MOLECULAR ONCOLOGY XXX (2015) 1–16

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/molonc

Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning

Stephanie N. Dorman[a], *Katherina Baranova*[a], *Joan H.M. Knoll*[b,c,d], *Brad L. Urquhart*[e], *Gabriella Mariani*[f], *Maria Luisa Carcangiu*[g], *Peter K. Rogan*[a,d,h,i,*]

[a]*Department of Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada*
[b]*Department of Pathology and Laboratory Medicine, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada*
[c]*Molecular Diagnostics Division, Laboratory Medicine Program, London Health Sciences Centre, ON, Canada*
[d]*Cytognomix Inc., London, ON, Canada*
[e]*Department of Physiology and Pharmacology, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada*
[f]*Department of Medical Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
[g]*Department of Diagnostic and Laboratory Pathology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*
[h]*Department of Computer Science, University of Western Ontario, London, ON, Canada*
[i]*Department of Oncology, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada*

ARTICLE INFO

Article history:
Received 20 July 2015
Accepted 31 July 2015
Available online ■

Keywords:
Paclitaxel
Gemcitabine
Resistance
Drug sensitivity
Genomic profiles
Breast cancer

ABSTRACT

Increasingly, the effectiveness of adjuvant chemotherapy agents for breast cancer has been related to changes in the genomic profile of tumors. We investigated correspondence between growth inhibitory concentrations of paclitaxel and gemcitabine (GI50) and gene copy number, mutation, and expression first in breast cancer cell lines and then in patients. Genes encoding direct targets of these drugs, metabolizing enzymes, transporters, and those previously associated with chemoresistance to paclitaxel (n = 31 genes) or gemcitabine (n = 18) were analyzed. A multi-factorial, principal component analysis (MFA) indicated expression was the strongest indicator of sensitivity for paclitaxel, and copy number and expression were informative for gemcitabine. The factors were combined using support vector machines (SVM). Expression of 15 genes (*ABCC10*, *BCL2*, *BCL2L1*, *BIRC5*, *BMF*, *FGF2*, *FN1*, *MAP4*, *MAPT*, *NKFB2*, *SLCO1B3*, *TLR6*, *TMEM243*, *TWIST1*, and *CSAG2*) predicted cell line sensitivity to paclitaxel with 82% accuracy. Copy number profiles of 3 genes (*ABCC10*, *NT5C*, *TYMS*) together with expression of 7 genes (*ABCB1*, *ABCC10*, *CMPK1*, *DCTD*, *NME1*, *RRM1*, *RRM2B*), predicted gemcitabine response with 85% accuracy. Expression and copy number studies of two independent sets of patients with known responses were then analyzed with these models. These included tumor blocks from 21 patients that were treated with both paclitaxel and gemcitabine, and 319 patients on paclitaxel and anthracycline therapy. A new paclitaxel SVM was derived from an 11-gene subset since data for 4 of the original genes was

* *Corresponding author.*
E-mail address: progan@uwo.ca (P.K. Rogan).
http://dx.doi.org/10.1016/j.molonc.2015.07.006

unavailable. The accuracy of this SVM was similar in cell lines and tumor blocks (70–71%). The gemcitabine SVM exhibited 62% prediction accuracy for the tumor blocks due to the presence of samples with poor nucleic acid integrity. Nevertheless, the paclitaxel SVM predicted sensitivity in 84% of patients with no or minimal residual disease.

1. Introduction

Chemotherapeutic agents, such as paclitaxel and gemcitabine, are recommended to patients with developed metastases, basal-like breast cancer, and high-risk indications (premenopausal, ER/PR-negative, HER2-status, large tumors, or node-positive) (Cardoso et al., 2012; Oostendorp et al., 2011). There is currently no gold standard chemotherapy regimen (Cardoso et al., 2012; Oostendorp et al., 2011). Treatment selection is suggested to be individualized and should take into account clinical disease characteristics, treatment history, patient-related factors, and patient preference. However, resistance is one of the major barriers to successful therapy. In a recent study, breast cancer patient response rates to paclitaxel and gemcitabine after 6 cycles of chemotherapy were found to be only 50.0% and 78.6% respectively (Lee et al., 2014). This has motivated a number of groups to develop gene signatures aimed at predicting therapeutic response to these drugs in breast cancer patients (Gąsowska-Bodnar et al., 2014; Hatzis et al., 2011; He et al., 2014).

As in breast cancer patients, breast cancer cell lines show variable responses to growth inhibition by paclitaxel and gemcitabine (Daemen et al., 2013; Shoemaker, 2006). Cell lines mirror many of the pathological features of breast tumors, such as the intrinsic subtypes of breast cancer (Neve et al., 2006; Prat et al., 2013), and can be useful for testing anticancer therapy responses (Heiser et al., 2012). Daemen et al. (2013) employed random forest machine learning to assess genomic information from 70 breast cancer cell lines (including DNA sequence, gene copy number, gene expression, promoter methylation, protein expression, and the corresponding cell line response to 90 anti-cancer compounds) with the objective of establishing pretreatment signatures that predict response. The gene expression profile of the tumor subtype was found to be the most effective way to model response to therapy. However, many molecular signatures derived using genome-wide approaches are inconsistent between different data sets (Ein-Dor et al., 2006; Nilsson et al., 2009). This is partly due to the fact that deriving predictive gene models using thousands of genes risks overtraining, that is, fitting the noise rather than the actual gene signature in the data (Ein-Dor et al., 2006).

We recently defined a set of genes that are stable in gene expression and copy number in the majority (>90%) of breast cancer tumors (Park et al., 2012). The stable gene set is composed of genes that are unmutated in the majority of tumors. Interestingly, many stable gene products were found to be targets of paclitaxel and gemcitabine. We examine the possibility that genomic differences in expression, copy number or mutation in these genes may be related to GI50. Rather than a genome-wide approach to predict sensitivity to paclitaxel and gemcitabine (eg. employed by Daemen et al. (2013)), we analyze stable and linked unstable genes in pathways that determine their disposition (Figure 1).

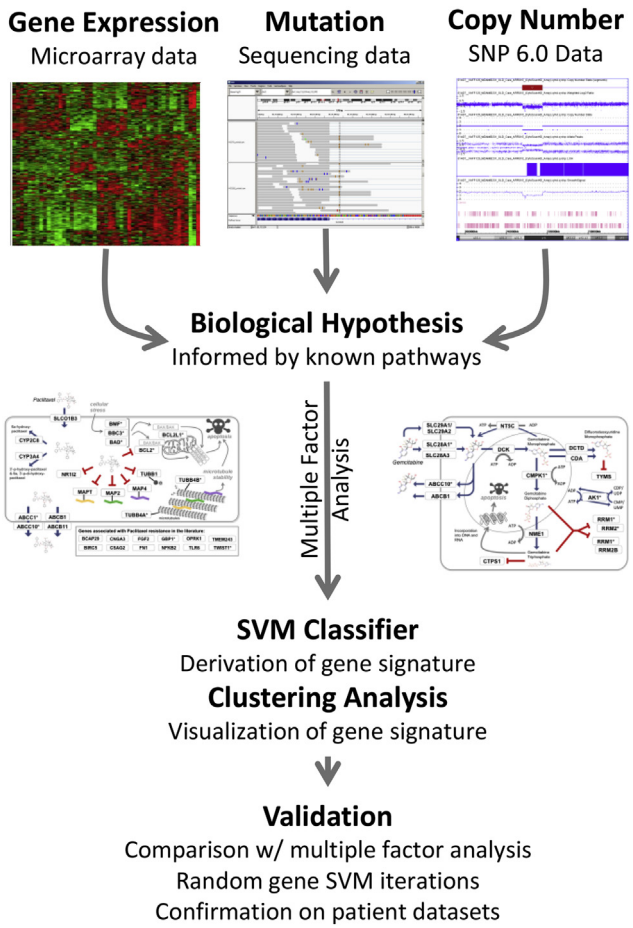

Figure 1 – Workflow to derive gene signatures. Gene sets were derived for paclitaxel and gemcitabine based on known drug pathways, metabolism, and genes previously implicated in resistance. A multiple factor analysis was completed for each gene to determine which data types (gene expression, copy number, and mutation data) were correlated with the growth inhibitory values for paclitaxel and gemcitabine. Gene expression values were used to derive the paclitaxel SVM classifier, and both gene expression and copy number were used for the gemcitabine SVM. Cell lines were then clustered on optimized gene sets to visualize stratification of tumor subtype and sensitivity. The SVM classifiers were validated using random gene iterations to determine the significance of the classification accuracy, and patient data sets to ensure robustness of the models derived.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

Gene panels were established based on biological and experimental studies of paclitaxel and gemcitabine metabolism. Paclitaxel binds to the β subunit of tubulin (*TUBB1*), inhibiting microtubule formation during mitosis (Jordan and Wilson, 2004). It also binds *BCL2*, which induces programmed cell death (Ferlini et al., 2009). Paclitaxel is now also recognized to target microtubule-associated proteins 2 (*MAP2*), 4 (*MAP4*) and Tau (*MAPT*) (McGrogan et al., 2008), as well as the xenobiotic receptor (*NRI12*, or PXR) (Harmsen et al., 2009). *SLCO1B3* transports paclitaxel into cells (Smith et al., 2005), and it is exported by *ABCB1* (P-glycoprotein), multidrug resistance-associated proteins *ABCC1* (Heijn et al., 1997) and *ABCC10* (Chen et al., 2003), and the bile salt export pump *ABCB11* (Lecureur et al., 2000). Other genes previously implicated as contributing to paclitaxel resistance include *TMEM243* (Duan et al., 2004), *BCAP29* (Rao et al., 2008), *GBP1* (Duan et al., 2006), *TLR6* (Kaczanowska et al., 2013), *NFKB2* (Tantivejkul et al., 2005), *FGF2* (Carmo et al., 2011), *BIRC5* (Lu et al., 2009), *TWIST1* (Hong et al., 2011), *FN1* (Xing et al., 2008), *OPRK1* (Duan et al., 2005), *CSAG2* (Duan et al., 1999), and *CNGA3* (Duan et al., 2005). Additionally, genes expressed in breast tissue involved in paclitaxel metabolism were included: *CYP2C8* and *CYP3A4* (Preissner et al., 2010), as well as stable genes in pathways of known direct targets (Park et al., 2012): *BAD*, *BBC3*, *BCL2L1*, *BMF*, *TUBB4A* (Kavallaris, 2010), and *TUBB4B* (Kavallaris, 2010).

Gemcitabine, a deoxycytidine analog, is transported into the cell by *SLC29A1* (Marcé et al., 2006), *SLC29A2*, *SLC28A1* (Mackey et al., 1999), and *SLC28A3* (Govindarajan et al., 2009). The prodrug is then phosphorylated by *DCK*, *CMPK1*, and *NME1* to gemcitabine diphosphate and triphosphate (Ueno et al., 2007). These active forms are incorporated into DNA, which halts replication and cell growth (Plunkett et al., 1995). Gemcitabine di- and triphosphate target ribonucleotide reductase (*RRM1*, *RRM2*, and *RRM2B*), and inhibit DNA synthesis (Mini et al., 2006). An alternative metabolite, difluorodeoxyuridine monophosphate, which is derived by cytidine deaminase (*CDA*) or dCMP deaminase (*DCTD*), inhibits thymidylate synthetase (*TYMS*), resulting in apoptosis (Ueno et al., 2007).

We examine the hypothesis that genomic differences in genotypes, expression and copy number of these genes explain concentration-dependent growth inhibition by gemcitabine and paclitaxel. We then use machine learning to stratify the relative contributions of different genes to chemoresistance, by identifying corresponding genomic signatures at the transcriptional and genomic level in both cell line and patient data.

