Citation: CPT: Pharmacometrics & Systems Pharmacology (2013) 2, e35; doi:10.1038/psp.2013.11
© 2013 ASCPT All rights reserved 2163-8306/12

www.nature.com/psp

ORIGINAL ARTICLE

# Metasignatures Identify Two Major Subtypes of Breast Cancer

Q Duan[1], Y Kou[1], NR Clark[1], S Gordonov[1,2] and A Ma'ayan[1]

**Genome-wide expression data from tumors and cell lines in breast cancer, together with drug response of cell lines, open prospects for integrative analyses that can lead to better personalized therapy. Drug responses and expression data collected from cell lines and tumors were used to generate tripartite networks connecting clusters of patients to cell lines and cell lines to drugs, to connect drugs to patients. Various approaches were applied to connect cell lines to tumor clusters: a standard method that uses a biomarker gene set, and new methods that compute metasignatures for transcription factors and histone modifications given upregulated genes in cell lines or tumors. The results from the metasignature analysis identify two major clusters of patients: one enriched for active histone marks and one for repressive marks. The tumors enriched for activation marks are correlated with poor prognosis. Overall, the analyses suggest new patient clustering, discover dysregulated pathways, and recommend individualized use of drugs to treat subsets of patients.**

*CPT: Pharmacometrics & Systems Pharmacology* (2013) **2,** e35; doi:10.1038/psp.2013.11; advance online publication 27 March 2013

Histopathologists have been attempting to define a standard taxonomy for breast cancer based on morphology since the 1960s. Outcome prediction and clinical decisions are still commonly made by histological profiling and various other clinical parameters.[1] However, the introduction of gene expression microarrays more than a decade ago promised that the use of quantitative assessment of all genes, rather than hisopathological subjective assessments or measurements of the expression level of single genes or proteins, would offer a more precise determination of the continuous tumor biology and outcome in breast cancer. Genome-wide profiling of breast cancer tumors and cell lines are continually used to identify dysregulated biological processes that characterize the etiology and progression of the disease at the molecular level.[2–4] cDNA microarrays, DNA and RNA sequencing, and proteomics and metabolomics technologies rapidly uncover the heterogeneity of this disease. Such heterogeneity, coupled with variable response to therapy, motivated subcategorization of patients with breast cancer into various subtypes,[5,6] and more recently subtypes of subtypes.[2] Of note, the molecular profiling of breast cancer tumors has identified upregulation of particular receptors within different patient subgroups such as the estrogen receptor, the progesterone receptor, and the epidermal growth factor receptor 2 (HER2),[7] which serve as personalized subtype-specific biomarkers and drug targets. Today, ~70% of patients receive adjuvant therapy, and decisions are increasingly made based on subtype classification. Adjuvant therapies are those given in addition to the main treatment, e.g., after surgical removal of the tumor. Adjuvant therapies can be radiotherapy or systemic therapies such as chemotherapy, or drug-targeted therapies. Clinical implementation of subcategorization of breast cancer tumors into the five subtypes; luminal A, luminal B, HER2+, normal-like, and basal are currently established. These subtypes are continually refined and are correlated with clinical outcome such as tumor metastatic propensity, response to various drugs, and patient-survival expectancy. Commercial and noncommercial biomarker diagnostic tools that rely on the combined expression of biomarker gene sets such as OncotypeDX (21 genes),[8] Veridex (76 genes),[9] MammaPrint (70 genes),[10] PAM50,[11] and others[12,13] demonstrated clinical applicability in identifying tumor subtypes, and are used routinely for selecting therapeutic regimens. Drug-induced gene expression signatures identified in panels of cancer cell lines can potentially enhance such approaches to elicit more precise and personalized therapies that would lead to better clinical outcomes.[14] Cell lines are more easily amenable than tumors from patients for measuring gene-expression changes induced by many drugs and their combinations at different concentrations and time points. In this direction, recent studies reported drug–response profiles for many breast cancer cell lines.[15,16] For instance, Heiser *et al.*[15] reported the responses of 55 cell lines to 77 drugs and combined the results from such screen to basal geneexpression profiles of these cell lines to explain the molecular mechanisms potentially responsible for the differential response of specific cells to specific drugs. The study by Heiser *et al.* showed that gene expression profiles can correlate with susceptibility of specific cell lines to particular classes of drugs. However, the next step, linking cell lines to patient tumors, is not trivial. Some of the challenges are dealing with the high dimensionality of the data including temporal stage of the tumors, time points after drug treatment of the cell lines, drug-concentration considerations, platform-dependent variability, and biological differences between tumors from patients and cell lines, including cell-media factors and the immortalized transformation from tumors into cell lines. Mindful of these issues, we developed new methods to (i) identify clusters

The first four authors contributed equally to this study.

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence: Avi Ma'ayan (avi.maayan@mssm.edu)

Received 29 October 2012; accepted 11 February 2013; advance online publication 27 March 2013. doi:10.1038/psp.2013.11

of patients across large cohorts; (ii) link clusters of patients to cell lines; and (iii) visualize networks that connect drugs to cell lines based on drug-sensitivity measures, and cell lines to patient tumors based on gene expression similarity or metasignature similarity. Metasignatures are identified by performing gene-list enrichment analyses using a transcription factor gene-set library created from the ChEA database,[17] or a histone modification gene-set library created by processing data from the Roadmap Epigenomics.[18] A schematic diagram of the overall approach is depicted in **Figure 1**. Consequently, this approach aims to facilitate the integration of such data for the ultimate purpose of guiding subtype-specific drug selection for individual patients for personalized and precision medicine. Along the way, the analysis identified discrepancies between our current classification of cell lines and patient tumors, suggesting two major subtypes of patient clusters in breast cancer. In addition, the analysis also uncovered molecular mechanisms of dysregulated pathways in specific subsets of patients and linked those pathways and clusters to expected outcome.

