Breast Cancer Res Treat (2012) 131:765–775
DOI 10.1007/s10549-011-1457-7

PRECLINICAL STUDY

bc-GenExMiner: an easy-to-use online platform for gene prognostic analyses in breast cancer

Pascal Jézéquel · Mario Campone · Wilfried Gouraud · Catherine Guérin-Charbonnel · Christophe Leux · Gabriel Ricolleau · Loïc Campion

Received: 2 February 2011 / Accepted: 14 March 2011 / Published online: 31 March 2011

Abstract Gene prognostic meta-analyses should benefit from breast tumour genomic data obtained during the last decade. The aim was to develop a user-friendly, web-based application, based on DNA microarrays results, called "breast cancer Gene-Expression Miner" (bc-GenExMiner) to improve gene prognostic analysis performance by using the same bioinformatics process. bc-GenExMiner was developed as a web-based tool including a MySQL relational database. Survival analyses are performed with R statistical software and packages. Molecular subtyping was performed by means of three single sample predictors (SSPs) and three subtype clustering models (SCMs). Twenty-one public data sets have been included. Among the 3,414 recovered breast cancer patients, 1,209 experienced a pejorative event. Molecular subtyping by means of three SSPs and three SCMs was performed for 3,063 patients. Furthermore, three robust lists of stable subtyped patients were built to maximize reliability of molecular assignment. Gene prognostic analyses are done by means of univariate Cox proportional hazards model and may be conducted on cohorts split by nodal (N), oestrogen receptor (ER), or molecular subtype status. To evaluate independent prognostic impact of genes relative to Nottingham Prognostic Index and Adjuvant! Online, adjusted Cox proportional hazards models are performed. bc-GenExMiner allows researchers without specific computation skills to easily and quickly evaluate the in vivo prognostic role of genes in breast cancer by means of Cox proportional hazards model on large pooled cohorts, which may be split according to different prognostic parameters: N, ER, and molecular subtype. Prognostic analyses by molecular subtype may also be performed in three robust molecular subtype classifications.

Electronic supplementary material The online version of this article (doi:10.1007/s10549-011-1457-7) contains supplementary material, which is available to authorized users.

P. Jézéquel · W. Gouraud · C. Guérin-Charbonnel
Unité Mixte de Génomique du Cancer, Hôpital Laënnec, Bd J. Monod, 44805 Nantes-Saint Herblain Cedex, France

P. Jézéquel (✉) · G. Ricolleau
Département de Biologie Oncologique, Centre de Lutte Contre le Cancer René Gauducheau, Bd J. Monod, 44805 Nantes-Saint Herblain Cedex, France
e-mail: p-jezequel@nantes.fnclcc.fr

M. Campone
Service d'Oncologie Médicale, Centre de Lutte Contre le Cancer René Gauducheau, Bd J. Monod, 44805 Nantes-Saint Herblain Cedex, France

M. Campone · W. Gouraud · C. Guérin-Charbonnel · L. Campion
INSERM U892, IRT-UN, 8 Quai Moncousu, 44007 Nantes Cedex, France

W. Gouraud · C. Guérin-Charbonnel · L. Campion
Unité de Biostatistique, Centre de Lutte Contre le Cancer René Gauducheau, Bd J. Monod, 44805 Nantes-Saint Herblain Cedex, France

C. Leux
Service d'épidémiologie et de Biostatistiques, Pôle d'information Médicale, d'évaluation et de Santé Publique, Hôpital Saint Jacques, CHU Nantes, 85 rue Saint Jacques, 44093 Nantes, France

Keywords Prognostic analysis · Breast cancer · Genomic data · Molecular subtype · Web tool

Abbreviations

AE	Any event
AOL	Adjuvant! Online
AR	Any relapse
D	Death
ER	Oestrogen receptor
GEO	Gene expression omnibus
GES	Gene-expression signature
IHC	Immunohistochemistry
MR	Metastatic relapse
MRD	Metastatic relapse or death
MSP	Molecular subtype predictor
N	Nodal
NPI	Nottingham prognostic index
RMSPC	Robust molecular subtype predictor classification
RSCMC	Robust subtype clustering model classification
RSSPC	Robust single sample predictor classification
SCM	Subtype clustering model
SSP	Single sample predictor

Introduction

Since 2000, transcriptome of breast cancer tumours has been explored by means of DNA microarrays. These studies and the ones that followed mostly aimed at identifying subtype, prognostic and predictive gene-expression signatures (GES). Once GES had been found, the data sets, if used again, often served for external validation of new genomic studies, or were merged in large "in silico" studies, which aimed at identifying new GES. But this wealth of annotated genomic values might be exploited in different ways than the projects' initial purposes, and notably in a more basic perspective. Concerning single genes, prognostic meta-analyses might be conducted based on breast cancer patient genomic data. In addition, prognostic informativity of genes could be evaluated in pooled subtyped cohorts (e.g., nodal status [N], oestrogen receptor [ER] status, molecular subtypes: luminal A, luminal B, basal-like, HER2+, normal breast-like) [1, 2]. Meta-analyses are designed to overcome the low sample size typical to microarray experiments and yield more valid and informative results than each experiment separately. Large numbers of data and independent data sets, which represent various preanalytical (tumour dissection, freezing delay, storage conditions) and analytical (different cohorts and genomic protocols [RNA isolation, probe preparation and labelling, hybridisation, microarray platforms]) conditions, would greatly improve the robustness of such prognostic analyses performed with the same bioinformatics process [3]. Hence, this kind of data-mining should quickly give an answer to researchers about in vivo prognostic informativity of a gene of interest in breast cancer. In other words, in silico evaluation could greatly help scientists in their quest for biomarkers. For example, a bench finding or an intuitive hypothesis could be reinforced or could find a rapid beginning of response by using such a tool. However, in the last case, the results would have to be bench-validated by means of quantitative reverse-transcription PCR. The aim was to develop a user-friendly, web-based application called "breast cancer Gene-Expression Miner" (bc-GenExMiner) (http://bcgenex.centregauducheau.fr) to facilitate the mining of published breast cancer annotated genomic data. This current version, 2.0, easily and quickly allows researchers without specific computation skills to evaluate the in vivo prognostic informativity of genes of interest in breast cancer by means of Cox regression model. Furthermore, bc-GenExMiner 2.0 encompassed molecular subtype assignment of 3,063 breast cancer patients included in the database. This study is original in the way that it presents results of molecular subtype determination based on three single sample predictors (SSPs) and three subtype clustering models (SCMs) for a large cohort of patients, and proposes robust lists of molecular subtyped patients to increase performance of molecular assignment and downstream gene prognostic analyses.

