Technical Report BCF-SIB 2007-1, 11 March 2007
Bioinformatics Core Facility, Swiss Institute of Bioinformatics
Lausanne, Switzerland

Integrative analysis of gene-expression profiles: toward a unified understanding of breast cancer subtyping and prognosis signatures

Pratyaksha Wirapati[1], Susanne Kunkel[1], Darlene R. Goldstein[1,2], Pierre Farmer[1,3], Sylvain Pradervand[4], Benjamin Haibe-Kains[5,6], Christine Desmedt[5], Thierry Sengstag[1,3], Frédéric Schütz[1], Martine Piccart[5], Christos Sotiriou[5] & Mauro Delorenzi[1,3]

[1]Swiss Institute of Bioinformatics, University of Lausanne, 1015 Lausanne, Switzerland. [2]Institute de Mathématiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. [3]National Centres of Competence in Research (NCCR) Molecular Oncology, Swiss Institute for Experimental Cancer Research, 1066 Epalinges, Switzerland. [4]DNA Array Facility, Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. [5]Translational Research Unit, Jules Bordet Institute, 121 Boulevard de Waterloo, 1000 Brussels, Belgium. [6]Machine Learning Group, Computer Science Department, Université Libre de Bruxelles (ULB), 1050 Brussels, Belgium.
Correspondence should be addressed to P.W. (Pratyaksha.Wirapati@isb-sib.ch).

Breast cancer subtyping and prognosis have been extensively studied by gene-expression profiling, resulting in disparate signatures with little overlap in their constituent genes. The roles of individual genes in a signature, the equivalence of various signatures and their relation to conventional prognostic factors are still unclear. Here, we analyzed publicly available expression data from 2833 breast tumors to uncover consistent patterns across independent cohorts and microarray platforms. The common thread unifying various signatures was revealed using coexpression modules associated with important processes in breast cancer. These modules were used to consolidate tumor subtyping. The low-proliferative subset of ER+/ERBB2- tumors were found to have risk of relapse low enough to be spared aggressive chemotherapy. Previously published prognostic signatures were dissected by characterizing the prognostic values and modular coexpression of their constituent genes. When applied to the dataset collection, most of these signatures showed similar prognostic power largely attributable to proliferation genes. These signatures concordantly assigned the low-proliferative subset of ER+/ERBB2- tumors to the low-risk group, recapitulating the classifications by coexpression modules. This study introduced a framework for uncovering consistent relationships in diverse gene-expression datasets and dissecting expression signatures.

1 Breast cancer is the disease most extensively studied by gene-expression profiling of primary tumors from patient populations[1–21]. However, the results are still fragmented. Disparate signatures had been proposed, either directly from breast cancer expression profiles,[1,3,4,11,15,16,22] or translated from model systems[8,23], with little agreement in the constituent genes. The relationship between "intrinsic subtypes" defined by cluster analysis[13] and prognostic signatures defined by associations with patient outcome needs to be clarified. Furthermore, the roles of individual genes in a signature and their biological interpretation are often unclear[24,25]. Fan *et al.*[26] recently compared the prognostic ability of the intrinsic subtypes and several prognostic signatures. They noted concordance in the risk classification, suggesting potential equivalence between some of these signatures. However, this study was limited to only one dataset and did not fully elucidate how the different genes were related to one another and to biological processes known to be prognostic, such as proliferation[15,27]. Lastly, since conventional biomarkers and clinical variables have been used extensively for prognosis[28–30], their redundancy or complementarity to expression signatures needs to be investigated.

2 The main concern with gene-list discrepancies among various signatures is that the biological and clinical importance of the genes might not be real[25], but artifacts of study design and analysis[31]. Although it is possible that some of the disagreements are caused by a few falsely identified genes, disparate gene lists may also arise when the number of truly prognostic genes is much larger than the number of genes in the signatures[32], which is often designed to be parsimoniously small for practical applications. Thus, arbitrary subsets of the true gene list may be selected as equivalent signatures. In this case, the gene-list discrepancy is not a problem in itself. Still, it is important to identify genes that are artifacts and to understand the equivalence using a more comprehensive gene list. An example of such list is the set of mainly proliferation genes associated with histological grade in breast cancer[15], whose expressions are also highly correlated with one another. Several prognostic signatures[1,3,11,22,33] contain genes from this list, and a signature made from eleven genes

overlapping with the 70-gene signature [1] remained strongly prognostic [15].

3 Here, we analyze the functioning of various signatures not only by comparing their prognostic performance but also by characterizing their constituent genes and how they contribute to the prognostic power of the whole signatures. The concept of "coexpression modules" (comprehensive lists of genes with highly correlated expression) was used extensively to reveal the common thread connecting various subtyping and prognostic signatures, as well as conventional prognostic factors. In addition to the strongly-prognostic proliferation module, we identified modules associated with other important processes in breast cancer: estrogen receptor (ESR1 or ER) signaling, amplification of epidermal growth factor receptor tyrosine kinase 2 (ERBB2 or *her2/neu*), tumor invasion and immune response. Meta-analytical approaches were used to uncover relationships that are consistent in a large collection of public datasets [1–21], and thus unlikely to be artifacts of specific cohorts or microarray platforms.

RESULTS

Compilation of public breast cancer datasets

4 We collected publicly available datasets from journal articles and repositories such as Gene Expression Omnibus (GEO) and ArrayExpress, selecting those with medium to large sample size (Table 1). Since publications sometimes used the same patients, datasets with unique patients were introduced (identified by the "dataset symbols" in Table 1) by merging some original datasets or removing redundant patients. The collection includes datasets produced on whole-genome microarrays, small diagnostic arrays and RT-PCR panels, totaling 2865 expression profiles. Small numbers of non-malignant samples (normal breast tissue or fibroadenoma) are present in some datasets. Almost all malignant tumors are invasive ductal carcinoma.

5 The hybridization probes were remapped to Entrez gene identifiers (GeneID)[34] through sequence alignment against the well-curated subset of RefSeq mRNA sequence database. The numbers of distinct GeneIDs obtained for each dataset are shown in Table 1. Amongst the genomic arrays, only 1963 genes were present in all platforms. To avoid discarding useful information about many genes, meta-analyses were performed on the union of all 17198 genes. Summary statistics of absent genes were considered as missing values.

Dataset heterogeneity and meta-analytical approach

6 The design of the analysis and the interpretation of the results should take into account the heterogeneity of patient characteristics of the datasets. Summaries of important clinical variables are shown in figure 1. The distributions of age at diagnosis, ER status, tumor grade and tumor size are similar, with few exceptions; while those of lymph-node status and adjuvant treatment are more heterogeneous. By design, some datasets (such as TRANSBIG and EMC) consist entirely of untreated, lymph-node negative patients. Of note is the wide variation of survival profiles across studies (Figure 1b), which can be partly explained by the patient selection criteria, such as advanced carcinoma (mostly large tumor size) in STNO.

7 Pooling patients from heterogeneous datasets to treat them as if they were from a single cohort may result in false associations due to "Simpson's paradox" [35]. Therefore, we chose to stratify all analyses by dataset and to combine only summary statistics (such as Z-scores of regression models) [36]. This approach also circumvents the problem in combining potentially incommensurable expression measures from different microarray datasets. The Z-scores are not affected by arbitrary shifting or scaling of the expression data matrix of each dataset.

Prototype-based coexpression module analysis

8 To identify coexpression modules associated with specific biological processes, we devised a supervised approach where a handful of "prototype" genes were selected based on biological knowledge about breast cancer [30] and previous results of expression studies. Expression values of these prototype genes are then used simultaneously as explanatory variables in regression models, to group other genes according to their coexpression with the respective prototype (**Methods**). We considered five key processes: estrogen receptor signaling, ERBB2 amplification, proliferation, invasion and immune response. The genes chosen as their prototypes were, respectively, ESR1, ERBB2, AURKA (aurora-related kinase 1; also known as STK6 or STK15), PLAU (urokinase-type plasminogen activator; uPA) and STAT1 (signal transducer and activator of transcription 1). Other choices of well-known genes for the prototypes, such as GATA3-GRB7-CCNB2-MMP11-MX1, did not affect the overall conclusions of this study.

9 Using the meta-analysis scheme outlined in Figure 2a, we identified genes associated with each prototype (**Supplementary Table 1**). The coexpression patterns of the genes are shown by two example heatmaps in Figure 2b (see **Supplementary Result 2** for complete results). Each module contains highly correlated or anticorrelated genes, as shown by the vertical color patterns. The annotation of the modules shows that they correspond well to the expected biological processes (see details in **Supplementary Result 2.4**).

