Breast Cancer Res Treat (2012) 133:37–47
DOI 10.1007/s10549-011-1683-z

PRECLINICAL STUDY

A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response

Oscar Krijgsman · Paul Roepman · Wilbert Zwart · Jason S. Carroll · Sun Tian · Femke A. de Snoo · Richard A. Bender · Rene Bernards · Annuska M. Glas

Received: 29 April 2011 / Accepted: 13 July 2011 / Published online: 4 August 2011

Abstract Classification of breast cancer into molecular subtypes maybe important for the proper selection of therapy, as tumors with seemingly similar histopathological features can have strikingly different clinical outcomes. Herein, we report the development of a molecular subtyping profile (BluePrint), that enables rationalization in patient selection for either chemotherapy or endocrine therapy prescription. An 80-Gene Molecular Subtyping Profile (BluePrint) was developed using 200 breast cancer patient specimens and confirmed on four independent validation cohorts ($n = 784$). Additionally, the profile was tested as a predictor of chemotherapy response in 133 breast cancer patients, treated with T/FAC neoadjuvant chemotherapy. BluePrint classification of a patient cohort that was treated with neoadjuvant chemotherapy ($n = 133$) shows improved distribution of pathological Complete Response (pCR), among molecular subgroups compared with local pathology: 56% of the patients had a pCR in the Basal-type subgroup, 3% in the MammaPrint Low-risk, Luminal-type subgroup, 11% in the MammaPrint High-risk, Luminal-type subgroup, and 50% in the HER2-type subgroup. The group of genes identifying Luminal-type breast cancer is highly enriched for genes having an Estrogen Receptor binding site proximal to the promoter-region, suggesting that these genes are direct targets of the Estrogen Receptor. Implementation of this profile may improve the clinical management of breast cancer patients, by enabling the selection of patients who are most likely to benefit from either chemotherapy or from endocrine therapy.

Keywords Breast cancer · Estrogen receptor · Molecular subtyping · Luminal · Basal · HER2

O. Krijgsman · P. Roepman · S. Tian · R. Bernards · A. M. Glas (✉)
Research and Development, Agendia BV, Amsterdam, The Netherlands
e-mail: annuska.glas@agendia.com

Present Address:
O. Krijgsman
Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands

W. Zwart · J. S. Carroll
Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK

F. A. de Snoo
Department of Medical Affairs, Agendia BV, Amsterdam, The Netherlands

R. A. Bender
Department of Medical Affairs, Agendia Inc, Irvine, CA, USA

R. Bernards
Division of Molecular Carcinogenesis, Center for Biomedical Genetics and Cancer, Genomics Center, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Introduction

Recent studies have shown that the classification of breast cancer into molecular subtypes is largely represented by the estrogen receptor (ER), progesterone receptor (PR), and HER2 status of the tumor: Basal-like breast cancers correlate best with ER-negative, profile may improve the-negative tumors [1, 2]; Luminal-like cancers are ER-positive [3], while HER2-positive cancers have a high expression of the HER2 gene [4]. Breast cancer can be classified into molecular subtypes by simple hierarchical clustering of breast tumors, according to their gene expression patterns [5].

While this classification system has been developed without consideration of patient survival rates, the different molecular subtypes of breast cancer have different prognoses: Luminal-like tumors have a more favorable outcome, and basal-like, and HER2 subgroups are more sensitive to chemotherapy [6–8]. To date, molecular classification has not played a major role in treatment decisions. However, given the role, that these subtypes have played in clinical trials (e.g. PARP inhibitors or platinum agents and taxanes for triple-negative cancers [9, 10]), they are likely to play an important role in future clinical practice.

Concordance between the immunohistochemistry (IHC) determined receptor status, and the molecular subtype suggests that, molecular profiles represent oncogenic processes that are driven by pathways in which ER, PR, and HER2 play pivotal roles [2, 3, 11–13]. It is, therefore, likely that the use of gene expression arrays will enable the identification of previously unappreciated subtypes of breast cancer that differ in clinical outcomes.

Molecular classification of breast tumors by IHC, or through determination of gene activity by measuring mRNA levels for single genes (e.g. TargetPrint), rely on the presence of protein and mRNA, respectively, but neither assay determines whether the protein, or mRNA is functional in making full length and functional receptor proteins. As such, both methods have an inherent uncertainty in predicting whether a tumor is truly positive for functional ER, PR, or HER2 protein [14]. One method to circumvent this problem is to develop gene signatures that measure the expression of groups of genes that correlate with the presence of the gene(s) of interest [15, 16].

Here we report the development of an 80-gene profile for the classification of breast cancer into three molecular subtypes.

Materials and methods

Patient selection

A total of 1,212 patient specimens from 6 different studies were analyzed (an overview of the different cohorts used in this study can be found in Table 1). Cohort 1, described in van de Vijver et al. [17], was used for the development (cohort 1a) and initial validation (cohort 1b) of the molecular subtyping profile. Cohort 1a samples ($n = 200$) were selected for their concordance between classification based on their ER, PR, and HER2 status by immunohistochemistry (IHC), and by TargetPrint microarray based single-gene readout (see below). Cohort 1b samples ($n = 95$) had a discordance between IHC and TargetPrint ER, PR, or HER2 determination.

Cohort 2 consisted of 274 early-stage breast cancer samples from a consecutive series of patients seen at the Netherlands Cancer Institute and treated with adjuvant tamoxifen monotherapy [18]. Cohort 3($n = 100$) was a group of patients from the RASTER trial [19].

Additional validation of the profile was performed using two publicly available data sets: cohort 4, $n = 159$, and cohort 5, $n = 251$ (Table 1) [20, 21].

The last cohort (cohort 6, Table 1), consisting of publicly available expression data from breast cancer patients ($n = 133$), was used to determine the response to T/FAC neoadjuvant chemotherapy for patients sub-divided by molecular subtype [22].

