometrioid endometrial cancer patients into risk levels using somatic mutations. *Gynecol. Oncol.* **142**, 150–157 (2016).
24. Abdallah, R., Chon, H. S. & Gonzalez, B. J. Gene expression and prediction of complete cytoreduction in ovarian cancer. *Obstet. Gynecol.* **123**(Suppl 1), 89S (2014).
25. Gonzalez Bosquet, J., Marchion, D. C., Chon, H., Lancaster, J. M. & Chanock, S. Analysis of chemotherapeutic response in ovarian cancers using publically available high-throughput data. *Cancer Res.* **74**(14), 3902–3912 (2014).
26. Marchion, D. C. *et al.* Gene expression data reveal common pathways that characterize the unifocal nature of ovarian cancer. *Am. J. Obstet. Gynecol.* **209**(576), e1–e16 (2013).
27. Newtson, A. M., Devor, E. J. & Gonzalez, B. J. Prediction of epithelial ovarian cancer outcomes with integration of genomic data. *Clin. Obstet. Gynecol.* **63**, 92–108 (2020).
28. Santillan, M. K. *et al.* Collection of a lifetime: a practical approach to developing a longitudinal collection of women's healthcare biological samples. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **179**, 94–99 (2014).
29. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2020). http://www.R-project.org/.
30. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
31. Miller, M. D. *et al.* An integrated prediction model of recurrence in endometrial endometrioid cancers. *Cancer Manag. Res.* **11**, 5301–5315 (2019).
32. Reyes, H. D. *et al.* Differential DNA methylation in high-grade serous ovarian cancer (HGSOC) is associated with tumor behavior. *Sci. Rep.* **9**, 17996 (2019).
33. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
36. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
37. Badr, E., ElHefnawi, M. & Heath, L. S. Computational identification of tissue-specific splicing regulatory elements in human genes from RNA-seq data. *PLoS ONE* **11**, e0166978 (2016).
38. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
40. Kuilman, T. *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* **16**, 49 (2015).
41. Sun, Z. *et al.* UClncR: ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Sci. Rep.* **7**, 14196 (2017).
42. Haas, B. J. *et al.* STAR-fusion: fast and accurate fusion transcript detection from RNA-seq. *bioRxiv* https://doi.org/10.1101/120295 (2017).
43. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
44. Subramanian, J. & Simon, R. Overfitting in prediction models—Is it a problem only in high dimensions?. *Contemp. Clin. Trials* **36**, 636–641 (2013).
45. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
46. Simon, R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.* **23**, 7332–7341 (2005).
47. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
48. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
49. Kim, H. S. *et al.* Significance of preoperative serum CA-125 levels in the prediction of lymph node metastasis in epithelial ovarian cancer. *Acta Obstet. Gynecol. Scand.* **87**, 1136–1142 (2008).
50. Nassir, M. *et al.* The role of HE4 for prediction of recurrence in epithelial ovarian cancer patients-results from the OVCAD study. *Tumour. Biol.* **37**, 3009–3016 (2016).
51. Bandiera, E. *et al.* Serum human epididymis protein 4 and risk for ovarian malignancy algorithm as new diagnostic and prognostic tools for epithelial ovarian cancer management. *Cancer Epidemiol. Biomark. Prev.* **20**, 2496–2506 (2011).