10 The correlated expression measures in a module provide redundant information about the module's overall expression in a tumor. They can be summarized into a single number by averaging (**Methods**). We called the resulting value a "module score". Figure 3a examines the relationship between the module scores and some clinical variables. We see the expected associations between estrogen score and ER status, between proliferation score and histological grade, as well as between immune-response score and lymphocytic infiltration according to pathological data.

11 Subsequent analyses are focused on the estrogen, ERBB2-amplification and proliferation modules to clarify several important aspects of breast cancer subtyping and prognosis. Detailed analyses involving the invasion (PLAU) and immune-response (STAT1) modules, which are not essential for supporting our conclusions, will be reported elsewhere.

Module scores for tumor subtyping

12 "Subtyping" refers to tumor classification according to naturally existing clusters, typically using hierarchical cluster analysis. In breast cancer, several versions of subtypes have been developed [10,12–14]. The most well-known [13], the "intrinsic subtypes", divide breast tumor into five groups: "basal-like", *her2/neu* (ERBB2), luminal A, luminal B and "normal-like". More recently, Kapp *et al.* [37] redefined tumor subtyping using pairs of genes, such as BCMP11/ABCC11 (shown in ESR1 and ERBB2 modules of Figure 2), and found that only three subtypes were consistently observed, corresponding to the combina-

tions of conventional marker status ER-/ERBB2-, ERBB2+ and ER+/ERBB2-. Intriguingly, while luminal A and B subtypes are both parts of the ER+/ERBB2- subtype under this scheme, they have very different survival [13,26].

13 We re-examined tumor subtyping using our module scores as the variables. Instead of using exploratory methods such as hierarchical cluster analysis, we applied the more rigorous Gaussian mixture models [38] to identify natural clusters of tumors. The dot histograms of the estrogen and ERBB2-amplification scores (Figure 3a) show significant bimodality that is consistent across datasets (see **Supplementary Result 3** for complete datasets and bimodality tests). Surprisingly, when the two scores are combined (Fig. 3b) we see only three clusters, instead of four clusters that would have been observed if the two scores had been independent (**Supplementary Result 4**). The relative positions of the clusters are reproducible across datasets and the cluster with high ERBB2-amplification score showed intermediate levels of estrogen score. For brevity, the clusters will be subsequently referred to as tumor type 1, 2 and 3. They respectively correspond to the intrinsic subtypes of basal-like, *her2* and combined luminal A/B (see dataset UNC, NKI and STOCK in Figure 3b). They can also be related to the subtypes of Kapp *et al.* [37] by the estrogen and ERBB2-amplification scores. Our approach extends the lists of genes that are specifically expressed in each subtype (see ESR1 and ERBB2 modules in Figure 2b). Furthermore, type-1 tumors can be defined in a positive manner by the overexpression of genes such as LMO4, FOXC1 and EGFR, instead of by the absence of ER and ERBB2 expression.

14 Figure 3c shows that while type-1 and type-2 tumors have mostly high proliferation scores; type-3 tumors have a wide range of values, encompassing the low values of normal breast tissue (see dataset UNC) and the high values typical for tumor type 1 and 2 (compare to heatmap patterns in Figure 2b). The luminal A and B subdivisions of the intrinsic subtypes correspond respectively to high and low proliferation score within type 3. However, we do not see natural clustering in the distribution of proliferation score. The slight bimodality in figure 3a is the result of pooling the subtypes and is not observed within type 3 alone. This lack of sharp distinction in the proliferation level has been noted previously [15,39] and may have contributed to inconsistent proposals for subdividing type-3 tumors [10,12–14]. Unimodal module scores may still be useful for subdividing tumors, using cutoffs determined by clinical utility. However, it is useful to distinguish tumor subdivisions according to intrinsic multimodality, which may reflect a discrete or switch-like underlying biological process. We use the term “subtypes” for the latter kind of subdivisions.

15 The relationship between module scores and some gene mutations can also be examined. Almost all BRCA1-mutated tumors are confined to type 1 (dataset NKI in Figure 3b,c), confirming the hypothesis that type-1 (“basal-like”) tumors are phenocopies of BRCA1-mutated tumors [18]. This is also supported by the strong overexpression of LMO4, a suppressor of BRCA1 function [40], in type-1 tumors (shown anticorrelated with ESR1 in Figure 2b). p53 mutation is examined in dataset UPP (Figure 3b,c). Although the mutations appear less frequently in type 3, their occurrence is more directly associated with the proliferation score than to the subtypes. Both p53 signatures proposed by Miller *et al.* [4] and Troester *et al.* [41] contain genes from the estrogen and proliferation modules.

Prognostic value of module scores

16 The attractiveness of the 70-gene signature [1] for clinical applications comes from the ability to identify a group with good survival rate (>90% in 5 years) that is acceptable for sparing the patients from aggressive chemotherapy [19]. In this section, we investigated whether classification based on the easily interpretable module scores could achieve such clinical relevance. Although ER-negative and ERBB2-positive tumors are known to have poor prognosis [29,30], the relatively better prognosis of ER-positive tumors is not good enough [15] and a subset of ER-positive tumors with much lower risk can be identified by expression-based tumor grade. Here, we adapted this knowledge to the three subtypes and proliferation score. Type 3 tumors were subdivided into low- and high-proliferative groups according to the median value of the proliferation score within the subtype. This cutoff classifies about one third of the total samples as low-risk, similar to the proportions made by other prognostic signatures [1,3]. The cutoff is approximately between the upper range for normal breast samples and the lower range for type 1 and 2 tumors (Figure 3c). Few tumors (less than 5%) in type 1 and 2 have proliferation score lower than this cutoff. We denote the low- and high-proliferative type-3 tumors by 3L and 3H, respectively.

17 Figure 4a,c show Kaplan-Meier analysis of more than 2000 patients in the collection, grouped into type 1, 2, 3H and 3L (see **Supplementary Result 5** for all combinations between subtypes and proliferation levels). Group 3L has much better overall survival than the rest, 94% in 5 years. The survival curves of type 1 and 2 drop faster than that group 3H in the first five years or so, although later the curves are rejoined. The survival differences between group 1, 2 and 3H do not affect chemotherapy decision (their risks are still too high), and therefore they are pooled into the ‘poor’ prognosis group, in contrast to the ‘good’ 3L group.

18 The consistency of the prognostic value across datasets is demonstrated by the forest plots in Figure 4b,d, where the analysis results of individual datasets are concisely summarized by the 5-year survival estimates and hazard ratios between the ’good’ and ’poor’ groups. Although the 5-year survival estimates vary between datasets, the relative survival and the hazard ratios are more consistent. Thus, the heterogeneous patient selection processes that led to the cohort-specific baseline risks (Figure 1b) may have biased the risk of the ’good’ or ’poor’ groups equally.

19 The interactions between the module-based risk groups and conventional prognostic variables are illustrated by paired cross-classifications in Figure 4e and tested in multivariable Cox regression analysis in Figure 4f. The module-based classification adds a strong prognostic effect over all other factors. Confirming previous studies [15,39], the effect of histological grade is much reduced, and can be explained by the refinement of intermediate grade (G2) into two groups with very different survival (Figure 4e). ER status remains significant in multivariate analysis. However, this is due mostly to the refinement of the ‘poor’ group (Figure 4e), which does not affect chemotherapy decisions. More substantial refinements are shown by lymph-node status and tumor size. Combining these two factors and the module-based classification allocates a larger proportion of patients (nearly 50%) into the low-risk category (Figure 4g).

20 The observation that type-2 tumors tend to have intermediate estrogen score raises a question whether it is meaningful to subdivide this subtype according to ER status. In this dataset collection, ERBB2+/ER- and ERBB2+/ER+ groups were not prognostically different (**Supplementary Result 6**). Additionally, others have noted that ERBB2+/ER+ tumors did not respond to tamoxifen therapy [42], unlike ERBB2-/ER+ tumors.

Dissecting prognostic signatures

21 Fan *et al.*[26] noted the similarity of the performance and patient classifications of several prognostic signatures. Here, we performed more detailed and extensive analysis to understand how disparate gene lists may give rise to potentially equivalent prognostic signatures. We assessed several important published signatures (Table 2), and two new signatures (CCYC and ZCOX). Four of the signatures (p53-32, CSR, GGI-128 and CCYC) were not identified through associations with patient outcome and were initially not meant to be prognostic.

