BMC Bioinformatics

RESEARCH ARTICLE Open Access

Consistent metagenes from cancer expression profiles yield agent specific predictors of chemotherapy response

Qiyuan Li[1,2†], Aron C Eklund[1†], Nicolai J Birkbak[1], Christine Desmedt[3], Benjamin Haibe-Kains[4], Christos Sotiriou[3], W Fraser Symmans[5], Lajos Pusztai[6], Søren Brunak[1], Andrea L Richardson[7*] and Zoltan Szallasi[1,8*]

Abstract

Background: Genome scale expression profiling of human tumor samples is likely to yield improved cancer treatment decisions. However, identification of clinically predictive or prognostic classifiers can be challenging when a large number of genes are measured in a small number of tumors.

Results: We describe an unsupervised method to extract robust, consistent metagenes from multiple analogous data sets. We applied this method to expression profiles from five "double negative breast cancer" (DNBC) (not expressing ESR1 or HER2) cohorts and derived four metagenes. We assessed these metagenes in four similar but independent cohorts and found strong associations between three of the metagenes and agent-specific response to neoadjuvant therapy. Furthermore, we applied the method to ovarian and early stage lung cancer, two tumor types that lack reliable predictors of outcome, and found that the metagenes yield predictors of survival for both.

Conclusions: These results suggest that the use of multiple data sets to derive potential biomarkers can filter out data set-specific noise and can increase the efficiency in identifying clinically accurate biomarkers.

Background

Microarray gene expression profiling provides an unbiased, comprehensive view of an entire molecular system, and is well suited to identify the relevant factors that define the cancer phenotype. However, the success of this method can be impeded by problems arising from the parallel measurements of tens of thousands of gene expression levels sampled in a far lower number of tumor specimens, typically a few hundred at most. Two specific problems have impacted cancer research: First, overfitting has produced several seemingly promising diagnostic patterns that have not been verifiable in independent studies [1,2]. Second, redundant information in the form of strongly correlated genes has led to the repeated "discovery" of diagnostic patterns detecting a single robust phenomenon, such as the cell proliferation pattern that is prognostic in estrogen receptor (ER) positive breast cancer [3]. One approach to these problems is to reduce the dimensionality of the data by combining (usually correlated) genes into a small number of metagenes.

Several gene combinations have been used to characterize the cancer phenotype [4-7]. For example, the linear combination of proliferation associated genes and estrogen regulated genes provides a better predictor of outcome in tamoxifen treated ER-positive breast cancer than does either class of genes alone [8]. Although several supervised methods to find biologically relevant linear gene combinations are available, finding such predictive metagenes in an unsupervised fashion remains a challenge [5,9]. In breast cancer, expression profiles can easily discriminate between ER-negative and ER-positive tumors, which have very different clinical behavior. For this reason it is also easy, but not clinically useful, to develop trivial predictors of outcome in cohorts of mixed ER subtype. Within the ER-positive subgroup, several predictors of response to chemotherapy have been described [10-12]. However, supervised methods have not yielded highly accurate

* Correspondence: arichardson@partners.org; zszallasi@chip.org
† Contributed equally
[1]Center for Biological Sequence Analysis, Department of Systems Biolology, Technical University of Denmark, 2800 Lyngby, Denmark
[7]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA
Full list of author information is available at the end of the article

predictors of chemotherapy response in DNBC [3,13,14]. This molecularly and clinically distinct subset of breast cancers represents approximately 20-25% of all breast cancers and can be treated only with chemotherapy. About 25-30% of these cancers respond favorably to treatment, but the remainder has very poor survival despite current best therapies [15].

Here we describe an unsupervised method to derive metagenes by leveraging the consistent expression patterns found in multiple gene expression data sets of the same cancer subtype. Our approach is based on the postulate that analogous microarray data sets, such as those from patient cohorts selected under similar criteria, are representative collections from a larger population "expression space". In this expression space, individual samples are robustly separated by a set of metagenes, some of which may be clinically relevant. However, each individual data set may be adulterated by sampling artifacts and with data set specific noise. Therefore, our approach is to derive metagenes that are consistently observed in several cohorts and are likely representative of the entire population. By first identifying metagenes in an unsupervised fashion, and then evaluating association between the metagenes and clinical outcome, we reduce the risk of overfitting.

Using this method we derived metagenes from expression profiles of DNBC, stage III ovarian cancer and early stage lung cancer, respectively. Then we verified the association of these metagenes with clinical outcome in independent validation cohorts of the three cancer types.

Results

Derivation of DNBC-specific consistent expression indices (CEIs)

We created a reference data set of DNBC from five previously published breast cancer cohorts that were all profiled on the same microarray platform (HG-U133A) and were without neoadjuvant drug response data [3,16-21] (Additional file 1). From a total of 1037 tumors we identified a subset of 218 DNBC based on expression levels of ESR1 and ERBB2 [3,4,22-24] (Additional file 2).

First, we used principal component analysis (PCA) as an unsupervised method to identify a subset of genes representing highly variable patterns in DNBC expression profiles. In PCA, each principal component (PC) is defined by a vector of gene expression weights. We hypothesize that the between-sample variability of tumor is driven by a finite number of biological effects, which are summarized into the principal components. Hence a finite number of components will explain the majority of the variation of the data matrices. Therefore, we define the likelihood as the fraction of total variance that is explained by the given number of principal components. For each individual data set, we performed PCA and used the Bayesian information criterion (BIC) to select a set of 3-6 PCs that best represent the predominant variation in the data without including components that are likely to represent noise (Figure 1a, b; see methods). We expected to find any clinically relevant information enriched in these top PCs, since as the variance diminishes it becomes more difficult to distinguish signal from noise. For each reference data set, we distilled the PCs to include only the genes with a substantial contribution, as determined by the correlation between gene expression levels and PC scores across all samples. Hierarchical clustering of these distilled PCs revealed six distinct groups, or *consistent principal components* (CPCs), with at least two members. We identified 108 genes with a substantial contribution to at least two PCs in any of these clusters, hypothesizing that these genes are likely to capture consistent biologically-relevant information about DNBC (*CPC genes*) (Figure 1c).

To validate the consistency of these CPC genes, we collected four independent DNBC data sets and subjected them to PCA using only the 108 CPC genes [13,25-27]. As result, the first and the second principal components of the CPC genes are highly consistent across the four test data sets, suggesting that these genes correspond to conserved biological variation in DNBC (Figure 2a). When we applied this gene set to the ER-positive HER2-negative subset of the same cohorts, we found that the resulting top PCs were distinct from those of the DNBC samples (Figure 2b). Thus, the CPC genes represent a specific type of variation of gene-expression within DNBC, which is highly conserved in multiple different cohorts.

