ROCplot.org: Validating predictive biomarkers of chemotherapy/hormonal therapy/anti-HER2 therapy using transcriptomic data of 3,104 breast cancer patients

János T. Fekete[1] and Balázs Győrffy[1,2]

[1] 2nd Department of Pediatrics, Semmelweis University, Budapest, Hungary
[2] MTA TTK Lendület Cancer Biomarker Research Group, Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungary

Systemic therapy of breast cancer can include chemotherapy, hormonal therapy and targeted therapy. Prognostic biomarkers are able to predict survival and predictive biomarkers are able to predict therapy response. In this report, we describe the initial release of the first available online tool able to identify gene expression-based predictive biomarkers using transcriptomic data of a large set of breast cancer patients. Published gene expression data of 36 publicly available datasets were integrated with treatment data into a unified database. Response to therapy was determined using either author-reported pathological complete response data ($n = 1,775$) or relapse-free survival status at 5 years ($n = 1,329$). Treatment data includes chemotherapy ($n = 2,108$), endocrine therapy ($n = 971$) and anti-human epidermal growth factor receptor 2 (HER2) therapy ($n = 267$). The transcriptomic database includes 20,089 unique genes and 54,675 probe sets. Gene expression and therapy response are compared using receiver operating characteristics and Mann–Whitney tests. We demonstrate the utility of the pipeline by cross-validating 23 paclitaxel resistance-associated genes in different molecular subtypes of breast cancer. An additional set of established biomarkers including TP53 for chemotherapy in Luminal breast cancer ($p = 1.01E\text{-}19$, AUC = 0.769), HER2 for trastuzumab therapy ($p = 8.4E\text{-}04$, AUC = 0.629) and PGR for hormonal therapy ($p = 8.6E\text{-}05$, AUC = 0.7), are also endorsed. The tool is designed to validate and rank new predictive biomarker candidates in real time. By analyzing the selected genes in a large set of independent patients, one can select the most robust candidates and quickly eliminate those that are most likely to fail in a clinical setting. The analysis tool is accessible at www.rocplot.org.

Key words: chemotherapy, hormonal therapy, targeted therapy, ROC, relapse-free survival, molecular subtypes

Abbreviations: ABCB1: ATP binding cassette subfamily B member 1; ABCB11: ATP binding cassette subfamily B member 11; AUC: area under the curve; BBC3: BCL2 binding component 3; BCAP29: B cell receptor-associated protein 29; BCL2: BCL2 apoptosis regulator; BCL2L1: BCL2 like 1; BMF: BCL2 modifying factor; CMF: cyclophosphamide, methotrexate, fluorouracil; CNGA3: cyclic nucleotide gated channel alpha 3; CSAG2: CSAG family member 2; CYP2C8: cytochrome P450 family 2 subfamily C member 8; CYP3A4: cytochrome P450 family 3 subfamily A member 4; ERBB2: erb-b2 receptor tyrosine kinase 2; ESR1: estrogen receptor 1; FAC: fluorouracil, adriamycin, cytoxan; FDR: false discovery rate; FEC: fluorouracil, epirubicin, cyclophosphamide; FN1: fibronectin 1; GBP1: guanylate binding protein 1; GEO: Gene Expression Omnibus; HER2: human epidermal growth factor receptor 2; MAP2: microtubule-associated protein 2; MAP4: microtubule-associated protein 4; MDR: multiple drug resistance; MIAME: minimum information about a microarray experiment; MKI67: marker of proliferation Ki-67; OPRK1: opioid receptor kappa 1; pCR: pathological complete response; PGR: progesterone receptor; RFS: relapse-free survival; ROC: receiver operating characteristic; SLCO1B3: solute carrier organic anion transporter family member 1B3; TNBC: triple-negative breast cancer; TP53: tumor protein p53; TUBB1: tubulin beta 1 class VI; TUBB4: tubulin beta 4A class IVa; TWIST1: twist family bHLH transcription factor 1

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Research, Development and Innovation Office, Hungary; **Grant numbers:** KH-129581, NVKP_16-1-2016-0037, 2018-1.3.1-VKE-2018-00032

DOI: 10.1002/ijc.32369

History: Received 9 Feb 2019; Accepted 17 Apr 2019; Online 24 Apr 2019.

Correspondence to: Balázs Győrffy, MD, PhD, DSc, Lendület Cancer Biomarker Research Group, MTA TTK, Institute of Enzymology, Magyar tudósok körútja 2, H-1117, Budapest, Hungary, E-mail: gyorffy.balazs@ttk.mta.hu

What's new?

While several online tools capable of delivering a prognostic prediction for breast cancer already exist, no such genome-wide biomarker validation tool is available to evaluate and compare predictive biomarker candidates. Here, the authors combine multiple datasets to establish a sufficiently large breast cancer cohort with transcriptomic, anticancer treatment, and clinical response data. Then, they establish a framework capable of studying new candidate genes by mining this database and demonstrate the robustness of the pipeline by cross-validating an established set of resistance-associated genes. The novel online platform provides an easily accessed resource for researchers to validate and rank future biomarker candidates.

Introduction

A biomarker is a scientifically supported analytical tool with a clinically useful significance. A biomarker can be measured in a test system and has a recognized characteristic that enables researchers to use it for support in making decisions in pharmacology, physiology or toxicology. Today, the Food and Drug Administration (FDA) encourages the use of biomarkers to increase the efficacy of new drugs.[1] In cancer treatment, two major types of biomarkers can be implemented. Prognostic biomarkers are able to predict patient survival and predictive biomarkers are able to predict the response to a selected anticancer therapy.[2]

Depending on the molecular and pathological characteristics of a tumor and the projected survival of the patient, systemic therapy of breast cancer can include chemotherapy, hormonal therapy, and molecular targeted therapy. Each of these therapies is supported by prognostic and predictive biomarkers. Estrogen receptor and progesterone receptor are the most important predictive biomarkers to select those eligible for hormonal therapy.[3] Molecular targeted therapy is given to those whose tumors harbor an amplification or overexpression of the erb-b2 receptor tyrosine kinase 2 (ERBB2)/human epidermal growth factor receptor 2 (HER2) receptor.[4] Multigene tests can provide support to select those most benefiting from chemotherapy (for a comprehensive review, see Ref. 2).

Response to an anticancer agent depends on pharmacological (dose, pharmacokinetics and localization of the tumor) and cellular factors. This second group of cellular factors can be further subdivided according to three major mechanisms of action. First, the intracellular drug concentration can be decreased in cases where transmembrane transport systems are activated[5] or when the agent is intracellularly metabolized.[6] Second, an altered interaction between the drug and the target can lead to lower efficiency of a given drug.[7] Third, a change in the cellular response, including mutations and expression changes in genes related to cell cycle, DNA repair[8] and apoptosis[9] can also allow cancer cells to evade the effects of systemic anticancer therapy.

In the last decade, several online tools capable of delivering a prognostic prediction were developed for breast cancer. Almost all of these tools include a platform linking survival and gene expression (see KM-plotter,[10] PROGgene,[11] GenExMiner,[12] APPEX,[13] KMexpress[14] or PPISURV[15]). Other tools use miRNA expression (miRpower,[16] BreastMark[17]) or estimate survival for a given patient (RecurrenceOnline[18]). However, to date, no such genome-wide biomarker validation tool has been made available to evaluate and compare predictive biomarker candidates.

In our study, we report an online platform enabling the identification of predictive biomarkers in breast cancer. Multiple transcriptome-level gene expression datasets were integrated into a single database containing 3,104 breast cancer patients with treatment and response data. Responder and nonresponder patients are compared *via* two diverse statistical approaches. In the second part of the project, prediction results delivered by the pipeline are used to validate clinically used and previously proposed biomarker candidates.

