Supporting Information

Chen et al. 10.1073/pnas.1701512114

SI Materials and Methods

Transcriptomic Analysis of SB Mammary Tumors. Total RNA of SB mammary tumors was extracted using TRIzol (Invitrogen) following the manufacturer's instructions. For each sample, 100 ng of total RNA was reverse transcribed to produce cDNA, which was subsequently used as a template to create biotin-labeled amplified RNA (aRNA). The aRNA then was fragmented and hybridized to Affymetrix Mouse Genome 230 2.0 Arrays (Affymetrix) for 16 h at 45 °C using GeneChip Hybridization oven with rotation at 60 rpm. Arrays then were washed and stained using the FS450_0001 fluidics protocol and scanned using an Affymetrix 3000 7G̅ scanner.

Identification of Transposon Insertion Sites. The cloning and mapping of transposon insertion sites was done via splinkerette PCR using the 454 GS-Titanium Sequencer (Roche Applied Science) and the NCBI mouse genome assembly m37, as described previously (1). The GKC method was used for CIS determination using multiple kernel scales (widths of 30, 50, 75, 120, and 240 Kb), as described previously (1–3).

Data Preprocessing of Mouse Mammary Tumor and Human Breast Carcinoma Panels. Mouse mammary tumors on the Affymetrix Mouse 430A_2/Affymetrix Mouse 430_2 platform and human breast carcinomas and cell line data on the Affymetrix U133A/U133plus2 platform were downloaded from Array Express and Gene Expression Omnibus website (GEO) (https://www.ncbi.nlm.nih.gov/geo). Robust Multichip Average (RMA) normalization was performed on each dataset, and the normalized data were subsequently combined for mouse mammary tumor and human breast carcinoma panels. The panels then were standardized separately using ComBat software (4) to remove the batch effect. The panel of mouse mammary tumors ($n = 394$) included a subset of our SB mammary tumors ($n = 35$), E-TABM-684 ($n = 12$) (5), E-TABM-997 ($n = 9$) (5), GSE10193 ($n = 7$) (6), GSE13221 ($n = 6$) (7), GSE13230 ($n = 7$) (7), GSE14226 ($n = 12$) (8), GSE14753 ($n = 3$) (9), GSE15119 ($n = 13$) (10), GSE15904 ($n = 126$) (11), GSE20406 ($n = 75$) (12), GSE20465 ($n = 25$) (13), GSE22150 ($n = 8$) (14), GSE6246 ($n = 3$) (15), GSE6596 ($n = 3$) (16), GSE9355 ($n = 46$) (17), and GSE9447 ($n = 4$) (18). The panel of human breast carcinomas ($n = 1{,}345$) included GSE12276 ($n = 204$), GSE19615 ($n = 115$), GSE21653 ($n = 266$), GSE23177 ($n = 116$), GSE23593 ($n = 50$), GSE26639 ($n = 226$), GSE3744 ($n = 47$), GSE5460 ($n = 127$), GSE5764 ($n = 10$), GSE6532 ($n = 87$), GSE5764 ($n = 20$), and GSE9195 ($n = 77$). The related detailed information is provided in Dataset S6.

Subtypes of Mouse Mammary Tumors and Human BCs. For subtype identification of mouse mammary tumors, an SD of 1.35 was applied to filter out less variable genes from the mouse mammary tumor panel. The remaining 1,028 Mouse 430A_2 probes, corresponding to 835 genes, were used for subtype identification. The consensus clustering method then was applied on the mouse mammary tumor panel to identify the four major subtypes (Euclidean distance, $k = 7$). ABC subtype signature (3) and single-sample gene set enrichment analysis (ssGSEA) (19) were used to predict the subtypes of the human BC panel samples.

BC Gene-Expression Data Collection and Characteristics. We downloaded the expression values of 11 U133A BC "cohort collections" (hereafter denoted simply as cohorts) from the Gene Expression Omnibus (GEO) website (https://www.ncbi.nlm.nih.gov/geo). Each cohort consists of either a single microarray dataset or several small datasets merged together (batches). In total we collect 11 cohorts including 23 batches (as seen in Table S1). Each batch includes the raw perfect match/mismatch (PM/MM) data of 22,215 transcripts (represented by ~18,000 genes).

The number of patients with primary BC tumors in each cohort varies from 64 to 508; the smallest batch belongs to GSE12276 (the batch ID number is 18) with only two patient samples. For each patient k, $k = 1, \ldots, K$, the following clinical characteristics are recorded: survival time t in years (time to relapse), event e (0 for nonrelapse at time t; 1 for relapse at time t), cancer subtype (ERBB2, luminal A, luminal B, normal-like, basal-like, no subtype, HER2, and claudin-low), Elston tumor grade (1–3), ER status (ER^+ and ER^-), PGR status (PGR^+ and PGR^-), HER status (positive and negative), nymph node status (positive and negative), p53 mutation status (mutation or wild type), age at diagnosis (ranging from 24 to 88 y), and tumor size in mm (ranging from 0 to 8.2). Table S1 presents the summary characteristics of the 11 cohorts (the rest of the variables are not recorded in more than half of the cohorts). Notice that the three datasets GSE12276, GSE6532, and GSE7390 consist of five merged batches, and the summary characteristics are estimated across all the five batches.

Normalization, Data Integration, and Batch-Effect Correction. Each batch is independently background corrected, normalized, and summarized by the RMA (20). This task is performed using the *Affy* R package. The normalized data from all 23 different batches (in total 2,333 samples) are merged, and ComBat (4) corrects for batch effects. The term "batch effect" refers to the unwanted nonbiological variation observed across the multiple batches caused by data processing by different technicians, at different sites, processing times, and protocol variations. In our work all these factors can possibly be significant, and ComBat detects and removes them. This removal is a major analysis step to keep only biologically meaningful data variation in the search for clinical subgroups specified by the molecular pathways networks and the molecular features that link patient subpopulations to treatments clinical biomarkers. ComBat is implemented via the R package sva. The performance of the algorithm and the properties of the batch-corrected data are visualized by quality control analysis via the R package arrayqualitymetrics. Package arrayqualitymetrics performs a series of tests, summary measures, and plots for visualization of outlier samples using only the expression data. Obvious outliers (samples failing many tests) should be removed from further analysis, but one should be very cautious (conservative) in removing a sample. In this way we generate our full dataset consisting of the normalized and batch effect-corrected expression levels of 22,215 Affymetrix U133A transcripts, measured for 2,333 patient samples (no outliers were found by arrayqualitymetrics). Following the experimental design described in the main text, we keep only the expression data of the 126 genes of interest.

Hierarchical Clustering Microarray Data and Identification of the Mouse and Human Tumor Subtypes. For hierarchical clustering of the mouse ($n = 394$, Affymetrix Mouse430A) and human ($n = 1{,}345$, Affymetrix, U133P2) BC transcriptional profiles, the samples are grouped according to their tumor subtypes. For human BCs, the samples were arranged in the following subtype order: basal-like ($n = 229$), claudin-low ($n = 40$), ERBB2 ($n = 197$), luminal-A ($n = 420$), luminal-B ($n = 400$), and normal-like ($n = 59$). For mouse BCs, the samples were arranged in the

following subtype order: mesenchymal (n = 105), Neu (n = 89), ductal (n = 121), glandular (n = 69), and unclassified (n = 10). Only probe sets corresponding to the 126 hBCSG genes were selected for both the mouse and human expression dataset. Because each gene of interest may be represented by more than one Affymetrix probe set, we selected a representative probe set for each gene. The probe set with the highest median expression across all samples was selected as the representative probe set for each gene of interest. Before unsupervised clustering, the expression matrix was preprocessed via median centering and normalization for both rows and columns. Unsupervised hierarchical clustering was used to cluster the genes (via the Kendall's tau similarity metric and average linkage method) while maintaining the order of the columns. To address the sample imbalance between the mouse (n = 394) and human (n = 1,345) expression datasets, we performed random sampling of the human BC datasets so that the sample numbers could match those from the mouse BC datasets. Previous observations from the clustered heatmap involving 394 mouse BC datasets and 1,345 human BC datasets revealed similar heatmap patterns between (*i*) the human basal-like and claudin-low subtypes and the mouse mesenchymal subtype, (*ii*) the human *ERBB2* subtype and the mouse *Neu* subtype, (*iii*) the human luminal-A subtype and the mouse ductal subtype, and (*iv*) the human luminal-B and normal-like subtypes and the mouse glandular subtype. Therefore, we sampled 105 human BC samples from the basal-like and claudin-low subtypes, 89 samples from the *ERBB2* subtype, 121 samples from the luminal-A subtype, and 69 samples from the luminal-B and normal-like subtypes. As previously described, the expression matrix was preprocessed, and unsupervised hierarchical clustering was used to cluster the genes (via Kendall's tau similarity metric and average linkage method) while maintaining the order of the columns. Five independent random sampling procedures were performed to assess the stability of the heat map clusters. Clustering was implemented via Gene Cluster 3.0 and visualized in Java Tree View.