22 The prognostic power of each gene in the union of 17198 genes (Table 1) is characterized by calculating the meta-analytical *Z*-score of gene-by-gene Cox regression (analogous to the procedure in Figure 2a), referred to as "*Z*-survival" scores. The coexpression module analysis described above produced five *Z*-scores for each gene, corresponding to the coexpression with the respective prototype. The scores are referred to as as *Z*-ESR1, *Z*-ERBB2, and so on. Scatter plots relating prognostic power and coexpression with the prototypes are shown in Figure 5a. Many genes are significantly associated with survival even under a stringent Bonferroni multiple testing correction. More importantly, the association with survival is strongly correlated with the association with the AURKA proliferation prototype. Of the 524 genes with significant Z-survival, 340 (65%) are most strongly coexpressed with AURKA, 75 (14%) with ESR1, 2 (0.6%) with ERBB2, 14 (2.7%) with PLAU, 8 (1.5%) with STAT1 and 84 (16%) with none of the prototypes.

23 The scatter plots can be used to characterize a signature by highlighting its constituent genes (Figure 5b; showing only *Z*-AURKA due to limited space). As shown by the *Z*-survival values, many of the genes used in various signatures are confirmed to be individually prognostic in the whole dataset collection. However, some genes have low *Z*-survival, possibly because they are artifacts (due to sampling error or biases in the original single-cohort study), or because the genes are not meant to be prognostic (such as in p53-32, CSR and CCYC). In accordance to the trend of the whole genome, signature genes that are strongly prognostic tend to be coexpressed with the proliferation prototype AURKA. Interestingly, the wound-healing signature CSR, which was claimed to have been cleared of cell-cycle genes[43], still contains such genes (for instance MYBL2, CENPN and MCM3).

24 The performance of the signatures was tested on the entire data collection. Not all genes of a signature can be mapped to a given platform. However, all signatures suffer from this problem and their relative performance can still be compared, as well as their robustness in cross-platform applications. A simple and uniform algorithm is used to compute a prediction score for all signatures (**Methods**). The original gene-specific parameters or direction of effects were not readjusted to maximize the performance, and thus the procedure validates not only the gene lists but also the roles of individual genes. The prediction score cutoffs were chosen such that all signatures assigned 33% of each cohort to the low-risk group, effectively fixing the cost of treatment to compare the actual risks on an equal footing.

25 Most signatures are robust to cross-platform applications and show similar performance (Figure 6a). The variations in performance between datasets is larger than those between signatures within a dataset, suggesting stronger influence from cohort characteristics than from signature differences. ZCOX, derived meta-analytically from this collection, is not substantially better than others. Signatures p53-32 and EMC-76 show slightly worse performance, which can be explained by the higher proportion of genes that are not individually prognostic (Figure 5b). The superior performance of EMC-76 in the EMC dataset (where it was derived) is not reproduced in other datasets, as indicated by the total performance.

26 To investigate the role of proliferation genes, we split each signatures into two "partial signatures": one with only proliferation genes (operationally defined by $|Z\text{-AURKA}| > 10$) and the other with the complementary non-proliferation genes (Figure 6b and c). When only proliferation genes are used, the total performance is not degraded and even improved for some signatures (p53-32 and EMC-76). On the other hand, the non-proliferation partial signatures typically show degraded performance. Interestingly, the non-proliferation parts of NKI-70 and EMC-76 show superior performance over other signatures in the dataset NKI and EMC, respectively. In contrast, the proliferation parts these signatures show reduced performance in their own datasets, but their total performance is not (and even improved in the case of EMC-76). These examples show that proposed signatures may contain genes that are unnecessary or even detrimental to their performance.

27 The average pairwise concordance of the patient assignments into risk groups is 79% (85% if p53-32 and EMC-76 were excluded), Among proliferation-only partial signatures, the concordance is 87%. Combining the signatures could not improve the performance (**Supplementary Result 9**), as expected from the high concordance in their classifications. These results extend the findings of Fan *et al.*[26] to a much larger sample size and for several additional signatures. More importantly, the equivalence of various signatures can be understood by looking at the risk classifications on the plots of proliferation score versus the subtypes (Figure 6d). Most signatures identify the low-proliferative subset of type-3 (ER+/ERBB2-) tumors as low risk. This is similar to the action of the module-based predictor and of the intrinsic subtype (Figure 3 and 4).

DISCUSSION

28 The recent Microarray Quality Control (MAQC) project[44], established the technical consistency of expression profiling technologies. This finding is complemented by our results, which revealed consistent biological and clinical relationships (such as coexpression, clustering patterns, survival associations and signature performance) across independent breast cancer cohorts, despite the diverse study designs and methodologies. Meta-analytical approaches are valuable not only for increasing the sample size (thus decreasing sampling error artifacts) but also in reducing the contribution of strong but non-reproducible associations caused by platform- or cohort-specific biases. Unlike an earlier meta-analysis of diverse cancer expression data[45], our study focused on breast cancer, but investigated in greater detail several important aspects of the disease (such as subtyping, tumor proliferation and prognosis) and their interconnections.

29 We have found the concept of coexpression modules to be a versatile tool for unifying disparate results. Although coexpression does not imply direct physical interactions, the highly correlated genes in a module can be considered surrogate markers of one another and of the same underlying transcriptional process. Thus, coexpression is more appropriate for understanding the equivalence of signatures than functional annotations of the genes. It is noteworthy that the grouping of genes in the 21-gene RT-PCR signature of Paik *et al.*[22] (signature ONC-16) largely agrees with our modules, suggesting that similar prognostic systems can be designed using many possibilities of alter-

native genes. Coexpression modules can also be used to dissect signatures, revealing the parts that are essential. The strong correlation of expression within a module allows summarizing the module's overall expression by simple averaging. These module scores concisely characterize a tumor by a handful of quantitative measures with straightforward interpretation. Lastly, the concordance of patient outcome prediction of various signatures can be elegantly interpreted in terms of a few module scores (Figure 6d).

30 Although it may be argued that microarray signatures are merely alternative ways to monitor well-known processes such as proliferation or estrogen receptor signaling, their results are not perfectly concordant to conventional variables. For example, although the proliferation module score and histological grade both aim to measure cell proliferation, the former is more informative. We also observe that type-2 (ERBB2+) tumors have intermediate estrogen module score, which is not obvious from the traditional ER and ERBB2 marker status combination. Thus, using many genes from a coexpression module may provide a more accurate quantitation of a whole transcriptional process than using a single-gene markers or histopathological variable.

31 Blamey[28] distinguished prognostic factors into those related to the extent of tumor progression (such as lymph-node status and tumor size) and those related to the intrinsic aggressiveness (such as mitotic rate and growth receptors). Histological grade was found to already contain the prognostic information of other intrinsic factors, and only factors of tumor progression (lymph-node status and tumor size) had additional prognostic values. Our results recapitulate these observations. The proliferation score already contains the poor prognosis information attributable to various sources: ERBB2 amplification, type-1 phenotype (with or without BRCA1 mutation), p53 mutation, or yet unknown factors specifically affecting half of type-3 tumors. Thus, proliferation can be considered as the downstream effector process of other factors of intrinsic aggressiveness; while lymph-node status and tumor size influence the outcome through their own independent paths.

32 Although the downstream variables cover most of the prognostic information, knowledge about the upstream processes is important for selecting and developing treatments. Genes in the proliferation module are already targeted by several chemotherapeutic agents[46], but less harmful drugs are more desirable. Type-3 tumors are treatable to some extent by hormone therapy[47] (targeting ESR1 signaling), and type-2 tumors by trastuzumab[48] (targeting ERBB2). However, drugs specific to type-1 tumor are not yet established. Furthermore, unresponsiveness to existing drugs warrants further search for alternative targets, possibly compensatory genes in the same pathway.

33 In summary, this study unified various results of previous gene-expression studies in breast cancer. Methodologically, we provided a new framework, also applicable to other diseases, for utilizing heterogeneous microarray datasets to uncover consistent biological relationships and to consolidate proposed signatures. Biologically, we identified comprehensive gene lists that improve our understanding of the breast cancer transcriptome, as well as providing new candidates for biomarkers and therapeutic targets. Clinically, we revealed connections between traditional prognostic factors, expression-based subtyping and prognostic signatures that should increase our confidence in practical applications of gene-expression signatures.