## RESULTS

### Stratification of patient tumors and cell lines

We first reanalyzed the classical gene expression data from the 122 breast cancer Stanford/Norway (S/N) patient tumor microarray study.[7] From this study, we only considered the authors' defined 453-probe signature used to stratify samples into the five known subtypes. From these 453 genes, we identified a 55-gene biomarker set that best stratified the five tumor subtypes using analysis of variance with $P < 0.00001$ cutoff after Benjamini–Hochberg correction. This 55-gene biomarker set was created using a similar approach applied to extract the widely used PAM50[11] biomarker gene set. PAM50 and our 55-gene biomarker gene set share nine genes: *CCNE1*, *ERBB2*, *ESR1*, *FOXA1*, *FOXC1*, *GRB7*, *KRT5*, *NAT1*, and *SLC39A6*. To visualize the heterogeneity of breast cancer tumors from the S/N data set, the biomarker set of genes was used to generate principal component analysis and hierarchical clustering plots (**Supplementary Figure S1** online). The plots show that the tumors segregate well based on subtype. However, the luminal B subtype

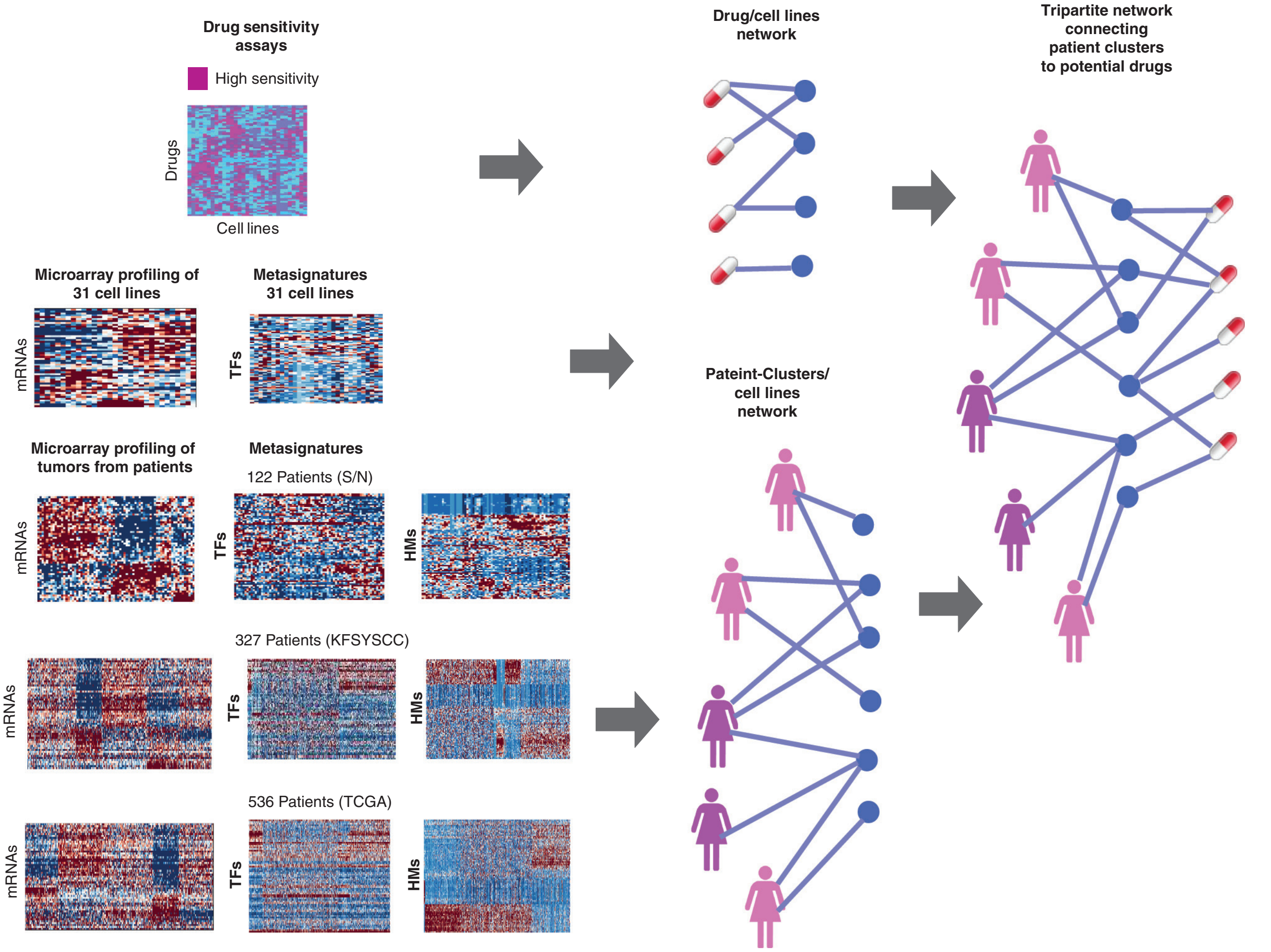
Drug sensitivity assays
High sensitivity
Drugs
Cell lines
Drug/cell lines network
Tripartite network connecting patient clusters to potential drugs
Microarray profiling of 31 cell lines
mRNAs
Metasignatures 31 cell lines
TFs
Pateint-Clusters/ cell lines network
Microarray profiling of tumors from patients
Metasignatures
122 Patients (S/N)
mRNAs
TFs
HMs
327 Patients (KFSYSCC)
mRNAs
TFs
HMs
536 Patients (TCGA)
mRNAs
TFs
HMs

**Figure 1** Schematic of the pipeline for the generation of integrated tripartite networks of patient clusters, cell lines, and drugs. Drug/cell-line correlation matrix was integrated with gene expression or metasignatures matrices from cell lines and from three studies that profiled gene expression in tumors from patients. From each expression or metasignature matrix, clusters of patients were identified and mapped to individual cell lines, whereas cell lines were connected to the drugs that show most potency in inhibiting their proliferation. KFSYSCC, Koo Foundation Sun-Yat-Sen Cancer Center; TCGA, The Cancer Genome Atlas. S/N, Stanford/Norway.

is slightly mixed with the ERBB2+ subtype. The hierarchical clustering plot shows that *STARD3*, a lipid transporter, and *MED1*, a transcriptional regulator, are uniquely upregulated in the ERBB2+ subtype; a set of 16 genes (center cluster) are upregulated in the basal tumor subtype; whereas 38 genes (top cluster), one of which being *ESR1*, are found to be predominantly upregulated in the luminal A subtype.

Next, we examined the ability of the 55-gene biomarker set to stratify patients from two other large-scale studies that profiled breast cancer tumors from individual patients: the Koo Foundation Sun-Yat-Sen Cancer Center (KFSYSCC) study (327 patients)[19] and The Cancer Genome Atlas (TCGA) study (536 patients).[2] In addition, gene expression and drug–response data were extracted from the supplementary data of the study by Heiser *et al*.[15] A subset of 31 cell lines was chosen because those cell lines have both gene expression and drug–response data. Of the 55 biomarker genes, 52 from the S/N study were found in the other sources of expression data. Therefore, this 52 gene set was used as the final biomarker set (**Supplementary Table S1** online). This 52-gene biomarker set, identified from the S/N tumors, can segregate patients based on their known subtype in the independent KFSYSCC and TCGA tumor cohorts similar to previously published biomarker gene sets that significantly overlap with our set (**Supplementary Table S2** online). The principal component analysis plots show that the biomarker set produced subtype-specific groups, with the luminal cell lines grouping closer to the ERBB2+ tumors (**Supplementary Figure S1a–d** online). It is known that ERBB2+ have luminal origin but contain a higher genomic ERBB2 copy number. The differential-expression profiles show consistency among the proportion of patients classified into the various subtypes for KFSYSCC and TCGA; whereas showing that the basal subtype is the most distinct in differential expression. The biomarker set produced similar differential-expression patterns in the cell lines, which were independently categorized into their four previously defined subtypes (**Supplementary Figure S1d** online).