Methods

Database construction

Breast cancer datasets were identified in Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/gds) using the GEO platform IDs "GPL96" (for HG-U133A), "GPL570" (for HG-U133 Plus 2.0), "GPL571" (for HG-U133A_2) and the keywords "breast," "cancer" and "therapy." Datasets with less than 30 samples were excluded at this stage (a few datasets included more than 30 specimens initially, but some of the samples dropped out because only a reduced amount of patient samples were actually useful for our study). We selected the above platforms because they are widely used and because they use the exact same probes to measure the same genes. For genes with multiple probes, we used Jetset[19] to select the most reliable probe set (http://www.cbs.dtu.dk/biotools/jetset/).

The raw CEL files were MAS5 normalized in the R statistical environment (http://www.r-project.org) using the Affy Bioconductor library. A second scaling normalization was performed to set the mean expression of the 22,277 identical probe sets in each array to 1,000 to reduce batch effects due to different mean targets during normalization of the three human genome arrays (Supporting Information Fig. S1).

Repeatedly published arrays ($n = 46$) were identified by searching for identical expression values. Of these arrays, only the first one was retained in the final database.

Quality control

Each included study fulfilled the Minimum Information About a Microarray Experiment (MIAME) criteria—description of extraction protocol, hybridization protocol, scan protocol and data processing.

Quality control of the gene chips was performed as described previously.[18] In brief, each array was examined for background intensity, scaling factor, percentage of present calls, bioB-, bioC-, bioD-, cre-, dap-, lys-, phe-, thr- and tryp-spikes, and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and beta-actin 3′–5′ ratio. Arrays with more than one parameter outside of the 95% range across all arrays were excluded from further analysis.

Some of the datasets used samples from randomized clinical trials. As the reliability of clinical data collected in these studies can be superior to those not from a clinical trial, we marked each of these studies.

Receptor status determination

Patients were assigned into molecular subtypes based on the expression of estrogen receptor 1 (ESR1), HER2 and marker of proliferation Ki-67 (MKI67). For ESR1, the probe set 205225_at was used with a cutoff value of 500, for the HER2 receptor, the probe set 216836_s_at was used with a cutoff value of 4,800 and for MKI67, the probe set 212021_s_at was used with a cutoff value of 470.[18] We compared the gene-array based receptor designations to the IHC-based receptor designations in patients where IHC data was available and found substantial agreement in both the RFS and pathological complete response (pCR) cohorts (Supporting Information Table S1).

Statistical methods

First, the patients are assigned to two cohorts (responder and nonresponder) based on their clinical characteristics. Patients with neoadjuvant chemotherapy were classified according to pathological response as published by the authors. In this, instead of four cohorts (progressive disease, stable disease, partial response and complete response), we have assigned all patients into two cohorts, including those where no residual histological evidence of the tumor remains after chemotherapy (responders) and all other patients with residual tumor tissue (nonresponders).

Patients with adjuvant therapy were classified into two cohorts based on survival status at 5-year follow-up. In this case, expression of the gene in patients relapsed before 5 years is compared to the expression of the gene in patients surviving over 5 years. Patients censored before 5 years are excluded.

The two cohorts are compared using Mann–Whitney test or Receiver Operating Characteristic test in the R statistical environment (www.r-project.org) using Bioconductor libraries (www.bioconductor.org). Statistical significance was set at $p < 0.05$.

Validation of established markers and discovery of new candidates

First, we evaluated a set of commonly referenced predictive biomarkers including progesterone receptor (PGR), HER2 and tumor protein p53 (TP53).

A more exciting validation analysis was executed to confirm previously published paclitaxel resistance biomarkers. To this end, we used an established set of 31 genes.[20] From the 31-gene panel, we analyzed 29 genes because the Affymetrix chip does not have probes to CSAG family member 2 (CSAG2) and because the expression of tubulin beta 4A class IVa (TUBB4A) and TUBB4B are combined in the TUBB4 probe set. The study of *Dorman et al.* was selected because of two major advantages: first, they used support vector machines, an approach different from our analysis pipeline. Second, the genes identified in their analysis were independently validated using tumor blocks from a panel of 340 independent patients. In this instance, the receiver operating characteristic (ROC) plotter was used to compute pathological complete response (pCR) based classifications for each of the proposed genes in each of the molecular subtypes. To this end, we used all of the samples in which the clinical file confirmed the administration of paclitaxel as a chemotherapy agent.

Finally, we performed the analysis across all genes in triple-negative breast cancer (TNBC) patients to identify new biomarker candidates of chemotherapy resistance specifically in this cohort. For this, ESR1- and HER2-negative patients were designated as TNBC, HER2-positive and ESR1-negative patients were designated as HER2, ESR1-positive and MKI67-negative patients were designated as Luminal A, and all remaining samples were designated as Luminal B patients.

Results

Database—pCR dataset

Processing of the GEO samples is summarized in Figure 1*a*. Overall, 5,476 breast cancer patients were identified in GEO with relapse data. After exclusion of repeatedly published arrays and samples measured using a different platform, 3,756 patients from 20 datasets remained. Sufficient clinical data were available for 1,775 patients from 16 datasets (Fig. 1*b*). Aggregate clinical characteristics of the pCR samples are presented in Table 1*a* and Figure 1*c*.

Database—RFS dataset

Overall, 9,013 breast cancer patients with follow-up data were identified in GEO. Of these patients, 3,070 samples from 28 datasets represent unique samples measured by the Human Genome Arrays. Sufficient clinical data, including relapse-free survival (RFS) time and treatment data, were available for 1,329 patients from 20 datasets (Fig. 1*b*). Aggregate clinical characteristics of the RFS samples are presented in Table 1*b* and Figure 1*c*.

Treatment cohorts

Most patients in the pCR cohort have received cytotoxic chemotherapy, including a regimen containing an anthracycline ($n = 1{,}626$) or taxane ($n = 1{,}213$). Smaller cohorts comprise patients with cyclophosphamide, methotrexate, fluorouracil