2. Materials and methods

2.1. Data acquisition

Growth inhibition (GI50), copy number, gene expression, and exome sequencing data were obtained from the supplementary data of Daemen et al. (2013). GI50s ($-\log_{10}$M, where *M* is the drug concentration required to inhibit cell line growth by 50%) for paclitaxel were available for 49 cell lines and GI50s for gemcitabine were available for 47 cell lines. Supplementary Table 1 indicates the cell lines used and Supplementary Table 2 indicates the gene, gene product names and their respective drug disposition functions. Supplementary Methods I & II describe copy number and variant calling, results of which are shown in Supplementary Tables 3 and 4. Log_2 normalized gene expression data were derived from Affymetrix Gene Chip Human Exon 1.0 ST arrays. Replication studies performed to re-measure and confirm GI50s, verify copy number and mutation data for a subset of the cell lines are outlined in Supplementary Methods III. Figure 1 is an overview of the complete workflow used.

2.2. Cell lines

Cell lines were composed of 10 basal, 9 claudin-low, 25 luminal, and 5 normal-like subtypes. Cell lines were designated resistant, if their GI50 was <8.0 for paclitaxel and <7.0 for gemcitabine, respectively. The threshold values for distinguishing sensitive from resistant cell lines were based on median GI50s for each particular drug (7.99 and 7.13, for paclitaxel and gemcitabine). Daemen et al. (2013) classified cell lines by comparing mean GI50s. We used median GI50, which is not impacted to the same extent by outlier cell lines.

2.3. Multiple factor analysis (MFA)

MFA was used to relate each cell line GI50 according to sets of genomic variables (Abdi and Valentin, 2007). The 44 (gemcitabine) or 45 (paclitaxel) breast cancer cell lines (Supplementary Table 1) were treated as separate individuals. MFA was carried out with the R library "FactoMineR" (Lê et al., 2008), with GI50s, gene expression, copy number, mutation status (if the gene contained 1 or more mutations), and 31 and 18 genes associated with paclitaxel and gemcitabine activity, respectively, as input.

2.4. Support vector classification

A binary support vector machine (SVM) was trained with the Statistics Toolbox in MATLAB (Natick, MA) using fitcsvm (linear kernel function) and then tested with a leave-one-out cross-validation (using 'crossval' and 'leaveout' options). The SVM was trained on the cell lines and explanatory gene variables deemed relevant from the MFA: expression data for the paclitaxel SVM, and copy number and expression data for the gemcitabine SVM. The input data consisted of measurements from all genes used in the MFA. Sequential backward feature selection was performed for feature optimization (Dash and Liu, 1997) to minimize the percentage of misclassified cell lines (classification error) returned from the leave-one-out cross validation (Supplementary Methods IV). Genes that did not reduce or change the classification error were removed from the SVM (one at a time). This procedure was iterated until further gene removal lead to a higher classification error (stopping criterion). The SVM used to analyze patient expression and outcomes of paclitaxel treatment excluded 2 of the 49 cell lines (184A1 and 184B5) from the feature selection process and the subsequent validation. This was done because these lines were not derived from breast cancer tissue, but rather normal tissue that was transformed by exposure to benzo(a) pyrene. By contrast with the SVM, a partial-least squares regression was not effective in relating genomic findings to paclitaxel response (Supplementary Methods V).

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

The hinge loss was also determined for the subset of genes included in the final SVMs. Hinge loss applies a linear penalty for misclassified data according to their distance from the hyperplane. The loss function is represented by Equation (1) where $y_j = \{-1,1\}$ and $f(X_j)$ is the score, i.e. hyperplane distance, for cell line j:

$$L = \max\left(0, 1 - y_j f(X_j)\right) \quad (1)$$

2.5. Applying the cell line SVM to patient data

Formalin fixed, paraffin embedded (FFPE) tumor samples were obtained from the Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy), from leftover material available after diagnostic procedures in consented patients (Musella et al., 2015). Samples obtained were from patients that were first treated with paclitaxel (or in a small number of cases docetaxel) and carboplatin, and then subsequently gemcitabine, upon development of resistance. Clinical information was available as to whether the patients responded to each of the drugs (paclitaxel and gemcitabine). Tumor and control normal tissues were analyzed for expression and copy number of SVM genes, respectively, by real-time reverse-transcriptase polymerase chain reaction (qRT-PCR) and real time PCR (qPCR, methods described in Supplementary Methods, Section VI). The cell line-based SVM models were used to predict patient sample drug responses in a blinded manner. Two SVM models were trained for paclitaxel and gemcitabine: one using the normalized gene expression values, and the other using expression values binned into 10 categories, using the Matlab function: quantile(X,10). Binning was performed because amplifiable RNA template concentration in FFPE blocks is not known precisely, because it is subject to long term degradation and reactivity (Antonov et al., 2005; Fleige and Pfaffl, 2006). Expression measurements were obtained for 11 genes from the paclitaxel SVM, and 6 genes for the gemcitabine SVM. The SVM was trained on the cell line data with these reduced gene sets. Predicted and actual responses were compared, and odds risk ratios (contingency analysis) were calculated (GraphPad Prism, San Diego, California).

Patient data were also obtained from GEO Accession GSE25066, in which expression levels of tumors that were treated with taxane and anthracycline chemotherapy were reported (Hatzis et al., 2011). Expression levels for the paclitaxel SVM genes (except *BMF* and *CSAG2*, which were not measured) were extracted for those patients treated with paclitaxel (n = 319). In cases with multiple probe sets per gene, expression levels were averaged. The SVM predictions were then related to response to therapy and residual cancer burden class for each patient.

2.6. Clustering cell lines and patients using expression values of the SVM gene subsets

The unsupervised, hierarchical clustering function 'clustergram' in Matlab was used to cluster cell lines and patient data (described in 2.5) according to gene expression values included in the optimized SVM. Expression values were normalized by row so the mean expression of each gene across individuals was 0, and the standard deviation was 1. Clustering was performed by individuals and genes, and dendrograms are displayed for each dimension that indicate relatedness based on their lengths and hierachical branching.

3. Results

3.1. Multiple factor analysis

MFAs were performed using GI50s of 49 cell lines, and genomic measurements of 31 and 18 genes related to paclitaxel and gemcitabine activity from an existing data set (Daemen et al., 2013). We re-confirmed measurements for a subset of the cell lines to ensure consistency between cell line sources (see Supplementary Results I). MFAs were assessed by statistics generated by the program, FactoMineR (Lê et al., 2008). Relationships were stratified by the correlation between the variable and GI50, the RV coefficient (a multivariate generalization of the squared Pearson correlation coefficient), the position of variables on the correlation circle, and the representation quality of each variable group in the first two dimensions (cos2 values). These criteria were used to classify each gene as having a "strong relationship", "relationship", "possible relationship" or "no relationship" to GI50 (see Supplementary Table 5 for the thresholds for each class). Examples of correlation circles and individual factor maps for *MAPT* (paclitaxel) and *DCTD* (gemcitabine) are illustrated in Supplementary Results II.

MFA revealed "strong relationships" between paclitaxel GI50 and copy number and/or gene expression for 11 genes, consisting of both negative relationships (diminished copy number and gene expression [-] for *CYP2C8, CYP3A4, NR1I2* (previously known as PXR), *TLR6*, and *TUBB1*) and positive relationships (increased copy number and gene expression [+] for *BBC3, BCL2L1, BMF, CNGA3, MAPT*, and *TUBB4B*) with increased chemoresistance (Supplementary Table 6 lists all MFA measurements). The gemcitabine set revealed strong associations between resistance and *ABCB1* (+), *DCTD* (−), and *SLC28A1* (+) gene expression as well as strong relationships for *ABCC10* (+) and *CDA* (+) copy number (Supplementary Table 7). The MFA results for paclitaxel (gene expression results only) and gemcitabine treatment (copy number and gene expression), in the respective pathway contexts, are summarized in Figure 2.

Point mutation status was based on 74 deleterious coding mutations (Supplementary Table 4) that were predicted to be damaging (Ng and Henikoff, 2003) or to affect mRNA splicing (Shirley et al., 2013; Viner et al., 2014). Point mutations predicted to be damaging demonstrated strong relationships in *ABCB1* (n = 4, in 2 cell lines) to paclitaxel resistance and in *SLC28A3* (n = 3, in 2 cell lines) to increased sensitivity to gemcitabine. The limited number of cell lines with mutations in these genes cannot be effectively incorporated into machine learning models, and point mutation results were not included in these analyses.

3.2. Support vector machine (SVM) learning

A binary SVM was employed to develop a predictive multigene classification of genomic signatures for resistance to these

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

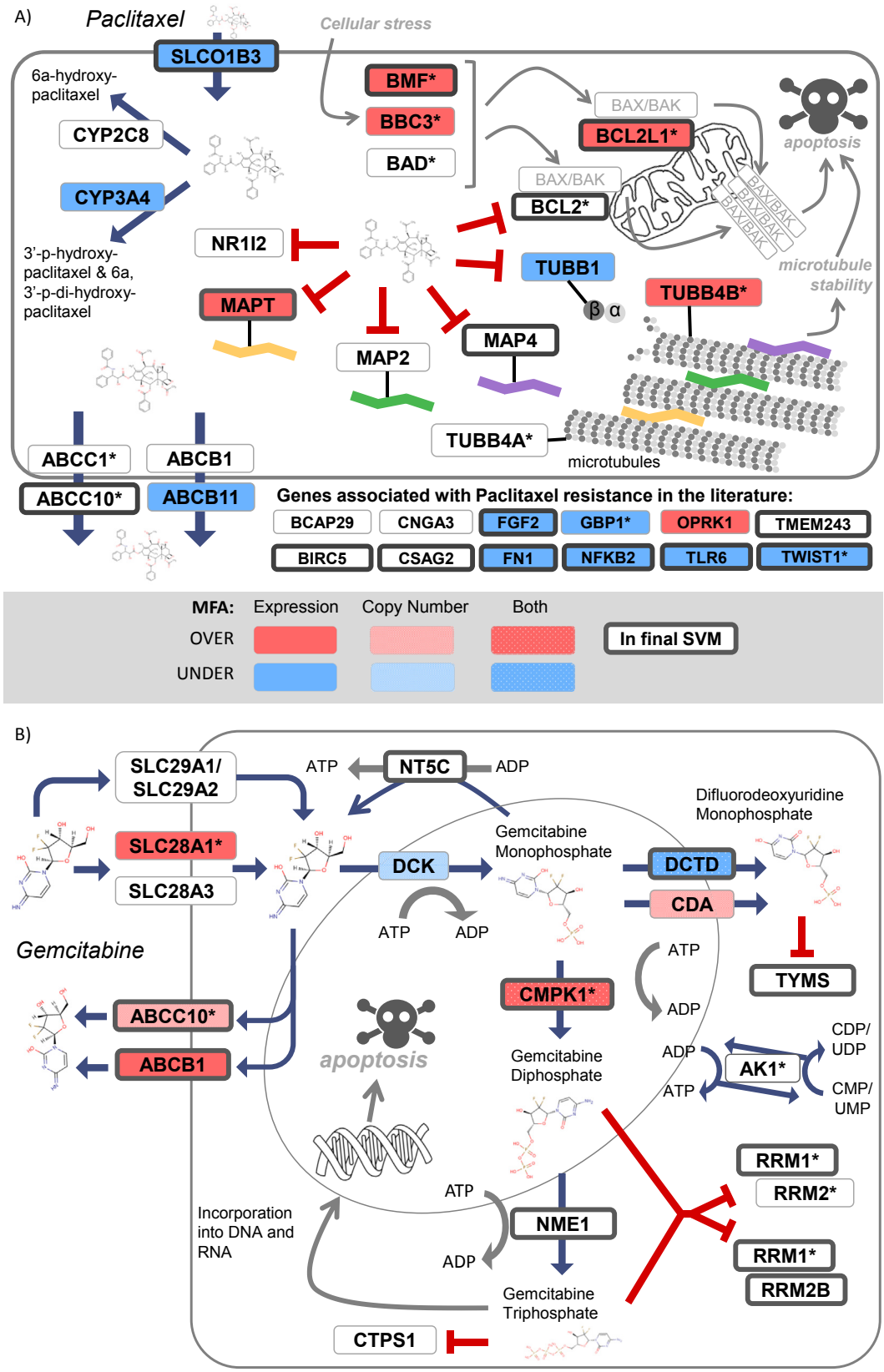

Figure 2 – Genes associated with paclitaxel (A) and gemcitabine (B) mechanism of action (direct targets, metabolizing enzymes), genes previously associated with resistance, and stable genes in the biological pathways targets. Genes with an asterisk (*) are stable genes (Park et al., 2012). Genes highlighted in red showed a positive correlation (within dimension 1 and/or dimension 2) between gene expression or copy number, and resistance in the MFA, whereas genes highlighted in blue demonstrated a negative correlation. Genes outlined in dark grey are those included in the final predictive model that was derived using the SVM. Red T-shaped bars indicate the genes that paclitaxel directly binds/inhibits. Genes outlined in light grey (ie. BAX/BAK) were not included in the analysis because they were not stable genes in the *BCL2* pathway.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