METHODS

34 **Probe annotation and gene matching** Hybridization probes were mapped to Entrez GeneID[34] through sequence alignment against RefSeq mRNA in the well-curated (NM) subset, similar to the approach by Shi *et al.*[44]. Affymetrix and Agilent probe sequences were obtained from the manufacturer. For other platforms (cDNA or Agilent with unavailable probe sequences), the mapping was done by retrieving the GenBank sequences corresponding to the probes. Alignment is done using BLAT[49]. Alternate transcripts (different RefSeq with the same GeneID) were considered to be the same "gene". RefSeq version 21 (2007.01.21) and Entrez database version 2007.01.21 were used. When multiple probes were mapped to the same GeneID, the one with the highest variance in a particular dataset was selected to represent the GeneID.

35 **Preprocessing of expression values** We used the normalized expression measures ($\log_2$ of intensity in single-channel platforms or $\log_2$ratio in dual-channel platforms) published by the original studies. Because our meta-analytical approach combines summary statistics instead of expression data, normalization across datasets was not necessary. Within each dataset, missing values were imputed using the mean of present values for the same gene.

36 **Identifying coexpression modules** The expression levels of the prototype genes on the $\log_2$ scale were used as explanatory variables in multiple regression with Gaussian error model, using the following equation (gene symbols stand for their log expression and coefficients are omitted for clarity):

$$Y_i = \text{ESR1} + \text{ERBB2} + \text{AURKA} + \text{PLAU} + \text{STAT1}$$

where the response variable Y_i is the expression of gene i. This model is fitted separately for each gene i in the array. The association between gene i and prototype j (in the presence of or conditional on all other prototypes) is tested using the t-statistic for each coefficient. Because the t-statistics for different datasets have different degrees of freedom, we put them all on the same scale by transforming to the corresponding cumulative probabilities and then to Z-scores using the inverse standard normal cumulative distribution function.

37 The linear model above was fitted separately to each gene in each dataset, and the Z-scores were combined meta-analytically over K studies using the "inverse normal method"[36]:

$$\overline{Z}_{ij} = \sum_{k=1}^{K} Z_{ijk} / \sqrt{K_i}.$$

where i, j and k are indices for genes, tested regression terms and dataset, respectively. K_i is the number of datasets where the gene i is present (that is, any platform missing the gene is ignored). Due to the large sample size, the conservative Bonferroni-corrected p-value of less than 0.05 is achieved for many coefficients (for a test of any departure from linear independence, $\beta_{ij} = 0$). To select genes that are most strongly associated with the prototypes, we use a more stringent criterion of $|Z| \geq 16$ instead, which is well above $|Z| \approx 5$, that corresponds to a corrected p-value of 0.05.

38 Some genes have more than one prototype satisfying the criteria above. To make the genes in a module more specific, we introduced a "uniqueness" criterion:

$$U_{ij} = Z_{ij}^2 / \sum_q Z_{iq}^2$$

where q is the index over all prototypes. Thus a gene i is a part of module j when $Z_{ij} \geq 16$ and $U_{ij} > 0.5$.

39 **Module scores** For a specific dataset, the module score is computed for each sample as:

$$\text{module score} = \sum_i w_i x_i / \sum_i |w_i|$$

where x_i is the expression of a gene in the module that is present in the dataset's platform. w_i is either +1 or -1 depending on the sign of the Z-score of the association with the prototypes. The denominator is used so that the range of the module score roughly corresponds to the typical range of log expression values. Mean centering is performed on each gene prior to module score calculation, to make the score roughly centered around zero. However, these scaling and centering are only for convenience in displaying the scores. Our analyses do not assume commensurability of module scores between datasets.

40 **Clustering and multimodality tests** Gaussian mixture models[38] with equal variance for all clusters were fitted. In the case of two-dimensional data, diagonal covariance matrices were used, allowing for dimension-specific variances. For testing multimodality, we used the likelihood ratio test statistics between the fitted model for the tested number of components, k, versus the alternative model with $k-1$ components. The null distribution was generated by parametric bootstrapping from the fitted alternative model.

41 Each tumor was automatically classified as type 1, 2 or 3 using the posterior probability of membership in the clusters. Although there were a few ambiguous cases with low probabilities shared among clusters (such as points in between clusters in Figure 3b), for simplicity all tumors were strictly categorized into either one type using the criterion of maximum probability.

42 **Survival analysis** Detailed treatment of survival endpoints and time units can be found in **Supplementary Result 1.1**. Because not all studies reported the complete data on all endpoints, we only showed results for one or combined endpoints. Figure 5 and 6 were based on metastasis-free survival (if available) or overall survival. Complete results for each type of endpoints can be found in **Supplementary Result 1, 5-8**.

43 Survival curves and 5-year survivals in forest plots were based on Kaplan-Meier estimates, with the Greenwood method for computing the 95% confidence intervals[50]. Hazard ratios between two groups were calculated using Cox regression. Stratified Cox regression was used to compute total hazard ratios in forest plots and multivariate analysis in Figure 4f, using the dataset as the stratum indicator, thus allowing for different baseline hazard functions between cohorts.

44 In multivariate analysis, tumor size is categorized as small or large using the 2-cm cutoff because some datasets provide only the category instead of the continuous length measurements. Histological grade is dichotomized into grade 1 versus 2+3 because it is the most clinically relevant for identifying the low-risk subset (contrasting 1+2 versus 3 will give, respectively, intermediate and high risk groups not appropriate for chemotherapy decision).

45 Cox regression was used to compute gene-by-gene Z-survival scores, treating the log expression measures as continuous explanatory variables. The Z-score was based on the signed square-root of the deviance (two times the log likelihood ratio). These Z-scores were combined across datasets using the same meta-analytical formula as used for coexpression module analysis, described above.

46 **Cell-cycle periodicity** We used datasets from various cell cycle experiments reported in Whitfield *et al.*[46]. The periodicity of expression was scored using linear models with a pair of cosine and sine functions as the explanatory variable, with the frequency corresponding to the estimated cell-cycle periodicities in the respective experiments. The F-ratio test statistics were converted to p-values with the appropriate degrees of freedom, and then converted to Z-scores using the inverse standard normal distribution function. The Z-scores from different experiments were combined meta-analytically as described above.

47 **Cross-platform applications of signatures** Only genes in the signatures that can be mapped to GeneID were used. A prediction score was computed for each signature, using a linear combination similar to the formula for module score above. Gene-specific weights (coefficients, correlations, or other measures) from the original studies were used (or, if not available, +1 or -1 depending on the original up- or down-regulation of each gene). Genes absent in a platform were ignored. The scales of the prediction scores were not comparable between signatures. For each cohort and signature combination, a cutoff was chosen based on the percentile of the scores (33% for the results in Figure 6).

48 The signature EMC-76 had different gene lists for ER- and ER+ tumors, and were applied accordingly (except in proliferation-only partial signatures that were applied to the all patients regardless of the ER status because the ER- signature does not contain proliferation genes).

49 The signature ZCOX was identified from the data collection by selecting the top ranking genes according to survival association. The criterion was an absolute value of Z-score greater than 6 (shown in Figure 5b). To avoid bias in the performance assessment, we used a "leave-one-dataset-out" cross-validation: the test dataset (whose performance was reported) was excluded from the calculation of the meta-analytical Z-scores for selecting the tested signature.

Acknowledgments

This work was supported by the European Commission Framework Programme VI (FP6-LSHC-CT-2004-503426) (P.W., F.S.); by the National Center of Competence in Research Molecular Oncology of the Swiss National Science Foundation (M.D., T.S., S.K.); by the MEDIC Foundation (P.F., C.S.); and by the Belgian National Foundation for Cancer Research, FNRS (B.H-K, C.D., C.S.). We thank Sandra Flores-Urushima for initial work in dataset collection. We dedicate this work to Hans-Martin Schultze, who started the preliminary analysis but succumbed prematurely to metastasis of melanoma.

Competing Interests Statement

None declared.

Authors' Contribution

All authors contributed to the preparation of the manuscript. P.W., M.D. and C.S. designed the overall study. S.K., T.S. and P.W. compiled and curated the datasets. D.R.G., P.W., and M.D. designed the statistical approaches. P.W., M. D., S.P., F.S., B.H-K and C.D. performed the computational analysis. P.F. and P.W. developed the prototype genes and biological interpretation. M.P. and C.S. provided expertise in clinical breast oncology.

1. van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
2. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
3. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).
4. Miller, L. D. *et al.* 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 (2005).
5. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* **7**, R953–R964 (2005).
6. Calza, S. *et al.* Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res.* **8**, R34 (2006).
7. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
8. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).
9. Korkola, J. E. *et al.* Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis. *Cancer Res.* **63**, 7167–7175 (2003).
10. Hu, Z. *et al.* The molecular protraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96 (2006).
11. Naderi, A. *et al.* A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* **epub**, 28 August (2006).
12. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874 (2001).
13. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418–8423 (2003).
14. Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA* **100**, 10393–10398 (2003).
15. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).