To assess similarity in gene expression profiles between the cell lines and the tumor groups, we correlated the mRNA expression of the cell lines with those of each patient cluster using the biomarker gene set (**Figure 2a**). Clustering the correlations between cell lines and tumor clusters revealed subtype-specific expression similarity and enabled matching of patient subtypes to cell-line subtypes. For the mRNA-expression analysis, we found that the basal subtype tumors have similar expression profiles to the basal subtype cell lines, as expected. However, we also found strong similarity of basal classified tumors to claudin-low cell lines, suggesting that claudin-low cell lines and basal subtyping of tumors may be similar at the transcriptome level. Of note, we found that ERBB2+ tumors are more similar to the AU-565, SK-BR-3, HCC1954, and UACC-893 cell lines relative to the other luminals and ERBB2+ cell lines, suggesting that not all ERBB2+ cell lines can be matched to ERBB2+ patients equally well. The higher correlation between the normal-like tumors and the basal subtype cell-lines HCC70 and HCC1187 is also consistent with what is expected. Once we have identified the various clusters of patients and cell lines, we can match them and connect them to the drugs that show most potency in growth inhibition and apoptosis for these cell lines.

## Computing metasignatures

As an alternative, we developed a different approach to identify clusters of patients and connect clusters of patients to cell lines. We first compared the normalized gene expression levels for each gene in each group of patients or cell lines to identify the genes that are highly expressed in a specific patient or cell line. For this, we computed the average and SD for each gene in each data set and then converted expression levels to *z*-scores that reflect deviation from the average expression. We then selected the genes that are highly expressed in each patient or cell line using the $P < 0.01$ cutoff to create a gene-set library stored in a gene matrix transpose format (**Supplementary Table S3** online). The next step is computing gene-list enrichment analysis using the Fisher's exact test for enrichment for transcription factors or histone modifications. Each list of highly expressed genes from each individual patient or cell line is compared for overlap with lists of genes identified to be regulated by mammalian transcription factors as determined by ChIP-seq data collected for the ChEA database,[17] or regulated by a histone modification as determined by the ChIP-seq-based experiments performed by the Roadmap Epigenomics Consortium.[18] We have created a gene-set library from the Roadmap Epigenomics by processing 644 ChIP-seq experiments to identify putative target genes for 27 histone-modification types measured in various human cell types. The results from the enrichment analysis that compared the upregulated genes in the patients and cell lines with the ChEA or histone modification gene-set libraries are matrices in which the columns represent patients or cell lines, whereas the rows represent enrichment terms (**Supplementary Table S4** online). The negative log of the overlap *P* value enrichment score can be considered as transcription factor or histone modification pseudo-activity and as such we name these matrices metasignatures. We followed exactly the same analysis as described above for mRNAs using the metasignatures instead of the mRNA values. Now the biomarker set is not made of genes, but of the most informative transcription factors or histone modification ChIP-seq experiments that overlap with the highly expressed genes in subsets of cell lines or patients (**Supplementary Figures S2** and **S3** online).

The results from such metasignature analysis attempt to divide the patients and cell lines into five subtypes. However, it appears that the patients are actually naturally divided into two relatively even major subtypes: tumors that are mostly regulated by Myc and RUNX1 and the activation mark H3K4ME3 histone modification; whereas the second group is made of tumors that are mostly regulated by Suz12 and the repressive histone mark H3K27ME3. This is consistent with our knowledge that Suz12 is a part of the polycomb repressive complex (PRC2) that is responsible for the H3K27ME3 modification,[20] which is found near suppressed genes. On the other hand, the H3K4ME3 is known to be an activation mark[21] and its association with Myc is also known. The segregation into two subtypes is most profound for the TCGA and KFSYSCC data sets, both showing a similar pattern that is consistent with the S/N data set. However, the S/N data set also shows a strong Smad signature for