Methods

System implementation

bc-GenExMiner is a web-based tool powered by Apache with a MySQL relational database storage. Dynamic web interfaces were written in PHP v5 and JavaScript. The website requires a HTML 4.0-compliant browser with JavaScript enabled but does not require any particular visual plug-in tool. Statistical analyses are performed with the R statistical software (v2.9.2) and packages: rmeta (v2.14) and survival (v2.35-4) [4–6].

Data selection

We looked for publicly available breast cancer gene expression data sets with clinical information, including prognosis, in repositories such as Gene Expression Omnibus (GEO), ArrayExpress and Stanford microarray database, author's individual web pages and in articles, selecting those with a medium to large sample size [7–9]. Patients who received neoadjuvant chemotherapy and microdissected samples were not included. Treatment was not retained as an inclusion criterion for bc-GenExMiner prognostic analyses because it was rarely mentioned in the

different studies and very few cohorts had the same therapeutic process.

Once a list of studies satisfying the criteria was obtained, patients included in those studies were compared in order to avoid duplicates. In case of similarities, the data set that included fewer patients was rejected. In order to keep the data up-to-date and provide a powerful tool, gene symbol annotations have been updated before the launches of bc-GenExMiner v1.0, v1.1, and v2.0 and will be updated every 6 months. Affymetrix chips were updated with tables downloadable on manufacturer's web site. Other chips were checked manually using NCBI Reference Sequence (RefSeq), Ensembl Transcript ID or Entrez GeneID, according to what was provided when the array was deposited. SOURCE (http://source.stanford.edu/) and Ensembl (http://www.ensembl.org/) were used for request of new annotations.

We also retrieved a genomic study including HER2 immunohistochemistry (IHC) data to evaluate HER2 molecular subtype assignment [10].

Data-pre-processing

Because the tool was designed for users without special skills in bioinformatics, we proposed "ready-to-use" data. Before being log2-transformed, non-Affymetrix platform data were ratio-normalised and Affymetrix raw CEL data were MAS5-normalised in the Affymetrix Expression Console (v1.1.1) (see supplementary data for details). Finally, in order to merge all studies data and create pooled cohorts, we converted studies data to a common scale (median equal to 0 and standard deviation equal to 1) [11].

Molecular subtyping

bc-GenExMiner proposes prognostic analyses in breast cancer molecular subtypes: luminal A, luminal B, basal-like, HER2+, and normal breast-like. Many molecular subtype predictors (MSPs) exist but assignment of patient to a particular subtype may be dependent on MSP used [12]. To take into account these discrepancies, and because to date no gold standard MSP has been defined, we decided to propose and perform analyses based on six different MSPs: three SSPs and three SCMs. The three SSPs (500 gene centroids by Sorlie and colleagues, 306 gene centroids by Hu and coworkers, and 50 gene centroids by Parker and colleagues) were used as described by Weigelt et al. [12–15]. The three SCMs were computed with the function "subtype.cluster" of the R package genefu (v.1.0.9), which fits the SCM as published in Desmedt et al. and Wirapati et al. studies [16–18]. Details about MSPs computation are provided in supplementary data. To overlay the MSP gene lists and the ones of the different cohorts chips, gene symbols were used [12]. When multiple probes corresponded to the same gene symbol, their median value was taken. All molecular subtypes were determined based on data from the database, i.e., after centring and scaling.

Statistical tests

Targeted analysis

A targeted analysis performs gene-based survival analyses for each cohort separately and all cohorts pooled together. The bc-GenExMiner user chooses one to ten different genes (actualised gene symbol or Affymetrix probeset ID) to analyse and selects population (N and ER statuses) and prognostic event criteria. When the analysis is launched, the prognostic impact of each chosen gene is evaluated by means of univariate Cox proportional hazards model, and a forest plot and Kaplan–Meier curves (for the median-split pool) are performed. Furthermore, to evaluate independent prognostic impact of genes relative to the well-established breast cancer prognostic indexes, Nottingham Prognostic Index- (NPI)- and Adjuvant! Online (AOL)-adjusted Cox proportional hazards models are systematically performed [19, 20].

When analysing all cohorts pooled together, Cox proportional hazards models are stratified by cohort.

Exhaustive analysis

In an exhaustive analysis, the user chooses one gene and when the analysis is launched, univariate Cox proportional hazards model, stratified by cohort, is performed on each of the 45 possible pools corresponding to every combination of population (N [$n = 3$] and ER [$n = 3$] statuses) and prognostic event criteria ($n = 5$) to assess the prognostic impact of the chosen gene.