Data for Identification of Survival-Significant Genes. We downloaded the expression values of 11 U133A BC cohort collections from the GEO website. Each cohort may consist of a single microarray dataset or several small datasets merged together (batches). In total, the 11 cohorts included 23 batches. Each batch included the raw PM/MM data of 22,215 transcripts (represented by about 18,000 genes). Each batch was independently background corrected, normalized, and summarized by the RMA (20). Detailed information about the identification of the survival genes is provided in Dataset S7. Available clinical information, including Elston tumor grades; ER, PR, HER2, and lymph node status; and tumor size (≥2 cm; <2 cm) also was collected for some patients. DFS data (time, event) for all patients were collected for all 2,333 patient tumor samples (Dataset S6). We also analyzed the normalized data from the Agilent mRNA expression microarrays and the corresponding OS data for 226 BC patients in the TCGA who received systemic therapy (hormone therapy, chemotherapy, and combine therapy). The BC samples have been histologically classified as invasive ductal carcinoma (IDC). Expression and clinical datasets are available in Dataset S9*A*.

1D-DDg Method. The 1D-DDg method aims to identify an optimal gene-expression cutoff value that most significantly stratifies the patient cohort into two survival significant subgroups (21, 22). 1D-DDg is a univariate analysis performed for a single continuous random variable (e.g., for a gene expression or microarray probeset signal value). First, patients are ranked in ascending order based on the quantity of this continuous variable. Using a cutoff value for a prognostic variable (e.g., gene expression), patients are then stratified into two subgroups represented by two Kaplan–Meier survival curves. The prognostic significance is quantified via the log-rank and/or Wald statistics tests, bootstrap-defined confidence intervals, and FDR estimation. The minimum size of the smallest subgroup is controlled by the 1D-DDg, which reduces the imbalance in subgroup size and stabilizes the prediction outcome.

2D-DDg Prognostic Method. The 2D-DDg method is an extension of the 1D-DDg method for the case of variable pairs (21, 22). The patients are represented as points on a 2D plane where the two axes are for the two variables of interest. The patients are stratified via identification of two prognostic cutoff variables (e.g., gene-expression values) on the axes, one for each variable. Two orthogonal lines, including these cutoff values as the points, split the plane into five possible distinct subdomains. Optimal cutoff values are selected that best stratify the patients into two statistically distinct prognostic subgroups. The statistical significance is quantified via the log-rank and/or Wald statistics test and FDR estimation.

SWVg Method. SWVg is an automatic method of prognostic feature selection and disease risk prediction that allows the construction of an optimized, multivariable, prognostic classifier (22, 23). To construct a multivariable, prognostic signature, the SWVg selects a set of the most statistically significant prognostic variables and optimizes the list of selected prognostic variables to stratify the BC patients further into two, three, or more risk groups using the binarized patient risk-class data that separated the patients into two (or more) subgroups according to their risk of disease development. Such input data could be provided by 1D-DDg and/or 2D-DDg and include weighted variables that reflect the relative importance and significance of each prognostic variable with respect to the others. This information is used to construct a decision rule and to assign a patient to one of the risk subgroups. Also, SWVg optimizes the number of prognostic variables via minimization of the −log function of the log-rank statistics P values, where the paired K-M functions are compared.

Functional Analysis of Gene Lists. PANTHER (www.pantherdb.org) and DAVID tools (https://david.ncifcrf.gov/) were used to identify gene functional annotation terms that are significantly enriched in particular gene lists with all human genes as the background. A list of NCBI Entrez gene IDs was generated for each dataset and was used as the input into PANTHER and DAVID software. DAVID software calculates a modified Fisher's exact P value to demonstrate GO terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) molecular pathway enrichment, where P values less than 0.05 after Benjamini multiple test correction are considered to be strongly enriched in the annotation category. PANTHER calculates P values to demonstrate GO enrichment and PANTHER Pathways, where P values less than 0.05 after Bonferroni multiple test correction are considered to be strongly enriched in the annotation category.

MetaCore Significant Pathways. Protein interactions network were constructed according to protein–protein interactions available in the GeneGo MetaCore platform (https://portal.genego.com/) Their Refseq gene symbols are used in MetaCore to generate the most significant pathways (at FDR 5%) in which these genes are involved. Based on this analysis, we classify the genes into different biological processes.

1. Mann KM, et al.; Australian Pancreatic Cancer Genome Initiative (2012) Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proc Natl Acad Sci USA* 109(16):5934–5941.
2. de Ridder J, Uren A, Kool J, Reinders M, Wessels L (2006) Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLOS Comput Biol* 2(12):e166.
3. March HN, et al. (2011) Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat Genet* 43(12):1202–1209.
4. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.
5. Zvelebil M, et al. (2013) Embryonic mammary signature subsets are activated in Brca1-/- and basal-like breast cancers. *Breast Cancer Res* 15(2):R25.
6. Flowers M, et al. (2010) Pilot study on the effects of dietary conjugated linoleic acid on tumorigenesis and gene expression in PyMT transgenic mice. *Carcinogenesis* 31(9): 1642–1649.
7. Lukes L, Crawford NP, Walker R, Hunter KW (2009) The origins of breast cancer prognostic gene expression profiles. *Cancer Res* 69(1):310–318.
8. Wertheim GB, et al. (2009) The Snf1-related kinase, Hunk, is essential for mammary tumor metastasis. *Proc Natl Acad Sci USA* 106(37):15855–15860.
9. Kuraguchi M, Ohene-Baah NY, Sonkin D, Bronson RT, Kucherlapati R (2009) Genetic mechanisms in Apc-mediated mammary tumorigenesis. *PLoS Genet* 5(2):e1000367.
10. Eilon T, Barash I (2011) Forced activation of Stat5 subjects mammary epithelial cells to DNA damage and preferential induction of the cellular response mechanism during proliferation. *J Cell Physiol* 226(3):616–626.
11. Andrechek ER, et al. (2009) Genetic heterogeneity of Myc-induced mammary tumors reflecting diverse phenotypes including metastatic potential. *Proc Natl Acad Sci USA* 106(38):16387–16392.
12. Rosa-Rosa JM, et al. (2010) Deep sequencing of target linkage assay-identified regions in familial breast cancer: Methods, analysis pipeline and troubleshooting. *PLoS One* 5(4):e9976.
13. Schoenherr RM, et al. (2011) Proteome and transcriptome profiles of a Her2/Neu-driven mouse model of breast cancer. *Proteomics Clin Appl* 5(3-4):179–188.
14. Hebbard L, et al. (2011) Control of mammary tumor differentiation by SKI-606 (bosutinib). *Oncogene* 30(3):301–312.
15. Klein A, et al. (2005) Gene expression profiling: Cell cycle deregulation and aneuploidy do not cause breast cancer formation in WAP-SVT/t transgenic animals. *J Mol Med (Berl)* 83(5):362–376.
16. Klein A, et al. (2007) Comparison of gene expression data from human and mouse breast cancers: Identification of a conserved breast tumor gene set. *Int J Cancer* 121(3):683–688.
17. Li Z, et al. (2007) ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. *Cancer Cell* 12(6):542–558.
18. Hedlund M, Ng E, Varki A, Varki NM (2008) alpha 2-6-Linked sialic acids on N-glycans modulate carcinoma differentiation in vivo. *Cancer Res* 68(2):388–394.
19. Verhaak RG, et al.; Cancer Genome Atlas Research Network (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17(1):98–110.
20. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193.
21. Motakis E, Ivshina AV, Kuznetsov VA (2009) Data-driven approach to predict survival of cancer patients: Estimation of microarray genes' prediction significance by Cox proportional hazard regression model. *IEEE Eng Med Biol Mag* 28(4):58–66.
22. Motakis E, Kuznetsov VA (2009) Genome-scale identification of survival significant genes and gene pairs. Wcecs 2009: *World Congress on Engineering and Computer Science*, Vols I and Ii, pp 41–46.
23. Tang Z, Ow GS, Thiery JP, Ivshina AV, Kuznetsov VA (2014) Meta-analysis of transcriptome reveals let-7b as an unfavorable prognostic biomarker and predicts molecular and clinical subclasses in high-grade serous ovarian carcinoma. *Int J Cancer* 134(2):306–318.

