Pathway-centric generalizable computational framework uncovers pathway markers governing chemoresistance across cancers

By

Nusrat Jahan Epsi

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

School of Health Professions

Rutgers, The State University of New Jersey

August 2020

APPROVAL PAGE

Dr. Antonina Mitrofanova, Dissertation Committee Chair and Advisor

Date:

Dr. Shankar Srinivasan, Dissertation Committee Member

Date:

Dr. Fredrick D. Coffman, Dissertation Committee Member

Date:

TABLE OF CONTENTS

ABSTRACT

Background: Despite recent advances in discovering a wide array of novel chemotherapy agents, identification of patients with poor and favorable treatment response prior to treatment administration remains a major challenge in clinical oncology and cancer management.

Methods: We have developed a genome-wide systematic computational framework to uncover an interplay between transcriptomic and epigenomic mechanisms that elucidate the complexity of chemotherapy response in cancer patients. Our approach integrates transcriptomic (i.e., mRNA expression) and epigenomic (i.e., DNA methylation) patient profiles to uncover molecular pathways with significant alterations on transcriptomic and epigenomic levels that can distinguish favorable from poor treatment response.

Results: We have tested our approach on patients with lung adenocarcinoma who received a carboplatin and paclitaxel combination chemotherapy (i.e., carboplatin-paclitaxel), a standard-of-care for treating advanced lung cancer. Our integrative approach identified seven molecular pathways with significant alterations on transcriptomic and epigenomic levels that distinguish favorable from poor carboplatin-paclitaxel response, including *chemokine receptors bind chemokines, mRNA splicing, G alpha (s) signalling events, immune network for IgA production*, etc. We have demonstrated that these pathways can classify patients based on their risk to developing carboplatin- paclitaxel resistance in an independent patient cohort (log-rank p-value = 0.0081) and their predictive ability is independent of and is not affected by *(i)* signatures of lung cancer aggressiveness, and *(ii)* commonly utilized covariates, such as age, gender,

and disease stage at diagnosis (adjusted hazard ratio = 14.0). To demonstrate generalizability of our approach, we have applied our algorithm across additional chemotherapy regimens (i.e., cisplatin-vinorelbine, oxaliplatin-fluorouracil) and cancer types (i.e., lung squamous cell carcinoma, and colorectal adenocarcinoma); and have demonstrated our method's ability to accurately predict patients' treatment response.

Conclusions: We propose that our approach can be utilized to identify transcriptomic and epigenomic altered pathways implicated in primary chemoresponse and effectively classify patients who would benefit from specific chemotherapy regimens or are at risk of resistance, which will significantly improve personalized therapeutic strategies and informed clinical decision making.

ACKNOWLEDGEMENTS

I would like to extend my sincere and heartfelt gratitude to all the individuals who have helped me in this scientific endeavor.

I am profoundly grateful to my academic advisor and my mentor, Dr. Antonina Mitrofanova. I feel privileged to have had the opportunity to learn from and interact with such a brilliant paragon, and I feel honored to have done so. Her supervision has made this a thoughtful, enjoyable, and incredible academic journey. She showed me how far I could push myself, that I am capable of meeting challenges I never before thought possible.

Throughout my academic journey, I feel very fortunate to have been surrounded by wonderful, dedicated, and accomplished individuals. I would like to extend my most sincere gratitude to my dissertation committee members. I am very grateful to Dr. Shankar Srinivasan for his endless support, enthusiastic encouragement, and academic guidance. I am also thankful to Dr. Frederick Coffman for his brilliant suggestions for strengthening this dissertation, and for the pointers to the literature and other fruitful discussions on this work. Also, I would like to thank Dr. Suril Gohel for being part of my dissertation committee, a thesis reader, who always provides constructive insights throughout this research.

I would like to acknowledge my wonderful lab mates which I had the pleasure to meet throughout the years specially Ahmed, Sarra, Zach, Vincent, and Daniel for many great conversations, both research and non-research related. In particular, I would like to

thank Sukanya, her warm friendship and brilliant but wacky mind that made the lab a very exciting place!

Further, I would also like to acknowledge the support I have received from the NJ Chapter of the Healthcare Information Management Systems Society (NJHIMSS), New Jersey Health Foundation, and Rutgers School of Health Professions.

On a personal note, I am grateful to my parents who instilled in me the ability to be confident and taught me the virtue of humility. They always held high expectations, and at the same time they trusted me with the choice and the responsibility to meet their standards. They always motivated and encouraged me to pursue my dreams which is a result I am here today. I am deeply appreciative of my Bhai and Chotto, who did everything possible by being continuously supportive and constantly encouraged me through the doctoral process. The unconditional love, encouragement, and unwavering support my family and my beloved pets have offered, and the sincerity with which it has been offered, has meant more to me than any mere words on paper could ever express.

Finally, I would also like to express my deep gratitude to all the other individuals that helped me in many ways throughout my study, I am sincerely grateful.

DEDICATION

To Bhai for his unwavering support and unconditional love, who gave me the courage to chase my dreams and has always been my shield.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NSCLC	Non-small cell lung carcinoma
SCLC	Small cell lung carcinoma
LUAD	Lung Adenocarcinoma
RNA	Ribonucleic acid
mRNA	Messenger RNA
DNA	Deoxyribonucleic acid
pathCHEMO	uncovering transcriptomic and epigenomic **p**athways of **CHEMO**resistance
CHEMO	Chemotherapy
TCGA	The Cancer Genome Atlas
AUROC	Area Under the ROC
ROC	Receiver Operating Characteristic
GDC	Genomics Data Commons
GSEA	Gene Set Enrichment Analysis
NES	Normalized Enrichment Score
LE	Leading edge
MSigDB	Molecular Signatures Database
C2	Gene sets representing canonical pathways
t-SNE	t-Distributed Stochastic Neighbor Embedding
LOOCV	Leave-one-out cross-validation
Epi2GenR	Epigenomic and Genomic mechanisms of treatment Resistance
PRES	Personalized REgimen Selection
SVM	Support Vector Machine
LUSC	Lung Squamous Cell Carcinoma
COAD	Colorectal Adenocarcinoma
FOLFOX	FOL-Folinic Acid, F-Fluorouracil, OX-Oxaliplatin
CP	Carboplatin Paclitaxel
CV	Cisplatin Vinorelbine
FDR	False Discovery Rate
TSS200	200 base-pairs upstream of TSS
TSS1500	1500 base-pairs upstream of Transcription Start Site
5'UTR	5' Untranslated Region
3'UTR	3' Untranslated Region
PKA	Protein Kinase A
EMT	Epithelial-Mesenchymal Transition
LCNEC	Large Cell Neuroendocrine Carcinoma

CHAPTER I

1 INTRODUCTION

1.1 Background, Literature Review and Statement of the Problem

Despite recent advances in diagnosis and treatment, the five-year survival rate in lung cancer (17.7%) is lower than in colon (64.4%), breast (89.7%), and prostate (98.9%) cancer combined[1], which represents a major cause of cancer-related mortality and morbidity in both men and women in the United States[1]. Historically lung cancers were divided into Small cell lung carcinoma (SCLC) and Non-small cell lung carcinoma (NSCLC)[2]. *(i)* SCLC (10-15% of all lung cancers) is an aggressive neuroendocrine tumor consisting of small tumor cells deriving from epithelial and neuroendocrine cells. These are disseminated diseases in which a central tumor mass is usually never found, but which have more favorable prognoses overall than non-small cell lung cancers[3]. *(ii)* NSCLC is a heterogeneous aggregate of malignancies, its represents approximately 80-85% of all lung cancers, carry a worse prognosis, and are subdivided into three major subtypes: lung adenocarcinoma (40% of all lung cancers), squamous cell carcinoma (25% of all lung cancers), and large cell carcinoma (10% of all lung cancers)[4].

Lung adenocarcinoma often occurs in the outer or peripheral areas of the lungs, and frequently develops at multiple sites in the lungs and spread throughout the alveolar

surface[5]. Patients with lung adenocarcinoma in advanced stages (and most recently, in earlier stages) are subjected to standard-of-care adjuvant chemotherapy[6-8] and despite improved survival for a group of patients [9-12], nearly 50% of the patients develop resistance to the administered treatment, which results in advanced metastatic progression and lethality [13,14]. The majority of patients with LUAD lack clinically actionable mutations and are commonly administered a doublet chemotherapy (i.e., platinum-based chemotherapy often combined with plant alkaloids and/or antimetabolites) to improve response rates and survival[15-18]. Most recently, treatment for LUAD has also included immune checkpoint inhibitors, yet they are not curative for most patients[19]. The heterogeneity of response to the standard-of-care therapies and rapidly emerging treatment resistance remain major challenges in lung cancer management. Prioritization of patients based on their risk of developing resistance prior to therapy administration would improve disease course and enhance informed clinical decision making at large.

Resistance to chemotherapy treatment is known to occur through several mechanisms: *(i)* defect in DNA repair mechanisms, including increased nucleotide excision repair, or loss of mismatch repair[20,21]; *(ii)* defect in apoptosis pathway[22-24], including mutation in p53[25,26]; *(iii)* drug transporter ABC (ATP binding cassette) including *P-gp* (P-glycoproteins) and *MDRs* (Multi-drug resistance associated proteins) involved in the efflux of chemotherapeutic drug[27,28]; *(iv)* alteration of enzyme expression (e.g., glutathione, metallothionein) enhance drug sequestration[29-32]; *(v)* alteration of drug target by mutation [33]; and *(vi)* EGFR-mediated *PI3K/Akt* and *NF-kB* pathway dysregulation [34]. Even though these mechanisms have been widely investigated, effective prioritization of

patients for specific chemotherapy regimens has remained a central challenge in clinical oncology.

In recent years, several successful attempts[35-44] have improved classification of LUAD based on markers of overall disease aggressiveness, including mutations in oncogenes (*EGFR* [35], *KRAS*[36], *MET* [41]); proto-oncogenes (*AKT1*[42], *ERBB2*[37], *BRAF*[38]), and tumor suppressor genes (*TP53*[39], *PTEN*[40], *CDKN2A*[43], *STK11*[44]). Despite being successful as prognostic markers of LUAD aggressiveness, they have not been associated with the complexity of therapeutic response yet[45,46], suggesting that more complex mechanisms might be at play in this malignancy.

Recently, multiple transcriptomic and epigenomic alterations have been highlighted to play a role in primary and secondary chemoresistance across various cancer types[47-52]. For example, studies focused on transcriptomic alterations have demonstrated that: *MDR1* amplification is implicated in acquired resistance to anthracyclines, vinca alkaloids, and other antineoplastic chemotherapies in breast cancer[47]; over-expression of dihydrodiol dehydrogenase enzyme is central in resistance to cisplatin in ovarian cancer[48]; and higher genomic instability due to p53 inactivation is essential in resistance to platinum-based chemotherapy in ovarian cancer[49]. In parallel, epigenomic-centered studies have demonstrated that: genome-wide hypermethylation is implicated in resistance to antineoplastic fotemustine in melanoma[50]; hypermethylation of *DKK3* leads to docetaxel resistance in non-small cell lung cancer[51]; and hypomethylation of *MIR663A* induce cyclophosphamide and docetaxel resistance in breast cancer[52]. Given the success of individual transcriptomic and epigenomic determinants of chemoresponse, a systematic genome-wide investigation of the interplay

between transcriptomic and epigenomic mechanisms implicated in resistance can provide valuable predictive markers of predisposition to chemotherapy failure.

In the last decade, several computational methods have been successfully applied to understand cancer initiation and progression through integration of transcriptomic and epigenomic data, including correlation of mRNA expression and DNA methylation and/or copy number variations[53-55], linear regression connecting DNA methylation sites and mRNA expression of the site-harboring genes[56], network-based integration of mRNA expression and DNA methylation and/or copy number variations[57-59]. Even though successful in identifying clinically relevant signatures of disease progression, these methods have not yet fully explored the interplay between transcriptomic and epigenomic mechanisms altered in molecular pathways implicated in chemo response, which would shed light on complex molecular mechanisms that govern therapeutic resistance.

1.2 Research Objectives

In this work, we develop a generalizable computational framework to identify molecular pathways altered on transcriptomic (i.e., mRNA expression) and epigenomic (i.e., DNA methylation) levels that govern resistance to chemotherapy. We name our approach **pathCHEMO** – uncovering transcriptomic and epigenomic **path**ways implicated in **CHEMO**resistance. Our overall idea is that pathways that are altered on both mRNA expression and DNA methylation levels are more likely to capture complex relationships implicated in therapeutic resistance and overcome noise present in any single experiment or data type (see the example of pathway alterations on both transcriptomic and epigenomic levels in Figure. 1).

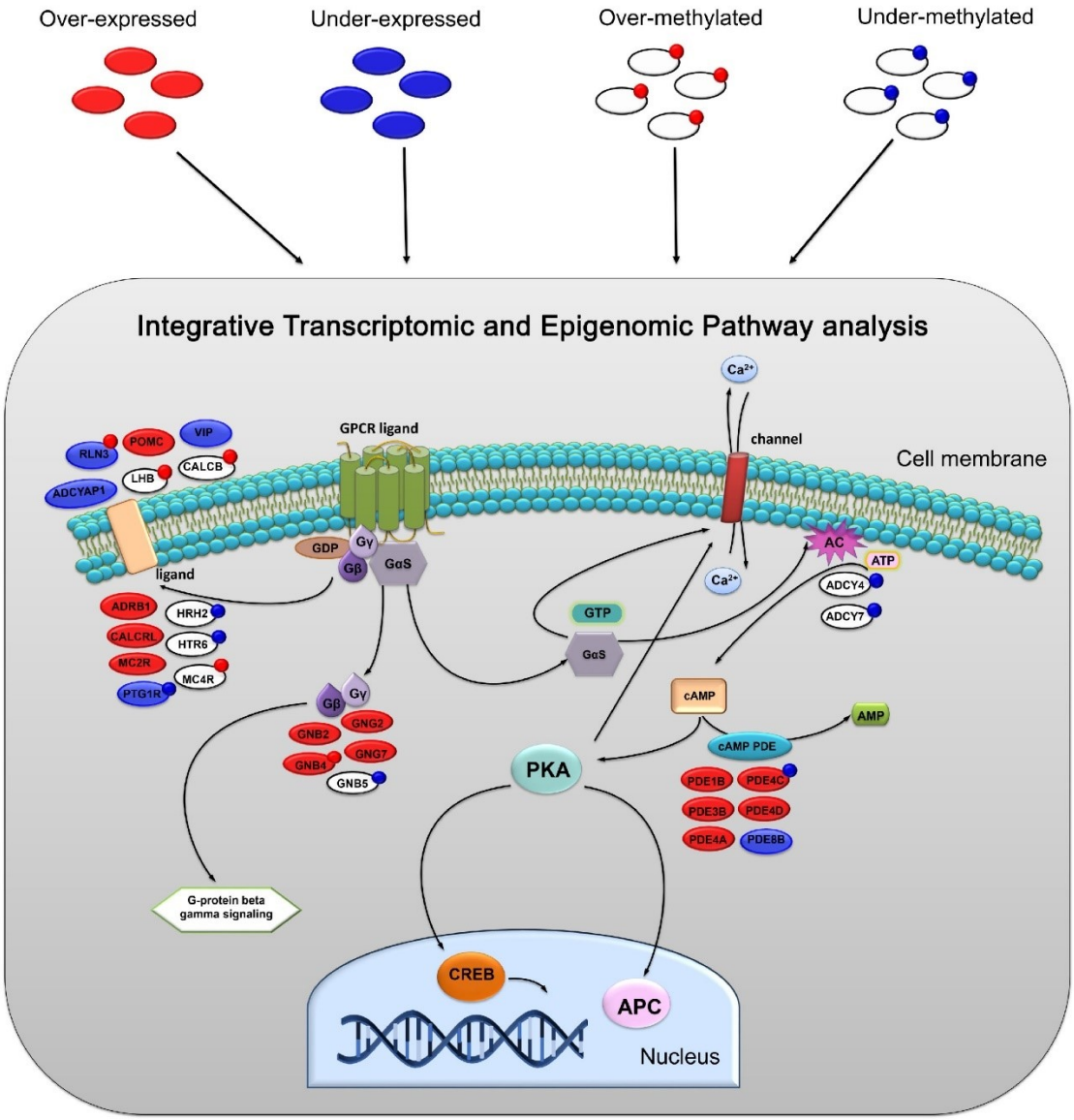


Figure 1. Schematic representation of pathway altered on both transcriptomic and epigenomic levels.

Pathway genes affected on transcriptomic and epigenomic levels in *G alpha (s) signalling events* pathway are represented by ovals, where their colors correspond to either over-expression (*red*), under-expression (*blue*) or no differential expression (*white*). Small satellite circles represent over-methylation (*red*) or under-methylation (*blue*).

In addition, our approach provides several important advantages that tackle complexity of treatment response. First, it uncovers molecular pathways altered on both transcriptomic and epigenomic levels, which increases the likelihood to identify functionally relevant alterations. Second, these pathways can be utilized as effective markers of primary chemoresistance to predict patients with poor and favorable response, even prior to therapy administration. Finally, it uncovers molecular pathways, rather than single determinants, thus providing potential functional candidates for therapeutic intervention to preclude or overcome resistance.

1.3 Overview of the Study

Motivated by the need for markers of chemoresponse in lung cancer, we analyze profiles of patients with LUAD from The Cancer Genome Atlas (TCGA-LUAD)[60], which received adjuvant standard-of-care chemotherapy (i.e., a combination of platinum-based carboplatin and plant alkaloid paclitaxel). pathCHEMO identifies seven molecular pathways altered on transcriptomic and epigenomic levels that differentiate patients with poor and favorable carboplatin-paclitaxel response. We demonstrate that the activity of these pathways as well as their representative read-out genes, can serve as molecular markers to identify patients at risk of resistance to carboplatin-paclitaxel in an independent patient cohort[18] (log-rank p-value = 0.0081, hazard ratio = 10) and can predict the risk of resistance to carboplatin-paclitaxel combination for new patients (i.e., through leave-one-out cross-validation). We also confirm significant non-random predictive ability of our identified seven candidate pathways, when compared to seven pathways selected at random (random model p-value < 0.007) and show that our

approach outperforms other commonly utilized methods (e.g., linear regression, support vector machine, and random forest) in identifying patients at risk of resistance to chemotherapy (Area Under the Receiver Operating Characteristics (AUROC) = 0.98)[56,61,62]. In addition, we demonstrate that our model is independent of, and is not affected by commonly used covariates (i.e., age, gender, and disease stage at diagnosis) and by the known signatures of lung cancer aggressiveness (adjusted hazard ratio = 14, hazard p-value = 0.03).

Finally, we extend our approach to additional chemo combinations (i.e., a combination of platinum-based cisplatin and plant alkaloid vinorelbine, and a combination of platinum-based oxaliplatin and antimetabolite agent fluorouracil) and additional cancer types (i.e., lung squamous cell carcinoma and colorectal adenocarcinoma)[17,63] and demonstrate accuracy and general applicability of our approach (log-rank p-value < 0.03, hazard ratio > 3.5 across cancer types and chemotherapy-regimens). We propose that our model can be used to pre-screen patients and prioritize them for specific chemotherapy treatments.