Next we used factor analysis (FA) to distill the information in the CPC genes into six biologically relevant metagenes (Figure 1d, e). FA can be considered an extension of PCA in which an additional rotation maximizes variance of the gene weights. This additional rotation step results in a more even distribution of variance among components than does PCA alone. In general, FA is often preferred when the goal of the analysis is to understand and explain the structure in the data [28]. Using only the CPC genes in the combined reference data sets, we identified six factors that together explained 57% of the variance in the CPC genes (Additional file 3). In order to estimate the contribution of these factors in other data sets, we defined six *consistent expression indices* (CEIs) based on the sign of the non-trivial gene weights from each factor; thus each CEI comprises between 23 and 80 of the CPC genes (Additional file 3). At this point the CEIs were finalized, and in all subsequent analysis the CEIs were applied to the data sets without further adjustment. Thus, the CEIs were derived entirely from expression data, without consideration of any functional annotation or clinical outcome.

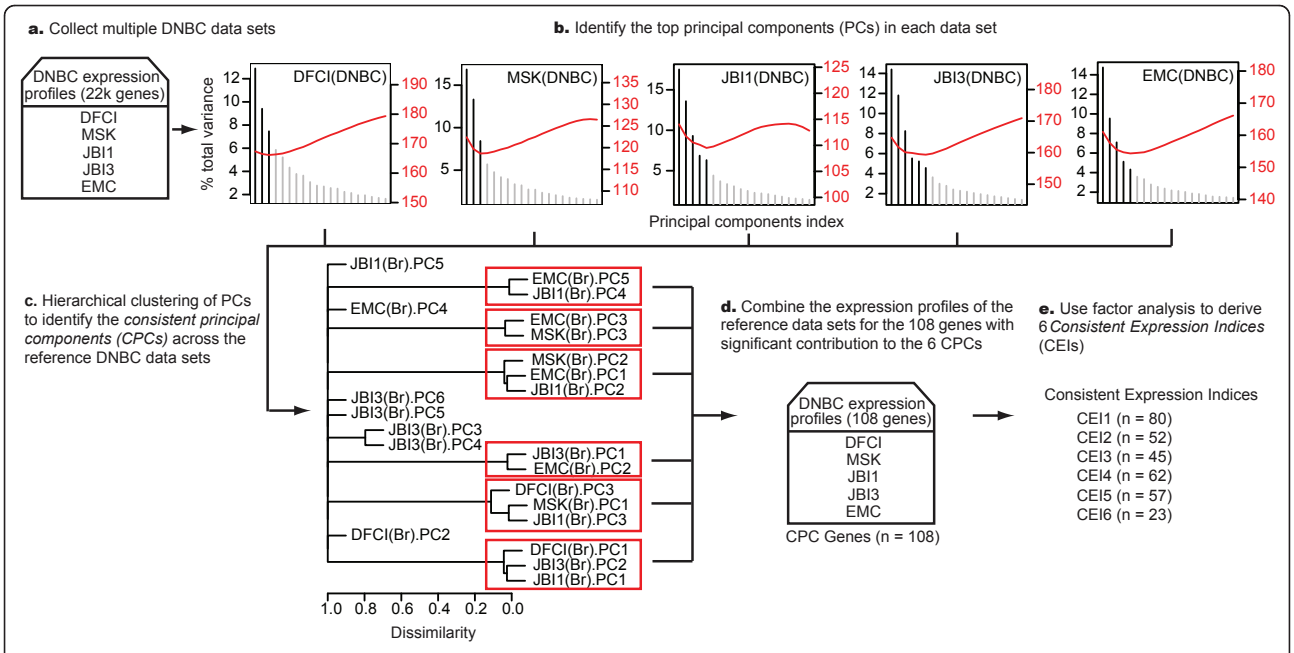

Figure 1 Schematic of CPC analysis and CEI derivation, showing results from DNBC.

Association between CEIs and clinical outcome in double-negative breast cancer

We hypothesized that the six CEIs, which account for highly conserved biological variation among DNBC cases in the five reference data sets, are also associated with certain clinical phenotypes of the tumors. We investigated whether the CEIs were predictive of response to specific treatment regimens in four independent test cohorts in which expression profiles were obtained from DNBC samples prior to neoadjuvant therapy (Table 1). Two of these cohorts, MDA1 [26] and MDA/MAQC [13], were similar: the samples were

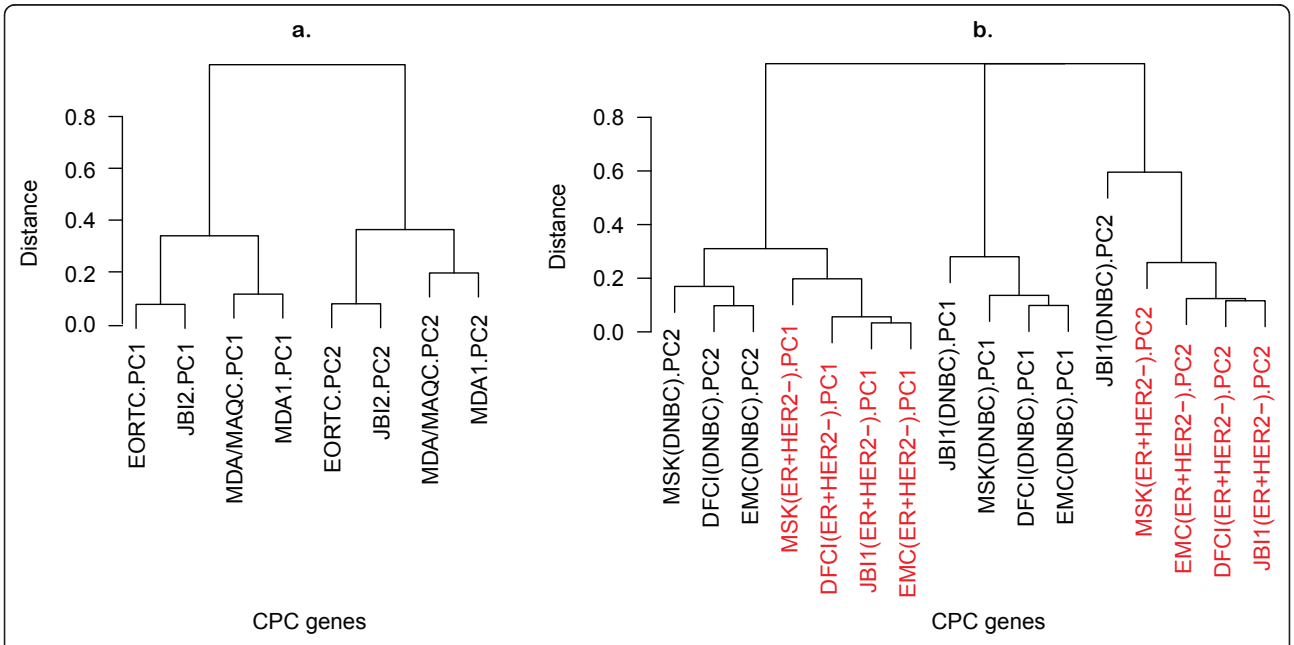

Figure 2 The CPC genes yield consistent, subtype-specific PCs in gene expression data sets. In each panel, PCA was performed separately on each data set using only the CPC genes, and the resulting first and second PCs from each data set were compared by hierarchical clustering. (**a**) The first two PCs of the 108 CPC genes in the DNBC subset of four validation data sets. (**b**) The first two PCs of the 108 CPC genes in the DNBC subset (black) and the ER-positive HER2-negative subset (red) of the four validation data sets.

