Molecular Genetics and Metabolism 128 (2019) 45–52

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

Review article

Multigene signatures of responses to chemotherapy derived by biochemically-inspired machine learning

Peter K. Rogan

Departments of Biochemistry, Oncology, and Computer Science, University of Western Ontario, London, ON N6A 2C1, UK

ARTICLE INFO

Keywords:
Chemotherapy response
Machine learning
Gene signatures
Breast carcinoma
Bladder carcinoma
Cancer cell lines
Patient validation
Gene expression
Copy number

ABSTRACT

Pharmacogenomic responses to chemotherapy drugs can be modeled by supervised machine learning of expression and copy number of relevant gene combinations. Such biochemical evidence can form the basis of derived gene signatures using cell line data, which can subsequently be examined in patients that have been treated with the same drugs. These gene signatures typically contain elements of multiple biochemical pathways which together comprise multiple origins of drug resistance or sensitivity. The signatures can capture variation in these responses to the same drug among different patients.

1. Background

Current pharmacogenetic analysis relates genotypes of various individual genes to their impact on transport, biotransformation, or disposition of drugs in patients. However, in cancer chemotherapy, unrelated gene products significantly contribute to the overall cellular and clinical responses by impacting elements of other biochemical pathways that respond to these agents in tumors [1]. The effects of multiple genes, termed gene signatures, have been used to predict chemotherapy response in cell lines using differential gene expression (DGE) as well as machine learning techniques (ML; [2,3]). These involve measurements of combinations of expression (GE) and/or DNA copy number (CN) levels as surrogates for cancer cell growth ([4–8]; Fig. 1).

Machines learn to classify by means of loss functions. This method evaluates how well a specific algorithm models the given data. If predictions deviate too much from the actual results in the training data, the loss function generates a large number. Hinge loss is one type of loss function that maximizes the impact or weight of training data distant from the threshold that distinguishes the drug sensitive from resistant classes of patients or cell lines using support vector machines (SVMs). When the same type of loss is compared among many ML models (each consisting of a different combination and weights of genes), a lower loss indicates a better predictive model.

DGE gene signatures for drug response have been based on on the average differences in gene co-expression between sensitive and resistant tumor tissues among a set of patients [9]. These signatures have traditionally selected genes based on the largest overall changes in expression levels among 2 (or more) groups of patients (for example, complete pathological remission vs recurrent disease). For a selected gene, the variance among members of the same group can be large, resulting in overlap in the expression levels between the groups. While DGE maximizes differences between the mean expression levels of the groups, there is often considerable overlap in overall expression over the quartiles adjacent the mean, resulting in only a subset of individuals exhibiting significant differences between classes. By contrast, highly weighted genes in the best performing ML models can exhibit a lower degree of overlapping expression between sensitive and resistant categories, due to lower coefficients of variation among these classes. Another distinction between DGE and ML approaches is that in DGE, while individual expression values share information with chemotherapy response, many of these genes may reveal similar information, and are often redundant in the signatures themselves. The process of selecting genes, ie. features, for ML-based models attempts to minimize redundant features. This reduces a source of noise in the data and mitigates against overfitting of the resultant signature to a particular dataset of cell lines or tumors.

Direct comparisons of DGE and the ML models can be challenging because the impact of different gene combinations is not additive in non-linear models. Correlation of statistical test results with weights of ML model features may impacted by the order of gene removal when determining misclassification accuracy corresponding to the weights of individual genes. Equally weighted genes in the DGE signatures can

E-mail address: progan@uwo.ca.

https://doi.org/10.1016/j.ymgme.2019.08.005
Received 13 May 2019; Received in revised form 9 August 2019; Accepted 16 August 2019
Available online 19 August 2019