Microarray gene expression data

Pre-processed and normalized Agilent 22 K dual-color expression data from cohort 1 was downloaded from http://www.rii.com, and duplicate dye-swap hybridizations were combined into a single log-ratio expression value per probe per sample. Samples from cohorts 2 and 3 were analyzed on Agilent arrays according to manufacturer's protocols. Expression data was quantified using Feature Extraction software. Pre-processed and normalized Affymetrix U133A and U133B gene expression data from cohorts 4 and 5 were available at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) with accession numbers GSE1456 (cohort 4) and GSE3494 (cohort 5). Affymetrix data for cohort 6 was publicly available and downloaded from http://bioinformatics.mdanderson.org/pubdata.html. Microarray expression datasets were imported into R/Bioconductor software (www.bioconductor.org) for further analysis. Affymetrix datasets were normalized by median scaling to represent the expression ratio distributions of cohorts 1–3.

ER, PR, and HER2 status by IHC and TargetPrint

The ER, PR, and HER2 status of cohorts 1, 2, and 3 were determined by IHC and by TargetPrint (Agendia; see overview in Table 1) [14, 23]. Detailed procedures for centrally performed IHC and TargetPrint of the samples from cohorts 1–3 have been previously reported [14].

As shown in Table 1, locally determined IHC status of ER was available for cohorts 5 and 6, PR status was available for cohort 5, and HER2 status was available for cohort 6. Receptor status of the publically available datasets was downloaded from their respective GEO websites.

Identification of an 80-gene Molecular Subtyping Profile (BluePrint)

The 200 samples (cohort 1a) with concordant ER, PR, and HER2 status were used for supervised training to identify gene expression profiles specific for three molecular subtyping classes: Basal-type (triple-negative) tumors, Luminal-type

Table 1 Overview and characteristics of the training and validation cohorts

	1[a]		2	3[b]	4	5	6
Cohort #							
Samples (*n*)	295		274	100	159	251	133
Array	Agilent 22 k		Agilent 44 K	Agilent 44 K	Affymetrix U133	Affymetrix U133	Affymetrix U133
Reference	van de Vijver [17]		Kok [18]	Bueno-de-Mesquita [19]	Pawitan [20]	Miller [21]	Hess [22]
Subsets*	1a	1b					
Samples (*n*)	200	95					
Purpose	Training	Validation	Independent validation	Independent validation	in silico validation	in silico validation	in silico validation chemo-response
MSP class							
Hormone receptor-positive (%)	60	74	78	49	67	60	62
Triple-negative (%)	16	16	8	16	17	30	20
HER2-positive (%)	24	11	14	35	16	10	18
TargetPrint							
ER-positive (%)	77	80	80	71			
PR positive (%)	72	43	47	61	NA	NA	NA
HER2 positive (%)	24	7	13	39			
IHC (+CISH for HER2)							
ER-positive (%)	73	78	84	68		85	62
PR positive (%)	72	49	46	49	NA	76	NA
HER2 positive (%)	24	13	13	38		NA	25

Note: Not all percentages sum to 100 because of rounding

[a] Cohort 1 was divided into training cohort 1a that consisted of samples with concordant TargetPrint and IHC based receptor classification, and into validation cohort 1b that consisted of samples with a discordance between IHC and TargetPrint based classification

[b] Samples within cohort 3 have been selected to include approximately two-thirds hormone positive samples and one-third HER2 positive samples

(hormone receptor-positive) tumors, and HER2-positive breast tumors. Using a threefold cross validation (CV) procedure, we identified the genes that best discriminate between the three molecular subtypes. Within each CV iteration, two-sample Welch *t*-tests were performed on a randomly selected set of 133 of the 200 training samples to score all genes for their differential expression among the three classes. Genes were ranked according to their absolute *t*-statistics and the threefold CV procedure was repeated a hundred times. Next, the 100 gene ranking scores were combined into a single ranking per gene, and the minimal number of genes with optimal performance was determined using a leave-one-out CV on all 200 training samples. Optimal performance was achieved with a total of 80 unique genes (Table 2).

Next, a nearest-centroid classification model was built utilizing the 80-gene profile, in a fashion similar to that described previously [15, 24, 25]. Cohort 1a was used to establish a Basal-type centroid profile (based on 28 genes), a Luminal-type profile (based on 58 genes), and a HER2-type profile (based on 4 genes). For all additional samples and for cohort 1a using a leave-one-out CV, a correlation index was calculated between the sample's 80-gene profile and each of the three MSP centroids.

ChIP-seq analysis and intersections with the Luminal-type gene signature

ERα ChIP-seq data from proliferating MCF-7 breast cancer cells were used from a publicly available dataset [26]. The transcription start sites of the Luminal-type gene set were determined, after which, the presence of Estrogen Receptor binding sites was analyzed within a window of ∓ 20 kb from the transcription start site. The sequence track was visualized using the UCSC genome browser (http://genome.ucsc.edu/). Peak intensity was determined from the tag count.

The set of 58 Luminal genes was analyzed through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com). The genes were queried against the Ingenuity Knowledge Base and interactions between the 58 Luminal genes and ER were identified.

Table 2 Overview of reporter genes for Basal-like, Luminal-like, and HER2-like molecular subgroups