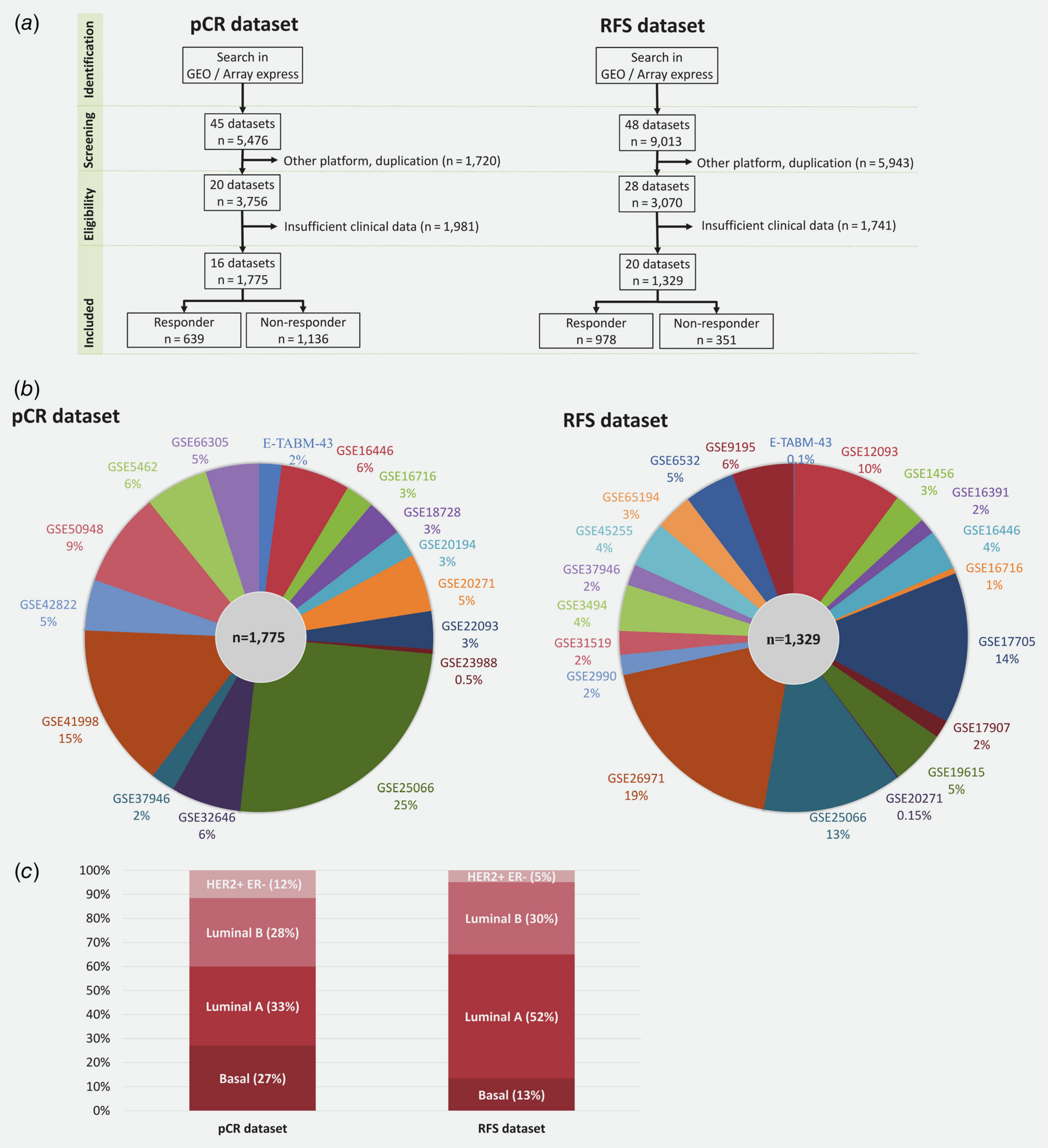

Figure 1. Overview of database setup. Pipeline used to select samples to be included in the pathological response (pCR dataset) and in the relapse-free survival (RFS dataset) cohorts (*a*), proportion of datasets included in each cohort (*b*) and distribution of molecular subtypes (*c*). [Color figure can be viewed at wileyonlinelibrary.com]

(CMF; *n* = 156), fluorouracil, epirubicin, cyclophosphamide (FEC; *n* = 303) or fluorouracil, adriamycin, cytoxan (FAC; *n* = 248) protocols and patients with ixabepilone (*n* = 136), lapatinib (*n* = 65) and trastuzumab (*n* = 186) treatments. A minor group of patients was administered hormonal therapy (aromatase inhibitors).

Two-thirds of patients in the RFS cohort have received a hormonal therapy (*n* = 907). The most common chemotherapy

Table 1. Overview of datasets included in the analysis with available response data (*a*) and survival and follow-up at 5 years (*b*)

(*a*)

Dataset	Platform	Reference (PMID)	Year	Sample size	Age	Outcome (responder/ nonresponder)	Grade (1/2/3)	Nodal status (0/1)	Molecular subtype (Basal/LumA/LumB/HER2+)
E-TABM-43	HG-U133A	17,388,661	2007	37	48.51 ± 12.5	11/26	–/14/23	–	7/14/10/6
GSE5462	HG-U133A	17,885,619	2007	104	–	74/30	–	–	–/101/3/–
GSE16716	HG-U133A	20,676,074	2010	47	54.38 ± 11.2	18/29	–/13/34	2/6	4/6/24/13
GSE18728	HG-U133_PLUS_2	20,012,355	2010	61	–	23/38	–	–	14/35/10/2
GSE20194	HG-U133A	20,064,235	2010	45	51.91 ± 11.5	7/38	–/8/26	9/27	9/10/19/7
GSE20271	HG-U133A	20,829,329	2010	96	50.75 ± 10.3	12/84	5/30/41	38/57	21/18/48/9
GSE16446	HG-U133_PLUS_2	21,422,418	2011	114	–	16/98	2/20/87	52/62	83/3/4/24
GSE22093	HG-U133A	21,191,116	2011	62	50.03 ± 11.3	25/37	2/19/39	18/18	20/7/28/7
GSE23988	HG-U133A	21,191,116	2011	8	49.38 ± 9.8	7/1	–/3/4	1/7	2/2/4/–
GSE25066	HG-U133A	21,558,518	2011	448	49.72 ± 10.5	85/363	28/157/230	151/294	128/121/194/5
GSE32646	HG-U133_PLUS_2	22,320,227	2011	115	51.49 ± 10.3	27/88	16/78/21	32/83	23/49/30/13
GSE37946	HG-U133A	22,460,789	2012	40	47.90 ± 11.0	23/17	–/10/29	32/8	4/2/25/9
GSE42822	HG-U133A	23,158,478	2012	84	49.53 ± 8.9	36/48	–/20/49	29/52	18/36/13/17
GSE41998	HG-U133A_2	23,340,299	2013	270	48.41 ± 10.6	201/69	–	–	113/122/15/20
GSE50948	HG-U133_PLUS_2	24,443,618	2014	156	51.67 ± 9.9	53/103	–/67/86	–	28/37/46/45
GSE66305	HG-U133_PLUS_2	26,245,675	2015	88	–	27/61	–	–	6/22/32/28
Total				1,775	50.04 ± 10.59	639/1136	53/439/669	364/612	480/585/505/205

(*b*)

Dataset	Platform	Reference (PMID)	Year	Sample size	Age	Outcome (responder/ nonresponder)	Grade (1/2/3)	Nodal status (0/1)	Molecular subtype (Basal/LumA/LumB/HER2+)
GSE1456	HG-U133A	16,280,042	2005	40	–	26/14	–/7/30	–	16/2/21/1
GSE3494	HG-U133A	16,141,321	2005	57	–	35/22	–	–	–/36/21/–
GSE2990	HG-U133A	16,478,745	2006	25	–	18/7	13/–/12	–	–/20/5/–
E-TABM-43	HG-U133A	17,388,661	2007	1	–	0/1	–	–	–/–/1/–
GSE6532	HG-U133A	17,401,012	2007	62	60.72 ± 10.3	48/14	–/41/–	39/16	–/47/13/2
GSE9195	HG-U133_PLUS_2	18,498,629	2008	76	64.30 ± 9.2	67/9	14/20/24	40/36	1/65/10/–
GSE12093	HG-U133A	18,821,012	2009	134	–	122/12	–	–	–/103/31/–
GSE16391	HG-U133_PLUS_2	19,573,224	2009	20	59.90 ± 8.8	10/10	–/11/9	9/11	–/14/6/–
GSE16716	HG-U133A	20,064,235	2010	7	54.43 ± 9.7	5/2	–/–/7	–/7	1/–/2/4
GSE17705	HG-U133A	20,697,068	2010	186	–	158/28	–	105/76	5/95/86/–
GSE17907	HG-U133_PLUS_2	20,932,292	2010	22	53.95 ± 11.9	12/10	2/4/13	8/9	–/–/5/17
GSE19615	HG-U133_PLUS_2	20,098,429	2010	68	53.18 ± 11.4	57/11	19/19/30	34/34	11/34/16/7
GSE20271	HG-U133A	20,829,329	2010	2	63.50 ± 13.4	–/2	–/1/1	–	–/–/1/1

(Continues)

Table 1. Overview of datasets included in the analysis with available response data (*a*) and survival and follow-up at 5 years (*b*) (Continued)