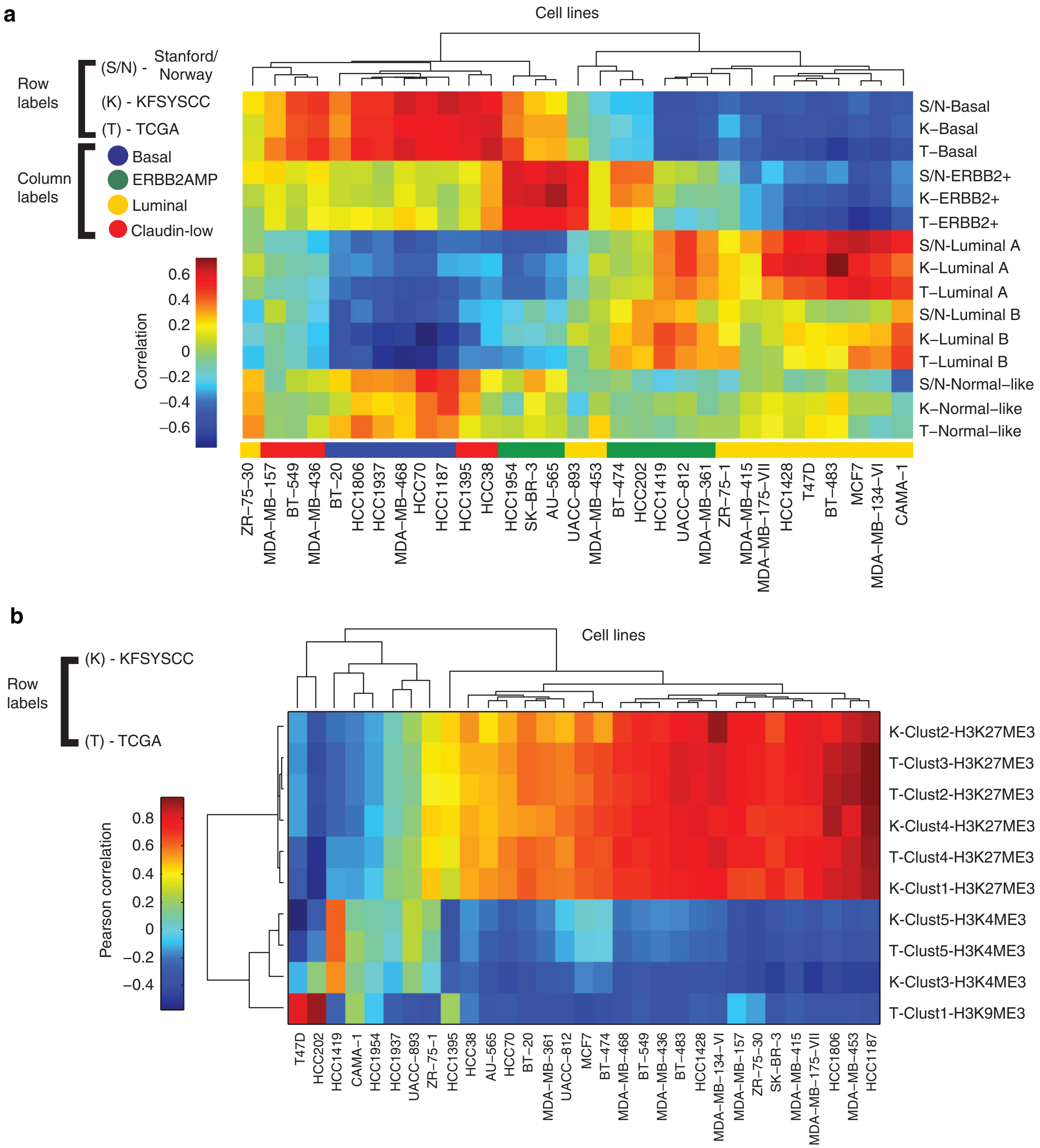

**Figure 2** Connecting cell lines to clusters of patients. (**a**) Correlation of coexpressed patient tumor clusters and cell lines using the biomarker gene set or (**b**) the supervised metasignature approach applied using the histone modification gene-set library. KFSYSCC, Koo Foundation Sun-Yat-Sen Cancer Center; TCGA, The Cancer Genome Atlas.

the basal tumors, which are only weakly detected in the TCGA and KFSYSCC data sets. Correlation with the cell lines (**Figure 2a,b** and **Supplementary Figure S4** online) shows that most cell lines match the Suz12-PRC cluster of patients; whereas the HCC1419 cell line is the only line that matches the Myc-H3K4ME3 signature. Of note, the TCGA data set has a small group of patients that show enrichment for H3K9ME3. These patient clusters correlate with the T47D and HCC202 cell lines (**Figure 2b**).

The biomarker set approach is supervised and tries to divide patients into their expected five categories. Hence, the analysis is biased toward previous classification of the patients from the S/N study into five subtypes. To test the metasignature approach without this bias, we also created hierarchical clustering plots using the metasignatures of KFSYSCC and TCGA without the filtering step of identifying a biomarker set based on previously defined subtypes (**Figure 3a,b** and **Supplementary Figures S5–S8** online). This

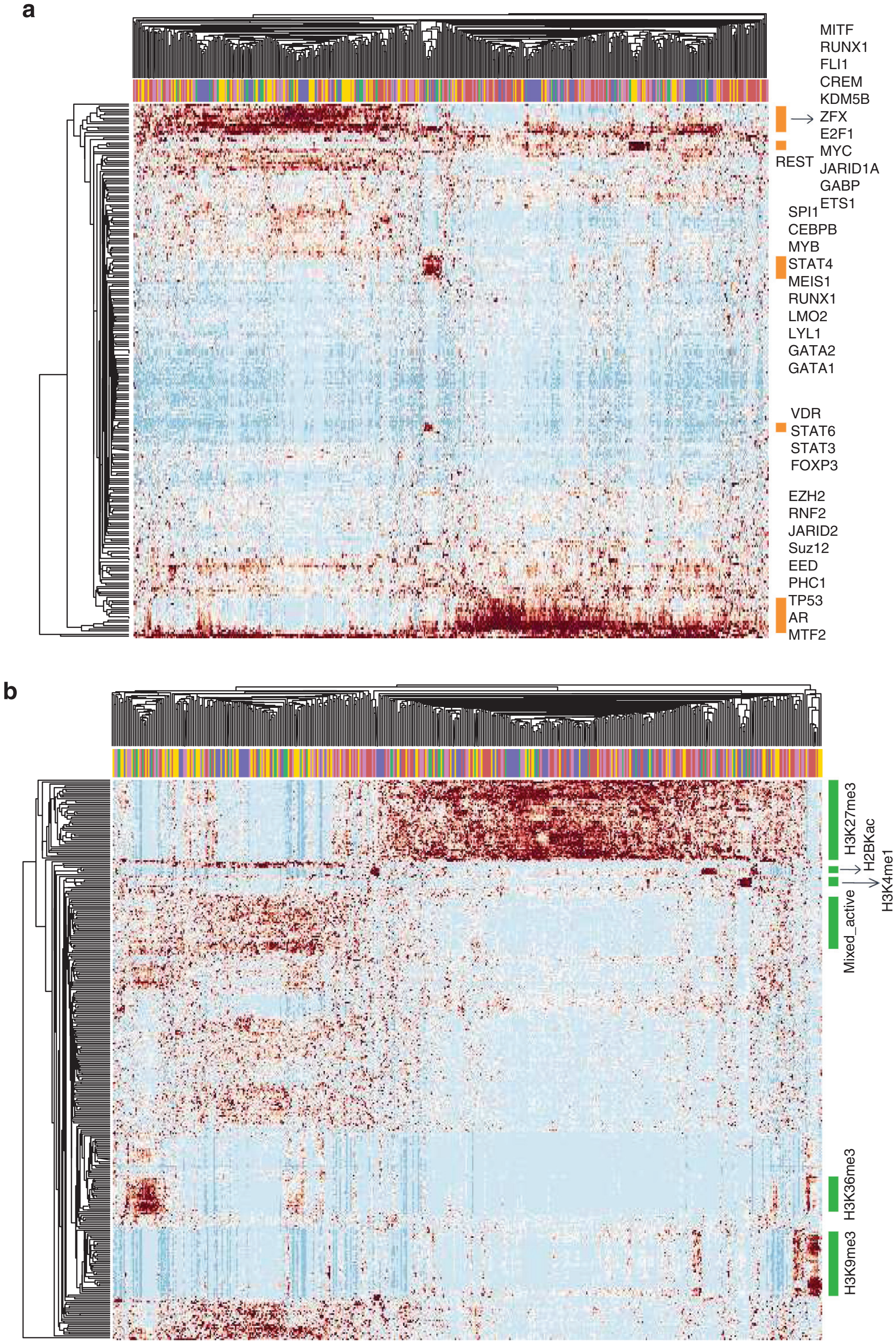

**Figure 3** Unsupervised metasignature clustering of The Cancer Genome Atlas patients using the (**a**) ChEA or (**b**) histone modifications gene-set libraries. Patients are colored based on the classifications determined by the supervised mRNA approach according to their classified subtype using the 52 biomarker gene set.