Luminal		Luminal, *cont.*		Basal		HER2	
NM_000663	ABAT	NM_006864	LILRB3	NM_145186	ABCC11	NM_004448	**ERBB2***
NM_001609	ACADSB	NM_015541	**LRIG1**	NM_001609	ACADSB	NM_001030002	**GRB7***
NM_024722	ACBD4	NM_005375	**MYB**	NM_002285	AFF3	NM_033419	PERLD1
NM_001124	ADM	NM_000662	**NAT1***	NM_006408	**AGR2**	NM_153694	SYCP3
NM_002285	AFF3	NM_000909	NPY1R	NM_000044	**AR**		
NM_000633	**BCL2***	NM_007083	NUDT6	NM_206925	**CA12**		
NM_003766	BECN1	NM_017830	OCIAD1	NM_144575	**CAPN13**		
NM_000060	BTD	NM_032521	PARD6B	NM_031942	**CDCA7**		
NM_003939	BTRC	NM_000926	**PGR***	NM_001267	CHAD		
NM_206925	CA12	NM_203453	PPAPDC2	NM_005794	DHRS2		
NM_207310	CCDC74B	NM_020820	PREX1	NM_000125	ESR1		
NM_004358	CDC25B	NM_032918	**RERG**	NM_004496	**FOXA1***		
NM_014246	**CELSR1**	NM_173079	RUNDC1	NM_001453	**FOXC1***		
NM_001408	CELSR2	NM_002964	**S100A8**	NM_001002295	**GATA3**		
NM_001267	CHAD	NM_020974	**SCUBE2**	NM_014668	GREB1		
NM_016138	COQ7	NM_003108	SOX11	NM_019600	**KIAA1370**		
NM_003462	**DNALI1**	NM_145006	SUSD3	NM_177433	MAGED2		
NM_021814	**ELOVL5**	NM_153365	TAPT1	NM_024101	**MLPH***		
NM_000125	**ESR1***	NM_015130	TBC1D9	NM_002444	MSN		
NM_001002295	GATA3	NM_024549	TCTN1	NM_018728	**MYO5C**		
NM_017786	GOLSYN	NM_024817	**THSD4**	NM_033419	PERLD1		
NM_014668	GREB1	NM_144686	TMC4	NM_175887	PRR15		
NM_024827	HDAC11	NM_032376	TMEM101	NM_138393	**REEP6**		
NM_002115	HK3	NM_021103	**TMSB10**	NM_178568	RTN4RL1		
NM_000191	HMGCL	NM_198485	TPRG1	NM_004694	**SLC16A6**		
NM_002184	IL6ST	NM_152376	UBXD3	NM_015417	SPEF1		
NM_005544	IRS1	NM_018478	DBNDD2	NM_015130	TBC1D9		
NM_033426	KIAA1737	NM_006113	**VAV3**	NM_024817	THSD4		
NM_005733	**KIF20A**	NM_005080	**XBP1**				

Genes indicated in **bold** are also present in the intrinsic gene set as originally reported by Perou et al. [5]; genes that also have an asterisk (*) are genes that are also present in the PAM50 gene set [28]

Results

An 80-gene Molecular Subtyping Profile

We used the TargetPrint assay [14] to quantify ER, PR, and HER2 mRNA levels in a training cohort of 295 breast cancer samples. We then used IHC/CISH to measure ER, PR, and HER2 protein levels in the same 295 samples for the same three receptors. Employing the 200 samples with concordant ER, PR, and HER2 status by IHC and TargetPrint mRNA readout (Table 1), a Molecular Subtyping Profile (MSP) was developed using a supervised training method. By using only concordant samples, we sought to capture ER, PR, or HER2 regulated processes more reliably and robustly as compared with the use of each assay individually. Gene expression profiles were identified specific for three molecular subtypes: Basal-type (triple-negative) tumors, Luminal-type (hormone receptor-positive) tumors, and HER2-type (HER2-positive) breast tumors. Using a threefold Cross Validation (CV) procedure, we identified 80 genes that best discriminated the three molecular subtypes (Table 2; Fig. 1a, see "Materials and methods" for details).

Testing of the profile on the 95 samples from cohort 1b (discordant for ER, PR, and HER2 by IHC/CISH and TargetPrint, details in Table 1 and "Materials and methods") showed that, the Molecular Subtyping Profile was more concordant with the TargetPrint status than with ER, PR, and HER2 status as determined by IHC. For example, of the nine IHC/CISH HER2-positive samples, only one was classified as MSP HER2-type, compared with all (7/7) of the TargetPrint HER2-positive samples. Similarly, only three of the seven IHC based triple-negative samples were classified as Basal-type by MSP, while eleven of the

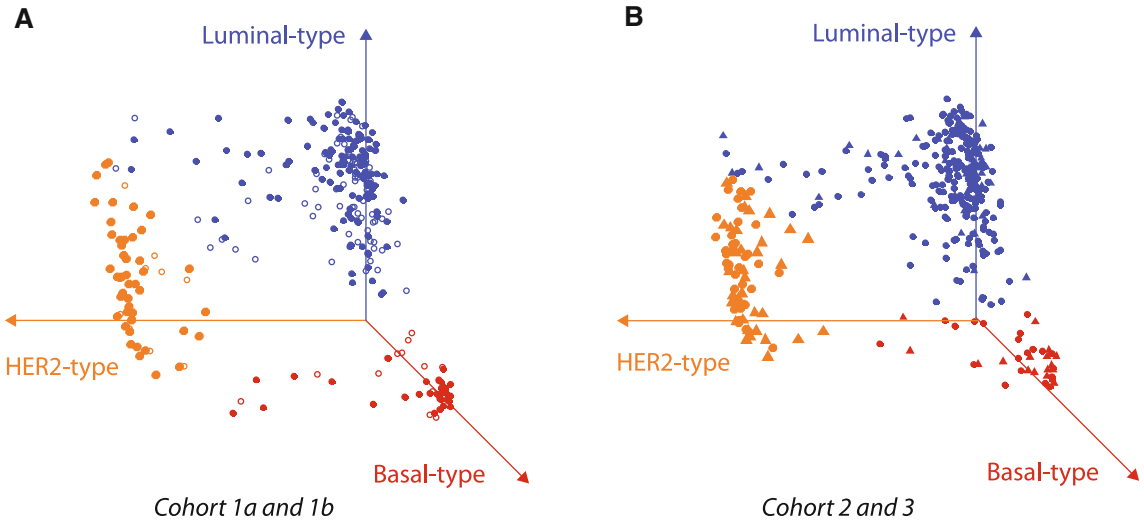

Fig. 1 Molecular Subtyping Profile (MSP) of training and validation cohorts. 3D-scatterplots are shown for **a** cohorts 1a (*solid circles*) and 1b (*open circles*), and for **b** validation cohorts 2 (*circles*) and 3 (*triangles*). The *x*, *y*, and *z*-axis, respectively, show the HER2-positive type, Basal-type, and Luminal-type scores of all samples. Samples are colored according to their MSP outcome, in which, Luminal-types are shown in *blue*, Basal-types in *red*, and Her2-types in orange

thirteen TargetPrint triple-negative samples showed a MSP Basal-type classification.

