Supplementary Figure Legend

Supplementary Figure 1. Quantification of the normalized photon flux measured following incubation with D-luciferin of sorted cells (ALDEFLUOR-positive and ALDEFLUOR-negative) infected with the luciferase vector from three different cell lines (**A**, HCC1954; **B**, MDA-MB-453; **C**, SUM159). To assess the efficiency of the lentivirus infection different number of cells were plated for each cell lines (100; 1,000; 10,000). An identical level of luciferase infection was detected for the ALDEFLUOR-positive and ALDEFLUOR-negative population of the three different cell lines tested.

Supplementary Figure 2. Representative flow cytometry analysis of ALDH enzymatic activity of 33 BCLs. The ALDEFLUOR assay was performed as described in the material and methods section. Cells incubated with ALDEFLUOR substrate (BAAA) and the specific inhibitor of ALDH, DEAB, were used to establish the baseline fluorescence of these cells (left chart) and to define the ALDEFLUOR-positive region (right chart). Data represent mean ± SD.

Supplementary Figure 3. ALDEFLUOR-positive cells are enriched in mammosphere-forming capacity. ALDEFLUOR-positive cells sorted from three different BCLs (SUM149, SUM159, 184A1) were enriched in sphere-initiating cells generated by 10,000 cells plated compared to ALDEFLUOR-negative cells.

Supplementary Fig. 4. Representative flow cytometry analysis of ALDH enzymatic activity in MDA-MB-453 (with DEAB **Aa,** without DEAB **Ab**) and SUM159 cells (with DEAB **Ba,** without DEAB **Bb**) before injection in NOD/SCID mice. The ALDEFLUOR assay was performed as described in the material and methods section. **Ac, Bc.** The ALDEFLUOR-positive population was capable of generating tumors in NOD/SCID mice which recapitulated the phenotypic heterogeneity of the initial tumor. Data represent mean ± SD.

Supplementary Fig. 5.**The ALDEFLUOR-positive cell population from BrCa-MZ-01 cell lines has cancer stem cell properties whereas the ALDEFLUOR-negative population contains proliferative progenitor cells. A-B**. Representative flow cytometry analysis of ALDH enzymatic activity in BrCa-MZ-01 cells. **C-G**.The ALDEFLUOR-positive and -negative populations were capable of tumor generation. Tumors generated by the ALDEFLUOR-positive population reconstituted the phenotypic heterogeneity of the initial tumor upon serial passages (**C-E**) whereas the ALDEFLUOR-negative population gave rise to tumors containing only ALDEFLUOR-negative cells that could not be transplanted more than three times (**F-G**). **H-I**.Tumor growth curves were plotted for different numbers of cells injected (**H**: 100,000 cells, 10,000 cells, and 1,000 cells and I: 50,000 cells, 5,000 cells, and 500 cells) and for each population (**H**: ALDEFLUOR-positive, ALDEFLUOR-negative, unseparated and **I**: ALDEFLUOR-positive cells from tumors generated from ALDEFLUOR-positive cells (+/+), ALDEFLUOR-negative cells from tumors generated from ALDEFLUOR-positive

cells (+/-), and ALDEFLUOR-negative cells from tumors generated from ALDEFLUOR-negative cells (-/-). The latency and size of tumor formation correlated with the number of ALDEFLUOR-positive cells implanted whereas the ALDEFLUOR-negative cells gave rise to slowly growing tumors that showed decreasing tumor growth upon serial passages, with no growth following three passages. (**H,I**).

Supplementary Fig. 6. Classification of the ALDEFLUOR-positive and ALDEFLUOR-negative populations isolated from breast cell lines based on the "cancer stem cell signature". Hierarchical clustering of 16 samples based on a 413-gene expression signature. Each row of the data matrix represents a gene and each column represents a sample. Expression levels are depicted according to the color scale shown at the bottom. Note the separation between ALDEFLUOR-positive (red names) and negative samples (black names) with the 413 genes for 15 out of the 16 samples. Some genes included in the signature are referenced by their HUGO abbreviation as used in 'Entrez Gene' (Genes down-regulated in the ALDEFLUOR-positive populations are labeled in green and genes up-regulated in the ALDEFLUOR-positive populations are labeled in red).

Supplementary Fig. 7. Classification of the ALDEFLUOR-positive and negative populations isolated from breast cell lines based on the 49-genes signature. A. Hierarchical clustering of 16 samples according to the mRNA

expression levels of the 49 genes of the signature. Each row of the data matrix represents a gene and each column represents a sample. Expression levels are depicted according to the color scale shown at the bottom. The overexpressed genes in ALDEFLUOR-positive samples are at the bottom (IL8RA is underlined). Note in A the perfect separation between all the positive samples (red names) that clustered together on the left part of the dendrogram and the negative samples (blue names) on the right part. Estimation of the accuracy of prediction of the molecular signature using Leave-one-out cross validation (**B-C**). **B**. Graphical representation of the validation where each sample was excluded one by one and classified with the LDA analysis. ALDEFLUOR-positive (red dots) and ALDEFLUOR-negative samples (black dots) are well classified using this model except for one. **C**. Statistical validation of the model with a concordance rate of 88% between the predicted and the observed class of a sample and a significant correlation using Fisher-exact test.

Supplementary Figure 8. Validation of gene expression results by quantitative RT-PCR. To validate our gene expression data measured by DNA microarrays, we analyzed in a set of five breast cancer cell lines, sorted for the ALDEFLUOR phenotype, the level of mRNA expression of five discriminator genes overexpressed in ALDEFLUOR-positive populations (*CXCR1/IL8RA, FBXO21, NFYA, NOTCH2* and *RAD51L1*) by quantitative RT-PCR. The quantitative RT-PCR expression level of *NFYA, NOTCH2*, and *RAD51L1* are presented in this figure. The increase of mRNA expression level in the

ALDEFLUOR-positive population compared to the ALDEFLUOR-negative population for these three genes is statistically significant for 3 out of 5 cell lines for *NOTCH2*, for 2 out of 5 cell lines for *NFYA*, and for 3 out of 5 cell lines for *RAD51L1* (* $p<0.05$).

Supplementary Fig. 9. Overlap between the cell population sorted with the ALDEFLUOR assay and the cell population expressing CXCR1 protein. Cells expressing CXCR1 are contained in the ALDEFLUOR-positive population. The ALDEFLUOR-positive and -negative population from four different breast cell lines (HCC1954, SUM159, MDA-MB-453, BrCa-MZ-01) were isolated by FACS, fixed, and analyzed for the expression of CXCR1 protein by immunostaining and FACS analysis. ALDEFLUOR-positive cells were highly enriched in CXCR1-positive cells compared to the ALDEFLUOR-negative population.

Supplementary Fig. 10. Detection of metastasis in three different breast cancer cell lines (HCC1954, MDA-MB-453, SUM159) utilizing the bioluminescence imaging software (**First line**: Mice facing down; **Second line**: Mice facing up). Mice inoculated with ALDEFLUOR-positive cells developed several metastasis localized at different sites (bone, muscle, lung, soft tissue) and displayed a higher photon flux emission than mice inoculated with unseparated cells, which developed no more than one metastasis per mouse. In contrast, mice inoculated with ALDEFLUOR-negative cells developed only an occasional small metastasis, which was limited to lymph nodes