Table 1 DNBC-derived CEIs are associated with tumor response to neoadjuvant chemotherapy in DNBC cohorts

				AUC					
cohort	regimen	patients	responders	CEI1	CEI2	CEI3	CEI4	CEI5	CEI6
EORTC	FEC	37	16	0.73^{R}*	0.57	0.51^{R}	0.61	0.56	0.54
MDA1	TFAC	27	13	0.78 **	0.62	0.77**	0.61^{R}	0.53	0.61
MDA/MAQC	TFAC	30	9	0.77*	0.66	0.78*	0.62^{R}	0.58	0.54
DFCI2	P	24	4	0.73	0.72^{R}	0.50	0.52^{R}	0.52^{R}	0.57^{R}
JBI2	E	43	4	0.85^{R}*	0.73^{R}	0.53	0.88**	0.58^{R}	0.72

Each CEI was evaluated as a univariate predictor of pathological complete response or residual disease using the area under the ROC curve (AUC). Chemotherapy regimens are indicated: A, doxorubicin; C, cyclophosphamide; E, epirubicin; F, 5-fluorouracil; P, either cisplatin or carboplatin; T, either paclitaxel or docetaxel. The CEIs were derived from four independent DNBC cohorts not shown in this table. * $P < 0.05$; ** $P < 0.01$. R: AUC is estimated based on association to residual disease (RD).

acquired by fine needle aspiration, and the patients received paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide (TFAC). In contrast, the two other data sets were derived from core biopsies; one cohort, EORTC, received fluorouracil, epirubicin and cyclophosphamide (FEC) [25], whereas the other cohort, JBI2, received only epirubicin [27] (Table 1).

We evaluated the association between pathologic complete response (pCR) and each of the six CEIs using area under the receiver operating characteristic (ROC) curves (AUC). In the MDA1 data set we observed a strong positive association between CEI1, CEI3 and pCR (AUC = 0.78, $P = 0.005$ for CEI1, AUC = 0.77, $P = 0.009$ for CEI3, Table 1). Similar associations were also observed in the second TFAC data set, MDA/MAQC (AUC = 0.77, $P = 0.02$ for CEI1, AUC = 0.78, $P = 0.001$ for CEI3, Table 1, Figure 3a, b).

In the two cohorts in which patients received neoadjuvant chemotherapy without taxane, we found CEI1 is significantly associated with residual disease (RD), a typical poor pathological response (AUC = 0.73, $P = 0.01$ in EORTC, AUC = 0.85, $P = 0.02$ in JBI2). On the other hand, there is no detectable association between CEI3 and response to either FEC or epirubicin treatment (Table 1, Figure 3c, d). These associations between CEIs and pathological responses in the validation cohorts was stronger than any we observed using published predictors [25,26] or using predictors we derived using conventional methods (Additional file 4).

Since pathological response to chemotherapy is based only on short-term follow-up, we also examined the association of these CEIs and long-term clinical outcome after chemotherapy. In a pooled DNBC cohort of 236 patients for which follow-up data is available (Additional file 1), of all the six CEIs, we found that binary classification based on CEI5 was significantly associated with disease-free survival of patients who received adjuvant chemotherapy within 10 years of follow-up ($HR = 2.70$, $P = 0.026$, Figure 4).

To test whether the CEIs were simply capturing known metagenes, we compared the six CEIs with 38 signatures reflecting tumor-associated biological processes or infiltrating cell types [25]. We used a meta-analysis based on seven data sets and found CEI1 was negatively correlated with ER/luminal-basal metagenes and ERBB2-molecular apocrine tumor metagenes; whereas CEI3 was positively correlated with the proliferation/AURKA metagene (Additional file 5). We also observed other correlations: CEI3 negatively correlated with the stroma and adipocyte metagenes. However, none of these metagenes was reported to hold similarly strong and consistent predictive power in the original studies as that of CEI1 and CEI3 [25] (Additional file 4). This may suggest that synergistic effects of multiple biological processes are more deterministic of the response to therapy than any single ones. In addition, CEI5 and CEI6 were not correlated with any of the known metagenes. Therefore, these two CEIs may reflect some biological processes relevant to DNBC but not yet described as such in any previous study.

Comparison with existing methods

In order to compare the performance of the CPC approach to existing algorithms, we assessed several supervised and unsupervised methods for their ability to generate metagenes predictive of treatment response.

For supervised methods, we first selected genes that are significantly associated with pathological response to taxane-based neoadjuvant therapy in the MDA1 data set based on Pearson's correlation coefficients, diagonal linear discrimination analysis [26,29], student's t-test, Wilcoxon's rank sum test, or nearest shrunken centroids [30]. We validated the predictive power of these metagenes in two other cohorts, MDA2 and EORTC. Metagenes based on Pearson correlation coefficients and nearest shrunken centroids yielded consistently significant predictions in the test data sets whereas the rest of the methods did not (Additional file 4). However, the predictive power represented by the area under the curves (AUCs) of all gene-by-gene methods decrease in the validation cohorts, suggesting overfitting..