Confirmation of MSP in independent patient cohorts

Next, the classification of breast cancer samples into three molecular subtypes using the developed 80-gene MSP was performed using two independent patient cohorts with a total of 374 patient specimens (cohorts 2–3, Table 1). Of the 374 patients, 39 (10%) were classified as Basal-type, 263 (71%) were classified as Luminal-type, and 64 (19%) as HER2-type (Table 1; Fig. 1b).

Statistical analysis to validate the presence of the three MSP classes in the independent cohorts was performed using the in-group proportion (IGP) statistic, which is defined to be the proportion of samples in a group, whose nearest neighbors are also in the same group, and can be considered as a measure of the robustness and reproducibility of the MSP profile in independent series [27]. The MSP classification of all the three subtypes was highly preserved in cohorts 2 and 3, respectively, with an IGP of 0.98 and 0.88 for Luminal-type, 0.87 and 0.94 for Basal-type, and 0.78 and 0.89 for HER2-type classifications (Table 3).

We also confirmed the MSP on microarray data generated by other investigators using Affymetrix arrays (cohort 4–5, Table 1). All 80 genes could be matched to the Affymetrix data using gene symbol or RefSeq annotation. Despite the use of different platforms, and the fact that the data had been generated by other investigators, our *in silico* analysis shows similar MSP outcome distributions as observed in both the training and in-house independent cohorts with 73 out of 410 (18%) samples classified as Basal-type, 244 (60%) as Luminal-type, and 93 (23%) as HER2-type (Table 1). Calculated IGP scores confirmed the reproducibility of the MSP classes on the two *in-silico* data sets (Table 3).

Table 3 In group proportion statistics (IGP) for the MSP classes across the independent cohorts.

MSP class	In-group proportion (IGP)				
	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6
Hormone receptor-positive (Luminal-type)	0.98	0.88	0.92	0.91	0.91
Triple-negative (Basal-type)	0.87	0.94	0.92	0.68	0.96
HER2-positive (HER2-type)	0.78	0.89	0.56	0.68	0.75

IGP is defined as the proportion of samples in a group, whose nearest neighbors are also in the same group, and can be considered as a measurement for the robustness and reproducibility of identified cluster/classes across independent data sets [28]

Note: An ideal cohort has an IGP value of 1.0

Comprehensive breast cancer classification by MammaPrint and MSP

The 295 samples from cohort 1 have previously been stratified by risk of development of distant metastasis by MammaPrint [17]. Here we have combined the MammaPrint based prognosis, and the developed MSP to subtype the breast cancer samples into 4 groups: MammaPrint Low-risk/Luminal-type, MammaPrint High-risk/Luminal-type, Basal-type, and HER2-positive. The great majority (92%) of MammaPrint Low-risk samples were Luminal-type by MSP, while the High-risk samples were more equally distributed across the MSP classes with 46% Luminal-type, 26% Basal-type, and 28% HER2-type. The combined MammaPrint and MSP classification was confirmed in cohort 6, for which, chemotherapy responsiveness data was available (see below). Eighty-six percent of the Low-risk samples were Luminal-type compared with 53% of the High-risk samples.

Comparison of the 80-gene profile and the "intrinsic gene set"

Molecular classification of breast cancer by the so-called "intrinsic gene set", as first identified by Perou et al., was based on hierarchical clustering without consideration of ER, PR, and/or HER2-associated tumor biology [5]. This set has been reduced in number and is currently available as "PAM50" [28]. In contrast, the BluePrint MSP has been developed with ER, PR, and HER2 status as a starting point. The 80 BluePrint genes and their contribution to distinguishing the different subtypes is provided in Table 2, in which, we also indicated the overlapping genes with the intrinsic gene set and the PAM50 gene set. Figure 2 is a Venn diagram directly comparing the genes in the BluePrint MSP, the original intrinsic gene set, and the genes present in the PAM50 profile.

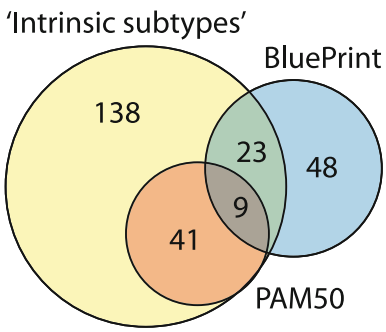

Fig. 2 Comparison of genes in the three different subtyping profiles; BluePrint, intrinsic gene set, and PAM50. Figure depicting a Venn diagram directly comparing the genes in the BluePrint MSP, the original intrinsic gene set, and the genes present in the PAM50 profile

Direct comparison between the two methods is possible, as cohort 1 samples have also been assessed using the "intrinsic gene set" (Table 4) [29]. Despite the different strategies, classification by the MSP 80-gene profile is in agreement with the classification based on the "intrinsic gene set" with an overall concordance of 92%. The "intrinsic gene set", Normal-like group was disregarded for this analysis, since, it is like an artifact of having a high percentage of normal breast cells in the specimens of the original study [28, 30, 31].

Agreement analysis with inclusion of the Normal-like class resulted in a concordance of 83%. Of note, within this analysis, we underestimated the concordance with Perou classification as the MSP has no Normal-like counterpart, and consequently, all Normal-likes are discordant between both methods. The MSP classified 2 out of the 29 "Normal-like samples" as Basal-type, 23 as Luminal-type, and 4 as HER2-type (Table 4).