Analysis by molecular subtype

In a molecular subtype analysis, patients are pooled according to their molecular subtypes, based on three SSPs and three SCMs, and on three supplementary robust molecular subtype classifications consisting on the intersections of the 3 SSP and of the three SCM classifications: only patients with concordant molecular subtype assignment for the three SSPs (robust SSP classification [RSSPC]), for the three SCMs (robust SCM classification [RSCMC]), or for all MSPs (robust MSP classification [RMSPC]), are kept. The bc-GenExMiner user chooses one gene and univariate Cox proportional analysis, stratified by cohort, is performed for this gene for each of the different molecular subtypes populations. Kaplan–Meier curves are also computed.

An illustration of the gene expression according to molecular subtypes (as determined by the different

predictors [robust or not]) is displayed by means of gene expression maps. For robust classifications, a gene expression table is also given, indicating for each subtype the proportion of patient with low, intermediate, and high gene expression, gene expression values being beforehand split in order to form three equal groups.

Agreement studies

Cohen's kappa (κ) coefficient was used to measure agreement between molecular subtype assignments [21]. Interpretation of κ values is as follows: 0.01–0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, and 0.81–1.00 = almost-perfect agreement [22].

Results

Data content

All data sets were downloaded from publicly available websites. Twenty-one public data sets, encompassing 8 different microarray platforms (commercial and academic) and 12 DNA chips have been included in bc-GenExMiner v2.0 database (Supplementary Table s1). We excluded redundant studies. Update of 1325, 1359 and 2320 gene names was done for v1.0, v1.1, and v2.0, respectively. Clinicopathologic characteristics retained for bc-GenExMiner prognostic analyses were: N status, ER status, NPI, and AOL score. The smallest and the biggest cohorts included 58 and 401 patients, respectively. Among the 3,414 recovered non-redundant breast cancer patients, 1,209 experienced a pejorative event. Five kinds of disease evolution were listed from selected studies and used for further targeted and exhaustive prognostic analyses. These were metastatic relapse (MR), any relapse (AR) (first pejorative event represented by local, regional, or distant relapse), death (D), metastatic relapse or death (MRD), and any event (AE) (first pejorative event represented by any relapse or death). Molecular subtype prognostic analyses are performed with AE as pejorative event. A summary of clinicopathological data of the whole cohort is displayed in Table 1.

Table 1 Clinicopathologic characteristics of 3,414 breast cancer patients included in bc-GenExMiner v2.0

Variable	*n*
Age at diagnosis	
Median	57
Range	24–93
ND[a]	1292
Nodal status	
Positive	1112
Negative	1855
ND[a]	447
Oestrogen receptor status	
Positive	2169
Negative	625
ND[a]	620
NPI status	
1	371
2	577
3	182
ND[a]	2284
AOL status[b]	
Median	70
Range	9–99
ND[a]	2498
Event status	
No event	2191
Metastatic relapse	723
Any relapse[c]	1066
Death	520
Metastatic relapse or death	942
Any event[d]	1209
ND[a]	14

[a] Not determined, [b] 10-year probability of relapse (%), [c] first pejorative event represented by local, regional, or distant relapse, [d] first pejorative event represented by any relapse or death

Molecular subtyping

In order to determine molecular subtypes of the different studies' patients based on each MSP, MSP genes present on the different chips used in those studies were checked. Two studies were excluded because too few MSP genes (<70%) were present to reliably determine the patients molecular subtypes (Supplementary Table s2). For SCMGENE, one gene out of the three was missing for two studies. Nineteen studies, representing 3,063 breast cancer patients for SSPs, SCMOD1 and SCMOD2, and 3,027 for SCMGENE, were thus kept for further analyses.

In bc-GenExMiner selected population, molecular subtype distribution varies in function of MSP (Table 2; Fig. 1). This result was previously found by Weigelt et al. [12], whose study focused on the three SSPs applied on four unpooled breast cancer cohorts representing a total of 832 patients. In this study, percentage means of the three SSP distributions arranged the different subtypes in the decreasing order of frequency: 27% for luminal A, 18.5% for basal-like, 16.5% for luminal B, 14.4% for normal breast-like, 11.8% for HER2+, and 11.8% for unclassified patients, i.e., patients who are not correlated enough with any of the molecular suptype centroids to be assigned to a specific subtype. We showed that subtype assignment also

Table 2 Molecular subtyping of 3,063 breast cancer patients according to six molecular subtype predictors

MSP[a]	Basal-like		HER2+		Luminal A		Luminal B		Normal breast-like		Unclassified	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Sorlie's SSP[b]	450	14.7	352	11.5	903	29.5	363	11.9	409	13.4	586	19.1
Hu's SSP[b]	673	22.0	264	8.6	731	23.9	518	16.9	464	15.1	413	13.5
Parker's SSP[b]	578	18.9	476	15.5	840	27.4	633	20.7	447	14.6	89	2.9
RSSPC[c]	388	–	103	–	443	–	116	–	210	–	–	–
	ER−/HER2−		HER2+		ER+/HER2− low proliferation		ER+/HER2− high proliferation		Unclassified	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
SCMOD1[d]	499	16.3	673	22.0	873	28.5	883	28.8	135	4.4
SCMOD2[d]	565	18.4	362	11.8	918	30.0	928	30.3	290	9.5
SCMGENE[d]	903	29.8	478	15.8	789	26.0	798	26.4	59	1.9
RSCMC[e]	318	–	207	–	424	–	399	–	–	–
RMSPC[f]	274	–	76	–	194	–	53	–	–	–

[a] Molecular subtype predictor, [b] single sample predictor, [c] robust SSP classification based on patients classified in the same subtype with the three SSPs, [d] subtype clustering model, [e] robust SCM classification based on patients classified in the same subtype with the three SCMs, [f] robust MSP classification based on patients classified in the same subtype with the six MSPs

varies in function of SCM. Order of SCM subtype distribution was different from that of SSP: 28.5% for ER+/HER2− high proliferation, 28.2% for ER+/HER2− low proliferation, 21.5% for ER-/HER2−, 16.5% for HER2+, and 5.3% for unclassified patients.