(*b*)

Dataset	Platform	Reference (PMID)	Year	Sample size	Age	Outcome (responder/ nonresponder)	Grade (1/2/3)	Nodal status (0/1)	Molecular subtype (Basal/LumA/LumB/HER2+)
GSE25066	HG-U133A	21,558,518	2011	171	50.62 ± 11.1	61/110	6/58/94	45/126	60/38/69/4
GSE26971	HG-U133A	21,807,638	2011	250	–	229/21	–	110/100	5/201/43/1
GSE16446	HG-U133_PLUS_2	21,422,418	2011	50	–	27/23	–/9/36	24/26	38/2/2/8
GSE31519	HG-U133A	22,220,191	2011	29	–	9/20	–/–/21	19/9	18/–/9/2
GSE37946	HG-U133A	22,460,789	2012	26	47.88 ± 12.3	16/10	–	–	3/1/16/6
GSE45255	HG-U133A	23,618,380	2013	57	53.12 ± 11.1	37/20	5/23/28	31/26	7/16/31/3
GSE65194	HG-U133_PLUS_2	25,848,952	2015	46	–	41/5	–	–	13/13/11/9
Total				1,329	54.92 ± 11.94	978/351	59/193/305	464/476	178/687/399/65

agents include anthracyclines ($n = 383$) and taxanes ($n = 237$). Smaller groups of patients were treated with trastuzumab ($n = 50$) and CMF ($n = 66$). Treatment cohorts with less than 50 patients were excluded from further analysis. A summary of the different treatment regimens, as well as the proportion of patients in each of these cohorts, is presented in Figure 2.

Server setup

We have established a webpage for automated analysis of future biomarker candidates. The PHP-based homepage runs an R server in the background and enables mining the database via Mann–Whitney or ROC analysis (Fig. 3*a*) using either the pathological response data or RFS at 5 years (Fig. 3*b*). Clinical variables (grade, nodal status, receptor status and molecular subtype) are implemented as filters when selecting any combination of these, then only samples with available information for each parameter are included in the analysis.

Samples collected in a clinical trial include the datasets GSE16446 (clinical trial ID: NCT00017095, NCT00336791), GSE41998 (NCT00455533), GSE50948 (ISRCTN86043495), GSE66305 (NCT00429299) and GSE16391 (NCT00004205). An additional filter enables to run the analysis using these patients only ($n = 628$ for the pCR dataset).

ROC also gives a numerical representation of the classifier performance when providing the "area under the curve" (AUC) value. An AUC of 0.5 corresponds to no classification power at all, and an AUC value of 1 denotes a perfect biomarker. In addition to a *p*-value and an AUC value, the ROC analysis also enables researchers to determine the strongest cutoff capable of best discriminating between responder and nonresponder patients (Fig. 3*c*). In case, the user enters multiple genes, then false discovery rate (FDR) is computed for each of the genes and a table is displayed showing the results at the FDR cutoffs of 20, 10 and 5%. The page is registration-free and can be accessed at www.rocplot.org.

Validation analyses

First, we analyzed a set of established biomarkers, including TP53 for chemotherapy in Luminal breast cancer ($p = 1.01\text{E-}19$, AUC = 0.769), HER2 for trastuzumab therapy ($p = 8.4\text{E-}04$, AUC = 0.629) and PGR for hormonal therapy ($p = 8.6\text{E-}05$, AUC = 0.7).

Second, we validated a set of paclitaxel-resistance markers. In this step, each of the biomarker candidates was checked in the pCR cohort. Of the 29 total genes, 23 genes reached significance (Table 2). We have uncovered 16 significant genes in Basal, 15 genes in Luminal A, 8 genes in Luminal B and 5 genes in HER2 + ER− subgroups.

The best-performing genes in the Basal samples were BCL2 modifying factor (BMF; $p = 0.023$, AUC = 0.688), B cell receptor-associated protein 29 (BCAP29; $p = 0.028$,

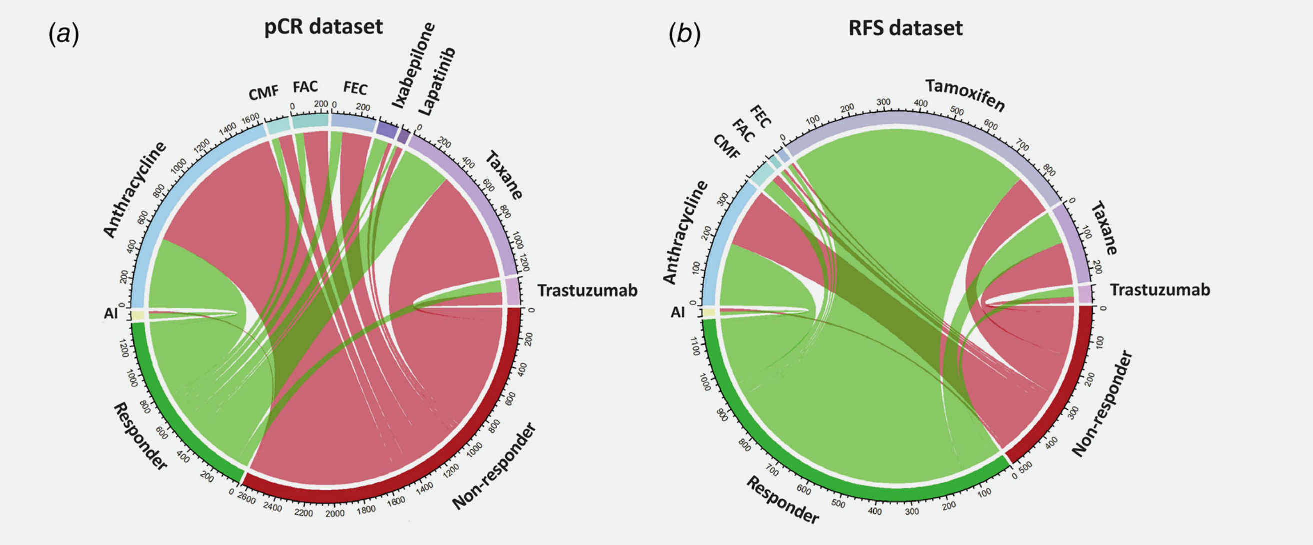

Figure 2. Treatment and response data. The circos plots summarize patient distribution for samples included in the pathological response dataset (pCR dataset—*a*) and in the relapse-free survival-based dataset (RFS dataset—*b*). The width of the connecting lines is proportional to the relative number of patients. [Color figure can be viewed at wileyonlinelibrary.com]

