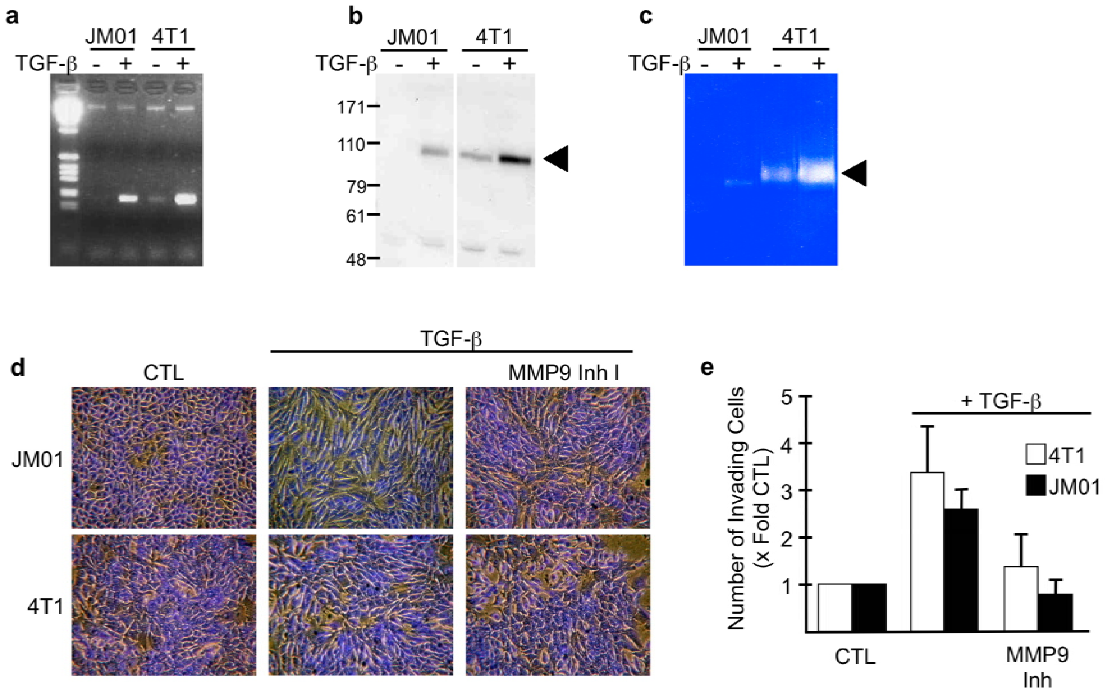
**Supplementary Information**

# Identification of high quality cancer prognostic markers and metastasis network modules

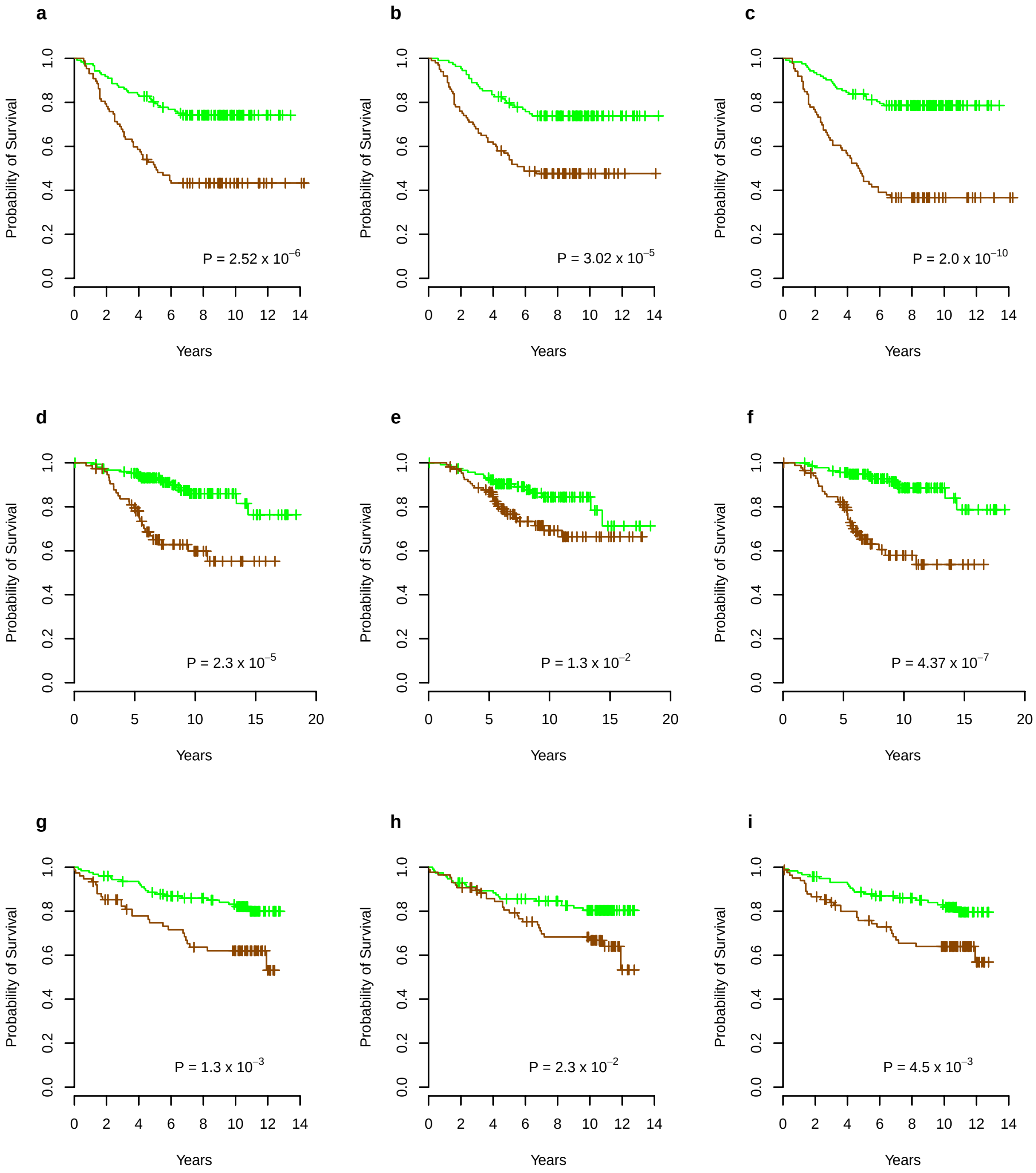
Jie Li, Anne EG Lenferink, Yinghai Deng, Catherine Collins, Qinghua Cui, Enrico O. Purisima, Maureen D O'Connor-McCourt, and Edwin Wang