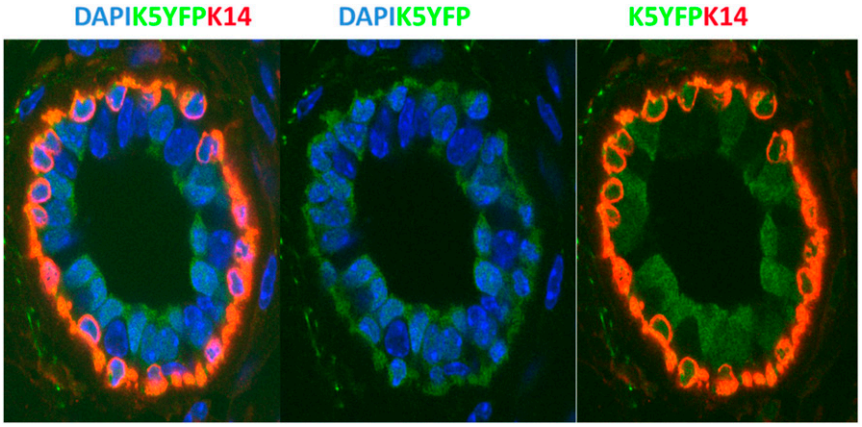


Fig. S1. Early activation of K5-Cre in mammary epithelial progenitor cells. Most cells originate from K5/K14 progenitors, as lineage traced by K5cre × floxed RosaYFP. Shown is a traverse section of a gland from a 4.5 week-old mouse. DAPI staining is indicated by blue (nuclei). YFP, driven by the K5 promoter, is indicated by green; K14-immunolabeling is red.

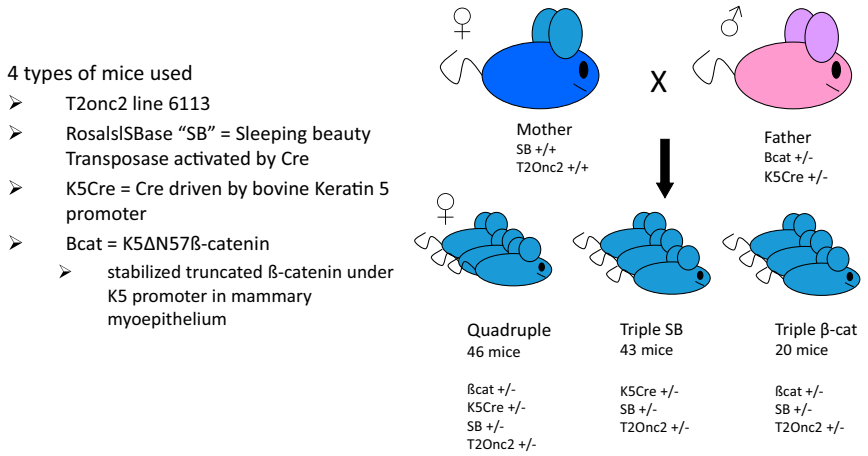


Fig. S2. The crossing scheme for generating triple and quadruple transgenic mice. The four types of mice used areT2onc2 line 6113; RosalslSBase (SBtransposase inserted in the Rosa locus); K5-Cre (Cre driven by bovine keratin 5 promoter); and Bcat (K5ΔN57β-catenin; ΔN57β-catenin is a stabilized truncated β-catenin).

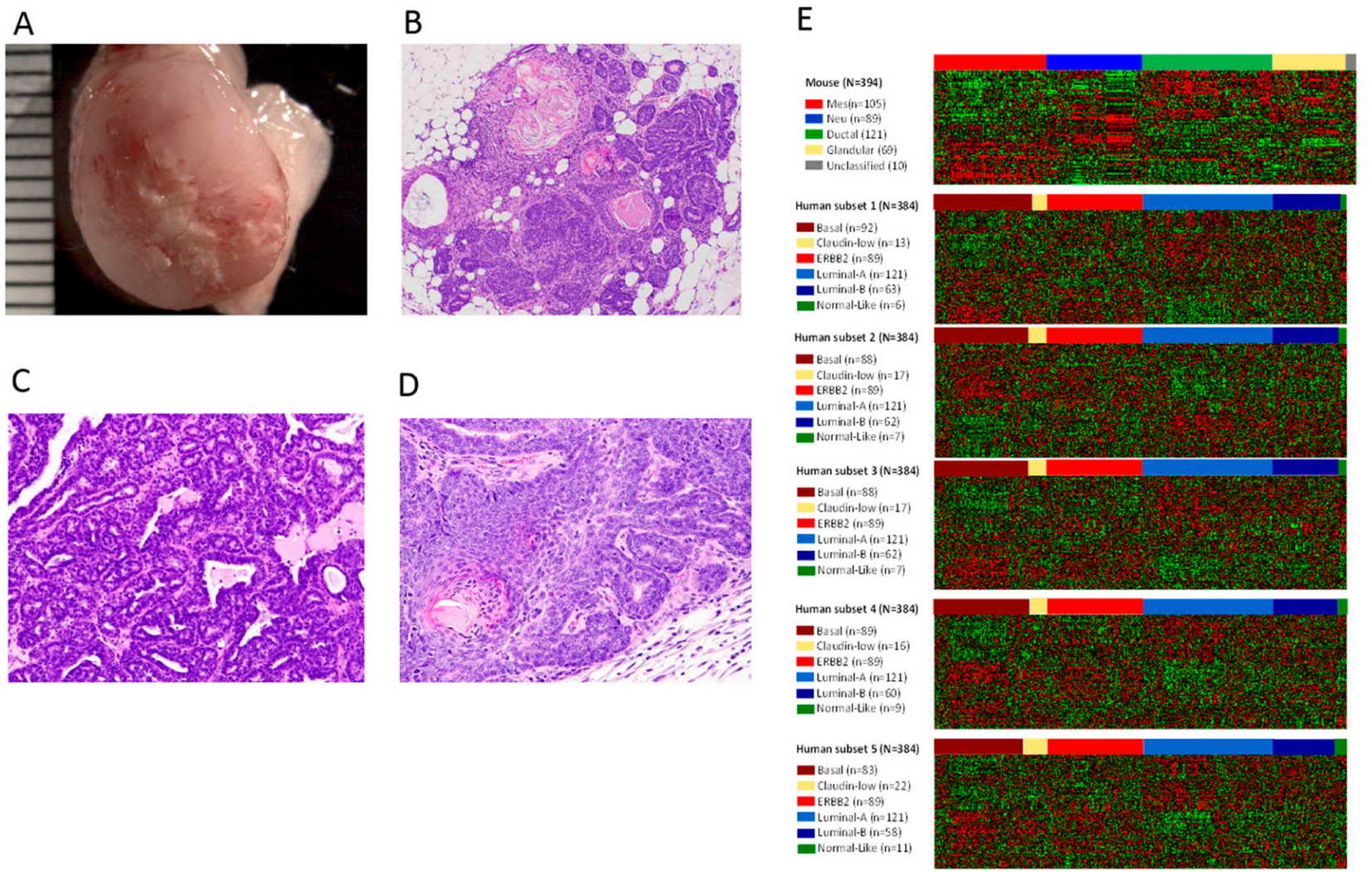


Fig. S3. Histopathology of mammary tumors. (*A*) Image of a whole mammary tumor induced by SB and N-terminally stabilized truncated β-catenin. (*B*) Squamous metaplasia. (*C*) Adenocarcinoma. (*D*) Adenosquamous carcinoma. (*E*) BC subtype classification using BCSG is comparable between mouse mammary tumors and human BC. Heatmaps of mRNA expression from mouse mammary tumor and human breast tumor (split into 5 subsets for equal-sized comparisons).