16. Ma, X.-J. *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**, 607–616 (2004).
17. Farmer, P. *et al.* Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* **24**, 4660–4671 (2005).
18. Richardson, A. L. *et al.* X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9**, 121–132 (2006).
19. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* **98**, 1183–1192 (2006).
20. Foekens, J. A. *et al.* Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J. Clin. Oncol.* **24**, 1665–1671 (2006).
21. Espinosa, E. *et al.* Breast cancer prognosis determined by gene expression profiling: a quantitative reverse transcriptase polymerase chain reaction study. *J. Clin. Oncol.* **23**, 7278–7285 (2005).
22. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
23. Chang, H. Y. *et al.* Robustness, scalability and integration of wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. USA* **102**, 3738–3743 (2005).
24. Jenssen, T.-K. & Hovig, E. Gene-expression profiling in breast cancer. *Lancet* **365**, 634–635 (2005).
25. Massagué, J. Sorting out breast-cancer signature. *New Engl. J. Med.* **356**, 294–297 (2007).
26. Fan, C. *et al.* Concordance among gene-expression-based predictors for breast cancer. *New Engl. J. Med.* **355**, 560–569 (2006).
27. Whitfield, M. L., George, L. K., Grant, G. D. & Perou, C. M. Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).
28. Blamey, R. W. The design and clinical use of the Nottingham Prognostic Index in breast cancer. *The Breast* **5**, 156–157 (1996).
29. Fitzgibbons, P. L. *et al.* Prognostic factors in breast cancer: College of American Pathologists Consensus Statement 1999. *Arch Pathol Lab Med* **124**, 966–978 (2000).
30. Esteva, F. J. & Hortobagyi, G. N. Prognostic molecular markers in early breast cancer. *Breast Cancer Res.* **6**, 109–118 (2004).
31. Ioannidis, J. P. A. Microarrays and molecular research: noise discovery? *Lancet* **365**, 454–455 (2005).
32. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178 (2004).
33. Teschendorff, A. E. *et al.* A consensus prognostic gene expression classifier for er positive breast cancer. *Genome Biology* **7**, R101 (2006).
34. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acid Res.* **33**, D53–D58 (2005).
35. Altman, D. G. & Deeks, J. J. Meta-analysis, Simpson's Paradox and the number needed to treat. *BMC Med. Res. Methodol.* **2**, 3 (2002).
36. Hedges, L. V. & Olkin, I. *Statistical methods for meta-analysis*. Academic Press, London (1985).
37. Kapp, A. V. *et al.* Discovery and validation of breast cancer subtypes. *BMC Genomics* **7**, 231 (2006).
38. McLachlan, G. & Peel, D. *Finite mixture models*. Wiley, New York (2000).
39. Ivshina, A. V. *et al.* Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* **66**, 10292–10300 (2006).
40. Sum, E. Y. M. *et al.* The LIM domain protein LMO4 interacts with the cofactor CtIP and the tumor suppresor BRCA1 and inhibits BRCA1 activity. *J. Biol. Chem.* **277**, 7849–7856 (2002).
41. Troester, M. A. *et al.* Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* **6**, 276 (2006).
42. Dowsett, M. *et al.* Benefit from adjuvant tamoxifen therapy in primary breast cancer patients according oestrogen receptor, progesterone receptor, EGF receptor and HER2 status. *Ann. Oncol.* **17**, 818–826 (2006).
43. Chang, H. Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biology* **2**, 206–214 (2004).
44. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151–1161 (2006).
45. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional prolifes of neoplastic transformation and progression. *Proc. Acad. Natl. Sci. U. S. A.* **101**, 9309–9314 (2004).
46. Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
47. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* **351**, 1451–1467 (1998).
48. Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med* **353**, 1673–1684 (2005). Clinical Trial.
49. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
50. Therneau, T. M. A package for survival analysis in S. Technical report, Mayo Foundation (1999).

Table 1 Publicly available gene-expression data from breast cancer studies

Dataset symbol	No. of arrays	Institution	Reference	Platform	Data source	No. of GeneIDs
Genomic platforms						
NKI	337	Nederlands Kanker Instituut	van't Veer *et al.* [1], van de Vijver *et al.* [2]	Agilent	author's website	13120
EMC	286	Erasmus Medical Center	Wang *et al.* [3]	Aff. U133A	GEO:GSE2034	11837
UPP	249	Karolinksa Institute (Uppsala)	Miller *et al.* [4], Calza *et al.* [6]	Aff. U133A,B	GEO:GSE4922	15684
STOCK	159	Karolinska Institute (Stockholm)	Pawitan *et al.* [5], Calza *et al.* [6]	Aff. U133A,B	GEO:GSE1456	15684
DUKE	171	Duke University	Huang *et al.* [7,8]	Aff. U95Av2	author's website	8149
UCSF	161+8	UC San Francisco	Korkola *et al.* [9]	cDNA	author's website	6178
UNC	143+10	University of Carolina	Hu *et al.* [10]	Agilent HuA1	author's website	13784
NCH	135	Nottingham City Hospital	Naderi *et al.* [11]	Agilent HuA1	AE:E-UCON-1	13784
STNO	115+7	Stanford Univ./Norwegian Radium Hosp.	Sorlie *et al.* [12,13]	cDNA	author's website	5614
JRH1	99	John Radcliffe Hospital	Sotiriou *et al.* [14]	cDNA	journal's website	4112
JRH2	61	John Radcliffe Hospital	Sotiriou *et al.* [15]	Aff. U133A	GEO:GSE2990	11837
MGH	60	Massachusetts General Hospital	Ma *et al.* [16]	Agilent	GEO:GSE1379	11421
expO	239	International Genomic Consortium	http://www.intgen.org	Aff. U133v2	GEO:GSE2109	16634
TGIF1	49	EORTC trial 10994	Farmer *et al.* [17]	Aff. U133A	GEO:GSE1561	11837
BWH	40+7	Brigham and Women's Hospital	Richardson *et al.* [18]	Aff. U133v2	GEO:GSE3744	16634
Small diagnostic platforms						
TRANSBIG	253	TRANSBIG Consortium	Buyse *et al.* [19]	Agilent	AE:E-TABM-77	1052
EMC2	180	Erasmus Medical Center	Foekens *et al.* [20]	Aff. (custom)	GSE3453	86
HPAZ	96	Hospital La Paz, Madrid	Espinosa *et al.* [21]	RT-PCR	paper's appendix	61
Total	2865	= 2833 carcinomas		No. of the union of all GeneIDs:		17198
		+ 32 non-malignant breast tissues		No. of GeneIDs common to genomic platforms:		1963

- Abbreviations: No. = number, GEO: = Gene Expression Omnibus accession, AE: = ArrayExpress accession, Aff. = Affymetrix
- Dataset UNC, STNO, UCSF and BWH include a small number of normal breast or fibroadenoma tissue samples.

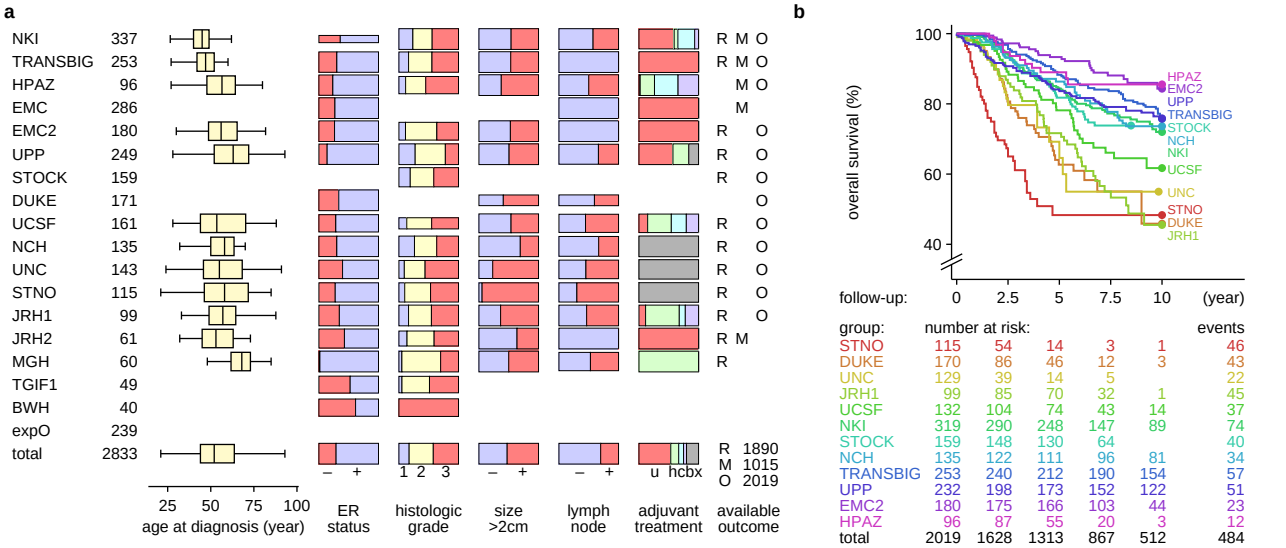

Figure 1 Summaries of patient characteristics for each dataset. a) Distribution of important clinical variables, shown by boxplots for continuous variables and by colored bars showing proportions for categorical variables. The heights of the bars correspond to proportion of non-missing values. For adjuvant treatment, u = untreated, h = hormone therapy, c = chemotherapy, b = both, x = unspecified. For 'available outcome', the reported endpoints are: R = any relapse (unspecified), M = distant metastasis and O = overall survival. The numbers of patients with available follow-up data for each type of outcome are shown on the "total" line. b) Kaplan-Meier curves showing the heterogeneity of survival when the patients were stratified by the cohorts.