# Table of Contents

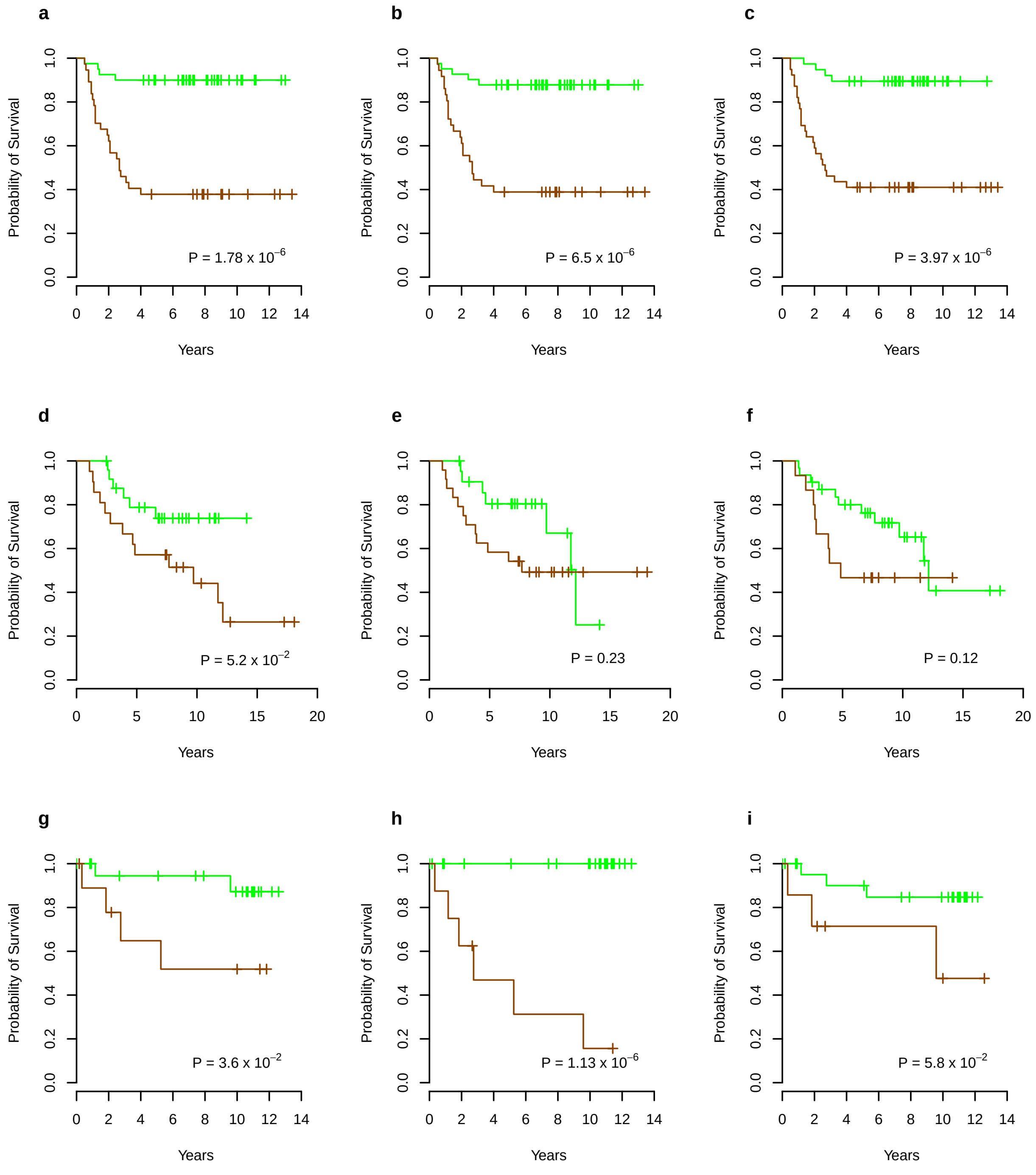
# Supplementary Figures



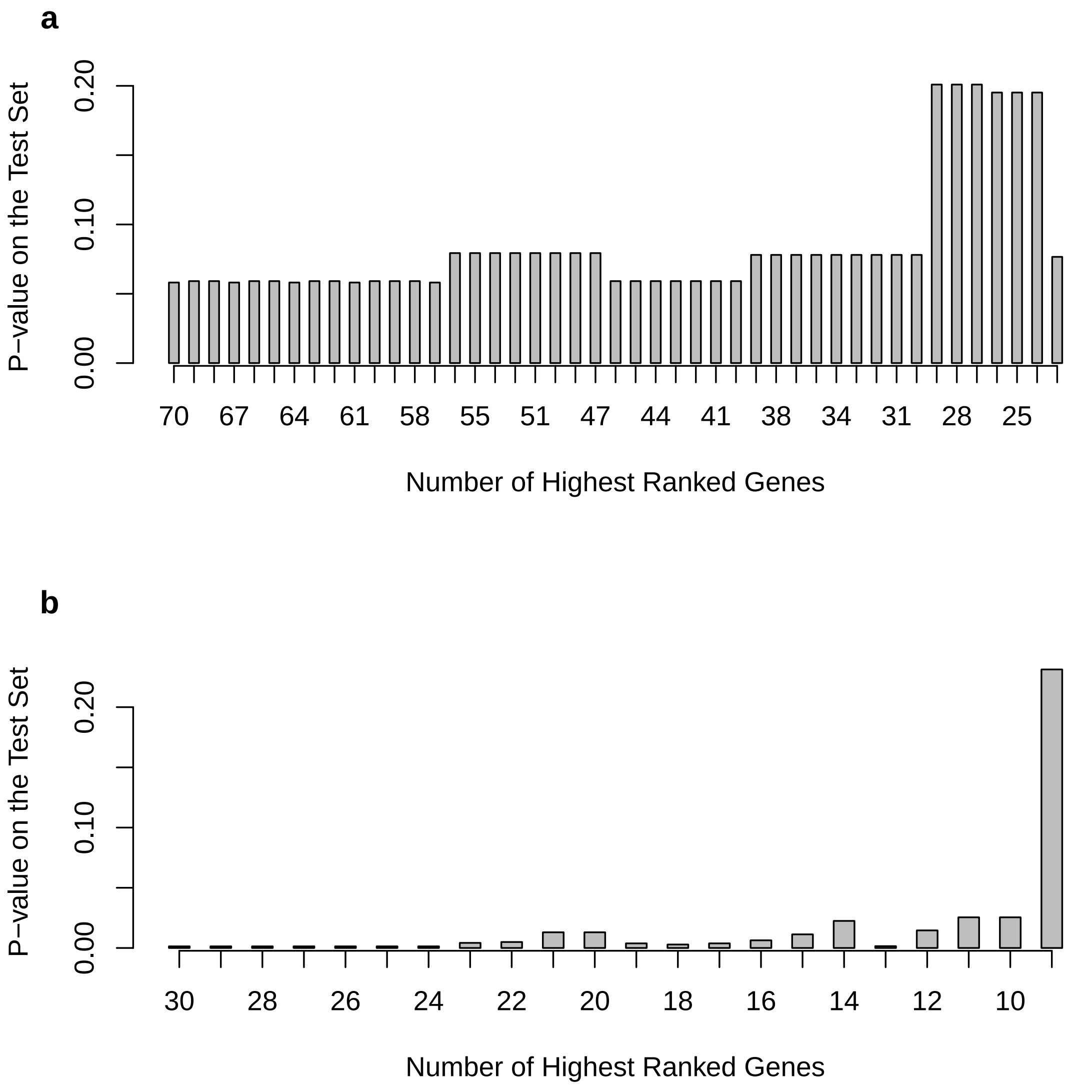
**Supplementary Figure S1 | MMP-9 activity plays a role in the invasiveness of BRI-JM01 and 4T1 mammary tumor cell lines.** We used an EMT (epithelial to mesenchymal transition) microarray dataset generated from BRI-JM01 mouse mammary cancer cell line treated with TGF-β[24, 25] and extracted ~100 secreted and cell surface proteins. Using the screening method[26] a survival analysis for these proteins was preformed using the breast cancer microarray datasets[27, 28]. The expression levels of seven genes, i.e., matrix metalloproteinase 9 (MMP9) were significantly correlated with patient survival. Here we validated the role of MMP9 in the BRI-JM01 cells and used the MMP9 secreting 4T1 mouse mammary tumor cell line[29] as a positive control. (**a**) RTPCR confirmed that TGF-β induces an upregulation of MMP9 expression in both cell lines. Additional western blot (**b**) and zymogram (**c**) analysis of the conditioned medium of TGF-β treated BRI-JM01 and 4T1 cells showed that MMP9 (◀) is secreted and enzymatically active, respectively. (**d**) A functional role of MMP9 in EMT was demonstrated by exposing BRI-JM01 (top panels) and 4T1 (bottom panels) cells to TGF-β in the presence or absence of a specific MMP9 inhibitor (MMP9 Inh I) which significantly inhibited the TGF-β induced morphology change in BRI-JM01, and to a lesser extent in 4T1 cells (magnification 40x). In addition, the results of a Transwell invasion assay (**e**) showed that TGF-β induced BRI-JM01 (white bars) and 4T1 (black bars) cells to penetrate and transgress a Matrigel barrier. The presence of MMP9 inhibitor significantly reduced invasiveness and confirms MMP9's role in the invasive character of BRI-JM01 cells undergoing EMT as a result of TGF-β exposure. Results are shown as the average (+/- SEM) of two independent experiments carried out in triplicate.