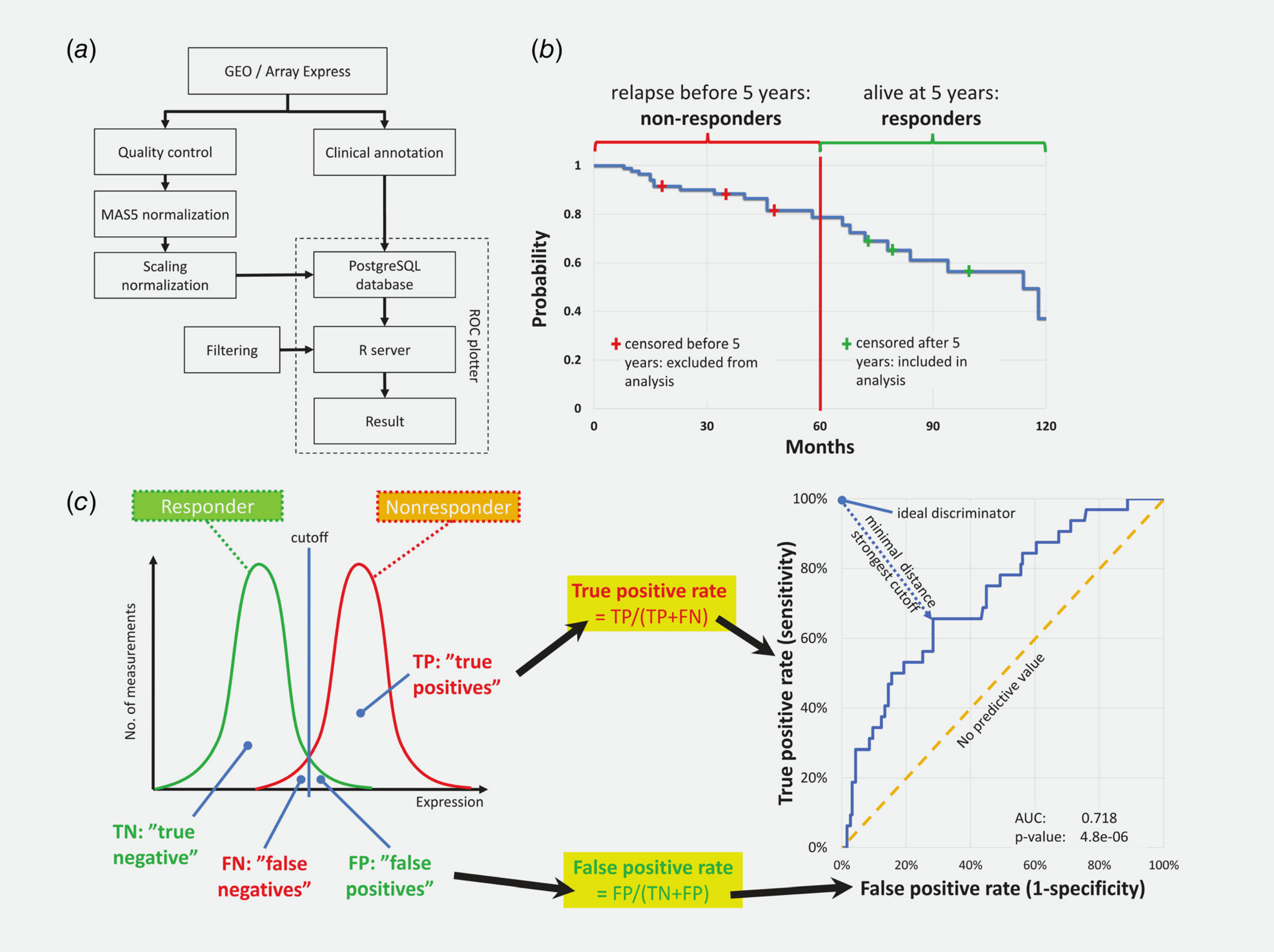

Figure 3. Overview of the statistical analysis. The applied computational pipeline (*a*), the designation of patients into responder and nonresponder cohorts using relapse-free survival at 5 years (*b*) and a quick guide for the interpretation of the ROC analysis results (*c*). [Color figure can be viewed at wileyonlinelibrary.com]

Table 2. Of 29 previously published paclitaxel resistance-related genes 21 were significant when evaluated in the different molecular subtypes in the pathological complete response dataset

			Basal			Luminal A			Luminal B			HER2+ ER-		
Affymetrix ID	Symbol	Approved name	*n*	AUC	ROC *p*-value	*n*	AUC	ROC *p*-value	*n*	AUC	ROC *p*-value	*n*	AUC	ROC *p*-value
209993_at	ABCB1	ATP binding cassette subfamily B member 1	**256**	**0.582**	**0.024**	302	0.511	0.796	378	0.545	0.208	**123**	**0.652**	**3.73E-03**
208288_at	ABCB11	ATP binding cassette subfamily B member 11	**256**	**0.647**	**5.81E-05**	**302**	**0.594**	**0.025**	378	0.517	0.645	123	0.592	0.078
213485_s_at	ABCC10	ATP binding cassette subfamily C member 10	256	0.516	0.669	302	0.576	0.069	**378**	**0.595**	**7.91E-03**	123	0.505	0.927
211692_s_at	BBC3	BCL2 binding component 3	**256**	**0.609**	**2.88E-03**	**302**	**0.618**	**4.95E-03**	378	0.565	0.067	**123**	**0.615**	**0.028**
230150_at	BCAP29	B cell receptor associated protein 29	**57**	**0.682**	**0.028**	103	0.693	0.056	**98**	**0.696**	**4.04E-03**	**74**	**0.643**	**0.039**
207005_s_at	BCL2	BCL2, apoptosis regulator	**256**	**0.586**	**0.018**	**302**	**0.799**	**9.61E-13**	378	0.538	0.291	123	0.571	0.173
215037_s_at	BCL2L1	BCL2 like 1	**256**	**0.615**	**1.56E-03**	**302**	**0.8**	**8.21E-13**	378	0.506	0.866	123	0.584	0.109
226530_at	BMF	Bcl2 modifying factor	**57**	**0.688**	**0.023**	103	0.661	0.112	98	0.563	0.357	74	0.599	0.152
207261_at	CNGA3	Cyclic nucleotide gated channel alpha 3	**256**	**0.611**	**2.37E-03**	**302**	**0.754**	**1.41E-09**	378	0.548	0.180	**123**	**0.668**	**1.41E-03**
208147_s_at	CYP2C8	Cytochrome P450 family 2 subfamily C member 8	**256**	**0.587**	**0.018**	**302**	**0.614**	**6.75E-03**	**378**	**0.604**	**3.59E-03**	123	0.589	0.089
205998_x_at	CYP3A4	Cytochrome P450 family 3 subfamily A member 4	**256**	**0.62**	**1.03E-03**	**302**	**0.675**	**3.19E-05**	378	0.529	0.425	123	0.597	0.064
212464_s_at	FN1	Fibronectin 1	256	0.544	0.228	**302**	**0.692**	**4.84E-06**	378	0.504	0.902	123	0.591	0.082
202270_at	GBP1	Guanylate binding protein 1	**256**	**0.621**	**9.12E-04**	**302**	**0.789**	**6.06E-12**	**378**	**0.6**	**5.35E-03**	123	0.535	0.506
225540_at	MAP2	Microtubule-associated protein 2	57	0.622	0.142	103	0.599	0.329	**98**	**0.644**	**0.034**	74	0.585	0.217
200836_s_at	MAP4	Microtubule-associated protein 4	256	0.541	0.258	**302**	**0.654**	**2.50E-04**	378	0.511	0.769	123	0.532	0.541
209636_at	NFKB2	Nuclear factor kappa B subunit 2	**256**	**0.587**	**0.018**	302	0.568	0.104	378	0.507	0.846	123	0.579	0.131
207202_s_at	NR1I2	Nuclear receptor subfamily 1 group I member 2	**256**	**0.613**	**1.95E-03**	**302**	**0.661**	**1.27E-04**	378	0.529	0.418	**123**	**0.622**	**0.020**
229944_at	OPRK1	Opioid receptor kappa 1	57	0.640	0.092	103	0.600	0.321	**98**	**0.646**	**0.033**	74	0.557	0.413
206354_at	SLCO1B3	Solute carrier organic anion transporter family member 1B3	**256**	**0.631**	**3.30E-04**	**302**	**0.732**	**3.06E-08**	378	0.522	0.533	123	0.595	0.071
204215_at	TMEM243	Transmembrane protein 243	**256**	**0.595**	**9.12E-03**	**302**	**0.697**	**2.68E-06**	**378**	**0.581**	**0.024**	123	0.551	0.332
230690_at	TUBB1	Tubulin beta 1 class VI	57	0.536	0.662	103	0.535	0.716	**98**	**0.646**	**0.032**	74	0.612	0.105
212664_at	TUBB4	Tubulin, beta 4 class IV	**256**	**0.613**	**1.95E-03**	**302**	**0.72**	**1.59E-07**	378	0.522	0.531	123	0.573	0.162
213943_at	TWIST1	Twist family bHLH transcription factor 1	256	0.528	0.445	**302**	**0.65**	**3.60E-04**	378	0.545	0.205	123	0.545	0.388

Significant genes are marked by bold text.