drugs (Furey et al., 2000). Based on MFA results, data types orthogonal to GI50 were excluded from the SVM (see section 2.4 for details). The classification error of the SVM model was minimized by removing genes, i.e. features, which did not improve accuracy by leave-one-out cross-validation. This feature selection process is illustrated in Figure 3-I. The optimized SVM was then trained, respectively, on 15 gene variables for paclitaxel (49 cell lines) and 10 variables for gemcitabine (44 cell lines). Gene expression values from *ABCC10*, *BCL2*, *BCL2L1*, *BIRC5*, *BMF*, *FGF2*, *FN1*, *MAP4*, *MAPT*, *NKFB2*, *SLCO1B3*, *TLR6*, *TMEM243*, *TWIST1*, and *CSAG2* comprised the final set of features used to train the SVM for classification of paclitaxel sensitivity. For gemcitabine, both gene expression values (from *ABCB1*, *ABCC10*, *CMPK1*, *DCTD*, *NME1*, *RRM1*, *RRM2B*) and copy number data (from *ABCC10*, *NT5C*, *TYMS*) were used in the final SVM. The distance of each cell line value from the SVM hyperplane that distinguishes the degree of sensitivity or resistance was plotted against the corresponding GI50 (Supplementary Results III). The trained SVMs misclassified 9 of 49 (18%) cell lines for paclitaxel and 7 of 44 (16%) for gemcitabine, which is comparable to or more accurate than other approaches (Ma et al., 2004, demonstrated 19% misclassification for tamoxifen). Partitioning by histological subtype did not improve the classification accuracy; a single variable SVM model based on subtype misclassified 30% of cell lines for paclitaxel and 45% for gemcitabine (Supplementary Table 8). The feature-optimized SVM outperformed the signature derived from the initial set of genes, which misclassified resistance/sensitivity of 36% cell lines for paclitaxel and 64% for gemcitabine treatments. In addition, multi-gene MFA analyses of the final SVM gene sets demonstrate that the individual factor maps of the resistant and sensitive cell lines segregate to a greater degree than MFAs based on the initial gene sets, which were indistinguishable (Supplementary Results IV). These differences were larger for gemcitabine than paclitaxel.

To assess the individual impacts of a gene on SVM accuracy, each gene remaining in the optimized SVM was removed, and the misclassification rate was redetermined

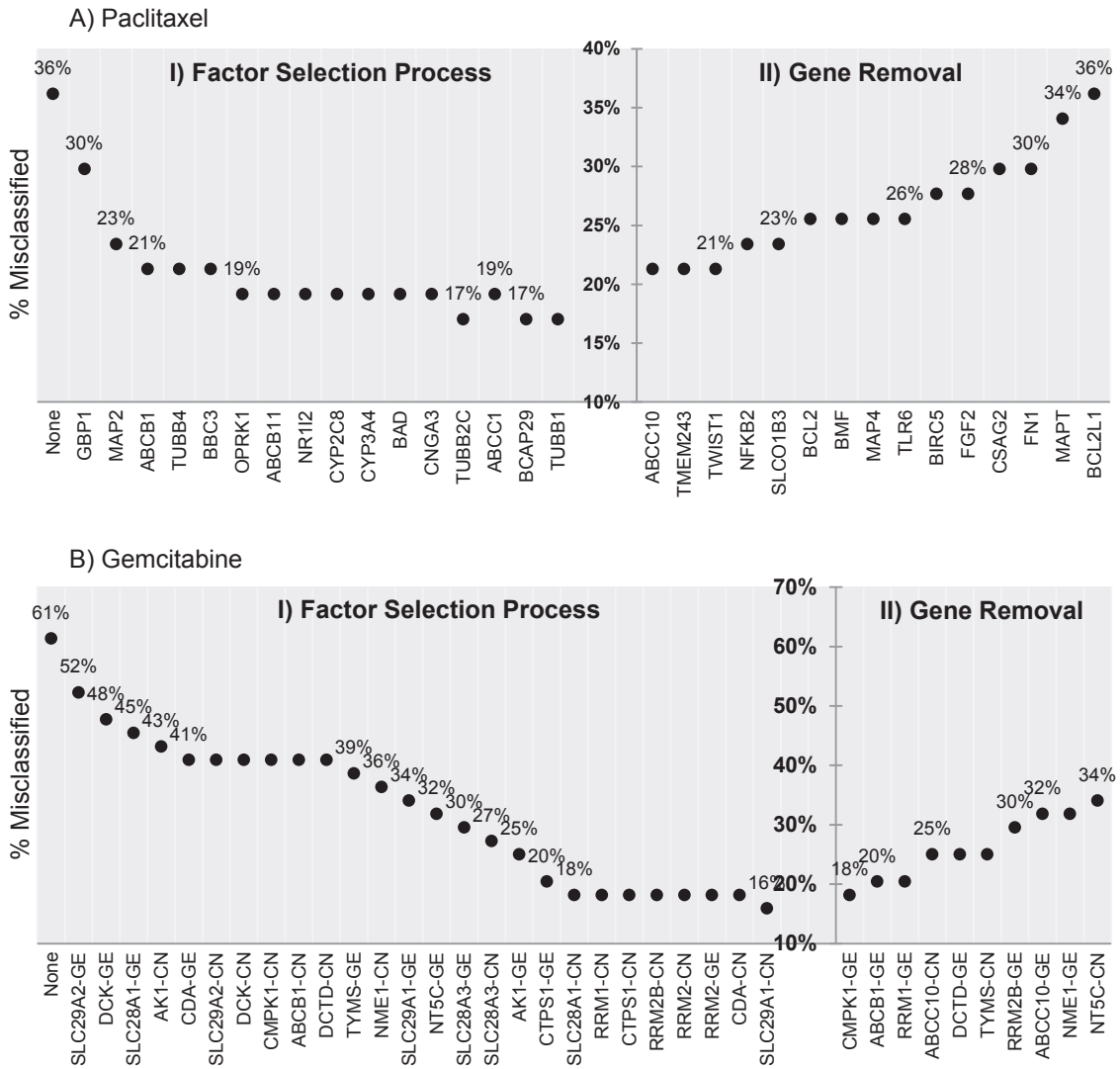

Figure 3 – Effect of the removal of each gene on the percent of cell lines misclassified during the SVM feature selection process to determine the most predictive gene set (left panels AI and BI). The right panels (AII and BII) demonstrate the increase in the percent of cell lines misclassified when the expression of genes in the inferred, optimal gene set are subsequently eliminated from the SVM.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

(Figure 3A-II). *BCL2L1* and *MAPT* had the highest predictive value for paclitaxel sensitivity, with misclassification rates of 36% and 34%, respectively, when eliminated (compared to 21–30% for the other genes). It is notable that the MFA also showed strong associations with decreasing *MAPT* or *BLC2L1* expression and increasing paclitaxel sensitivity. *BCL2L1* is a member of the Bcl-2 family and is involved in regulation of apoptosis (Ferlini et al., 2009). Additional apoptotic regulators, such as *BMF* and *BCL2*, also appear in our SVM results, as paclitaxel is known to trigger apoptosis through these pathways (Kutuk and Letai, 2008). The loss of *MAPT* in breast cancer cells has been shown to sensitize those cells to the action of paclitaxel (Bhat and Setaluri, 2007), which is supported by our analysis.

For gemcitabine, removing *NT5C* copy number, *NME1* gene expression, *ABCC10* gene expression, and *RRM2B* gene expression had the largest effects, by respectively increasing misclassification rates to 34%, 32%, 32%, and 30% (Figure 3B-II). *NT5C* is located on 17q25.1 a region associated with cancer (Fukino et al., 1999). Allelic imbalances in *TYMS* have previously been hypothesized to be involved in drug resistance in renal cell carcinoma (Colavito et al., 2009) and *ABCC10* has been associated with drug resistance (Hopper-Borge et al., 2009). *NME1* is a known metastasis suppressor gene which may have great prognostic value (Shoushtari et al., 2011). *RRM2B* and *RRM1* have been suggested to be associated with gemcitabine resistance (Aye et al., 2014) and have been shown to be overexpressed in a gemcitabine-resistant pancreatic cancer cell line (Wang et al., 2015).

3.3. *Applying the cell line-trained SVM to patient data*

Formalin fixed paraffin embedded (FFPE) tissue blocks were obtained from patients that were treated with paclitaxel and gemcitabine, and whose response to both drugs are known. Gene expression measurements for 11 genes from the paclitaxel SVM, and gene expression (6 genes) and copy number (CN; 3 genes) from the gemcitabine SVM were obtained using qRT-PCR and qPCR (Supplementary Tables 9 and 10). Gene expression measurements were not obtained for *BMF*, *CSAG2*, *SLCO1B3*, *TWIST1* (paclitaxel), and *ABCB1* (gemcitabine), as no amplification was observed in these samples by 40 cycles. The absence of amplification in these genes was related their low levels of expression in breast cancer tissue (Supplementary Results V – A). In cases where qRT-PCR showed no amplification for a specific sample out of the genes measured, the highest cycle run was used as the Ct value for that gene. Older samples, on average, had lower numbers of genes with successful measurements (Supplementary Results V – B).

SVMs were trained using the cell line data with a reduced set of 11 (paclitaxel – *ABCC10, BCL2, BCL2L1, BIRC5, FGF2, FN1, MAP4, MAPT, NFKB2, TLR6,* and *TMEM243*) and 9 (gemcitabine – *ABCC10, CMPK1, DCTD, NME1, RRM1, RRM2B, ABCC10*-CN, *NT5C*-CN, and *TYSM*-CN) gene values, which corresponded to the measurements obtained from the FFPE tissue block studies. These SVMs were then applied to the FFPE tissue sample data to predict their sensitivity to paclitaxel and gemcitabine (see Supplementary Table 11 for full FFPE sample predictions). The paclitaxel SVM predicted drug sensitivity with 71% accuracy (Table 1), which was similar to a leave-one-out analysis on the cell line data, which classified cell lines with 70.2% accuracy (using the reduced 11-gene subset). Patients who were treated with docetaxel were excluded from this summary because the trained SVM only predicted cell line response to docetaxel with 57% accuracy (misclassified 19/44, based on GI50s). Docetaxel and paclitaxel GI50s for all cell lines were correlated only to a limited extent ($R^2 = 0.722$), consistent with the possibility that there might potentially be differences in mechanisms of drug metabolism and resistance between these drugs. The gemcitabine SVM did not perform as well on the patient sample data as it did on the cell line leave-one-out analysis, which was 79.6% accurate (using the reduced 9-gene subset). The gemcitabine SVM derived using binned expression values predicted patient response with 62% accuracy, however, 72% accuracy was achieved for samples with gene expression measurements available for at least 4 of the 6 genes.