As no gold standard MSP exists, bc-GenExMiner prognostic analyses are performed on the six molecular subtype distributions and on three supplementary ones represented by: RSSPC based on patients always classified in the same subtype with the three SSPs ($n = 1{,}260$), RSCMC based on patients always classified in the same subtype with the three SCMs ($n = 1{,}348$), and RMSPC based on patients always classified in the same subtype with the six MSPs ($n = 597$) (Table 2). Subtype specificity increases in such new classifications. Hence, we can advance that results obtained by means of RSSPC and RSCMC are more robust compared to the ones obtained with any of the three SSPs or SCMs, respectively. Biological relevance of this hypothesis is proven by the following points. (1) Statistical analysis confirmed that expression of proliferation gene *UBE2C* is lower in the robust part of luminal A subtype, which is considered as a low proliferative subtype (P [Wilcoxon] $< 10^{-4}$), than in non robust part of luminal A subtype (determined by patients classified as luminal A by a reference MSP but not classified in luminal A subtype by RMSPC), (2) The same positive result was obtained for negative ER measured by IHC and robust part of basal-like subtype compared with its non robust part (P [Fisher] $< 10^{-4}$), (3) And finally, agreement of HER2 status measured by IHC definitely confirmed the robust classification of breast cancer patients in RSSPC, RSCMC and RMSPC lists (Supplementary Table s3).

We observed a better agreement between MSPs belonging to the same MSP group (SSPs or SCMs) than between MSP belonging to different MSP groups (Supplementary Table s4). Overall, concordance between different MSPs for classification of breast cancers into molecular subtypes seems to be modest.

Only fair to moderate agreement ($\kappa = 0.37$–0.58) was observed between assignments made by SCMGENE, when unclassified and normal breast-like patients were excluded. The small number of genes ($n = 3$) included in this MSP might be responsible for weak agreement. For this reason, the following discussion will not take into account this MSP.

Substantial to almost-perfect concordance ($\kappa = 0.71$–0.84) was noted for classification of the basal-like subtype by all three SSPs and two SCMs (SCMOD1 and SCMOD2) (Supplementary Table s5). For assignment of HER2+, luminal A and luminal B subtype classes, agreement between distinct MSPs is moderate to substantial ($\kappa = 0.41$–0.73).

As described by Weigelt et al. [12], we observed that stability of molecular subtype classification by means of the three SSPs was inconsistent and was better for basal-like subtype. In this study, SCMs gave similar results.

Survival analyses demonstrated that subtype classification defines different prognostic groups as described in previous studies (Fig. 2). We observed a common pattern for the different curves, and a better separation between the different subtypes for RSSPC and RSCMC curves (Fig. 2a4, b4). As determined by means of SSPs or SCMs, the group with the best prognosis was luminal A, other subtypes (HER2+, basal-like, and luminal B) displayed a

