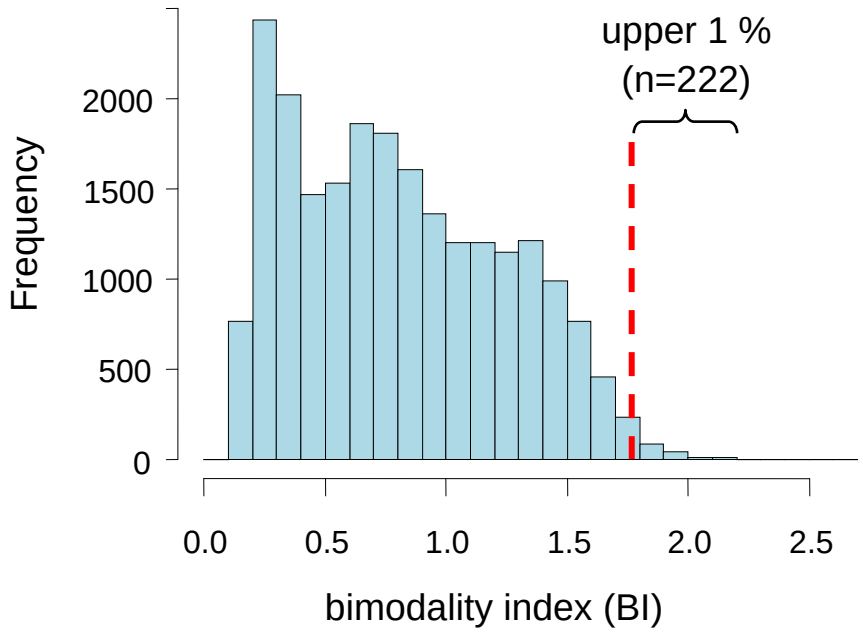
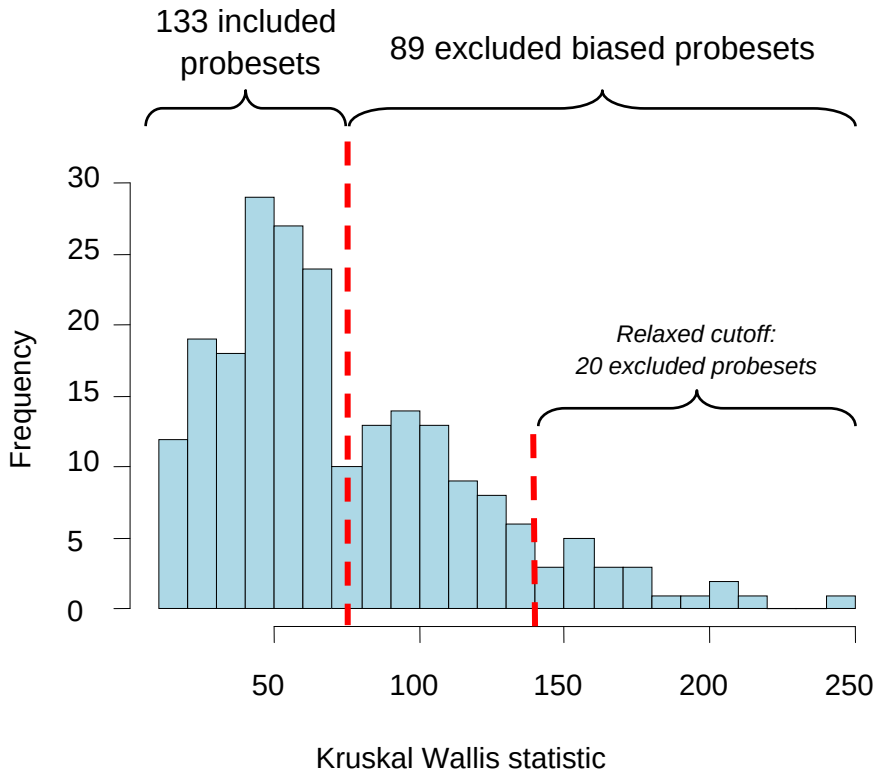