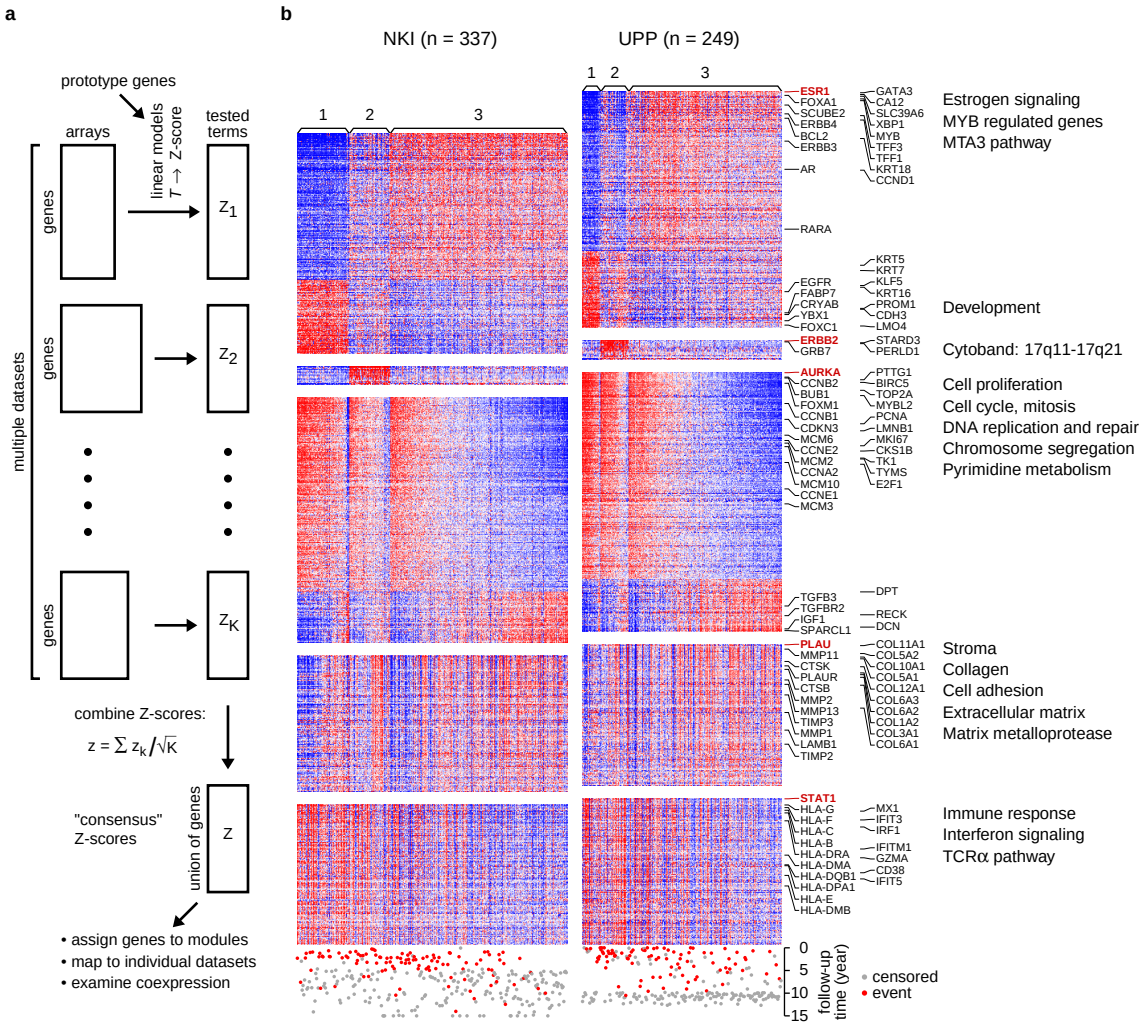

Figure 2 Coexpression module analysis. *a*) Outline of the meta-analytical procedure for identifying genes consistently coexpressed with the prototype genes in multiple datasets. *b*) Heatmaps demonstrating the coexpression modules in two example datasets. The rows are genes, grouped according to the modules, whose gene symbol of the prototypes are shown in red. Within each group, the genes are ranked according to their *Z*-scores of association with their respective prototype. The columns are tumors, grouped according to subtype 1, 2, or 3 (see the section "Module scores for tumor subtyping"). Within each subtype the tumors are sorted according to the average expression of the proliferation (AURKA) module. Genes within a module show strong correlated or anticorrelated expression with the prototype gene. The names of several well-known genes and the annotation from various sources are shown to the right of each module, illustrating that the coexpression patterns correspond to coherent biological processes. Underneath the heatmaps are patient's survival data

Table 2 Prognostic signatures

Signature symbol	Reference	Associated variables in gene selection procedure	Number of genes	
			original probes	mapped to geneID
ONC-16	Paik *et al.* [22]	biological knowledge; refined by patient outcome	16	16
NKI-70	van't Veer *et al.* [1]	patient outcome	70	52
EMC-76	Wang *et al.* [3]	patient outcome, stratified by ER-status	60+16	48+12
NCH-70	Naderi *et al.* [11]	patient outcome	70	69
CON-52	Teschendorff *et al.* [33]	patient outcome, consensus	52	50
p53-32	Miller *et al.* [4]	p53 mutation	32	19
CSR	Chang *et al.* [43]	fibroblast core serum response	512	457
GGI-128	Sotiriou *et al.* [15]	histological grade	128	98
CCYC	this study, Whitfield *et al.* [46] datasets	periodic expression in cell cycle progression	NA	126
ZCOX	this study	patient outcome, meta-analysis of all datasets	NA	113