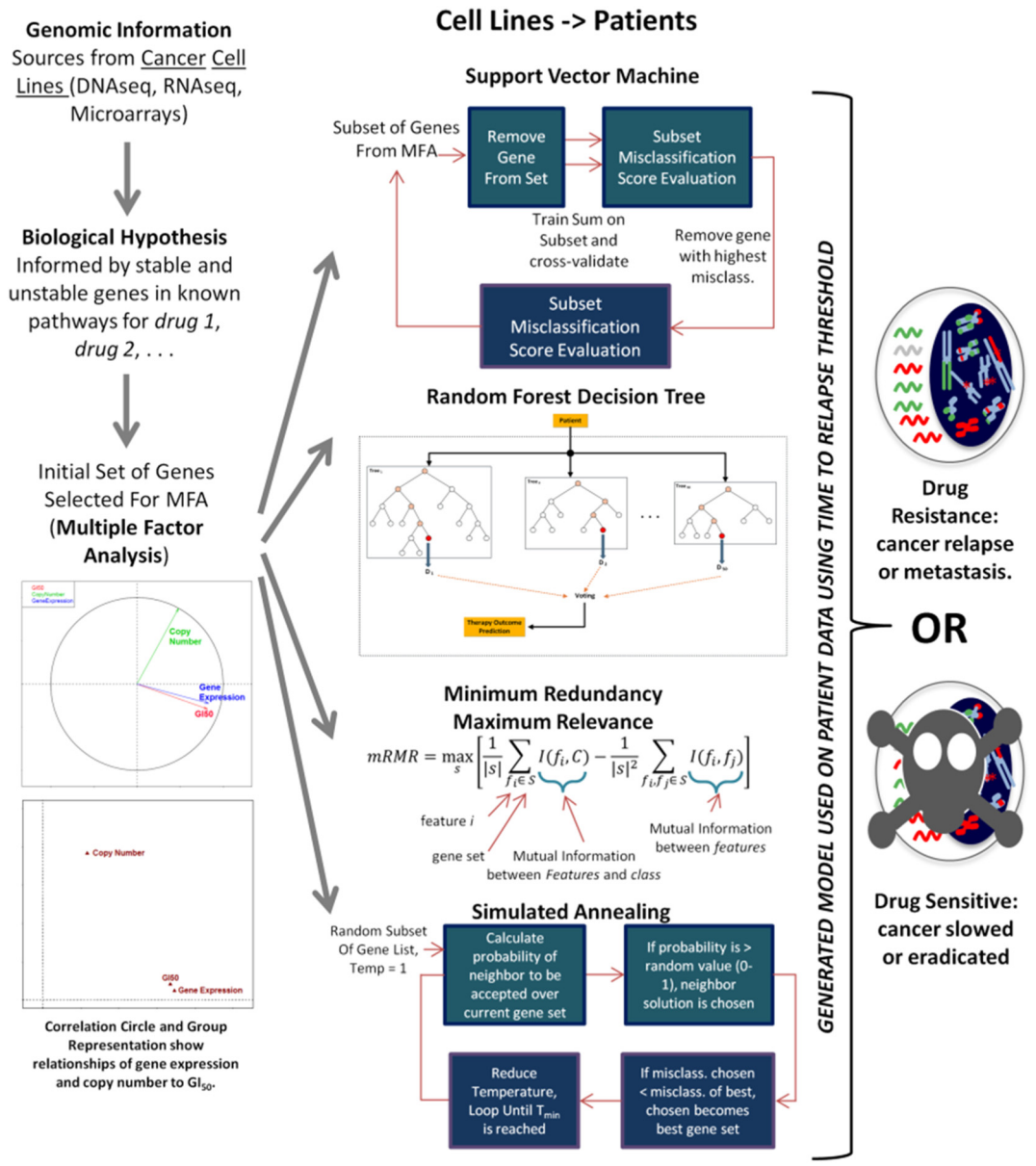

Fig. 1. Biochemically-inspired gene signature workflow. Genes biologically relevant to drug response were collected through examination of the published literature and public databases. Association of GE and/or CN with GI_{50} was tested in drug-treated 49 breast [8] and 18 bladder carcinoma [7] cell lines. Multiple Factor Analysis shows either positive or inversely correlated relationships between the GI_{50} and GE or CN as coincident or opposing vectors. These genes were then used to build models for the drug using the following machine learning techniques: Support Vector Machine Learning, Random Forest Decision Trees, Minimum Redundancy Maximum Relevance wrapper for feature selection, and Simulated Annealing. Those models were subsequently used on patient data to predict drug response.

cause genes with significantly overlapping expression levels to dilute the contributions of those with discrete distributions, affecting the overall performance of these signatures in discriminating between the classes of drug response.

The effectiveness of adjuvant chemotherapy agents has been related to changes in the profile of tumor gene expression [10–15]. Genomic signatures for chemotherapy (CT) response using supervised ML for breast cancer (BRCA) have been developed by discriminating sensitive from resistant cancer cell lines at the median concentration of drug that inhibits growth of these lines by half (GI_{50}; [3,16–19]). GI_{50} is an excellent measure of drug effectiveness and proxy for clinical outcomes, since it is a holistic approach that incorporates effects on direct targets as well as comprehensive cellular response, including other biochemical pathways [16]. This output is what is effectively being modeled based on GE and/or CN using biochemically-inspired ML approaches.

Genes with biological relevance to drug response were identified based on evidence in published literature and public cancer drug databases. GE and/or CN values were then filtered using available BRCA and bladder cancer (BLC) cell line microarray or next generation sequencing data by Multiple Factor Analysis (MFA), a statistical method, similar to principal component analysis, that quantifies relationships between GI_{50} and other variables. The genes eligible to be selected as ML features have GE and/or CN values that either correlate directly or inversely with GI_{50}. One optimal set of gene features classifies cancer cell lines as either resistant or sensitive to drugs using their respective GE and/or CN values [4]. In general, the median GI_{50} theshold of cell lines has the highest positive predictive value for distinguishing between these classes [5]. ML models select the GE and CN features and weights with the lowest possible classification error of drug resistance and sensitivity [20,21] by feature selection. ML approaches for predicting drug response have included: (a) Support Vector Machines (SVM; [22]), which selects genes that best discriminate classes along a

multidimensional hyperplane consisting of gene features, (b) Random Forest classifier (RF; [23]), a type of decision tree that votes for the most frequently selected subsets of genes [5,15] that separate the classes, and (c) a Minimum Redundancy Maximum Relevance (mRMR; [24]) wrapper that incrementally adds features that maximize the mutual information between gene features and classes, while keeping the redundancy between gene features at a minimum level [6]. CT signatures have also been derived by Simulated Annealing (SA; [25]), which minimizes errors in ML models by iteratively selecting gene combinations that incrementally improve classification accuracy. The performance of derived signatures are also compared with other published gene signatures as well as models derived from randomly selected genes - to assess significance of the null hypothesis [4].

2. Application of cell line-based gene signatures to CT patients

ML signatures [2] were used for prognosis of BRCA tumor response to paclitaxel and gemcitabine, which was more accurate than previous hierarchical clustering of mean differences in expression [26]. Gene signatures were derived by either by backward, forward and complete feature selection with SVMs [5], then tested by cross-validation with cell lines excluded from the original signatures. Cell line-derived signatures were validated for paclitaxel, tamoxifen, methotrexate, 5-fluorouracil, epirubicin, and doxorubicin with a 2000 BRCA patient dataset [3,27] with multiple ML approaches. These SVM- and RF-derived signatures were also assessed for their ability to provide prognosis of response, using thresholds based on the duration of progression free survival (PFS) of patients. This approach retrospectively identified gene signatures that may have been useful in guiding selection of specific CT agents for treatment [3]. These models exhibited higher accuracy than traditional signatures based upon DGE. ML signatures for particular chemotherapeutic drugs performed best on patient datasets where the patients received that specific chemotherapeutic drug. However, cross-application of ML models for individual drugs in patients with other types of cancer were found to be less accurate for prognosis [4], as well as in patients receiving drug combinations [5].

The performance of other ML approaches were evaluated for the MFA-derived gene set at different PFS or time to tumor progression (TTP) thresholds. mRMR feature selection of a paclitaxel-based SVM classifier based on expression of *ABCB11, ABCC1, BAD, BBC3* and *BCL2L1* was 79% accurate in 53 CT patients. By contrast, an RF classifier produced a gene signature (*ABCB11, ABCC1, BAD, BCL2, CYP2C8, CYP3A4, MAP4, MAPT, NR1I2, TUBB1, GBP1, OPRK1*) that gave prognoses of > 3 year survival with 82.4% accuracy in 420 hormone therapy (HT) patients. A similar RF gene signature showed 79.6% accuracy in 504 patients treated with CT and/or HT. These results suggest that tumor gene expression signatures refined by ML techniques can be useful for prognosis of PFS and TTP after drug therapies.

3. Example: gene signature of cisplatin response

Cisplatin covalently cross-links adjacent purines in the genome, which elicits increased DNA damage [28] and anti-oxidant scavenging responses at the highest levels of resistance. There are differences in expression of G1/S DNA Damage Checkpoint, Base Excision Repair, and Nucleotide Excision Repair genes that distinguish sensitive and resistant BLC cell lines treated with cisplatin. MFA for cisplatin on BRCA and BLC cell lines (Fig. 2), shows GI_{50} to be strongly correlated with GE of *ATP7B, BCL2L1, CDKN2C, GSTP1, MSH2, MAP3K1, MT1A, MT2A, MT4, NFKB2, SLC22A5, SLC22A7, SLC22A11, SLC22A12, SLC22A15, SLC31A2, SNAI1,* and *TLR4,* and gene copy number of *CFLAR, FOS,* and *NFKB1*. GE of *GSTO1, MAPK13, MT3, PPKAA2, PRKCA, PRKCB, SLC22A10, SLC22A13* and *TP63* and CN of *MAPK3* and *SLC22A20* were also correlated, but at lower significance. Results were consistent with published studies [29–31], however many candidate genes previously associated with cisplatin response were not correlated based on GI_{50}. The ML-based gene signature was 95% accurate in classifying 41 cell lines with the median GI_{50} as the threshold distinguishing sensitivity from resistance.