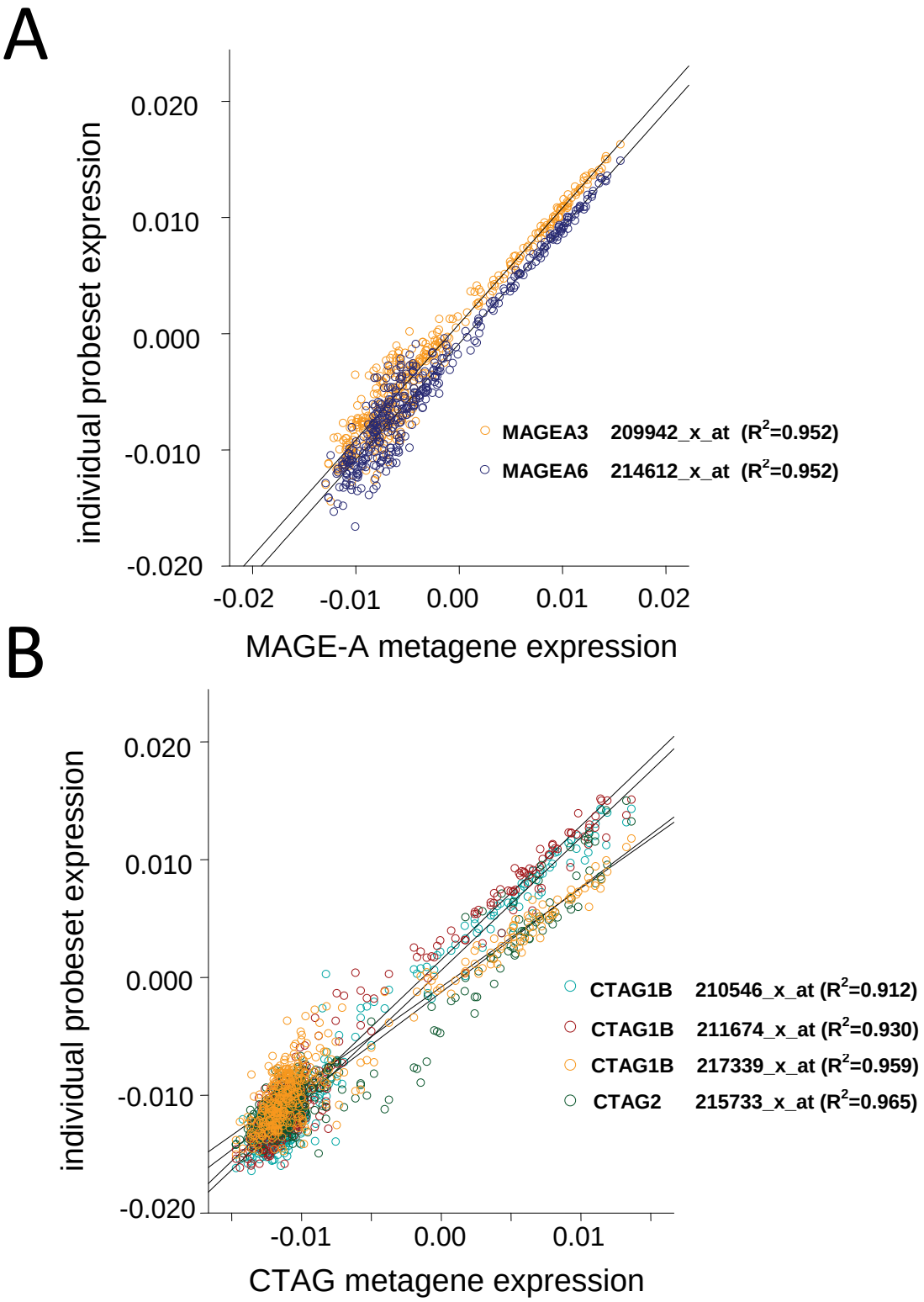
Supplementary Figure S1: Distribution of bimodality index values

The distribution of the bimodality index values of all 22,283 Affymetrix probesets in the dataset of 394 triple negative breast cancer samples is shown. The vertical red dashed line delineates the upper 1 % of 222 probesets with BI values > 1.768.



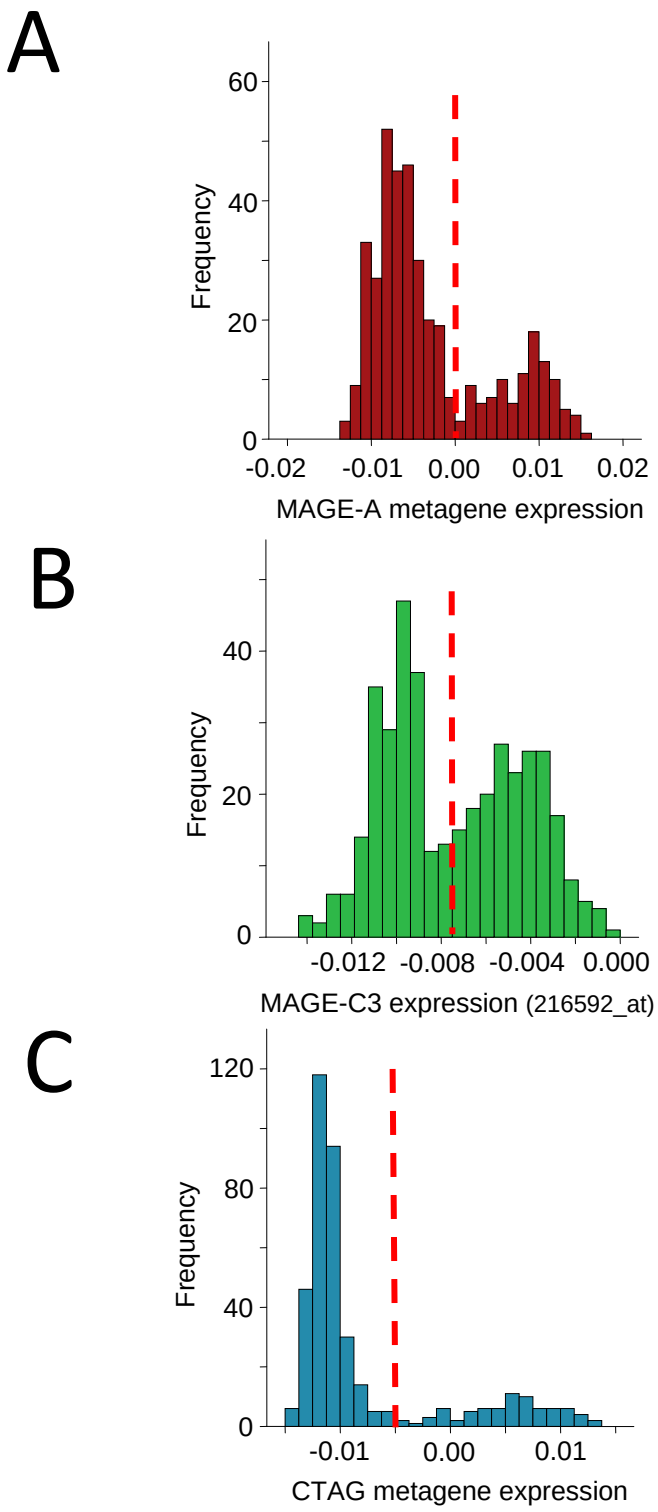
Supplementary Figure S2: Distribution of Kruskal Wallis statistics as marker for biased probesets