Ten year Distant Metastasis Free Survival (DMFS) of patients in cohort 1, demonstrated a 4% higher survival for patients with BluePrint Luminal-type/MammaPrint Low-risk (86%, 95%CI, and 79–93%) compared with patients having Luminal A-like with the intrinsic gene set (82%, 95%CI, and 75–89%); whereas, the Luminal-type/MammaPrint High-Risk patients had a similar survival (51%, 95%CI, and 40–65%) compared with the Luminal B-like patients (49%, 95%CI, and 36–67%). Interestingly, investigation of five year DMFS indicated a 95% survival (95%CI, 91–99%) for BluePrint Luminal-type/MammaPrint Low-risk patients compared with 89% (95%CI, 83–95%) for Luminal A-like patients. Due to relatively small datasets, these differences do not reach statistical significance. The DMFS for patients classified by BluePrint as HER2-positive (60%, 95%CI, and 48–74%), or as triple-negative (61%, 95%CI, and 46–80%) were also equal compared with the DMFS rates for the patients classified according to the intrinsic gene set: HER2-likes (58%, 95%CI, and 43–78%) and basal-like (60%, 95%CI, and 47–78%).

Estrogen receptor signaling identifies Luminal-type breast cancer

We have previously reported the genome-wide mapping of Estrogen Receptor alpha binding sites in MCF-7 breast cancer cells using ChIP-seq [26]. The transcription start sites of the 58 genes in the gene set that identifies the Luminal-type subgroup were determined, after which, the presence of Estrogen Receptor binding sites was analyzed within a window of $\mp$ 20 kb from the transcription start site. A 20 kb window was previously established as an optimal distance to capture regulatory regions [32]. The sequence track was visualized using the UCSC genome

Table 4 Comparison of MSP with classification using the "intrinsic gene set" on cohort 1

Molecular subtyping by MSP	Clustering of patients by "intrinsic gene set"					
	Luminal A	Luminal B	Basal-like	Her2-like	"Normal-like"	Total
Low-risk MammaPrint Luminal-type	**83**	**9**	0	0	14	106
High-risk MammaPrint Luminal-type	**34**	**39**	2	0	9	84
Basal-type	0	0	**44**	0	2	46
HER2-type	6	7	7	**35**	4	59
Total	123	55	53	35	29	**295**

The number in bold indicates the number of samples for which the classification by both methods is in agreement

browser (http://genome.ucsc.edu/). Peak intensity was determined from the tag count. Figure 3a shows an example of a gene (*SUSD3*) which has multiple Estrogen Receptor α binding sites within 20 kb of its transcription start site (black bar) and one gene (*TPRG1*) lacking ERα binding sites in its promoter. We then performed the same analysis for all 58 genes that together identify Luminal-type breast tumors (Fig. 3b). We found that, 32 out of the 58 Luminal profile genes have ERα binding sites adjacent to the transcription start sites (55%), whereas, only ~28% of all RefSeq genes (~24,000) have ERα binding sites within the same distance window. This result indicates that, the genes that identify Luminal-type breast cancer are significantly enriched for having ERα binding sites in close proximity to its transcription start site ($P = 1.2e{-}5$), and are likely to be regulated by Estrogen Receptor α activity. Figure 3c depicts the interaction network between the 58 Luminal genes and Estrogen Receptor. Seventeen of these genes are proven to have interactions with ER using Ingenuity Pathway Analysis.

We conclude that the Luminal-type subgroup of breast cancer most likely represents tumors that depend on Estrogen Receptor signaling and, hence, are likely to respond to endocrine therapy.

Chemotherapy response

Publicly available microarray expression data with full clinical history from a neo-adjuvant clinical study, allowed *in silico* analysis of the 80-gene MSP as a predictor of pathological Complete Response (pCR) [22]. MSP readout was determined on tumor samples from 133 breast cancer patients who were treated with neo-adjuvant T/FAC chemotherapy (cohort 6, Table 1). Within this cohort, 20% ($n = 27$) were classified as Basal-type, 62% ($n = 82$) as Luminal-type, and 18% ($n = 24$) as HER2-type, with an IGP of 0.96 for Basal-type, 0.91 for Luminal-type, and 0.75 for HER2-type (Table 3). The overall pCR of this patient cohort was 26%, and differed substantially among the subgroups (Fig. 4). PCR was observed in 9% of all Luminal-type samples and, importantly, only in 3% of MammaPrint Low-Risk/Luminal-type and 11% in MammaPrint High-Risk/Luminal-type, in 50% of the HER2-type samples, and in 56% of the Basal-type samples. These numbers change when the samples are stratified according to the intrinsic gene set, and when stratified with local IHC/FISH assessment. The combined use of MammaPrint with BluePrint improves the distribution of pCR rates for the different molecular subgroups, respectively lower in the Luminal A group and higher pCR rates in the Basal and HER2 group (Table 5).

Discussion

In this article, we describe the development and validation of an 80-gene Molecular Subtyping Profile (MSP) named "BluePrint" that classifies breast cancer patients into Basal-type, Luminal-type, and HER2-type subgroups. The profile was developed in a supervised training method, using samples with concordant ER, PR, and HER2 status by IHC and single-gene readout ensuring the capture of ER/PR/HER2-regulated processes, and development of a more reliable and robust test, than a single-gene read-out by IHC or mRNA measurement. This rational based approach is different from previously defined subtypes based on hierarchical clustering [5]. The classification was validated on gene expression data from 917 samples, in which, the separation of the three subgroups was clearly maintained, indicating the robustness of the profile and the reproducible differences among the subgroups.