For unsupervised methods, we pooled the five DNBC data sets and subjected it to independent component

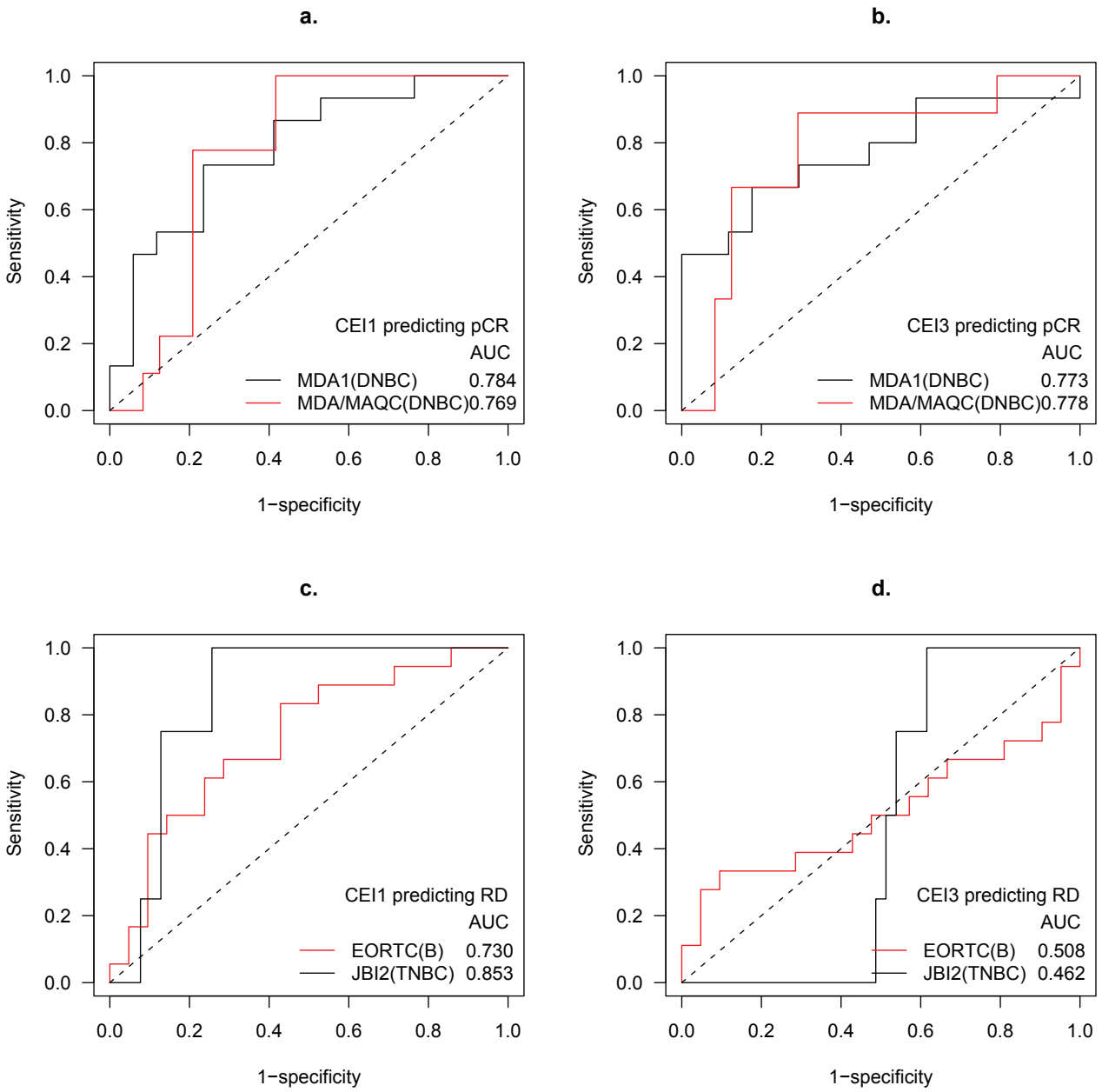

Figure 3 High CEI1 and CEI3 scores are associated with agent-specific response to neoadjuvant therapy. DNBC patients were given neoadjuvant TFAC (MDA1, MDA/MAQC), FEC (EORTC) or epirubicin only (JBI2). ROC curves indicate the association between (**a**) high CEI1 or (**b**) high CEI3 and pathological complete response (pCR) to taxane-based chemotherapy; and (**c**) high CEI1 or (**d**) high CEI3 and non-pCR to non-taxane-based chemotherapy.

analysis (ICA) [31] or sparse principal component analysis (SPCA) [32]. Three of the six top ICA components were predictive of pathological response in MDA1 and MDA2 data sets; and three of the six top SPCA components were predictive of pathological response in MDA1 and JBI2 data sets; whereas with the same number of components, consistent expression indices were predictive in four cohorts. More importantly, these methods produced less consistent results in terms of their predictive power in the two cohorts with similar treatment regimen. None of the components derived by ICA and SPCA, predicted the pathological response in the two taxane-based neoadjuvant trials (MDA1 and MDA/MAQC) in a consistent fashion. In particular, the third and fifth independent components (ICA3 and ICA5) predicted outcome the opposite direction, high values predicting favorable response in one and unfavorable response in the other cohort (Additional file 6).