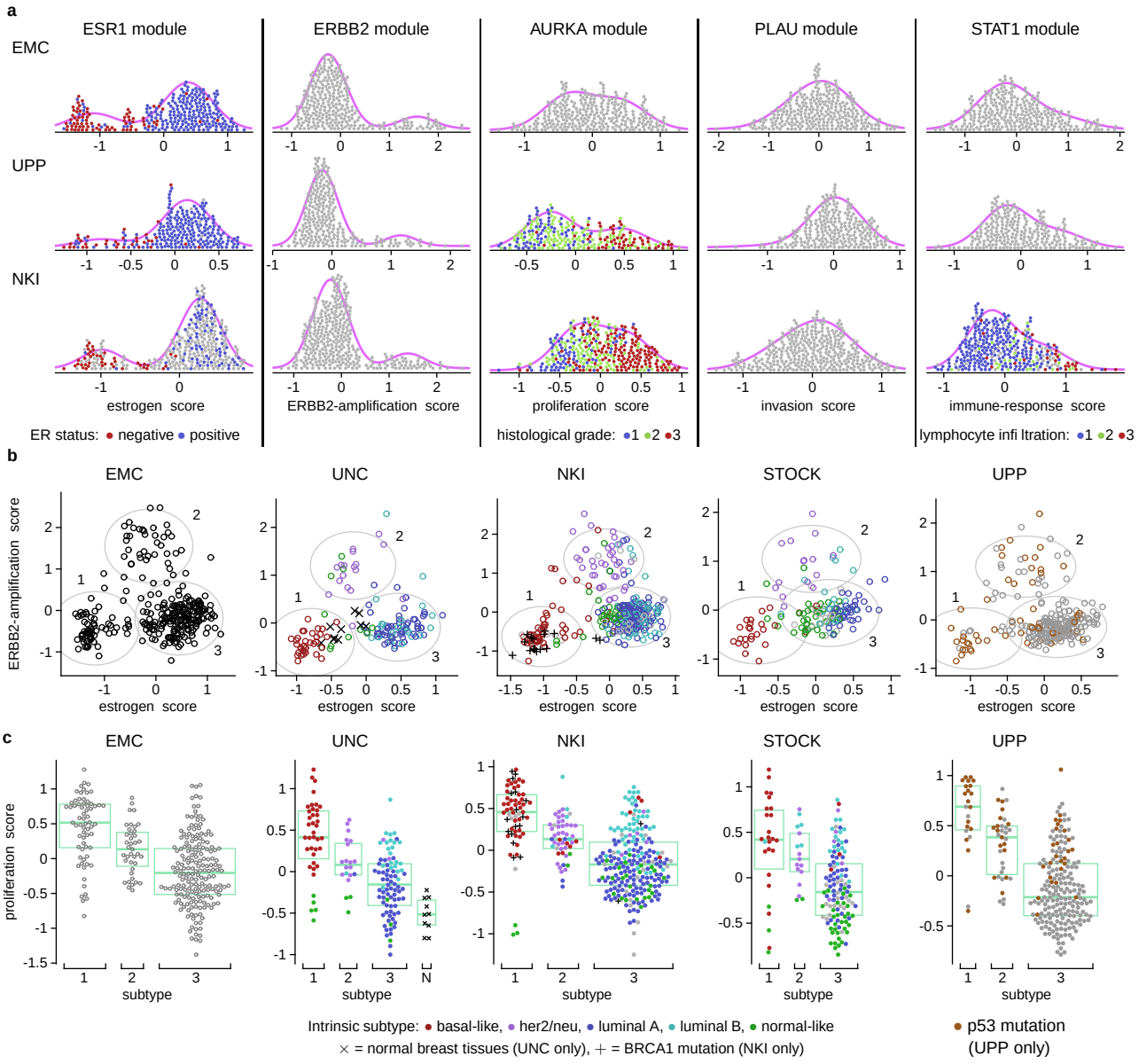

Figure 3 The distributions of module scores and some tumor characteristics. a) Dot histograms showing the distributions of the module scores (columns) for example datasets (rows). Colored dots correspond to the status of the relevant pathological variables; while gray dots correspond to missing information. The estrogen, proliferation and immune-response scores are associated with, respectively, ER protein status, histological grade and lymphocytic infiltration. The purple curves are fitted Gaussian mixture densities with two components. The estrogen and ERBB2-amplification scores show significant bimodality. b) Joint distribution between the estrogen and ERBB2-amplification scores in example datasets. Clusters are identified by Gaussian mixture models with three components. The ellipses correspond to the 95% cumulative probability around the cluster centers. The clusters are designated as tumor type 1, 2 and 3. Type-2 (ERBB2-amplified) tumors show intermediate estrogen scores. c) Dot histograms showing dependence of proliferation score on the subtypes. The median and quartiles for each group are shown by the box plot. Type 1 and 2 show high proliferation scores; while type 3 shows a wide range of proliferation scores. In panel *b* and *c*, the distributions of the intrinsic subtypes (colored dots), BRCA1 mutations and p53 mutations are shown in datasets where they are available.

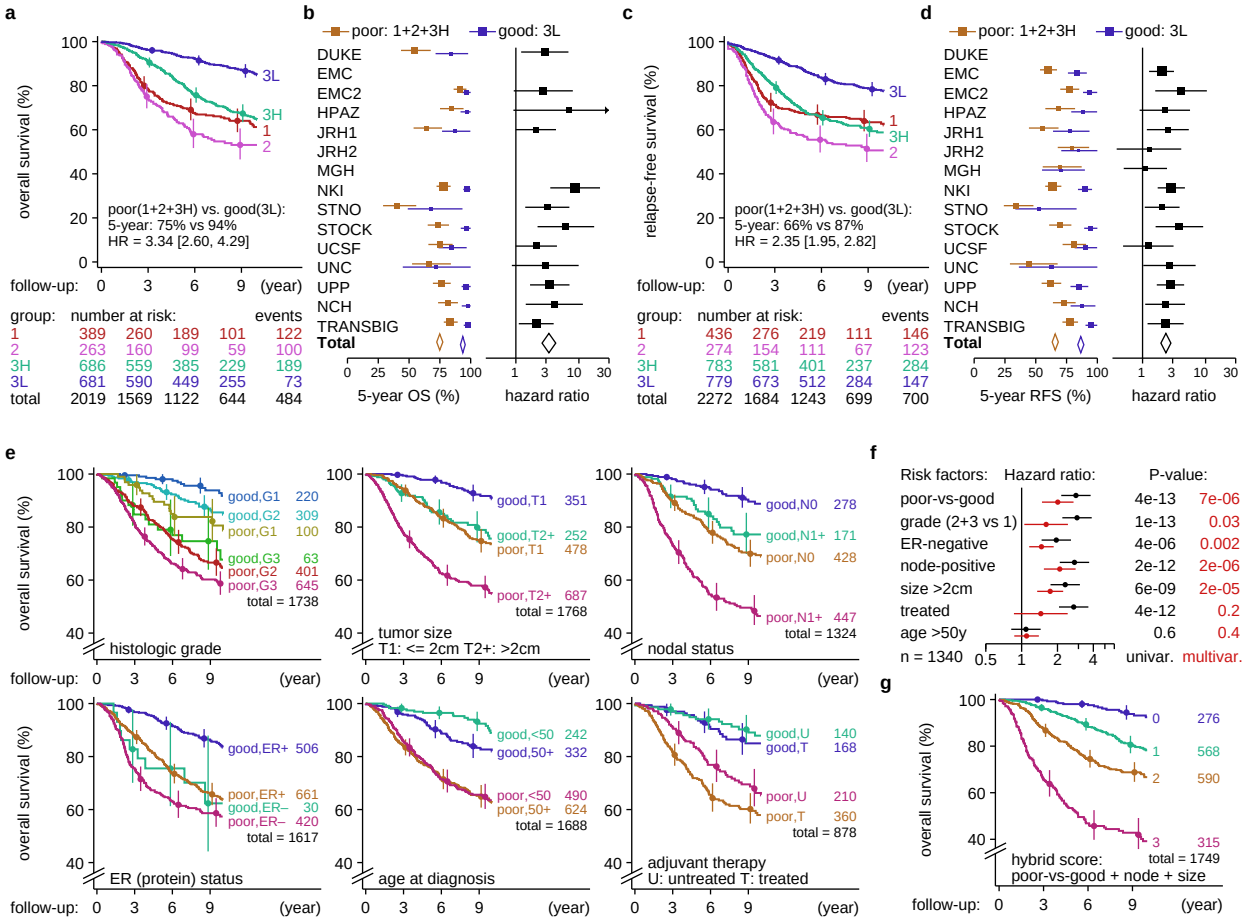

Figure 4 Survival analysis of groups based on module scores. *a*) Kaplan-Meier analysis of patient groups. Type 3 is split into 3L and 3H (low and high proliferation, respectively). Vertical bars on the curves are 95% confidence intervals for the Kaplan-Meier survival estimates. *b*) Forest plots showing the 5-year survival estimates and hazard ratios of individual datasets. The length of horizontal bars and the width of the diamonds of the "Total" correspond to 95% confidence intervals. Missing bars are unavailable data (see Figure 1). Panel *c*) and *d*) are analogous to *a*) and *b*) using metastasis-free survival data when available, or relapse-free survival otherwise. *e*) Survival analysis of "good" versus "poor" stratified by several conventional clinical variables. *f*) Multivariate analysis in patients where all the variables are available. *g*) Prognosis obtained by combining the module-based groups ("poor" vs "good"), lymph-node status and tumor size. Each variable is encoded as 1 or 0, and their sum is used as a prediction score, shown next to each curve. The groups with the score equals to zero or one seems to keep their favorable outcome, while their proportion is increased to (276+568)/1749, or 48%.