An SVM signature containing DNA repair genes: *CHEK2, NEIL1, PNKP, POLD1, POLD2, POLD3, POLE, POLR2H, PSMA2, PSMC2,* and *RFC2* expression was used to predict prognosis to segregate a set of 30 BLC patients with advanced disease (NCBI GEO database: dataset GSE5287) [5]. Long term survival was clustered according to good (Cluster 1), intermediate (Cluster 2), and poor (Cluster 2.1). These results were consistent for an independent set of BLC patients (GEO: GSE31684; Fig. 3).

Previously, the threshold that distinguished drug resistance and sensitivity was the median GI_{50} value, which consistently has among the highest positive predictive value in different patient datasets [5,6]. However, at different GI_{50} thresholds, signatures are obtained that can preferentially distinguish the genes contributing to the highest vs. lowest levels of drug resistance. GI_{50}-thresholded ML models were derived by minimizing either misclassification or a log-loss function to evaluate the distribution of selected genes and model accuracy. Log-loss penalizes false classifications, whose value ranges from zero (or completely accurate) to 1 (or completely inaccurate). The overall distribution of genes across various GI_{50} thresholds exhibited both similarities and differences with gene signatures derived by minimizing classification errors at the median GI_{50} threshold by backwards feature selection. However, varying these thresholds produces imbalances between the numbers of sensitive and resistant cell lines, which can affect the performance of ML models at extreme GI_{50} thresholds [32,33]. An important question is whether the genes contributing to drug responses are consistent among different cell lines, each with their own unique GI_{50} values. Different ML gene signatures were obtained by shifting the GI_{50} threshold, which changed the labels of the resistant and sensitive cell lines. After feature selection, the compositions of the corresponding gene signatures for each threshold were compared. Finally, ensemble averaging of all of these optimized SVMs, each derived for different GI_{50} thresholds, was used to create a single aggregated, threshold-independent signature with fewer independent features (i.e., a composite gene signature). Aggregated, threshold-independent models generated for each of the platin drugs at different GI50 thresholds, ie. by ensemble machine learning, classified bladder cancer patients with similar accuracy (50–63% disease free; 48–73% recurrent). Although the compositions of the component GI50-thresholded signatures emphasize different genes and pathways, their overall performance for discriminating sensitivity from resistance (between 12 and 24 months post-treatment) were similar across the set of patient data.

Kinase genes (*MAPK3* and *MAP3K1*) and apoptotic family members (*BCL2*, *BCL2L1*) were the most consistently represented in different cisplatin signatures at multiple GI_{50} thresholds,; error-prone and base-excision DNA repair genes were also present, but were less frequent (Fig. 4). The kinase genes were more concentrated in signatures associated with increased sensitivity to the drug, whereas *BCL2* and *BCL2L1* were more ubiquitous and found at all threshold levels. The error prone polymerases *POLD1* and *POLQ* were more frequently detected in gene signatures with lower sensitivity thresholds, while the flap endonuclease *FEN1* tended to be present in signatures that distinguished high resistance levels. By contrast, thresholded gene signatures for carboplatin-related genes commonly contained the apoptotic family member *AKT1*, transcription regulation genes *ETS2* and *TP53*, as well as cell growth factors *VEGFB* and *VEGFC*, although the latter were less common at lower sensitivity thresholds. Common oxaliplatin-related genes included the transporters *SLCO1B1* and *GRTP1* (but not *SLC47A1*), transcription-related genes *NFE2L2*, *PARP15* and *CLCN6*, and metabolism-related genes. These analyses showed certain pathways that contributed to higher levels of resistance, whereas others are prevalent at lower levels of sensitivity.

