Comparison of mRNA- and IHC-determined ER- and ErbB2-status

To compare the mRNA- and IHC-determined ER- and ErbB2-status, we applied the following procedure. We applied a Gaussian mixture model onto the expression data for ESR1 and ERBB2 recorded using probe sets 205225_at and 216836_s_at respectively as described by Lehmann and colleagues (J Clin Invest, 2011). The Gaussian mixture model allows the identification of multimodality in gene expression data. Put in other words, if the distribution of the gene expression data is composed of a mixture of discrete Gaussian distributions, the Gaussian mixture model allows the identification of these discrete distributions. When applied onto the expression data for ESR1 and ERBB2, we clearly identified a bimodal expression profile corresponding to the presence of two subpopulations in our study population, each defined by the presence or absence of either ER or ErbB2 expression. Supplementary Figure 1A and B demonstrates the distribution for ESR1 and ERBB2 in histogram format (grey bars). On top of the histogram, the real density distribution (blue line) and the theoretical density distribution derived from the Gaussian mixture model (black line) are fitted. The distribution of ESR1- and ERBB2-expression according to the Gaussian mixture model in each subpopulation is indicated using dashed red and green lines. The expression value for ESR1 and ERBB2 corresponding to the point were both distributions overlap (arrow) is used as threshold to dichotomize samples according to the presence (positive) or absence (negative) of ESR1- or ERBB2-expression. The resulting mRNA-based classification is then compared to the classification of the same samples according to the immunohistochemical data for ER and ErbB2. Results thereof are provided in
Supplementary Table 1. We observe a 90% concordance between the mRNA- and immunohistochemistry-based classifications.

*Performance of the normalization procedure*

To investigate the performance of the regression normalization in removing differences in gene expression between the different data sets based on the use of different technology platforms included in the study, we performed PCA on the samples before and after the normalization step. The results are shown in Supplementary Figure 2. Prior to the regression normalization (A), the samples in the 2D scatter plot representation are clearly grouped according to the site of inclusion (red: IPC, green: MDA, blue: TCRU), whereas after the regression normalization (B), no clustering according to inclusion site is observed. To investigate whether the regression normalization maintained biological variation present in the data set, we classified each of the samples in our data set according to the molecular subtypes with the SSP-algorithm and compared this classification with an ER activation score obtained using the signature published by Gatza et al (20). Results are demonstrated in a boxplot format in Supplementary Figure 2C. Kruskal-Wallis testing revealed a significant difference in ER activation between the subgroups with the highest ER activation in the samples belonging to the luminal subtypes and the lowest ER activation observed in the triple-negative subtypes (Basal-like and Claudin-Low). The data presented here suggest that the regression normalization was able to remove technical differences in gene expression while maintaining the information relevant to biological differences.
Determination of heterogeneity in IBC and nIBC breast cancers through silhouette score analysis

As described in the body of the manuscript, we used a silhouette score analysis approach to investigate common themes in gene expression within the IBC and nIBC tumor phenotype separately as well as for both tumor phenotypes combined. We randomly selected 100 samples from the three groups of interest. On this selection, unsupervised hierarchical cluster analysis (UHCA) using Manhattan distance as distance measure and Ward linkage as the dendrogram drawing method was performed only for those genes having a standard deviation greater than 2. UHCA was followed by a cluster robustness analysis using the silhouette algorithm. The silhouette algorithm divides the dendrogram into an increasing number of clusters. We chose to analyze a maximum number of 20 clusters. For each increase in the number of clusters, the algorithm generates a score, the magnitude of which is proportional to the robustness of the identified clusters. This analysis was repeated 10 times, resulting in series of 10 silhouette scores for each number of clusters. These series of 10 silhouette scores for each 1-unit increase in cluster number are depicted in boxplot format in Supplementary Figure 3. For each data set (i.e. IBC, nIBC and combined BC) a different panel is provided (A through C). The number of clusters in each group of interest (IBC, nIBC or combined) was then determined by sequentially comparing the silhouette scores (silhouette scores for 2 clusters vs. silhouette scores for 3 clusters, silhouette scores for 3 clusters vs. silhouette scores for 4 clusters and so on). The result of this comparison is provided in Supplementary Figure 4 (black: IBC, red: nIBC and green: combined). The X-axis indicates for which clusters the silhouette scores are compared. Each point represents the P-value when comparing the silhouette scores obtained for n clusters with the silhouette scores
obtained for \( n + 1 \) clusters. Thus, since the lowest number of clusters equals 2, the first point in each series represents the comparison of silhouette scores for 2 clusters with the silhouette scores obtained for 3 clusters. The grey horizontal lines indicate a \( P \)-value of 0.05 (solid grey line) and 0.1 (dashed grey line). The number of clusters equals the highest amount of clusters for which the silhouette score significantly drops when increasing the amount of clusters with one unit. In case of the IBC series (black line), this happens the number of clusters is increased from 4 to 5. For nIBC (red line) and the combined data set (green line) this happens, respectively, when of the number of clusters is increased from 6 to 7 and 7 to 8. The silhouette scores for the complete data sets are provided in Supplementary Figure 3D (black: IBC, red: nIBC, green: combined). What is readily observable is the similar variation in the silhouette scores when analyzing the complete data sets or subsets of 100 patients, strengthening the silhouette analysis outlined above. The blue line represents the silhouette scores obtained when executing the algorithm on 137 randomly selected nIBC samples. The blue and the red line follow a very similar course indicating that the difference in heterogeneity observed between IBC and nIBC (respectively 4 clusters and 6 clusters) is not solely attributable to a difference in sample size (137 vs. 252).