CHAPTER II

2 METHODS

2.1 Lung Adenocarcinoma patient cohorts

For this study, LUAD patient cohorts were obtained from publicly available data sources (Supplementary Table 1), which include The Cancer Genome Atlas-Lung Adenocarcinoma (TCGA-LUAD)[64], *Tang et al.* (GSE42127)[18], *Der et al.* (GSE50081)[65], and *Zhu et al.* (GSE14814)[17] datasets. The primary LUAD patient cohort, utilized for reconstruction of transcriptomic and epigenomic signatures of chemoresistance, was obtained from The Cancer Genome Atlas (TCGA-LUAD) project[64] and downloaded from the Genomics Data Commons database (GDC, https://portal.gdc.cancer.gov/) on February 2017. Clinical information (i.e., clinical file, follow-up, and treatment data) for these datasets were obtained from the TCGA GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive/).

2.2 Gene expression and DNA methylation analysis

For RNA-seq analysis, we normalized and stabilized variance for raw RNA-seq counts using *DESeq2*[66] R package. DNA methylation values for each site were reported as β (Beta) values, which were subsequently converted to *M*-values as suggested in[67] when parametric analysis was utilized, using *beta2m* function in Lumi[68] R package.

2.3 Defining signatures of chemotherapy response

To determine molecular characteristics that differ between poor response and favorable response, we defined signatures of treatment response on transcriptomic (i.e., differential expression) and epigenomic (i.e., differential methylation) levels between poor response and favorable response patient groups using two-sample two-tailed Welch t-test (*t.test* function in R)[69] in R studio version 3.3.2[70], such that differential expression signature was defined as a list of genes ranked on their differential expression (i.e., t-test values) and differential methylation signature was defined as a list of genes based on the differential methylation of the corresponding site (i.e., t-test values). We coupled this analysis with signatures defined based on a fold change and obtained similar results. For DNA methylation signature, we performed analysis two ways: selected one CpG site per gene through the coefficient of variation analysis, where a site with the highest coefficient of variation was selected for each gene; and considered all CpG sites for signature reconstruction, yielding similar results.

2.4 Transcriptomic and epigenomic pathway enrichment analysis

To identify molecular pathways altered on transcriptomic and epigenomic levels (Figure 1), we first performed pathway enrichment analysis on differential expression signature and differential methylation signature (as in Figure 2). For this, we used the comprehensive C2 pathway database[71] (http://software.broadinstitute.org/gsea/msigdb), which includes 833 pathways from REACTOME[72], KEGG[73], and BIOCARTA[74] databases, and implemented pathway enrichment analysis using Gene Set Enrichment

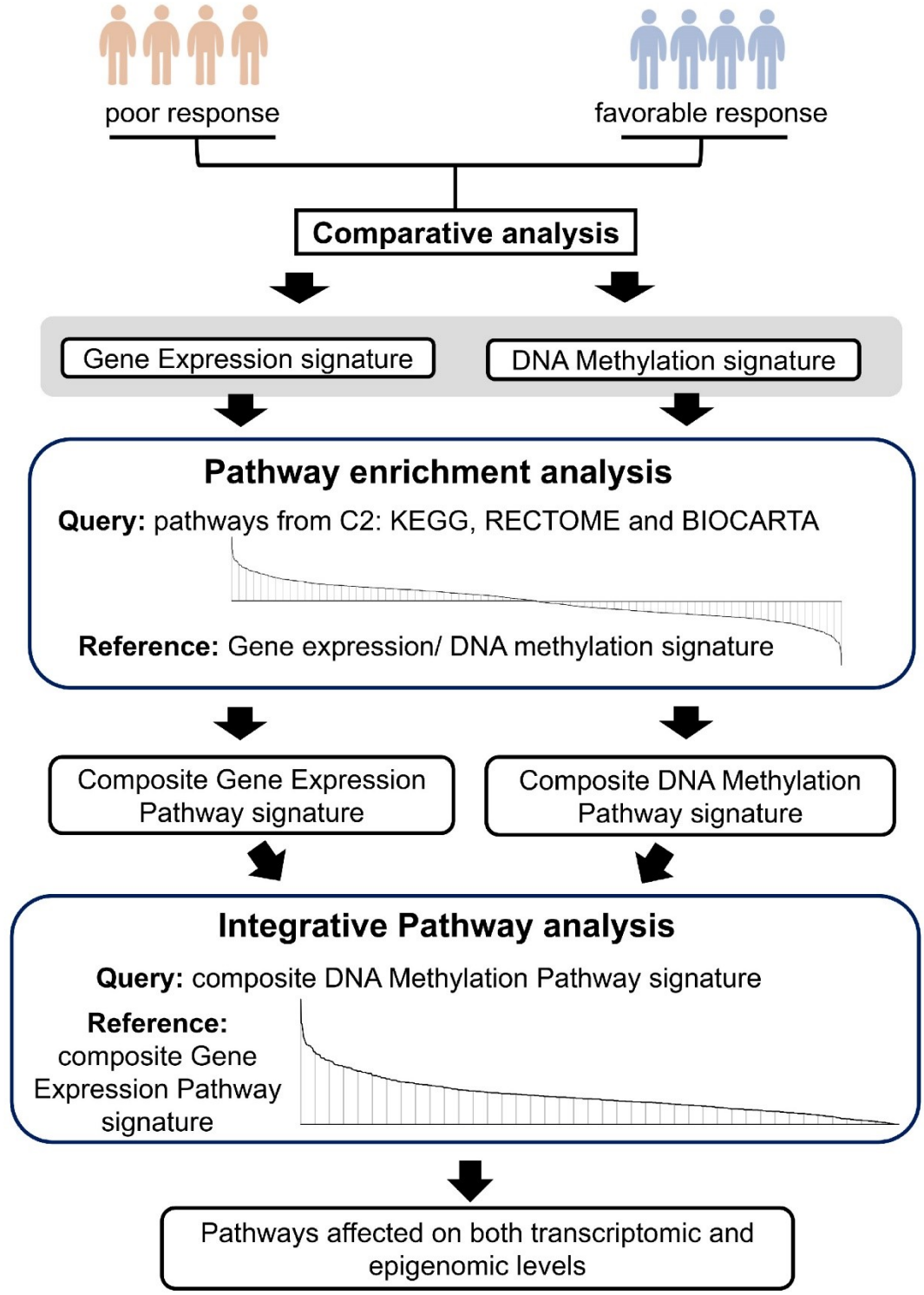


Figure 2. Schematic flow representation of pathCHEMO

Analysis (GSEA)[75], where differential expression and differential methylation signatures were used as a reference and collection of genes from each pathway was used as a query gene set. Normalized Enrichment Scores (NESs), and p-values were estimated using 1,000 gene permutations. This analysis estimated NESs for each of the 833 pathways, which reflects the extent to which each pathway is enriched in the treatment response signature and defines a so-called pathway activity. Positive NES would reflect pathway enrichment in the over-expressed part of the signature (e.g., majority of pathway genes being over-expressed) and negative NES would reflect pathway enrichment in the under-expressed part of the signature (e.g., majority of pathway genes being under-expressed). We refer to such pathway enrichment analysis as signed as it considers over- and under-expression of genes (with direction). Signed pathway enrichment analysis was performed on the differential methylation signature of treatment response in the similar manner.

Further, to overcome limitations of such (i.e., signed) pathway enrichment analysis, which assumes that the pathway will be enriched only if majority of genes in the pathway are changed in the same direction (i.e., either over-expressed or under-expressed, but not both), we performed absolute valued analysis. For this, the pathway enrichment analysis was run on the absolute valued differential expression signature, where signature t-stat values are absolute valued to collapse positive and negative signature tails, as was previously done in[76]. In this case, positive NESs reflect enrichment in the differentially expressed part of the signature (which includes both over-expressed and under-expressed genes) and negative NESs reflect enrichment in the non-differentially expressed part of the signature (and are therefore not considered). This absolute valued pathway enrichment analysis discovers pathways whose genes might be

changed in both directions (both over-expressed and under-expressed) as it estimates the enrichment in the differentially expressed tail of the signature (irrespective of sign). Such absolute valued pathway enrichment analysis defined NESs for each of 833 pathways, as above. Absolute valued pathway enrichment analysis was performed on the differential methylation signature of treatment response in the similar manner (Supplementary Figure 2).

The next essential step was to then integrate NESs from signed and absolute valued pathway enrichment analysis so that for each pathway a final integrative NES was defined as an NES with the lowest p-value between signed and absolute valued pathway analyses (note, that negative NES values for absolute valued analysis are not considered as they reflect enrichment in the non-changed part of the signature). The advantage of such integration is two-fold: it captures pathways whose genes are strictly over-expressed or under-expressed in each pathway, and whose genes are changed in both directions (i.e., such pathway would contain genes that are over-expressed and genes that are under-expressed), thus increasing the probability to identify functionally relevant molecular determinants. Similar logic applies to the methylation signatures. Such integration of signed and absolute valued NESs defined composite expression pathway signature and composite methylation pathway signature (Supplementary Figure 3).

2.5 Transcriptomic and epigenomic pathway integration

We have employed GSEA to compare composite expression pathway signature and composite methylation pathway signature to identify pathways that are affected on both transcriptomic and epigenomic levels (i.e., belong to the leading edge from the

GSEA analysis). To assure that we can identify pathways which are over-expressed and under-methylated; under-expressed and over-methylated; differentially expressed and differentially methylated etc., each pathway signature was ranked based on the absolute values of their NESs and used for subsequent GSEA comparative analysis.

For this pathway-based GSEA, we utilized composite expression pathway signature as a reference signature and top pathways from the composite methylation pathway signature as a query pathway set. To accurately define query pathway set, which should assure strongest enrichment between pathway signatures, we varied the threshold for the query pathway set between 0.001 and 0.05 (width of each step = 0.005) and estimated the strength of enrichment between the two signatures at each threshold. Since GSEA is a probabilistic algorithm, for each threshold, GSEA was run 100 times and average NES for the enrichment was reported. Threshold with the highest average NES then reflects the optimal threshold which corresponds to the strongest enrichment between the composite expression pathway signature and the composite methylation pathway signature, used for subsequent analysis. GSEA analysis between the composite expression pathway signature and the composite methylation pathway signature at the optimal threshold identified a set of pathways (e.g., for carboplatin-paclitaxel response LUAD, we identified 28 pathways) of treatment response altered on both transcriptomic and epigenomic levels.

One of the limitations of the pathways from the C2 collection is that they often represent a parent-child relationship, where a parent pathway (e.g., cell cycle) would encompass all genes in its child pathways (i.e., cell cycle phase). Such overlap produces data redundancy and can result in model overfitting as the same pathways are fit in the

model repeatedly. To overcome this limitation and to eliminate pathways with heavy overlaps, we performed Fisher Exact Test (*fisher.test* function in R) and compared leading edge genes for each pair of pathways from our analysis (e.g., for all 28 pathways, resulting in (28 choose 2 = 378) comparisons). From each group of parent-children pathways which shared a large number of overlapping genes, we selected one representative pathway with the NES corresponding to the lowest p-value, which defined a final set of pathways (e.g., for carboplatin-paclitaxel response LUAD, we identified seven pathways) maximally non-overlapping non-redundant pathways used for subsequent analysis.

2.6 Comparing expression and methylation predictive ability

To examine if, in our candidate pathways, both data types (i.e., mRNA expression or DNA methylation) have equivalent ability to predict therapeutic response, we compared the performance of candidate pathways utilizing their activity levels based on expression only and activity levels based on methylation only, separately. To compare pathway performances based on each data type, we first scaled both expression and methylation data matrices (i.e., *z-scored* on genes or sites) in the discovery (i.e., TCGA-LUAD) cohort, which defined single-sample differential expression and single-sample differential methylation signatures, respectively (Supplementary Figure 4). Each sample was then used for signed and absolute valued pathway enrichment analysis (separately for expression and for methylation, as above), where each single-sample signature was used as a reference and genes from each of seven candidate pathways were used as a query set thus producing a pathway activity signature for each patient. These single-sample

expression and methylation pathway signatures were then used to evaluate predictive ability of seven pathways (for expression and methylation, separately), using logistic regression modeling followed by ROC analysis. The logistic regression analysis was done using *glm* function and ROC analysis was done using *pROC*[77] and *ggplot2*[78] package in R.

2.7 Validation and robustness in independent clinical cohorts

To evaluate clinical significance of the candidate molecular pathways, we examined their ability to predict patients at risk of chemoresistance in an independent clinical cohort from the *Tang et al.*[18] dataset, and used survival status during the clinical study (1996 to 2007) as a clinical endpoint (time to event or follow-up was estimated between the start of carboplatin-paclitaxel treatment and death or follow-up, respectively; maximum time to event/follow-up is 2,567 days). First, we estimated activity levels of the candidate pathways in the independent clinical *Tang et al.* cohort on a single-sample level, as above. The activity levels (i.e., NESs) of the candidate pathways were then subjected to t-Distributed Stochastic Neighbor Embedding (t-SNE) clustering[79] (implemented through *Rtsne*[80] package in R), a non-linear dimensionality reduction technique which chooses two similarity measures between pairs of points of high dimensional input space and low-dimensional embedding space. First, it constructs a probability distribution over the pairs of high dimensional space (i.e., 7-dimension in our case) in such a way that similar points are exhibited by nearby instances, while dissimilar points are exhibited by distant instances. Second, it constructs a similar probability distribution over the points in low-dimensional embedding space and tries to minimize

the Kullback-Leibler divergence[81] (i.e., KL divergence) between the high dimensional data and low dimensional anticipated data at each point. Therefore, patients with similar pathway activity levels will be anticipated as nearby instances while patients with dissimilar pathway activity levels will be anticipated as dissimilar instances. The advantage of t-SNE lies in its ability to reduce dimensions from seven (maximum possible in our analysis) to two and effectively identify groups of patients that share similar pathway activity levels. This analysis stratified patients into two groups: a group with overall increased composite pathways' activities and a group with overall decreased composite pathways' activities. We then evaluated if these patient groups differ in their response to carboplatin-paclitaxel treatment using Kaplan-Meier survival analysis[82], and Cox proportional hazards model[83] via *survival*[84], *ggplot2*[78] and *survminer*[85] R packages.

In order to evaluate if a random set of pathways can perform as well as our identified seven pathways, we compared the predictive ability of our seven candidate pathways to the predictive ability of seven pathways selected at random. For this, we built a random model, where seven pathways were selected at random and their activity levels were utilized to stratify patients based on their treatment response, with subsequent evaluation using Kaplan-Meier survival analysis. Random selection was done 10,000 times and the empirical p-value was estimated as a number of times Kaplan-Meier log-rank p-value for seven candidate molecular pathways outperformed the results at random. We have also employed a second random model, where we evaluated the effect of selecting random patient groups.

Finally, to estimate the accuracy with which our model can predict treatment response for a new incoming patient, we simulated this process using leave-one-out

cross-validation (LOOCV)[86]. In LOOCV, one patient is removed; and the model is trained on the rest of the patients. The patient that was removed is considered as a new incoming patient, subjected to predictive analysis, and is assigned a risk of developing resistance. This process was repeated for all patients. We implemented the predictive model for LOOCV using generalized linear modeling (e.g., utilizing multivariable logistic regression) through *glm*[87] function and *ggplot2*[78] package in R.

2.8 Comprehensive comparative analysis

To assess advantages of our approach, we have compared its predictive performance to other commonly utilized approaches, including linear regression modeling, support vector machine, and random forest; and evaluated if our approach can be affected by commonly used covariates or known signatures of lung cancer aggressiveness.

To demonstrate the advantages of our approach over other commonly utilized methods, we compared its performance: first, to *Panja et al.*[56] method, Epigenomic and Genomic mechanisms of treatment Resistance (Epi2GenR), which utilized linear regression to integrate DNA methylation and mRNA expression data; second, to *Zhong et al.*[62] method, based on support vector machine (SVM) algorithm which utilized mRNA expression patient profiles; and finally, to *Yu et al.*[61] method, Personalized REgimen Selection (PRES) method, based on random forest machine learning approach which utilizes mRNA expression patient profiles. We followed the selection and cross-validation techniques suggested in each of the above publications to carefully compare their performance to our approach. Epi2GenR utilized the same signature as utilized in

our study. To apply SVM and PRES correctly, we split our validation set into 70:30 proportion subsets, where 70% of the validation set were used for model training and 30% for model validation. Predictive ability of the identified candidates from each of these methods was evaluated using ROC, Kaplan-Meier survival, and hazard ratio analyses through *survival*[84], *survcomp*[88], and *survminer*[85] packages in R.

Next, we evaluated if any of commonly used covariates (i.e., age, gender, and disease stage at diagnosis) and known signatures of lung cancer aggressiveness (from *Larsen et al.*[89], *Beer et al.*[90], and *Tang et al.*[18] described above) can predict therapeutic response or can affect predictive ability of the identified seven candidate pathways. For this, we utilized the multivariable Cox proportional hazards model[83] (using *coxph* function in R) and stratified Kaplan-Meier survival analysis through survival, and *survminer* packages in R.

2.9 Pathway activity read-outs

To identify pathway read-outs, we looked for genes inside each pathway, which were altered on transcriptomic and/or epigenomic levels (i.e., belong to the leading edge from the pathway enrichment analysis); correlated with pathway activity levels (i.e., correlation between NESs and a candidate gene across all patients, measured by Pearson correlation, *cor.test* function in R); and associated with carboplatin-paclitaxel response (i.e., Cox proportional hazards model through *coxph* in R, using likelihood-ratio test as reliable for small sample sizes[91]). Likelihood-ratio test p-values were then combined with Pearson correlation p-values using Fisher's method (*metap* R package) and utilized to

make final gene selection. Visualization of the resulting read-outs was done using Cytoscape[92].

2.10 Model generalizability

To test the generalizability of our model, we applied our method to additional chemotherapy combinations (i.e., cisplatin-vinorelbine and oxaliplatin-fluorouracil) and additional cancer types (i.e., lung squamous cell carcinoma and colorectal adenocarcinoma) (Supplementary Table 2-4). In particular, we investigated response to: cisplatin (platinum-based alkylating chemotherapy) and vinorelbine (non-platinum based plant alkaloid chemotherapy) response in lung adenocarcinoma (LUAD); cisplatin-vinorelbine response in lung squamous cell carcinoma (LUSC); and oxaliplatin (platinum-based alkylating chemotherapy), fluorouracil (antimetabolite chemotherapy) and, folinic acid (chemotherapy protective drug often given with fluorouracil to improves the binding; also known as leucovorin) (i.e., FOLFOX) response in colorectal adenocarcinoma (COAD).

For signature development, we utilized primary tumor samples from TCGA-LUAD/TCGA-LUSC/TCGA-COAD (n = 8), for patients without neo-adjuvant treatment (i.e., no pre-treatment), who received adjuvant chemotherapies of interest and were further monitored for new tumor events (as defined above).

For clinical validation of response to cisplatin-vinorelbine combination in LUAD we utilized the *Zhu et al.* patient cohort [17] (GSE14814), which included LUAD tumors obtained at surgery (n = 39), treated with adjuvant cisplatin-vinorelbine chemotherapy. In this cohort, lung cancer-related death was used as a clinical endpoint and time to event

was calculated between the start of cisplatin-vinorelbine treatment and lung-cancer related death (for patients with this event) or to follow-up (for censored patients), with maximum time to event/follow-up 3,390 days.