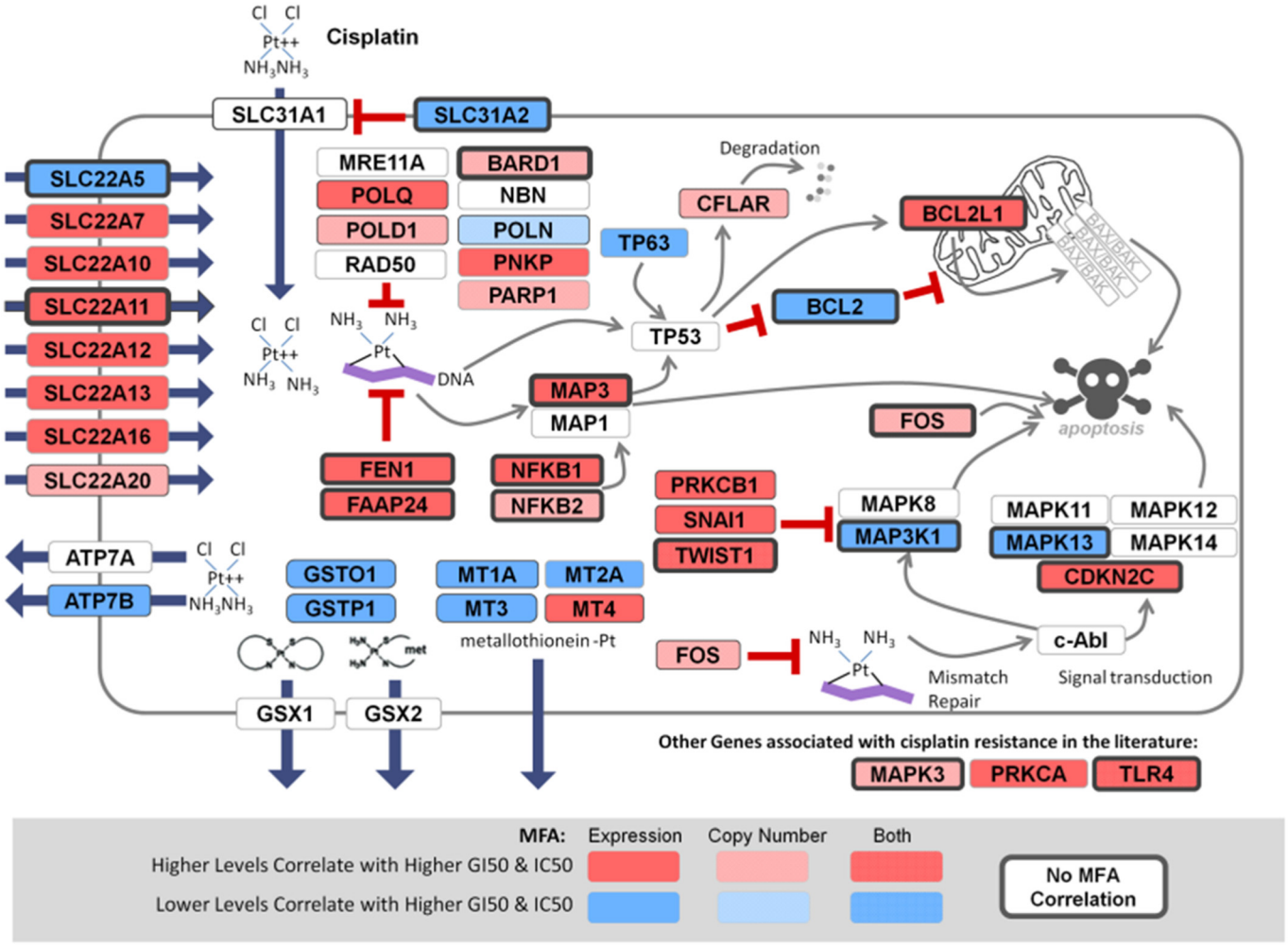

Fig. 2. Schematic of cisplatin sensitivity and resistance genes. Genes are indicated according to their role in drug metabolism. Red boxes indicate genes with a positive correlation between gene expression or copy number to GI_{50} and IC_{50} using Multiple Factor Analysis. Blue demonstrates a negative correlation. Genes circumscribed by bolded rectangles with rounded edges were selected by ML for inclusion in the gene signature. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Rational expansion of gene candidates in signatures

The best ML signatures correctly identify the CT responses of the majority of cell lines and patients according to GI_{50} threshold; however, some individuals remain misclassified. Aside from limitations in the computational approach, it is also reasonable to consider limitations on existing knowledge regarding the complement of genes associated with drug responses, i.e. the published literature is an incomplete source of information about all of the gene products that affect CT response.

We have described ML approaches that derive multigene signatures that predict drug response containing gene features that are weighted to optimally distinguish sensitive from resistant classes of cell lines and patients. As a consequence, certain input features will be eliminated or emphasized in performing this classification, but these do not reveal previously unknown genes that contribute to the response. Extending signatures to include other features from the same or related pathways is a foundational strategy that is based on established biochemical mechanisms and relationships revealed from systems biology. Many of these relationships have not been elucidated, however the principles of these interactions are well established. Feedback regulation involves the use of a reaction product to regulate its own or a another reaction. Interacting or multi-subunit protein complexes would also be subject to this type of constraint. These types of regulation tend to affect activity other elements in the same pathways, however these processes are not exclusive to any particular pathway or set of interactions. They comprise ubiquitous network motifs in all kinds of molecular interaction networks, such as for example metabolic networks, regulatory modules or signaling networks [34]. Negative feedback, which counteracts external perturbations, can cause oscillating behavior, but also has a stabilizing effect. Negative feedback may endow cells with robustness to internal and external perturbations and play a major role in maintaining homeostasis [35–37]. Positive feedback is also critical in cellular decision processes, by producing ultrasensitivity and prolonged responses to a transient external signal [37–39]. While positive and negative feedback loops are ubiquitous in biochemical signaling pathways, an additional effect called quasi-bistability can contribute to observed responses to different stimuli. Quasi-bistability allows monostable systems to maintain two distinct states upon receipt of a transient input, which may be related to positive feedback loops [40].

The discovery of other genes whose GE and/or CN contribute to chemotherapy response can be guided by interactions with the gene products present in existing signatures. Gene products within the same or linked biochemical pathways may contribute to drug response through direct or indirect feedback, and through epistatic relationships to existing genes in these signatures. Expansion of the set of candidate signature gene products is therefore based on their established proximity in biochemical pathways, direct interactions, and regulatory relationships [41] to gene products in the original signatures. Although

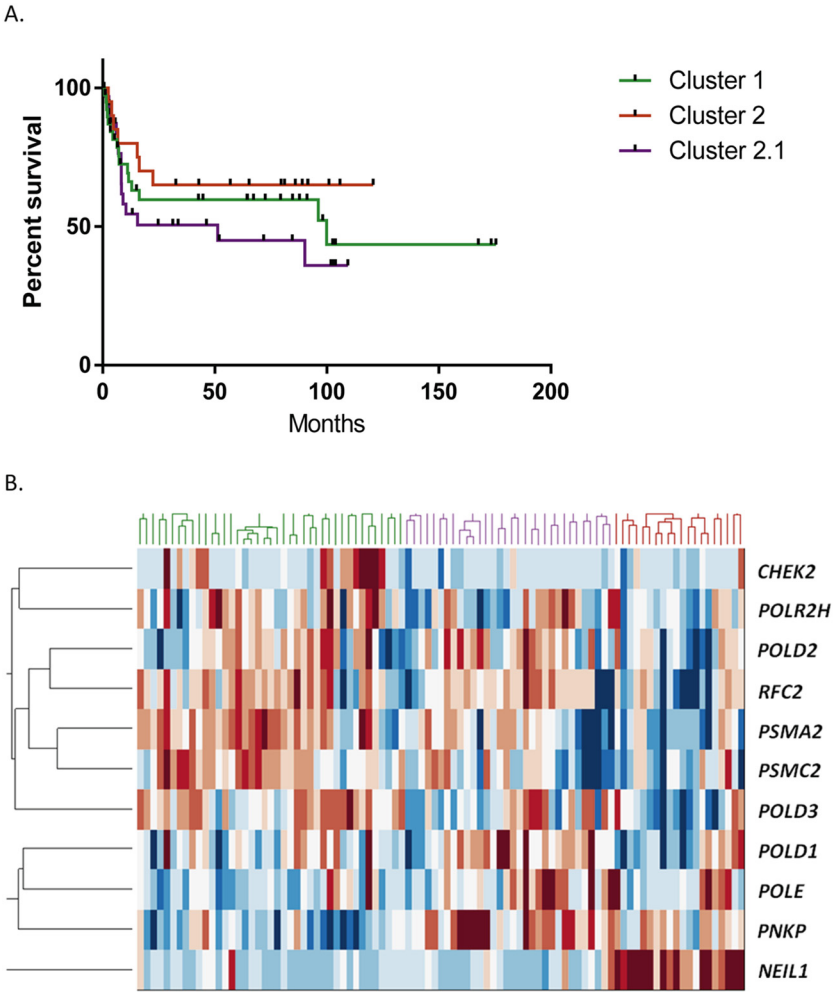

Fig. 3. Clustering cisplatin-treated patients based on expression of damage response genes in signature.
Of 146 genes involved in G1/S DNA damage checkpoints ($N = 61$), base excision repair ($N = 35$), and nucleotide excision repair ($N = 50$), 11 showed significant differences in gene expression between 4 sensitive and 6 resistant bladder cell lines[10]. These 11 genes were selected for unsupervised clustering of expression in treated patients in NCBI GEO datasets, GSE5287 and GSE31684. (A) indicates differences in survival for patients in GSE31684 with poor (2.1), intermediate [1], and good [2] prognostic outcomes (percent survival) in a Kaplan-Meier plot based on expression of these genes. (B) shows the heatmap for these clusters by patient (GSE31684; $n = 144$; red = high, blue = low expression). In this case, the differences between the outcome categories were not statistically significant. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