Identification and validation of the algorithm to identify transcription factor activation

A list of 234 transcription factors and their predicted targets was retrieved from the tfacts-website (www.tfacts.org) (24). For each of the 234 tested transcription factors, the transcription factor activation score was calculated per sample by subtracting the median expression level from the transcription factor non-target
genes from the median expression level from the transcription factor targets. As such, the activation score will be small when the difference in median expression between the transcription factor targets and non-targets is small and vice versa. A small difference is indicative for the inactivity of the involved transcription factor. To validate the transcription factor algorithm used in this paper we retrieved 27 publicly available gene expression data sets from cell lines treated for perturbation of in total 40 different transcription factors. We applied the algorithm and evaluated whether or not the activation profiles for the involved transcription factors predicted by the expression of their target genes are different between the treated and control conditions. The results are provided in Supplementary Table 2. Each unique data set is considered as a single experiment, but for some experiments multiple transcription factors were evaluated. For some transcription factors we evaluated the algorithm twice, specifically for those transcription factors for which our algorithm failed to correctly predict the transcription factor activation at least once (ER and CREB; indicated in bold), the algorithm was evaluated twice. For each experiment we also indicated the published reference in which the used data set was discussed. As the number of samples per experiment is often low, we not only evaluated the significance of the prediction (P-Value) but also the predicted effect (activation or inhibition). For the 53 comparisons performed, we failed to correctly predict the effect in only 2 cases (error rate: 96%). For those experiments for which the transcription factor activation was predicted correctly (N=51), we observed significance in 75% of the time at a P-value of 0.1. Given these data, we state that the algorithm to predict activation of transcription factors using gene expression data is validated. Nevertheless, caution is warranted, as we did not test all of the transcription factors available in our algorithm.
Validation of the molecular subtype-independent IBC/nIBC signature

In order to identify probe sets with an IBC-specific expression component, we adopted the following strategy. We split the entire data set into a training set of 250 samples and a validation set of 139 samples. In the training set, differentially expressed probe sets were identified using regression models, which allowed modeling the expression data for each probe set in function of the variation in both the tumor phenotype (IBC vs. nIBC) and the molecular subtype within our data set. As such, a series of regression coefficients was obtained for each probe set (=dependent variable), one for each independent variable in the model (IBC, nIBC, Basal-like, HER2-enriched, Claudin-Low, Luminal A, Luminal B and Normal-like). These regression coefficients are a measure for the extent to which the variation in probe set expression is determined by each of the independent variables, with low values indicating limited effects and vice versa. The modeling of the probe set expression in terms of the variation in tumor phenotype and molecular subtype allowed us to distinguish the differences in gene expression based on molecular subtype-specific differences from differences in gene expression due to tumor phenotype. A schematic outline of the analysis performed is provided in Supplementary Figure 5A. Two groups of probe sets were identified, one group with uniquely IBC-specific probe sets (N=79) and one group with both molecular subtype- and IBC-specific expression components (N=412). Both the full list of probe sets with an IBC-specific expression component (N=491) and the list of probe sets with uniquely IBC-specific expression components were validated.

To analyze to what extent both lists of probe sets with IBC-specific expression components truly reflects the biology of IBC, we performed a global test on the
validation set. We observed significant global expression differences between IBC and nIBC for both evaluated lists of probe sets with a slightly better performance for the list of uniquely IBC-specific probe sets (Observed \( Q = 51.677, \ P < 0.001 \) vs. Observed \( Q = 50.893, \ P < 0.001 \)). The sample plots (barplot format) resulting from this analysis are provided in Supplementary Figure 5B. The sample plot shows a color-coded bar (IBC: yellow; nIBC: purple) and a reference line for each sample. The bar shows the influence of each sample on the test statistic \( Q \). Samples that are similar in expression profile to samples of the same phenotype (IBC vs. nIBC) carry evidence in favor of a significant \( P \)-value and therefore have a positive influence on the test statistic \( Q \). Samples with negative influence on the test statistic \( Q \) carry evidence against a significant \( P \)-value as they are relatively similar in expression profile to samples with a different phenotype. Marks on the bars show the standard deviation of the influence under the null hypothesis for those samples, which are more than one standard deviation away from zero. The bars are ordered according to the tumor phenotype and within each tumor phenotype according to the magnitude of the influence of the sample on the test statistic \( Q \).