For clinical validation of response to cisplatin-vinorelbine combination in lung squamous cell carcinoma (LUSC) we utilized a different subset of patients from *the Zhu et al.* patient cohort[17] (GSE14814), which were patient with LUSC whose tumors were obtained at surgery (n = 26) and who were treated with adjuvant cisplatin-vinorelbine chemotherapy. In this cohort, lung cancer-related death was used as a clinical endpoint and time to event was calculated between the start of cisplatin-vinorelbine treatment and lung-cancer related death (for patients with this event) or to follow-up (for censored patients), with maximum time to event/follow-up 3,318 days.

Finally, for validation of FOLFOX combination in colorectal adenocarcinoma (COAD) we utilized *Marisa et al.* patient cohort[63] (GSE39582), which includes COAD tumors obtained at surgery (n = 23), treated with adjuvant FOLFOX chemotherapies. In this cohort, relapse-free survival (i.e., where relapse was defined as locoregional or distant recurrence) was used as a clinical endpoint and time to event was calculated between the start of FOLFOX treatment to relapse (for patients with this event) or to follow-up (for censored patients), with maximum time to event/follow-up 2,790 days.

To investigate pathways overlaps, we employed Fisher Exact Test (fisher.test function in R) on the leading edge genes from the transcriptomic and epigenomic pathways (i.e., genes that contribute to the enrichment of biological pathways in corresponding signatures). All resulting p-values are corrected for multiple hypotheses testing using FDR.

CHAPTER III

3 RESULTS

3.1 pathCHEMO Overview

We have developed a genome-wide computational approach pathCHEMO that integrates mRNA expression and DNA methylation patient profiles to identify pathways altered on both transcriptomic and epigenomic levels (as demonstrated in Figure 1) that differentiate poor from favorable response to chemotherapy-regimens. Here, we briefly outline the major steps of our integrative algorithm (Figure 2). Step1: our algorithm identifies two groups of patients, which will be used to define a chemotherapy response signature: patients that failed a specific chemotherapy-regimens (e.g., developed metastasis within 1 year after therapy administration), and patients with favorable chemotherapy response (e.g., remained disease-free for more than 2 years after chemotherapy administration). Step 2: it compares transcriptomic (mRNA expression) and epigenomic (DNA methylation) profiles between these two groups of patients, which define differential transcriptomic signature and differential epigenomic signature of chemoresponse. Step 3: Such signatures are then individually subjected to signed and absolute valued pathway enrichment analyses, which are then integrated and define molecular pathways affected in either one direction (i.e., containing either over-expressed or under-expressed genes) or both directions (i.e., containing both over-expressed and

under-expressed genes) enriched in the transcriptomic signature, and similarly pathways affected in either one direction or both directions on the epigenomic level, enriched in the epigenomic signature. Step 4: These transcriptomic and epigenomic pathway signatures are then integrated to define a set of pathways that control both transcriptomic and epigenomic programs disrupted in resistance. Step 5: Such candidate pathways and their read-out genes are subjected to validation studies, where they are evaluated for their ability to predict therapeutic response in independent patient cohorts, through multivariable survival analysis. Step 6: Finally, the identified pathways are used to assign individual risk of resistance for new incoming patients.

3.2 Defining molecular signatures of chemoresponse

We tested our approach to evaluate response to standard-of-care doublet chemotherapy, which contains carboplatin and paclitaxel (i.e., carboplatin-paclitaxel), in LUAD patients. For this, we have analyzed clinical and molecular profiles of patient with LUAD in the TCGA clinical cohort[60]. To study primary resistance to this chemo combination, we specifically selected primary tumors from patients that did not receive any neoadjuvant therapy, were treated with adjuvant carboplatin-paclitaxel chemo regimen, and were further monitored for disease progression (n = 14) (Supplementary Table 1). Each patient that received carboplatin-paclitaxel was evaluated for his/her time to tumor relapse defined as time between the start of carboplatin-paclitaxel administration and a new tumor event (defined as tumor re-occurrence, local or distant metastases). To accurately uncover signal that differentiates poor from favorable treatment response, we employed an extreme-responder analysis, widely utilized by us[56,76,93] and others[94,95], where two groups of patients with drastically different treatment response (i.e., favorable

response and poor response) are compared for differences in their molecular profiles to capture the most prominent molecular signal. To assure that the comparison groups are balanced with respect to initial age, gender, disease stage at diagnosis (i.e., initial disease aggressiveness), smoking status etc., we performed stratified sub-sampling (which identifies patient groups with similar distributions for these variables) and identified

Table 1. Clinical profiles of carboplatin-paclitaxel treated patients from the TCGA-LUAD cohort (n = 8).

Treatment response	Patient ID	Time to event or follow-up (days)	Age	Gender	Disease stage at diagnosis	Smoking status	Observed treatment related event or follow-up
poor response	6712	116	71	male	IIA	4	new tumor event
	5051	122	42	female	IIIA	4	new tumor event
	6979	138	59	female	IIB	3	new tumor event
	A4VP	153	66	female	IIIA	4	new tumor event
favorable response	4666	744	52	female	IV	4	no event, follow-up
	5899	784	58	male	IIA	2	no event, follow-up
	1678	1,120	70	female	IIB	3	no event, follow-up
	1596	2,031	55	male	IIB	2	no event, follow-up

Notes: NA = not available.
Smoking status: 1 = lifelong non-smoker (< 100 cigarettes smoked in Lifetime), 2 = current smoker (includes daily smokers and non-daily smokers (or occasional smokers), 3 = current reformed smoker for > 15 years, 4 = current reformed smoker for ≤ 15 years, 5 = current reformed smoker, duration not specified, and 6 = smoking history not documented.

patients that experienced relapse within 1 year of carboplatin-paclitaxel start (i.e., poor response, $n = 4$); and patients that did not experience any events for more than 2 years (i.e., favorable response, $n = 4$) (Table 1).

To uncover a complex interplay between transcriptomic and epigenomic mechanisms implicated in response to chemotherapy, we compared poor response and favorable response groups based on their mRNA expression and DNA methylation profiles using two-sample two-tailed Welch t-test[69] and re-confirmed with fold change (as described in the Methods section), which defined carboplatin-paclitaxel response differential gene expression signature and carboplatin-paclitaxel response differential methylation signature.

3.3 Integrative analysis identified pathways of resistance

To understand molecular mechanisms that govern chemoresponse, we next sought to identify molecular pathways that control transcriptomic and epigenomic signatures of carboplatin-paclitaxel resistance (Figure 1). For this, we subjected the carboplatin-paclitaxel response differential expression signature and carboplatin-paclitaxel response differential methylation signature to pathway enrichment analysis using the comprehensive C2 pathway database[71] (which includes 833 pathways from REACTOME[72], KEGG[73], and BIOCARTA[74] databases). Pathway enrichment was performed using Gene Set Enrichment Analysis (GSEA)[75]. This analysis estimated Normalized Enrichment Score (i.e., NES) for each of the 833 pathways, which reflects the extent to which each pathway is enriched in the treatment response signature, also referred to as pathway activity. A list of 833 pathways ranked by their enrichment (i.e., NESs) in the carboplatin-paclitaxel response differential expression signature defined

carboplatin-paclitaxel response differential expression pathway signature and a list of 833 pathways ranked by their enrichment (i.e., NESs) in the carboplatin-paclitaxel response methylation signature defined carboplatin-paclitaxel response differential methylation pathway signature (as described in the Methods section). To account both for the pathways that have majority of their genes affected in the same direction (e.g., majority of genes being either over-expressed or under-expressed) and pathways that have genes affected in different directions: some genes affected in one direction (e.g., over-expressed) and some in an opposite direction (e.g., under-expressed), we have performed both signed and absolute valued pathway enrichment analysis with their subsequent integration (as described in the Methods section), which defined carboplatin-paclitaxel response composite expression pathway signature and carboplatin-paclitaxel response composite methylation pathway signature.

Further, to define interplay between complex mechanisms implicated in chemoresistance, we sought to identify molecular pathways that are affected on both transcriptomic (i.e., mRNA expression) and epigenomic (i.e., DNA methylation) levels and which would capture pathway genes affected: only on transcriptomic level, only on epigenomic level, or both levels (as in Figure 1). To achieve this goal (Figure 3a), we compared the carboplatin-paclitaxel response composite expression pathway signature (as a reference) and carboplatin-paclitaxel response composite methylation pathway signature (as a query pathway set) using GSEA (the threshold for the query pathway set at p-value ≤ 0.001 was selected as in Figure 3b, as described in the Methods section), which identified seven molecular pathways with significant alterations on both transcriptomic and epigenomic levels (GSEA NES = 2.75, p-value < 0.001) (Figure 3c,

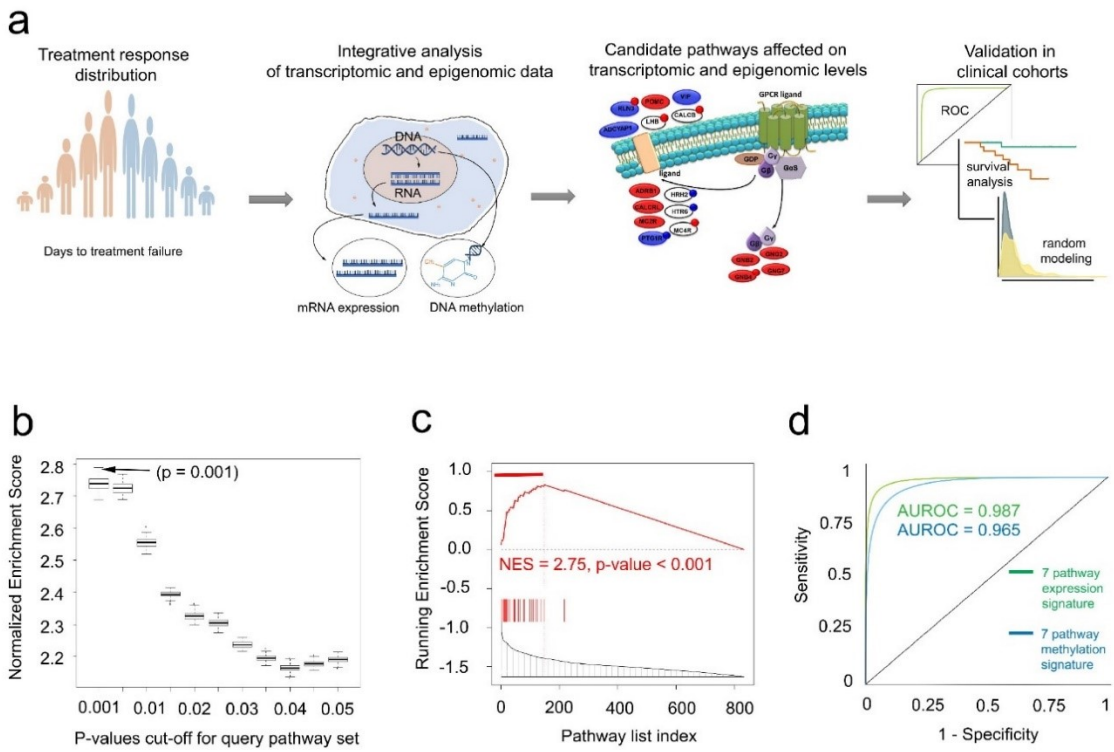


Figure 3. Integrative genome-wide transcriptomic and epigenomic analysis identifies candidate molecular pathways of chemotherapy response.

(a) Schematic representation of the integrative transcriptomic and epigenomic analysis: *(i)* patients are defined by their response to chemotherapy, *(ii)* analysis of transcriptomic and epigenomic patient profiles, *(iii)* integrative transcriptomic and epigenomic analysis identifies candidate pathways affected on both transcriptomic and epigenomic levels, and *(iv)* multi-modal validation of candidate pathways. **(b)** Box and whisker plot depicting p-value cutoff for query carboplatin-paclitaxel response composite methylation pathway signature (x-axis) and NESs from the corresponding GSEA comparison between composite methylation and expression pathways signatures (y-axis), based on analysis in TCGA-LUAD patient cohort. Arrow indicated optimal p-value threshold, which results in the strongest GSEA enrichment. **(c)** GSEA comparing carboplatin-paclitaxel response composite expression pathway signature (reference) and carboplatin-paclitaxel response composite methylation pathway signature (query, NES $p \leq 0.001$), based on analysis in TCGA-LUAD patient cohort. Horizontal red bar indicates leading edge pathways altered on both transcriptomic and epigenomic levels. NES and p-value were estimated using 1,000 pathway permutations. **(d)** ROC analysis comparing ability of the seven candidate pathways to predict carboplatin-paclitaxel where their activity is defined based on their expression values (*green*) or methylation values (*blue*). AUROC is indicated.

as described in the Methods section). These pathways included *chemokine receptors bind chemokines*, *mRNA splicing*, *G alpha (s) signalling events, intestinal immune network for IgA production*, *metabolism of proteins*, *RNA degradation*, and *cell cycle mitotic.*

To confirm that these identified seven molecular pathways are robust to the choice of the statistical methods used to define treatment response signatures, we have also performed our analysis using signatures defined using all DNA methylation sites and using non-parametric tests. First, we defined differential methylation signature with all DNA methylation sites considered (Supplementary Figure 1a). Second, we defined differential methylation signature using fold change (Supplementary Figure 1b). Finally, we defined both differential expression and differential methylation signatures using fold change (Supplementary Figure 1c). Analyses using all of these signatures identified the same seven candidate pathways (GSEA NES > 2.45, p-value < 0.001), demonstrating robustness of our analysis regardless of the signature choice.

To investigate if mRNA expression or DNA methylation carries more weight in the predictive ability of our seven candidate pathways, we have performed Receiver Operating Characteristic (ROC) analysis[96] based on pathway activities in each patient sample (i.e., through single-sample pathway analysis, as described in the Methods section), defined on either expression levels or methylation levels of the pathway genes (as described in the Methods section). The predictive ability was measured using Area under ROC (AUROC), which reflected how well each data type separates poor response and favorable response patients in the TCGA-LUAD patient cohort (the AUROC value of 0.5 indicates random predictor and 1 indicates a perfect predictor). Our analysis demonstrated that both expression levels (AUROC = 0.987) and methylation levels

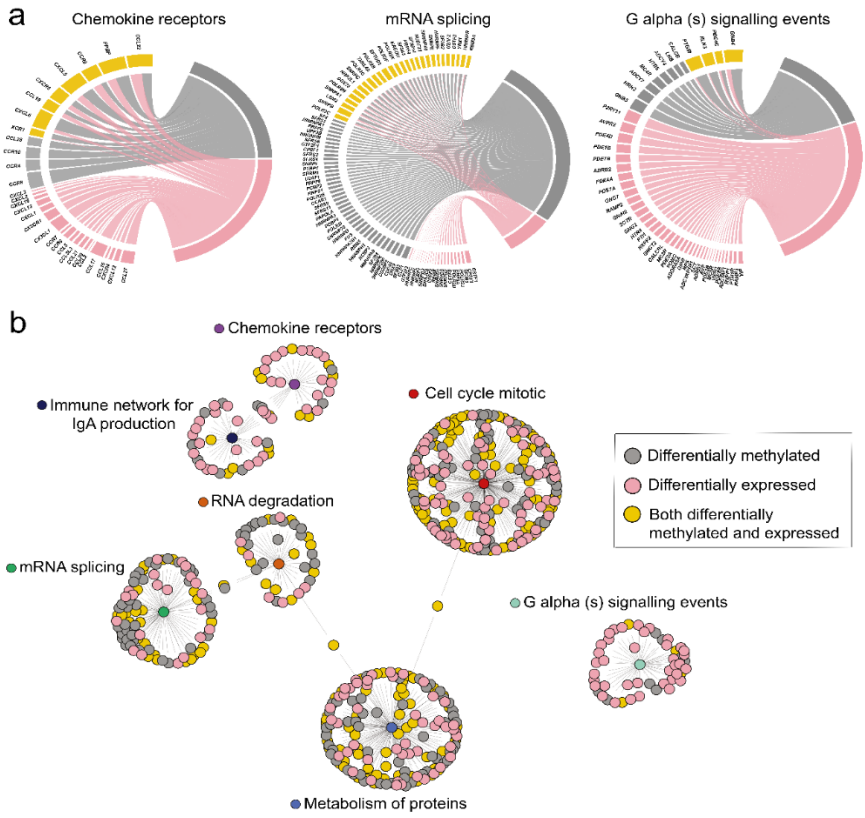


Figure 4. Transcriptomic and epigenomic alterations in candidate pathways of carboplatin-paclitaxel response.

(a) Representative molecular pathways altered on both transcriptomic and epigenomic levels, visualized through *circlize* R package. Genes from the leading edge in each pathway are represented as differentially expressed (*pink*), methylated (*grey*) and both differentially expressed and methylated (*yellow*). Width of each connecting line is proportional to the extent of differential expression and differential methylation. Pathways are depicting as follows: *(i) chemokine receptors bind chemokines* pathway (19 differentially expressed genes, 4 differentially methylated genes, and 8 differentially expressed and methylated genes), *(ii) mRNA splicing* pathway (21 differentially expressed genes, 39 differentially methylated genes, and 28 differentially expressed and methylated genes), and *(iii) G alpha (s) signalling events* pathway (37 differentially expressed genes, 8 differentially methylated genes, and 4 differentially expressed and methylated genes). **(b)** In the seven candidate pathway network visual representation, nodes correspond to the genes, which are connected to central pathway-membership circles (i.e., indicating pathway membership). Gene colors describe differential expression (*pink*), differential methylation (*grey*) and both differential expression and methylation (*yellow*). Network was constructed with *ggnetwork* R packages.

(AUROC = 0.965) of seven candidate pathways are highly predictive of poor response vs. favorable response separation (Figure 3d), indicating that they both can be used to identify patients at risk of developing chemoresistance.