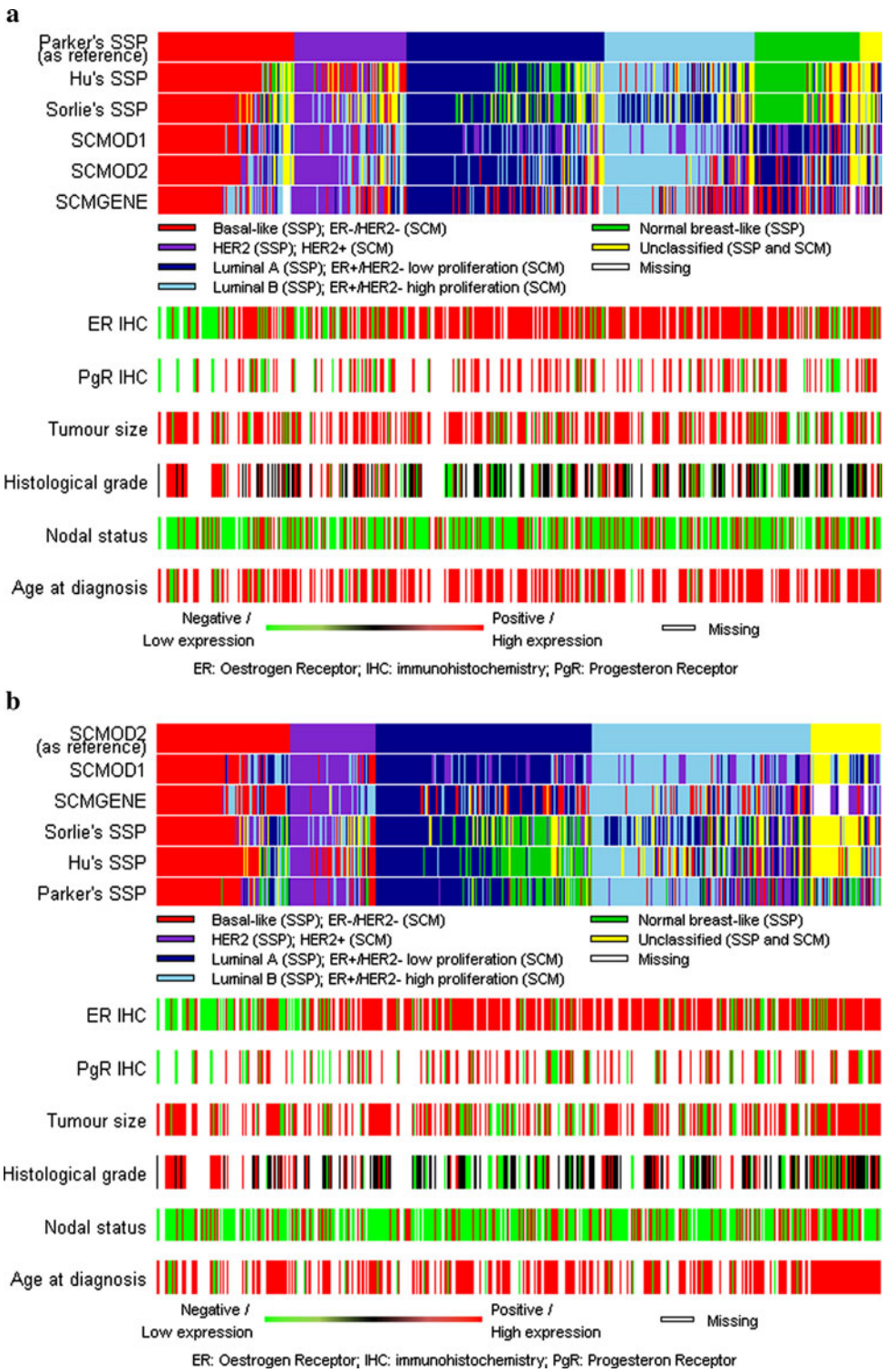


Fig. 1 Molecular subtype predictors (MSPs) classification comparison and clinicopathologic data distribution of breast cancer patients included in bc-GenExMiner using **a** Parker's single sample predictor (SSP) ($n = 2{,}974$) and **b** SCMOD2 as reference ($n = 2{,}773$)

worse prognosis. Use of MR or AE as prognostic events gave the same results and RMSPC curves confirmed this order (Fig. 2, Supplementary Figure s1 and s2). In previous studies, worst prognostic groups were HER2+ and basal-like; luminal B being intermediate between luminal A and these last subtypes [20].

Data analyses

bc-GenExMiner proposes to evaluate prognostic informativity of genes. Whichever the type of analysis chosen (targeted analysis, exhaustive analysis, and analysis by molecular subtype), the answer to the user's question

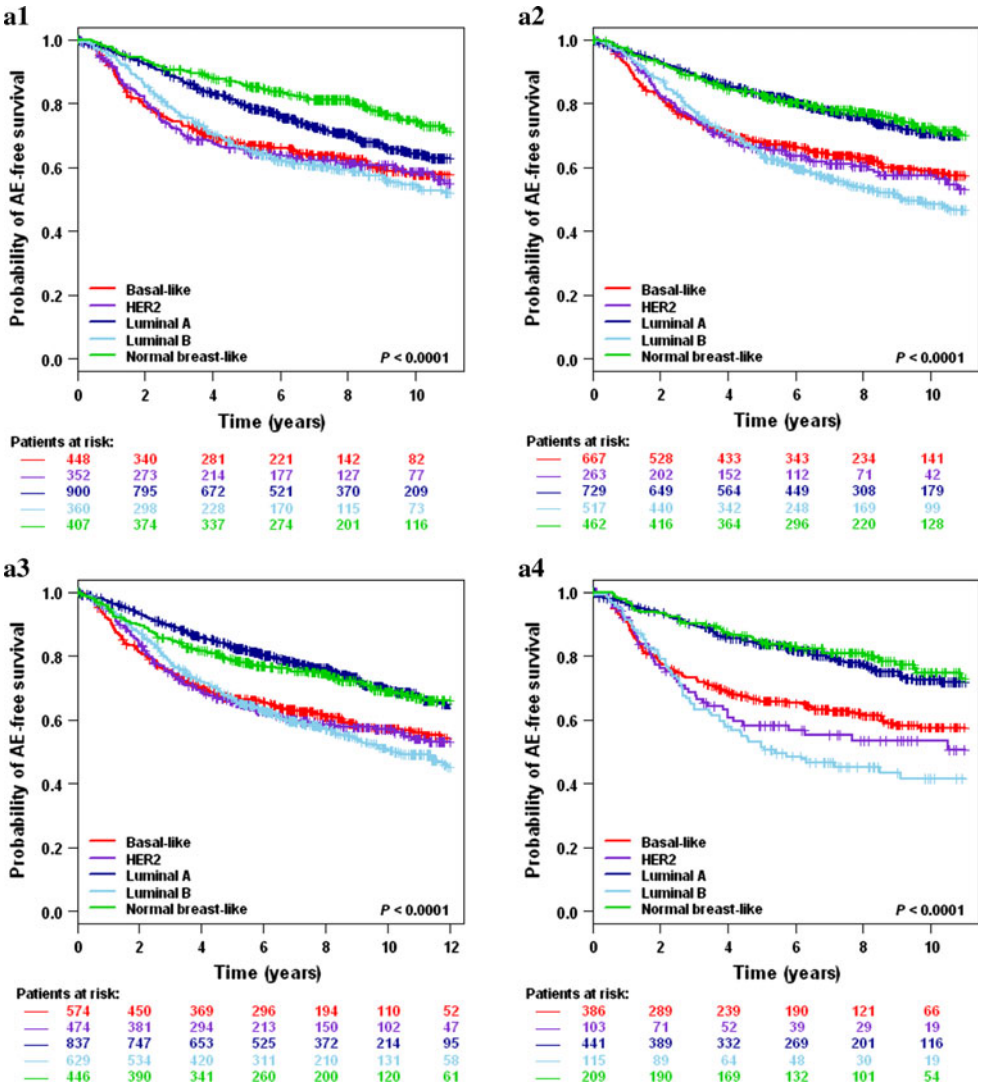


Fig. 2 "Any event" (AE)-free survival Kaplan–Meier curves (compared by means of log-rank test) of breast cancer patients assigned to the molecular subtype classes using Sorlie's (**a1**), Hu's (**a2**), and Parker (**a3**) single sample predictors (SSPs), intersection of patients classified in the same subtype with the three SSPs (**a4**), SCMOD1 (**b1**), SCMOD2 (**b2**), and SCMGENE (**b3**) Subtype Clustering Models (SCMs); and intersection of patients classified in the same subtype with the three SCMs (**b4**)

appears quickly and very easily after a few "clicks," without needing any programming skill.