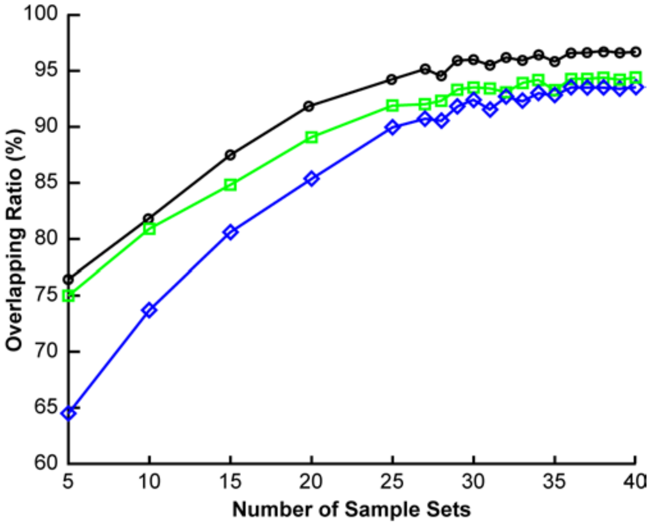
**Supplementary Figure S2 | Kaplan-Meier curves of the risk groups for the ER+ patients with 10-year disease-free survival.** Green and dark orange curves represent low- and high-risk groups, respectively. (**a**), (**b**) and (**c**) represent the National Research Council (NRC) gene signatures, NRC-1, -2 and -3 tested in the Wang cohorts, respectively. (**d**), (**e**) and (**f**) represent the NRC-1, -2 and -3 tested in the Chang cohorts, respectively. (**g**), (**h**) and (**i**) represent the NRC-1, -2 and -3 tested in the Miller cohorts, respectively. *P*-values were obtained from the log-rank test. ER+ indicates that the breast tumors are estrogen receptor positive.



**Supplementary Figure S3 | Kaplan-Meier curves of the risk groups for the ER- patients with 10-year disease-free survival.** Green and dark orange curves represent low- and high-risk groups, respectively. (**a**), (**b**) and (**c**) represent the National Research Council (NRC) gene signatures, NRC-7, -8 and -9 tested in the Wang cohorts, respectively. (**d**), (**e**) and (**f**) represent the NRC-7, -8 and -9 tested in the Chang cohorts, respectively. (**g**), (**h**) and (**i**) represent the NRC-7, -8 and -9 tested in the Miller cohorts, respectively. *P*-values were obtained from the log-rank test. ER- indicates that the breast tumors are estrogen receptor negative.



**Supplementary Figure S4 | Survival testing of the rank genes by one-by-one removal of the genes from the bottom of the lists.** (**a**) The list of rank-70 genes. (**b**) The list of the ranked genes of the National Research Council (NRC) gene signature, NRC-1 gene signature. *P*-values were obtained from the log-rank test. The Miller dataset was used for the testing.



**Supplementary Figure S5 | Relations of the size of the random datasets and the gene overlapping ratio for generating gene signatures.** Two distinct 1 million of randomly generated gene sets (RDSs, each set contains 30 genes) from cell cycle genes (see the main text) were used to run the Multiple Survival Screening (MSS) algorithm using different sizes of the random datasets. After each run, the genes were ranked based on the MSS. The number of the overlapped genes was determined from the top 30 ranked genes derived from each of the two millions of the RDSs. The overlap ratio equals to the number of the overlapped genes divided by 30. Gray, green and blue lines represent National Research Council (NRC) gene signatures, NRC-1, -2 and -3, respectively.