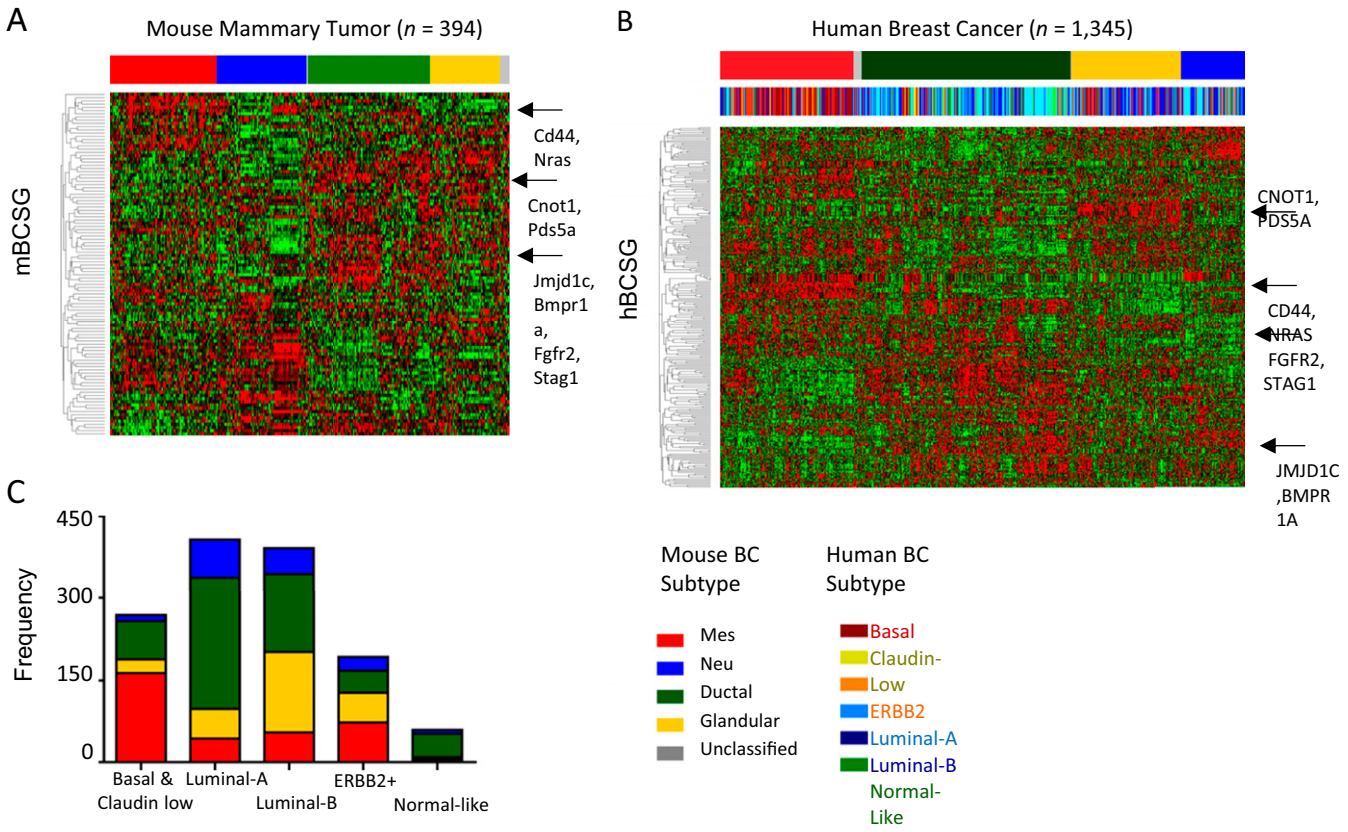


Fig. S4. BCSGs are differentially expressed in mouse and human BC subtypes. (*A* and *B*) Heatmaps of unsupervised hierarchical clustering in mouse (*A*) and human (*B*) gene expression datasets of mBCSGs and hBCSGs, respectively. Red indicates high expression, and green indicates low expression. Color bars above the heat maps show the mouse mammary tumor subtypes and the gene-associated distribution of human molecular subtypes, respectively. Selected marker genes representing tumor subtypes are labeled. (*C*) Bar plot showing the frequency of corresponding mouse and human BC subtypes.

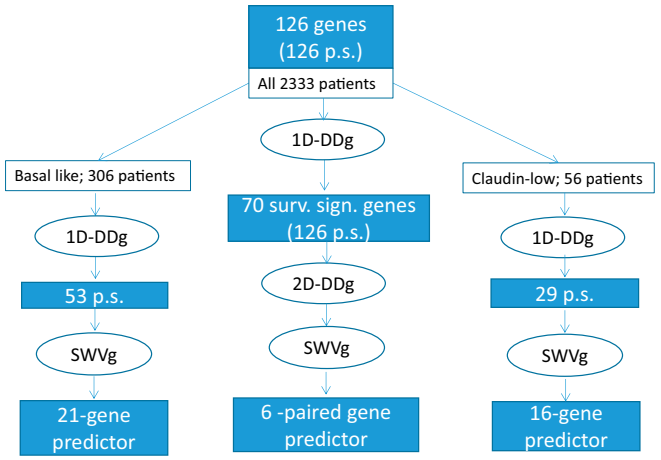


Fig. S5. Workflow for survival prediction and risk stratification analysis. SWVg is the end-point method for this workflow. To construct the six-gene-pair prognostic BC signature, the method selected a set of the 12 most statistically significant prognostic variables and optimized the selected prognostic variable list in the six prognostic variables (six-gene-pair binary variables) to stratify the BC patients further into two or three risk groups. The 12 genes were used as the input dataset for the identification of six gene pairs in 2D-DDg prognostic classification of the patients, and the 2D-DDg results were used consequently as the input dataset for the SWVg analysis.