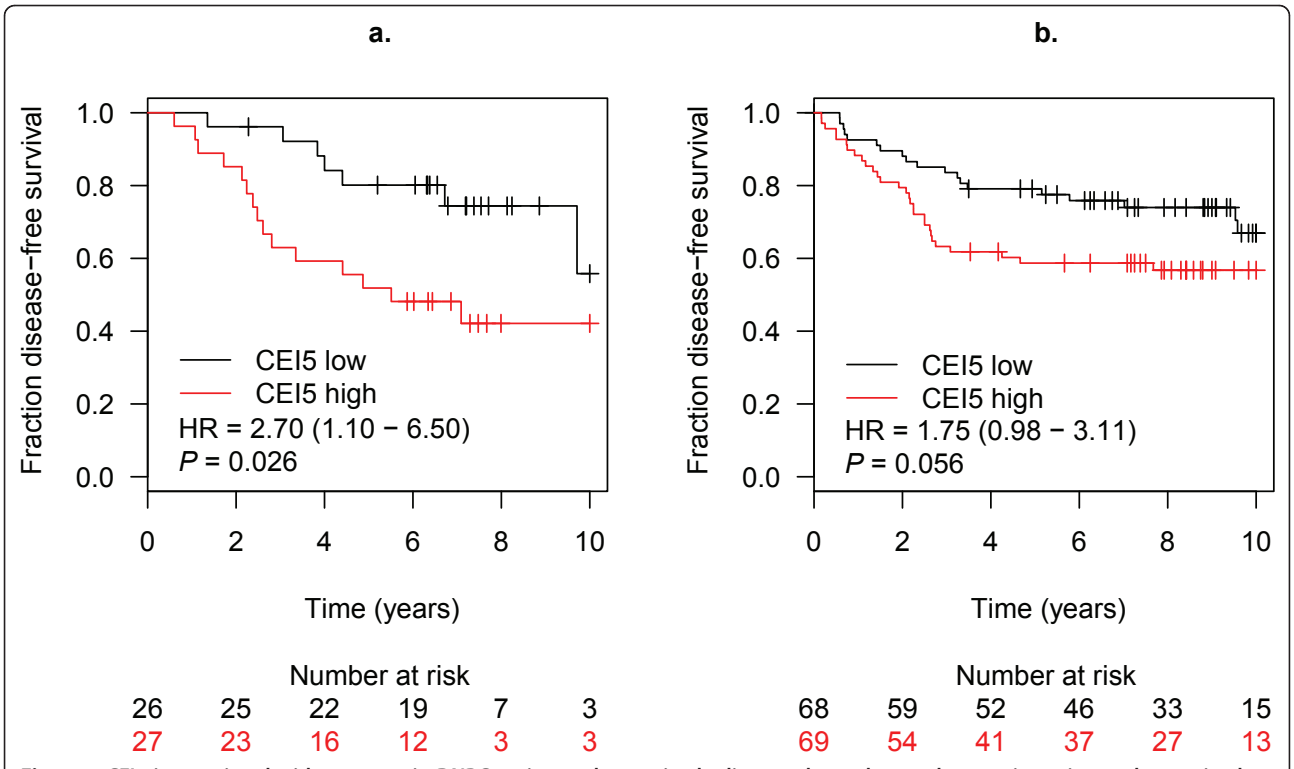

Figure 4 CEI5 is associated with outcome in DNBC patients who received adjuvant chemotherapy but not in patients who received no adjuvant chemotherapy. The EMC, JBI1, GIS, KUH, UCSF and NKI cohorts were combined, and patients were grouped according to subtype and presence of adjuvant therapy. Within each group, patients were stratified according to median of CEI scores, and disease-free survival was compared. (**a**) CEI5 in adjuvant treated tumors; (**b**) CEI5 in tumors without adjuvant therapy.

Other cancer types

ER-positive HER2-negative breast cancer

The ER-positive HER2-negative tumor is another major subtype of breast cancer and differs from DNBC in both transcriptional and genomic features [4]. Since some of the DNBC-derived CEIs may capture consistent biological variations common to both subtypes, we examined the association between the DNBC-derived CEIs and clinical outcome in ER-positive HER2-negative subsets of the validation cohorts. In a pooled cohort of 858 ER-positive HER2-negative tumors [9,21,33-36], binary classification based on CEI3 was significantly associated with disease-free survival in tamoxifen-treated patients (*HR* = 3.20, *P* = 0.016) as well as in patients not given tamoxifen treatment (*HR* = 1.8, *P* = 0.0004) (Additional file 7). Compared to DNBC, where CEI3 was associated with only pathological response to TFAC therapy but not long-term clinical outcome, the prognostic power of CEI3 in ER-positive HER2-negative tumors suggests that the same biological process, proliferation, may have different effects in the two different subtypes, which is concordant with previous translational studies performed in ER-positive tumors [3,37,38].

Ovarian cancer

Ovarian cancer is represented in only a limited number of microarray data sets and to the best of our knowledge there are no two analogous ovarian cancer data sets for which the same type of clinical outcome data is publicly available. Therefore, this type of cancer offered an opportunity to test our proposition that clinically relevant predictors can be extracted from data sets not associated with (and trained on) clinical outcome data.

We tested whether the CEIs derived from three stage III ovarian cancer data sets, EXPO‡, AOC and DU [39-41], predict treatment response or clinical outcome in other independent ovarian cancer cohorts (Additional file 3). In the BIDMC cohort [42], CEI1 derived from ovarian cancer was significantly associated with overall survival in 5 years after chemotherapy (HR = 8.36, *P* = 0.011, Additional file 8). Additionally, in the CRUK cohort [43], in which patients were assigned randomly to two groups treated with either paclitaxel or carboplatin monotherapy, CEI2 was associated with good response (pCR) to paclitaxel (AUC = 0.82, *P* = 0.02) but with poor response (RD) to carboplatin (AUC = 0.78, *P* = 0.09, Figure 5).

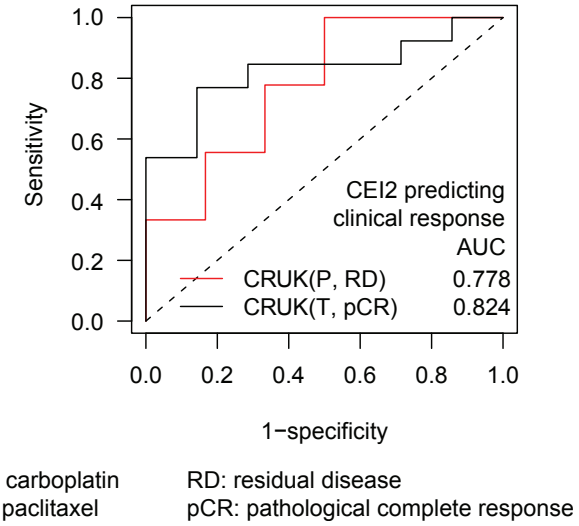

Figure 5 Consistent expression indices derived from stage III ovarian cancer are associated with treatment response. Ovarian cancer-derived CEI2 predicted pathological complete response (pCR) to paclitaxel monotherapy and non-pCR to carboplatin monotherapy in CRUK ovarian cancer cohort.

Lung adenocarcinoma

Finally, we turned our attention to lung adenocarcinoma, for which at least five microarray data sets are publicly available [39,44]. In a recent multi-site blinded validation study, at least eight gene expression based survival predictors were tested in two validation data sets, but none of these predicted clinical outcome in stage I cases in more than one data set unless clinical covariates were included [44]. Therefore, we applied the same strategy to early stage lung cancer. In order to test our method within the same analytical framework of the original study we applied a cross-validation approach in the four lung cancer cohorts by extracting CEIs from each combination of three cohorts (using early stage samples only) and testing for association between these lung cancer-derived CEIs and outcome in the remaining cohort (for stage I only). In three of the four rounds of the validation, at least one of the CEIs were significantly predictive of outcome in stage I lung cancer in the validation cohort, without the use of further clinical variables and without any training on outcome (Additional file 8). Furthermore, we derived four CEIs from all four lung cancer data sets (early stage only, Additional file 3) and tested them on a fifth independent lung cancer cohort [39] and found that CEI1 was predictive of 5-year overall survival in stage I samples (HR = 7.73, P = 0.034, Additional file 8).

To understand the biology underlying the predictive power of these CEIs, we tested for enrichment of Gene Ontology (GO) annotations for biological processes in the CPC genes. For the CPC genes of the DNBC derived CEIs, the most enriched GO categories included immune and inflammatory response. For the lung cancer derived CEIs, the top categories included digestion, response to external stimulus, and oxidation/reduction (Additional file 9). While the GO category analysis did not provide an easy interpretation of the observed predictive power of clinical behavior, a literature analysis identified several genes that were linked to specific chemotherapy response or resistance mechanisms, including GPX3 [45], HPGD [46], AKR1C1, and AKR1C2 [47].