# Supplementary Tables

**Supplementary Table S1 Parameters used for identifying gene signatures using MSS**

| Gene signature | GO term | Number[1] | RDS size[2] | Training size[3] |
|---|---|---|---|---|
| NRC-1 | Cell cycle | 78 | 69 (43:26) | 209 (129:80) |
| NRC-2 | Immune response | 67 | 69 (43:26) | 209 (129:80) |
| NRC-3 | Apoptosis | 96 | 69 (43:26) | 209 (129:80) |
| NRC-4 | Cell adhesion | 87 | 23 (8:15) | 44 (13:31) |
| NRC-5 | Cell cycle | 80 | 23 (8:15) | 44 (13:31) |
| NRC-6 | Cell motility | 72 | 23 (8:15) | 44 (13:31) |
| NRC-7 | Apoptosis | 75 | 40 (25:15) | 77 (50:27) |
| NRC-8 | Cell adhesion | 69 | 40 (25:15) | 77 (50:27) |
| NRC-9 | Cell growth | 60 | 40 (25:15) | 77 (50:27) |

**Notes:** The table lists gene groups define by GO (Gene Ontology) terms, number of random microarray datasets (RDS) and the ratios of "good" and "bad" tumors in each RDS used for running the Multiple Survival Screening (MSS) algorithm to generate the NRC (National Research Council) gene signatures. "Good" and "bad" tumor patients represent the breast cancer patients whose tumors have not recurred and recurred, respectively, within 10 years after the removal of their primary tumors. [1]Number of genes of the gene group used for screening of the NRC signatures; [2]Number of samples in each random dataset (ratio of the "good" and "bad" samples); [3]Number of total samples in the original training set (ratio of the "good" and "bad" samples).

**Supplementary Table S2 Datasets used in training and testing of the NRC signatures**

| Dataset | Number of Samples | Number of ER+ Samples | Number of ER- Samples |
|---|---|---|---|
| GSE2034 (Wang cohort) | 286 | 209 | 77 |
| Chang cohort | 295 | 226 | 69 |
| GSE3494 (Miller cohort) | 236 | 205 | 31 |
| GSE11121 | 200 | 156 | 44 |
| GSE1456 | 159 | 131 | 28 |
| GSE9195 | 77 | 77 | 0 |
| GSE6532 | 293 | 250 | 43 |
| GSE7378 | 54 | 54 | 0 |
| GSE12093 | 136 | 136 | 0 |
| Total | 1736 | 1444 | 292 |

**Notes:** Microarray datasets of breast tumor samples were used for training and testing of the NRC (National Research Council) gene signatures. Datasets are represented by NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) IDs. The Chang dataset was obtained from Chang et al., PNAS, 102:3738, 2005. Chang cohort is cDNA arrays, while others are Affymetrix arrays. ER+ and ER- indicate that the breast tumors are either estrogen receptor positive or negative.

**Supplementary Table S3 *P*-value of NRC signatures and meta-gene signature on three datasets**

| Sample | ER+ | | | |
|---|---|---|---|---|
| | NRC-1 | NRC-2 | NRC-3 | NRC-1,2,3-Meta |
| Wang cohort | $2.52 \times 10^{-6}$ | $3.02 \times 10^{-5}$ | $2.0 \times 10^{-10}$ | $1.46 \times 10^{-8}$ |
| Chang cohort | $2.30 \times 10^{-5}$ | $1.30 \times 10^{-2}$ | $4.37 \times 10^{-7}$ | $2.38 \times 10^{-5}$ |
| Miller cohort | $1.30 \times 10^{-3}$ | $2.30 \times 10^{-2}$ | $4.5 \times 10^{-3}$ | $1.60 \times 10^{-3}$ |

**Notes:** NRC (National Research Council) gene signatures and the meta-gene signatures (by appending genes from different sets of NRC signatures) were applied to the three breast patient cohorts to predict “good” and “bad” tumors. “Good” and “bad” tumor patients represent the breast cancer patients whose tumors have not recurred and recurred, respectively, within 10 years after the removal of their primary tumors. values were obtained from the log-rank test. Lower *P*-values represent better predictions. NRC-1,2,3-Meta (meta-gene signature) is the collection of the genes from NRC-1, -2, and -3. ER+ indicates that the breast tumors are estrogen receptor positive.

# Supplementary Methods

**Pseudo-code for the Multiple Survival Screening (MSS) algorithm.** Below is a detailed description of the algorithm shown schematically in Figure 2a of the main text. The actual programs and examples for running MSS are available as Supplementary Software.

1. Generate the survival gene pool.
    a. Analyze the gene expression data in the training set for 10-year disease-free survival as implemented previously[26].
    b. Use fuzzy clustering to classify the samples into 2 classes. Genes whose *P*-values are less than a cut off value (0.01 or 0.05) are regarded as survival genes.
2. Classify the members of the survival gene pool using the functional annotation-clustering tool[30].
    a. Assign survival genes to several non-exclusive gene groups based on selected GO (Gene Ontology) terms closely related to the development of cancer, such as cell cycle, apoptosis, immunological response and so on. The groups are called GO-term-defined gene sets.
    b. Retain only the gene sets whose size satisfies: 50 < size < 100.
3. For each GO-term-defined gene set retained in Step 2, generate 1 million random gene sets (RGSs) each containing 30 genes from the GO-term-defined gene set.
4. Generate *m* random datasets (RDSs) from the training set, maintaining the same ratio of "good" and "bad" tumors as in the original training set.
5. Screen the GO-term-defined gene sets.

```
For each GO-term-defined gene set
    For $i$=1:m
        For $j$=1:1000000
            Calculate the P-value of the $j$th RGS as a signature for survival for the $i$th RDS
            If (P-value < 0.05) $p_{i,j} = 1$ else $p_{i,j} = 0$
        End
    End
End
```