To identify biased probesets whose expression is associated with specific primary datasets the standard Kruskal Wallis test was applied. The distribution of the Kruskal Wallis statistics comparing the expression of each of the 222 bimodal genes with the dataset vector is shown. A cutoff of 75 was derived from the distribution excluding 89 biased probesets. In addition a relaxed cutoff of 140 is also shown which excluded only 20 probesets with strongest bias and was used in control experiments.



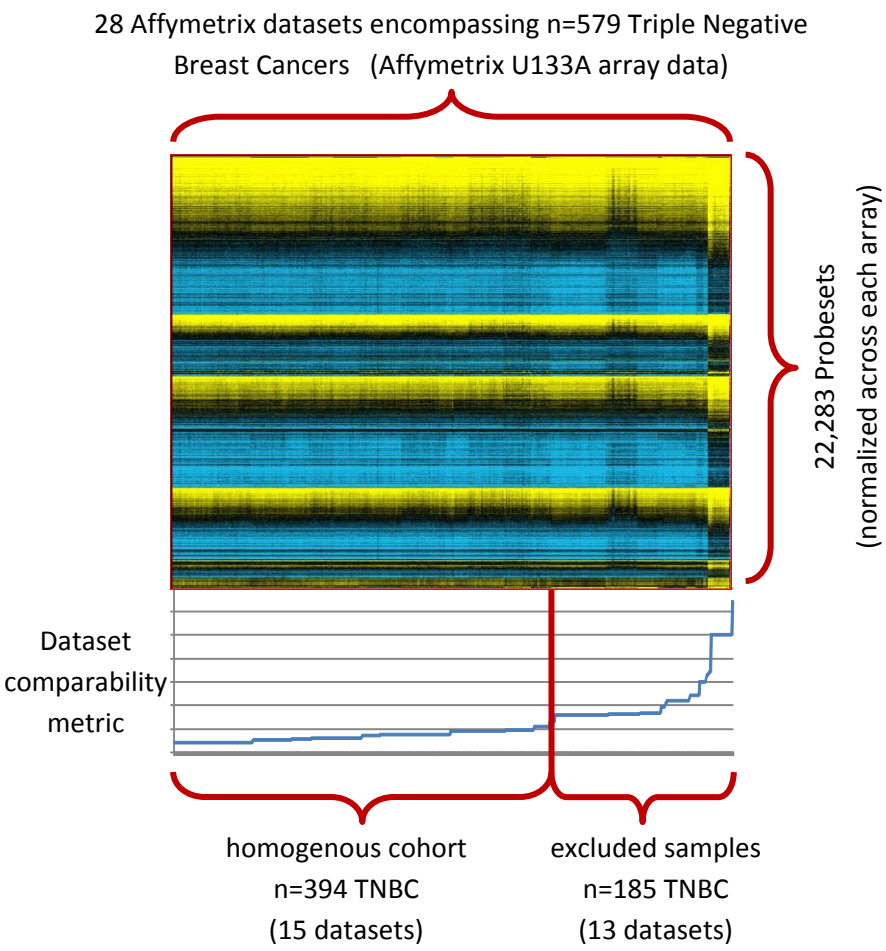
Supplementary Figure S3: Correlation between individual probesets and metagenes for MAGE and CTAG.