Gene expression measurements and clinical data were also obtained for 319 patient samples who were treated with paclitaxel and anthracycline chemotherapy (Hatzis et al., 2011). Gene expression data was not available for two genes from the paclitaxel SVM (*BMF* and *CSAG2*), which were two of the 4 genes that could not be measured in the FFPE samples. Consequently, the same 11-genes used for the FFPE samples were applied to the data from Hatzis et al., 2011. SVM predictions were compared with the clinical outcome – whether the patient had recurrent disease (RD) or complete pathological response (pCR, see Table 2 for a summary, and Supplementary Table 12 for all predictions). The SVM predicted sensitivity in 52 of the 63 patients (84%) that showed pCR. All patients that showed complete pathological response exhibited no or minimal residual disease (residual cancer burden [RCB] class 0/1; Symmans et al., 2007), although some patients within this subset did not respond to therapy. This group of patients (RCB 0/1) may derive the greatest benefit from the paclitaxel SVM analysis. The SVM did not perform as well in predicting resistance, miscategorizing 135 patients of the 257 with RD (52.5%) as sensitive. However, performance of the SVM exceeded that of the 512-gene signature

Table 1 – Using the SVM to predict patient response from archived FFPE tissue.

	Paclitaxel		Gemcitabine	
	NORM	10 bins	NORM	10 bins
No. of accurate predictions	12	12	9	13
Total	17[a]	17	21	21
Percent accurate	71%	71%	43%	62%
Odds ratio	5.83	6.00	3.00	3.33
P-value[b]	0.1534	0.1534	0.5333	0.3615

a 4 patients were treated with docetaxel instead of paclitaxel, and were not included in this summary.
b Fisher's exact test. Gene expression values were either normalized (NORM) or binned into 10 categories (10 bins), as described in the methods. Please refer to Supplementary Table 11 for all FFPE clinical response and prediction data.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

described in Hatzis et al., 2011 for both sensitive and resistant patients. The odds ratio of the 11-gene SVM was 4.484 (Fisher's exact test, $p < 0.0001$), compared to the odds ratio of 3.181 of the predictive signature described in that study (Fisher's exact test, $p < 0.0001$).

3.4. *Clustering cell line and patient data based on SVM gene subsets*

Two distinct groups emerge from unsupervised clustering using the SVM gene set for paclitaxel in the cell line data (Figure 4A). The left cluster (highlighted in light grey) corresponds with the luminal subtype, and the right corresponds to a mix of basal, claudin-low, and normal-like subtypes. The proportions of resistant (71% of the left cluster) vs. sensitive (58% of the right cluster) cell lines are not statistically significant ($\chi^2 = 3.67$, 1 degree of freedom, $p = 0.056$). Cell lines clustered using the gemcitabine SVM gene expression values display at least two distinct clusters that do not correspond to any subtype(s), but, stratify according to gemcitabine sensitivity (73%; left) or resistance (69%; right) (Figure 4B, chi-statistic = 10.75, $p = 0.001$, d.f. = 1). Clustering of the FFPE derived samples was not as strong as a consequence of limited sample numbers and lack of expression measurements for every gene in every sample (Supplementary Results VI – A). Nevertheless, clustering of expression in these samples mirrored the cell line data based on results for *MAPT* and *BCL2* (for paclitaxel) and *DCTD* (for gemcitabine).

Unsupervised clustering of expression data from Hatzis et al. (2011), using the paclitaxel SVM gene set distinguished patients according to the proportions of those free of distant relapse (Figure 5 and Supplementary Results VI – B). These clusters are partially distinguished by *MAPT* and *BCL2* expression (Figure 5A, the "low *MAPT*" cluster is indicated in purple, "high *MAPT*" in green). *MAPT* and *BCL2* are both components of the PAM50 Breast Cancer Intrinsic Classifier. Their expression patterns segregate into luminal and basal subtypes to a large extent. Low *MAPT* expressing luminal subtypes were observed to have significantly worse prognoses than higher *MAPT* expressing luminal tumors in the patient dataset ($p < 0.05$, Supplementary Results VII). The gene signature described by Hatzis et al. (2011) predicted treatment "sensitivity" and "insensitivity" accurately within the low *MAPT* cluster, where "sensitive" patients exhibit significantly longer times to distant relapse (Figure 5C, $p = 0.0013$, log rank test). However, this was not the case for the high *MAPT* cluster, as the proportion free of distant relapse between two predicted groups did not differ significantly ($p = 0.10$, log-rank test).

Table 2 – SVM predictions on 319 patients treated with paclitaxel from Hatzis et al. (2011).

	Cell line 11-gene SVM		Hatzis "Rx" prediction	
	RD	pCR	RD	pCR
All RCB classes				
Predicted insensitive	119	10	186	28
Predicted sensitive	138	52	71	34
Odds ratio	4.484		3.181	
P-value[a]	<0.0001		<0.0001	
RCB class 0/1 only				
Predicted insensitive	11	10	10	28
Predicted sensitive	19	52	20	34
Odds ratio	3.011		0.6071	
P-value[a]	0.0359		0.3673	

RD = recurrent disease (designated "insensitive" patient response), pCR = pathological complete response (designated as "sensitive" patient response), RCB = residual class burden (as described in Symmans et al., 2007).
a p-values were determined using a Fisher's exact test. Please refer to Supplementary Table 12 for all predictions and patient information.

3.5. *Significance of SVM classification accuracy*

To assess the significance of the derived SVM, we selected 100,000 random sets of 15 genes from a set of expression values (to compare to the paclitaxel SVM) and 10 genes from a set of copy number and expression values (gemcitabine SVM) for 23,030 genes. Only 0.14% of paclitaxel and 0.01% of gemcitabine random gene combinations exceeded the classification accuracy of the derived SVMs. (Figure 6 – A/C). The hinge loss, which increases based on the misclassified object's distance to the hyperplane, was 0.64 for the paclitaxel SVM and 0.66 for the gemcitabine SVM (optimal is close to zero). Among the random gene combinations, the likelihood of deriving SVMs with equal or lower scores was 1.45% and 0.83% for paclitaxel and gemcitabine, respectively (Figure 6 – B/D). Thus, the accuracy of the SVMs achieved for both drugs were not likely due to random chance ($p < 0.05$ in all cases, Table 3).

Nearly all of the high performance random gene set combinations appear to be statistical artifacts. Analysis of 10,000 random gene selections found 18 combinations with lower paclitaxel misclassification response rates. All 18 signatures were unique (2 transcripts occurred twice) and transcript combinations were dominated (24%) by alternative splice variants and expressed pseudogenes. None of the random gene combinations were significantly associated with known biological pathways. Six of the random signatures contained ≥10 gene expression values in the patient data. None of these signatures predicted paclitaxel sensitivity, except one set containing *WWP1*, which has previously been suggested to be a prognostic indicator in breast cancer (Nguyen Huu et al., 2008). This signature (and one based on *WWP1* expression alone) predicted more patients to be sensitive to paclitaxel (Hatzis et al., 2011) than our derived SVM. Similar numbers of patients predicted to be sensitive by both SVM models exhibited complete remission (52 vs. 55), however the *WWP1*-based SVM predicted sensitivity in a greater number of non-responders (n = 178) than our derived SVM (n = 138) and misclassified 41% of the cell lines. For the gemcitabine response, the SVM of a single random gene set had a lower misclassification rate than our derived SVM. The genes in this set were unrelated to gemcitabine metabolism, with 9 of 10 SVM variables exhibiting copy number changes, two of which involved non-coding RNA genes.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

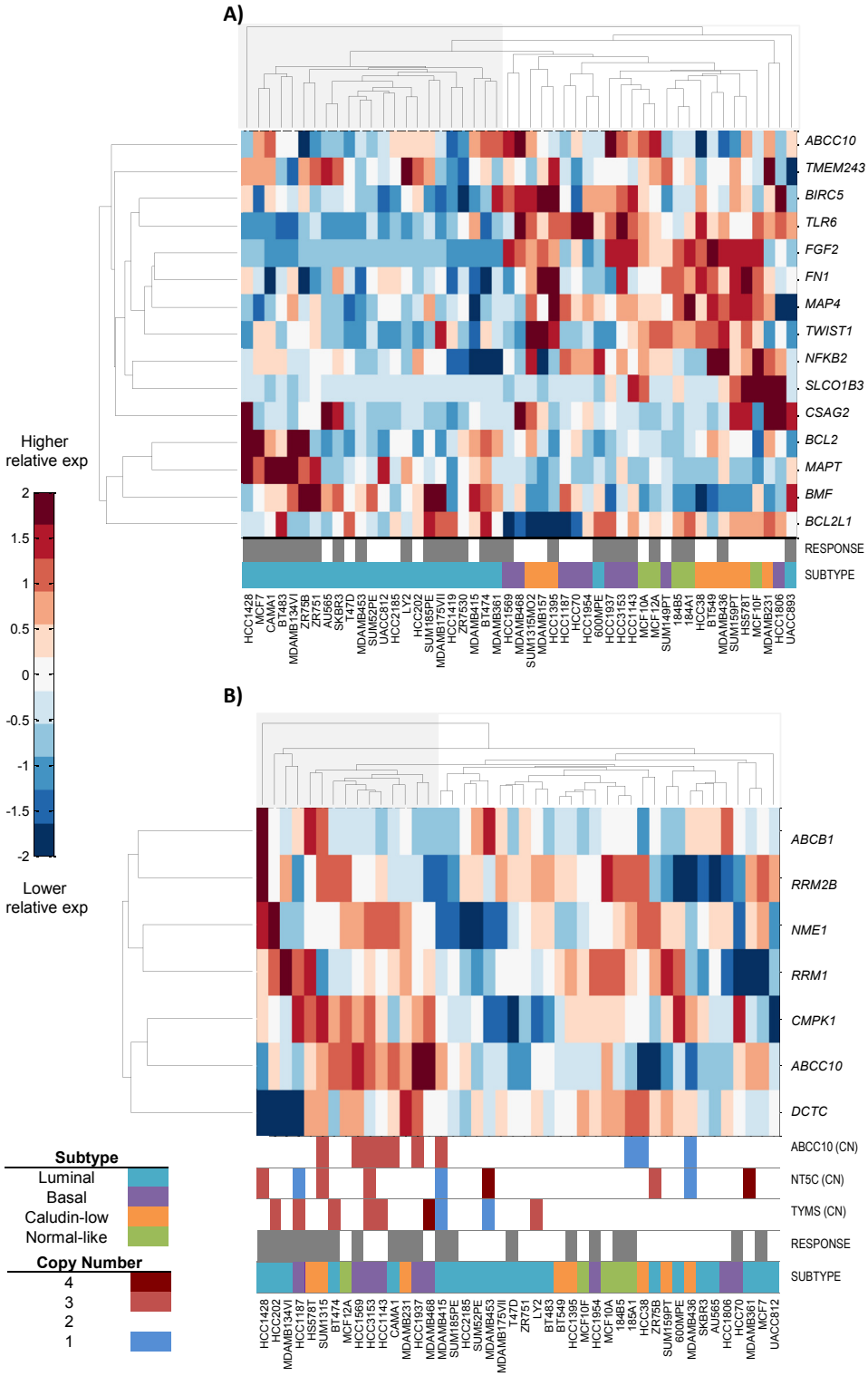

Figure 4 – Expression heatmap of the paclitaxel and gemcitabine SVM derived genes for the tested cell lines. Each row represents a gene and each column a cell line. Red indicates higher expression and blue represents lower expression, as shown by the color bar on the left. 'Resistant' cell lines are colored grey and 'sensitive' cell lines are colored white in the row labeled 'response'. Cell lines are labeled by subtype and copy number according to the legends. Clustering was done based on the similarity of each cell line's expression profile in the 1st (column) dimension and each gene's expression profile in the 2nd (row) dimension. The dendrograms on the top and left indicate the relatedness of each cell line and gene by the length and subdivision of the branches, with deeper branches indicating a stronger relationship and branches in the same 'tree' being more closely related to each other than data in other 'trees'. A) A section of the dendrogram for paclitaxel is shaded grey to indicate a cluster composed entirely of luminal cell lines and a higher proportion of resistant cell lines. The other section is white to indicate a cluster with very few luminal cell lines and a higher proportion of sensitive cell lines. B) A section of the dendrogram for gemcitabine is shaded grey to indicate a cluster composed of a higher proportion of resistant cell lines. The other section is white to indicate a cluster with a higher proportion of sensitive cell lines.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