analysis identified the two major clusters as above, and new distinct small clusters of patients in the TCGA and KFSYSCC data sets. It is interesting that the TCGA and KFSYSCC data sets show very similar patterns of global metasignatures even though the studies used various microarray platforms. For example, a set of patients enriched with the previously assigned luminal A subtype is found to be highly enriched for upregulated genes regulated by the RE1-silencing

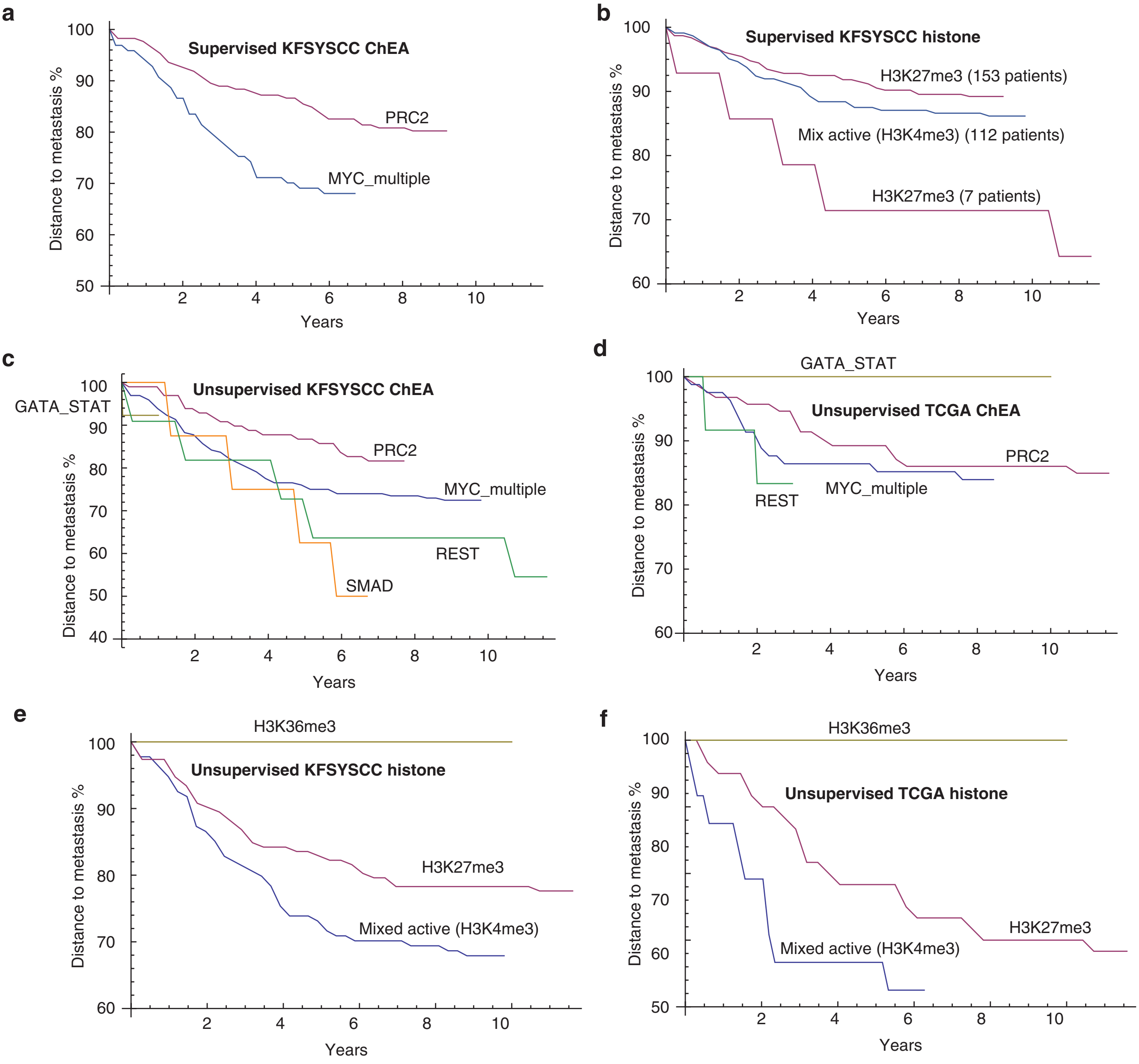

**Figure 4** Kaplan–Meier survival curves applied to the identified clusters using the metasignature supervised and unsupervised methods. (**a**) The two major clusters identified using the ChEA metasignatures in the KFSYSCC data set. (**b**) The two major clusters identified using the histone modification metasignatures in the KFSYSCC data set, as well as a smaller cluster of seven patients. (**c–f**) Survival curves for the unsupervised clustering applied to the TCGA and KFSYSCC data sets. The major clusters are the lines with the more refined fluctuations. These clusters correspond to clusters shown in Figure 3 and **Supplementary Figures S1–S6** online. KFSYSCC, Koo Foundation Sun-Yat-Sen Cancer Center; TCGA, The Cancer Genome Atlas.

transcription factors (Figure 3a for TCGA, and **Supplementary Figure S5** online for KFSYSCC). Another distinct cluster is enrichment for factors that include STAT3, 4, and 6, and MYB, and CEBPB, and GATA 1 and 2.

Overall and globally, the patients are divided into two major groups: the MYC group that also includes the RUNX1, E2F1 transcription factors, and the second group, which includes Suz12 and P53 as key enriched transcription factors. The previous assignment of patients into their designated clusters is highly mixed but definitely not random because small clusters of patients all belong to the same subtype, one of the five established subtypes. The histone-modification metasignatures applied to the TCGA and KFSYSCC data also contain distinct small clusters of patients, but divide the cohorts into the two main groups (Figure 3b for TCGA, and **Supplementary Figure S6** online for KFSYSCC). The correlation of the unsupervised TCGA and KFSYSCC metasignature with the unsupervised metasignatures computed for the cell lines shows high similarity to the results with the supervised approach. Most cell lines highly correlate with the Suz12/P53/H3K27ME3 metasignatures of patients and only the HCC1419 correlates with the active marks-enriched patients.

Some other cell lines, i.e., T47D, HCC202, HCC1937, and HCC1954 appear to have a unique correlation with small subsets of patients (**Supplementary Figures S9** and **S10** online). These patients may benefit from targeted therapies tailored specifically for them.