Shown are scatter plots comparing the individual probesets and the respective metagenes (means of all probesets). In (***A***) the results for the two probesets for members of the melanoma antigen family A (MAGEA3 and MAGEA6) are shown and in (***B***) the results of the three probesets for CTAG1B and the probeset for CTAG2



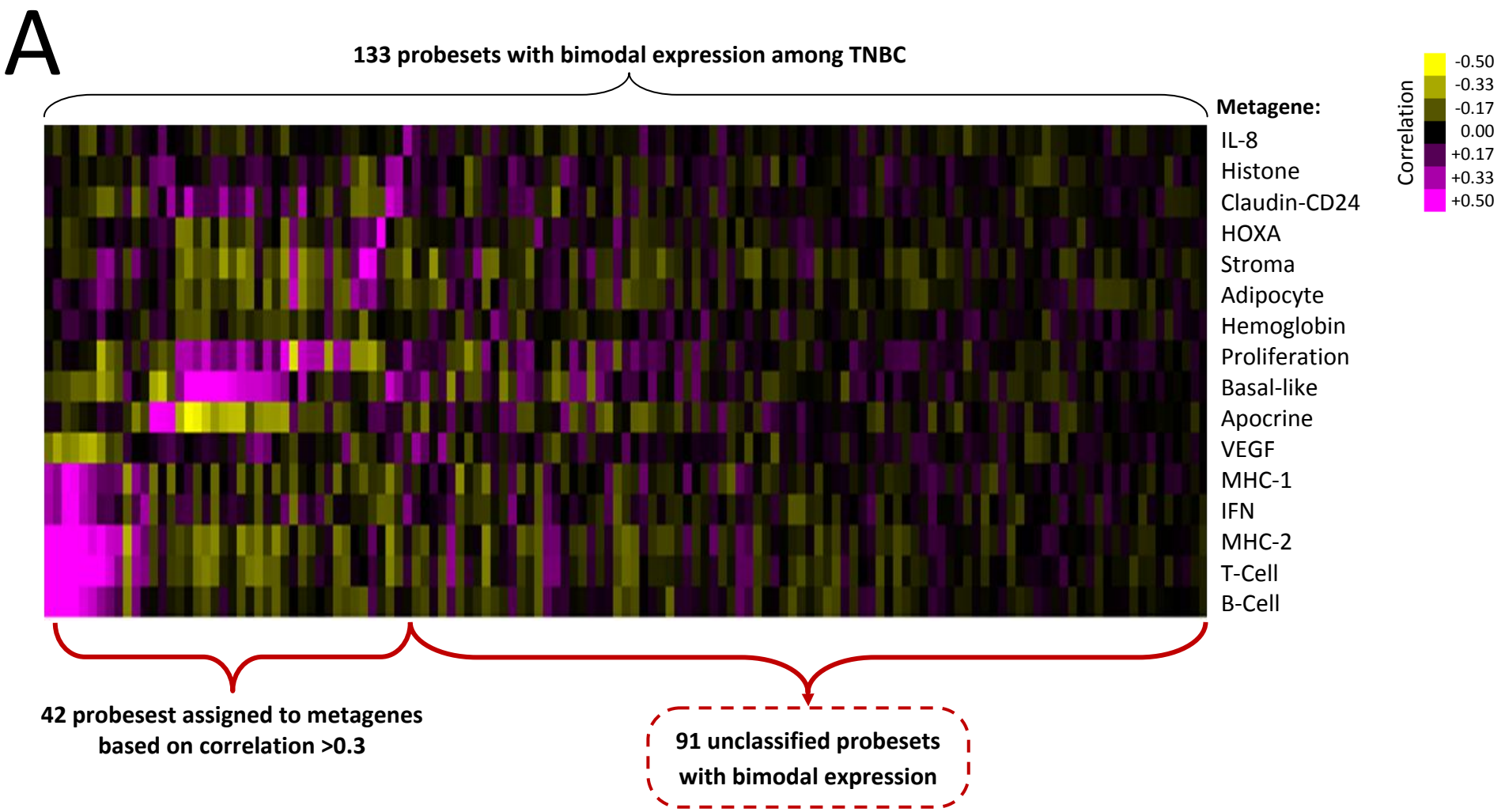
Supplementary Figure S4: **Distribution of expression values of the metagenes for MAGE-A, MAGE-C3 and CTAG among TNBC samples.**

The distribution of the expression values of the MAGE-A, MAGE-C, and CTAG metagenes in the dataset of 394 triple negative breast cancer samples is shown in (A), (B), and (C), respectively. The vertical dashed red lines denote the cutoff values used for Kaplan-Meier analyses (0.00, -0.0075, and -0.005 in *A*,*B*, and *C*, respectively).