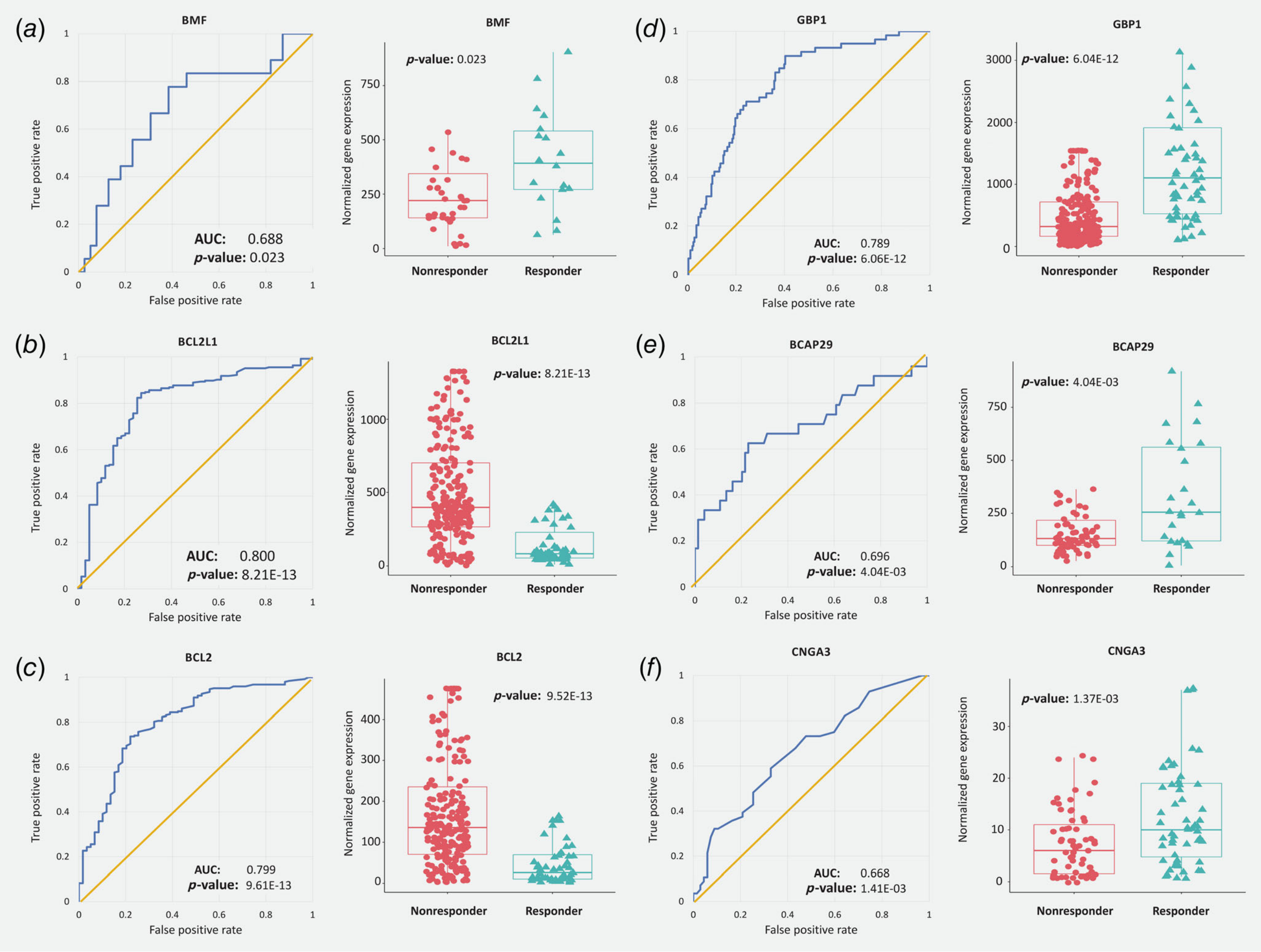

Figure 4. ROC curves and box-plots of top genes validated for Paclitaxel resistance by molecular subtype: BMF in Basal (*a*), BCL2, BCL2L1 and GBP1 in Luminal A (*b*–*d*), BCAP29 in Luminal B (*e*) and CNGA3 in HER2-enriched (*f*). [Color figure can be viewed at wileyonlinelibrary.com]

AUC = 0.682) and ATP binding cassette subfamily B member 11 (ABCB11; *p* = 5.8E-05, AUC = 0.647). In Luminal A tumors, the most significant genes include BCL2 apoptosis regulator (BCL2; *p* = 9.6E-13, AUC = 0.799), BCL2 like 1 (BCL2L1; *p* = 8.2E-13, AUC = 0.8) and guanylate binding protein 1 (GBP1; *p* = 6.1E-12, AUC = 0.789). In Luminal B, the strongest genes include BCAP29 (*p* = 4.04E-03, AUC = 0.695), tubulin beta 1 class VI (TUBB1; *p* = 0.032, AUC = 0.646) and

Table 3. Top 10 new biomarker candidates of chemotherapy response in the TNBC subtype (*n* = 473)

Affymetrix ID	Symbol	Approved name	AUC	ROC *p*-value
200959_at	FUS	FUS RNA binding protein	0.699	3.30E-16
203276_at	LMNB1	Lamin B1	0.698	2.20E-16
215905_s_at	SNRNP40	Small nuclear ribonucleoprotein U5 subunit 40	0.694	4.20E15
202416_at	DNAJC7	DnaJ heat shock protein family (Hsp40) member C7	0.693	4.20E-15
218733_at	MSL2	MSL complex subunit 2	0.678	7.70E-13
200773_x_at	PTMA	Prothymosin alpha	0.677	5.60E-13
204415_at	IFI6	Interferon alpha inducible protein 6	0.676	7.80E-13
40850_at	FKBP8	FKBP prolyl isomerase 8	0.675	2.30E-12
204166_at	SBNO2	Strawberry notch homolog 2	0.674	3.20E-12
202785_at	NDUFA7	NADH ubiquinone oxidoreductase subunit A7	0.671	5.80E-12

The analysis was performed across all genes in the pCR cohort.

opioid receptor kappa 1 (OPRK1; $p = 0.033$, AUC = 0.646). In HER2-positive samples, the genes with the highest correlation to resistance include cyclic nucleotide gated channel alpha 3 (CNGA3; $p = 1.4\text{E-}03$, AUC = 0.668), the classical multiple drug resistance (MDR) gene ATP binding cassette subfamily B member 1 (ABCB1; $p = 3.7\text{E-}03$, AUC = 0.652) and BCAP29 ($p = 0.039$, AUC = 0.643). The ROC plots and the mean plots for the best performing genes are presented in Figure 4.

Finally, we run the analysis for all available samples designated as TNBC ($n = 473$). This exploration was performed in the pCR cohort. The analysis was performed across all genes and the 10 strongest new biomarker candidates are presented in Table 3.

Discussion

In our study, we had two major aims: first, to combine available datasets to establish a sufficiently large breast cancer cohort with transcriptomic and clinical response data, including information regarding the systemic anticancer therapy administered to these patients. Second, we aimed to establish a framework capable of validating and ranking new candidate genes by mining this database. We also performed a validation analysis for established biomarkers to corroborate the reliability of our approach.