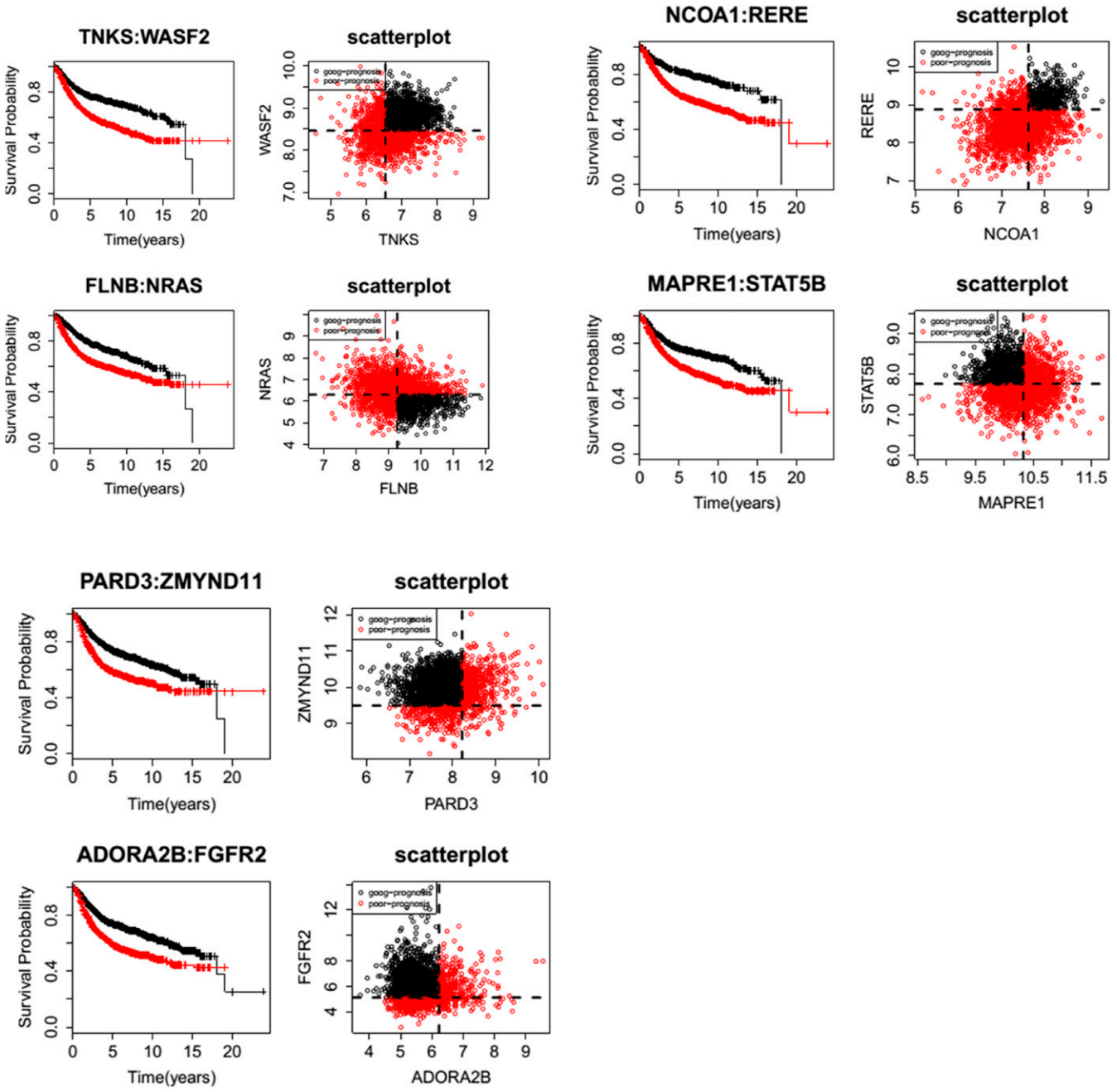


Fig. S6. The prognostic ability of each pair from the six-gene-pair BCSG prognostic signature.

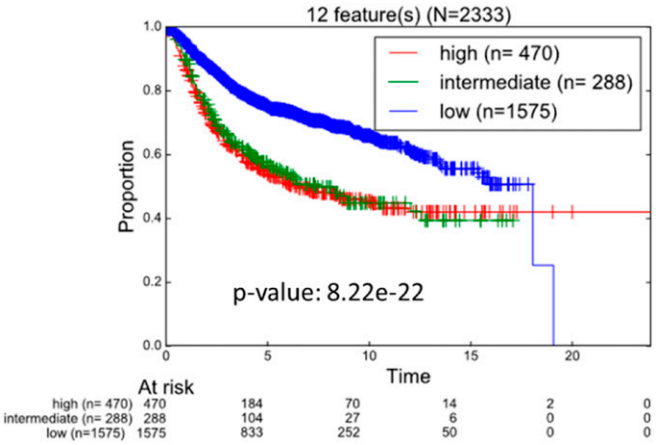


Fig. S7. The 12 genes of the six-gene-pair BCSG prognostic classifier, used 1D-DDg results as input data for SWVg stratification, can stratify 2,333 BC patients into three risk subgroups, however low- and intermediate-risk subgroups are not statistically differentiated. These 12 genes were used as the input dataset for the identification of six gene pairs in 2D-DDg prognostic classification of the patients. The 2D-DDg results then were used as the input dataset for SWVg analysis, which provides high confidence discrimination of the patients onto three risk subgroups (Fig. 5*A*).

SWV analysis:	
Dataset	p-value
Shuffled 1	2.22E-11
Shuffled 2	9.70E-18
Shuffled 3	2.64E-15
Shuffled 4	3.83E-17
Shuffled 5	8.91E-08

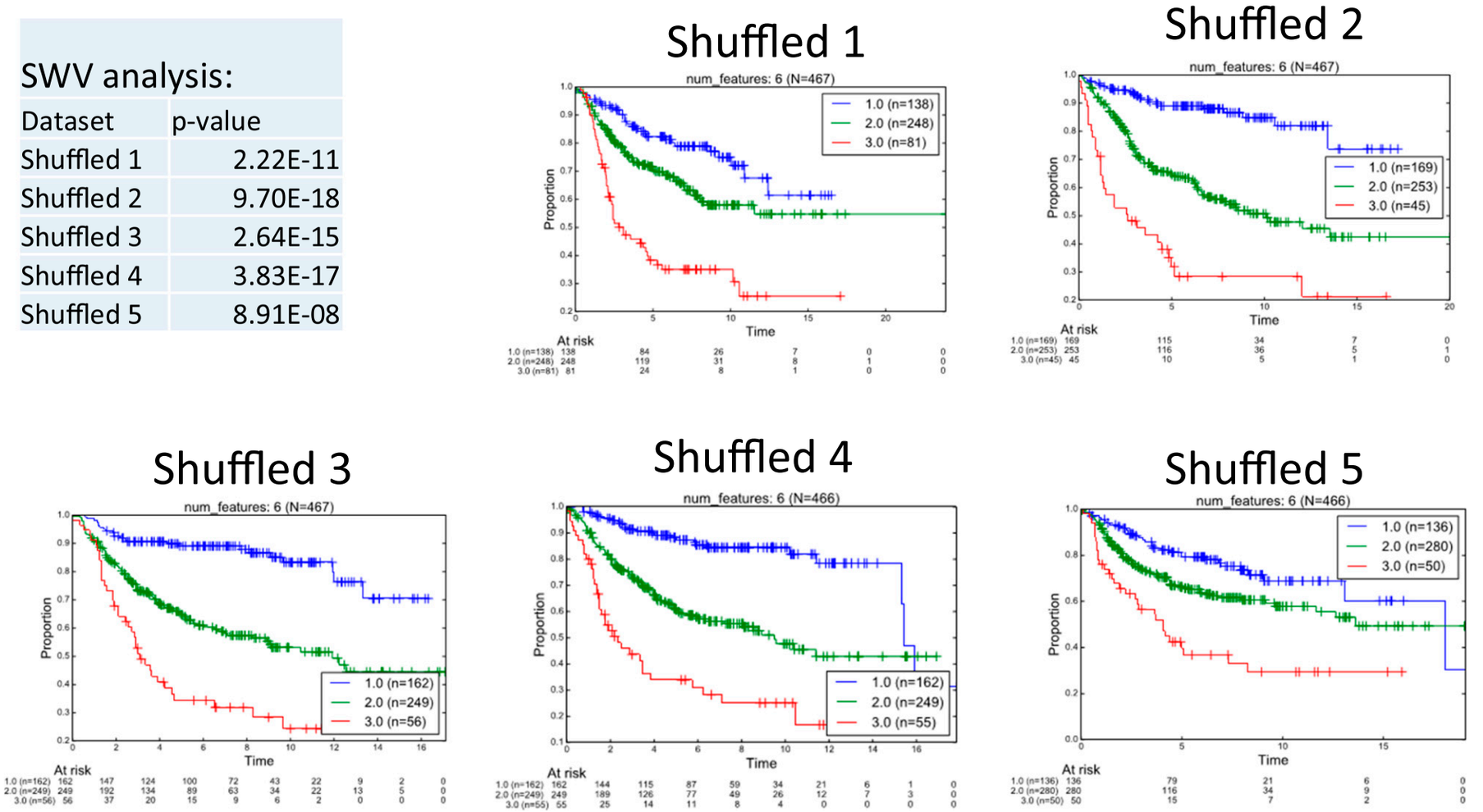


Fig. S8. The reproducible and robust prognostic ability of the six-gene-pair BCSG prognostic signature.