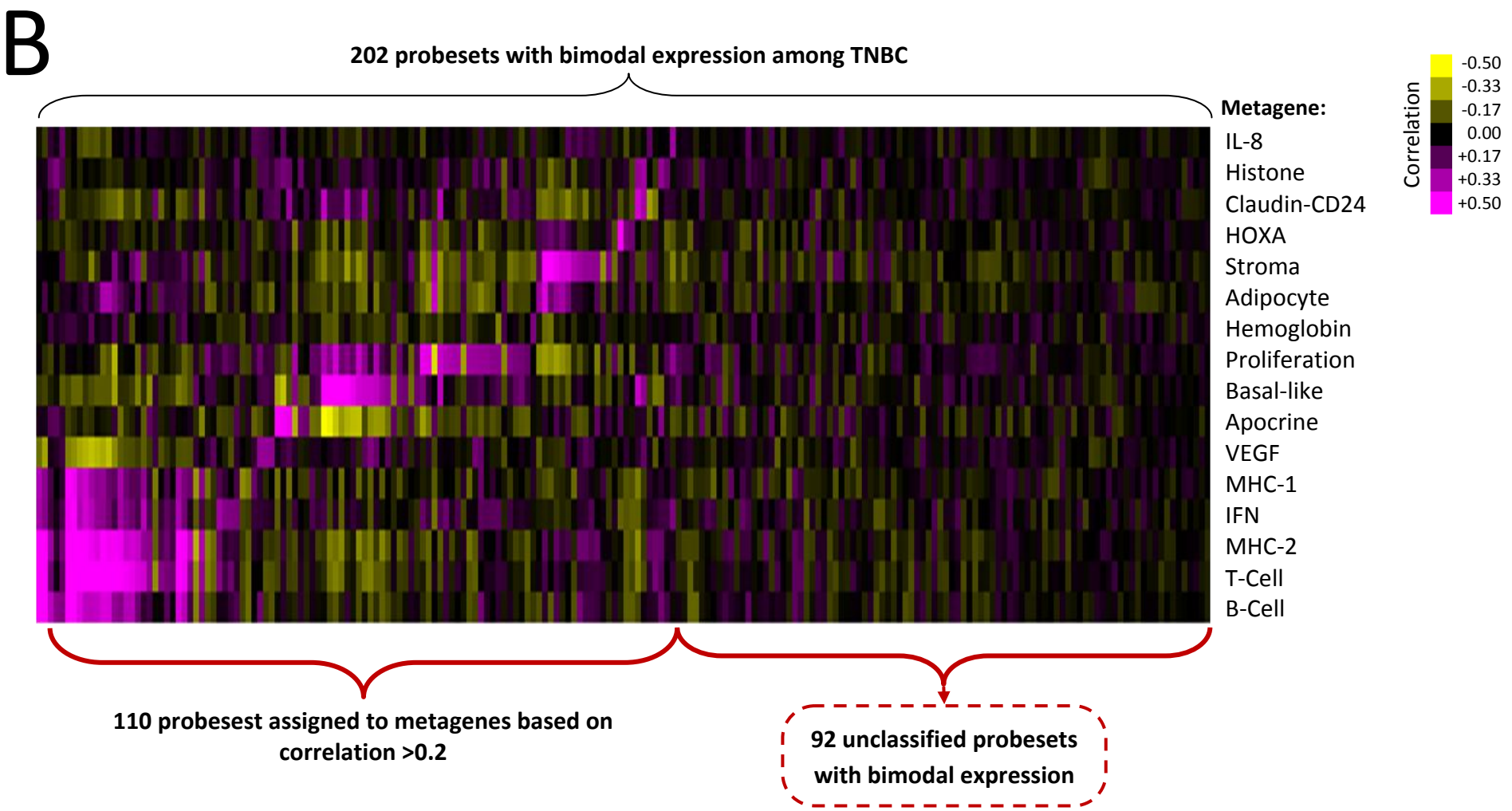


Supplementary Figure S5: **Selection of 394 TNBC with homogenous microarray data from multiple datasets based on dataset comparibility**