First, we have executed the pipeline for the most widely used clinical biomarkers, including HER2 and PGR. In each setting, only patients who actually received targeted therapy (for HER2) and hormonal therapy (for PGR) were included. For HER2, the results confirmed the predictive role of HER2 expression for RFS. Notably, most of the HER2-positive patients included in the database did not receive anti-HER2 therapy. This points to the still limited accessibility of samples with anti-HER2 therapy. Furthermore, the proportion of patients included in these therapies is also limited due to delayed administration, even in developed countries such as the US.[21] We have to note that the proportion of patients designated as HER2-positive also depends on the used cutoff. Current ASCO recommendations are in favor of lowering the cutoff percentage for ESR1 positivity,[22] and such a trend might also be feasible for HER2 status determination. However, the limited number of patients in the trastuzumab- and lapatinib-treated cohorts did not enable us to investigate this hypothesis.

Progesterone receptor (PGR) is an estrogen-regulated gene,[23] and its used to support the selection of patients for hormonal therapy has been questioned several times. A study of more than 155,000 women from the SEER registry has uncovered a declining trend of estrogen-negative, PGR-positive patients, possibly hinting at an improvement in receptor diagnosis accuracy.[24] In the United Kingdom, the National Institute for Health and Clinical Excellence recommendations do not have included PGR since 2009.[25] In our pipeline, one of the reasons for setting the cutoff to 60 months to discriminate responder and nonresponder patients was the StGallen/NCCN-recommended 5-year length of hormonal therapy.[26] Thus, the used cutoff enables us to identify those who progress during these initial 5 years. As PGR delivered a high significance in our analysis, our results support the continued utilization of PGR to select those who are eligible for hormonal therapy.

We also used the pipeline to validate previously published paclitaxel resistance biomarker candidates. In this analysis, multiple transport genes were upregulated including ABCB1, ABCB11, CNGA3 and SLCO183. ABCB1 (also called MDR1 or PGP) is one of the most widely investigated genes linked to multidrug resistance. ABCB11 encodes a sister gene of PGP, and ABCB11 transfectants display resistance against Taxol but no other chemotherapy agents.[27] Solute carrier organic anion transporter family member 1B3 (SLCO1B3) encodes a member of the organic anion transporter family. This gene's overexpression confers an antiapoptotic advantage against chemotherapy treatments by blocking the transcription of TP53.[28] Lessening the intracellularly available drug molecules is also the mechanism of action of the metabolic enzymes cytochrome P450 family 2 subfamily C member 8 (CYP2C8) and cytochrome P450 family 3 subfamily A member 4 (CYP3A4), both of which display higher expression in the resistant patients and have previously been linked to paclitaxel resistance.[29,30]

Taxenes disrupt microtubule function and genes involved in microtubule setup and assembly, including TUBB1, TUBB4, microtubule-associated protein 2 (MAP2)[31] and microtubule-associated protein 4 (MAP4),[32] therefore have a critical role in resistance. TUBB1 was recently described as the gene with the most frequently altered and amplified isoforms in breast cancer.[33] TUBB4 had higher expression in an MCF7 cell line engineered to withstand paclitaxel treatment by administering gradually increasing concentrations of the drug.[34] Ultimately, the intent of any chemotherapy agent is to send the damaged cancer cell into apoptosis. Thus, genes involved in the cell cycle (fibronectin 1 [FN1], twist family bHLH transcription factor 1 [TWIST1] and GBP1) and apoptosis (BCL2, BCL2L1, BCAP, BMF and BCL2 binding component 3 [BBC3]) play critical roles in resistance against these agents. We can confirm previous *in vitro* observations showing lower expression of TWIST1 and FN1 in relation to paclitaxel resistance.[35] The BCL2 and BCL2L1 (BCL-xL) genes reached high significance in both Basal and Luminal A samples. Overexpression of BCL2 was recently linked to paclitaxel resistance in cell lines.[36] Tumors with higher BCL2L1 expression had shorter RFS times.[37] Previously, knockdown of BBC3 (PUMA) reduced paclitaxel-induced apoptosis in T47D cells.[38] Overall, in our study, we can confirm multiple previous *in vitro* observations linking different genes to paclitaxel resistance; we have summarized the mechanisms for the significant genes in Figure 5.

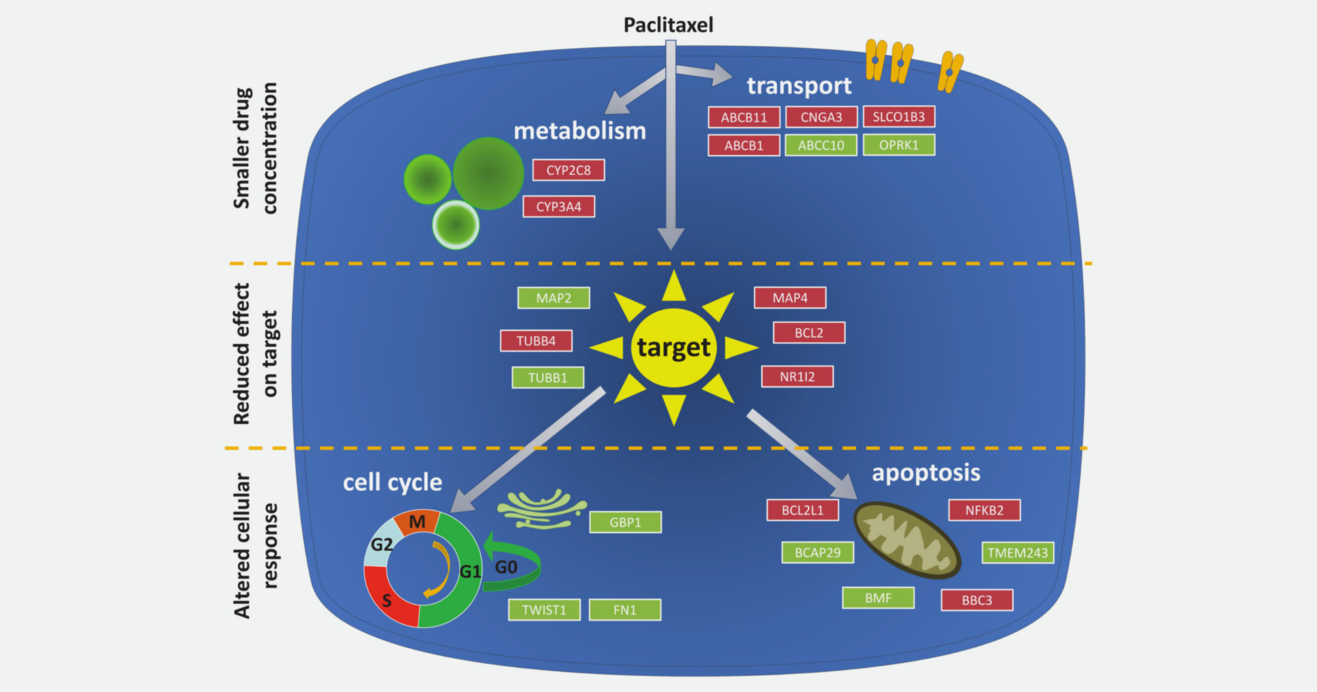

Figure 5. Summary of the biological functions of the validated paclitaxel resistance-associated genes. Genes highlighted with red were upregulated, whereas genes highlighted with green were downregulated in resistant patient samples. [Color figure can be viewed at wileyonlinelibrary.com]

We have to mention some limitations of our analysis. First, most patients included in the database have received multiple agents. This finding makes it more complicated to link a given gene to a response against a selected agent. At the same time, today, almost no patients receive a monotherapy. Unfortunately, this limits the chance of unearthing a large-scale dataset with monotherapy in the near future.