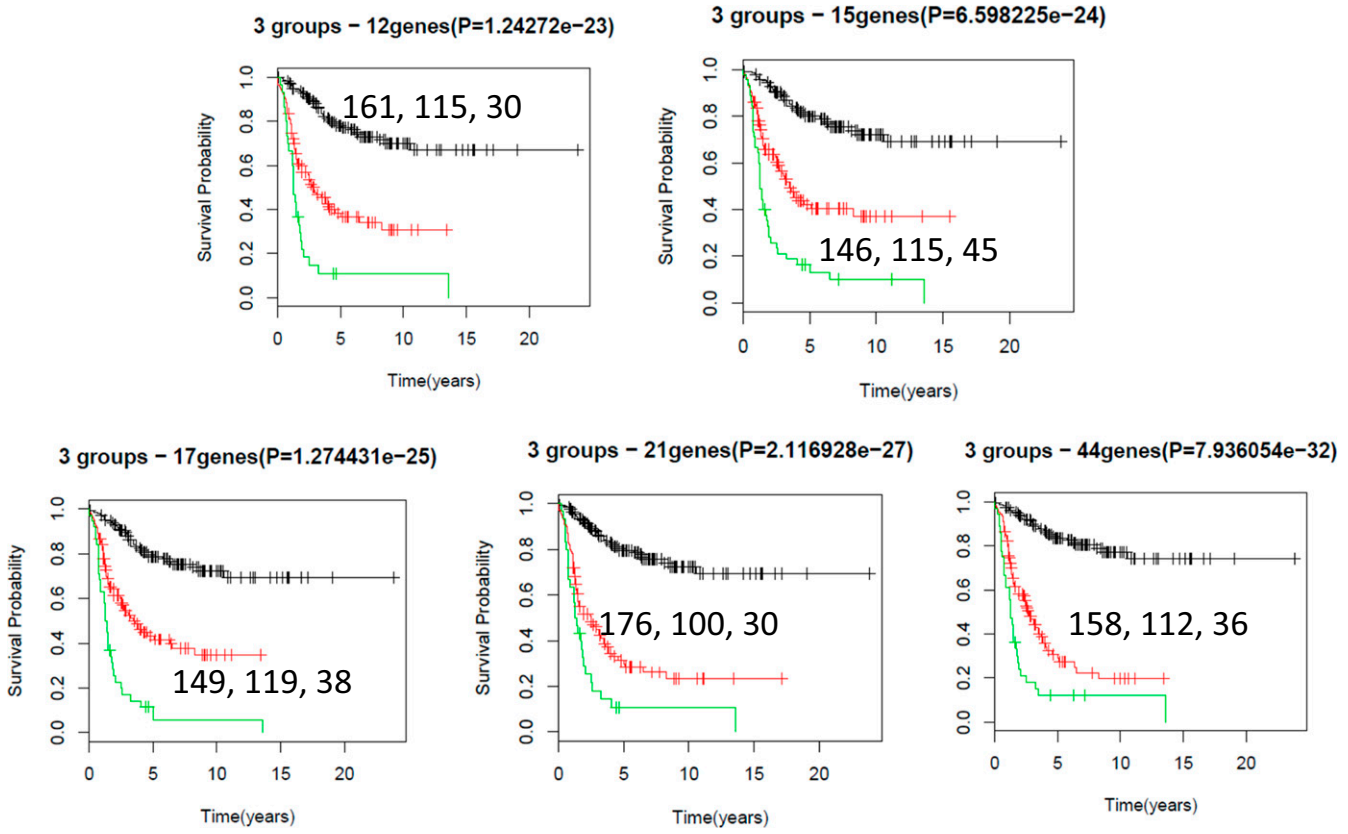


Fig. S9. Optimization of the number of genes required for the SWVg prognosis model for patients classified as having the basal-like BC subtype.

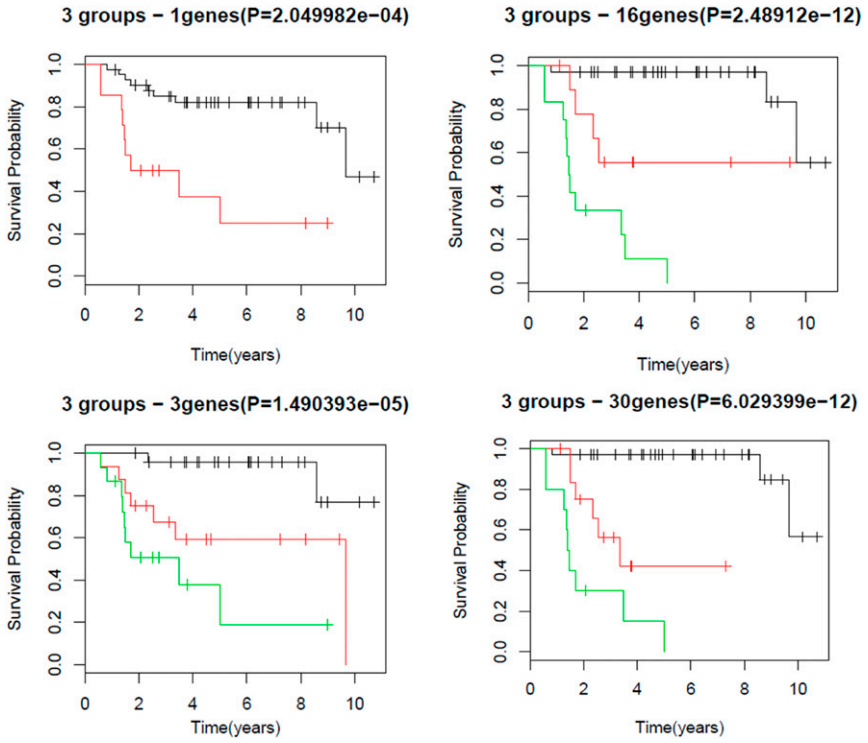


Fig. S10. Optimization of the number of genes required for the SWVg prognosis model for patients classified as the claudin-low BC subtype.