6. For each RGS, j, calculate the fraction of RDSs for which it is predictive.

$$Fs_j = \sum_i^m p_{i,j} / m$$

7. Calculate the number of times a gene, k, is a member of an RGS that is predictive at least 90% of the time.

$$Fg_k = \sum_{j}^{1000000} \varphi(k,j)\theta(Fs_j)$$

where $$\phi(p,j) = \begin{cases} 1 & \text{if gene } k \in \text{RGS } j \\ 0 & \text{otherwise} \end{cases}$$

and $$\theta(Fs_j) = \begin{cases} 1 & \text{if } Fs_j > 0.9 \\ 0 & \text{otherwise} \end{cases}$$

8. For each GO-term-defined gene set, rank the genes in the set according to their $Fg_k$ and retain the top 30.
9. Re-run steps 5 to 8 using another 1 million RGSs distinct from the first run. For each GO-term-defined gene set, if duplicate runs yield the same top 28 genes in step 8 then the top 30 genes from the second run will be used as a gene signature.

**The testing algorithm for gene signatures.** To validate the performance of the National Research Council (NRC) gene signatures, we tested them in the training and testing datasets using the leave-one-out method. We combined the outcomes of 3 gene signatures to classify samples. For each dataset which contains n samples, the algorithm of testing for ER+ gene signatures is described below. This is a more detailed description of the algorithm shown schematically in Figure 2b of the main text.

For *i*=1:n (n is the number of samples)

1. Extract feature vectors from a given gene signature (such as NRC-1).
   a) $\mathbf{V}_{\mathbf{i,j}}$= ($g_{1,i,j}$, $g_{2,i,j}$, …, $g_{30,i,j}$) where $g_{k,i,j}$ is the expression value of the *k*th gene of NRC-j in the *i*th sample (k=1, 2, … 30, j=1, 2, 3).
   b) Calculate $\mathbf{V}_{\mathbf{i,j,g}}$ , the feature vector of shrunken class centroids extracted from the *j*th set of signature genes (NRC-j) and the relapse-free patient samples (excluding the *i*th one) and using PAMR method[31].
   c) Calculate $\mathbf{V}_{\mathbf{i,j,b}}$ , the feature vector of shrunken class centroids extracted from the *j*th set of signature genes (NRC-j) and metastasis patient samples (excluding the *i*th one).
2. Classify the *i*th sample using the *j*th signature genes.
   a) Cor($\mathbf{V}_{\mathbf{i,j,x}}$,$\mathbf{V}_{\mathbf{i,j}}$) = Pearson correlation coefficient between $\mathbf{V}_{\mathbf{i,j,x}}$ and $\mathbf{V}_{\mathbf{i,j}}$.
   b) if Cor($\mathbf{V}_{\mathbf{i,j,g}}$,$\mathbf{V}_{\mathbf{i,j}}$) >= Cor($\mathbf{V}_{\mathbf{i,j,b}}$,$\mathbf{V}_{\mathbf{i,j}}$) assign the *i*th sample to the relapse free group else assign to the metastasis group.
3. Classify the *i*th sample combining the outcomes of the three sets of the gene signatures.
   a) if Cor($\mathbf{V}_{\mathbf{i,j,g}}$,$\mathbf{V}_{\mathbf{i,j}}$) >= Cor($\mathbf{V}_{\mathbf{i,j,b}}$,$\mathbf{V}_{\mathbf{i,j}}$) (j=1, 2, 3) assign the *i*th sample to the low-risk group
   b) else if Cor($\mathbf{V}_{\mathbf{i,j,g}}$,$\mathbf{V}_{\mathbf{i,j}}$) < Cor($\mathbf{V}_{\mathbf{i,j,b}}$,$\mathbf{V}_{\mathbf{i,j}}$) (j=1, 2, 3) assign to the high-risk group
   c) else assign to the intermediate-risk group.
4. For the samples in the high-risk group determined by NRC-1, -2 and -3, we further classify them using NRC-4, NRC-5 and NRC-6 by running Steps 1-3. If a sample is assigned to the

relapse free group by NRC-4, -5 and -6, the sample will be appended to the intermediate-risk group determined by NRC-1, -2 and -3. If not, the sample is assigned to a new high-risk group.

End

Application of the algorithm to testing of the ER- gene signatures involves running through steps 1-3 using NRC-j (j = 7, 8, 9) instead of 1 to 3 Also, due to the small size of the data sets, the intermediate- and high-risk groups are combined to form the high-risk group.

## Comments on the MSS algorithm

**Subgroups of tumor samples using PGS profiles.** In breast cancer, ER status has been used as a clinical feature to classify the samples into ER+ and ER- subgroups. However, certain tumor types have no clear subgroup classification using either clinical features or molecular feature (i.e., gene expression profiles). In this situation, we propose to use the PGS (passed gene sets) profiles of the random training datasets (RDSs) to reveal features (i.e., clinical or molecular features) that could classify the samples into subgroups for MSS.