Discussion

We have presented a method to extract metagenes that consistently distinguish among individual double-negative breast cancers in multiple gene expression data sets. We found a strong association between three of the six CEIs and the efficacy of various neoadjuvant treatments in DNBC. This association was stronger than that of previously published predictors and suggests that these gene sets reflect important biological processes that influence sensitivity to chemotherapy. Importantly, different CEIs were predictive of different regimens. Furthermore, some CEIs were predictive only in DNBC and not in ER-positive tumors.

An attractive feature of the method presented here is that it is unsupervised; i.e. the CEIs are derived without information about clinical response or outcome. This holds particular importance for cancer types with only a few existing clinical outcome matched microarray based cohorts [48]. In the case of cancer types of higher incidence and easier access to clinical material (e.g. breast, lung), multiple analogous cohorts complete with clinical outcome data, often up to six or seven independent data sets, are available for supervised analysis to identify individually informative genes. These genes could then be combined into multi-gene prediction models and independently validated on the various cohorts. In the case of other cancer types (pancreas, prostate, etc.), lower incidence, difficulties with obtaining appropriate RNA material, or the specific clinical course of the disease results in a lack of clinical outcome matched microarray data sets. In such cases a method that is able to extract potential outcome predictors without training on outcome data may provide a potential solution. Given the observation that CEIs may already hold predictive value without being fitted to the actual clinical outcome, CPC-based methods may extract testable predictors from microarray data without matched clinical outcome, and the few outcome matched microarray cohorts could then be used for independent validation. For example, prostate cancer is represented by at least fourteen microarray cohorts, but only three of these have clinical outcome published as well [49-52].

Although biological functions of the CEIs can be partially understood by methods such as GO analysis, our

knowledge about these genes still remains very limited. There might be several reasons for this. First, many of the genes listed in the CEIs have not been investigated in detail for direct involvement in drug resistance mechanisms. Second, drug resistance might be the result of a distinct but complex biological feature which involves a concert of relevant biological mechanisms, such as increased expression of multidrug resistance genes, low proliferation rate, and the combination of these mechanisms might be best quantified by common upstream and downstream markers that reflect the expression level the relevant biological mechanisms. In general, it is desirable for clinical predictors to be associated with uniquely identifiable biological mechanisms so as for therapeutic targetability. However, we emphasize that our approach was designed to overcome the failure of single gene, single biological mechanism prediction of clinical outcome [53]. We aimed at determining and testing the utility of the most robust and consistent information in high throughput data sets, which is more likely to capture the most comprehensive and dominant biological variations in human tumors rather than any single unique biological process from limited prior knowledge.

The predictors presented in this paper would need to be refined before introduction into clinical practice. Currently each CEI comprises up to 235 genes, a number that might be impractical for a clinical test such as multiple quantitative PCR. Also, treatment decisions are dichotomous; a patient either receives a particular treatment or does not. Therefore, the most useful clinical tests have decision thresholds, which will need to be determined for the CEIs and will need to be validated in independent cohorts to establish the sensitivity and specificity of a future treatment response test.

Conclusion

The approach we described in this analysis is well-suited to identify linear gene combinations that express consistent variations in a set of independent but biologically similar datasets, regardless of the observed clinical outcome. The ability of these metagenes to predict response to chemotherapy has been evaluated in completely independent set of cohorts. Unlike other existing unsupervised methods, by mandating the consistency of the weights of genes in the loading matrix, the consistent principal components are more likely to yield reproducible predictive power.

Methods

Data sets

All microarray data sets used in this study were previously published and are available from several public data repositories, except for the BIDMC ovarian cancer data set, which was obtained from the authors [42]. Each microarray data set was processed with RMA [54]. For each cohort, a list of samples used in the analysis is provided in Additional file 1.

To determine the double-negative breast cancer (DNBC, not expressing ESR1 or HER2), we clustered each data set based on the probe levels of ESR1 and HER2 using the Partitioning Around Medoids (PAM) algorithm. The DNBC is determined by the cluster with consistent low expression of both genes.

Consistent Principal Components Analysis

For each of the reference data sets independently, we computed the coefficient of variation (CV) based on the anti-logarithm of RMA probe levels and kept probe sets with a CV greater than one and less than 1000; thus we selected 614 to 1714 probe sets from each data set. Next we performed PCA on these highly variable probe sets in each data set, and selected an optimal number k of top PCs by the minimum of the BIC:

$$\mathrm{BIC} = n\ln\left(\frac{v}{n}\right) + k\ln(n)$$

Here, n is the number of samples, k is the number of components selected, and v is the unexplained variance which equals the residual sum of squares, given by:

$$v = \sum_{i=1}^{p} \sigma_i^2 - \sum_{j=1}^{k} \omega_j^2$$

Here, σ_i is the standard deviation of probe set i, p is the number of probe sets, and ω_j is the standard deviation explained by PC j (equal to the square root of the j'th eigenvalue). For each PC, we calculated the Pearson correlation coefficient (PCC) between its component scores and the expression level of each probe set and the significance of the correlation is assessed by Student's t-test. Probe sets with a $P < 0.01$ for PCC were selected to represent the PC. After the selection, each PC contains 42 to 211 representative probe sets.

To compare PCs derived from various data sets, we defined the following measure of the dissimilarity between PCs i and j:

$$D_{ij} = (1 - J_{ij}) \times (1 - C_{ij})$$

Where J_{ij} is the Jaccard index (the ratio between size of the intersection and the size of the union of the representative probe sets of component i and j) and C_{ij} is the cosine correlation coefficient between the weights of the common representative probe sets of component i and j.

We used this distance function to perform average linkage hierarchical clustering on the selected PCs from

all reference data sets. For each distinct cluster, we selected the set of genes found in at least two members.

Factor analysis and CEI calculation

We retrieved the RMA expression profile of the CPC genes from the reference data sets. When a gene was represented by multiple probe sets, we selected the probe set with largest standard deviation to represent that gene. For each of the expression matrices retrieved, we computed the standard z-scores for each gene and merged the matrices into one.

We performed factor analysis of the merged z-scores using the "varimax" rotation and with the number of factors set to six [28]. For each factor we estimated the gene coefficients using the least-square method. Coefficients with an absolute value below 0.1 were set to zero, and the signs of the coefficients were used as the gene weights in the corresponding CEI.

Prediction and prognosis

The ROC curves were based on individual CEI scores and treatment response. We calculated the area under the curve (AUC) using the trapezoidal rule [55] and estimated statistical significance using the Wilcoxon rank sum test. Survival curves were generated using the Kaplan-Meier method. Hazard ratios were estimated for 5 year or 10 year follow-up by Cox regression in which the patients were stratified into two groups of equal size according to the median of the CEI score. Statistical significance was estimated using the log rank test.

Further details are available in Additional file 2.

