Supplementary Table 1: List of mouse and human primers used in this study.

Supplementary Table 2: List of the splicing variants that were predicted to be differentially expressed by performing paired comparisons of primary tumors. The Official gene symbol, Official gene name, and several pieces of information (Entrez Gene ID, GenBank Accession, RefSeq ID, UniProt…) are provided to allow retrieving each gene in various databases. In addition, the FASTDB gene ID and the differentially expressed alternative exon(s) as annotated in FAST DB (http://www.fast-db.com/) are provided to allow retrieving the mouse alternative exons that are differentially expressed in the 4T1 mouse model of tumor progression. Finally, the comparison in which each alternative exon is predicted to be differentially expressed is indicated.

Supplementary Table 3: List of the splicing variants that were predicted to be differentially expressed by performing a primary tumor group comparison. Similar information as that described in supplementary Table 2 is provided. In addition, the FASTDB gene ID and the differentially expressed alternative exon(s) as annotated in FAST DB (http://www.fast-db.com/) are provided, as well as the "Mouse Transcript IDs" to allow retrieving the mouse alternative exons and splicing variants that are differentially expressed in the 4T1 mouse model of tumor progression. Fold change and associated p-value for each predicted differentially expressed exons are provided, as well as whether the corresponding exons are already annotated as alternative exons in FAST DB ("FAST DB annotation").

In addition, a cross indicates conserved events that correspond to events annotated in FAST DB for both mouse and human ("Conservation"). The transcript IDs of the conserved

human splicing variants ("Human transcript ID") are provided to allow the analysis of these transcripts in human samples.

Furthermore, a cross indicates tissue-specific events ("Tissue Specificity"). For that purpose, we used the same statistical group analysis on the splicing index values as described for the 4T1 model to search for alternative exons expressed in a tissue-specific manner thanks to the data set collected by Affymetrix (www.affymetrix.com/support/technical/sample_data/exon_array_data.affx/). Among the 78 genes with differentially expressed exons between tumors that disseminated (4TO7, 4T1) or not (67NR, 168FARN) into the lungs, 30 were exons differentially expressed in a collection of normal tissues.

Supplementary Figure S1: Immediately following surgery, the tumor samples were flash frozen and stored in liquid nitrogen until RNA extraction. The samples were examined histologically for the presence of at least 60% tumor cells. The patients (mean age, 59 yrs; range, 35-91 yrs) met the following criteria: primary unilateral breast carcinoma for which complete clinical, histological, and biological data were available; no radiotherapy or chemotherapy before surgery. The histological type and steroid-hormone receptor status of each tumor, as well as the number of positive axillary nodes, were established at the time of surgery. The series consisted of 70% estrogen receptor-positive and 61% lymph node-positive tumors. The malignancy of infiltrating carcinomas was scored according to Bloom and Richardson's histoprognostic system as follows: 15 grade I, 49 grade II, and 34 grade III tumors. The median follow-up was 8.2 years (range 1.4-16.2). Thirty-two patients relapsed at distant sites.

Supplementary Figure S2: For each tumor type, the number of analyzed mice, the number of animals with visible metastasis, the mean number (and range) of lung metastasis, and the median size (and range) of lung metastasis are indicated.

Supplementary Figure S3: Percentage of common events that were found in different comparisons involving the 67NR, the 168FARN, the 4T07, or the 4T1 samples.

Many alternative events of the 67NR vs. 4T07 comparison were found in the 67NR vs. 4T1 comparison. Indeed, 19 % of the alternative events involving the 67NR samples were common to the 67NR vs. 4T07 and the 67NR vs. 4T1 comparisons, but not to the 67NR vs. 168FARN comparison (67NR, Supplementary Figure 3). Meanwhile, only 4 % of the events were common to the 67NR vs. 168FARN and 67NR vs. 4T07 comparisons and only 2 % of the events were common to the 67NR vs. 168FARN and 67NR vs. 4T1 comparisons (67NR, Supplementary Figure 3).

Similarly, 19 % of the events involving the 168FARN samples were common to the 168FARN vs. 4T07 and the 168FARN vs. 4T1 comparisons but not to the 67NR vs. 168FARN comparison (168FARN, supplementary Figure 5), while only 3 % of the events were common to the 67NR vs. 168FARN and 168FARN vs. 4T07 comparisons and only 6 % of the events were common to the 67NR vs. 168FARN and 168FARN vs. 4T1 comparisons (168FARN, Supplementary Figure 3). Furthermore, 21 % of the events involving the 4T07 were common to the 67NR vs. 4T07 and 168FARN vs. 4T07 comparisons (4T07, Supplementary Figure 3) and 26 % of the events involving the 4T1 were common to the 67NR vs. 4T1 and 168FARN vs. 4T1 comparisons (4T1, Supplementary Figure 3). These observations supported the possibility that a large proportion of alternatively expressed exons may be common to the 67NR and 168FARN when compared to the 4T07 and 4T1 samples.

Supplementary Figure S4: Number of genes and range of *p*-values used with the Ingenuity Pathways Analysis “High Level Functions” feature to identify genes that were most significant to the data set.