A second limitation is the rather limited number of patients in some of the treatment arms. We plan to extend the database as new studies are published; thus, we will be able to gradually increase the validation power of the analysis tool. Similarly, we can expect large RNA-seq datasets to be published in the near future. Unfortunately, the Cancer Genome Atlas, the largest cohort published until today, has neither pathological response data nor RFS data.

A third limitation is the different quality of the included studies. Although each study fulfilled the MIAME criteria and the array quality control, this only focus on the technical issues. Clinical trials provide high-quality patient records—and here we took account of five datasets which used patients from different clinical trials. An additional filter was built into the online platform to enable the user to use exclusively these samples in the analysis.

In summary, we established a large transcriptomic database that includes treatment data and expression data of more than 20,000 genes from 3,104 samples. We used pathological response data or RFS time at 5 years to assign the patients into response cohorts. We demonstrated the robustness of the analysis pipeline by cross-validating an established set of resistance-associated genes. The online platform at www.rocplot.org provides an easily accessed resource for researchers to mine the database and to validate and rank future biomarker candidates.

Acknowledgements

The use of the computational infrastructure of Pázmány Péter University, provided within the National Bionics Program, is gratefully acknowledged.

References

1. Khleif SN, Doroshow JH, Hait WN, et al. AACR-FDA-NCI cancer biomarkers collaborative consensus report: advancing the use of biomarkers in cancer drug development. *Clin Cancer Res* 2010; 16:3299–318.
2. Gyorffy B, Hatzis C, Sanft T, et al. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res* 2015; 17:11.
3. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* 1992;339:71–85.
4. Hortobagyi GN. Trastuzumab in the treatment of breast cancer. *N Engl J Med* 2005; 353:1734–6.
5. Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer* 2002;2:48–58.
6. van Kuilenburg AB. Dihydropyrimidine dehydrogenase and the efficacy and toxicity of 5-fluorouracil. *Eur J Cancer* 2004;40:939–50.
7. Pommier Y. Topoisomerase I inhibitors: camptothecins and beyond. *Nat Rev Cancer* 2006;6: 789–802.

8. Martin LP, Hamilton TC, Schilder RJ. Platinum resistance: the role of DNA repair pathways. *Clin Cancer Res* 2008;14:1291–5.
9. Abdul-Ghani R, Serra V, Gyorffy B, et al. The PI3K inhibitor LY294002 blocks drug export from resistant colon carcinoma cells overexpressing MRP1. *Oncogene* 2006;25:1743–52.
10. Gyorffy B, Lanczky A, Eklund AC, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* 2010;123:725–31.
11. Goswami CP, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinformatics* 2013;3:22.
12. Jezequel P, Frenel JS, Campion L, et al. Bc-GenExMiner 3.0: new mining module computes breast cancer gene expression correlation analyses. *Database* 2013;2013:bas060.
13. Kim SK, Hwan Kim J, Yun SJ, et al. APPEX: analysis platform for the identification of prognostic gene expression signatures in cancer. *Bioinformatics* 2014;30:3284–6.
14. Chen X, Miao Z, Divate M, et al. KM-express: an integrated online patient survival and gene expression analysis tool for the identification and functional characterization of prognostic markers in breast and prostate cancers. *Database* 2018;2018.
15. Antonov AV, Krestyaninova M, Knight RA, et al. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene* 2014;33: 1621–8.
16. Lanczky A, Nagy A, Bottai G, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res Treat* 2016;160: 439–46.
17. Madden SF, Clarke C, Gaule P, et al. BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. *Breast Cancer Res* 2013;15:R52.
18. Gyorffy B, Benke Z, Lanczky A, et al. RecurrenceOnline: an online analysis tool to determine breast cancer recurrence and hormone receptor status using microarray data. *Breast Cancer Res Treat* 2012;132:1025–34.
19. Li Q, Birkbak NJ, Gyorffy B, et al. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics* 2011;12:474.
20. Dorman SN, Baranova K, Knoll JH, et al. Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol Oncol* 2016;10:85–100.
21. Gallagher CM, More K, Kamath T, et al. Delay in initiation of adjuvant trastuzumab therapy leads to decreased overall survival and relapse-free survival in patients with HER2-positive non-metastatic breast cancer. *Breast Cancer Res Treat* 2016;157:145–56.
22. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 2010;28:2784–95.
23. Kastner P, Krust A, Turcotte B, et al. Two distinct estrogen-regulated promoters generate transcripts encoding the two functionally different human progesterone receptor forms a and B. *EMBO J* 1990;9:1603–14.
24. Dunnwald LK, Rossing MA, Li CI. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Res* 2007;9:R6.
25. Yarnold J. Early and locally advanced breast cancer: diagnosis and treatment National Institute for health and clinical excellence guideline 2009. *Clin Oncol* 2009;21:159–60.
26. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen international expert consensus conference on the primary therapy of early breast cancer 2017. *Ann Oncol* 2153;2018:29.
27. Childs S, Yeh RL, Hui D, et al. Taxol resistance mediated by transfection of the liver-specific sister gene of P-glycoprotein. *Cancer Res* 1998;58: 4160–7.
28. Lee W, Belkhiri A, Lockhart AC, et al. Overexpression of OATP1B3 confers apoptotic resistance in colon cancer. *Cancer Res* 2008;68: 10315–23.
29. Garcia-Martin E, Pizarro RM, Martinez C, et al. Acquired resistance to the anticancer drug paclitaxel is associated with induction of cytochrome P450 2C8. *Pharmacogenomics* 2006;7:575–85.
30. Harmsen S, Meijerman I, Beijnen JH, et al. Nuclear receptor mediated induction of cytochrome P450 3A4 by anticancer drugs: a key role for the pregnane X receptor. *Cancer Chemother Pharmacol* 2009;64:35–43.
31. Zheng S, Shi L, Zhang Y, et al. Expression of SNCG, MAP2, SDF-1 and CXCR4 in gastric adenocarcinoma and their clinical significance. *Int J Clin Exp Pathol* 2014;7:6606–15.
32. McGrogan BT, Gilmartin B, Carney DN, et al. Taxanes, microtubules and chemoresistant breast cancer. *Biochim Biophys Acta* 2008;1785:96–132.
33. Nami B, Wang Z. Genetics and expression profile of the tubulin gene superfamily in breast cancer subtypes and its relation to Taxane resistance. *Cancer* 2018;10:E274.
34. Banerjee A. Increased levels of tyrosinated alpha-, beta(III)-, and beta(IV)-tubulin isotypes in paclitaxel-resistant MCF-7 breast cancer cells. *Biochem Biophys Res Commun* 2002;293:598–601.
35. Duan Z, Lamendola DE, Duan Y, et al. Description of paclitaxel resistance-associated genes in ovarian and breast cancer cell lines. *Cancer Chemother Pharmacol* 2005;55:277–85.
36. Duran GE, Wang YC, Moisan F, et al. Decreased levels of baseline and drug-induced tubulin polymerisation are hallmarks of resistance to taxanes in ovarian cancer cells and are associated with epithelial-to-mesenchymal transition. *Br J Cancer* 2017;116:1318–28.
37. Flores ML, Castilla C, Avila R, et al. Paclitaxel sensitivity of breast cancer cells requires efficient mitotic arrest and disruption of Bcl-xL/Bak interaction. *Breast Cancer Res Treat* 2012;133: 917–28.
38. Kutuk O, Letai A. Displacement of Bim by Bmf and Puma rather than increase in Bim level mediates paclitaxel-induced apoptosis in breast cancer cells. *Cell Death Differ* 2010;17:1624–35.