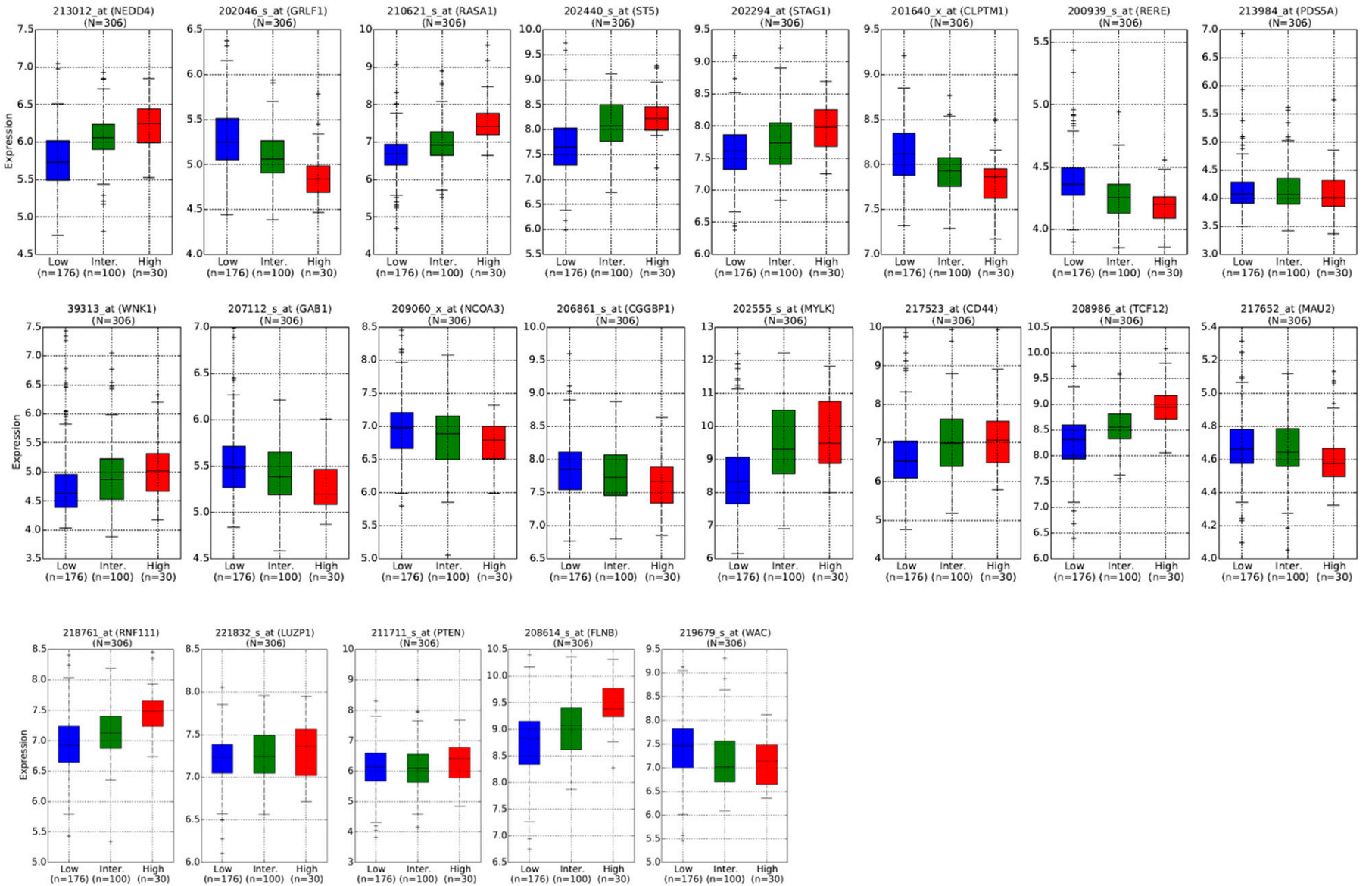


Fig. S11. Individual gene-expression variations in the low-, intermediate-, and high-risk subgroups defined by the 21-gene basal-like tumor subtype classifier.

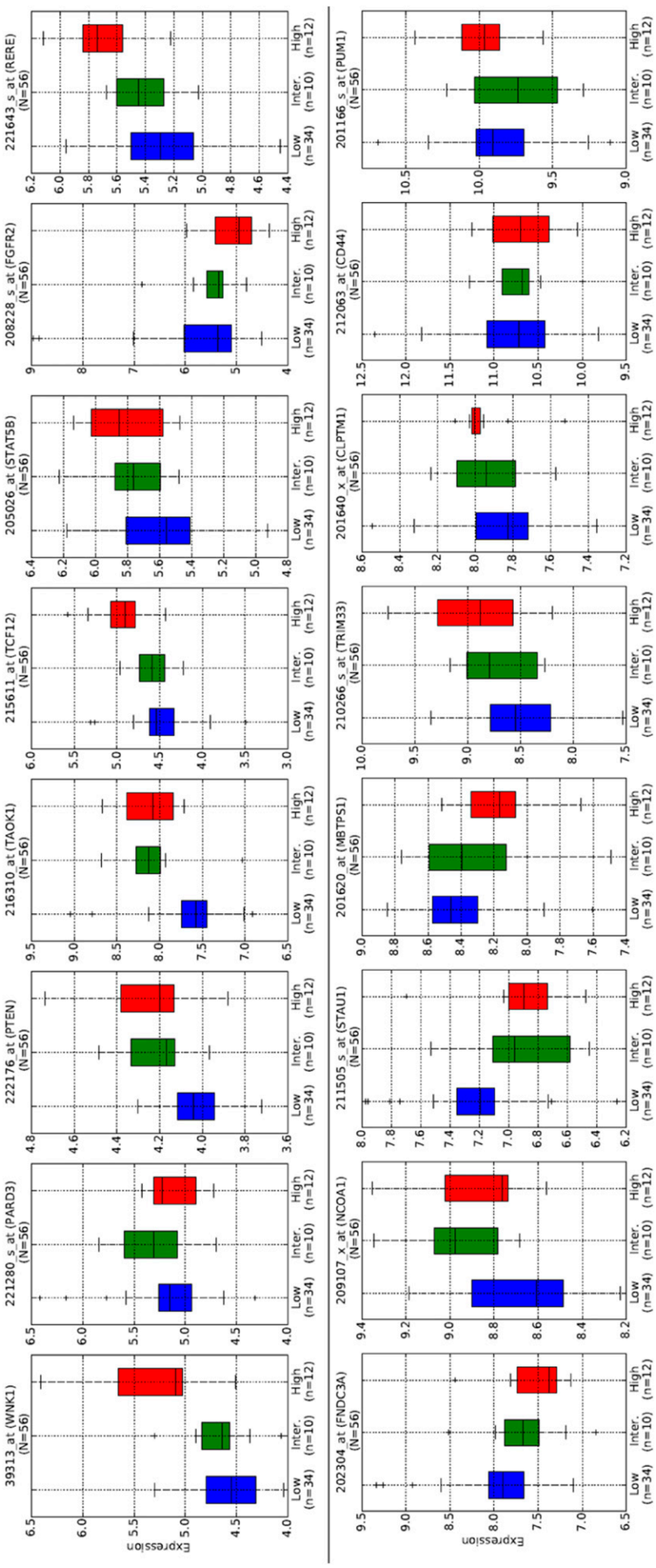


Fig. S12. Individual gene-expression variations in the low-, intermediate-, and high-risk subgroups defined by the 16-gene claudin-low BC subtype classifier.

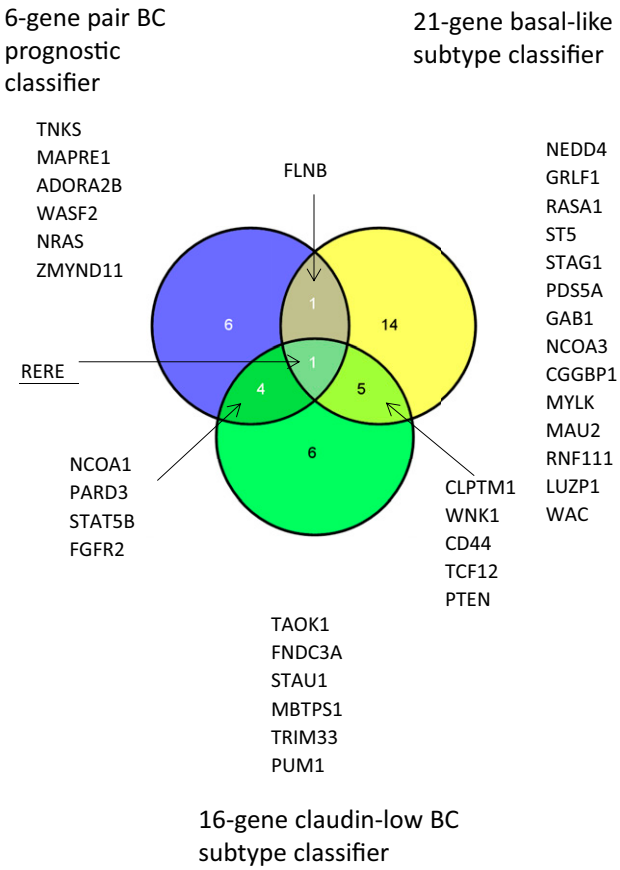


Fig. S13. Venn diagram of the 12 genes of the six-gene-pair BCSG prognostic signature, 21-gene basal-like BC subtype risk stratification signature, and 16-gene claudin-low BC subtype risk stratification signature.

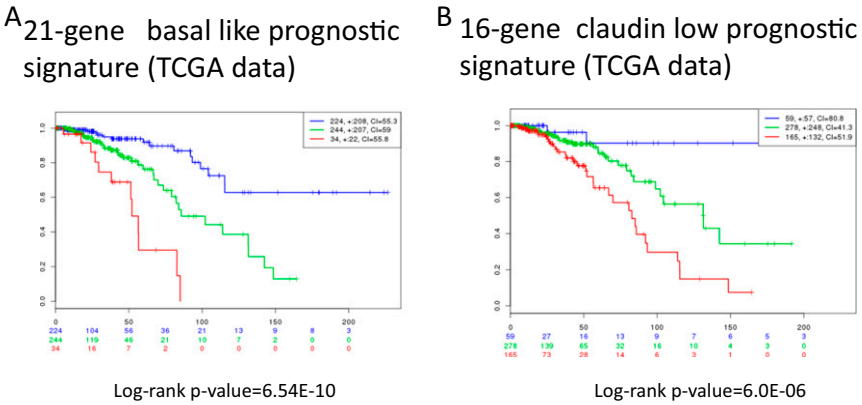
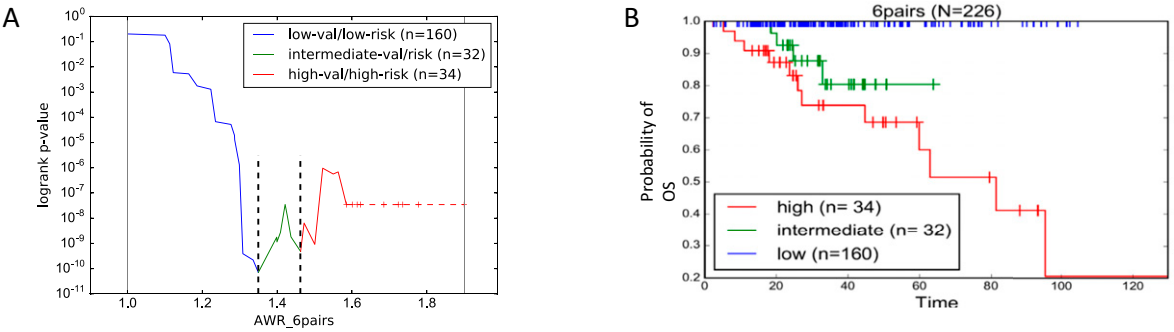