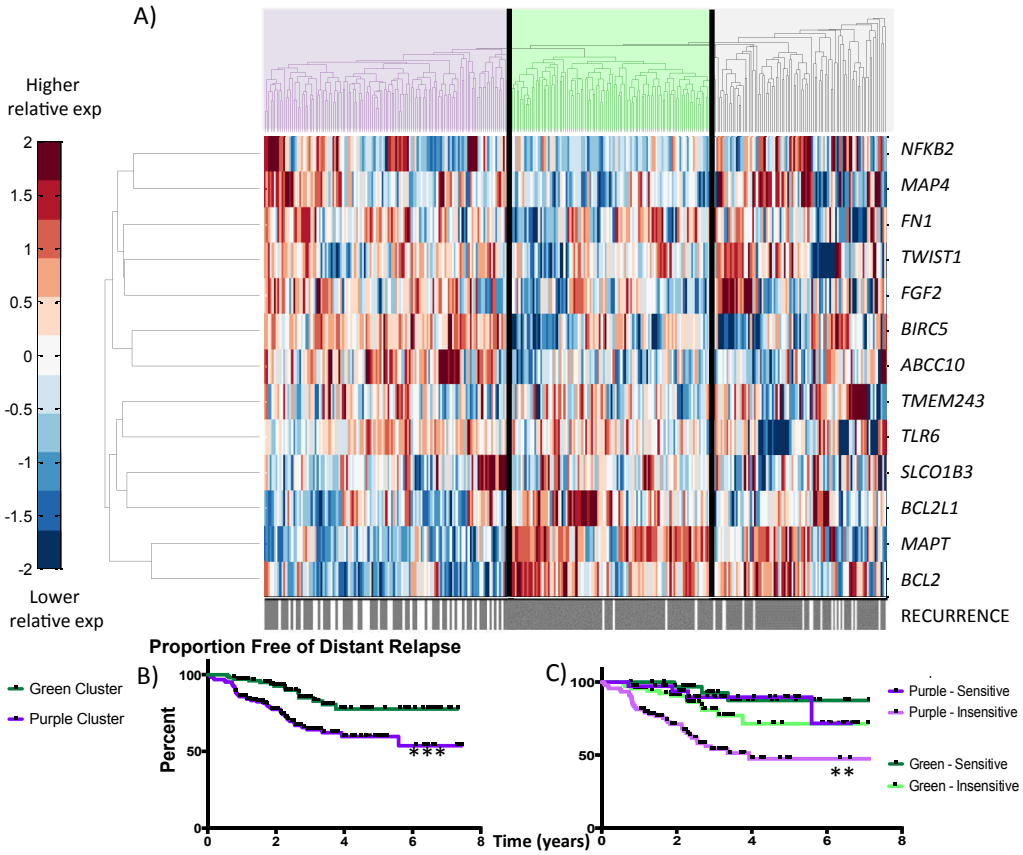

Figure 5 – A) Expression heatmap of the paclitaxel SVM derived genes for 319 tumor samples (Hatzis et al., 2011). See Figure 4 for heat map labeling and diagram details. A section of the dendrogram on the top is shaded purple to indicate a cluster of tumors (63% basal) with a significantly worse outcome assessed by the proportion free of distant relapse curves (shown in B). Another section is shaded green (83% luminal) with significantly better outcomes. The cluster shaded gray (22% basal, 53% luminal) can be clustered independently with similar stratification by subtype and outcome (Supplemental Information VI). C) The Hatzis et al., 2011 gene signature performs very well in the purple cluster and poorly in the green, based on the Kaplan–Meier curves constructed on each subset using their published labels ("insensitive" and "sensitive").

We have developed a online web-based SVM calculator (http://chemotherapy.cytognomix.com) to predict chemosensitivity from normalized gene expression and copy number (gemcitabine) values. The user may either input their own values or populate them from available patient or cell line results that determined the accuracy of the SVM.

3.6. *Translation of signature to other cancer types*

To mitigate tissue-specific effects, we rederived SVM models specific to lung cancer (lung) and hematopoietic and lymphoid tissue cancer (hematopoietic) cell lines using expression data from the broad institute (www.broadinstitute.org/ccle/home; "CCLE_Expression_2012-09-29.res" and "CCLE_NP24.2009_profiling_2012.02.20.csv"). Lung and hematopoietic tissue types were chosen because they contained the highest number of cell lines with expression and paclitaxel GI50s. The final lung SVM contained 14 genes, and classified cell lines with 72% accuracy (Supplementary Results VIII – A). The final hematopoietic SVM was composed of 8 genes, and classified cell lines with 75% accuracy (Supplementary Results VIII – B). Four genes were present in all three (breast, lung and hematopoietic) cancer cell line SVMs (*BMF*, *FGF2*, *TMEM243*, and *TWIST1*), and 8 genes were eliminated from all of the SVMs (*ABCB11*, *BBC3*, *CNGA3*, *CYP2C8*, *CYP3A4*, *NR1I2*, *TUBB4A*, and *TUBB4B*; Supplementary Results VIII – C). MFAs using the Lung and Hematopoietic SVM gene sets do not show the same degree of segregation between resistant and sensitive cell lines as the breast SVM (Supplementary Results VIII – D/E).

4. Discussion

This paper describes the development of genomic signatures using support vector machines that can predict breast cancer tumor response to paclitaxel and gemcitabine. We used a biologically-driven approach to identify a meaningful group of genes whose expression levels and copy number may be useful in guiding selection of specific chemotherapy agents during patient treatment. Previous studies have derived associations between the genomic status of one or more genes and tumor response to certain therapies (Duan et al., 2003; Glinsky et al., 2005; Hatzis et al., 2011; Ma et al., 2004; Rajput et al., 2013; van't Veer et al., 2002). Correlations between single gene expression and tumor resistance (Duan et al., 2003,

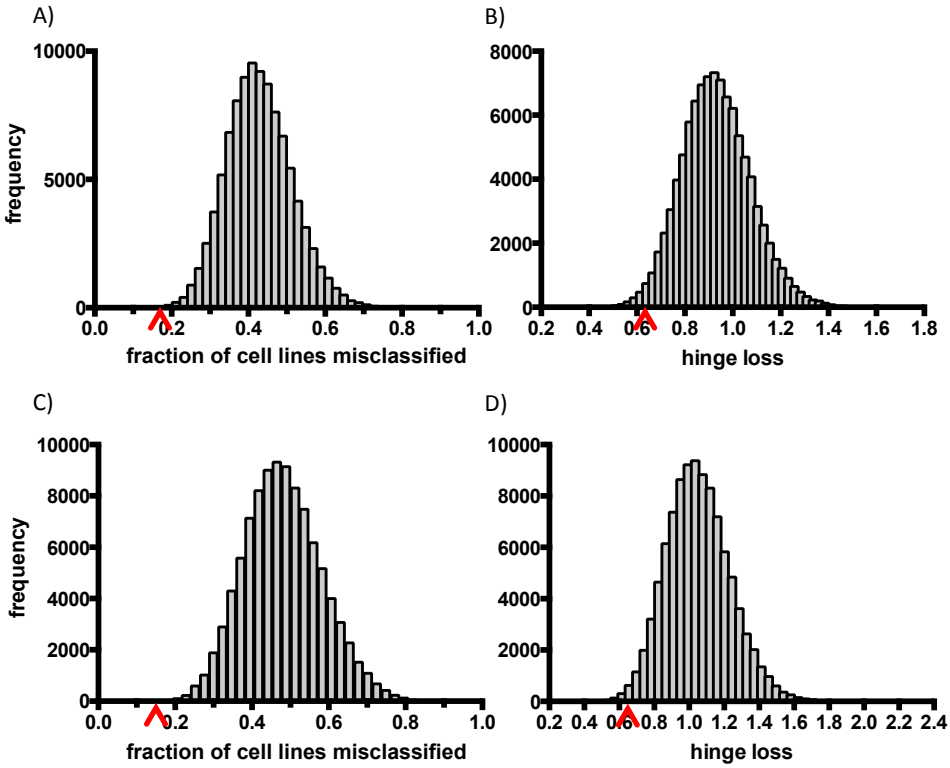

Figure 6 — The fraction of cell lines misclassified (A/C) and hinge loss scores (B/D) were measured on SVMs derived using randomly selected gene sets. 15-gene (to compare to the paclitaxel SVM, A/B) or 10-gene (to compare to the gemcitabine SVM, C/D) values were randomly selected from an initial set of 23,030 genes and used to derive SVMs. The performance of 100,000 iterations of the random signatures are plotted in the above histograms. The hinge loss scores for the paclitaxel and gemcitabine final SVM gene subsets lie in the lowest 2nd (paclitaxel, z-score −2.0, $p < 0.05$ one-sided) and 1st percentiles (gemcitabine, $z = -2.16$, $p < 0.05$) of the data. Expression alone was used for the 15-gene sets (A/B). Copy number and expression were used for the 10-gene sets (C/D). The red arrow heads indicates where the optimized paclitaxel and gemcitabine SVM gene signatures are found in the distribution.

1999) do not take into account multiple mechanisms of resistance or assess interactions between multiple genes. ABC transporter overexpression has long been shown to confer resistance, but enzymatic or functional inhibition has not substantially improve patient response to chemotherapy (Samuels et al., 1997).

Multi-gene analytical approaches have previously been successful in deriving prognostic gene signatures for metastatic risk stratification (Oncotype DX™, MammaPrint®), subtypes (PAM50), and efforts have been made to predict chemotherapy resistance (Hess et al., 2006; Hatzis et al., 2011). Given the complexity of genomic changes and the fundamental

Table 3 — SVM performance using randomly selected genes based off 100,000 iterations.

	Minimum	Maximum	Average	Standard deviation	Drug SVM	z-score	p-value	No. random SVM ≤ drug SVMs[a]
Percent misclassification of cell lines in leave-one-out analysis								
15-gene[b]	12.2%	83.7%	42.7%	8.8%	18.4%	−2.78	0.0027	141
10-gene[c]	12.2%	90.2%	48.0%	10.5%	15.9%	−3.06	0.0011	10
Hinge loss score								
15-gene	0.39	1.66	0.93	0.14	0.64	−2.04	0.0207	1453
10-gene	0.30	2.02	1.05	0.18	0.66	−2.16	0.0153	826

Misclassification rates and hinge loss scores were determined from SVMs derived using 100,000 random combinations of gene expression and copy number values from 23,0303 genes. The minimum, maximum, average, and standard deviations of each 100,000 iterations were determined, and compared to the pacliataxel and gemcitabine SVMs ("drug SVM").

a The number of random gene combinations with equal or lower misclassification rates or hinge loss scores compared to the drug SVMs.

b Random selection of 15 gene expression values were compared to the paclitaxel SVM.

c Random selection of 10 gene expression or copy number values were compared to the gemcitabine SVM.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

biological differences among the intrinsic subtypes of breast cancer (Cancer Genome Atlas Network, 2012; Dorman et al., 2014), this approach has advantages over analysis of isolated genes. Reasonable gene signatures associated with breast cancer outcome can be obtained by chance alone (Venet et al., 2011), however our results show that such signatures are especially rare. Gene signatures derived without reference to the underlying mechanisms of chemotherapy response do not capture meaningful biological results (Drier and Domany 2011).

Our approach started with a focused biologically-relevant initial gene set, rather than taking a genome-wide approach. The derived signatures were demonstrated to significantly outperform random selected combinations of genes in prediction of sensitivity and resistance. The random gene sets may be statistical artifacts, as they were not enriched for any biological relevant pathways, and included expressed pseudogenes. The compositions of these other gene sets were distinct from the set used to derive the SVM and another 20-gene signature for taxane sensitivity (He et al., 2014).