### Validation of newly identified clusters using distance to metastasis

The identification of new clusters of patients using the metasignature approach can be validated if it provides clear classification of patients with respect to observed outcome. For this, we analyzed the time-to-metastasis-event data available for both the TCGA and KFSYSCC data sets to evaluate the survival curves for each cluster identified by the metasignature approach. The results show clear and consistent division in expected outcome for the two major classes of patients: the Suz12/P53/H3K27ME3-enriched tumors have better prognosis than the MYC/RUNX1/H3K4ME3-enriched tumors (**Figure 4**). In addition, the STAT3/GATA/H3K36ME3 cluster shows very good prognosis with almost no recurrence events (**Figure 4d**–**f**), whereas the RE1-silencing transcription factor and SMAD-enriched cluster have very poor prognosis (**Figure 4c**). **Figure 4** only shows recurrence curves that are statistically significantly different (paired log-rank test, $P < 0.05$).

### Integrated network visualization of patient tumors, cell lines, and drugs

Next, we processed the drug–response data for the 31 cell lines treated with 77 drugs from the study by Heiser *et al.*[15] Response was quantified as the concentration of the drug needed to inhibit 50% of cell growth ($IG_{50}$). The concentrations were converted into sensitivity measures by taking the $-\log_{10}(IG_{50})$; this means that higher values correspond to higher sensitivity of a cell line to a drug. Finally, to provide a condensed, integrated view of the connections between the independent data sets and data types, we created tripartite networks that capture the connections between gene expression signature, or metasignatures, from the patients and cell lines with drug–response data for the 31 cell lines treated with 77 drugs. These data sets were integrated into tripartite graphs illuminating the indirect relationships between patient clusters and drugs (**Figure 5** and **Supplementary Figure S11** online).

The tripartite network created from the supervised mRNA approach automatically identified the luminal A cell lines HCC1428, BT-483, and MCF7. The CAMA-1 cell line was clustered with the luminal B clusters of patients and two ERBB+ cell lines: HCC202 and HCC1419. These two ERBB+ cell lines are appropriately sensitive to ERBB-signaling inhibitors. However, these inhibitors are predicted to work less well on the normal-like clusters of patients that are also connected to two ERBB+ cell lines. Although most cell lines are sensitive to chemotherapies that target microtubules, each identified cluster of patients and their associated cell lines are connected to different targeted therapies: e.g., heat shock protein inhibitors are predicted to work best for the luminal A cluster.

The tripartite networks created from the supervised (**Figure 5b** and **Supplementary Figure S11a** online) and unsupervised (**Supplementary Figure S11b** online) metasignature approaches show a consistent but clearer picture. The clusters of patients divide into two main groups with more cell lines connected to the Suz12/H3K27ME3 patients. These cell lines are more sensitive to the chemotherapies. Targeted therapies including kinase inhibitors such as those targeting EGFR and ERRB2, or PI3K or mTOR, are connected to the few ERBB+ cell lines and their corresponding patient clusters. The MEK inhibitor GSK1120212 is most specific for the HCC202 cell line, which is most similar to the H3K9ME3 cluster, suggesting these subgroup of patients are likely to benefit mostly by using this drug (**Figure 5b**).

## DISCUSSION

In this study, we developed a new method to cluster patients based on gene expression data. The method computes metasignatures for the upregulated genes in each patient based on a comparison across all patients. It would be interesting to also look at downregulated genes' metasignatures. The results from the metasignature analysis challenge current views of subtypes in breast cancer. It suggests two broad categories with a few additional distinct subtypes made of few patients. Low levels of trimethylation at lysine 27 have been previously associated with poor prognosis.[22] The fact that only few cell types match the Myc/ERBB2+ signature is surprising and could be due to issues with our computational settings, but can also challenge current dogmas in the field. If our analysis is correct, it suggests that more cell types from this type should be developed. It would be interesting to see if different clustering of patients will emerge when robust proteomics approaches become feasible. In this study, we also developed a network-based approach for integrating and visualizing gene expression similarity between patient tumors and cell lines, together with *in vitro* drug–response data. The network condenses, prioritizes, and connects heterogeneous data types to enable matching individual patients to potential treatments. Future work can prioritize drug combinations by also including drug-induced gene expression signatures collected from breast cancer cell lines.[23]

There are currently 53 drugs approved by the US Food and Drug Administration for use in breast cancer. Many are derivatives of the same drugs and many are chemotherapies targeting cell replication by DNA damage, microtubule polymerization disruption, or protein synthesis. Few targeted drugs exist, and these targets are mainly from the EGFR/ERRB2 or the ESR1 pathways. Broadly, our analysis suggests that the tumors with metasignatures enriched for the repressive marks Suz12/H3K27ME3 would benefit more from chemotherapies targeting microtubule polymerization disruption, whereas tumors with metasignatures enriched for active marks Myc/H3K4ME3 are more likely to benefit from targeted therapies such as those directed at the EGFR/ERRB2 pathway and PI3K/AKT pathway. There are many more experimental drugs that are pathway specific and these are currently being tested for both growth-inhibition response and global gene expression in many cell lines. It is expected that the results from such studies will lead to better specific therapeutics with fewer adverse events.[24]

One of the shortcomings of the metasignature approach is that the ChEA and histone modification gene set library