Fig. S14. Allocation of 502 BC patients from the TCGA database into three risk subgroups using SurvXpress software. (*A*) The 21-gene basal-like subtype risk stratification classifier genes stratify basal-like BC patients of the TCGA database into three risk groups. (*B*) The 16-gene claudin-low subtype risk stratification classifier genes can stratify claudin-low BC patients of TCGA database into three risk subgroups.



Visual presentation (1D-DDg –like) of SWVg analysis result.
On axis X: gene expression average (AWR) for 6 gene pairs classifier. Three patient subgroups were separated by two cut-off values.

1-st (low) expression cut-off value: 1.3497 (inclusive of right)

2-nd (intermediate expression cut-off value: 1.463 (inclusive of right)

high: >1.463

Results (K-M plot) of survival prediction based on SWVg prediction.

3 risk groups: multivariate p-value
7.25E-11

pairwise p-value

	high	intermediate	low
high	1	0.274775	1.72E-12
intermediate	0.274775	1	5.84E-06
low	1.72E-12	5.84E-06	1

Fig. S15. The SWVg-derived six-gene-pair prognostic classifier separates 226 TCGA BC patients receiving postsurgery systemic therapy (hormone therapy, chemotherapy, or combined therapy) into three risk subgroups [low risk (n = 160), intermediate risk (n = 32), and high risk (n = 34)]. (*A*) Visual presentation of two cutoff values, based on the –log(P value) function of the 1D-DDg–like SWVg results. The first cutoff value (left vertical direct line) separates the low-risk patient's subgroup (on the left) from two other patients. The second cutoff value (right vertical direct line) separates the intermediate-risk patient's subgroup from the high-risk patient's subgroup. (*B*) Results of the survival prediction subgrouping of the patients based on the specified SWVg analysis. Low-risk < 1.3497 < intermediate-risk subgroup < 1.463 < high-risk subgroups (see panel *A*).

Table S1. Clinical characteristics of the breast cancer patients in the 11 cohorts (23 batches)

GEO label	No. of batches	Batch IDs	No. of patients	% relapse	[min(t),max(t)]	Ratio of [N,LA] to [ERBB2, LB, basal]	% G1	% G3
GSE1456	1	1	159	25.1	[0.18,8.49]	0.96	19.0	41.4
GSE2034	1	3	286	37.4	[0.16,14.25]	0.74	N.d.	N.d.
ETABM158	1	8	129	30.2	[0,14.20]	0.63	11.2	51.6
GSE19615	1	11	115	12.1	[0.08,7.33]	0.65	20.0	55.6
GSE11121	1	12	200	23.0	[0.08,20.00]	0.80	14.5	17.5
GSE31519	1	14	64	35.9	[0.25,10.00]	0.45	N.d.	N.d.
GSE12276	5	15–19	204	97.5	[0.35,7.61]	0.53	N.d.	N.d.
GSE9195	1	25	77	16.8	[0.57,11.29]	0.61	24.1	41.3
GSE6532	5	27–31	393	36.5	[0.27,13.55]	0.75	23.9	23.7
GSE7390	5	32–36	198	47.4	[0.72,18.21]	0.79	15.1	43.1
GSE25066	1	46	508	21.8	[0,7.43]	0.65	6.7	54.7

N.d., no data are available for the variable and dataset.

Dataset S1. Histological review of the tumor samples used for sequencing

Dataset S1

Dataset S2. Characteristics of CIS

Dataset S2

Dataset S3. The 126 mouse and human BC gene candidates from 129 mouse CIS loci

Dataset S3

Dataset S4. Somatic mutations in BCSGs identified in databases of somatic mutations in human BC tissues and cell lines

Dataset S4

Dataset S5. GO and Pathway Enrichment Analysis of 126 hBCSG (enrichment analyses held on 7 October 2014)

Dataset S5

Dataset S6. Datasets of mouse and human BCSGs

Dataset S6

(*A*) Tumor sample IDs and data sources. (*B*) Integrated set of 394 gene-expression microarrays of mouse mammary tumors and the tumor subtypes. Presented are 220 annotated probe sets. (*C*) Integrated sets of 1,345 gene-expression microarrays of human BC and the BC subtypes. Batch-effect was corrected. Normalized and log-transformed microarray data were used. Gene-expression values were presented by 280 microarray probe sets. (*D*) Integrated sets of the 2,333 Affymetrix U133A microarrays of BC patients with supported clinical characteristics including DFS data. (*E*) Clinical data and tumor subtype information supporting the Affymetrix U133A expression microarray probe sets data of tumor samples from 2,333 BC patients.

Dataset S7. Survival prediction analysis: 1D-DDg, 2D-DDg, and SWVg

Dataset S7

(*A*) The 126 Affymetrix U133A probe set selected by the 1D-DDg method at Wald statistics $P < 0.01$. (*B*) Summary statistics 1D-DDg–selected genes of prooncogenic and suppressor-like prognostic genes; $P < 0.01$. (*C*) The six-gene-pair prognosis classifier: initial data and gene annotation. (*D*) Affymetrix U133A microarray and available clinical data for the basal-like and claudin-low BC subtypes. (*E*) Prognostically significant genes for the basal-like subtype analyzed with 1D-DDg. In this analysis, the minimum number of patients in each group was >30. (*F*) Prognostically significant genes for the claudin-low subtype analyzed with 1D-DDg. In this analysis, the minimum number of patients in each group was >10. (*G*) Prognostically significant genes for the basal-like subtype analyzed with the SWVg method (three groups). (*H*) Prognostically significant genes for the claudin-low subtype analyzed with the SWVg method (three groups). (*I*) Unique and common gene symbols in our three prognostic signatures. (*J*) Analyzed BC signatures.

Dataset S8. Microarray and clinical data for 2,333 BC patients and the results of 1D-DDg, 2D- DDg, SWVg, univariate, and multivariate analyses

Dataset S8

(*A*) Integrated data (master table). (*B*) Univariate Cox proportional hazard regression for all 2,333 patients. (*C*) SWVg six-gene-pair prognostic signature. Multivariate Cox proportional hazard regression for all 2,333 patients. (*D*) SWVg 21-gene prognostic signature. Multivariate Cox proportional hazard regression for basal patients.

Dataset S9. Univariate and multivariate analyses of 226 TCGA patients treated with hormone therapy, chemotherapy, or combined adjuvant therapy

Dataset S9

(*A*) The 226 TCGA patients were treated with hormone therapy, chemotherapy, and combined hormone and chemotherapy. Integrated data (Master table). (*B*) Summary of the univariate (1) and multivariate (2) Cox proportional hazard regression model. (*C*) 1D-DDg analysis. (*D*) 2D-DDg analysis. (*E*) 1D-DDg–based (modified) SWVg. The three risk groups based on the six-gene-pair classifier. (*F*) Univariate analysis. (*G*) Multivariate analysis. (1) Data for multivariate analysis. (2) Results of multivariate analysis. (3) Characteristics of multivariate survival analysis by Cox proportional hazard regression model.

Dataset S10. Therapeutic drug analysis for hBCSGs via MetaCore (direct interaction, version 6.12 build 42289)

Dataset S10