software has been developed and applied to automate this process, the extent to which particular pathways, interactions or forms of regulation are preferentially incorporated into the expanded signatures has not yet been established. Another limitation is that the resources used for pathway expansion of existing signatures do not currently account for steady-state vs. dynamic differences between interactions based indirectly on gene expression changes and actual biochemical regulation of the drug targets and metabolites.

Neighboring genes may be related either through direct interaction or by substrate-product dependency within that same or adjacent

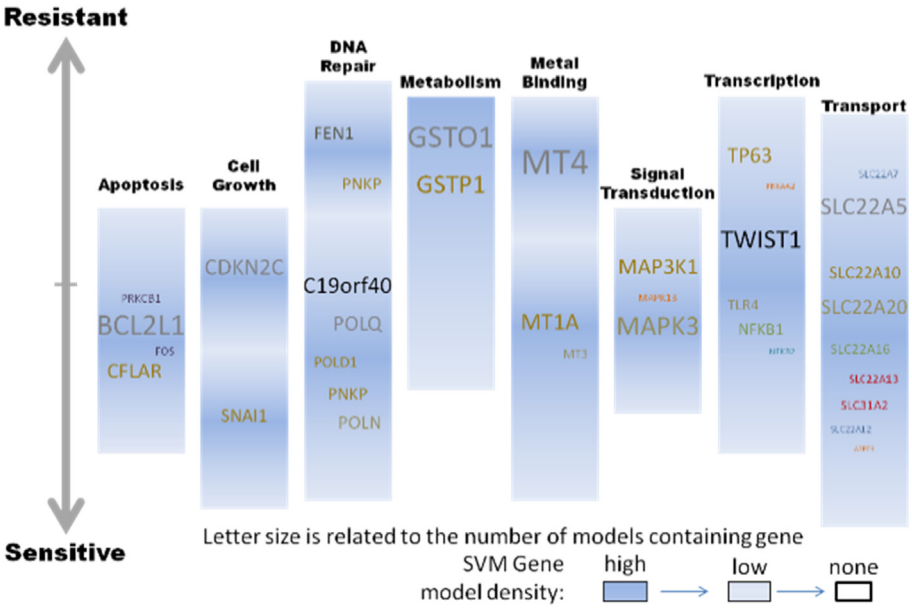

Fig. 4. Ensemble of gene signatures for cisplatin at different GI_{50} thresholds. Each column represents the contributions of genes (number of genes by pathway: 4 apoptosis, 2 cell growth, 9 DNA repair, 2 metabolism, 3 metal binding, 3 signal transduction, 7 transcription and 9 transport) at different GI_{50} levels. Font size corresponds to frequency of gene in different gene signature models.

biochemical pathways, with adjacency defined as the number of pathway nodes separated from the original signature gene. This involves referencing a comprehensive, updated digital source of biochemical reaction and interaction pathways, for example, from Pathway Commons (http://www.pathwaycommons.org/), which includes the Reactome, ConsensusPathDB, and KEGG databases. To discover either previously unknown genes, or replace existing genes and create higher accuracy signatures, well-established functional relationships between original and novel genes are leveraged. This integrated resource enables efficient computational searching of hierarchical and other relationships between interacting gene products and reaction dependences of initial signature genes. Candidates are selected based on known relationships (regulatory, associations) with gene products included in gene signatures, and systematically, based on network proximity to those in the current signature. By systematically traversing the pathway graph for each signature gene, candidates are selected for another round of machine learning based on GE and/or CN features that exhibit significant direct or inverse correlations to GI_{50} values (Pearson correlation coefficient: $-0.8 < r > 0.8$). Neighbors for each signature gene are present at each level of the biochemical and interaction network. Inclusion of neighboring genes can be influenced by more distant node interactions that interact with gene products encoded in the original signatures. Pathway expansion can produce large numbers of potential gene candidates that are correlated with GI_{50} (Table 1). To minimize false discovery of novel signature genes, we have limited permissible distances between original and derivative genes to a depth of up to 3 nodes.

Analysis of extended cisplatin signatures based on manual testing and inclusion of genes from a single adjacent pathway node demonstrated that two new gene neighbors improved accuracy of prognosis of remission, and one gene improved prognosis of recurrent disease. Candidate immediate neighboring genes to expand cisplatin signatures (Table 1) revealed genes in neighboring biochemical pathways that, when introduced, marginally improved accuracy (2–4%, depending on the gene) of Cis1, the best performing published signature [7].

Expanded signatures for many drugs are still in the process being analyzed, however results for tyrosine and serine kinase inhibitors (sunitinib, sorafenib, lapatinib, erlotinib, imatinib and gefitinib) are currently the most mature. Initial ML-based classifiers were originally based on backwards feature selection of response genes from published literature on these drugs. Except for Sunitinib, the performance of all of these ML models appears to have been improved by extension of the biochemical network to include pathway neighbors. In the best performing datasets, overall classification accuracies for prognosis of patient response were 73% for erlotinib, 75% for lapatinib, 85% for sorafenib, 83% for sunitinib, 66% for imatinib, and 53% for gefitinib (however the latter dataset consisted of only 11 patients, of which sensitivity was correctly predicted in 4 of 4 patients and resistance in 2 of 7). Interestingly, the genes in the extended signatures replace most, but not all, of the literature-based classifiers. The improved signatures contain genes that are 1 or 2 nodes from the original gene from which it was associated. Generally, they are each composed of unique sets of genes (there is little overlap between the compositions of signatures for different kinase inhibitors).

Many additional genes are likely to be found with strong correlations to GI_{50} levels or other response measures. Inclusion of these additional features (in either the same signature or in the number of high performing signatures) could increase susceptibility of ML models to overfitting of the training data, which would falsely inflate their accuracy. Several strategies can be implemented to reduce the number of additional gene features, thereby minimizing the likelihood of model overfitting, including: (a) limiting genes to pathways that are represented predominantly in the original signatures with highest performance at extreme responses (most or least chemo-resistant), (b) filtering epistatic genes that merge gene features present in the same pathway based on evidence that they both exert the overlapping regulatory effects on a single pathway product or at the same node in a pathway, and (c) ensemble averaging of multiple gene signatures with equivalent performance can define consensus ML models with fewer features. The number of distant extant nodes in pathway networks can also be pruned to mitigate overfitting of genomic data. Features can also be prioritized by limiting the new genes to pathways that make the most significant contributions to CT resistance. These strategies could improve reproducibility and prognostic accuracy compared to the original gene signatures using new data sources.