We further evaluated a topological structure of transcriptomic and epigenomic alterations within each identified pathway. Firstly, we examined to which extent genes from each pathway were affected on transcriptomic or on epigenomic levels (Figure 4a, and Supplementary Figure 5) and have observed that seven pathways exercised different patterns of transcriptomic and epigenomic alterations. For example, majority of genes from *G alpha (s) signalling events* pathway were altered on their mRNA level (i.e., Figure 4a, nodes in *pink*) while genes from the *mRNA splicing* pathway were heavily altered on DNA methylation level (Figure 4a, nodes in *grey*) and on both mRNA expression and DNA methylation levels (Figure 4a, nodes in *yellow*). Secondly, we examined connectivity within and between the pathway genes, where an edge within the pathway corresponds to the pathway membership and connecting edge between pathways shows shared genes and demonstrated that our candidate pathways share little overlap (Figure 4b). Finally, we examined differentially methylated sites harbored in genes from the seven pathways and evaluated their regions/locations on the genome (Figure 5a), where regions were defined as TSS200 (i.e., 200 base pairs upstream of transcription start site, TSS), TSS1500 (i.e., 1500 base pairs upstream of TSS200), 5'UTR, 1st exon, gene body,

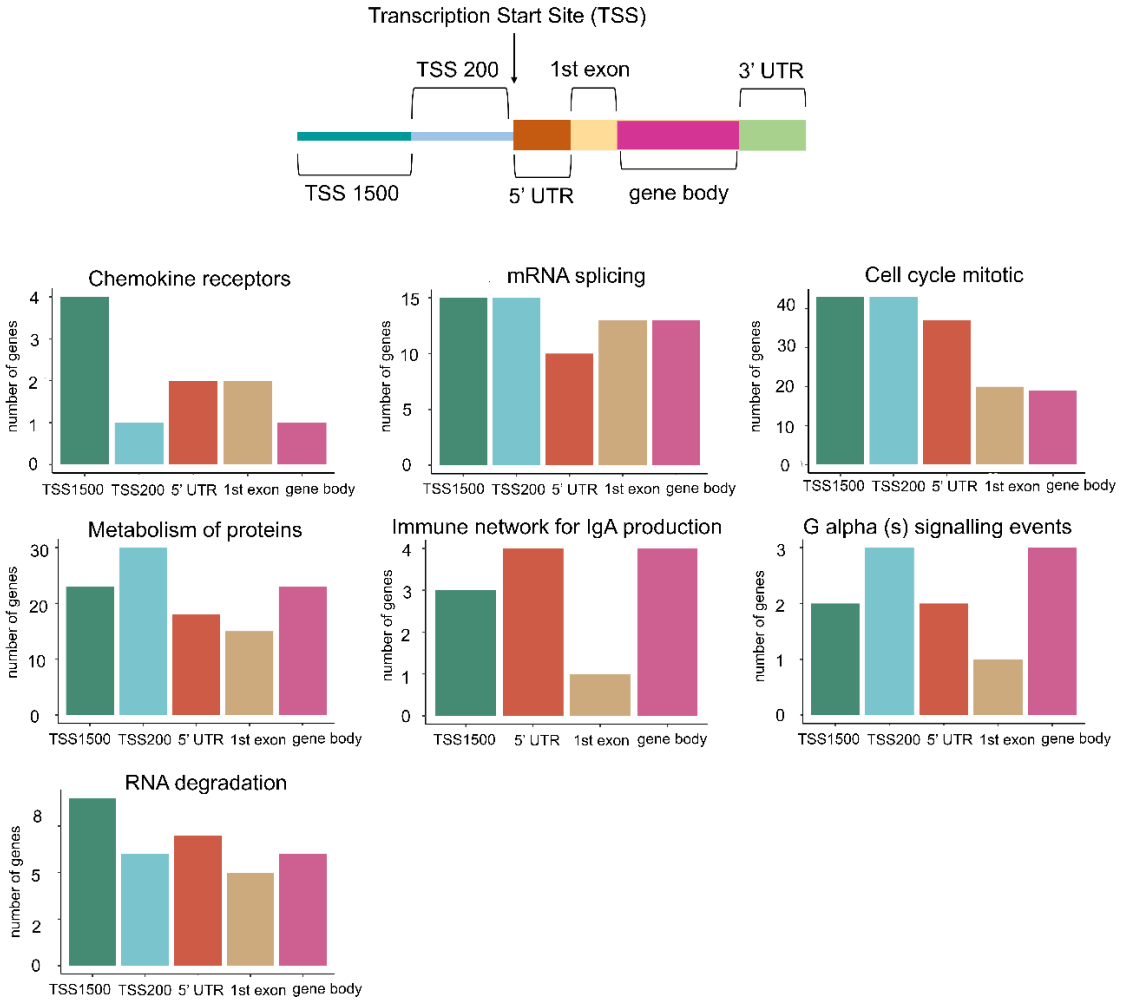


Figure 5. Region-based analysis of differentially methylated sites in seven candidate pathways.

(a) Schematic representation of regions (TSS1500, TSS200, 5'UTR, first exon, gene body, and 3'UTR) used to profile differentially methylated sites in the HumanMethylation450 array. **(b)** Bar plot representation of region distribution for pathway genes harboring differentially methylated sites.

and 3'UTR. In fact, the majority of pathways have methylated sites overrepresented in TSS200+TSS1500 regions, indicating a possible interaction with the transcription machinery binding at the promoter/enhancer regions[97]. An exception was *Immune*

network for IgA production pathway, whose sites were heavily enriched in the gene body, indicating their potential interaction with alternative splicing machinery[98] (Figure 5b).

3.4 Validation in independent patient cohorts

Our next essential step was to evaluate if the candidate molecular pathways can stratify patients based on the risk of failing chemotherapy in an independent, non-overlapping patient cohort (Figure 6a). For this, we first considered a *Tang et al*. cohort[18] (Supplementary Table 1) from the University of Texas MD Anderson Cancer Center, which contains LUAD primary tumor samples obtained at surgery (n = 39) collected between 1996 to 2007, followed by treatment with carboplatin and a taxane (e.g., paclitaxel) and monitored for further disease progression for 11 years. In this cohort, survival status during the clinical study (1996 to 2007) was used as a clinical endpoint and time to this event was calculated between the start of carboplatin-paclitaxel treatment to death (for patients with this event) or to follow-up (for censored patients). Similar to the analysis above, we evaluated activity levels of seven candidate pathways in each patient sample (i.e., through single-sample pathway analysis, as described in the Methods section) and employed t-Distributed Stochastic Neighbor Embedding (t-SNE) clustering[79], which stratified patients into two groups based on pathway activity levels (Figure 6b): one group with increased composite pathways' activities (*orange*) and one group with decreased composite pathways' activities (*green*). We then subjected these patient groups to Kaplan-Meier survival analysis and Cox proportional hazards model (Fig. 4c), which demonstrated that these groups had a significant difference in their

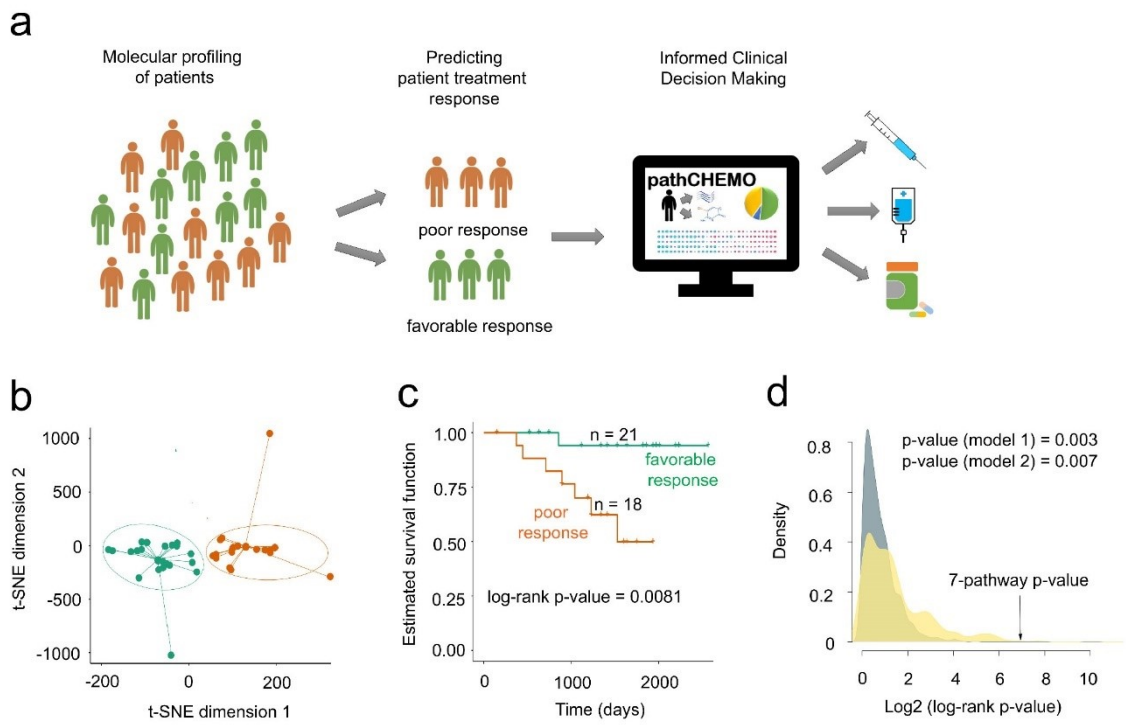


Figure 6. Candidate molecular pathways stratify patients based on response to carboplatin-taxane in an independent cohort.

(a) Validation strategy, as follows: *(i)* molecular transcriptomic and epigenomic profiling of patients, *(ii)* predicting patients' risk of developing chemoresistance, and *(iii)* informed clinical decision making based on patients personalized risks. ***(b)*** t-SNE clustering of lung adenocarcinoma patients treated with carboplatin-taxane (e.g., paclitaxel) from the *Tang et al.* validation cohort (n = 39 biologically independent patient samples), based on activity levels of seven candidate pathways. Among two groups green group (n = 21 biologically independent patient samples) corresponds to patients with low composite activity levels of candidate pathways and orange group (n = 18 biologically independent patient samples) corresponds to patients with high composite activity levels of candidate pathways. ***(c)*** Kaplan-Meier survival analysis to estimate difference in response to carboplatin-taxane (e.g., paclitaxel) between two patient groups is identified in **(b)**. Log-rank p-value and number of patients in each group are indicated. ***(d)*** Two random models indicate non-random predictive ability of our model in the *Tang et al.* validation cohort: random model 1 (*steel-blue*) is defined based on to seven pathways selected at random, and random model 2 (*goldenrod*) is defined based on to equally–sized patient groups selected at random.

response to carboplatin-paclitaxel (log-rank p-value = 0.0081, hazard ratio = 10) (as described in the Methods section).

To evaluate non-randomness of this result, we compared predictive ability of our candidate seven pathways to the predictive ability of seven pathways selected at random (as described in the Methods section), which demonstrated that ability of the candidate seven pathways to predict carboplatin-paclitaxel response is highly non-random compared to 10,000 randomly selected pathways (Figure 6d, random model 1: p-value = 0.003). We paralleled this analysis with evaluation if patient groups stratified by our model are different in their treatment response compared to patient groups chosen at random, which were shown to be highly non-random (Figure 6d, random model 2: p-value = 0.007).

Further, we simulated a situation when a new incoming patient is diagnosed with LUAD and needs to be assigned the risk of developing resistance to carboplatin-paclitaxel utilizing leave-one-out cross-validation (LOOCV)[86] in the *Tang et al.* validation cohort[18]. In LOOCV, one patient is removed, and the model is trained on the rest of the patients. Then the patient that was removed is subjected to predictive analysis and is assigned a risk of developing resistance (i.e., simulating a scenario of a new incoming patient). This process is repeated for all patients (as described in the Methods section). LOOCV analysis demonstrated that our model has high accuracy in as described in the Methods section predicting poor and favorable carboplatin-paclitaxel response for a new incoming patient (Figure 7a).

Finally, to determine that our candidate pathways specifically distinguish carboplatin-paclitaxel response and not disease aggressiveness, we have evaluated if the

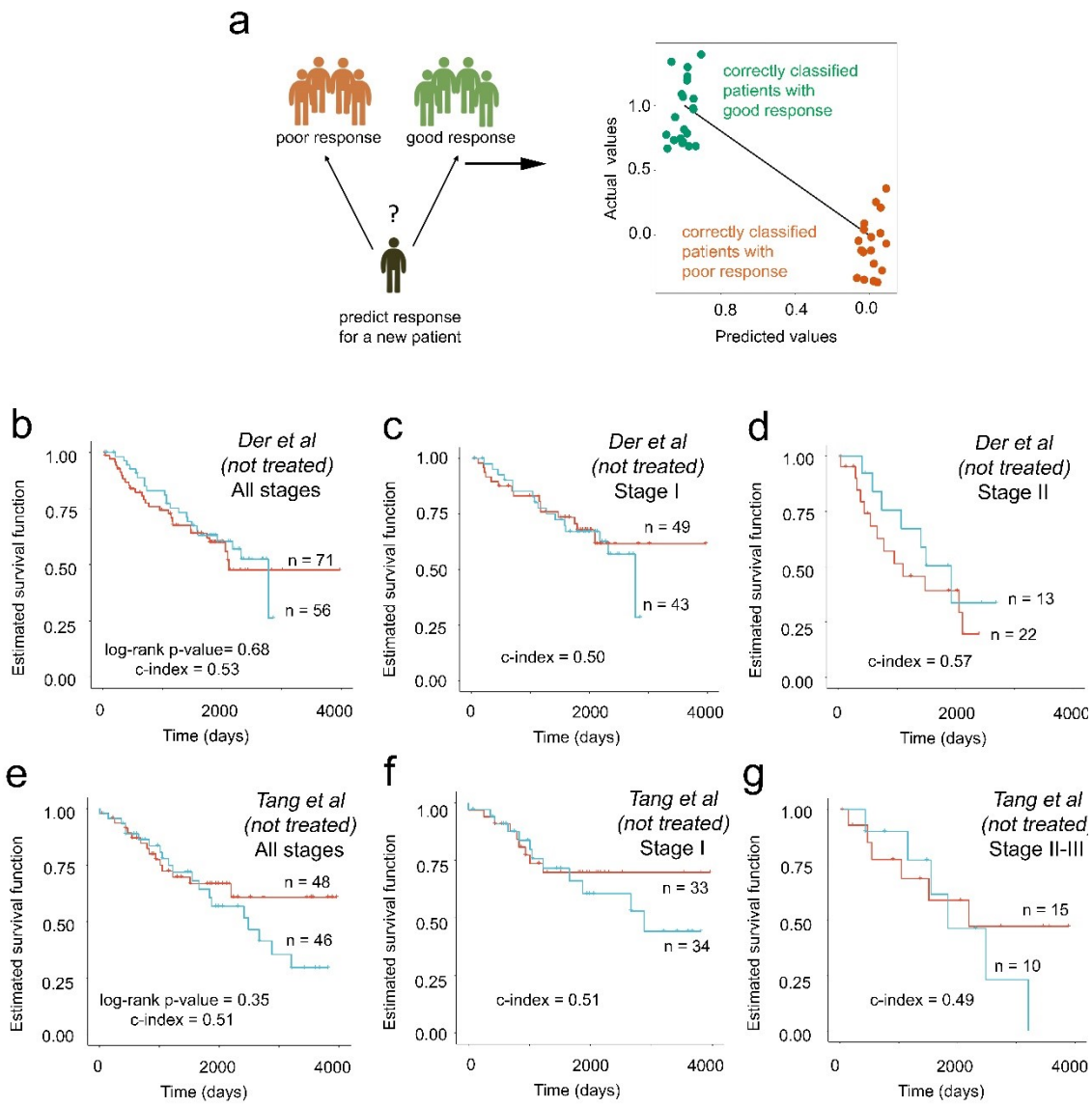


Figure 7. Candidate molecular pathways predict response to carboplatin-taxane and are not predictive of lung cancer aggressiveness.

(a) Leave-one-out cross-validation (LOOCV) in the *Tang et al*. (*n* = 39) validation cohort. Correctly predicted patients with favorable response to carboplatin-taxane (e.g., paclitaxel) (*green*) and patients with poor response to carboplatin-taxane (e.g., paclitaxel) (*orange*) are indicated. **(b-g)** Kaplan-Meier survival analysis shows no significant difference between untreated patients based on the overall lung cancer aggressiveness in **(b-d)** *Der et al*. (*n* = 127) and **(e-g)** *Tang et al*. (*n* = 94) observational (i.e., not treated) patient cohorts. Log-rank p-value and the number of patients in each group are indicated.

pathways can also separate patients based on their lung cancer aggressiveness. For this, we evaluated the predictive ability of our candidate pathways on the LUAD patient cohorts that did not receive any treatment after surgery (we used these cohorts as negative controls). These datasets (Supplementary Table 1) included: *Der et al*.[65] LUAD tumor samples (n = 127) collected through surgery between 1996 and 2005 at Princess Margaret Cancer Centre, and *Tang et al*.[18] provisional cohort, which includes LUAD tumor samples (n = 94) collected through surgery between 1996 and 2007 at The University of Texas MD Anderson Cancer Center. These negative control patient cohorts did not receive any subsequent treatment but were monitored for disease progression (for *Der et al*. lung cancer-related death was used as a clinical endpoint and for *Tang et al*. survival status during the clinical study (1996 to 2007) was used as a clinical endpoint). Kaplan-Meier survival analysis on these datasets demonstrated that our candidate seven pathways did not separate patients based on the disease progression in both unstratified and stratified (i.e., based on tumor stages) analyses *Der et al*. (Figure 7b-d, log-rank p-value = 0.68), and *Tang et al*. (Figure 7e-g, log-rank p-value = 0.35) and are in fact specific for carboplatin-paclitaxel response.

3.5 Comprehensive comparative analysis

To assess advantages of our approach, we have compared its predictive performance to other commonly utilized methods, including methods based on linear regression modeling, support vector machine (SVM), and random forest; and evaluated if our approach can be affected by commonly utilized covariates or known signatures of lung cancer aggressiveness.

First, to measure the advantage of our model over other commonly utilized methods, we have compared predictive performance of our model (as described in the Methods section) to *Panja et al.*[56] method, Epi2GenR, based on linear regression integration between DNA methylation and mRNA expression patient profiles, which

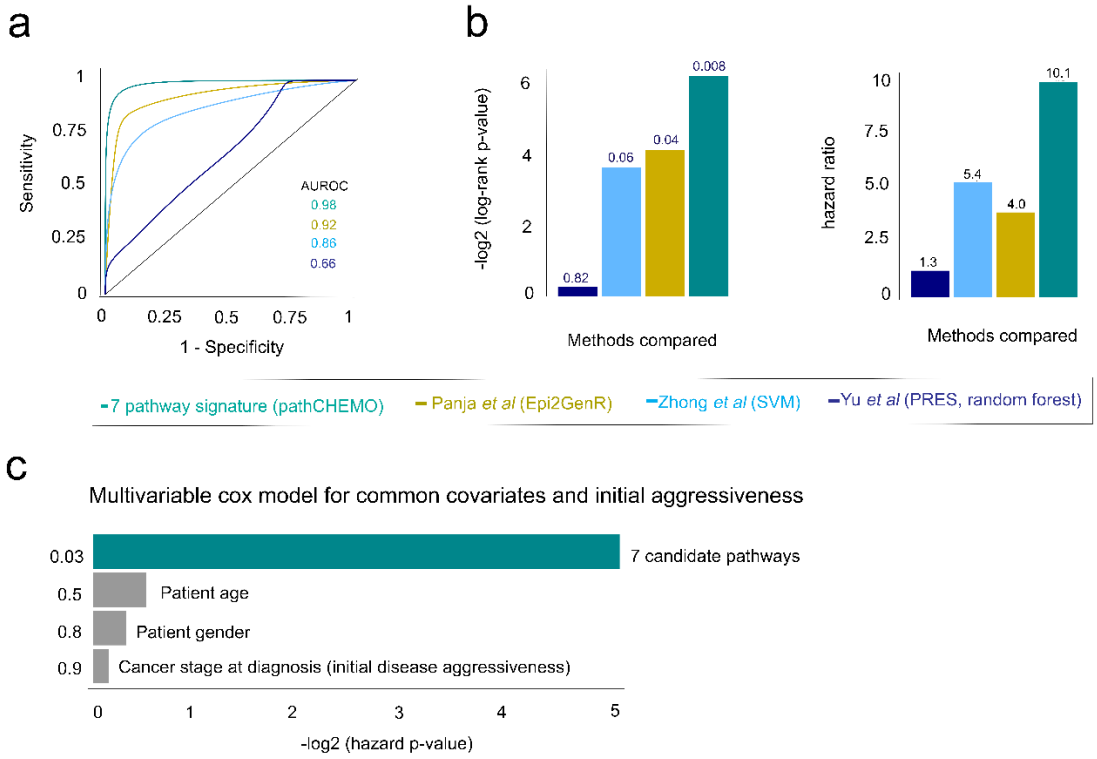


Figure 8. Comparative performance analysis confirms robust predictive ability of pathCHEMO.