Sub-stratification of the Luminal-like patients into A and B subtypes, is thought to be based on the expression of markers of tumor grade and/or proliferation, such as Ki67 [11, 12, 33]. The classification into Luminal A and B appears to have prognostic implications, and maybe useful in determining the need for chemotherapy, however, more accepted characteristics for therapeutic decision making including histological grade, proliferation, nodal status, peritumoral vascular invasion, tumor size, and risk stratification by multi-gene assays (such as the 70-gene MammaPrint or the 21-gene Oncotype assay) are preferred [34]. In the current study, sub-stratification of the Luminal-type

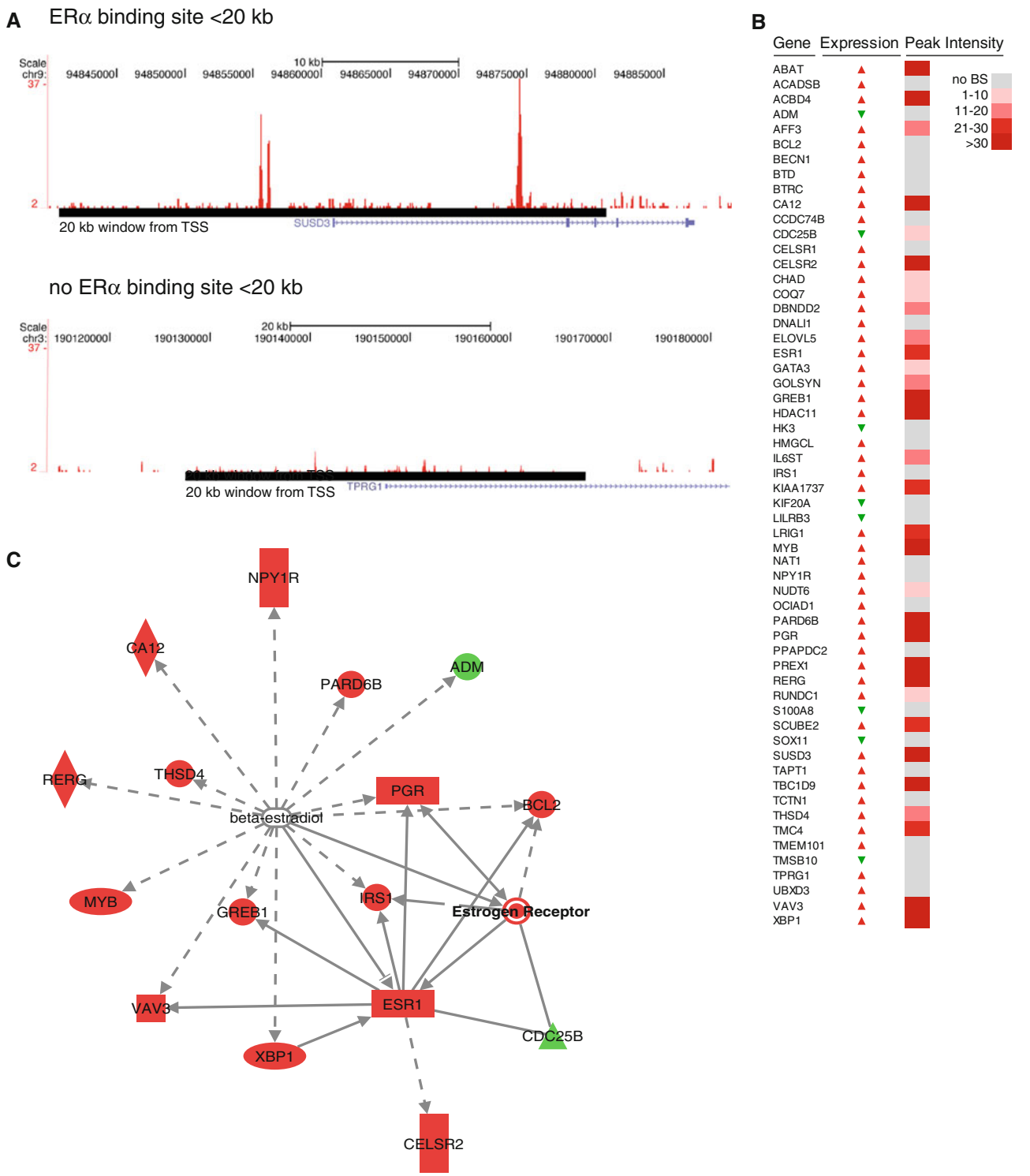

Fig. 3 Interaction network between the Luminal genes and the Estrogen Receptor. Luminal signature genes show enrichment for Estrogen Receptor binding sites close to their transcription start sites. Estrogen Receptor ChIP-seq analyses were performed from a publicly available dataset of proliferating MCF7 breast cancer cells, and the Estrogen Receptor binding site occupancy was determined within a 20 kb distance from the transcription start site of a gene from the Luminal signature. **a** Shows an example of the presence (*top panel*) or absence (*bottom panel*) of an Estrogen Receptor binding site within a 20 kb distance from the transcription start site. Estrogen Receptor binding site presence was determined for all genes in the Luminal signature and visualized in a heatmap (**b**). **c** Depicts the interaction network between the Luminal genes and Estrogen Receptor. Seventeen of the Luminal genes are proven to have interactions with ER. A *dashed line* indicates an indirect interaction and a solid line indicates direct interactions. *Red/Blue* indicates that the gene is up/down-regulated in the Luminal-type subgroup. The shape of the gene indicates the type of the protein encoded by this gene

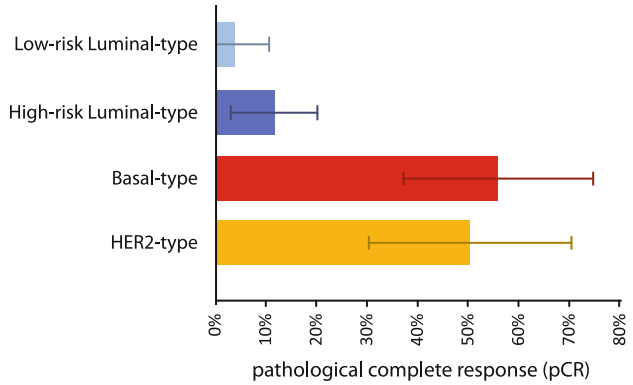

Fig. 4 Chemotherapy response differs among the MSP classes. Pathological Complete Response (pCR) rates are shown for cohort 6 Low-risk Luminal-, High-risk Luminal-, Basal-, and HER2-type breast cancers. *Error bars* indicate the 95%CI for the corresponding pCR proportions

group by MammaPrint (where MammaPrint Low-Risk/Luminal-type is equivalent to Luminal-like A and MammaPrint High-Risk/Luminal-like is equivalent to Luminal-like B) indicated a (non significant) improved Hazard Ratio for the combined use of MammaPrint and MSP compared with the sub-stratification of the "intrinsic gene set" into A and B subtypes.