52. Szpurek, D., Moszynski, R., Zietkowiak, W., Spaczynski, M. & Sajdak, S. An ultrasonographic morphological index for prediction of ovarian tumor malignancy. *Eur. J. Gynaecol. Oncol.* **26**, 51–54 (2005).
53. Szpurek, D., Moszyniki, R. & Sajdak, S. Clinical value of the ultrasound Doppler index in determination of ovarian tumor malignancy. *Eur. J. Gynaecol. Oncol.* **25**, 442–444 (2004).
54. Miller, M. D., Devor, E. J., Salinas, E. A., *et al.* Population substructure has implications in validating next-generation cancer genomics studies with TCGA. *Int. J. Mol. Sci.* **2019**, 20 (2019).
55. French, P. J. *et al.* Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res.* **67**, 5635–5642 (2007).
56. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
57. Tomar, T. *et al.* Methylome analysis of extreme chemoresponsive patients identifies novel markers of platinum sensitivity in high-grade serous ovarian cancer. *BMC Med.* **15**, 116 (2017).
58. Gyparaki, M. T. & Papavassiliou, A. G. Epigenetic pathways offer targets for ovarian cancer treatment. *Clin. Breast Cancer* **18**, 189–191 (2018).
59. Bonito, N. A., Borley, J., Wilhelm-Benartzi, C. S., Ghaem-Maghami, S. & Brown, R. Epigenetic regulation of the homeobox gene MSX1 associates with platinum-resistant disease in high-grade serous epithelial ovarian cancer. *Clin. Cancer Res.* **22**, 3097–3104 (2016).
60. Gloss, B. *et al.* ZNF300P1 encodes a lincRNA that regulates cell polarity and is epigenetically silenced in type II epithelial ovarian cancer. *Mol. Cancer* **13**, 3 (2014).
61. Xu, J. *et al.* Multidrug resistant lncRNA profile in chemotherapeutic sensitive and resistant ovarian cancer cells. *J. Cell. Physiol.* **233**, 5034–5043 (2018).
62. Sorrentino, A. *et al.* Role of microRNAs in drug-resistant ovarian cancer cells. *Gynecol. Oncol.* **111**, 478–486 (2008).
63. Liu, G., Yang, D., Rupaimoole, R., *et al.* Augmentation of response to chemotherapy by microRNA-506 through regulation of RAD51 in serous ovarian cancers. *J. Natl. Cancer Inst.* **2015**, 107 (2015).
64. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* 2nd edn. (Springer, 2009).
65. Salinas, E. A., Miller, M. D., Newtson, A. M., *et al.* A prediction model for preoperative risk assessment in endometrial cancer utilizing clinical and molecular variables. *Int. J. Mol. Sci.* **20**, 1205 (2019).
66. Picelli, S. Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol.* **14**, 637–650 (2017).
67. Aletti, G. D. *et al.* Quality improvement in the surgical approach to advanced ovarian cancer: the Mayo Clinic experience. *J. Am. Coll. Surg.* **208**, 614–620 (2009).
68. Heatplus: Heatmaps with row and/or column covariates and colored clusters (Karolinska Institutet, 2020). https://github.com/alexploner/Heatplus.
69. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis. use R!*,. 2nd edn. (Springer International Publishing, Cham, 2016):1 online resource (XVI, 260 pages 32 illustrations, 140 illustrations in color.

Acknowledgements

The authors would like to thank the Genomics Division of the University of Iowa Institute of Human Genetics for their assistance with this project, specifically Mary Boes and Garry Hauser (core facilities of the IIHG are funded in part by NIH/NCI P30CA086862). We are grateful as well to Dr. Donna Santillan, director of the Department of Obstetrics & Gynecology Women's Health Tissue Repository and Gynecologic Malignancy Bank, for assistance in assembling the Iowa endometrial and ovarian tumor cohorts. Also, we would like to thank 'TCGA Research Network' for generating, curating and providing high quality biological and clinical data.

Author contributions

J.G.B., A.M.N., E.J.D., and B.J.S. conceived of the project and contributed to study design; J.G.B., A.M.N., and E.J.D. performed data collection; J.G.B., and B.J.S. performed computational analyses; J.G.B., E.J.D., B.J.S., and T.A.B. analyzed and interpreted the data; J.G.B., K.W.T., K.K.L., and B.J.S. wrote the manuscript with input from all authors; J.G.B., E.J.D., A.M.N., B.J.S., D.P.B., M.J.G., M.E.M., T.A.B., K.W.T. and K.K.L. read and approved the final version of the manuscript.

Funding

Tumor samples were obtained under informed consent after approval by the University of Iowa Institutional Review Board: IRB# 200910784 and 200209010. This work was supported in part by the NIH grant R01 CA99908 and R01 CA184101 to Kimberly K. Leslie, and the basic research fund from the Department of Obstetrics & Gynecology at the University of Iowa. Also, was supported in part by the American Association of Obstetricians and Gynecologists Foundation (AAOGF) Bridge Funding Award.

Competing interests

K.W.T. is a cofounder of and holds and equity stake in Immortagen, Inc. All other authors: J.G.B., E.J.D., A.M.N., B.J.S., D.P.B., M.J.G., M.E.M., T.A.B., K.K.L. certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85256-9.

Correspondence and requests for materials should be addressed to J.G.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.