(a-b) Comparison of pathCHEMO (*turquoise*) to other commonly utilized methods, including *Panja et al*. Epi2GenR (*yellow*), *Zhong et al*. SVM (*light blue*), *Yu et al*. PRES random forest (*dark blue*) using **(a)** ROC analysis (with AUROC indicated) and **(b)** Kaplan-Meier and Cox proportional hazards model (with log-rank p-value and hazard ratio indicated) in *Tang et al*. validation cohort. **(c)** Multivariable Cox proportional hazards analysis demonstrating adjustment of seven candidate pathways for common covariates (i.e., age, gender, and stage at diagnosis). Hazard p-value is indicated.

identified 35 site-gene pairs as candidate markers of carboplatin-paclitaxel response.

Second, our model was compared to *Zhong et al.*[62] method based support vector machine

(SVM) analysis, which identified 104 candidate genes. Finally, our model was evaluated against *Yu et al*.[61] method PRES, based on random forest algorithm, which identified 3 candidates of carboplatin-paclitaxel response. We first compared ability of the identified candidates from each method to separate patients with poor and favorable carboplatin-paclitaxel response in the *Tang et al*. dataset using ROC analysis, which demonstrated advantage of pathCHEMO over other commonly utilized methods (Figure 8a, $AUROC_{pathCHEMO} = 0.98$, $AUROC_{Epi2GenR} = 0.92$, $AUROC_{SVM} = 0.86$, $AUROC_{PRES} = 0.66$). Furthermore, we compared ability of these methods to predict response to carboplatin-paclitaxel in the *Tang et al*. validation set (as above), through Kaplan-Meier survival analysis (Figure 8b: log-rank $p\text{-value}_{pathCHEMO} = 0.008$, log-rank $p\text{-value}_{Epi2GenR} = 0.04$, log-rank $p\text{-value}_{SVM} = 0.06$, log-rank $p\text{-value}_{PRES} = 0.82$) and Cox proportional hazards model (Figure 8b: $\text{hazard ratio}_{pathCHEMO} = 10.1$, $\text{hazard ratio}_{Epi2GenR} = 4.0$, $\text{hazard ratio}_{SVM} = 5.4$, $\text{hazard ratio}_{PRES} = 1.3$), which confirmed that pathCHEMO outperformed other commonly used methods in its ability to predict therapeutic response.

Second, to assure that our model is not affected by commonly utilized covariates (i.e., age, gender, and disease stage at diagnosis), we have evaluated their effect through multivariable (i.e., adjusted) Cox proportional hazards model[83] on the *Tang et al.* dataset (as described in the Methods section), which demonstrated that these covariates are not predictive of treatment response and do not affect predictive ability of our model (Figure 8c). Furthermore, to re-confirm this result we performed stratified Kaplan-Meier survival analysis; where we stratified the *Tang et al.* validation cohort into patient groups based on: age (< median age and >= median age); gender (i.e., female and male); and disease

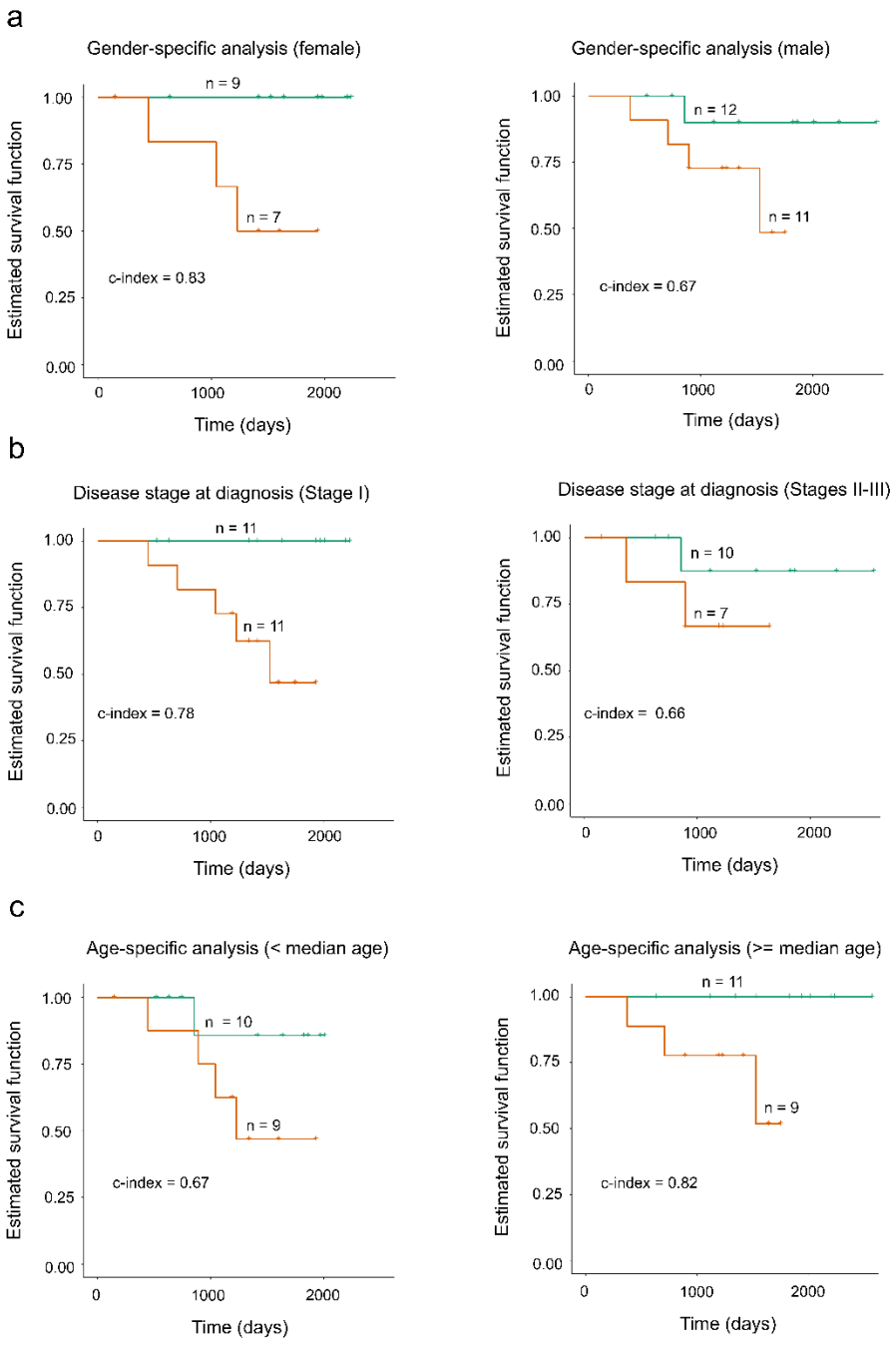


Figure 9. Stratified Kaplan-Meier survival analysis demonstrates independence of the candidate pathways from the common covariates.

Stratified Kaplan-Meier survival analysis in the *Tang et al.* patient cohort (n = 39) based on common prognostic covariates: **(a)** age-specific analysis (greater than and less than median age); **(b)** gender-specific analysis (female and male), and **(c)** diseases stage at diagnosis (I and II-III). C-index and number of patients in each group are indicated.

stage at diagnosis (stage I, and stages II and III), which confirmed that ability of our model to predict chemotherapy response does not depend on commonly utilized covariates and is indeed indicative of a therapeutic response to carboplatin-paclitaxel (Figure 9).

Finally, to assure that our model is not affected by markers of overall tumor aggressiveness, we tested if any known prognostic signatures of lung cancer aggressiveness can predict carboplatin-paclitaxel response or affect predictive ability of our model. For this, we first selected known prognostic signatures of lung cancer aggressiveness including: *Larsen et al*.[89] (54 prognostic markers); *Beer et al*.[90] (50 prognostic markers); and *Tang et al*.[18] (12 prognostic markers) (Figure 10) and utilized them in multivariable Cox proportional hazards model, as above. Our analysis demonstrated that these prognostic signatures were not predictive of carboplatin-paclitaxel response and did not affect the predictive ability of our seven candidate pathways (Figure 10).

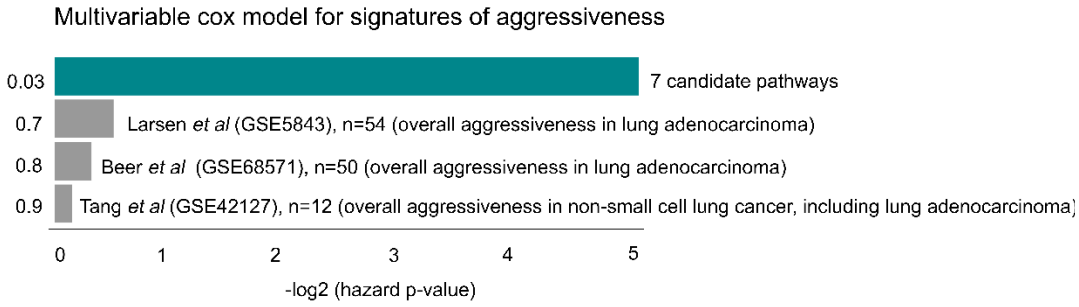


Figure 10. Comparative performance analysis of known markers of lung cancer aggressiveness confirms significant predictive ability of pathCHEMO.

Multivariable Cox proportional hazards analysis demonstrating adjustment of seven candidate pathways for signatures of lung cancer aggressiveness, including *Larsen et al*. (54 lung adenocarcinoma markers), *Beer et al*. (50 lung adenocarcinoma markers), and *Tang et al*. (12 non-small cell lung cancer markers). Hazard p-value is indicated.

3.6 Pathway activity read-outs

Molecular pathways are comprised of multiple genes, which complicate their clinical applicability as markers of treatment response. To tackle this limitation, we

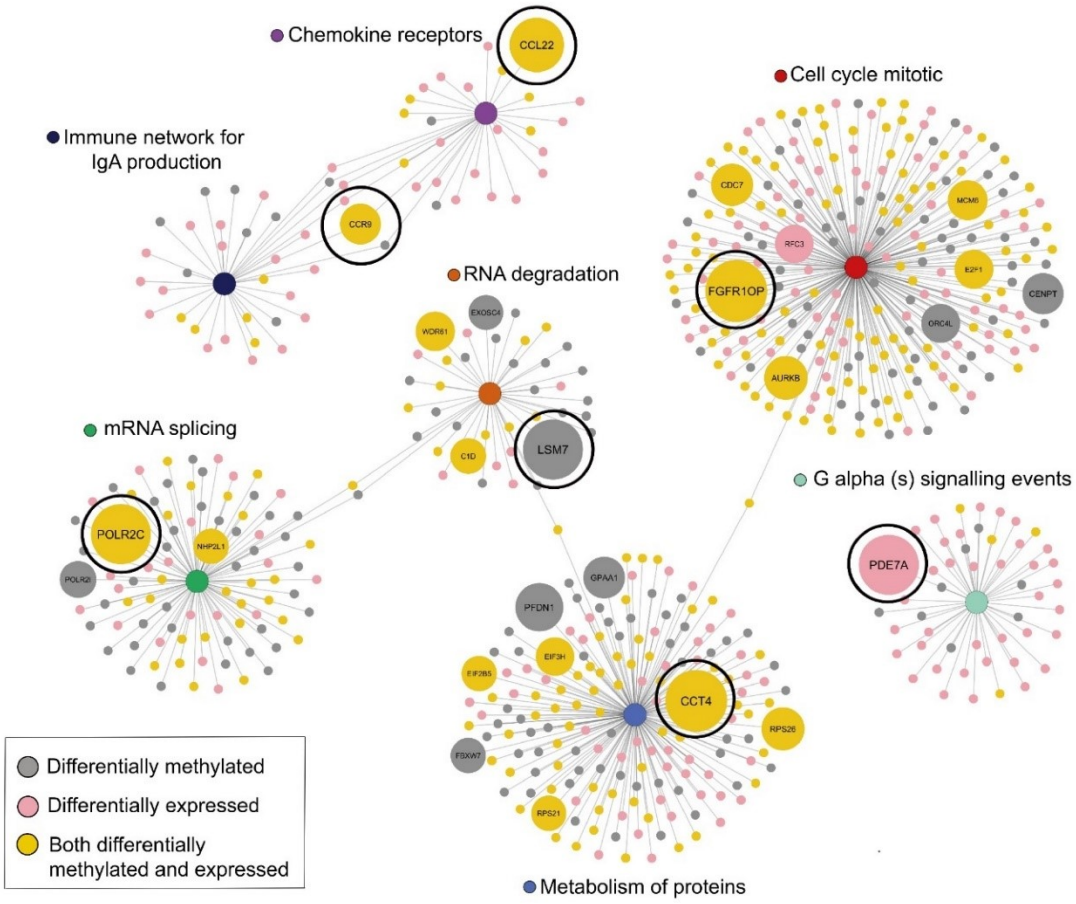


Figure 11. Network representation of candidate molecular pathways with their read-out genes.

Network representation of the candidate pathways, where leading edge genes correspond to nodes and their sizes indicates –log based 2 of Fisher's combined p-values (i.e., combining likelihood-ratio test p-value for association with treatment response and Pearson correlation p-value for correlation with pathway activity). Largest nodes correspond to readout genes, for each pathway. Gene colors depict differential expression (*pink*), differential methylation (*grey*), and both differential expression and methylation (*yellow*).

looked for genes which could serve as read-outs of pathway's activity implicated in therapeutic response. Specifically, we looked for genes inside each pathway, which were: first, altered on transcriptomic and/or epigenomic levels; second, correlated with pathway activity levels (i.e., NESs in each patient); and finally, associated with carboplatin-paclitaxel response (as described in the Methods section). This analysis identified seven read-out genes (i.e., *FGFR1OP, CCL22, CCR9, LSM7*, *PDE7A*, *CCT4*, and *POLR2C*), which: first, accurately reflected activity levels of their corresponding pathways; second, were associated with treatment response; and finally, achieved identical accuracy in predicting patients at risk of carboplatin-paclitaxel resistance (Figure 11, Table 2 , log-rank p-value = 0.0043, hazard ratio = 6.28). We propose that these seven read-out genes can be used as markers of carboplatin-paclitaxel response and can be easily adopted in the clinic.

3.7 Model generalizability

In order to test the general applicability of pathCHEMO, we applied our approach across additional chemotherapy combinations and cancer types. In particular, we extended pathCHEMO to: cisplatin-vinorelbine response in lung adenocarcinoma, cisplatin-vinorelbine response in lung squamous cell carcinoma, and folinic acid, fluorouracil, and oxaliplatin (i.e., FOLFOX) response in colorectal adenocarcinoma (Supplementary Table 2-4). First, we applied our approach to additional chemo-combination (i.e., cisplatin-vinorelbine) administered to lung adenocarcinoma (TCGA-LUAD) patients (Supplementary Table 2), which identified a set of three molecular pathways as markers of cisplatin-vinorelbine resistance (GSEA NES = 2.51, p-value <

0.001) (Figure 12a) and their corresponding read-out genes (Table 2). These pathways included *metabolism of nucleotides, actin Y*, and *ribosome* pathways. We validated these predictions using the *Zhu et al*.[17] cohort from the National Cancer Institute of Canada

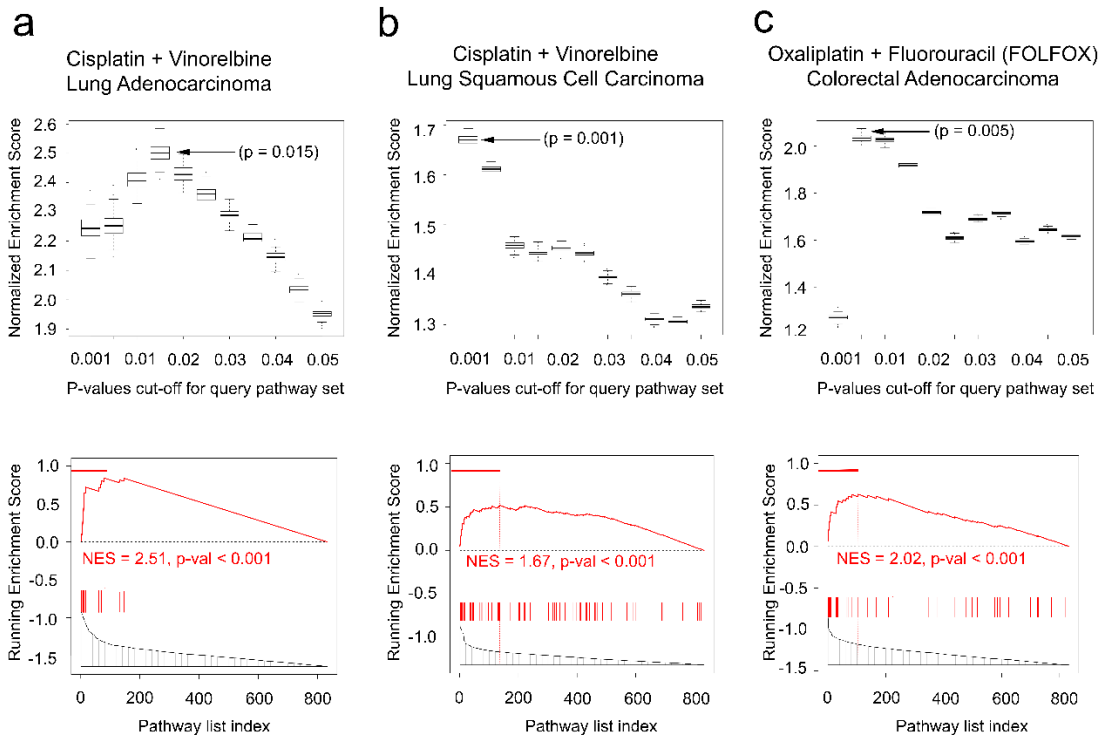


Figure 12. Identification of pathways of treatment resistance across chemo-regimens and cancer types.

pathCHEMO discovery in **(a)** cisplatin-vinorelbine treated lung adenocarcinoma patients (TCGA-LUAD), **(b)** cisplatin-vinorelbine treated lung squamous cell carcinoma patients (TCGA-LUSC), and **(c)** FOLFOX (folinic acid, fluorouracil, and oxaliplatin) treated colorectal adenocarcinoma patients (TCGA-COAD). *(top)* Box and whisker plots depicting p-value cutoff discovery for query treatment response composite methylation pathway signature (x-axis) and NESs from the corresponding GSEA comparison between treatment response composite methylation and expression pathways signatures (y-axis). Arrows indicate optimal p-value thresholds, resulting in the most significant GSEA enrichment. (*bottom*) GSEAs comparing indicated treatment response composite expression pathway signatures (reference) and indicated treatment response composite methylation pathway signatures (query). Horizontal red bars indicate leading edge pathways altered on both transcriptomic and epigenomic levels. NES and p-value were estimated using 1,000 pathway permutations.