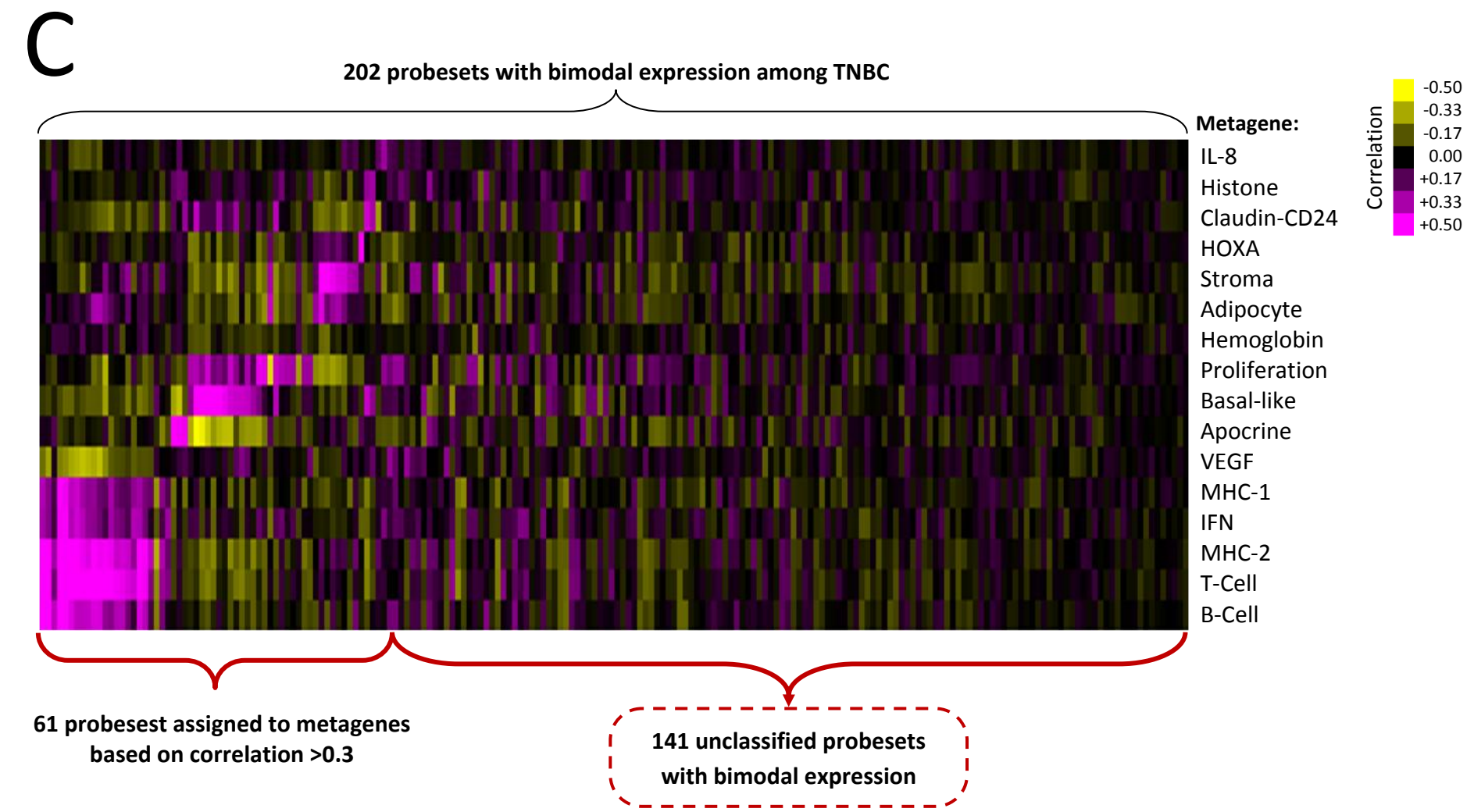
Triple negative breast cancers (TNBC, n=579) from 28 datasets were sorted by dataset according to a dataset comparability metric (horizontally). Shown are the full array data of normalized Affymetrix U133A microarrays. The 15 most comparable datasets encompassing n=394 TNBC samples were subsequently used as a homogenous cohort and the remaining 13 datasets (n=185 TNBC samples) excluded.



Supplementary Figure S6 (*continued*)



Supplementary Figure S6 (*continued*)

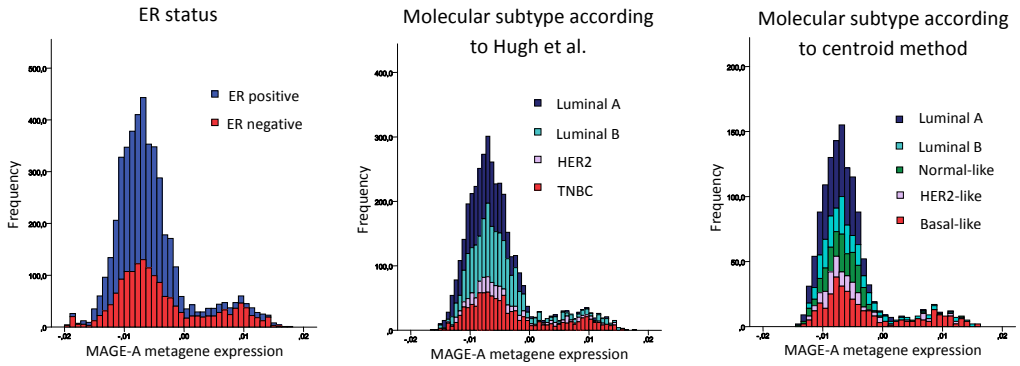


Supplementary Figure S6 (*continued*)

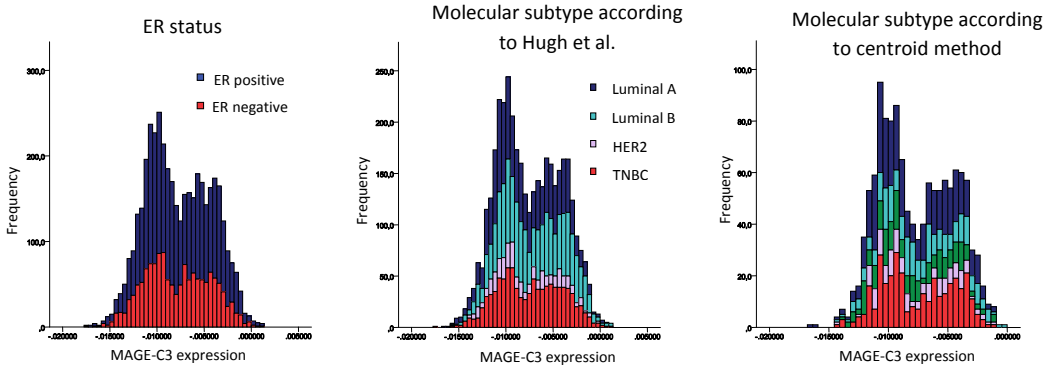
Supplementary Figure S6: Stability analysis through different thresholds for correlation to metagenes and dataset bias: Correlation matrices of probe sets with bimodal expression and 16 metagenes in TNBC using alternative cutoff values.