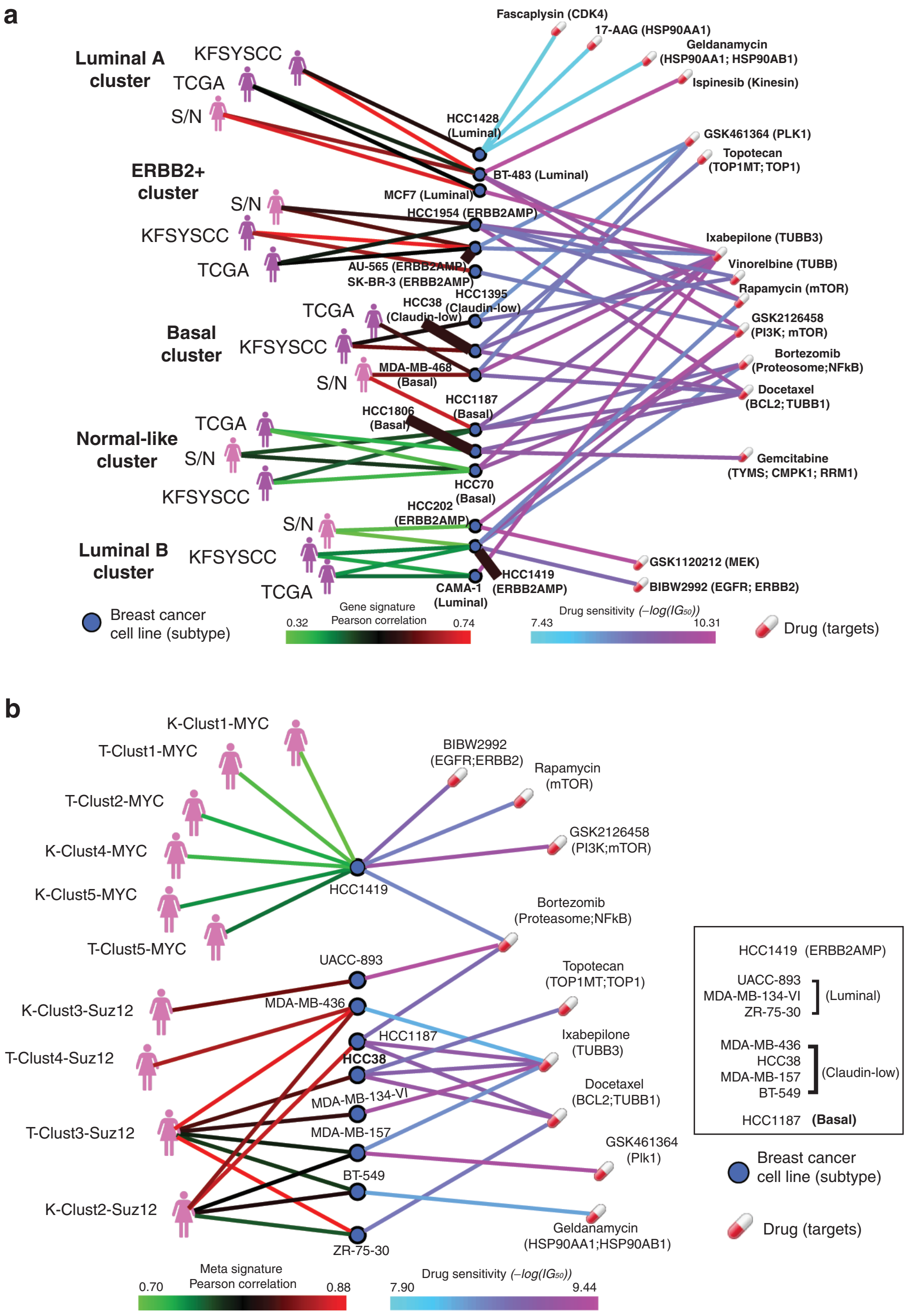

**Figure 5** Network that integrates gene expression and drug–response data to connect groups of patients, cell lines, and drugs. Edges between patient groups and cell lines are colored based on higher (red) or lower (green) expression correlation. Edges between cell lines and drugs are colored based on higher (magenta/purple) to lower (cyan) drug sensitivity. (**a**) On the basis of supervised mRNA expression; (**b**) On the basis of metasignature applied using the ChEA gene-set library. KFSYSCC, Koo foundation Sun-Yat-Sen Cancer Center; S/N, Stanford/Norway; TCGA, The Cancer Genome Atlas.

data sets are incomplete and come from many cell types; for ChEA, many of the ChIP-seq data are from mice. Within this data there might also be a bias for some specific cell types such as stem cells, which are highly represented in both data sets. Regardless, the advantage of the metasignature approach is that the results, besides providing a different level of clustering, suggest regulatory mechanisms specific for subtypes; these can serve as potential drug targets tailored for specific subtypes.

Currently, there are no clinical data available to validate the predictions made by our analyses. Clinical trials can be designed by classifying patients first into their respective subtypes, using various approaches, and then treating patients with the predicted drugs that match their subtype classification. Such an approach to clinical trials is increasingly becoming more accepted, but the gap between knowledge and practice is still wide.[25] Before clinical trials can be considered, drug responses of cell line need to be proven to be clinically appropriate and even before that, high-throughput drug responses for many cell-line studies need to show consistency across labs and publications. So far only two studies profiled many drugs across many cell lines in cancer,[5,16] and hopefully these two studies are consistent. More efforts in this direction and further validation in xenograft models, as well as considerations of side effects, could move us closer to testing the approach presented here in humans.

Although much is known about breast cancer, and the prognosis for this disease has substantially improved, there are other cancers with much worse prognoses, for which less is known and new therapeutics are desperately needed. Hence, data-integration approaches, such as the one presented here, may better fit these cancers. On the other hand, the method is data hungry, and less data are typically available for other cancers. Overall, the study is important for communicating ideas about data-integration opportunities and the types of analyses that gradually become more possible. However, conclusions about our findings need to be further confirmed by additional computational and experimental methods given that the approach has many limitations.

## METHODS

### Stratification of patient tumors and cell lines

Data from the S/N patient tumor gene expression microarray study (GEO accession GSE4335)[7] profiling 122 tumor samples from patients with breast cancer were reprocessed. Probes without a gene symbol or those belonging to multiple UniGene clusters as assessed by SOURCE (http://source.stanford.edu) were removed. Probes corresponding with the same gene symbol were averaged for each sample if the correlation between the probes was >0.7; otherwise the probe with the highest variance across samples was chosen, yielding the 453 unique gene biomarker set. Samples that exhibited close intrasubtype-cluster similarity were retained for further analysis ($n$ = 73 patients). From the 453 genes, genes that best stratified the five tumor subtypes, using analysis of variance with $P < 0.00001$ after Benjamini–Hochberg correction, were selected; thereby resulting in a 55-gene biomarker set. The $P$ value cutoff was empirically determined to yield the best stratification of tumors based on subtype. Gene expression data from a cohort of 327 fresh frozen tumors from patients with breast cancer diagnosed by the KFSYSCC were obtained from GEO (accession GSE20685).[19] All probes for the same gene symbols as for the S/N clones were then matched. Principle component analysis and hierarchical clustering plots were applied using MATLAB, Natick, MA.