Our analysis highlights the importance of the expression of genes encoding microtubule-associated proteins and apoptotic regulators in paclitaxel resistance (McGrogan et al., 2008; Tanaka et al., 2009; Wang et al., 2013). *MAPT* expression was significantly correlated with drug resistance, and both *MAPT* and *MAP4* were components of the optimized paclitaxel SVM gene set. In clustering analysis of both cell lines and patients, *MAPT* was differentially expressed between tumor clusters stratified by subtype and outcome (Figures 4 and 5). Our results confirm that apoptosis-related proteins, particularly *BCL2L1*, but also *BCL2*, *BMF*, and *BIRC5*, contribute to paclitaxel sensitivity (Flores et al., 2012). *BCL2L1*, *BCL2* and *BMF* were found to be stable in breast cancer tumors, reinforcing the notion that alterations in stable genes contribute to drug resistance (Park et al., 2012). Supplementary Table 13 describes genes analyzed in the context of their biological pathways and relevant literature.

The gemcitabine metabolic pathway has been well characterized (Alvarellos et al., 2014), however the critical genes have not been treated as an ensemble in conferring resistance (see Supplementary Table 14 for interpretation of the MFA results for all genes). The MFA analyses indicated gemcitabine genes predominately contribute to drug resistance through overexpression. For *DCTD*, however, underexpression is associated with increased resistance in the MFA analysis. *DCTD* deficiency causes an imbalance in the dNTP pool (Eriksson et al., 1984), which affects control of DNA replication. *DCTD* is inhibited by dFdCTP (a gemcitabine metabolite) through a mechanism by which gemcitabine exhibits self-potentiation (the reduction of competing natural metabolites) (Xu and Plunkett, 1992). Lower *DCTD* expression and as a consequence, activity, would reduce gemcitabine self-potentiation by altering the dNTP pool. This state is related to drug resistance, which was noticeably lower in 4 cell lines with increased resistance (HCC1187, HCC1428, HCC202, and MDAMB134VI). Like *DCTD*, *CDA* also catalyzes the conversion of gemcitabine monophosphate to difluorodeoxyuridine monophosphate (Figure 2B), and accounts for 90% of this conversion in the cell (Govindarajan et al., 2009). However, drug resistance was associated with CDA overexpression. Likewise, the ribonucleotide reductase subunits *RRM1* and *RRM2B* make significant contributions to the gemcitabine SVM. The *RRM1-RRM2B* complex is associated with mitochondrial genomic integrity (Bourdon et al., 2007) and *RRM2B* is necessary for nucleotide synthesis in DNA repair (Kuo et al., 2012). Changes in *RRM2B* expression could be associated with mitochondrial dysfunction, or may result from loss of p53 expression, which usually induces *RRM2B* expression (Tanaka et al., 2000).

The 11-gene paclitaxel SVM was able to classify FFPE patient samples we obtained and measured in our lab with similar accuracy to that of the cell lines. In addition, the same SVM model was able to predict complete pathological response on a second patient data set, with greater accuracy than the originally reported gene signature (Hatzis et al., 2011). The SVM performed particularly well for predicting drug-sensitive tumors with low or no minimal residual disease (Table 2). The SVM gene signature proved to be resilient as a diagnostic marker, as the performance was not compromised by the lack of expression data for 4 genes.

Unlike paclitaxel, gemcitabine was not used to treat patients in the study by Hatzis et al. (2011) or other publically available data sets. The SVM analysis on RNA expression and DNA copy number from the FFPE-derived tumor punches appeared to predict response more accurately when expression values were obtained for the majority of genes in the SVM. Obtaining high quality gene expression measurements from FFPE samples was especially difficult from older tissue blocks (Supplementary Results V – B) as previously noted (Choudhary et al., 2014). Consequently, the SVM analysis may be better suited for fresh-frozen tumor tissue or more sensitive gene expression analyses (such as mRNA sequencing). Missing data appeared to impact the gemcitabine SVM to a greater extent than the paclitaxel SVM, which may be due to the smaller number of gene measurements required for this SVM.

Including gene expression subtype in the SVM did not improve the classification accuracy even though subtype is known to contribute to tumor biology (Heiser et al., 2012). However, the two paclitaxel PAM50 genes (*MAPT* and *BCL2*) partially stratify the cell lines by subtype during unsupervised clustering (Figure 5). This is not the case in the gemcitabine gene set. In patient data, clustering by expression of the SVM genes also revealed statistically significant deterioration in outcome for low *MAPT* expressing luminal tumors (Supplementary Information VII).

Machine learning may be a fruitful approach in the selection of other chemotherapy agents. Translating our results to the assessment of human tumor samples (Gąsowska-Bodnar et al., 2014) confirmed our gene signature's relevance to predicting chemoresistance by SVM. However, both SVMs were not integrated because cell lines were treated with individual drugs, so predicting whether patient response to these drug interactions will be synergistic or antagonistic is difficult with our approach. In addition, while point mutations are well known contributors to chemoresistance of other drugs, this approach – for either SVM training or testing – is not conducive for prediction of chemosensitivity given the sparse number of observations for these types of mutations.

In cases without residual disease, the paclitaxel SVM was particularly effective in predicting which tumors would

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

show complete pathological response. Docetaxel is prescribed somewhat interchangeably (Crown et al., 2004; Hatzis et al., 2011; O'Shaughnessy et al., 2013) and both paclitaxel and docetaxel act through similar biological pathways (Oshiro et al., 2009). However, the performance of the paclitaxel SVM on patients treated with docetaxel was reduced. This SVM contains 8 paclitaxel resistance genes. Predictions of docetaxel sensitivity might be improved by rederiving a specific SVM using taxane pathway genes (Oshiro et al., 2009), and those known to be associated with resistance to doclitaxel (such as *CYP1B1* (Chang et al., 2015; Cui et al., 2015), *miR-141* or *EIF4E* (Yao et al., 2015), *DKK3* (Tao et al., 2015), *ABCB1* (Hansen et al., 2015; Kato et al., 2015), *BIRC5* (Ghanbari et al., 2014), *ABCC10* (Domanitskaya et al., 2014), *miR-452* (Hu et al., 2014), and *PAWR* (Pereira et al., 2013)). The approach that we have introduced could aid in rational selection of other therapeutic regimens that evade or at least minimize the effects of chemoresistance.

Funding

PKR is supported by the Canadian Breast Cancer Foundation, Canadian Foundation for Innovation, Canada Research Chairs Secretariat and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 371758-2009) and Cytognomix Inc. This work is patent pending. SND received fellowships from the Ontario Graduate Scholarship Program, the Pamela Greenaway-Kohlmeier Translational Breast Cancer Research Unit, and the CIHR Strategic Training Program in Cancer Research and Technology Transfer, and the Canadian Cancer Society Research Institute Travel Award. KB received support from the Department of Biochemistry Undergraduate Summer Research Program and the Dean's Undergraduate Research Opportunities Program, Schulich School of Medicine, Western University.

Acknowledgments

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET) and Compute/Calcul Canada. We thank Ben Shirley and Yancey Li for their assistance in developing the web-based chemotherapy SVM calculator. The authors would also like to thank Dr. Laura Heiser from the Oregon Health & Science University for providing the raw data from the cell line and genomic studies from the Daemen et al., 2013 study and Linda Jackson from the Department of Pathology, Western University, for helping prepare the FFPE tissue blocks for nucleic acid extractions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.molonc.2015.07.006.

REFERENCES

Abdi, H., Valentin, D., 2007. Multiple factor analysis (MFA). In: Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks (CA), pp. 657–663.

Alvarellos, M.L., Lamba, J., Sangkuhl, K., Thorn, C.F., Wang, L., Klein, D.J., Altman, R.B., Klein, T.E., 2014. PharmGKB summary: gemcitabine pathway. Pharmacogenet. Genomics 24, 564–574. http://dx.doi.org/10.1097/FPC.0000000000000086.

Antonov, J., Goldstein, D.R., Oberli, A., Baltzer, A., Pirotta, M., Fleischmann, A., Altermatt, H.J., Jaggi, R., 2005. Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. Lab. Investig. J. Tech. Methods Pathol. 85, 1040–1050. http://dx.doi.org/10.1038/labinvest.3700303.

Aye, Y., Li, M., Long, M.J.C., Weiss, R.S., 2014. Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. Oncogene. http://dx.doi.org/10.1038/onc.2014.155.

Bhat, K.M.R., Setaluri, V., 2007. Microtubule-associated proteins as targets in cancer chemotherapy. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 13, 2849–2854. http://dx.doi.org/10.1158/1078-0432.CCR-06-3040.

Bourdon, A., Minai, L., Serre, V., Jais, J.-P., Sarzi, E., Aubert, S., Chrétien, D., de Lonlay, P., Paquis-Flucklinger, V., Arakawa, H., Nakamura, Y., Munnich, A., Rötig, A., 2007. Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. Nat. Genet. 39, 776–780. http://dx.doi.org/10.1038/ng2040.

Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. http://dx.doi.org/10.1038/nature11412.

Cardoso, F., Harbeck, N., Fallowfield, L., Kyriakides, S., Senkus, E.on behalf of the ESMO Guidelines Working Group, 2012. Locally recurrent or metastatic breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann. Oncol. 23, vii11–vii19. http://dx.doi.org/10.1093/annonc/mds232.

Carmo, C.R., Lyons-Lewis, J., Seckl, M.J., Costa-Pereira, A.P., 2011. A novel requirement for Janus kinases as mediators of drug resistance induced by fibroblast growth factor-2 in human cancer cells. PloS One 6, e19861. http://dx.doi.org/10.1371/journal.pone.0019861.

Chang, I., Mitsui, Y., Fukuhara, S., Gill, A., Wong, D.K., Yamamura, S., Shahryari, V., Tabatabai, Z.L., Dahiya, R., Shin, D.M., Tanaka, Y., 2015. Loss of miR-200c up-regulates CYP1B1 and confers docetaxel resistance in renal cell carcinoma. Oncotarget 6, 7774–7787.

Chen, Z.-S., Hopper-Borge, E., Belinsky, M.G., Shchaveleva, I., Kotova, E., Kruh, G.D., 2003. Characterization of the transport properties of human multidrug resistance protein 7 (MRP7, ABCC10). Mol. Pharmacol. 63, 351–358.

Choudhary, A., Mambo, E., Sanford, T., Boedigheimer, M., Twomey, B., Califano, J., Hadd, A., Oliner, K.S., Beaudenon, S., Latham, G.J., Adai, A.T., 2014. Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel. BMC Med. Genomics 7, 62. http://dx.doi.org/10.1186/s12920-014-0062-0.

Colavito, D., Cartei, G., Dal Bianco, M., Stecca, A., Zustovich, F., Dalle Carbonare, M., Ragazzi, E., Farina, M., Colombrino, E., Leon, A., 2009. Thymidylate synthetase allelic imbalance in clear cell renal carcinoma. Cancer Chemother. Pharmacol. 64, 1195–1200. http://dx.doi.org/10.1007/s00280-009-0986-9.

Crown, J., O'Leary, M., Ooi, W.-S., 2004. Docetaxel and paclitaxel in the treatment of breast cancer: a review of clinical experience. The Oncologist 9 (Suppl 2), 24–32.

Cui, J., Meng, Q., Zhang, X., Cui, Q., Zhou, W., Li, S., 2015. Design and synthesis of new α-naphthoflavones as cytochrome P450

(CYP) 1B1 inhibitors to overcome docetaxel-resistance associated with CYP1B1 overexpression. J. Med. Chem. 58, 3534–3547. http://dx.doi.org/10.1021/acs.jmedchem.5b00265.

Daemen, A., Griffith, O.L., Heiser, L.M., Wang, N.J., Enache, O.M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J.E., Griffith, M., Hur, J.S., Huh, N., Chung, J., Cope, L., Fackler, M., Umbricht, C., Sukumar, S., Seth, P., Sukhatme, V.P., Jakkula, L.R., Lu, Y., Mills, G.B., Cho, R.J., Collisson, E.A., van't Veer, L.J., Spellman, P.T., Gray, J.W., 2013. Modeling precision treatment of breast cancer. Genome Biol. 14, R110. http://dx.doi.org/10.1186/gb-2013-14-10-r110.