PAM50, another profile for Molecular Subtyping, has recently been developed from the 4 different intrinsic gene sets that have followed since the first article on intrinsic subtyping was published [28]. This 50-gene profile classifies patients into Basal-like, HER2-enriched, Luminal A, and Luminal B subgroups. This profile has been developed specifically using qRT-PCR with a risk algorithm. The classification of patients into Luminal, HER2, and Basal subgroups by PAM50 or BluePrint is expected to have great similarity, given that the concordance with the original intrinsic gene set is >90%.

There is currently no "gold standard" for molecular subtyping of breast cancer. It is, therefore, unclear which method is best at classifying the Luminal, HER2, or Basal-like subtypes. For instance, although several investigators have used the term "triple- negative" and "basal-like" interchangeably, it should be noted that, these subtypes are not completely concordant, and that additional markers are needed to separate them [35, 36]. With standardization still pending, the 80-gene MSP Basal-type subgroup has been developed using a rational based model with concordant negative IHC/TargetPrint-assessed samples for ER, PR, and HER2.

Pathological Complete Response (pCR) in the neoadjuvant setting can be used as a surrogate measure of response to chemotherapy, and is associated with excellent long-term cancer-free survival [37–39]. We observe improved risk distribution, in response to neo-adjuvant treatment by Molecular Subtype as defined by our 80-gene profile compared with local pathology; with a pCR of 3% in the MammaPrint Low-Risk/Luminal-type samples, 11% in the MammaPrint High-Risk/Luminal-type samples, 50% in the HER2-type samples, and 56% in the Basal-type samples [2, 40].

We show here that, Luminal breast tumors are identified by a set of genes that is highly enriched for genes having ERα binding sites in close proximity to their transcription start sites. It is, therefore, very likely that these tumors are characterized by active ER signaling. Such active ER signaling is typical for tumors having a functional and active ER, which in turn suggests that such tumors respond to endocrine therapies. Thus, our data suggest that, Luminal-type breast tumors are most likely to respond to endocrine therapy, but not chemotherapy, whereas, Basal-type and HER2 type tumors are likely to respond better to chemotherapy (and HER2-targeted therapy for the HER2 types).

The clinical utility of molecular subtyping of breast tumors using expression profiling has remained unclear in part, because standardized methodology for classification has been lacking [41]. Medical practice has adopted the use

Table 5 Comparison of chemotherapy response in Hess et al. [22]

	Local pathology		Intrinsic Subtyping		BluePrint	
	n (%)	pCR (%)	*n* (%)	pCR (%)	*n* (%)	pCR (%)
Luminal (ER +/HER2−)	67 (50%)	7	30 (23%)	6	NA	
Luminal A (Low-risk Luminal-type)	NA		NA[a]		29 (22%)	3
Luminal B (High-risk Luminal-type)	NA		NA[a]		53 (40%)	11
Basal (ER−/HER2−)	32 (24%)	47	22 (17%	45	27 (20%)	56
HER2 (HER2 +)	33 (25%)	39	20 (15%)	45	24 (18%)	50
Normal-like	NA		10 (8%)	0	NA	
Not assessed	NA		51 (38%)	25	NA	

Subgroups are classified according to local pathological assessment, intrinsic subtyping, and BluePrint

Note: not all percentages sum to 100 because of rounding

[a] Hess et al. did not stratify the Luminal group into the subgroups A and B

of validated multi-gene assays, such as the 70-gene MammaPrint and 21-gene Oncotype profiles, for use in conjunction with standard clinico-pathologic risk assessment, in determining a role for adjuvant chemotherapy. The MSP profile described here will help in further establishment of a clinical correlation between molecular subtyping and treatment responses, especially in combination with MammaPrint; one biopsy sample is able to provide multiple answers. The addition of the BluePrint profile to the MammaPrint profile using the same platform indicates, how versatile the platform is, and how these two tests are merely a hint of the future possibilities provided by this technology; a multitude of profiles on one single biopsy that can be used for the optimal treatment decision. Moreover, the rational based method of BluePrint development, and the high percentage of genes in the profile with Estrogen Receptor binding sites in their promoters, proving our profile measures functionality of the estrogen receptor, argues that the BluePrint 80 gene profile is suited for medical practice, especially in clinical trial setting.

The notion that certain drugs may only be effective for patients of particular subtypes has already led to clinical trials evaluating drug responses in molecularly based subgroups. The outcome of these trials and future implementation of this knowledge may improve the clinical management of breast cancer patients, by enabling the physician to decide who is most likely to benefit from which chemotherapy before surgery. Further, it may supplement information already available from validated multi-gene assays, in helping the clinician decide on the best treatment for each patient.

Acknowledgments The authors thank Rob Pover, Maurice Pater, and Arenda Schuurman for sample and microarray processing; Cor Lieftink and Bram Gerritsen for statistical advice, and Lisette Stork-Sloots and Laura van 't Veer for fruitful discussions and helpful suggestions.

Conflicts of interest RAB is employed by Agendia Inc, all other authors, except WZ and JC, are employed by Agendia BV.