We used Chang's cohort (this set has rich clinical annotations of samples) as an example to illustrate the PGS-based method. We first generated PGS profiles for 72 RDSs of Chang's cohort (as described in the first part of Result section in the main text). Based on the PGS profiles, we could divide the RDSs into 2 groups (keeping the redundant samples in the groups): low PGS group (i.e., RDSs have less than 5-10% of the average PGSs across all RDSs) and high PGS group. We then conducted statistical tests (i.e., t-test for continuous data; Fisher's test for binary data) between the two groups using every clinical or molecular feature of the samples. We found that except for ER, other features did not show statistical trends for the differences between the two groups. For ER, the *P*-value was ~0.08, a modestly significant *P*-value. However, it suggests that ER could be a useful feature to group the samples before running MSS. We tested if ER is a good classifier. To do so, we randomly generated 6 RDSs in which all are ER+ samples, and another 6 RDSs in which all are ER- samples. We performed the survival screening on these RDSs using the $1 \times 10^5$ of RGSs (randomly generated gene sets) which were used to generate the PGS profiles for the 72 RDSs (see the main text). If one group has high PGSs and the other has low PGSs, it means that ER can be used to classify the samples into 2 subgroups, which, in turn, can be used for running MSS. As shown in the main text, ER is in fact a classifier for grouping samples for MSS.

**Determination of the appropriate number of genes for a gene signature.** The number of breast cancer signature genes ranges from several to a few hundred among the five well-known breast cancer gene signatures. Among these signatures, 70-gene signature and the 21-gene Oncotype DX are the most well-known ones. We decide to start from a larger set (70 genes) to refine the size of signature genes. Therefore, in the preliminary testing, we decided to use van 't Veer dataset[27], which was used to generate the 70-gene signature, to generate another 70-gene signature by running MSS. van 't Veer's 70-genes were selected from 231 modulated genes[27]. We used these 231 modulated genes to generate 1 million of RGSs, in which each RGS contains 70 genes. Using Wang dataset[28], we generated 72 RDSs. After running MSS, we took the top-ranked 70 genes, which have only a few genes overlapped with van 't Veer's 70-genes, to conduct survival testing in another independent dataset[32] (Miller dataset). In the tests, we systemically removed genes, one-by-one, from the bottom of the 70 gene list. Interestingly, we found that the *P*-values for the survival analyses were not changed until the gene set was reduced to the top 28 genes. Upon removing more genes from this list, the *P*-values suddenly increased and fluctuated (Supplementary Figure S4a). Based on these observations we decided to use 30 genes as the size of the NRC gene signatures. We also validated the gene size using NRC-1, -2, and -3 by systemically conducted the survival tests by removing genes, one-by-one, from the bottom of the genes in each signature. Removal of several genes from the bottom of the rank-lists did not affect the *P*-values of the survival tests in testing dataset (Supplementary Figure S4b), suggesting that 30 is a reasonable size.

**Gene group selection.** In the MSS algorithm, we suggested selecting gene groups based on GO terms associated with cancer hallmarks. Normally most of the GO terms contain 60-80 genes. We realize that there are likely a number of cancer hallmark genes that are not known yet. Hence, our reliance on GO term annotation may result in lacking some important genes in our signatures. Nevertheless, this does not detract from the predictive value of our GO-term-derived signatures. One strategy to widen our net is to augment the GO terms with other information, such as, text mining of genes based on co-occurrence with cancer hallmark genes or stem cell genes in the same sentences of the PubMed abstracts, or perhaps to include genes that physically or genetically interact with cancer hallmark genes or stem cell genes. These associated genes might then be used for running MSS.

**Number of random training datasets for MSS.** We use the NRC-1 signature to illustrate how we determined the size of random datasets. First we generated two sets (A and B) of 1 million RGSs from the cell cycle genes (see the main text) such that no any RGS is common between the two sets. Using the Wang dataset, we generated 5, 10, 15, 20, 25, 30, 35 and 40 sets of RDSs. For each set of RDSs, we ran MSS using the RGS Sets A and B and ranked the genes as described in MSS. Then we calculated gene overlapping ratio: the numbers of the common genes among the top 30 genes are divided by 30. We found that the top 29 genes are the same for the RDS sets with the sizes of 35 and 40. In order to further investigate these phenomena in details, we extend the same analyses by generating 14 sets of $RDS_i$ (i= 27 +j; j= 0, 1, 2,..13). To obtain robust results, such analyses have been conducted 100 times and the average gene overlapping ratios were calculated and plotted (Supplementary Figure S5). As shown in Supplementary Figure S5, gene overlapping ratios are stable when the RDS sets have the sizes above 30. We extended the same analysis to NRC-2 and -3. Similar results were obtained (Supplementary Figure S5). Based on these observations, we decided on the somewhat arbitrary value of 36 in our own production runs because we had 36 compute-nodes available to us to run our calculations in parallel.