Targeted analysis

A targeted analysis performs survival analyses on 1–10 chosen genes for the population corresponding to the selected criteria. Results are displayed in a table summarising Cox scores (*P* values, hazard ratios, 95% confidence interval, number of patients, and events) for each cohort fulfilling the selected criteria, separately and pooled together. The forest plot, Kaplan–Meier curves (for the median-split pool) and NPI- and AOL-adjusted Cox analyses results are also displayed.

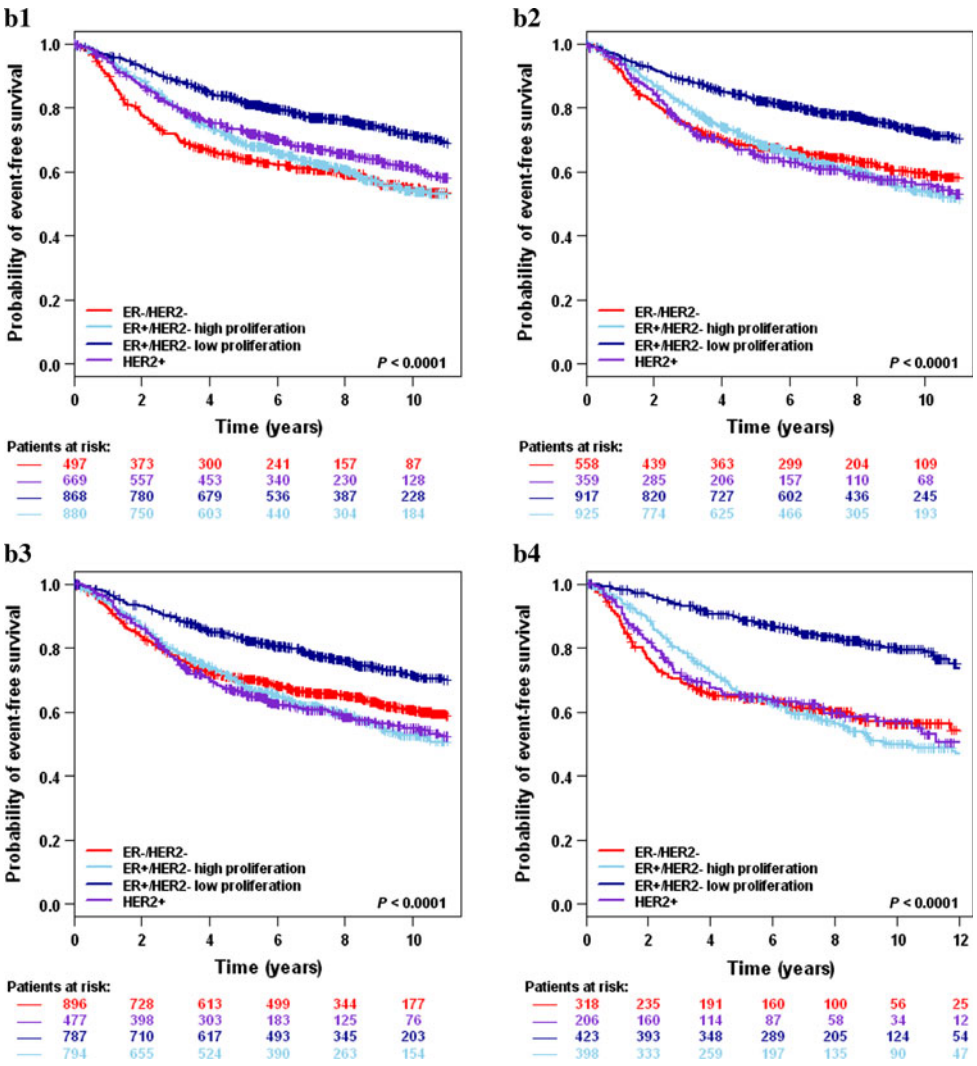
Exhaustive analysis

In an exhaustive analysis, univariate Cox proportional hazards model is performed on each of the 45 possible pools corresponding to every combination of population (N and ER statuses) and prognostic event criteria. Results are displayed in a table summarising Cox scores for all combinations of population.

Analysis by molecular subtype

In a molecular subtype analysis, univariate Cox proportional analysis and Kaplan–Meier curves are performed for each of the different molecular subtyped populations for the gene chosen by the user. Results are displayed in a table organised by molecular subtype and MSP. Number of patients unclassified is also displayed for each of the MSP. Kaplan–Meier curves can be accessed via links embedded in the table. In addition, gene expression maps are displayed according to different molecular subtyping classification methods, and notably to what we named: robust classifications. Genes whose expression is linked to particular molecular subtype demonstrate clear clusters of

Fig. 2 continued



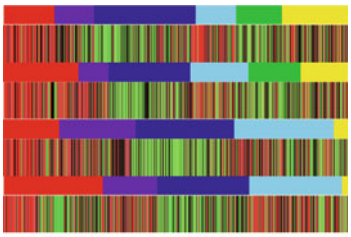
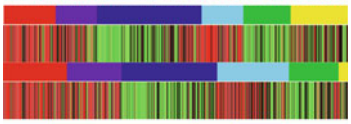
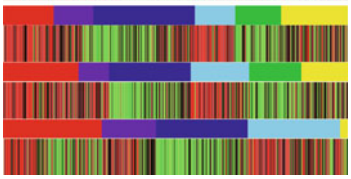
expression, especially in robust classifications. Percentages of gene with high, intermediate, and low expression per molecular subtype are given for robust classifications.

bc-GenExMiner biological validation

Complexity of bioinformatics process may distort genomic data, and downstream, statistics applied on these data may conduct to erroneous results. Validation of bc-GenExMiner may be tackled by different approaches focusing on biological relevance of annotated genomic data included in tool's database, and prognostic tests of actually validated breast cancer biomarkers.