Clinical Trials Group (Supplementary Table 2), which contains LUAD tumor samples (n = 39) collected through surgery, for patients that received adjuvant cisplatin-vinorelbine, and demonstrated that three candidate pathways can predict poor and favorable cisplatin-vinorelbine response in patients with LUAD (lung cancer-related death used as a clinical endpoint) using Kaplan-Meier survival analysis and Cox proportional hazards model (Figure 13a, log-rank p-value = 0.0048, hazard ratio = 3.64).

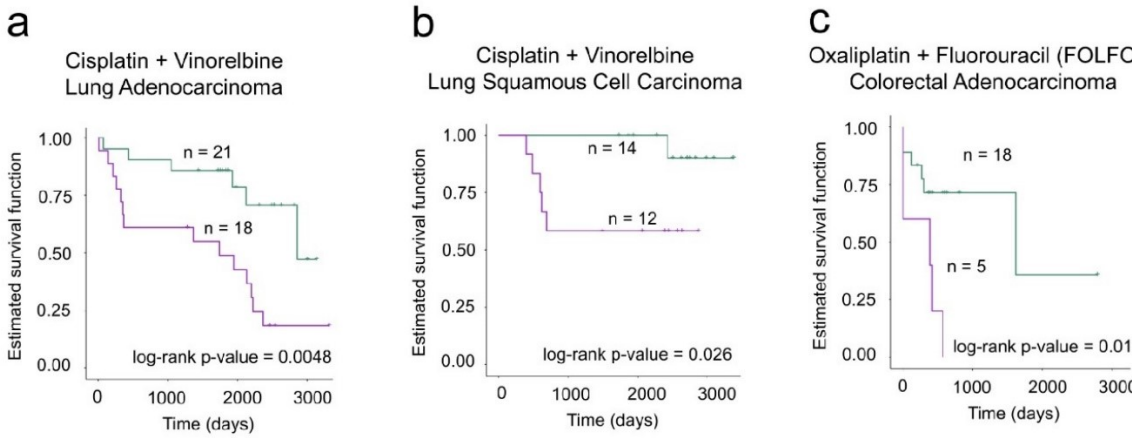


Figure 13. pathCHEMO accurately identifies pathways of treatment resistance across chemo-regimens and cancer types.

Treatment related Kaplan-Meier survival analysis in **(a)** cisplatin-vinorelbine treated lung adenocarcinoma (LUAD) patients in the Zhu et al. patient cohort (n = 39 biologically independent patient samples), **(b)** cisplatin-vinorelbine treated lung squamous cell carcinoma (LUSC) patients in the *Zhu et al*. patient cohort (n = 26 biologically independent patient samples), and **(c)** FOLFOX (folinic acid, fluorouracil, and oxaliplatin) treated colorectal adenocarcinoma (COAD) patients in the *Marisa et al.* patient cohort (n = 23 biologically independent patient samples), demonstrating ability of identified candidate pathways (for each analysis) to predict treatment response. Log rank p-value and number of patients in each group are indicated.

Next, we applied our approach to cisplatin-vinorelbine treated lung squamous cell carcinoma (TCGA-LUSC) patients (Supplementary Table 3) and identified a set of six

molecular pathways (GSEA NES = 1.67, p-value < 0.001) (Figure 12b) including *neuroactive ligand-receptor interaction, SLC-mediated transmembrane transport, transport of mature mRNA derived from an intron-containing transcript, cytokine-cytokine receptor interaction, DNA repair,* and *translation* pathways and their corresponding read-out genes (Table 2). We validated these predictions using the *Zhu et al*. patient cohort[17] (Supplementary Table 3), which contains LUSC primary tumor samples (*n* = 26) collected through surgery, for patients that received adjuvant cisplatin-vinorelbine treatment, and demonstrated that six candidate pathways can accurately predict poor and favorable cisplatin-vinorelbine response in patients with LUSC (lung cancer-related death used as clinical endpoint) (Figure 13b, log-rank p-value = 0.026, hazard ratio = 7.94).

Lastly, we applied our approach to patients with colorectal adenocarcinoma (TCGA-COAD) that received FOLFOX (i.e., folinic acid, fluorouracil, and oxaliplatin) combination (Supplementary Table 4), which identified five molecular pathways as markers of FOLFOX resistance (GSEA NES = 2.02, p-value < 0.001) (Supplementary Figure 8c). These pathways included *processing of capped intron containing pre mRNA, S phase, elongation and processing of capped transcripts, metabolism of proteins*, and *calcium signaling* pathways and their corresponding read-out genes (Table 2). We validated these predictions using an independent patient cohort, *Marisa et al*.[63] (Supplementary Table 4) from the French National Cartes d'Identité des Tumeurs (CIT), which contains COAD tumor samples (*n* = 23) collected through surgery followed by adjuvant treatment with FOLFOX monitored for further disease progression (i.e., defined as locoregional or distant recurrence) and demonstrated that five candidate pathways can predict poor and favorable FOLFOX response in patients with COAD (Figure 13c, log-

Table 2. Identified candidate pathways (carboplatin-paclitaxel treated LUAD, cisplatin-vinorelbine treated LUAD, cisplatin-vinorelbine treated LUSC, and FOLFOX (folinic acid, fluorouracil, oxaliplatin) treated COAD) readout, source, and contribution to cancer

Cancer types & treatments	Candidate pathways	Readout	Source	Contribution to cancer
LUAD_CP	*chemokine receptors bind chemokines*	*CCL22*		promotes bone metastasis in lung cancer[99]
	mRNA splicing	*POLR2C*		therapeutic target in breast cancer[100]
	G alpha (s) signalling events	*PDE7A*		prognostic marker of lung cancer[90]
	intestinal immune network for IgA production	*CCR9*		prognostic marker of non-small cell lung cancer[101], etoposide resistance in prostate cancer[102], cisplatin resistance in breast[103] and ovarian[104] cancers
	metabolism of proteins	*CCT4*		therapeutic target in lung cancer[105]
	RNA degradation	*LSM7*		diagnostic marker of thyroid cancer[106]
	cell cycle mitotic	*FGFR1OP*		prognostic biomarker and therapeutic target in lung cancer[107]
LUAD_CV	*metabolism of nucleotides*	*DTYMK*		therapeutic target for LKB1-deficient lung cancer[108]
	actin Y	*ARPC1A*		novel marker of pancreatic cancer[109]
	ribosome	*RPLP2*		prognostic marker in gynecologic tumor[110] and in gastric cancer[111]
LUSC	*cytokine-cytokine receptor interaction*	*CCL11*		biomarker of ovarian cancer[112]
	neuroactive ligand-receptor interaction	*GABRA1*		DNA methylation markers in colorectal cancer[113]
	DNA repair	*ERCC1*		prognostic marker in prostate[114], and bladder[115] cancer
	SLC-mediated transmembrane transport	*SLC44A4*		novel target for prostate and pancreatic cancer[116]
	translation	*RPL14*		molecular marker for esophageal squamous cell carcinoma[117]

Cancer types & treatments	Candidate pathways	Readout	Source	Contribution to cancer
LUSC	*transport of mature mRNA derived from an intron-containing transcript*	*U2AF1*		contributes to cancer progression[118]
COAD	*elongation and processing of capped transcripts*	*SF3B3*		therapeutic target for ER-positive breast cancer[119]
	processing of capped intron containing pre mRNA	*PRPF6*		tumor marker in colon cancer[120]
	metabolism of protein	*PFDN1*		promotes epithelial-mesenchymal transition (EMT) and lung cancer progression[121]
	S phase	*CDC25B*		prognostic marker in non-small cell lung cancer[122]
	calcium signaling	*MYLK3*		biomarker in ovarian cancer[123]

Notes: LUAD_CP = lung adenocarcinoma treated with carboplatin and paclitaxel; LUAD_CV = lung adenocarcinoma treated with cisplatin and vinorelbine; LUSC = lung squamous cell carcinoma treated with cisplatin and vinorelbine; COAD = colon adenocarcinoma treated with FOLFOX (folinic acid, fluorouracil, oxaliplatin); Source (fourth column): readout in each pathway are represented as differentially expressed (*pink*), methylated (*grey*) and both differentially expressed and methylated *(yellow*).

rank p-value = 0.01, hazard ratio = 6.21).

Interestingly, when evaluating overlaps between pathways across different chemo-treatments and cancers, we have noticed that even though some biological pathways might be overlapping, their overlapping genes exhibit totally different behaviors (e.g., are over-expressed for one chemo-regimen and are under-expressed for another etc.), thus demonstrating drastically different patterns of pathway dysregulations inherent for each specific chemo-regiment and for each cancer type.

Furthermore, we have identified readout genes and clinical utility for these identified molecular pathways, which demonstrated identical accuracy in predicting patients at risk of treatment resistance in lung adenocarcinoma (Table 2, log-rank p-value = 0.0027, hazard ratio = 4.64), lung squamous cell carcinoma (Table 2, log-rank p-value = 0.0004 , hazard ratio = 17.90), and colon adenocarcinoma (Table 2, log-rank p-value = 0.0039, hazard ratio = 6.251).

Taken together, these analyses demonstrate the general applicability of our method across various chemotherapy-regimens and cancer types and builds a foundation for our long-term goal to enhance personalized therapeutic advice and improve patient care and clinical decision support at large.

CHAPTER IV

4 DISCUSSION

We have introduced a systematic generalizable computational approach pathCHEMO to uncover molecular pathways that govern complex transcriptomic and epigenomic mechanisms implicated in chemotherapy response. Firstly, the distinguishing feature of pathCHEMO is in the identification of molecular pathways altered on both transcriptomic and epigenomic levels, which increases the likelihood of elucidating functionally relevant alterations. Secondly, the identified pathways constitute not only molecular markers for predictive analysis but also valuable candidates for therapeutic targeting to preclude or overcome resistance. Thirdly, our approach is generalizable and has been successfully applied to additional chemotherapy-regimens and cancer types, where it demonstrated the high accuracy of its predictions. Fourthly, pathCHEMO predicts patients at risk of developing resistance to specific chemotherapy, even prior to therapy administration, which builds a platform for optimal treatment planning and personalized therapeutic advice. Finally, to the best of our knowledge, pathCHEMO is the first computational predictive effort of its kind in chemotherapy resistance space, with near-term potential to improve informed clinical decision-making and cancer management.

We used pathCHEMO to elucidate mechanisms of resistance to carboplatin–paclitaxel chemotherapy in lung adenocarcinoma and identified seven molecular pathways implicated in resistance, including *chemokine receptors bind chemokines*, *mRNA splicing*, *G alpha (s) signalling events*, *intestinal immune network for IgA production*, *metabolism of proteins*, *RNA degradation*, and *cell cycle mitotic* pathways. Interestingly, paclitaxel resistance has been shown to be modulated by Hippo signaling pathway in breast cancer[124], which is directly activated by our candidate *G alpha (s) signalling events* pathway[125]. Furthermore, *chemokine receptors bind chemokines* pathway is directly associated[126] with cytokine and inflammatory response pathway, which modulates carboplatin resistance in ovarian cancer[127]. Finally, *cell cycle mitotic* pathway has been shown to be directly affected by paclitaxel[128] and carboplatin-paclitaxel[129,130] treatments in ovarian cancer. Thus, primary (i.e., before therapy administration) dysregulation in these pathways might affect drug mechanism of action and can be utilized to identify patients at risk of resistance.

Interestingly, one of the identified pathways, *G alpha (s) signalling events* pathway, is involved in mediation of extracellular signaling and activation of Protein Kinase A (PKA), a known player in cancer cell invasion and metastasis. Recently, PKA has been shown to play a central role in resistance to tamoxifen in breast cancer[131], and disease progression in prostate cancer[132]. PKA has been known to contribute to lung cancer tumorigenesis by interacting with RAS oncogenic pathway and promoting epithelial-mesenchymal transition (EMT) during hypoxia. Several recent studies have confirmed the role of EMT as a key player in acquired (i.e., caused by the treatment) resistance to chemotherapy including acquired resistance to gemcitabine in pancreatic

cancer[133], to paclitaxel in ovarian cancer[134], and to gefitinib in lung cancer[135], emphasizing importance of further investigating EMT as a mechanism of primary resistance to chemotherapy in lung adenocarcinoma.

4.1 Limitations and future directions

In addition to EMT, the development of neuroendocrine phenotype has been shown to be a major emerging player in acquired therapeutic resistance in lung cancer[136,137]. Recent studies have demonstrated that 50% of patients with metastatic lung adenocarcinoma, which were treated with erlotinib and acquired resistance to it, had a histological transformation to large cell neuroendocrine carcinoma (LCNEC), leading to increased metastatic burden and lethality[138,139]. Therefore, further investigation of the role of EMT and neuroendocrine markers and their interplay with transcriptomic and epigenomic molecular alterations are necessary for comprehensive understanding of complex mechanisms involved in resistance to chemotherapy and will contribute a central focus of our subsequent studies.

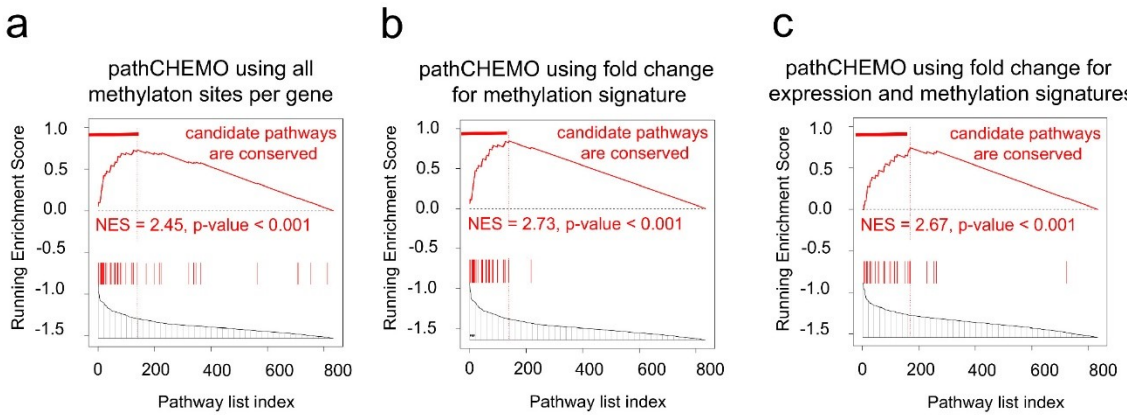
CHAPTER V

5 CONCLUSIONS

In summary, we have introduced a generalized applicable integrative transcriptomic and epigenomic approach that discovered molecular pathways implicated in chemotherapy resistance in lung adenocarcinoma. We recommend that these molecular pathways can *(i)* used as a predictive marker to accurately predict response to chemotherapy regimens across independent patient cohorts; *(ii)* prioritize patients who would benefit from chemotherapy regimens and patients at risk of resistance that should be offered personalized therapeutic advice for alternative regimen. Furthermore, we propose that such chemo-resistance biological pathways may serve as predictive markers to infer the potential efficacy of drug treatments in patients. Additionally, pathCHEMO should be applicable for identifying molecular pathway that point toward treatment response and improve personalized therapeutic advice in cancer and other diseases.

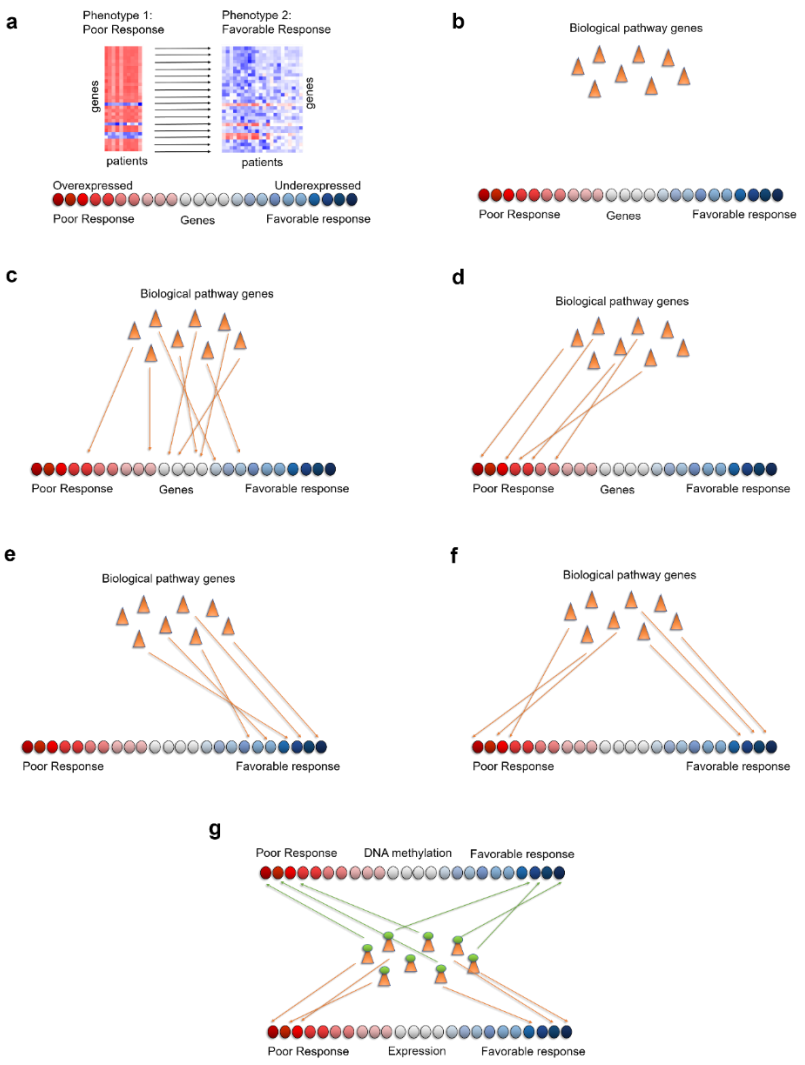
Appendix A

Supplementary Materials



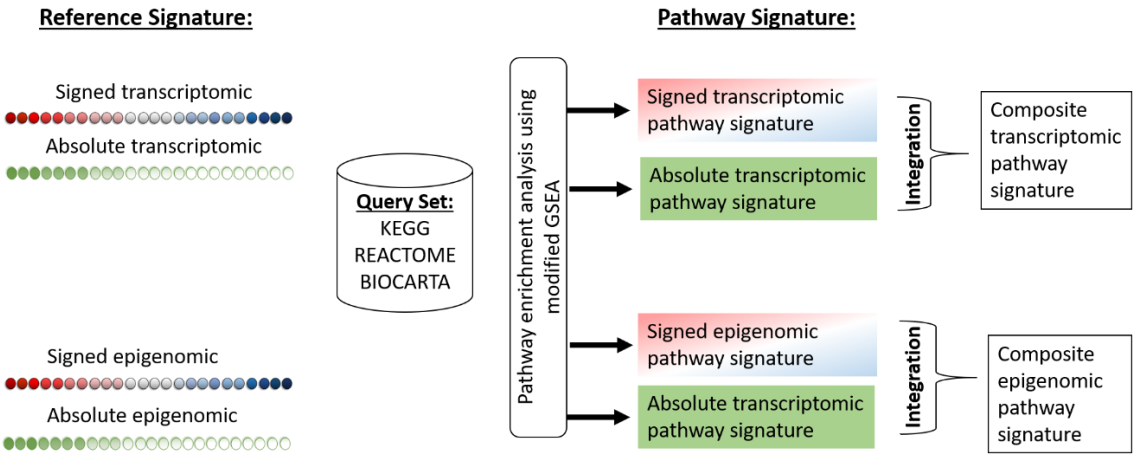
Supplementary Figure 1. Comparative testing of treatment response signatures demonstrates their robustness.