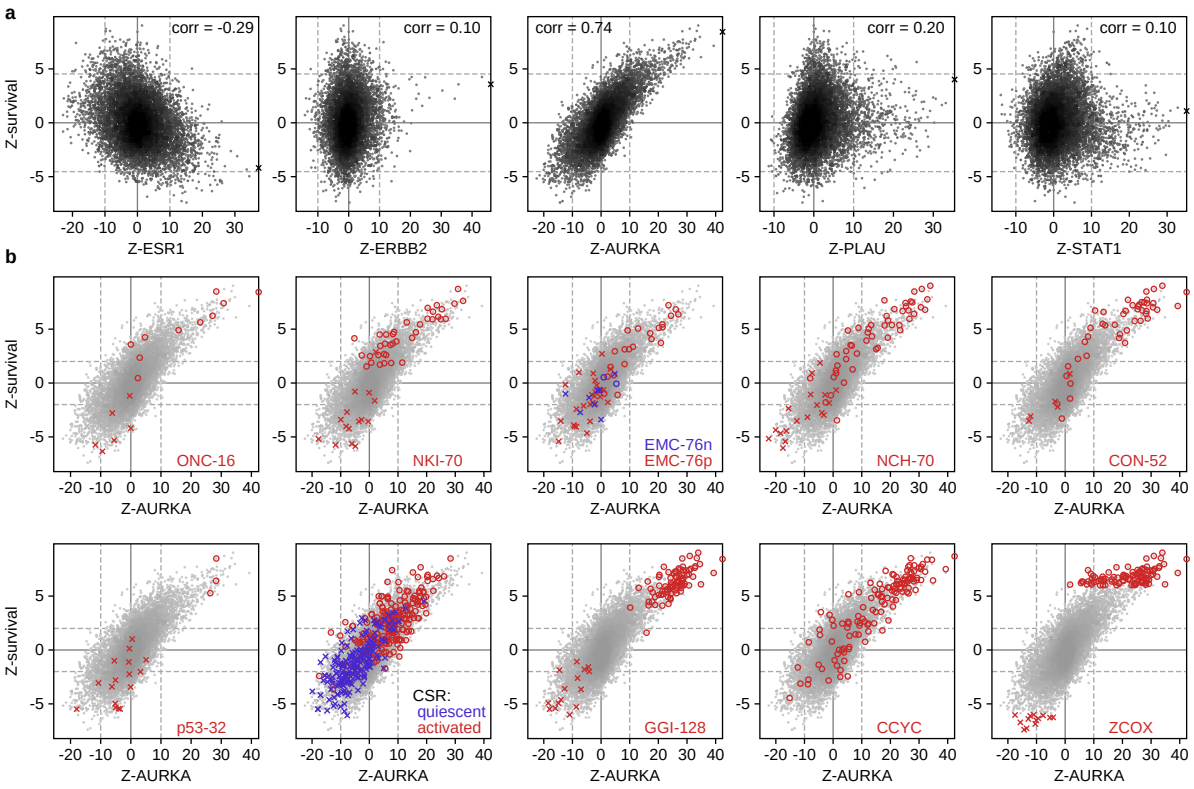

Figure 5 Prognostic power and modular association of individual genes. a) In each plot, the Z-score of survival association (vertical axis) is plotted against the Z-score of coexpression with a module prototype (horizontal axis) with their Pearson correlation coefficient indicated by "corr". The Z-survival of the prototype gene is shown by the cross on the right border. The band of (± 4.5) for Z-survival corresponds to Bonferroni correction for selecting from 17198 genes (at $p = 0.05$). The more stringent band of ± 10 is used for the Z-scores of coexpression. Strongly prognostic genes tend to be coexpressed with AURKA or, less frequently, ESR1. There is no general trend between Z-survival and Z-ERBB2, Z-PLAU or Z-STAT1. b) Genes from signatures in Table 2 are overlaid on the scatter plots of Z-survival versus Z-AURKA. Circles and crosses respectively indicate positive and negative effects in the original studies. For signature EMC-76, the ER-positive and ER-negative signatures are denoted as EMC-76p and EMC-76n, respectively. The band for Z-survival (± 2) is the single-test significance level at $p = 0.05$ (more appropriate if the genes are already known). Many, but not all, signature genes are confirmed to be prognostic in the dataset collection.

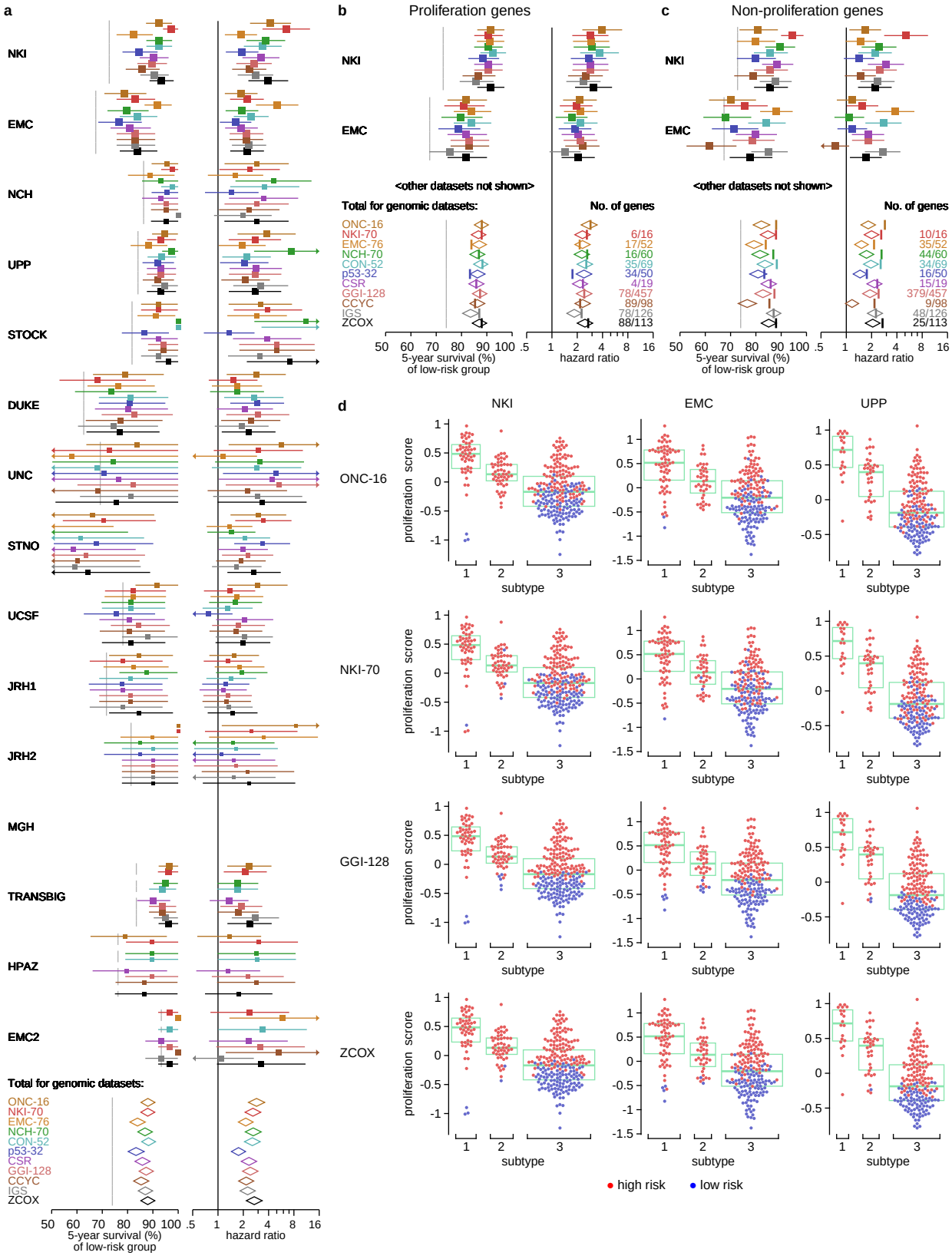

Figure 6 Signature comparison. *a*) The prognostic performance of the signatures are compared by the forest plots of hazard ratio and 5-year survival of the low-risk patients. 5-year survivals of the high-risk groups are not shown to avoid clutter. For reference, the 5-year survival of the whole (unstratified) cohort is shown by the vertical gray lines. Not all signatures can be mapped to TRANSBIG, HPAZ and EMC2, and therefore these datasets are excluded when calculating the totals. Most signatures show similar performance. *b*) Analogous analysis using partial signatures containing only proliferation genes. The total survival estimates and hazard ratios from panel *a* are replotted as vertical color bars for comparison. The performance of most signatures is not degraded, and even improved for p53-32 and EMC-76. *c*) As in panel *b*, but using the complementary subsets of non-proliferation genes. *d*) Patient classifications made by example signatures (rows) applied to example datasets (columns), showing that the different signatures are essentially detecting low-proliferative subset of type-3 (ER+/ERBB2-) tumors.