First, we evaluated the molecular subtyping relevance of bc-GenExMiner cohort. (1) IHC ER-negative patients belonged significantly to basal-like subtype (P [Fisher] $< 10^{-4}$), (2) Agreement study using IHC HER2 and HER2 molecular subtype proved that MSPs applied on bc-GenExMiner data performed a good subtype assignment, (3) We selected genes located in the close vicinity of *ERBB2* at 17q12–q21 (DNA plus- and minus-strand) reported to be co-amplified with *ERBB2* in breast cancer. To circumvent assignment bias, we only tested genes which were not included in any of the SSP gene lists by means of RSSPC. RSCMC and RMSPC could not be used as all of the genes were present in at least one of the SCM gene lists. Tested genes were: *MED1*, *STARD3*, *TCAP*, *PNMT*, *PGAP3*, *C17orf37*, *ORMDL3*, *PSMD3*, and *NR1D1*. For all of them, gene expression maps showed high expression for RSSPC HER2+ molecular subtype (mean "high expression" = 82.6%) (Supplementary Table s6). This result is in concordance with numerous ones, which showed that 17q12–q21 region is a hot spot of genes highly co-expressed in HER2-positive breast tumours [23], (4) We tested genes whose expression significantly varies between basal-like and luminal subtypes

Table 3 Gene expression maps and percentages of proliferation genes (*MKI67*, *AURKA*, and *UBE2C*) in function of molecular subtype predictors, which do not include these genes in their gene lists (molecular subtype predictor reference above gene expression map) (gene expression: low [*green*], intermediate [*black*], high [*red*]; molecular subtype: basal-like or ER−/HER− [*red*], HER2+ [*purple*], luminal A or ER+/HER2− low proliferation [*dark blue*], luminal B or ER+/HER2− high proliferation [*sky blue*], normal breast-like [*green*], unclassified [*yellow*])

Gene	MSP	Gene expression map	No. basal-like, HER2+ and luminal B patients	% gene expression in basal-like, HER2+ and luminal B subtypes			No. Luminal A patients	% gene expression in luminal A subtype		
				Low	Medium	High		Low	Medium	High
MKI67	Sorlie's		1165	15	29	56	903	46	38	16
	Hu's		1455	16	29	55	731	54	39	7
	SCMOD1		2055	20	36	44	873	63	28	9
	SCMGENE		2090	25	32	43	779	54	39	7
AURKA	Sorlie's		1147	11	29	60	895	48	37	15
	Parker's		1614	12	34	54	812	67	29	4
UBE2C	Sorlie's		1165	12	29	59	903	48	34	18
	Hu's		1455	8	31	61	731	63	31	6
	SCMGENE		2090	22	33	45	779	65	28	7

[24]. As described above, we retained genes which were not included in any of the MSP gene lists. Hence, gene expression maps and tables were established based on RMSPC for genes highly expressed in basal-like subtype (*MET*, *ETS1*, *KRT6A*, *KRT6B*, *ANXA8*, and *MMP9*) and in luminal subtypes (*PRLR* and *KRT19*) compared with the other subtype. bc-GenExMiner results were concordant with this previous study (Supplementary Table s7). Mean "high expression" of the six first genes was 58.2% in RMSPC basal-like subtype and 20.2% in RMSPC luminal subtypes. As expected, percentage of "high expression" was the inverse for the last two genes: 8.5 versus 41.5%. Despite the fact that Charafe-Jauffret gene lists were established based on breast cancer cell lines, bc-GenExMiner results, which are based on genomic data of undissected breast cancer tumours, were concordant with this previous study. In this example, stromal environment, which must be taken into account to understand breast cancer molecular physiopathology, does not influence expression of these genes, (5) NPI, which is linked to proliferation, is significantly lower in luminal A, which is considered as a low proliferative subtype, and higher in proliferative subtypes (luminal B, basal-like, and HER2+) (P [Wilcoxon] $< 10^{-4}$), (6) Expression of prototypic proliferation genes (*AURKA*, *MKI67*, and *UBE2C*) confirmed the proliferation relevance of subtyping methods. Proliferation genes' expression maps and percentages in function of the different MSPs, which do not include these genes in their gene lists, displayed a concordant expression with proliferative status of the different subtypes (low for luminal A and high for basal-like, HER2+, and luminal B) (Table 3), (7) Survival analyses demonstrated that subtype classification defines different prognostic groups as described in previous studies.

Second, we performed prognostic analyses (1) We tested genes whose proteins are used in clinical practise because of their prognostic informativity in breast cancer: *ESR1* (ER), *MKI67* (KI67), *PLAU* (uPA), and *Serpine 1* (PAI-1). bc-GenExMiner results are concordant with clinico-biological knowledge (Supplementary Figure s3). (2) We tested genes whose prognostic significance at the RNA level has been proven for particular molecular subtypes: *PLAU* for HER2+ and *STAT1* for basal-like [17, 25]. bc-GenExMiner results confirmed the prognostic informativity of *PLAU* for HER2+ subtype ($P = 0.0491$ for RMSPC) and a tendency for *STAT1* and basal-like ($P = 0.0818$ for RSSPC).