GSEAs comparing **(a)** treatment response composite expression pathway signature (reference) and treatment response composite methylation pathway signature constructed considering all CpG DNA methylation sites (query), **(b)** treatment response composite expression pathway signature (reference) and treatment response composite methylation pathway signature(query), where methylation signature was defined using fold change, and **(c)** treatment response composite expression pathway signature (reference) and treatment response composite methylation pathway signature (query), where both signatures were defined using fold change. Horizontal red bars indicate leading edge pathways altered on both transcriptomic and epigenomic levels. NES and p-value were estimated using 1,000 pathway permutations.

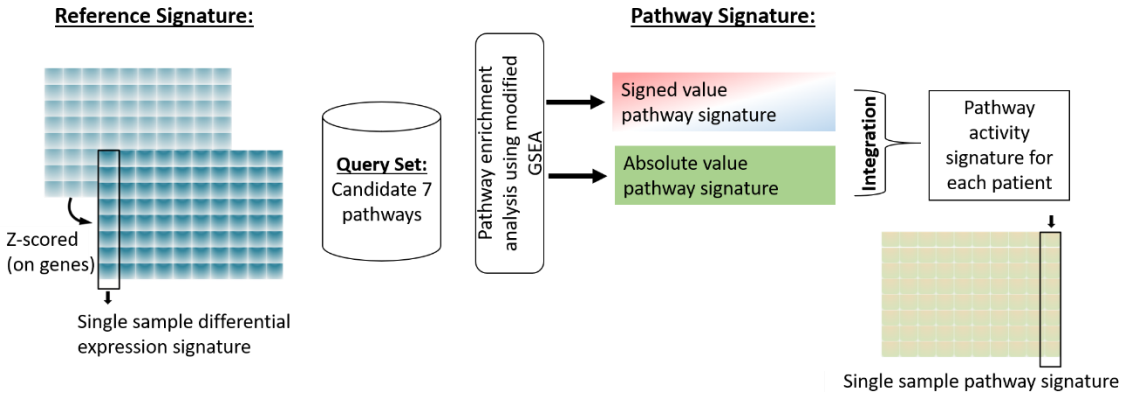


Supplementary Figure 2. Schematic representation of pathCHEMO

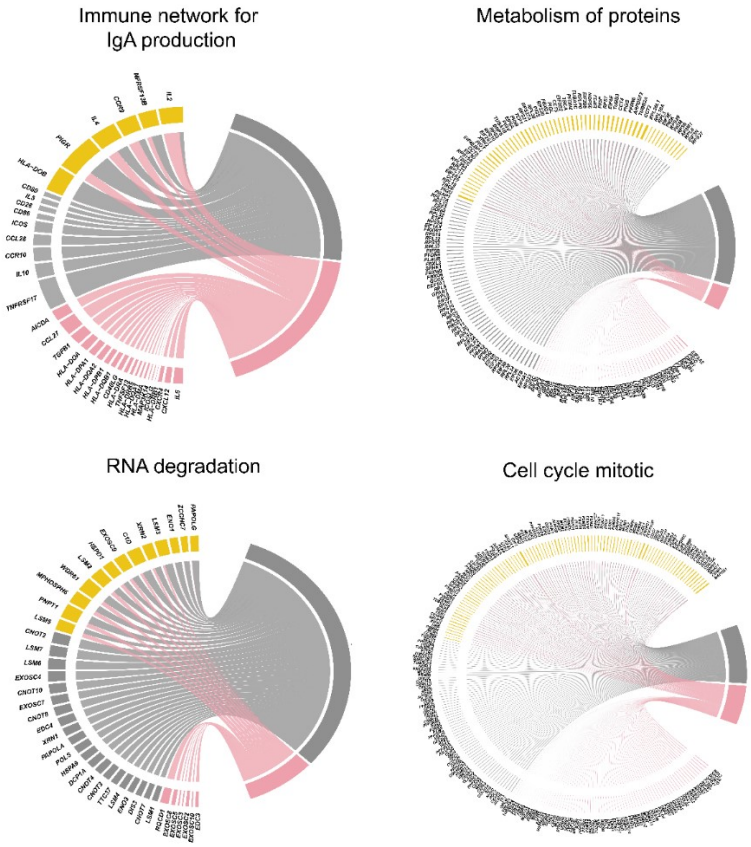
(a) schematic representation of molecular profiles (i.e., gene expression) for poor response and favorable response patients group (top) and by comparing this molecular profiles we define reference signature of treatment response which is a list of genes ranked by their differential expression (bottom) from most overexpressed (bottom left, red) to most under-expressed (bottom right, blue). **(b)** Schematic representation of pathway genes; **(c)** and how pathway genes map on the reference signature. In the majority of cases, we see those pathway genes fall all over the signature, but our main goal is to look for pathway that will map **(d)** over-expressed part of the signature; **(e)** under-expressed part of the signature and **(f)** both over and under expressed part of the signature; **(g)** Same logic applied to methylation, over-methylation, under-methylation and both over and under methylation.



Supplementary Figure 3. Schematic representation of integrative multi-omic pathway enrichment analysis.



Supplementary Figure 4. Schematic representation of single sample pathway enrichment analysis.



Supplementary Figure 5. Transcriptomic and epigenomic alterations in selected candidate molecular pathways of carboplatin-paclitaxel.

Representative molecular pathways altered on both transcriptomic and epigenomic levels. Genes from the leading edge in each pathway are represented as differentially expressed (*pink*), methylated (*grey*) and both differentially expressed and methylated (*yellow*). Width of each connecting line is proportional to the extent of differential expression and differential methylation. Pathways are depicting as follows: *(i) intestinal immune network for IgA production* pathway (20 differentially expressed genes, 9 differentially methylated genes, and 6 differentially expressed and methylated genes), *(ii) metabolism of proteins* pathway (47 differentially expressed genes, 53 differentially methylated genes, and 62 differentially expressed and methylated genes), *(iii) RNA degradation* pathway (7 differentially expressed genes, 21 differentially methylated genes, and 13 differentially expressed and methylated genes), and *(iv) cell cycle mitotic* pathway (75 differentially expressed genes, 64 differentially methylated genes, and 100 differentially expressed and methylated genes). Pathways were visualized using circlize[140] package in R.

Supplementary Table 1. Clinical and pathological features of lung adenocarcinoma patient cohorts treated with carboplatin-paclitaxel, used for discovery, validation, and negative controls.

	Signature discovery	Validation	Negative controls	
Description	TCGA	*Tang et al.* (treated)	*Tang et al.* (not treated)	*Der et al.* (not treated)
Accession #	TCGA-LUAD[64]	GSE42127[18]	GSE42127[18]	GSE50081[66]
Platform	Illumina HiSeq 2000 (mRNA expression) Illumina Infinium Human Methylation (HM450) array (DNA methylation)	Illumina HumanWG-6 v3.0 expression beadchip	Illumina HumanWG-6 v3.0 expression beadchip	Affymetrix Human Genome U133 Plus 2.0 Array
Patients	14	39	94	127
Sample collection	surgery	surgery	surgery	surgery
Histological subtype				
mixed	1	NA	NA	NA
acinar	1	NA	NA	NA
NOS	12	NA	NA	NA
Anatomic Site				
Left-Upper	5	NA	NA	NA
Left-Lower	2	NA	NA	NA
Right-Lower	1	NA	NA	NA
Right-Middle	2	NA	NA	NA
Right-Upper	4	NA	NA	NA
Gender				
Female	9	16	49	62
Male	5	23	45	65
Tumor Stage				
IA	NA	1	31	36
IB	1	21	36	56
IIA	1	1	5	7
IIB	4	5	11	28
IIIA	4	3	4	NA
IIIB	1	8	5	NA
IV	1	NA	1	NA
NA	2	NA	1	NA
Smoking Status				
1	2	NA	NA	NA
2	4	NA	NA	NA
3	3	NA	NA	NA
4	5	NA	NA	NA

Notes: NA = Not available, NOS = Not otherwise specified. Smoking status: 1 = lifelong non-smoker (<100 cigarettes smoked in Lifetime), 2 = current smoker (includes daily smokers and non-daily smokers (or occasional smokers), 3 = current reformed smoker for > 15 years, 4 = current reformed smoker for ≤ 15 years.

Supplementary Table 2. Clinical and pathological features of lung adenocarcinoma patient cohorts treated with cisplatin-vinorelbine, used for discovery and validation

	Signature discovery	Validation
Description	TCGA	*Zhu et al*.
Accession #	TCGA-LUAD[64]	GSE14814[17]
Platform	Illumina HiSeq 2000 (mRNA expression) Illumina Infinium Human Methylation (HM450) array (DNA methylation)	Affymetrix Human Genome U133A
Patients	8	39
Sample collection	surgery	surgery
Histological subtype		
mixed	6	NA
acinar	1	9
papillary	NA	5
mucinous	NA	1
lepidic	NA	1
solid	NA	9
NOS	1	14
Anatomic Site		
Left-Upper	2	NA
Left-Lower	NA	NA
Right-Lower	2	NA
Right-Middle	1	NA
Right-Upper	3	NA
Gender		
Female	5	20
Male	3	19
Tumor Stage (Pathological)		
IA	NA	8
IB	1	14
II	NA	NA
IIA	3	11
IIB	1	6
IIIA	2	NA
IIIB	NA	NA
IV	1	NA
Smoking Status		
1	1	NA
2	NA	NA
3	4	NA
4	3	NA

Notes: NA = Not available, NOS = Not otherwise specified.
Smoking status: 1 = lifelong non-smoker (<100 cigarettes smoked in Lifetime), 2 = current smoker (includes daily smokers and non-daily smokers (or occasional smokers), 3 = current reformed smoker for > 15 years, 4 = current reformed smoker for ≤ 15 years.

Supplementary Table 3. Clinical and pathological features of lung squamous cell carcinoma patient cohorts treated with cisplatin-vinorelbine, used for discovery and validation.

	Signature discovery	Validation
Description	TCGA	*Zhu et al*.
Accession #	TCGA-LUSC[141]	GSE14814[17]
Platform	Illumina HiSeq 2000 (mRNA expression) Illumina Infinium Human Methylation (HM450) array (DNA methylation)	Affymetrix Human Genome U133A
Patients	8	26
Sample collection	surgery	surgery
Histological subtype		
NOS	8	26
Anatomic Site		
Left-Upper	2	NA
Left-Lower	NA	NA
Right-Lower	4	NA
Right-Middle	1	NA
Right-Upper	1	NA
Gender		
Female	1	3
Male	7	23
Tumor Stage (Pathological)		
I	NA	13
IA	NA	NA
IB	2	NA
II	NA	13
IIA	1	NA
IIB	4	NA
IIIA	1	NA
IIIB	NA	NA
IV	NA	NA
Smoking Status		
1	NA	NA
2	NA	NA
3	2	NA
4	6	NA

Notes: NA = Not available, NOS = Not otherwise specified.
Smoking status: 1 = lifelong non-smoker (<100 cigarettes smoked in Lifetime), 2 = current smoker (includes daily smokers and non-daily smokers (or occasional smokers), 3 = current reformed smoker for > 15 years, 4 = current reformed smoker for ≤ 15 years.

Supplementary Table 4. Clinical and pathological features of colorectal adenocarcinoma patient cohorts treated with FOLFOX (folinic acid, fluorouracil, oxaliplatin), used for discovery and validation.

	Signature discovery	**Validation**
Description	TCGA	*Marisa et al*.
Accession #	TCGA-COAD[142]	GSE39582[63]
Platform	Illumina HiSeq 2000 (mRNA expression) Illumina Infinium Human Methylation (HM450) array (DNA methylation)	Affymetrix Human Genome U133 Plus 2.0 Array
Patients	8	23
Sample collection	surgery	surgery
Histological subtype		
Ascending Colon	1	NA
Cecum	2	NA
Descending Colon	1	NA
Sigmoid Colon	3	NA
NA	1	NA
Gender		
Female	4	8
Male	4	15
Tumor Stage (Pathological)		
I	NA	NA
IA	NA	NA
IB	NA	NA
II	NA	NA
IIA	1	2
IIB	NA	1
III	1	NA
IIIA	1	3
IIIB	4	3
IIIC	1	3
IV	NA	11

Notes: NA = Not available.

References

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA: a cancer journal for clinicians.* Jan 2017;67(1):7-30.

2. Hanna JM, Onaitis MW. Cell of origin of lung cancer. *Journal of Carcinogenesis.* 2013;12.

3. Byers LA, Rudin CM. Small cell lung cancer: where do we go from here? *Cancer.* Mar 1 2015;121(5):664-672.

4. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic proceedings.* May 2008;83(5):584-594.

5. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer.* Feb 2011;6(2):244-285.

6. Blum RH. Adjuvant Chemotherapy for Lung Cancer — A New Standard of Care. *New England Journal of Medicine.* 2004;350(4):404-405.

7. Arriagada R, Bergman B, Dunant A, Le Chevalier T, Pignon JP, Vansteenkiste J. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. *The New England journal of medicine.* Jan 22 2004;350(4):351-360.

8. Douillard JY, Rosell R, De Lena M, et al. Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. *The Lancet. Oncology.* Sep 2006;7(9):719-727.

9. Rapp E, Pater JL, Willan A, et al. Chemotherapy can prolong survival in patients with advanced non-small-cell lung cancer--report of a Canadian multicenter randomized trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Apr 1988;6(4):633-641.

10. Anderson H, Hopwood P, Stephens RJ, et al. Gemcitabine plus best supportive care (BSC) vs BSC in inoperable non-small cell lung cancer--a randomized trial with quality of life as the primary outcome. UK NSCLC Gemcitabine Group. Non-Small Cell Lung Cancer. *British journal of cancer.* Aug 2000;83(4):447-453.

11. Lilenbaum R, Villaflor VM, Langer C, et al. Single-agent versus combination chemotherapy in patients with advanced non-small cell lung cancer and a performance status of 2: prognostic factors and treatment selection based on two large randomized clinical trials. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer.* Jul 2009;4(7):869-874.

12. Depierre A, Chastang C, Quoix E, et al. Vinorelbine versus vinorelbine plus cisplatin in advanced non-small cell lung cancer: a randomized trial. *Annals of oncology : official journal of the European Society for Medical Oncology.* Jan 1994;5(1):37-42.

13. Merk J, Rolff J, Dorn C, Leschber G, Fichtner I. Chemoresistance in non-small-cell lung cancer: can multidrug resistance markers predict the response of xenograft lung cancer models to chemotherapy? *European Journal of Cardio-Thoracic Surgery.* 2011;40(1):e29-e33.

14. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Jul 20 2008;26(21):3543-3551.

15. Pfister DG, Johnson DH, Azzoli CG, et al. American Society of Clinical Oncology treatment of unresectable non-small-cell lung cancer guideline: update 2003. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Jan 15 2004;22(2):330-353.

16. Lilenbaum RC, Herndon JE, 2nd, List MA, et al. Single-agent versus combination chemotherapy in advanced non-small-cell lung cancer: the cancer and leukemia group B (study 9730). *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Jan 1 2005;23(1):190-196.

17. Zhu CQ, Ding K, Strumpf D, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Oct 10 2010;28(29):4417-4424.

18. Tang H, Xiao G, Behrens C, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Mar 15 2013;19(6):1577-1586.

19. Reck M, Rodriguez-Abreu D, Robinson AG, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *The New England journal of medicine.* Nov 10 2016;375(19):1823-1833.

20. Galluzzi L, Senovilla L, Vitale I, et al. Molecular mechanisms of cisplatin resistance. *Oncogene.* Apr 12 2012;31(15):1869-1883.

21. Olaussen KA, Dunant A, Fouret P, et al. DNA repair by ERCC1 in non-small-cell lung cancer and cisplatin-based adjuvant chemotherapy. *The New England journal of medicine.* Sep 7 2006;355(10):983-991.

22. Ikuta K, Takemura K, Kihara M, et al. Defects in apoptotic signal transduction in cisplatin-resistant non-small cell lung cancer cells. *Oncology reports.* Jun 2005;13(6):1229-1234.

23. Johnstone RW, Ruefli AA, Lowe SW. Apoptosis: a link between cancer genetics and chemotherapy. *Cell.* Jan 25 2002;108(2):153-164.

24. Dive C, Hickman JA. Drug-target interactions: only the first step in the commitment to a programmed cell death? *British journal of cancer.* Jul 1991;64(1):192-196.

25. Rusch V, Klimstra D, Venkatraman E, et al. Aberrant p53 expression predicts clinical resistance to cisplatin-based chemotherapy in locally advanced non-small cell lung cancer. *Cancer research.* Nov 1 1995;55(21):5038-5042.

26. Lin X, Howell SB. DNA mismatch repair and p53 function are major determinants of the rate of development of cisplatin resistance. *Molecular Cancer Therapeutics.* 2006;5(5):1239-1247.

27. Berger W, Setinek U, Hollaus P, et al. Multidrug resistance markers P-glycoprotein, multidrug resistance protein 1, and lung resistance protein in non-small cell lung cancer: prognostic implications. *Journal of cancer research and clinical oncology.* Jun 2005;131(6):355-363.

28. Young LC, Campling BG, Cole SP, Deeley RG, Gerlach JH. Multidrug resistance proteins MRP3, MRP1, and MRP2 in lung cancer: correlation of protein levels with drug response and messenger RNA levels. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Jun 2001;7(6):1798-1804.

29. Miyara H, Hida T, Nishida K, et al. Modification of Chemo-radiosensitivity of a Human Lung Cancer Cell Line by Introduction of the Glutathione S-transferase φ Gene. *Japanese journal of clinical oncology.* 1996;26(1):1-5.

30. Kasahara K, Fujiwara Y, Nishio K, et al. Metallothionein content correlates with the sensitivity of human small cell lung cancer cell lines to cisplatin. *Cancer research.* Jun 15 1991;51(12):3237-3242.

31. Lewis AD, Hayes JD, Wolf CR. Glutathione and glutathione-dependent enzymes in ovarian adenocarcinoma cell lines derived from a patient before and after the onset of drug resistance: intrinsic differences and cell cycle effects. *Carcinogenesis.* 1988;9(7):1283-1287.

32. Godwin AK, Meister A, O'Dwyer PJ, Huang CS, Hamilton TC, Anderson ME. High resistance to cisplatin in human ovarian cancer cell lines is associated with marked increase of glutathione synthesis. *Proceedings of the National Academy of Sciences.* 1992;89(7):3070-3074.

33. Kobayashi S, Boggon TJ, Dayaram T, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *The New England journal of medicine.* Feb 24 2005;352(8):786-792.