The revised, expanded signatures can then be derived by ML training on cancer cell line GE and/or CN data at specified GI_{50} thresholds. Pathway node partners of genes in the existing gene signatures are retained if they either reduce misclassification rates or improve classification of the cell lines. The resultant signatures, including network partner genes, are compared with the previously-derived signatures and evaluated with patient genomic data to assess whether prognostic accuracy have also been improved for different clinical response duration thresholds. Results are limited solely to genes that significantly contribute by MFA analysis to chemotherapy response. The relative contribution of each gene to the chemotherapy response is determined by computing misclassification or log loss after removal of this gene from the signature. Robust gene signatures can be determined for different GI_{50} thresholds with hinge loss functions that weight GE/CN values of cell lines with outlier GI_{50} values more highly relative to those with GI_{50} values close to the mean threshold.

Gene expression signature expansion does not have to be restricted to polyadenylated mRNAs that encode proteins. The complex taxonomy of the RNA universe includes many other non-coding (nc) species, such as miRNA, RNA Polymerase III-derived transcripts (tRNA, small nuclear RNA, sno- and piwi-RNA, and Alu, small nuclear RNAs), enhancer RNAs, long non-coding (lnc) RNAs, circular (circ) RNAs, and RNA polymerase I transcripts, for example rRNAs, which comprise the most predominant nucleic acids in cells. It is feasible to computationally derive expression signatures that are prognostic for drug response without necessarily understanding their functional or phenotypic basis. For example, non-coding (nc) RNAs have recently been used to derive accurate gene signatures for predicting BLCA and lung adenocarcinoma patient responses to platinum therapies based on expression data from The Cancer Genome Atlas [42].

Nevertheless, inclusion of lncRNA and circRNA evidence in gene signatures will likely require that these species be related extrinsic properties of drug response over a range of measured phenotypes. Transcripts with recurrent mutation hotspots in essential coding domains that would result in localized exon skipping, cryptic splice site activation and/or intron inclusion detected by analysis of RNASeq would be reasonable candidates to consider for such signatures.

It is notable that different ncRNA species are identified, sequenced and measured by different protocols. ncRNAs are still not well covered in microarray platforms, but are evident by analysis of RNASeq libraries. This introduces challenges into deriving metaRNA signatures

Table 1
Correlated genes eligible for cisplatin extended gene signature.

Minimum correlation with GI_{50}	Number of correlated genes by node[a]:			
	0	1	2	3
≥ 0%	14	2451	10,780	340
≥ 70%	14	377	2035	82
≥ 80%	10	241	1291	50
≥ 90%	5	120	621	23
≥ 95%	4	61	292	11

[a] Nodes correspond to the number of biochemical pathway steps separating a gene from the initial gene signature [Node 0; same pathway], which includes: *ATP7B*, *BARD1*, *BCL2*, *CDKN2C*, *ERCC2*, *FOS*, *MAP3K1*, *MAPK13*, *NFKB1*, *PNKP*, *POLQ*, *PRKCA*, *SLC22A5* and *SNAI1*.

comprising multiple species, since globally-normalized expression levels are necessary for the signatures to be reproducible among different datasets. Nevertheless, conventional RNASeq libraries contain cDNA sequences from polyadenylated circRNA and lncRNAs, and are likely be generated concurrently with mRNAs. circRNAs are byproducts of mRNA splicing resulting from excision of non-contiguous donor and acceptor joins, usually by alternative splicing. Although highly expressed, a unified explanation for the functions of most circRNAs has not been found. The sequence content of short-read RNASeq cDNA libraries from mutated transcripts of circRNA and lncRNA could be indistinguishable from those generated by genomically-encoded mutations. We have found that mutations at splice sites can affect exon definition, resulting in intron inclusion, or exon skipping, cryptic isoforms or combinations of these [43]. It seems likely that some intronic reads originate from circRNAs, which themselves may be derived from different genomic variants. Common features in ML-based gene signatures could be consolidated from multiple distinct, rare point mutations that produce the same altered circRNA isoforms.

While gene signatures can be derived by inclusion of lncRNA and circRNA evidence, their utility remains to be established. Functional relationships to tumor or cell line response to a drug are needed. Strong candidates would include genes with recurrent mutation hotspots generating circRNAs in essential coding domains that would result in localized exon skipping, cryptic splice site activation and/or intron inclusion detected by analysis of RNASeq.

5. Prognostic utility of cell line-derived signatures for patient CT responses

Gene signatures based on tumor GE and/or CN data have been used to determine if ML-based gene signatures for sensitivity are related to patient response. In the absence of GI_{50}, EC_{50}, IC_{50} or other ground truth measures of drug response, primary study endpoints of patient clinical trials are used as surrogate measures for chemotherapy response thresholds. These may vary among available datasets and can include PFS, TTP, time to treatment failure (TTF) or proportionate complete response (CR) [44,45]. Gene signatures of invasive breast cancer performed well at the 4 year PFS threshold in 68.7% patients and 3 year CR in 74% patients [5]. There are limitations on interpreting drug signatures based on sample size, heterogeneous pathology, biases in patient ascertainment or study design. Studies with small numbers of enrolled patients should be reported as categorical results. Analyses of CT response have shown that the duration of clinical follow-up metadata has been adequate for larger studies (eg. METABRIC and the Cancer Genome Atlas). While the precise timing of administration of chemotherapy post-diagnosis may confound results, it is difficult to assess the impact of this, though this may explain some signatures with limited prognostic value ($< 60\%$ accuracy).

Approaches that are free of cell line endpoints (such as GI_{50}) can be used to set thresholds using patient-derived outcomes [4]. For example, TTP can be used to distinguish resistant from sensitivie drug phenotypes. Patients will be categorized as either sensitive or resistant, based on tumor status at specific time thresholds until either death or relapse. Varying these time thresholds will identify the gene signature model(s) that optimally discriminate these categories, with disease-free patients exceeding the threshold classified as sensitive. Patients treated with either surgery alone or radiation can serve as negative controls [4], as their outcomes are not expected to be related to a particular CT signature. If prognosis predicted by the signature produces similar proportions of sensitive vs. resistant patients in the CT vs control groups, the signature could not be related to the tumor GE and/or CN profile. The signature accuracy may also reflect selection for proliferation of tumor sub-clones, consistent with resistance arising in a subset of cells with accumulated genetic changes that confer resistance [44].