Dash, M., Liu, H., 1997. Feature selection for classification. Intell. Data Anal, 131–156. http://dx.doi.org/10.3233/IDA-1997-1302.

Domanitskaya, N., Wangari-Talbot, J., Jacobs, J., Peiffer, E., Mahdaviyeh, Y., Paulose, C., Malofeeva, E., Foster, K., Cai, K.Q., Zhou, Y., Egleston, B., Hopper-Borge, E., 2014. Abcc10 status affects mammary tumour growth, metastasis, and docetaxel treatment response. Br. J. Cancer 111, 696–707. http://dx.doi.org/10.1038/bjc.2014.326.

Dorman, S.N., Viner, C., Rogan, P.K., 2014. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. Sci. Rep. 4, 7063. http://dx.doi.org/10.1038/srep07063.

Duan, Z., Brakora, K.A., Seiden, M.V., 2004. MM-TRAG (MGC4175), a novel intracellular mitochondrial protein, is associated with the taxol- and doxorubicin-resistant phenotype in human cancer cell lines. Gene 340, 53–59. http://dx.doi.org/10.1016/j.gene.2004.06.013.

Duan, Z., Duan, Y., Lamendola, D.E., Yusuf, R.Z., Naeem, R., Penson, R.T., Seiden, M.V., 2003. Overexpression of MAGE/GAGE genes in paclitaxel/doxorubicin-resistant human cancer cell lines. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 9, 2778–2785.

Duan, Z., Feller, A.J., Toh, H.C., Makastorsis, T., Seiden, M.V., 1999. TRAG-3, a novel gene, isolated from a taxol-resistant ovarian carcinoma cell line. Gene 229, 75–81.

Duan, Z., Foster, R., Brakora, K.A., Yusuf, R.Z., Seiden, M.V., 2006. GBP1 overexpression is associated with a paclitaxel resistance phenotype. Cancer Chemother. Pharmacol. 57, 25–33. http://dx.doi.org/10.1007/s00280-005-0026-3.

Duan, Z., Lamendola, D.E., Duan, Y., Yusuf, R.Z., Seiden, M.V., 2005. Description of paclitaxel resistance-associated genes in ovarian and breast cancer cell lines. Cancer Chemother. Pharmacol. 55, 277–285. http://dx.doi.org/10.1007/s00280-004-0878-y.

Ein-Dor, L., Zuk, O., Domany, E., 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. 103, 5923–5928. http://dx.doi.org/10.1073/pnas.0601231103.

Eriksson, S., Skog, S., Tribukait, B., Jäderberg, K., 1984. Deoxyribonucleoside triphosphate metabolism and the mammalian cell cycle. Effects of thymidine on wild-type and dCMP deaminase-deficient mouse S49 T-lymphoma cells. Exp. Cell Res. 155, 129–140.

Ferlini, C., Cicchillitti, L., Raspaglio, G., Bartollino, S., Cimitan, S., Bertucci, C., Mozzetti, S., Gallo, D., Persico, M., Fattorusso, C., Campiani, G., Scambia, G., 2009. Paclitaxel directly binds to Bcl-2 and functionally mimics activity of Nur77. Cancer Res. 69, 6906–6914. http://dx.doi.org/10.1158/0008-5472.CAN-09-0540.

Fleige, S., Pfaffl, M.W., 2006. RNA integrity and the effect on the real-time qRT-PCR performance. Mol. Aspects Med. 27, 126–139. http://dx.doi.org/10.1016/j.mam.2005.12.003.

Flores, M.L., Castilla, C., Ávila, R., Ruiz-Borrego, M., Sáez, C., Japón, M.A., 2012. Paclitaxel sensitivity of breast cancer cells requires efficient mitotic arrest and disruption of Bcl-xL/Bak interaction. Breast Cancer Res. Treat 133, 917–928. http://dx.doi.org/10.1007/s10549-011-1864-9.

Fukino, K., Iido, A., Teramoto, A., Sakamoto, G., Kasumi, F., Nakamura, Y., Emi, M., 1999. Frequent allelic loss at the TOC locus on 17q25.1 in primary breast cancers. Genes. Chromosomes Cancer 24, 345–350.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinforma. Oxf. Engl. 16, 906–914.

Gąsowska-Bodnar, A., Bodnar, L., Dąbek, A., Cichowicz, M., Jerzak, M., Cierniak, S., Kozłowski, W., Baranowski, W., 2014. Survivin expression as a prognostic factor in patients with epithelial ovarian cancer or primary peritoneal cancer treated with neoadjuvant chemotherapy. Int. J. Gynecol. Cancer Off. J. Int. Gynecol. Cancer Soc. 24, 687–696. http://dx.doi.org/10.1097/IGC.0000000000000108.

Ghanbari, P., Mohseni, M., Tabasinezhad, M., Yousefi, B., Saei, A.A., Sharifi, S., Rashidi, M.R., Samadi, N., 2014. Inhibition of survivin restores the sensitivity of breast cancer cells to docetaxel and vinblastine. Appl. Biochem. Biotechnol. 174, 667–681. http://dx.doi.org/10.1007/s12010-014-1125-6.

Glinsky, G.V., Berezovska, O., Glinskii, A.B., 2005. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J. Clin. Invest. 115, 1503–1521. http://dx.doi.org/10.1172/JCI23412.

Govindarajan, R., Leung, G.P.H., Zhou, M., Tse, C.-M., Wang, J., Unadkat, J.D., 2009. Facilitated mitochondrial import of antiviral and anticancer nucleoside drugs by human equilibrative nucleoside transporter-3. Am. J. Physiol. Gastrointest. Liver Physiol. 296, G910–G922. http://dx.doi.org/10.1152/ajpgi.90672.2008.

Hansen, S.N., Westergaard, D., Thomsen, M.B.H., Vistesen, M., Do, K.N., Fogh, L., Belling, K.C., Wang, J., Yang, H., Gupta, R., Ditzel, H.J., Moreira, J., Brünner, N., Stenvang, J., Schrohl, A.-S., 2015. Acquisition of docetaxel resistance in breast cancer cells reveals upregulation of ABCB1 expression as a key mediator of resistance accompanied by discrete upregulation of other specific genes and pathways. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.. http://dx.doi.org/10.1007/s13277-015-3072-4

Harmsen, S., Meijerman, I., Beijnen, J.H., Schellens, J.H.M., 2009. Nuclear receptor mediated induction of cytochrome P450 3A4 by anticancer drugs: a key role for the pregnane X receptor. Cancer Chemother. Pharmacol. 64, 35–43. http://dx.doi.org/10.1007/s00280-008-0842-3.

Hatzis, C., Pusztai, L., Valero, V., Booser, D.J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H., Martin, M., Cotrina, J., Gomez, H., Hubbard, R., Chacón, J.I., Ferrer-Lozano, J., Dyer, R., Buxton, M., Gong, Y., Wu, Y., Ibrahim, N., Andreopoulou, E., Ueno, N.T., Hunt, K., Yang, W., Nazario, A., DeMichele, A., O'Shaughnessy, J., Hortobagyi, G.N., Symmans, W.F., 2011. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305, 1873–1881. http://dx.doi.org/10.1001/jama.2011.593.

He, D.-X., Xia, Y.-D., Gu, X.-T., Jin, J., Ma, X., 2014. A 20-gene signature in predicting the chemoresistance of breast cancer to taxane-based chemotherapy. Mol. Biosyst. 10, 3111–3119. http://dx.doi.org/10.1039/c4mb00464g.

Heijn, M., Hooijberg, J.H., Scheffer, G.L., Szabó, G., Westerhoff, H.V., Lankelma, J., 1997. Anthracyclines modulate multidrug resistance protein (MRP) mediated organic anion transport. Biochim. Biophys. Acta 1326, 12–22.

Heiser, L.M., Sadanandam, A., Kuo, W.-L., Benz, S.C., Goldstein, T.C., Ng, S., Gibb, W.J., Wang, N.J., Ziyad, S., Tong, F., Bayani, N., Hu, Z., Billig, J.I., Dueregger, A., Lewis, S., Jakkula, L., Korkola, J.E., Durinck, S., Pepin, F., Guan, Y., Purdom, E., Neuvial, P., Bengtsson, H., Wood, K.W., Smith, P.G., Vassilev, L.T., Hennessy, B.T., Greshock, J., Bachman, K.E., Hardwicke, M.A., Park, J.W., Marton, L.J., Wolf, D.M., Collisson, E.A., Neve, R.M.,

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

Mills, G.B., Speed, T.P., Feiler, H.S., Wooster, R.F., Haussler, D., Stuart, J.M., Gray, J.W., Spellman, P.T., 2012. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc. Natl. Acad. Sci. U. S. A 109, 2724–2729. http://dx.doi.org/10.1073/pnas.1018854108.

Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, A.U., Dempsey, P.J., Rouzier, R., Sneige, N., Ross, J.S., Vidaurre, T., Gómez, H.L., Hortobagyi, G.N., Pusztai, L., 2006. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 24, 4236–4244. http://dx.doi.org/10.1200/JCO.2006.05.6861.

Hong, J., Zhou, J., Fu, J., He, T., Qin, J., Wang, L., Liao, L., Xu, J., 2011. Phosphorylation of serine 68 of Twist1 by MAPKs stabilizes Twist1 protein and promotes breast cancer cell invasiveness. Cancer Res. 71, 3980–3990. http://dx.doi.org/10.1158/0008-5472.CAN-10-2914.

Hopper-Borge, E., Xu, X., Shen, T., Shi, Z., Chen, Z.-S., Kruh, G.D., 2009. Human multidrug resistance protein 7 (ABCC10) is a resistance factor for nucleoside analogues and epothilone B. Cancer Res. 69, 178–184. http://dx.doi.org/10.1158/0008-5472.CAN-08-1420.

Hu, Q., Chen, W., Zhong, S., Zhang, J., Ma, T., Ji, H., Lv, M., Tang, J., Zhao, J., 2014. MicroRNA-452 contributes to the docetaxel resistance of breast cancer cells. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. 35, 6327–6334. http://dx.doi.org/10.1007/s13277-014-1834-z.

Jordan, M.A., Wilson, L., 2004. Microtubules as a target for anticancer drugs. Nat. Rev. Cancer 4, 253–265. http://dx.doi.org/10.1038/nrc1317.

Kaczanowska, S., Joseph, A.M., Davila, E., 2013. TLR agonists: our best frenemy in cancer immunotherapy. J. Leukoc. Biol. 93, 847–863. http://dx.doi.org/10.1189/jlb.1012501.

Kato, T., Mizutani, K., Kameyama, K., Kawakami, K., Fujita, Y., Nakane, K., Kanimoto, Y., Ehara, H., Ito, H., Seishima, M., Deguchi, T., Ito, M., 2015. Serum exosomal P-glycoprotein is a potential marker to diagnose docetaxel resistance and select a taxoid for patients with prostate cancer. Urol. Oncol.. http://dx.doi.org/10.1016/j.urolonc.2015.04.019

Kavallaris, M., 2010. Microtubules and resistance to tubulin-binding agents. Nat. Rev. Cancer 10, 194–204. http://dx.doi.org/10.1038/nrc2803.

Kuo, M.-L., Sy, A.J., Xue, L., Chi, M., Lee, M.T.-C., Yen, T., Chiang, M.-I., Chang, L., Chu, P., Yen, Y., 2012. RRM2B suppresses activation of the oxidative stress pathway and is up-regulated by p53 during senescence. Sci. Rep. 2, 822. http://dx.doi.org/10.1038/srep00822.

Kutuk, O., Letai, A., 2008. Alteration of the mitochondrial apoptotic pathway is key to acquired paclitaxel resistance and can be reversed by ABT-737. Cancer Res. 68, 7985–7994. http://dx.doi.org/10.1158/0008-5472.CAN-08-1418.

Lecureur, V., Sun, D., Hargrove, P., Schuetz, E.G., Kim, R.B., Lan, L.B., Schuetz, J.D., 2000. Cloning and expression of murine sister of P-glycoprotein reveals a more discriminating transporter than MDR1/P-glycoprotein. Mol. Pharmacol. 57, 24–35.