### Integrated network visualization of patient tumors, cell lines, and drugs

To establish edges in the network, the two patient tumor data sets and the cell-line data set were independently standardized by subtracting the median expression of each gene. Each patient sample in the S/N 73-sample data set was assigned to one of five tumor clusters, corresponding to five known breast cancer subtypes. The 327 KFSYSCC tumor samples were clustered using K-means into five clusters as well, independent of the S/N data. Five K-means-cluster-centroids were chosen under the assumption that the KFSYSCC samples also contained patient tumors of five subtypes, similar to the S/N subtype designations. Pearson correlations were then computed between the mean expression of the biomarker genes in each patient cluster and the biomarker genes in each of the 31 cell lines, yielding 31 cell lines × 10 = 310 comparisons between patient tumors and cell lines. The patient tumor/cell-line edges were extracted from this adjacency matrix of correlations between the 10 patient clusters (five from each study) and the 31 cell lines using the 52 biomarker genes. For each patient cluster, edges representing the top 5% of cell lines with the highest correlation between the cell line and patient cluster were retained. Edges were colored from green to red in gradient, signifying lower to higher correlations. The cell line/drug edges in the network were extracted from the adjacency matrix of sensitivity measures between the 77 drugs and the 31 cell lines. For each cell line, edges were drawn for the top 5% of drugs that the cell line is most sensitive to, where edges were colored in shades of cyan to magenta, signifying lower to higher sensitivity. The network was visualized using the yEd software (http://www.yworks.com) and customized MATLAB scripts.

### Processing the ChIP-seq data from the roadmap epigenomics

The histone modification gene-set library was created by processing experiments from the Roadmap Epigenomics.[18] All ChIP-seq experiments from this data set were applied on human cell lines with antibodies targeting 27 different histone modification marks. ChIP-seq data sets from the Roadmap Epigenomics project deposited to GEO database were analyzed and converted to gene sets with the use of the tool SICER.[26] For each experiment, an input control sample was matched according to the description provided. ChIP-seq experiments without matched controls input were not included. The resulting gene-set library contains 27 types of histone modifications for 64 human cell lines from various tissue origins.

### Calculation of *P* values for the significance of differences between Kaplan–Meier curves

We consider two groups of patients who experience events (metastases) at various times and may be censored (left the study or death) at any time. Let $j = 1,2,3,\ldots,J$ be the indexes labeling the distinct times of events in either group. Then let $N_{1j}$ and $N_{2j}$ be the number of patients “at risk” (not experienced an event and still in the study) at time $j$, and let $N_j = N_{1j} + N_{2j}$. Let the number of observed events at time $j$ in each group be labeled $O_{1j}$ and $O_{2j}$, respectively, with the total number $O_j = O_{1j} + O_{2j}$. We then make the null hypothesis: each group is identically distributed.

In this case, the number of observed events in the first group, $O_{1j}$, at any given time should be distributed according to the Hypergeometric distribution, with mean: $E_{1j} = \frac{O_j}{N_j} N_{1j}$ and variance:

$$V_{1j} = \frac{O_j\left(\frac{N_{1j}}{N_j}\right)\left(1-\frac{N_{1j}}{N_j}\right)\left(N_j - O_j\right)}{N_j - 1}$$

The log-rank statistic is then: $Z = \sum_{j=1}^{J}(O_{1j} - E_{1j}) / \sqrt{\sum_{j=1}^{J} V_j}$, which can then be compared with the standard Gaussian distribution to derive the *P* value.

**Acknowledgments.** This work is partially supported by National Institutes of Health grants R01GM098316-01A1, P50GM071558-03, R01DK088541-01A1, and U54HG006097-02S1 to A.M. We thank Mario Niepel from Harvard Medical School for useful discussions.

**Author contributions.** A.M., S.G., and N.R.C. wrote the manuscript; A.M., S.G., and N.R.C. designed the research. A.M., S.G., Q.D., Y.K., and N.R.C. performed the research; A.M., S.G., Q.D., Y.K., and N.R.C. analyzed the data.

**Conflict of interest.** The authors declared no conflict of interest.

## Study Highlights

### WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

✓ Genome-wide mRNA-expression profiling has been applied to profile breast cancer cell lines and large cohorts of tumors from patients, identifying five major subtypes.

### WHAT QUESTION DID THIS STUDY ADDRESS?

✓ The study combines drug–response data and basal gene expression data from cell lines and tumors to identify patient clusters and map combinatorial therapy for each cluster.

### WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

✓ The study identifies new clusters of patients using a metasignature approach. The approach identifies distinct clusters of patients that are strongly correlated with prognosis as well as pointing to specific transcriptional regulatory mechanisms at play for each cluster.

### HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS

✓ The study identified clusters of patients with breast cancer in a unique way that might be more powerful than current methods; it also rationally connects clusters to therapeutic options through computational analysis and data integration.

1. Korde, L.A. *et al.* Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. *Breast Cancer Res. Treat.* **119**, 685–699 (2010).
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
3. Stephens, P.J. *et al.*; Oslo Breast Cancer Consortium (OSBREAC). The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
4. Gray, J. & Druker, B. Genomics: the breast cancer landscape. *Nature* **486**, 328–329 (2012).
5. Perou, C.M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
6. Van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
7. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10869–10874 (2001).
8. Malo, T.L., Lipkus, I., Wilson, T., Han, H.S., Acs, G. & Vadaparampil, S.T. Treatment choices based on OncotypeDx in the breast oncology care setting. *J. Cancer Epidemiol.* **2012**, 941495 (2012).
9. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).
10. Cardoso, F., Van't Veer, L., Rutgers, E., Loi, S., Mook, S. & Piccart-Gebhart, M.J. Clinical application of the 70-gene profile: the MINDACT trial. *J. Clin. Oncol.* **26**, 729–735 (2008).
11. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8418–8423 (2003).
12. 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).
13. Galanina, N., Bossuyt, V. & Harris, L.N. Molecular predictors of response to therapy for breast cancer. *Cancer J.* **17**, 96–103 (2011).
14. Gunter, C. Cancer genomics: constructing a 'cancerpaedia'. *Nat. Rev. Genet.* **13**, 300 (2012).
15. Heiser, L.M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 2724–2729 (2012).
16. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
17. Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R. & Ma'ayan, A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
18. Bernstein, B.E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
19. Kao, K.J., Chang, K.M., Hsu, H.C. & Huang, A.T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11**, 143 (2011).
20. Young, M.D. *et al.* ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.* **39**, 7415–7427 (2011).
21. Balasubramanian, D. *et al.* H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island-containing start sites. *Genome Med.* **4**, 47 (2012).
22. Wei, Y. *et al.* Loss of trimethylation at lysine 27 of histone H3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers. *Mol. Carcinog.* **47**, 701–706 (2008).
23. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
24. Jenkins, S.L. & Ma'ayan, A. Systems pharmacology meets predictive, preventive, personalized and participatory medicine. *Pharmacogenomics* **14**, 119–122 (2013).
25. Vaidyanathan, G. Redefining clinical trials: the age of personalized medicine. *Cell* **148**, 1079–1080 (2012).
26. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K. & Peng, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).

Supplementary information accompanies this paper on the *Pharmacometrics & Systems Pharmacology* website (http://www.nature.com/psp)