Supplementary Figure S5:

A. RT-PCR analysis of splicing events for the RAI14, ADD3, FN1, ECT2, TPM2, CALU, MYO1B, HISPPD1, SSR3, SLC38A2, and ADAM33 genes as described in Figure 2. The RT-PCR results are representative of a second set of experiments out of three independent RT-PCR experiments performed with a different preparation of primary tumors.

B. RT-PCR analysis of alternatively spliced exons that were predicted to be differentially expressed in the (67NR, 168FARN) group compared to the (4T07, 4T1) group as described in Figure 3B. The RT-PCR results are representative of a second set of experiments out of three independent RT-PCR experiments performed with a different preparation of primary tumors.

Supplementary Figure S6: RT-PCR analysis of splicing events using RNAs prepared from the cell lines (67NR, 168FARN, 4T07, and 4T1) instead from the primary tumors described in Figure 2.

Supplementary Figure S7: Analysis of the *CLSTN1* gene. The intensity of the probes corresponding to CLSTN1 exon 11 (red bars) was higher in the 67NR compared to the 4T1 samples, while the intensity of the other CLSTN1 probes (dark bars) was similar in both samples. RT-PCR analysis demonstrated that, indeed, exon 11 was more frequently included in the CLSTN1 transcripts expressed in the 67NR than in the 4T1 primary tumors. Screen shots are from EASANA/FAST DB.

Supplementary Figure S8: Different numbers of PCR cycles were performed to co-amplify the splicing variants produced by the FGFR2, CDD, STRN3, KIAA1109, EPB41, TMEM16F, and CLSTN1 genes using RNA prepared from primary tumors, as indicated. The number of PCR cycles did not affect the splicing variant ratios.

Supplementary Figure S9: Human EPB41, KIAA1109, TMEM16F, and CLSTN1 alternatively spliced exons were differentially expressed across normal human tissues (1-breast; 2-cerebellum; 3-heart; 4-liver; 5-muscle; 6-spleen; 7-testis). The sequences of the human alternative exons corresponding to the mouse alternative exons differentially expressed in the 4T1 model are provided. As expected from the literature, additional EPB41 alternative transcripts were detected in the muscle, liver, and heart.

Supplementary Figure S10: Kaplan-Meier curves for 15-year outcome in breast cancer patients (n=104) based on splicing variant content of *KIAA1109*, *EPB41*, *TMEM16F*, and *CLSTN1* genes. Metastasis-free survival curves according to the splicing variant content (i.e., level of exon exclusion, exon inclusion, and inclusion:exclusion ratio) of *KIAA1109*, *EPB41*, *TMEM16F,* and *CLSTN1* genes. Abscise axis (months) and ordinate axis (% of metastasis-free survival).

All RT-qPCR reactions were performed using an ABI Prism 7700 Sequence Detection System (Perkin-Elmer Applied Biosystems) and the SYBR Green PCR Core Reagents kit (Perkin-Elmer Applied Biosystems). TATA-box-binding protein (TBP) transcripts were used as an endogenous RNA control, and each sample was normalized on the basis of its TBP content. All single cassette exons involved in the mouse splicing variants of EPB41, KIAA1109, TMEM16F, and CLSTN1 genes were identified in the homologous human genes and the corresponding primers were designed (Supplementary Table 1). To determine whether the splicing variants were associated with patient clinical outcome, metastasis-free survival distributions were estimated by the Kaplan-Meier method. To select cutoff expression levels to classify each patient in the two risk groups, sensitivity and specificity for all splicing variants were explored using a receiver-operating curve (ROC) analysis. Briefly, the area under the ROC curve (with 95% confidence interval [CI]) was calculated, and a test for the null hypothesis that

the area under the curve was 50% was performed. The ROC analysis provided the threshold expression value to balance sensitivity and specificity for detection of life threatening cancer, and this cut point was used in the Kaplan – Meier analysis. Kaplan – Meier curves estimate metastasis-free survival from 0 to 15 years after breast cancer diagnosis. Estimates of survival rates (with 95% CIs) are shown for the population with a high or a low expression level for each splice variant. The significance of differences between survival rates was ascertained using the log-rank test. The linear combination of all the ratios ("All ratios") was calculated as the sum of the KIAA1109 E+/E-, EPB41 E+/E-, CLSTN1 E-/E+, and TMEM16F E-/E+ ratios. The linear combination of all the splicing variants ("All variants") was calculated as the sum of weighted expression signals of all variants with their Cox's regression coefficient as the weight. All variants values were calculated based on the following formula:

All variants = $[A + \sum_{i=1}^{8} w_i x_i]$, where A is a constant, w_i is the standardized Cox's regression coefficient for the variants and x_i is the expression value of the variant (log scale). The threshold was determined from the ROC curve to ensure the highest sensitivity and specificity. The constant value A was chosen to center the threshold of the risk index to zero. Patients with positive values were classified into the high risk group and patients with negative values are classified into the low risk group.

Supplementary Figure S11: Kaplan-Meier curves for 15-year outcome in breast cancer patients (n=104) based on splicing variant content of *TPM2*, *FN1*, and *HISPPD1* genes. Metastasis-free survival curves according to the splicing variant content (i.e., level of exon exclusion, exon inclusion, and inclusion/exclusion ratio) of *TPM2*, *FN1*, and *HISPPD1* genes. Abscise axis (months) and ordinate axis (% of metastasis-free survival).