34. Kuroda H, Takeno M, Murakami S, Miyazawa N, Kaneko T, Ishigatsubo Y. Inhibition of heme oxygenase-1 with an epidermal growth factor receptor inhibitor and cisplatin decreases proliferation of lung cancer A549 cells. *Lung Cancer.* 2010;67(1):31-36.

35. Langevin SM, Kratzke RA, Kelsey KT. Epigenetics of Lung Cancer. *Translational research : the journal of laboratory and clinical medicine.* Jan 2015;165(1):74-90.

36. Rodenhuis S, Slebos RJ, Boot AJ, et al. Incidence and possible clinical significance of K-ras oncogene activation in adenocarcinoma of the human lung. *Cancer research.* Oct 15 1988;48(20):5738-5741.

37. Stephens P, Hunter C, Bignell G, et al. Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature.* Sep 30 2004;431(7008):525-526.

38. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. *Nature.* Jun 27 2002;417(6892):949-954.

39. Takahashi T, Nau MM, Chiba I, et al. p53: a frequent target for genetic abnormalities in lung cancer. *Science (New York, N.Y.).* Oct 27 1989;246(4929):491-494.

40. Forgacs E, Biesterveld EJ, Sekido Y, et al. Mutation analysis of the PTEN/MMAC1 gene in lung cancer. *Oncogene.* Sep 24 1998;17(12):1557-1565.

41. Kong-Beltran M, Seshagiri S, Zha J, et al. Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer research.* Jan 01 2006;66(1):283-289.

42. Carpten JD, Faber AL, Horn C, et al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature.* Jul 26 2007;448(7152):439-444.

43. Packenham JP, Taylor JA, White CM, Anna CH, Barrett JC, Devereux TR. Homozygous deletions at chromosome 9p21 and mutation analysis of p16 and p15 in microdissected primary non-small cell lung cancers. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Jul 1995;1(7):687-690.

44. Sanchez-Cespedes M, Parrella P, Esteller M, et al. Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer research.* Jul 01 2002;62(13):3659-3662.

45. Pao W, Wang TY, Riely GJ, et al. KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. *PLoS medicine.* Jan 2005;2(1):e17.

46. Khoo C, Rogers TM, Fellowes A, Bell A, Fox S. Molecular methods for somatic mutation testing in lung adenocarcinoma: EGFR and beyond. *Translational Lung Cancer Research.* Apr 2015;4(2):126-141.

47. Yang X, Uziely B, Groshen S, et al. MDR1 gene expression in primary and advanced breast cancer. *Laboratory investigation; a journal of technical methods and pathology.* Mar 1999;79(3):271-280.

48. Deng HB, Parekh HK, Chow K-C, Simpkins H. Increased expression of dihydrodiol dehydrogenase induces resistance to cisplatin in human ovarian carcinoma cells. *Journal of Biological Chemistry.* 2002;277(17):15035-15043.

49. Petty R, Evans A, Duncan I, Kurbacher C, Cree I. Drug resistance in ovarian cancer - the role of p53. *Pathology oncology research : POR.* 1998;4(2):97-102.

50. Christmann M, Pick M, Lage H, Schadendorf D, Kaina B. Acquired resistance of melanoma cells to the antineoplastic agent fotemustine is caused by reactivation of the DNA repair gene MGMT. *International journal of cancer.* Apr 01 2001;92(1):123-129.

51. Tao L, Huang G, Chen Y, Chen L. DNA methylation of DKK3 modulates docetaxel chemoresistance in human nonsmall cell lung cancer cell. *Cancer biotherapy & radiopharmaceuticals.* Mar 2015;30(2):100-106.

52. Hu H, Li S, Cui X, et al. The overexpression of hypomethylated miR-663 induces chemotherapy resistance in human breast cancer cells by targeting heparin sulfate proteoglycan 2 (HSPG2). *The Journal of biological chemistry.* Apr 19 2013;288(16):10973-10985.

53. Lokk K, Vooder T, Kolde R, et al. Methylation Markers of Early-Stage Non-Small Cell Lung Cancer. *PLOS ONE.* 2012;7(6):e39813.

54. Lockwood WW, Thu KL, Lin L, et al. Integrative Genomics Identified RFC3 as an Amplified Candidate Oncogene in Esophageal Adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Apr 1 2012;18(7):1936-1946.

55. Rhee J-K, Kim K, Chae H, et al. Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Research.* 2013;41(18):8464-8474.

56. Panja S, Hayati S, Epsi NJ, Parrott JS, Mitrofanova A. Integrative (epi) Genomic Analysis to Predict Response to Androgen-Deprivation Therapy in Prostate Cancer. *EBioMedicine.* 2018/04/12/ 2018.

57. Figueroa ME, Reimers M, Thompson RF, et al. An Integrative Genomic and Epigenomic Approach for the Study of Transcriptional Regulation. *PLOS ONE.* 2008;3(3):e1882.

58. Selamat SA, Chung BS, Girard L, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome research.* Jul 2012;22(7):1197-1211.

59. Ma X, Liu Z, Zhang Z, Huang X, Tang W. Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics.* 2017/01/31 2017;18(1):72.

60. TCGA. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* Jul 31 2014;511(7511):543-550.

61. Yu K, Sang QA, Lung PY, et al. Personalized chemotherapy selection for breast cancer using gene expression profiles. *Scientific reports.* Mar 3 2017;7:43294.

62. Zhong Q, Fang J, Huang Z, et al. A response prediction model for taxane, cisplatin, and 5-fluorouracil chemotherapy in hypopharyngeal carcinoma. *Scientific reports.* Aug 23 2018;8(1):12675.

63. Marisa L, de Reyniès A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine.* May 2013;10(5):e1001453.

64. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* Jul 31 2014;511(7511):543-550.

65. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including

stage IA patients. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer.* Jan 2014;9(1):59-64.

66. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology.* 2014;15(12):550.

67. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics.* 2010// 2010;11(1):587.

68. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics (Oxford, England).* Jul 1 2008;24(13):1547-1548.

69. Welch BL. The generalisation of student's problems when several different population variances are involved. *Biometrika.* 1947;34(1-2):28-35.

70. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. 2017.

71. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England).* Jun 15 2011;27(12):1739-1740.

72. Fabregat A, Sidiropoulos K, Garapati P, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* Jan 4 2016;44(D1):D481-487.

73. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999;27(1):29-34.

74. Nishimura D. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient.* 2001;2(3):117-120.

75. Subramanian A, Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P.

Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 2005(102):15545-15550.

76. Dutta A, Panja S, Virk RK, et al. Co-clinical Analysis of a Genetically Engineered Mouse Model and Human Prostate Cancer Reveals Significance of NKX3.1 Expression for Response to 5alpha-reductase Inhibition. *European urology.* Oct 2017;72(4):499-506.

77. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics.* 2011;12(1):77.

78. Wickham H. ggplot2: elegant graphics for data analysis. *J Stat Softw.* 2010;35(1):65-88.

79. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research.* 2008;9(Nov):2579-2605.

80. Maaten LVD. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* 2014;15(1):3221-3245.

81. Kullback S, Leibler RA. On Information and Sufficiency. *Ann. Math. Statist.* 1951/03 1951;22(1):79-86.

82. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association.* 1958/06/01 1958;53(282):457-481.

83. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological).* 1972;34(2):187-220.

84. Therneau T. A package for survival analysis in S. R package version 2.38. *Retrived from http://CRAN. R-project. org/package= survival.* 2015.

85. Kassambara A. Kosinski M. survminer: drawing survival curves using 'ggplot2'. R package version 0.2. 4. 2016.

86. Stone M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological).* 1974:111-147.

87. Chambers J, Hastie T, Pregibon D. Statistical Models in S. 1990; Heidelberg.

88. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics (Oxford, England).* Nov 15 2011;27(22):3206-3208.

89. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research.* May 15 2007;13(10):2946-2954.

90. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine.* Aug 2002;8(8):816-824.

91. Oliveira NL, Pereira CAdB, Diniz MA, Polpo A. A discussion on significance indices for contingency tables under small sample sizes. *PLOS ONE.* 2018;13(8):e0199102.

92. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research.* Nov 2003;13(11):2498-2504.

93. Zou M, Toivanen R, Mitrofanova A, et al. Transdifferentiation as a Mechanism of Treatment Resistance in a Mouse Model of Castration-Resistant Prostate Cancer. *Cancer discovery.* Jul 2017;7(7):736-749.

94. Unnikrishnan A, Papaemmanuil E, Beck D, et al. Integrative Genomics Identifies the Molecular Basis of Resistance to Azacitidine Therapy in Myelodysplastic Syndromes. *Cell reports.* Jul 18 2017;20(3):572-585.

95. Jiang P, Sellers WR, Liu XS. Big Data Approaches for Modeling Response and Resistance to Cancer Drugs. *Annual Review of Biomedical Data Science.* 2018;1(1):1-27.

96. Metz CE. Basic principles of ROC analysis. *Seminars in nuclear medicine.* Oct 1978;8(4):283-298.

97. Zhang XY, Ehrlich KC, Wang RY, Ehrlich M. Effect of site-specific DNA methylation and mutagenesis on recognition by methylated DNA-binding protein from human placenta. *Nucleic Acids Res.* Nov 11 1986;14(21):8387-8397.

98. Laurent L, Wong E, Li G, et al. Dynamic changes in the human methylome during differentiation. *Genome research.* Mar 2010;20(3):320-331.

99. Nakamura ES, Koizumi K, Kobayashi M, et al. RANKL-induced CCL22/macrophage-derived chemokine produced from osteoclasts potentially promotes the bone metastasis of lung cancer expressing its receptor CCR4. *Clinical & experimental metastasis.* 2006;23(1):9-18.

100. Grinchuk OV, Motakis E, Yenamandra SP, et al. Sense-antisense gene-pairs in breast cancer and associated pathological pathways. *Oncotarget.* Dec 8 2015;6(39):42197-42221.

101. Gupta P, Sharma PK, Mir H, et al. CCR9/CCL25 expression in non-small cell lung cancer correlates with aggressive disease and mediates key steps of metastasis. *Oncotarget.* Oct 30 2014;5(20):10170-10179.

102. Sharma PK, Singh R, Novakovic KR, Eaton JW, Grizzle WE, Singh S. CCR9 mediates PI3K/AKT-dependent antiapoptotic signals in prostate cancer cells and inhibition of CCR9-CCL25 interaction enhances the cytotoxic effects of etoposide. *International journal of cancer.* Nov 1 2010;127(9):2020-2030.

103. Johnson-Holiday C, Singh R, Johnson EL, Grizzle WE, Lillard JW, Jr., Singh S. CCR9-CCL25 interactions promote cisplatin resistance in breast cancer cell

through Akt activation in a PI3K-dependent and FAK-independent fashion. *World journal of surgical oncology.* May 3 2011;9:46.

104. Johnson EL, Singh R, Johnson-Holiday CM, et al. CCR9 interactions support ovarian cancer cell survival and resistance to cisplatin-induced apoptosis in a PI3K-dependent and FAK-independent fashion. *Journal of ovarian research.* Jun 17 2010;3:15.

105. Vishnubhotla P, Carr AC, Khaled A, Bassiouni R, Khaled AR. CT20p as a therapeutic for lung cancer with elevated chaperonin containing TCP1 (CCT) expression levels. *Journal of Clinical Oncology.* 2017;35(15_suppl):e23163-e23163.

106. Tomei S, Marchetti I, Zavaglia K, et al. A molecular computational model improves the preoperative diagnosis of thyroid nodules. *BMC cancer.* Sep 7 2012;12:396.

107. Mano Y, Takahashi K, Ishikawa N, et al. Fibroblast growth factor receptor 1 oncogene partner as a novel prognostic biomarker and therapeutic target for lung cancer. *Cancer science.* Dec 2007;98(12):1902-1913.

108. Liu Y, Marks K, Cowley GS, et al. Metabolic and functional genomic studies identify deoxythymidylate kinase as a target in LKB1-mutant lung cancer. *Cancer discovery.* Aug 2013;3(8):870-879.

109. Laurila E, Savinainen K, Kuuselo R, Karhu R, Kallioniemi A. Characterization of the 7q21-q22 amplicon identifies ARPC1A, a subunit of the Arp2/3 complex, as a regulator of cell migration and invasion in pancreatic cancer. *Genes, chromosomes & cancer.* Apr 2009;48(4):330-339.

110. Artero-Castro A, Castellvi J, García A, Hernández J, y Cajal SR, LLeonart ME. Expression of the ribosomal proteins Rplp0, Rplp1, and Rplp2 in gynecologic tumors. *Human pathology.* 2011;42(2):194-203.

111. Zhang Y-Z, Zhang L-H, Gao Y, et al. Discovery and validation of prognostic markers in gastric cancer by genome-wide expression profiling. *World journal of gastroenterology: WJG.* 2011;17(13):1710.

112. Levina V, Nolen BM, Marrangoni AM, et al. Role of eotaxin-1 signaling in ovarian cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Apr 15 2009;15(8):2647-2656.

113. Lee S, Oh T, Chung H, et al. Identification of GABRA1 and LAMA2 as new DNA methylation markers in colorectal cancer. *International journal of oncology.* 2012;40(3):889-898.

114. Jacobsen F, Taskin B, Melling N, et al. Increased ERCC1 expression is linked to chromosomal aberrations and adverse tumor biology in prostate cancer. *BMC cancer.* 2017;17(1):504.

115. Bellmunt J, Group OboSOG, Paz-Ares L, et al. Gene expression of ERCC1 as a novel prognostic marker in advanced bladder cancer patients receiving cisplatin-based chemotherapy. *Annals of Oncology.* 2007;18(3):522-528.

116. Mattie M, Raitano A, Morrison K, et al. The discovery and preclinical development of ASG-5ME, an antibody–drug conjugate targeting SLC44A4-positive epithelial tumors including pancreatic and prostate cancer. *Molecular cancer therapeutics.* 2016;15(11):2679-2687.

117. Huang XP, Zhao CX, Li QJ, et al. Alteration of RPL14 in squamous cell carcinomas and preneoplastic lesions of the esophagus. *Gene.* Jan 17 2006;366(1):161-168.

118. Palangat M, Anastasakis DG, Fei DL, et al. The splicing factor U2AF1 contributes to cancer progression through a noncanonical role in translation regulation. *Genes & development.* May 1 2019;33(9-10):482-497.

119. Gokmen-Polar Y, Neelamraju Y, Goswami CP, et al. Expression levels of SF3B3 correlate with prognosis and endocrine resistance in estrogen receptor-positive

breast cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* May 2015;28(5):677-685.

120. Adler AS, McCleland ML, Yee S, et al. An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth. *Genes & development.* May 15 2014;28(10):1068-1084.

121. Wang D, Shi W, Tang Y, et al. Prefoldin 1 promotes EMT and lung cancer progression by suppressing cyclin A expression. *Oncogene.* 10/03/online 2016;36:885.

122. Sasaki H, Yukiue H, Kobayashi Y, et al. Expression of the cdc25B gene as a prognosis marker in non-small cell lung cancer. *Cancer letters.* Nov 28 2001;173(2):187-192.

123. Phelps DL, Borley JV, Flower KJ, et al. Methylation of MYLK3 gene promoter region: a biomarker to stratify surgical care in ovarian cancer in a multicentre study. *British journal of cancer.* May 9 2017;116(10):1287-1293.

124. Lai D, Ho KC, Hao Y, Yang X. Taxol resistance in breast cancer cells is mediated by the hippo pathway component TAZ and its downstream transcriptional targets Cyr61 and CTGF. *Cancer research.* Apr 1 2011;71(7):2728-2738.

125. Yu FX, Zhang Y, Park HW, et al. Protein kinase A activates the Hippo pathway to modulate cell proliferation and differentiation. *Genes & development.* Jun 1 2013;27(11):1223-1232.

126. Chow MT, Luster AD. Chemokines in cancer. *Cancer immunology research.* Dec 2014;2(12):1125-1131.

127. Konstantinopoulos PA, Fountzilas E, Pillay K, et al. Carboplatin-induced gene expression changes in vitro are prognostic of survival in epithelial ovarian cancer. *BMC medical genomics.* 2008;1:59.

128. Chong T, Sarac A, Yao CQ, et al. Deregulation of the spindle assembly checkpoint is associated with paclitaxel resistance in ovarian cancer. *Journal of ovarian research.* 2018;11(1):27.

129. Bicaku E, Xiong Y, Marchion DC, et al. In vitro analysis of ovarian cancer response to cisplatin, carboplatin, and paclitaxel identifies common pathways that are also associated with overall patient survival. *British Journal Of Cancer.* 05/17/online 2012;106:1967.

130. Koussounadis A, Langdon SP, Harrison DJ, Smith VA. Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *Br J Cancer.* Jun 10 2014;110(12):2975-2984.

131. Michalides R, Griekspoor A, Balkenende A, et al. Tamoxifen resistance by a conformational arrest of the estrogen receptor alpha after PKA activation in breast cancer. *Cancer cell.* Jun 2004;5(6):597-605.

132. Merkle D, Hoffmann R. Roles of cAMP and cAMP-dependent protein kinase in the progression of prostate cancer: cross-talk with the androgen receptor. *Cellular signalling.* Mar 2011;23(3):507-515.

133. Elaskalani O, Razak NB, Falasca M, Metharom P. Epithelial-mesenchymal transition as a therapeutic target for overcoming chemoresistance in pancreatic cancer. *World journal of gastrointestinal oncology.* Jan 15 2017;9(1):37-41.

134. Kajiyama H, Shibata K, Terauchi M, et al. Chemoresistance to paclitaxel induces epithelial-mesenchymal transition and enhances metastatic potential for epithelial ovarian carcinoma cells. *International journal of oncology.* Aug 2007;31(2):277-283.

135. Rho JK, Choi YJ, Lee JK, et al. Epithelial to mesenchymal transition derived from repeated exposure to gefitinib determines the sensitivity to EGFR inhibitors in A549, a non-small cell lung cancer cell line. *Lung cancer (Amsterdam, Netherlands).* Feb 2009;63(2):219-226.

136. Sequist LV, Waltman BA, Dias-Santagata D, et al. Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Science translational medicine.* Mar 23 2011;3(75):75ra26.

137. Zhao J, Shao J, Zhao R, et al. Histological evolution from primary lung adenocarcinoma harboring EGFR mutation to high-grade neuroendocrine carcinoma. *Thoracic cancer.* Jan 2018;9(1):129-135.

138. Lim JU, Woo IS, Jung YH, et al. Transformation into large-cell neuroendocrine carcinoma associated with acquired resistance to erlotinib in nonsmall cell lung cancer. *The Korean journal of internal medicine.* Nov 2014;29(6):830-833.

139. Baglivo S, Ludovini V, Sidoni A, et al. Large Cell Neuroendocrine Carcinoma Transformation and EGFR-T790M Mutation as Coexisting Mechanisms of Acquired Resistance to EGFR-TKIs in Lung Cancer. *Mayo Clinic Proceedings.* 2017;92(8):1304-1311.

140. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England).* Oct 2014;30(19):2811-2812.

141. Comprehensive genomic characterization of squamous cell lung cancers. *Nature.* Sep 27 2012;489(7417):519-525.

142. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* Jul 18 2012;487(7407):330-337.

This page intentionally left blank