Additional material

Additional file 1: Summary of the tumor expression data sets used in this study. **(a)** Summary of all data sets used in this manuscript; **(b)** The number of DNBC samples from each data set used in each figure; **(c)** The number of ER-positive/Her2-negative breast cancer samples from each data set used in each figure; **(d)** The number of ovarian cancer samples from each data set used in each figure; **(e)** The number of lung cancer samples from each data set used in each figure.

Additional file 2: Supplementary methods. Supplementary methods.

Additional file 3: CEIs derived from three tumor types. CEIs derived from DNBC, Stage III ovarian cancer and early-stage lung cancer by consistent principal component analysis.

Additional file 4: AUCs and P values for prediction of TFAC response. Summary of AUCs and P values for prediction of TFAC response in DNBC using published metagenes and signatures derived using various supervised methods.

Additional file 5: Correlation between DNBC-derived CEIs and known metagenes. Colorgram showing the pooled Pearson correlation coefficients between DNBC-derived CEIs and known metagenes.

Additional file 6: AUCs for prediction of pathological response in five DNBC cohorts which received neoadjuvant chemotherapy of different regimens using various unsupervised methods. (a) CEIs derived from consistent principal components; (b) Components derived using independent component analysis; (c) Components derived using sparse principal component analysis. The pooled correlation coefficients were estimated from seven breast cancer data sets based on a meta-analysis.

Additional file 7: DNBC-derived CEI3 predict clinical outcome of in ER-positive HER2-negative breast cancer. (a) ER-positive HER2-negative samples which received endocrine or radio-therapy from the EMC, JBI1, GIS, KUH, UCSF and NKI cohorts; (b) ER-positive HER2-negative samples which received no systematic therapy.

Additional file 8: Validation of the association between CEIs and clinical outcomes in ovarian cancers and lung cancers. **(a)** Hazard ratios based on 5-year follow-up of three ovarian cancer-derived CEIs in the validation cohort (DU) based on univariate and multivariate Cox regression; **(b)** Summary cross-validation of CEIs derived from three early-stage lung cancer data sets and validated in the fourth for the association to clinical outcomes; **(c)** hazard ratios based on 5-year follow-up of seven lung cancer-derived CEIs in the validation cohort (DU) based on univariate and multivariate Cox regression.

Additional file 9: Gene Ontology (GO) annotation analysis. Gene Ontology of CEIs derived from **(a)** DNBC, from **(b)** stage III ovarian cancer, and from **(c)** early stage lung cancer.

Acknowledgements and Funding

This work was supported in part by the National Institutes of Health through grant 1PO1CA-092644-01 and by the Breast Cancer Research Foundation (ZS, ALR), by the Danish Council for Independent Research, Medical Sciences (FSS) (ZS, ACE), and by BioSim (NoE), FP6, LSHB-CT-2004-005137 (QL) and by the Harvard SPORE in breast cancer CA089393 (ZS, ALR).
We thank Wiktor Mazin for his suggestions, and Dimitrios Spentzos and Towia Libermann for providing the BIDMC data set.
expO data set was obtained from the International Genomic Consortium, http://www.intgen.org/expo/

Author details

[1]Center for Biological Sequence Analysis, Department of Systems Biolology, Technical University of Denmark, 2800 Lyngby, Denmark. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [3]Medical Oncology Department, Jules Bordet Institute, Brussels, 1000, Belgium. [4]Department of Biostatistics, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [5]Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA. [6]Department of Breast Medical Oncology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA. [7]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA. [8]Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology (CHIP@HST), Harvard Medical School, Boston, MA 02115, USA.

Authors' contributions

QL conceived the study, analyzed the data and helped draft the manuscript; ACE, NJB participated in the data analysis and helped draft the manuscript; CD, BH and CS contributed data and participated in the data analysis; WFS, LP contributed data; SB helped draft the manuscript; ALR contributed data and helped draft the manuscript; ZS conceived the study and drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 29 March 2011 Accepted: 28 July 2011
Published: 28 July 2011

References

1. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365(9458)**:488-492.
2. Fan X, Shi L, Fang H, Cheng Y, Perkins R, Tong W: **DNA microarrays are predictive of cancer prognosis: a re-evaluation.** *Clin Cancer Res* 2010, **16(2)**:629-636.

3. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C: **Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes.** *Clin Cancer Res* 2008, **14(16)**:5158-5165.
4. Sotiriou C, Pusztai L: **Gene-expression signatures in breast cancer.** *N Engl J Med* 2009, **360(8)**:790-800.
5. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871)**:530-536.
6. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98(24)**:13790-13795.
7. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo JH: **Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy.** *J Clin Oncol* 2004, **22(14)**:2790-2799.
8. 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: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351(27)**:2817-2826.
9. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M: **Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.** *J Natl Cancer Inst* 2006, **98(4)**:262-272.
10. Jansen MP, Foekens JA, van Staveren IL, Dirkzwager-Kiel MM, Ritstier K, Look MP, Meijer-van Gelder ME, Sieuwerts AM, Portengen H, Dorssers LC, Klijn JG, Berns EM: **Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling.** *J Clin Oncol* 2005, **23(4)**:732-740.
11. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC: **A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.** *Cancer Cell* 2004, **5(6)**:607-616.
12. Oh DS, Troester MA, Usary J, Hu Z, He X, Fan C, Wu J, Carey LA, Perou CM: **Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers.** *J Clin Oncol* 2006, **24(11)**:1656-1664.
13. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L: **Effect of training-sample size and classification difficulty on the accuracy of genomic predictors.** *Breast Cancer Res* **12(1)**:R5.
14. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavare S, Brenton JD, Ylstra B, Caldas C: **High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biol* 2007, **8(10)**:R215.
15. Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M, Cristofanilli M, Hortobagyi GN, Pusztai L: **Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.** *J Clin Oncol* 2008, **26(8)**:1275-1281.
16. Doane AS, Danso M, Lal P, Donaton M, Zhang L, Hudis C, Gerald WL: **An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen.** *Oncogene* 2006, **25(28)**:3994-4008.
17. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, Klijn JG, Larsimont D, Buyse M, Bontempi G, Delorenzi M, Piccart MJ, Sotiriou C: **Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.** *J Clin Oncol* 2007, **25(10)**:1239-1246.
18. Lu X, Lu X, Wang ZC, Iglehart JD, Zhang X, Richardson AL: **Predicting features of breast cancer with gene expression patterns.** *Breast cancer research and treatment* 2008, **108(2)**:191-201.
19. Matros E, Wang ZC, Lodeiro G, Miron A, Iglehart JD, Richardson AL: **BRCA1 promoter methylation in sporadic breast tumors: relationship to gene expression profiles.** *Breast cancer research and treatment* 2005, **91(2)**:179-186.
20. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: **X chromosomal abnormalities in basal-like human breast cancer.** *Cancer Cell* 2006, **9(2)**:121-132.
21. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365(9460)**:671-679.
22. Pusztai L, Ayers M, Stec J, Clark E, Hess K, Stivers D, Damokosh A, Sneige N, Buchholz TA, Esteva FJ, Arun B, Cristofanilli M, Booser D, Rosales M, Valero V, Adams C, Hortobagyi GN, Symmans WF: **Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors.** *Clin Cancer Res* 2003, **9(7)**:2406-2415.
23. Gong Y, Yan K, Lin F, Anderson K, Sotiriou C, Andre F, Holmes FA, Valero V, Booser D, Pippen JE Jr, Vukelja S, Gomez H, Mejia J, Barajas LJ, Hess KR, Sneige N, Hortobagyi GN, Pusztai L, Symmans WF: **Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study.** *Lancet Oncol* 2007, **8(3)**:203-211.
24. Kreike B, van Kouwenhove M, Horlings H, Weigelt B, Peterse H, Bartelink H, van de Vijver MJ: **Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas.** *Breast Cancer Res* 2007, **9(5)**:R65.
25. Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Becette V, Andre S, Piccart M, Campone M, Brain E, Macgrogan G, Petit T, Jassem J, Bibeau F, Blot E, Bogaerts J, Aguet M, Bergh J, Iggo R, Delorenzi M: **A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer.** *Nat Med* 2009, **15(1)**:68-74.
26. 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: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.** *J Clin Oncol* 2006, **24(26)**:4236-4244.
27. Li Y, Zou L, Li Q, Haibe-Kains B, Tian R, Desmedt C, Sotiriou C, Szallasi Z, Iglehart JD, Richardson AL, Wang ZC: **Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer.** *Nat Med* 2010, **16(2)**:214-218.
28. Bartlett M: **The statistical conception of mental factors.** *British Journal of Psychology (Statistics Section)* 1937, **28**:97-104.
29. Hastie T, Tibshirani R, Friedman J, Franklin J: **The elements of statistical learning: data mining, inference and prediction.** *The Mathematical Intelligencer* 2005, **27(2)**:83-85.
30. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA* 2002, **99(10)**:6567-6572.
31. Hyvärinen A, Karhunen J, Oja E: **Independent Component Analysis.** New York: Wiley; 2001.
32. Zou H, Hastie T, Tibshirani R: **Sparse Principal Component Analysis.** *Journal of Computational and Graphical Statistics* 2006, **2(15)**:22.
33. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10(6)**:529-541.
34. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, Wong JE, Liu ET, Bergh J, Kuznetsov VA, Miller LD: **Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer.** *Cancer Res* 2006, **66(21)**:10292-10301.
35. Pawitan Y, Bjohle 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, Wedren S, Bergh J: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Res* 2005, **7(6)**:R953-964.

36. 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: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347(25)**:1999-2009.
37. Chang J, Powles TJ, Allred DC, Ashley SE, Makris A, Gregory RK, Osborne CK, Dowsett M: **Prediction of clinical outcome from primary tamoxifen by expression of biologic markers in breast cancer patients.** *Clin Cancer Res* 2000, **6(2)**:616-621.
38. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, Goldstein DR, Piccart M, Delorenzi M: **Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.** *Breast Cancer Res* 2008, **10(4)**:R65.
39. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439(7074)**:353-357.
40. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, Traficante N, Fereday S, Hung JA, Chiew YE, Haviv I, Gertig D, DeFazio A, Bowtell DD: **Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome.** *Clin Cancer Res* 2008, **14(16)**:5198-5208.
41. **International Genomics Consortium.** [http://www.intgen.org/expo/].
42. Spentzos D, Levine DA, Kolia S, Otu H, Boyd J, Libermann TA, Cannistra SA: **Unique gene expression profile based on pathologic response in epithelial ovarian cancer.** *J Clin Oncol* 2005, **23(31)**:7911-7918.
43. Ahmed AA, Mills AD, Ibrahim AE, Temple J, Blenkiron C, Vias M, Massie CE, Iyer NG, McGeoch A, Crawford R, Nicke B, Downward J, Swanton C, Bell SD, Earl HM, Laskey RA, Caldas C, Brenton JD: **The extracellular matrix protein TGFBI induces microtubule stabilization and sensitizes ovarian cancers to paclitaxel.** *Cancer Cell* 2007, **12(6)**:514-527.
44. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, *et al*: **Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.** *Nat Med* 2008, **14(8)**:822-827.
45. Saga Y, Ohwada M, Suzuki M, Konno R, Kigawa J, Ueno S, Mano H: **Glutathione peroxidase 3 is a candidate mechanism of anticancer drug resistance of ovarian clear cell adenocarcinoma.** *Oncol Rep* 2008, **20(6)**:1299-1303.
46. Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M: **Relevance network between chemosensitivity and transcriptome in human hepatoma cells.** *Mol Cancer Ther* 2003, **2(2)**:199-205.
47. Wsol V, Szotakova B, Martin HJ, Maser E: **Aldo-keto reductases (AKR) from the AKR1C subfamily catalyze the carbonyl reduction of the novel anticancer drug oracin in man.** *Toxicology* 2007, **238(2-3)**:111-118.
48. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data.** *PLoS Biol* 2004, **2(4)**:E108.
49. Best CJ, Gillespie JW, Yi Y, Chandramouli GV, Perlmutter MA, Gathright Y, Erickson HS, Georgevich L, Tangrea MA, Duray PH, Gonzalez S, Velasco A, Linehan WM, Matusik RJ, Price DK, Figg WD, Emmert-Buck MR, Chuaqui RF: **Molecular alterations in primary prostate cancer after androgen ablation therapy.** *Clin Cancer Res* 2005, **11(19 Pt 1)**:6823-6834.
50. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39(1)**:41-51.
51. Gregg JL, Brown KE, Mintz EM, Piontkivska H, Fraizer GC: **Analysis of gene expression in prostate cancer epithelial and interstitial stromal cells using laser capture microdissection.** *BMC Cancer* 2010, **10**:165.
52. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL: **Gene expression profiling predicts clinical outcome of prostate cancer.** *J Clin Invest* 2004, **113(6)**:913-923.
53. Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB: **Independent component analysis: Mining microarray data for fundamental human gene expression modules.** *J Biomed Inform* 2010.
54. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2)**:249-264.
55. Burden RL, Faires JD: **Numerical Analysis.** Brooks/Cole;, 7 2000.

doi:10.1186/1471-2105-12-310
Cite this article as: Li *et al.*: Consistent metagenes from cancer expression profiles yield agent specific predictors of chemotherapy response. *BMC Bioinformatics* 2011 **12**:310.