Datasets contain different numbers of patients with both available meta- and genomic data for a single drug treatment or a combination therapy. For multiple, independent patient datasets treated with the same CT, a signature is considered validated in different sets of patients if the accuracies are comparable. The minimum number of patients required to validate a gene signature depends on the balance between the census of sensitive vs. resistant individuals. Smaller patient datasets should only be analyzed categorically, without reporting statistical significance analysis, since these may reveal only trends in the data. With 77 or more patients, sensitivity and resistance can be distinguished at $p < .01$ significance with 80% power, assuming each category to be equally prevalent. However, resistance to CT was 1.8 fold more common in the METABRIC breast cancer cohort [3]. With this level of imbalance, 84 patients are required (54 and 30 patients, respectively). Our paclitaxel gene signature [4,26] was also based on an imbalanced outcome data, with approximately 4.2 patients developing recurrent disease for each CR. To achieve significance, 124 patients for the signature were required (100 recurrent and 24 CR).

The analysis of expressed genetic polymorphisms in RNA sequencing data from tumors could also potentially identify clonal events that define subpopulations of cells with different drug response phenotypes [45]. If these subpopulations were identified prior to or early in treatment, it might assist in timely selection of appropriate therapies. Such bioinformatic analysis of deep sequencing or single-cell next generation sequencing data is likely to provide valuable insight into whether the output of prognostic signatures reflects homogeneous or heterogeneous cell populations. These results could be important in understanding whether therapies operate on only a subset or all cells. Where karyotypes are often mosaic and polyclonal, it is conceivable that correlation of genotypes with signature predictions in different cells from the same tumor might reveal the cellular origins of drug resistant phenotypes.

6. Summary

Gene signatures that assist decisions to treat cancer with chemotherapy have been clinically approved by regulators after analysis of large patient cohorts [46–50]. These tests have been endorsed as mainstream medical practice and for government reimbursement [51]. The objective of incorporating such studies into chemotherapy management has a number of advantages for patients, including the elimination of redundant genetic information, their transferability and prognostic value, and the benefits of integrating cellular responses that incorporate different mechanisms of resistance into a single prognostic signature. Genomic signatures could be used by oncologists to avoid the use of drugs with unfavorable signatures or by substituting other drugs expected to be effective in a patient, either individually or in combination. We envision panels of CT signatures for multiple drugs that can be evaluated from the complete transcriptome data of a tumor. Response profiles may be capable of providing a set of prognostic drug responses for each patient. Such a strategy could assist in decisions to treat individuals with approved secondary CT rather than standard first-line adjuvant therapies.

Acknowledgements

PKR is supported by The Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs, and CytoGnomix. Compute Canada and Shared Hierarchical Academic Research Computing Network (SHARCNET) provided high performance computing and storage facilities.

References

[1] N.I. Park, P.K. Rogan, H.E. Tarnowski, J.H. Knoll, Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy, Mol. Oncol. 6 (2012) 347–359.

[2] R. Maglietta, A. D'Addabbo, A. Piepoli, F. Perri, S. Liuni, G. Pesole, N. Ancona, Selection of relevant genes in cancer diagnosis based on their prediction accuracy, Artif. Intell. Med. 40 (1) (2007) 29–44.

[3] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th edition, Elsevier, 2009.

[4] S. Dorman, K. Baranova, J. Knoll, B. Urquhart, G. Marciani, M.-L. Carcangiu, P.K. Rogan, Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Mol. Oncol. 10 (2016) 85–100.

[5] E. Mucaki, K. Baranova, H.P. Quang, I. Rezaeian, D. Angelov, L. Ilie, A. Ngom, L. Rueda, P.K. Rogan, Predicting outcomes of hormone and chemotherapy in the molecular taxonomy of breast cancer international consortium (METABRIC) study by biochemically-inspired machine learning, F1000Research 5 (2017) 2124.

[6] J.Z. Zhao, E.J. Mucaki, P.K. Rogan, Predicting ionizing radiation exposure using biochemically-inspired genomic machine learning, F1000Research 7 (2018) 233.

[7] E.J. Mucaki, J.Z. Zhao, D.J. Lizotte, P.K. Rogan, Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning, Sig. Transduct. Target. Ther. 4 (1) (2019) 1.

[8] L.M. Heiser, A. Sadanandam, W.L. Kuo, S.C. Benz, T.C. Goldstein, S. Ng, W.J. Gibb, N.J. Wang, S. Ziyad, F. Tong, N. Bayani, Subtype and pathway specific responses to anticancer compounds in breast cancer, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 2724–2729.

[9] A. Daemen, O.L. Griffith, L.M. Heiser, N.J. Wang, O.M. Enache, Z. Sanborn, F. Pepin, S. Durinck, J.E. Korkola, M. Griffith, J.S. Hur, Modeling precision treatment of breast cancer, Genome Biol. 14 (2013) R110.

[10] C.D. Hurst, M.A. Knowles, Molecular subtyping of invasive bladder cancer: time to divide and rule? Cancer Cell 25 (2) (2014) 135–136.

[11] W. Choi, S. Porten, S. Kim, D. Willis, E.R. Plimack, J. Hoffman-Censits, B. Roth, T. Cheng, M. Tran, I.L. Lee, J. Melquist, Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy, Cancer Cell 25 (2) (2014) 152–165.

[12] R. Seiler, W. Choi, L.L. Lam, N. Erho, C. Buerki, E. Davicioni, G.N. Thalmann, D.J. McConkey, P.C. Black, Association of p53-ness with chemo-resistance in urothelial cancers treated with neoadjuvant gemcitabine plus cisplatin, J. Clin. Oncol. 33 (15) (2015) 4512.

[13] J.S. Lee, S.H. Leem, S.Y. Lee, S.C. Kim, E.S. Park, S.B. Kim, S.K. Kim, Y.J. Kim, W.J. Kim, I.S. Chu, Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors, J. Clin. Oncol. 28 (16) (2010) 2660–2667.

[14] G. Sjödahl, M. Lauss, K. Lövgren, G. Chebil, S. Gudjonsson, S. Veerla, O. Patschan, M. Aine, M. Fernö, M. Ringnér, W. Månsson, A molecular taxonomy for urothelial carcinoma, Clin. Cancer Res. 18 (12) (2012) 3377–3386.

[15] E. Blaveri, J.P. Simko, J.E. Korkola, J.L. Brewer, F. Baehner, K. Mehta, S. DeVries, T. Koppie, S. Pejavar, P. Carroll, F.M. Waldman, Bladder cancer outcome and subtype classification by gene expression, Clin. Cancer Res. 11 (11) (2005) 4044–4055.

[16] R.H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen, Nat. Rev. Cancer 6 (2006) 813–823.