Lee, S.-Y., Im, S.-A., Park, Y.H., Woo, S.Y., Kim, S., Choi, M.K., Chang, W., Ahn, J.S., Im, Y.-H., 2014. Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. Eur. J. Cancer Oxf. Engl. 1990 (50), 698–705. http://dx.doi.org/10.1016/j.ejca.2013.11.028.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25, 1–18.

Lu, J., Tan, M., Huang, W.-C., Li, P., Guo, H., Tseng, L.-M., Su, X., Yang, W.-T., Treekitkarnmongkol, W., Andreeff, M., Symmans, F., Yu, D., 2009. Mitotic deregulation by survivin in ErbB2-overexpressing breast cancer cells contributes to taxol resistance. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 15, 1326–1334. http://dx.doi.org/10.1158/1078-0432.CCR-08-0954.

Mackey, J.R., Yao, S.Y., Smith, K.M., Karpinski, E., Baldwin, S.A., Cass, C.E., Young, J.D., 1999. Gemcitabine transport in xenopus oocytes expressing recombinant plasma membrane mammalian nucleoside transporters. J. Natl. Cancer Inst. 91, 1876–1881.

Marcé, S., Molina-Arcas, M., Villamor, N., Casado, F.J., Campo, E., Pastor-Anglada, M., Colomer, D., 2006. Expression of human equilibrative nucleoside transporter 1 (hENT1) and its correlation with gemcitabine uptake and cytotoxicity in mantle cell lymphoma. Haematologica 91, 895–902.

Ma, X.-J., Wang, Z., Ryan, P.D., Isakoff, S.J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J.T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habin, K., Baer, T.M., Brugge, J., Haber, D.A., Erlander, M.G., Sgroi, D.C., 2004. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 5, 607–616. http://dx.doi.org/10.1016/j.ccr.2004.05.015.

McGrogan, B.T., Gilmartin, B., Carney, D.N., McCann, A., 2008. Taxanes, microtubules and chemoresistant breast cancer. Biochim. Biophys. Acta 1785, 96–132. http://dx.doi.org/10.1016/j.bbcan.2007.10.004.

Mini, E., Nobili, S., Caciagli, B., Landini, I., Mazzei, T., 2006. Cellular pharmacology of gemcitabine. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO 17 (Suppl 5), v7–12. http://dx.doi.org/10.1093/annonc/mdj941.

Musella, V., Callari, M., Di Buduo, E., Scuro, M., Dugo, M., Miodini, P., Bianchini, G., Paolini, B., Gianni, L., Daidone, M.G., Cappelletti, V., 2015. Use of formalin-fixed paraffin-embedded samples for gene expression studies in breast cancer patients. PloS One 10, e0123194. http://dx.doi.org/10.1371/journal.pone.0123194.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P.T., DeVries, S., Lapuk, A., Wang, N.J., Kuo, W.-L., Stilwell, J.L., Pinkel, D., Albertson, D.G., Waldman, F.M., McCormick, F., Dickson, R.B., Johnson, M.D., Lippman, M., Ethier, S., Gazdar, A., Gray, J.W., 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10, 515–527. http://dx.doi.org/10.1016/j.ccr.2006.10.008.

Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814.

Nguyen Huu, N.S., Ryder, W.D.J., Zeps, N., Flasza, M., Chiu, M., Hanby, A.M., Poulsom, R., Clarke, R.B., Baron, M., 2008. Tumour-promoting activity of altered WWP1 expression in breast cancer and its utility as a prognostic indicator. J. Pathol. 216, 93–102. http://dx.doi.org/10.1002/path.2385.

Nilsson, R., Björkegren, J., Tegnér, J., 2009. On reliable discovery of molecular signatures. BMC Bioinformatics 10, 38. http://dx.doi.org/10.1186/1471-2105-10-38.

Oostendorp, L.J.M., Stalmeier, P.F.M., Donders, A.R.T., van der Graaf, W.T.A., Ottevanger, P.B., 2011. Efficacy and safety of palliative chemotherapy for patients with advanced breast cancer pretreated with anthracyclines and taxanes: a systematic review. Lancet Oncol. 12, 1053–1061. http://dx.doi.org/10.1016/S1470-2045(11)70045-6.

O'Shaughnessy, J., Gradishar, W.J., Bhar, P., Iglesias, J., 2013. Nab-paclitaxel for first-line treatment of patients with metastatic breast cancer and poor prognostic factors: a retrospective analysis. Breast Cancer Res. Treat 138, 829–837. http://dx.doi.org/10.1007/s10549-013-2447-8.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006

Oshiro, C., Marsh, S., McLeod, H., Carrillo, M.W., Klein, T., Altman, R., 2009. Taxane pathway. Pharmacogenet. Genomics 19, 979–983. http://dx.doi.org/10.1097/FPC.0b013e3283335277.

Park, N.I., Rogan, P.K., Tarnowski, H.E., Knoll, J.H.M., 2012. Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy. Mol. Oncol. 6, 347–359. http://dx.doi.org/10.1016/j.molonc.2012.01.001.

Pereira, M.C., de Bessa-Garcia, S.A., Burikhanov, R., Pavanelli, A.C., Antunes, L., Rangnekar, V.M., Nagai, M.A., 2013. Prostate apoptosis response-4 is involved in the apoptosis response to docetaxel in MCF-7 breast cancer cells. Int. J. Oncol. 43, 531–538. http://dx.doi.org/10.3892/ijo.2013.1983.

Plunkett, W., Huang, P., Xu, Y.Z., Heinemann, V., Grunewald, R., Gandhi, V., 1995. Gemcitabine: metabolism, mechanisms of action, and self-potentiation. Semin. Oncol. 22, 3–10.

Prat, A., Karginova, O., Parker, J.S., Fan, C., He, X., Bixby, L., Harrell, J.C., Roman, E., Adamo, B., Troester, M., Perou, C.M., 2013. 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, 237–255. http://dx.doi.org/10.1007/s10549-013-2743-3.

Preissner, S., Kroll, K., Dunkel, M., Senger, C., Goldsobel, G., Kuzman, D., Guenther, S., Winnenburg, R., Schroeder, M., Preissner, R., 2010. SuperCYP: a comprehensive database on cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. Nucleic Acids Res. 38, D237–D243. http://dx.doi.org/10.1093/nar/gkp970.

Rajput, S., Volk-Draper, L.D., Ran, S., 2013. TLR4 is a novel determinant of the response to paclitaxel in breast cancer. Mol. Cancer Ther. 12, 1676–1687. http://dx.doi.org/10.1158/1535-7163.MCT-12-1019.

Rao, P.S., Bickel, U., Srivenugopal, K.S., Rao, U.S., 2008. Bap29varP, a variant of Bap29, influences the cell surface expression of the human P-glycoprotein. Int. J. Oncol. 32, 135–144.

Samuels, B.L., Hollis, D.R., Rosner, G.L., Trump, D.L., Shapiro, C.L., Vogelzang, N.J., Schilsky, R.L., 1997. Modulation of vinblastine resistance in metastatic renal cell carcinoma with cyclosporine A or tamoxifen: a cancer and leukemia group B study. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 3, 1977–1984.

Shirley, B.C., Mucaki, E.J., Whitehead, T., Costea, P.I., Akan, P., Rogan, P.K., 2013. Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. Genomics Proteomics Bioinformatics 11, 77–85. http://dx.doi.org/10.1016/j.gpb.2013.01.008.

Shoemaker, R.H., 2006. The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer 6, 813–823. http://dx.doi.org/10.1038/nrc1951.

Shoushtari, A.N., Szmulewitz, R.Z., Rinker-Schaeffer, C.W., 2011. Metastasis-suppressor genes in clinical practice: lost in translation? Nat. Rev. Clin. Oncol. 8, 333–342. http://dx.doi.org/10.1038/nrclinonc.2011.65.

Symmans, W.F., Peintinger, F., Hatzis, C., Rajan, R., Kuerer, H., Valero, V., Assad, L., Poniecka, A., Hennessy, B., Green, M., Buzdar, A.U., Singletary, S.E., Hortobagyi, G.N., Pusztai, L., 2007. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 25, 4414–4422. http://dx.doi.org/10.1200/JCO.2007.10.6823.

Tanaka, H., Arakawa, H., Yamaguchi, T., Shiraishi, K., Fukuda, S., Matsui, K., Takei, Y., Nakamura, Y., 2000. A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. Nature 404, 42–49. http://dx.doi.org/10.1038/35003506.

Tanaka, S., Nohara, T., Iwamoto, M., Sumiyoshi, K., Kimura, K., Takahashi, Y., Tanigawa, N., 2009. Tau expression and efficacy of paclitaxel treatment in metastatic breast cancer. Cancer Chemother. Pharmacol. 64, 341–346. http://dx.doi.org/10.1007/s00280-008-0877-5.

Tantivejkul, K., Loberg, R.D., Mawocha, S.C., Day, L.L., John, L.S., Pienta, B.A., Rubin, M.A., Pienta, K.J., 2005. PAR1-mediated NFkappaB activation promotes survival of prostate cancer cells through a Bcl-xL-dependent mechanism. J. Cell. Biochem 96, 641–652. http://dx.doi.org/10.1002/jcb.20533.

Tao, L., Huang, G., Chen, Y., Chen, L., 2015. DNA methylation of DKK3 modulates docetaxel chemoresistance in human nonsmall cell lung cancer cell. Cancer Biother. Radiopharm. 30, 100–106. http://dx.doi.org/10.1089/cbr.2014.1797.

Ueno, H., Kiyosawa, K., Kaniwa, N., 2007. Pharmacogenomics of gemcitabine: can genetic studies lead to tailor-made therapy? Br. J. Cancer 97, 145–151. http://dx.doi.org/10.1038/sj.bjc.6603860.

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536. http://dx.doi.org/10.1038/415530a.

Viner, C., Dorman, S.N., Shirley, B.C., Rogan, P.K., 2014. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. F1000Research 3, 8. http://dx.doi.org/10.12688/f1000research.3-8.v2.

Wang, C., Zhang, W., Fu, M., Yang, A., Huang, H., Xie, J., 2015. Establishment of human pancreatic cancer gemcitabine-resistant cell line with ribonucleotide reductase overexpression. Oncol. Rep. 33, 383–390. http://dx.doi.org/10.3892/or.2014.3599.

Wang, K., Deng, Q.-T., Liao, N., Zhang, G.-C., Liu, Y.-H., Xu, F.-P., Zu, J., Li, X.-R., Wu, Y.-L., 2013. Tau expression correlated with breast cancer sensitivity to taxanes-based neoadjuvant chemotherapy. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. 34, 33–38. http://dx.doi.org/10.1007/s13277-012-0507-z.

Xing, H., Weng, D., Chen, G., Tao, W., Zhu, T., Yang, X., Meng, L., Wang, S., Lu, Y., Ma, D., 2008. Activation of fibronectin/PI-3K/Akt2 leads to chemoresistance to docetaxel by regulating survivin protein expression in ovarian and breast cancer cells. Cancer Lett. 261, 108–119. http://dx.doi.org/10.1016/j.canlet.2007.11.022.

Xu, Y.Z., Plunkett, W., 1992. Modulation of deoxycytidylate deaminase in intact human leukemia cells. Action of 2′,2′-difluorodeoxycytidine. Biochem. Pharmacol. 44, 1819–1827.

Yao, Y.-S., Qiu, W.-S., Yao, R.-Y., Zhang, Q., Zhuang, L.-K., Zhou, F., Sun, L.-B., Yue, L., 2015. miR-141 confers docetaxel chemoresistance of breast cancer cells via regulation of EIF4E expression. Oncol. Rep. 33, 2504–2512. http://dx.doi.org/10.3892/or.2015.3866.

Please cite this article in press as: Dorman, S.N., et al., Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning, Molecular Oncology (2015), http://dx.doi.org/10.1016/j.molonc.2015.07.006