Based on these positive validation results, which demonstrated gene expression subtype and prognostic analyses relevancies, we can advance that bc-GenExMiner caught biological sense contained in annotated genomic data and preserved it from bioinformatics biases, even when data are merged in new cohorts, and that its results are pertinent.

Discussion

bc-GenExMiner was developed to give an answer to researchers about breast cancer prognostic in vivo role of genes based on genomic data, in the shortest time possible with aims of simplicity and usability. Many web tools devoted to molecular biology are freely accessible, or not, and offer numerous data-mining platforms, which often require time and particular knowledge. Unlike these tools, bc-GenExMiner is easy to use and requires no special skills (e.g., in bioinformatics): the user selects one or more genes (and population criteria) and, after a few clicks, knows the in vivo prognostic impact of the chosen gene(s). We think that this web tool will offer a valuable help to researchers for prognostic gene discovery in breast cancer.

Development of bc-GenExMiner included molecular subtyping of 3,063 patients. To our knowledge, this is one of the first times that this process was performed on such a large number of patients. As Weigelt et al. [19], we pointed out the inconsistency of molecular subtyping, and the fact that basal-like was the most stable subtype. MSPs definitely demonstrated a weak performance to robustly subtype breast cancer patients. We showed that use of stable MSP subtyped patients, by means of a meta-classification based on intersection of three SSPs and/or SCMs, certainly brings more robustness to molecular assignment but does not resolve all problems of specificity. That said, we advance that robust molecular classification cohorts are much more suitable for prognostic analyses than any of the three SSP- or SCM-subtyped cohorts.

bc-GenExMiner v2.0 represents the module of a more complex platform. To increase the robustness of the analyses, bc-GenExMiner will continuously be updated with new data sets. To this end, researchers are invited to deposit their own annotated breast cancer genomic data in the database. In a future phase, we will include annotated genomic data of other tumours. Transversal comparisons between results obtained for different tumours will pinpoint crucial cancer genes. Furthermore, new functions will contribute to the evolution and utility of the present web tool.

To conclude, we would like to insist on the need for automated genomic data-mining tools by paraphrasing Andrea Bild [26]: "Genomic era transforms biology from an observational science into a data-intensive quantitative science". And finally, we would like to add: The more easy-to-use these tools, the more they benefit to researchers.

Acknowledgments This study was supported by SANOFI-AVENTIS-France, PFIZER-France and GSK. These pharmaceutical companies did not have any role in the design of this study, or in the preparation of this manuscript. We thank Franck Poiron for technical assistance. We are grateful to Pascale Hillard for English revision of this manuscript.

Conflict of interest None.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406:747–752
2. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lønning P, Børresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869–10874
3. Brenton JD, Carey LA, Ahmed AA, Caldas C (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol 23:7350–7360
4. R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
5. Lumley T (2006) rmeta: Meta-analysis. R package version 2.14. http://CRAN.R-project.org/package=rmeta
6. Therneau T and original R port by Thomas Lumley (2009) Survival: Survival analysis, including penalised likelihood. R package version 2.35-4. http://CRAN.R-project.org/package=survival
7. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210
8. Parkinson H, Kapushensky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37(Database Issue):D868–D872
9. Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TB, Wymore F, Zachariah ZK, Sherlock G, Ball CA (2009) Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res 37(Database Issue):D898–D901
10. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gomez HL, Hortobagyi GN, Pusztai L (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol 24:4236–4244
11. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB (2008) Merging two gene-expression studies via cross-platform normalization. Bioinformatics 24:1154–1160
12. Weigelt B, Mackay A, A'hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS (2010) Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol 11:339–349
13. Sorlie T, Tibshirani R, Parker J, Hasties T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100:8418–8423
14. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H,

Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM (2006) The molecular portraits of breast tumors are conserved across microarray platform. BMC Genomics 7:96
15. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27:1160–1167
16. Haibe-Kains B, Bontempi G, Quackenbush JF et al (2010) genefu: Relevant functions for gene expression analysis, especially in breast cancer. R package version 1.0.9. http://CRAN.R-project.org/package=genefu
17. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 14:5158–5165
18. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schütz F, Goldstein DR, Piccart M, Delorenzi M (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 10:R65
19. Galea MH, Blamey RW, Elston CE, Ellis IO (1992) The Nottingham prognostic index in primary breast cancer. Breast Cancer Res Treat 22:207–219
20. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, Parker HL (2001) Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 19:980–991
21. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
22. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
23. Kauraniemi P, Kallioniemi A (2006) Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. Endocr Relat Cancer 13:39–49
24. Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adélaïde J, Cervera N, Fekairi S, Xerri L, Jacquemier J, Birnbaum D, Bertucci F (2006) Gene expression profiling of breast cancer cell lines identifies new basal markers. Oncogene 25:2273–2284
25. Urban P, Vuaroqueaux V, Labuhn M, Delorenzi M, Wirapati P, Wight E, Senn HJ, Benz C, Eppenberger U, Eppenberger-Castori S (2006) Increased expression of urokinase-type plasminogen activator mRNA determines adverse prognosis in ErbB2-positive primary breast cancer. J Clin Oncol 24:4245–4253
26. Bild AH, Parker JS, Gustafson AM, Acharya CR, Hoadley KA, Anders C, Marcorn PK, Carey LA, Potti A, Nevins JR, Perou CM (2009) An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. Breast Cancer Res 11:R55