**Cells, cell culture, antibodies and reagents.** BRI-JM01 cells were isolated, characterized and cultured as described[24]. Mouse mammary 4T1 tumor cells were obtained from, and cultured according to the instructions of the ATCC. Human recombinant TGF-β1 (R&D Systems) and MMP9 Inhibitor I (Calbiochem) were reconstituted according to the manufacturer’s instructions.

**Reverse Transcriptase Polymerase Chain reaction (RT-PCR)**. Cells were grown in 35 mm dishes for 24 h in the absence or presence of TGF-β1 (100 pM), and total RNA was isolated using the RNeasy mini kit (Qiagen) according to the manufacturer's instructions. First strand cDNA was synthesized in a final volume of 20 μl containing 500 ng of total RNA, 1 μL dNTPs (10 mM each) and 1 μL oligo dT (500 ng/ul). After 5 min at 65°C, samples were quickly chilled on ice. Four μL 5x first strand buffer and 2 μL DTT (0.1 M) was added. After 2 min at 42°C 1 μL Superscript II (200 units) was added followed by incubation at 42°C (50 min) and 70°C (15 min). The PCR reaction was carried out in a volume of 50 μL containing 2 μl of the first strand reaction, 1 μL (10 μM) of each of the primers specific for mouse MMP9 (5'-TGA-ATC-AGC-TGG-CTT-TTG-TG-3' and 5'-ACC-TTC-CAG-TAG-GGG-CAA-CT-3') or GAPDH (5'-ACC-ACA-GTC-CAT-GCC-ATC-AC-3' and 5'-TCC-ACC-ACC-CTG-TTG-CTG-TA-3'), 40 μL $H_2O$, 5 μL 10x buffer, 0.5 μL 10mM dNTP mix and 0.5 μL Taq Polymerase (5 U/ul). PCR reactions were carried out using the following conditions: 2 min at 95°C followed by 25 cycles with an annealing temperature of 55°C. RT-PCR products were evaluated on a 2% agarose gel.

**Western blot**. Cells grown in 35 mm dishes were treated for 24 h with or without TGF-β1 (100 pM). Conditioned medium was collected and 30 μL conditioned medium was resolved by SDS-PAGE (10%) under reducing conditions. Proteins were transferred to nitrocellulose, membranes were incubated with αMMP9 antibodies (1/5000, Cedarlane), and immunoreactive bands were visualized by chemiluminescence (Perkin-Elmer).

**Morphology assay**. BRI-JM01 and 4T1 cells were seeded in 12-well dishes and grown to 70% confluency. Cells were then treated for 24 h with 100 pM TGF-β1 in the absence or presence of 10 μM MMP9 Inhibitor I. Monolayers were washed with PBS and then fixed and stained with 0.2% crystal violet in anhydrous ethanol for 5 minutes at room temperature. Excess staining fluid was removed and wells were rinsed with tap water. Finally, images were captured using a Nikon CoolPix 995 digital camera mounted on a Leitz Labovert inverted microscope.

**Transwell invasion assay.** 24-well Biocoat Matrigel invasion chambers (8 μm; BD Biosciences) were used according to the manufacturer's instructions. Briefly, top chambers were seeded with $5 \times 10^4$ viable BRI-JM01 or 4T1 cells in cells specific culture medium containing 0.2% FBS, bottom chambers were filled with the same culture medium containing 10% Fetal Bovine Serum (FBS). TGF-β1 (100 pM) +/- MMP9 Inhibitor I (10 μM) was added to both compartments. After 24 h, non-invasive cells remaining in and on the Matrigel-coated membrane were removed with a cotton swab. Cells that migrated to the other side of the membrane were fixed and stained with 0.2% crystal violet in anhydrous ethanol. Migratory cells in four random fields of each membrane were counted using a light microscope at 200x magnification.

## Supplementary References

24. Lenferink,A.E., Magoon,J., Cantin,C., & O'Connor-McCourt,M.D. Investigation of three new mouse mammary tumor cell lines as models for transforming growth factor (TGF)-beta and Neu pathway signaling studies: identification of a novel model for TGF-beta-induced epithelial-to-mesenchymal transition. *Breast Cancer Res* **6**, R514-R530 (2004).

25. Lenferink,A.E. *et al.* Transcriptome profiling of a TGF-beta-induced epithelial-to-mesenchymal transition reveals extracellular clusterin as a target for therapeutic antibodies. *Oncogene* **29**, 831-844 (2010).

26. Cui,Q. *et al.* A map of human cancer signaling. *Mol. Syst. Biol.* **3**, 152 (2007).

27. van 't Veer,L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

28. Wang,Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679 (2005).

29. Tester,A.M., Ruangpanit,N., Anderson,R.L., & Thompson,E.W. MMP-9 secretion and MMP-2 activation distinguish invasive and metastatic sublines of a mouse mammary carcinoma system showing epithelial-mesenchymal transition traits. *Clin Exp Metastasis* **18**, 553-560 (2000).

30. Dennis,G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).

31. Tibshirani,R., Hastie,T., Narasimhan,B., & Chu,G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A* **99**, 6567-6572 (2002).

32. 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. U. S. A* **102**, 13550-13555 (2005).