Shown are colour representations of the correlation matrices of probe sets with a bimodal expression (horizontically) and the 16 metagenes for different molecular phenotypes in TNBC (vertically). Either the 133 probe sets with bimodal expression from Table 1 were used (***A***) or those 202 probe sets selected by the more relaxed Kruskal-Wallis filter for potential dataset bias (***B***,***C***). In (*B*) the same minimal cutoff of a correlation >0.2 for assignment of a probe set to a metagene as in Figure 2 was used. In (***A***) and (***C***) a more stringent cutoff of a correlation >0.3 for assignment was applied resulting in a larger number of "unclassified" bimodally expressed probe sets. In all figures probe sets are grouped according to the assigned metagene and sorted according to their correlation from left to right in decreasing order. Positive correlation values are represented by magenta and negative correlation values by yellow.

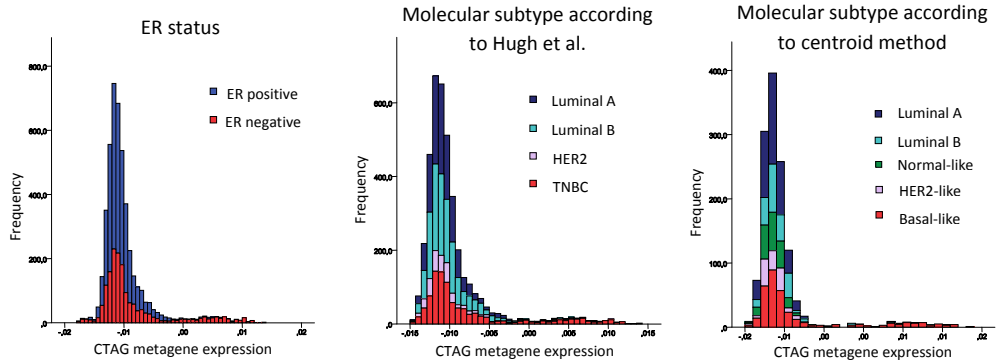
A



B



C

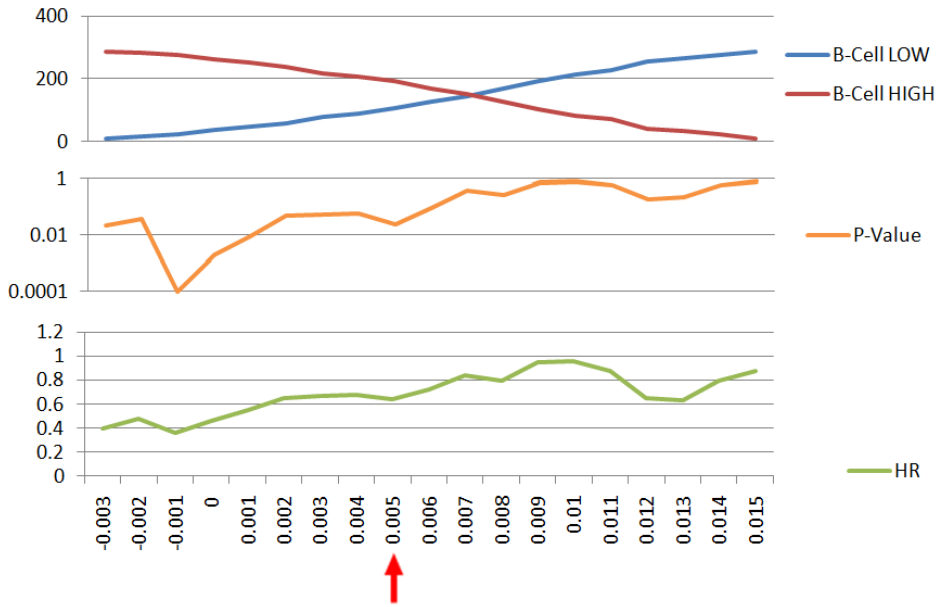


Supplementary Figure S7: Expression of MAGE-A, MAGE-C3 and CTAG metagenes in different molecular subtypes of breast cancer

A) The distribution of the expression values of the MAGE-A metagene in different molecular subtypes is shown. Samples were either stratified by ER status of the tumor or by molecular subtype according to the method of Hugh et al. (*J Clin Oncol*., 2009; 27:1168) or the centroid method using the intrinsic gene set of Hu et al. (*BMC Genomics*, 2006, 7:96).

B) The distribution of the expression values of MAGE-C3 in different molecular subtypes is shown. Samples were stratified as in (A).

