Supplemental Material and Methods

Primary tumor processing and implantation. Each mouse injected with primary tumors was supplemented in estrogen by implantation of intrascapular estrogen pellets (Innovative Research of America, Sarasota, FL, USA). Tumor growth was measured weekly using callipers. The animals were euthanized when the tumors were approximately 12 mm in the largest diameter (<1000 mm$^3$), to avoid tumor necrosis and in compliance with regulations for use of vertebrate animals in research. A portion of each tumor was fixed in formalin and embedded in paraffin for histological analysis. The rests of the tumor were for a part re-implanted into secondary mice, for another part flash frozen in liquid nitrogen for genomic analysis, and for the last part processed for ALDEFLUOR phenotype analysis. Animals without palpable outgrowth were euthanized 1 year after primary tumor injection and each fat pad injected was fixed in formalin and embedded in paraffin for histological analysis.

Immunohistochemistry. Antibodies used are listed below: Stem cell marker Aldehyde dehydrogenase 1 (ALDH1) (Mmab clone 44, BD Biosciences, 1/50), Estrogen receptor (ER) (Mmab, clone 6F11, Novocastra Laboratory,1/60), Tyrosine kinase receptor ERBB2 (Mmab, clone AO 485, Dakocytomation, 1/800), MIB1/Ki-67 (Mmab, clone Ki-67, Dakocytomation, 1/100), Progesterone receptor (PR) (Mmab, clone PgR 636, Dako Corporation, 1/80), and Tumor suppressor P53 (Mmab, clone DO-1, Immunotech, 1/4). IHC was done as previously described on 5 μm paraffin sections (1). Briefly, after deparaffinization, slides were pre-treated according to the supplier’s recommendations, transferred to a Dako autostainer. This was followed by
the use of a streptavidin/biotin kit (Dako, Trappes 78196, France). Diaminobenzidine (DAB) or 3-amino-9-ethylcarbazole (AEC) was used as chromogen. Double immunostaining with antibodies for detection of CD44 and CD24 was performed by an Autostainer (DAKO) using EnVision Flex Target Retrieval Solution High pH (DAKO) according to the manufacturer’s instructions. Antibodies for detection of CD44 (Clone 156-3C11, 1:400) and CD24 (Clone SN3b, 1:300) were purchased from Neomarkers (Fremont, CA). CD44 was detected with DAB and CD24 was detected using AEC. Sections counterstained with hematoxylin were evaluated by light microscopy by two independent observers. Immunoreactivities were scored mainly by measuring the percentage of positive cells in nucleus (ER, Ki67, PR, P53), cytoplasm (ALDH1A1) or membrane (ERBB2) of tumoral cells, with cut-off values in accordance with previous studies (1). For ERBB2, quantification was done using the Dako scale in accordance with ASCO/CAP 2007 recommendations (2). For double immunostaining CD44/CD24, we evaluated the percentage of cells expressing CD44 (membrane staining) and without CD24 protein expression (cytoplasm staining).

**Gene expression analysis.** Expression data were analysed by the RMA (Robust Multichip Average) method in R using Bioconductor and associated packages (3). RMA performed the background adjustment, the quantile normalization and finally the summarization of 11 oligonucleotides per gene. The molecular subtypes of tumours and PDXs were determined using the PAM50 predictor (4). Supervised analyses searched for genes differentially expressed between ALDEFLUOR-positive cell samples (N=8) and ALDEFLUOR-negative samples (N=8) were done using SAM analysis (Significance Analysis of Microarrays) for two-class
paired cases (5) with a FDR inferior to 5%. The robustness of the signature – thereafter designed the breast CSC gene expression signature (BCSC-GES) - was internally tested using PAM (Prediction Analysis of Microarrays) cross-validation (6).

To help in the interpretation, the list of differential genes was interrogated using the Ingenuity Pathway Analysis (IPA) software (version 5.5.1-1002; Ingenuity Systems, Rewood City, CA). Based on the BCSC-GES, samples were classified according to the Pearson correlation coefficient of their expression profile with the centroid of ALDEFLUOR-positive cell samples. Positive correlation defined “BCSC-like” samples, and negative correlation defined “non-BCSC-like” samples. To test the clinical correlations and the prognostic performance of this GES in breast cancer, we analyzed 13 public data sets collected from 11 publications (7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17), and from the UNC Microarray Database and our database (Table S4). When different publications included the same patients redundancy was eliminated, resulting in 2,609 different patient samples available for analysis. Before analysis, we mapped hybridization probes for the 837 genes across the different microarray platforms used across the series. When multiple probes were mapped to the same GeneID (EntrezGene identification number), the one with the highest variance in a particular dataset was selected to represent the GeneID. Analysis of each data set (using available normalized data) was done separately to guarantee a larger number of genes common with our signature. The samples were classified as “BCSC-like” or “non-BCSC-like” as described above. We searched for common expression patterns between the 836 up-regulated genes of our 837-gene signature and three murine stemness GES defined from embryonic (1787 genes up-regulated in stem cells), neural (2458 up-regulated genes) and hematopoietic (1977 upregulated genes) stem cells (18). After having matched these murine genes with
our human Affymetrix chip and eliminated the redundancy contained in these murine GES, we applied an enrichment test (hypergeometric test, FDR-corrected). A total of 19 genes were common to the four GES. We then derived a classifier based on the combined expression of these 19 genes by computing a metagene using principal component analysis (PCA) of the matrix of expression levels. The metagene was defined as the first eigenvector (i.e. direction of the largest variance of data) and was tested as continuous value for its prognostic performances.

Array comparative genomic hybridization analysis. We used a reference DNA was a pool of 13 normal male DNA. Scanning was done with Agilent Autofocus Dynamic Scanner (G2565BA, Agilent Technologies). Normalized log2 ratios were obtained from “Feature extraction” software (Agilent Technologies). Data generated by probes mapped to X and Y chromosomes were eliminated. The final dataset contained 234,583 unique probes covering 22,831 genes and intergenic regions according to the hg18/NCBI human genome mapping database (build 36.1). Data were analyzed using circular binary segmentation (CBS) as implemented in the DNA copy R/Bioconductor package with default parameters to translate intensity measurements in regions of equal copy number, each region being defined by at least five consecutive probes. Thus, each probe was assigned a segment value referred to as its “smoothed” value.

Statistical analysis. Correlations between sample groups and molecular parameters were calculated with the Fisher’s exact test or the one-way ANOVA for independent samples. A p-value <0.05 was considered significant. Data were summarized by
frequencies and percentages for categorical variables, and by median and range for continuous variables. To investigate associations among variables, univariate analysis was performed using non-parametric Wilcoxon rank sum test, Chi-Square test or Fisher’s exact test when appropriate. The correlation of CSC variation data obtained from both siRNA screen using either ALDEFLUOR phenotype or SFE was measured using Pearson’s rank correlation. A Metastasis-Free-Survival (MFS) rate was estimated by the Kaplan-Meier method using the first metastasis recurrence as first event. Patients without events were censored at the time of last follow-up or at the date of the death if they died for other reason than BC. All survival times were calculated from the date of BC diagnosis. Univariate analysis with binary variables available in more than 1700 patients: age (inferior or equal to 50 versus superior to 50), pN (positive versus negative), pT (pT2-3 versus pT1), grade (2-3 versus 1), IHC ER status (negative versus positive). Multivariate analysis was done using Cox's proportional hazard models with a backward stepwise selection of variables to minimize the Akaike Information Criterion.
Reference List