**IPA for genes with an IBC-specific expression pattern**

Genes with an IBC-specific expression component were subjected to analysis using IPA software. The gene list was subdivided in genes that were overexpressed or repressed in IBC. To calculate enrichment scores for canonical pathways, all probe sets on the Affymetrix HGU133A chip were used as background. The results are presented in Supplementary Table 6, which provides a list of all IBC-specific genes per significant canonical pathway as well as the nominal \( P \)-value for the gene set enrichment analysis.
To construct networks, only direct and indirect relationships that have been experimentally observed in humans in all data sources available to the software were considered. In addition, the molecules included in the networks were limited to human molecules or human endogenous chemicals. Network analysis was limited to 35 molecules per network and 25 networks per analysis. We identified 12 networks that were upregulated in IBC, 1 of which contained at least 20 focus molecules (i.e. probe sets included in our list of IBC-specific genes). Supplementary Figure 6A-C shows the three most significant activated networks. The molecules are color-coded red if the corresponding probe set is included in the list of upregulated probe sets. The intensity of the color reflects the magnitude of the fold change. Uncolored nodes are added by the software and are not represented in the list of 491 differentially expressed probe sets. Solid lines signify direct gene-gene interactions, whereas broken lines represent indirect relationships that may require secondary effectors not depicted in the network. All connections are supported by at least one published report or from canonical information stored in the Ingenuity Pathway Knowledge Base. Blunt arrowheads indicate inhibitory interactions; sharp arrow heads indicate activating interactions. We also identified 19 networks repressed in IBC, 3 of which contained at least 20 focus molecules. Supplementary Figure 7A-C shows the three most significant repressed networks. Here the molecules are color-coded green if the corresponding probe set is included in the list of downregulated probe sets. The meaning of the color-intensity, nodes and lines is similar as in Supplementary Figure 6. Supplementary Table 7 provides a listing of the networks shown in Supplementary Figure 6 and Supplementary Figure 7, the molecules included in these networks and, per network, the top enriched molecular functions.
To assess the prognostic value of the 79-gene signature in breast cancer, we applied the SVM-based classifier to expression data from 871 breast cancers available though publicly available data sets. A listing of included data sets is provided in Supplementary Table 3. Samples included in this analysis were derived from treatment-naïve patients with early, lymph node negative breast cancer. Application of the classification model allows defining two classes: “IBC-like” and “nIBC-like”. Both classes were compared with respect to Distant Metastasis-Free Survival (DMFS). Resulting Kaplan-Meier plots are provided in Supplementary Figure 8. Patients with “IBC-like” breast cancer have a significantly shorter DMFS-interval as compared to patients with “nIBC-like” breast cancer. We compared the prognostic value of the 79-gene signature in this series of samples against the several other parameters including patient age at diagnosis, histological grade, T-status and the classification according to the molecular subtypes. In univariate analysis all of these parameters, except for patient age at diagnosis, were associated with DMFS. In multivariate analysis, incorporating all variables associated with DMFS in univariate analysis, all variables except T-status were independently associated with DMFS. Details are provided in Supplementary Table 8. These data demonstrate that the 79-gene signature, which is composed of genes with a unique IBC-specific expression component, provides independent prognostic information in addition to the histological grade and the molecular subtype classification.
Supplementary Tables

Supplementary Table 1: Comparison of ER- and ErbB2-status determined using gene expression profiling and immunohistochemistry

Supplementary Table 2: Validation of the algorithm to identify transcription factor activation

Supplementary Table 3: Description of the public nIBC data sets

Supplementary Table 4: Molecular data of IBC and nIBC samples

Supplementary Table 5: List of 491 probe sets differentially expressed between IBC and nIBC samples in a molecular subtype-independent manner

Supplementary Table 6: IPA analysis of the list of 443 unique genes differentially expressed between IBC and nIBC

Supplementary Table 7: Networks identified by IPA analysis of the list of 443 unique genes differentially expressed between IBC and nIBC

Supplementary Table 8: Uni- and multivariate analyses for DMFS in the public pooled nIBC data set
Supplementary Figures

Supplementary Figure 1: Application of the Gaussian mixture model to ESR1- and ERBB2-expression data

Supplementary Figure 2: Normalization

Supplementary Figure 3: Silhouette Analysis

Supplementary Figure 4: Silhouette Analysis (P-Values)

Supplementary Figure 5: Validation of Limma-corrected IBC/nIBC signature

Supplementary Figure 6: Networks overrepresented in IBC

Supplementary Figure 7: Networks overrepresented in nIBC

Supplementary Figure 8: Prognostic value of the IBC/nIBC signature in nIBC