Supplementary Figure Legends

Figure S1. Average delay of growth of PDXs among passages (up to P5). PDXs growth kinetic (time to reach 500 mm^3) was mostly stable (CRCM 258, 181,168, 226, 237) or accelerated (CRCM 214, 216, 177, 174, 272) with serial passages.

Figure S2. DNA copy number variations found in the original primary tumors are well maintained in PDXs. High-resolution 244K CGH microarrays are shown to illustrate DNA copy-number changes relative to normal DNA. Each primary tumor (white box) and the corresponding PDX (black box) (and, in some cases PDXs that were serially passaged more than 8 times) are indicated on the left. A copy number of two (defined as normal) is indicated in black, a copy number greater than two (chromosomal gain or amplification) is showed in red, and a copy number of less than two (chromosomal loss or deletion) is shown in blue. The position of the copy number variant across the 22 autosomal chromosomes and the X chromosome is show at the bottom.

Figure S3. Example of stromal dilution in parental primary tumor CRCM184 PT compared to paired CRCM 184 X, generating a possible artefact in molecular subtype due to stromal dilution in human sample. (**Figure S3**).

Figure S4. ALDEFLUOR phenotype of our 20 PDX models. For each of our PDX, ALDEFLUOR phenotype was determined in early passages (P<2). On the *left panel*, examples of FACS flow charts with cells incubated with ALDEFLUOR substrate (BAAA) and the specific inhibitor of ALDH, DEAB, used to establish the baseline fluorescence of these cells and to define the ALDEFLUOR-positive region. All our PDXs presented an ALDEFLUOR-positive populations ranging from 0.2% to 12.3% as listed in table (*right panel*). Data represent mean ± SD.

Figure S5. Outgrowth kinetic of each sorted cell population isolated from three PDXs and injected in limited dilutions in fat pads of NSG mice. ALDEFLUOR-positive, -negative, and unselected populations from three PDXs (CRCM226 x, CRCM168 x, CRCM174 x) were injected in limited dilutions (5.10^4, 5.10^3, 5.10^2 cells) in fat pads of NSG mice. For each limited dilutions (*in column*) and each PDX models

(*in lane*) an outgrowth kinetic is represented with in red the ALDEFLUOR-positive population (A+), in green the unselected population (U) and in blue the ALDEFLUOR-negative population (A-). Data represent mean ± SD.

Figure S6. Univariate and multivariate analysis identified primary tumor engrafment as independent parameter associated with metastasis-free survival (MFS) in the series of 74 patients. Univariate analysis (upper table) identified pN, SBR Grade, and primary tumor engrafment as parameter associated with MFS. Multivariate analysis (lower table) with Cox proportional model identified only pN and primary tumor engraftment as independent parameters associated with MFS.

Figure S7. Identification of parameters that predict successful engraftment. Kaplan-Meier analysis show proportion of tumor outgrowth for different groups of primary tumor injected and stratified according to either Ki67, specimen type, presence of a CD44+/CD24- cell population, IBC status, or ERBB2 protein expression. Only a high Ki67 index predict significantly successful engraftment (LogRank test, p=0.003).

Figure S8. Global gene expression profiling of 53 PDXs. Hierarchical clustering of 53 PDXs and 13404 genes/ESTs based on mRNA expression levels. The dendrogram of samples represents overall similarities in gene expression profiles. Two large groups of samples are evidenced by clustering (separated by an orange line). Molecular subtype, Claudin-low subtype, and engraftment status of primary tumors injected are represented according to color ladders reported below the dendrogram.

Figure S9. Supervised analyses of aCGH data comparing copy number variations in 17 paired primary tumors/PDXs. Genomic segments are ordered on the X axis from chromosome 1 to Y and on the Y axis according to their association with primary tumors or PDXs (-log(p-value)). All the genomic segments analysed were not significantly associated with any of the two groups (primary tumors or PDXs) (p>0.01, red lines).

Figure S10. Representative network of genes isolated from the BCSC-GES and involved in DNA damage repair. This network was generated using Ingenuity® software. Node colors indicate genes from the BCSC-GES (*in blue*) and key node genes not present in the BCSC-GES (*in white*).

Figure S11. Representative network of genes isolated from the BCSC-GES and involved in cell cycle control. This network was generated using Ingenuity® software. Node colors indicate genes from the BCSC-GES (*in blue*) and key node genes not present in the BCSC-GES (*in white*).

Figure S12. The "common" stem cell core transcriptional program (CE-4SC) is associated with clinical outcome. A. Kaplan-Meier metastasis-free survival curves according to CE-4SC status. Tumors that present the CE-4SC gene (CE-4SC-positive) are significantly associated with a reduced metastasis-free survival (LogRank test, $p=4.86^{-6}$). **B.** Multivariate analysis with Cox proportional model identified CE-4SC as independent parameter associated with metastasis-free survival.

Figure S13. Functional validation of the CE-4SC in two different breast cancer cell lines (BCLs). Screening of a library of 57 siRNAs targeting the 19 identified genes, using the variation of the ALDEFLUOR-positive population as read out, was performed in SUM159 (Mesenchymal BCL; left panel) and in S68 (Luminal BCL; right panel). Each gene with at least one targeting siRNA construct that induces a variation of the number of ALDEFLUOR-positive cells over a threshold of two-fold the CSC proportion detected in the control was considered as a hit. In green, gene targeted by at least one siRNA construct that reduced significantly the CSC population, in red, gene targeted by at least one siRNA constructs that increased significantly the CSC population, in yellow gene targeted by two siRNAs inducing opposite effect, and in grey gene targeted by siRNA constructs that had no effect on the CSC population. Correlation between both screen are evaluated using Chi2-squared test (p=0.003). Data represent mean ± SD.

Figure S14. Knock-down of genes from the CE-4SC affect tumorsphere-forming efficiency (SFE) in S68 BCL. Comparison of CSC variation after gene knock-down

by ALDEFLUOR phenotyping (right) and SFE (left) using the 57 siRNAs constructs. Results for each siRNA (from top to bottom) are represented as opposite bars. Correlations are measured using Spearman’s rank correlation (ρ). Data represent mean ± SD.