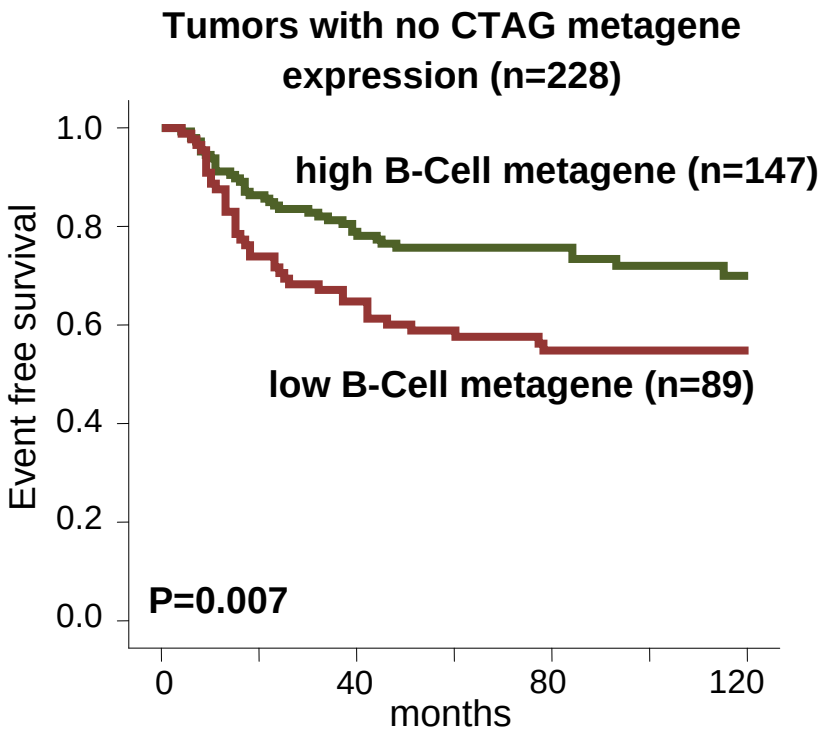
C) The distribution of the expression values of the CTAG metagene in different molecular subtypes is shown. Samples were stratified as in (A).



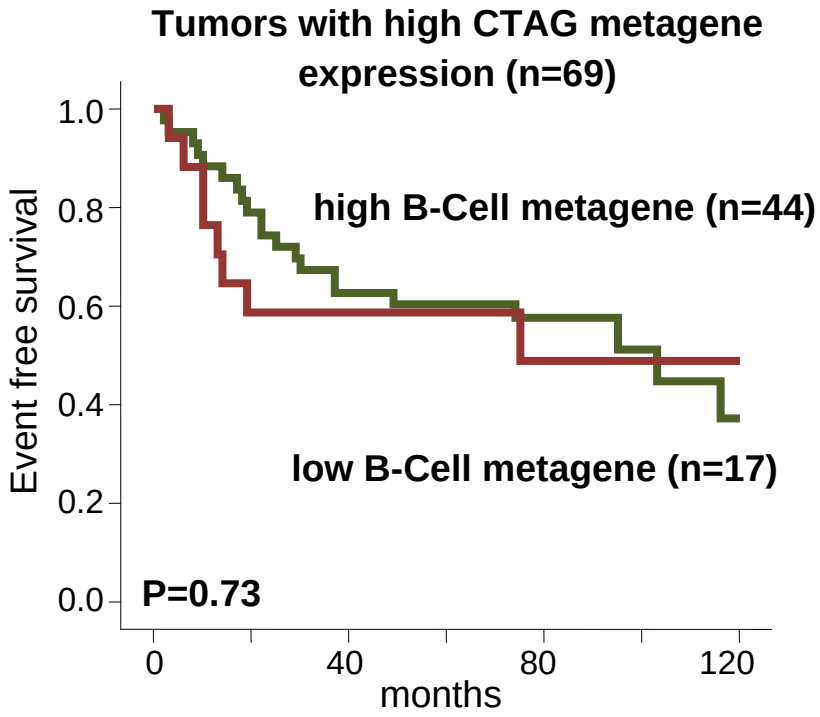
Supplementary Figure S8: Selection of a cutoff for dichotomizing of the B-Cell metagene

Univariate Cox regression analysis of event free survival was performed in the cohort of n=297 TNBC with follow up data using a dichotomized B-Cell metagene. The results of different cutoff values in steps of 0.001 are shown. The upper panel shows the number of samples in the two groups according to the used cutoff. The middle and lower panels show the P-Value and hazard ratio of the respective univariate Cox regression according to the applied cutoff. A cutoff of 0.005 (marked by a red arrow) was selected for all subsequent analyses which concurrently displayed (i) a low P-Value and (ii) mostly equally sized sample groups.

A



B



Supplementary Figure S9: **Prognostic value of B-Cell metagene expression in TNBC separately according expression of the CTAG metagene**

Kaplan-Meier analysis of event free survival of TNBC patients without (*A*) or with high CTAG metagene expression (*B*), respectively. Samples were stratified according to the expression of the B-Cell metagene as a surrogate marker for lymphocyte infiltration.