References

1. Rakha EA, Elsheikh SE, Aleskandarany MA, Habashi HO, Green AR, Powe DG, El-Sayed ME, Benhasouna A, Brunet JS, Akslen LA, Evans AJ, Blamey R, Reis-Filho JS, Foulkes WD, Ellis IO (2009) Triple-negative breast cancer: distinguishing between basal and nonbasal subtypes. Clin Cancer Res 15(7):2302–2310
2. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM (2007) The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res 13:2329–2334
3. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, van de Rijn M, Perou CM (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res 10:5367–5374
4. Kauraniemi P, Kallioniemi A (2006) Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. Endocr Relat Cancer 13:39–49
5. Perou CM, Sørlie 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
6. 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
7. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN, Pusztai L (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 11(16):5678–5685
8. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M, Cristofanilli M, Hortobagyi GN, Pusztai L (2008) Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol 26(8):1275–1281
9. Rakha EA, Reis-Filho JS, Ellis IO (2008) Basal-like breast cancer: a critical review. J Clin Oncol 26:2568–2581
10. Liang H, Tan AR (2010) Iniparib, a PARP1 inhibitor for the potential treatment of cancer, including triple-negative breast cancer. Drugs 13(9):646–656
11. Sotiriou C, Pusztai L (2009) Gene-expression signatures in breast cancer. N Engl J Med 360(8):790–800
12. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, Watson M, Davies S, Bernard PS, Parker JS, Perou CM, Ellis MJ, Nielsen TO (2009) Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst 101(10):736–750
13. Livasy CA, Karaca G, Nanda R, Tretiakova MS, Olopade OI, Moore DT, Perou CM (2006) Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. Mod Pathol 19(2):264–271
14. Roepman P, Horlings HM, Krijgsman O, Kok M, Bueno-de-Mesquita JM, Bender R, Linn SC, Glas AM, van de Vijver MJ (2009) Microarray-based determination of ER, PR and HER2 receptor status in breast cancer. Clin Cancer Res 15(22):7004–7011
15. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
16. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817–2826
17. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
18. Kok M, Koornstra RH, Mook S, Hauptmann M, Fles R, Glas AM, Jansen MP, Berns EM, Linn SC and Van't Veer LJ. Additional value of the 70-gene prognosis profile and levels of ER and PR in the prediction of outcome in tamoxifen-treated ER-positive breast cancer. 2011 (submitted)

19. Bueno-de-Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, Wu TS, Atsma D, Rutgers EJ, Brink G, Floore AN, Glas AM, Roumen RM, Bellot FE, van Krimpen C, Rodenhuis S, van de Vijver MJ, Linn SC (2007) Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). Lancet Oncol 8(12):1079–1087
20. Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7(6):R953–R964
21. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 102:13550–13555
22. 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, Gómez 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(26):4236–4244
23. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies. J Clin Oncol 23(36):9067–9072. Epub 2005 Sep 19
24. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, Witteveen AT, Rzyman W, Floore A, Burgers S, Giaccone G, Meister M, Dienemann H, Skrzypski M, Kozlowski M, Mooi WJ, van Zandwijk N (2009) An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clin Cancer Res 15(1):284–290
25. Glas AM, Kersten MJ, Delahaye LJ, Witteveen AT, Kibbelaar RE, Velds A, Wessels LF, Joosten P, Kerkhoven RM, Bernards R, van Krieken JH, Kluin PM, van'tVeer LJ, de Jong D (2005) Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. Blood 105(1):301–307
26. Ross-Innes CS, Stark R, Holmes KA, Schmidt D, Spyrou C, Russell R, Massie CE, Vowler SL, Eldridge M, Carroll JS (2010) Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. Genes Dev 24(2):171–182
27. Kapp AV, Tibshirani R (2007) Are clusters found in one data set present in another dataset? Biostatistics 8(1):9–31
28. 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(8):1160–1167
29. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355(6):560–569
30. Kapp AV, Jeffrey SS, Langerød A, Børresen-Dale AL, Han W, Noh DY, Bukholm IR, Nicolau M, Brown PO, Tibshirani R (2007) Discovery and validation of breast cancer subtypes. BMC Genomics 8(1):101
31. Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF (2006) Molecular classification of breast cancer: limitations and potential. Oncologist 11(8):868–877
32. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462(7269):58–64
33. Sotiriou C, Desmedt C (2006) Gene expression profiling in breast cancer. Ann Oncol 17(Suppl 10):x259–x262
34. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn HJ (2009) Panel members. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol 20(8):1319–1329
35. Bertucci F, Finetti P, Cervera N, Esterni B, Hermitte F, Viens P, Birnbaum D (2008) How basal are triple-negative breast cancers? Int J Cancer 123(1):236–240
36. Rakha EA, Reis-Filho JS, Ellis IO (2008) Basal-like breast cancer: a critical review. J Clin Oncol 26(15):2568–2581
37. Fisher B, Bryant J, Wolmark N, Mamounas E, Brown A, Fisher ER, Wickerham DL, Begovic M, DeCillis A, Robidoux A, Margolese RG, Cruz AB Jr, Hoehn JL, Lees AW, Dimitrov NV, Bear HD (1998) Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. J Clin Oncol 16(8):2672–2685
38. Kuerer HM, Newman LA, Smith TL, Ames FC, Hunt KK, Dhingra K, Theriault RL, Singh G, Binkley SM, Sneige N, Buchholz TA, Ross MI, McNeese MD, Buzdar AU, Hortobagyi GN, Singletary SE (1999) Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. J Clin Oncol 17(2):460–469
39. Rastogi P, Anderson SJ, Bear HD, Geyer CE, Kahlenberg MS, Robidoux A, Margolese RG, Hoehn JL, Vogel VG, Dakhil SR, Tamkus D, King KM, Pajon ER, Wright MJ, Robert J, Paik S, Mamounas EP, Wolmark N (2008) Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. J Clin Oncol 26(5):778–785
40. Straver ME, Glas AM, Hannemann J, Wesseling J, van de Vijver MJ, Rutgers EJ, Vrancken Peeters MJ, van Tinteren H, Van't Veer LJ, Rodenhuis S (2010) The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. Breast Cancer Res Treat 119(3):551–558
41. 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(4):339–349