[17] R.M. Neve, K. Chin, J. Fridlyand, J. Yeh, F.L. Baehner, T. Fevr, L. Clark, N. Bayani, J.P. Coppe, F. Tong, T. Speed, A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes, Cancer Cell 10 (2006) 515–527.

[18] A. Prat, O. Karginova, J.S. Parker, C. Fan, X. He, L. Bixby, J.C. Harrell, E. Roman, B. Adamo, M. Troester, C.M. Perou, Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes, Breast Cancer Res. Treat. 142 (2013) 237–255.

[19] M. Hafner, M. Niepel, M. Chung, P.K. Sorger, Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs, Nat. Methods 13 (2016) 521–527.

[20] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. (2006) 861–874.

[21] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinform. 21 (2005) 3940.

[22] S. Abe, Support Vector Machines for Pattern Classification, Springer, 2010.

[23] L. Breiman, Random forests, J. Mach. Learn. 45 (1) (2001) 5–32.

[24] C. Ding, H.C. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinforma. Comput. Biol. 3 (2) (2003) 185–205.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, SIGKDD Explor. 11 (1) (2009) 10–18.

[26] C. Hatzis, L. Pusztai, V. Valero, D.J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, M. Martin, A genomic predictor of response and survival following taxane- anthracycline chemotherapy for invasive breast cancer, JAMA 305 (2011) 1873–1881.

[27] C. Curtis, S.P. Shah, S.F. Chin, G. Turashvili, O.M. Rueda, M.J. Dunning, D. Speed, A.G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature. 486 (2012) 346–352.

[28] L. Galluzzi, L. Senovilla, I. Vitale, J. Michels, I. Martins, O. Kepp, M. Castedo, G. Kroemer, Molecular mechanism of cisplatin resistance, Oncogene 31 (2012) 1869–1883.

[29] H.B. Grossman, R.B. Natale, C.M. Tangen, V.O. Speights, N.J. Vogelzang, D.L. Trump, R.W.D. White, M.F. Sarosdy, D.P. Wood Jr., D. Raghavan, E.D. Crawford, Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer, N. Engl. J. Med. 349 (2003) 859–866.

[30] J.J. Meeks, J. Bellmunt, B.H. Bochner, N.W. Clarke, S. Daneshmand, M.D. Galsky, N.M. Hahn, S.P. Lerner, M. Mason, T. Powles, C.N. Sternberg, A systematic review of neoadjuvant and adjuvant chemotherapy for muscle-invasive bladder cancer, Eur. Urol. 62 (3) (2012) 523–533.

[31] J. Bellmunt, H. von der Maase, G.M. Mead, I. Skoneczna, M. De Santis, G. Daugaard, A. Boehle, C. Chevreau, L. Paz-Ares, L.R. Laufman, E. Winquist, Randomized phase III study comparing paclitaxel/cisplatin/gemcitabine and gemcitabine/cisplatin in patients with locally advanced or metastatic urothelial cancer without prior systemic therapy: EORTC intergroup study 30987, J. Clin. Oncol. 30 (10) (2012) 1107–1113.

[32] R. Batuwita, V. Palade, Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatic datasets learning, J. Bioinforma. Comput. Biol. 10 (4) (2012) 1250003.

[33] R. Batuwita, V. Palade, Class imbalance learning methods for support vector machines, Imbalanced Learning: Foundations, Algorithms and Applications, Wiley, 2012.

[34] U. Alon, An Introduction to Systems Biology - Design Principles of Biological Circuits. Math and Comput Biol Series, Chapman & Hall/CRC, London, 2006.

[35] R. Thomas, On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations, in: J. Della-Dora, J. Demongeot, B. Lacolle (Eds.), Numerical Methods in the Study of Critical Phenomena, Springer Series in Synergetics, vol. 9, Springer, Berlin, Heidelberg, 1981, pp. 180–193.

[36] J.-L. Gouzé, Positive and negative circuits in dynamical systems, J. Biol. Syst. 6 (21) (1998) 11–15.

[37] M. Freeman, Feedback control of intracellular signalling in development, Nature 408 (2000) 313–319.

[38] U. Alon, Network motifs: theory and experimental approaches, Nat. Rev. Genet. 8 (2007) 450–461.

[39] M.A. Savageau, Comparison of classical and autogenous systems of regulation in inducible operons, Nature 252 (5484) (1974) 546–549.

[40] A. Jensch, C. Thomaseth, N.E. Radde, Sampling-based Bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines, BMC Syst. Biol. 11 (2017) 11.

[41] E. van den Akker, B. Verbruggen, B. Heijmans, M. Beekman, J. Kok, E. Slagboom, M. Reinders, Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis, J. Integr. Bioinform. 8 (2) (2011) 188.

[42] Y. Zhu, Y. Zhao, S. Dong, L. Liu, L. Tai, Y. Xu, Systematic identification of dysregulated lncRNAs associated with platinum-based chemotherapy response across 11 cancer types, Genomics (2019), https://doi.org/10.1016/j.ygeno.2019.07.007.

[43] B.C. Shirley, E.J. Mucaki, P.K. Rogan, Pan-cancer repository of validated natural and cryptic mRNA splicing mutations, F1000Research 7 (2019) 1908.

[44] C.G. Mullighan, L.A. Phillips, X. Su, J. Ma, C.B. Miller, S.A. Shurtleff, J.R. Downing, Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia, Science 322 (2008) 1377–1380.

[45] K. Yizhak, F. Aguet, J. Kim, J.M. Hess, K. KÜBLER, J. Grimbsby, R. Frazer, H. Zhang, N.J. Haradhvala, D. Rosebrock, D. Livitz, X. Li, E. Arich-Landkof, N. Shoresh, C. STEWART, A.V. Segre, P.A. Branton, P. Polak, K.G. Ardlie, G. Getz, RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues, Science 364 (2019) (eaaw0726).

[46] E.D. Saad, A. Katz, Progression-free survival and time to progression as primary end points in advanced breast cancer: often used, sometimes loosely defined, Ann. Oncol. 20 (3) (*2019*) 460–464.

[47] K.L. Lloyd, I.A. Cree, R.S. Savage, Prediction of resistance to chemotherapy in ovarian cancer: a systematic review, BMC Cancer 15 (2015) 117.

[48] L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, Proc. Natl. Acad. Sci. 103 (2006) 5923–5928.

[49] R. Nilsson, J. Björkegren, J. Tegnér, On reliable discovery of molecular signatures, BMC Bioinform. 10 (2009) 38.

[50] S.C. Smith, A.S. Baras, J.K. Lee, D. Theodorescu, The COXEN principle: translating signatures of *In vitro* chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer, Cancer Res. 70 (2009) 1753–1758.

[51] H. Varmus, The transformation of oncology, Science 352 (2016) 123.