Published in final edited form as:
*Mol Cancer Res*. 2010 July ; 8(7): 961–974. doi:10.1158/1541-7786.MCR-09-0528.

# Exon-level microarray analyses identify alternative splicing programs in breast cancer

**Anna Lapuk**[1,*], **Henry Marr**[1], **Lakshmi Jakkula**[1], **Helder Pedro**[6], **Sanchita Bhattacharya**[1], **Elizabeth Purdom**[3], **Zhi Hu**[1], **Ken Simpson**[5], **Lior Pachter**[6], **Steffen Durinck**[1], **Nicholas Wang**[1], **Bahram Parvin**[1], **Gerald Fontenay**[1], **Terence Speed**[3,5], **James Garbe**[1], **Martha Stampfer**[1], **Hovig Bayandorian**[7], **Shannon Dorton**[1], **Tyson A. Clark**[2], **Anthony Schweitzer**[2], **Andrew Wyrobek**[1], **Heidi Feiler**[1], **Paul Spellman**[1], **John Conboy**[1], and **Joe W. Gray**[1,4]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, One cyclotron Road, Berkeley, CA 94720, USA

[2]Affymetrix Inc., 3420 Central Expy, Santa Clara, CA 95051, USA

[3]Department of Statistics, University of California at Berkeley, 367 Evans Hall #3860, Berkeley, CA 94720–3860, USA

[4]University of California San Francisco, 2340 Sutter St., San Francisco, CA 94143, USA

[5]The Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria, 3050, Australia

[6]Department of Mathematics, University of California at Berkeley, 1081 Evans Hall Berkeley, CA 94720-3840

[7]Department of Physics, University of California at Berkeley, 366 LeConte Hall MC 7300 Berkeley, CA 94720-7300

## Abstract

Protein isoforms produced by alternative splicing (AS) of many genes have been implicated in several aspects of cancer genesis and progression. These observations motivated a genome-wide assessment of AS in breast cancer. We accomplished this by measuring exon level expression in 31 breast cancer and nonmalignant immortalized cell lines representing luminal, basal and claudin-low breast cancer subtypes using Affymetrix Human Junction Arrays (HJAY). We analyzed these data using a computational pipeline specifically designed to detect AS with a low false positive rate. This identified 181 splice events representing 156 genes as candidates for AS. RT-PCR validation of a subset of predicted AS events confirmed 90%. Approximately half of the AS events were associated with basal, luminal or claudin-low breast cancer subtypes. Exons involved in claudin-low subtype-specific AS were significantly associated with the presence of evolutionarily conserved binding motifs for the tissue-specific Fox2 splicing factor. siRNA knockdown of Fox2 confirmed the involvement of this splicing factor in subtype specific AS. The subtype specific AS detected in this study likely reflects the splicing pattern in the breast cancer progenitor cells in which the tumor arose and suggests the utility of assays for Fox-mediated AS in cancer subtype definition and early detection. These data also suggest the possibility of reducing the toxicity of protein-targeted breast cancer treatments by targeting protein isoforms that are not present in limiting normal tissues.

**Corresponding authors:** Joe Gray. Life Sciences Division, Lawrence Berkeley National Laboratory, One cyclotron Road, Berkeley, CA 94720, USA. Phone (510) 495-2438, Fax (510) 495-2535, JWGray@lbl.gov, Anna Lapuk. Vancouver Prostate Centre, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada. Phone (604)875-4111 x62445, alapuk@prostatecentre.com.
*present address Vancouver Prostate Centre, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada

### Keywords

breast cancer subtypes; alternative splicing; Fox2 splice factor; exon-level junction microarrays; Illumina sequencing

## Introduction

Breast cancer is a heterogeneous disease that demonstrates considerable variability in response to existing therapies. Recent advances in genome characterization and transcriptome profiling techniques have defined distinct molecular subtypes of breast cancer that differ in biological characteristics and clinical outcome. Subtypes defined through analysis of transcriptional profiles have been designated basal, luminal A and luminal B, *ERBB2*, normal (1,2) and more recently, claudin-low(3,4). Like basal tumors, claudin-low tumors are generally triple negative (*ER*−, *PR*−, *ERBB2*−). However they uniquely express low levels of tight and adherens junction genes including *Claudin 3* and *E-cadherin* and often highly express markers associated with epithelial to mesenchymal transition (EMT) (3,4). Definition of breast cancer subtypes is important to efforts to improve prognostic and predictive markers and to identify new therapeutic targets. The transcript-level measures of gene expression used for subtype definition so far have been important in these areas but may be incomplete indicators of gene function or cellular phenotype because they fail to account for important differences in RNA structure generated by alternative RNA processing events. Alternative transcription initiation or termination events frequently alter the coding capacity at the N- or C-terminal ends of proteins, whereas alternative pre-mRNA splicing of cassette exons can alter expression of functionally important internal domains. In fact, recent studies of human transcriptome suggest that more than 90% of human genes are processed to produce alternative transcript isoforms via one of these mechanisms (5,6) and it is becoming clear that AS is important in the development of the pathophysiology of many human cancers (7).

Information about AS in cancer comes from cDNA sequencing, exon level microarray analysis and RNA sequencing using massively parallel sequencing techniques (5,6,8). A recent assessment of estrogen receptor positive (*ER*+) breast cancer using high throughput RT-PCR identified a number of alternative splicing events that differed between tumors and normal tissue (9). In this report we assess AS in a panel of 26 breast cancer cell lines representing three different tumor subtypes representing the luminal, basal and claudin-low subtypes in primary breast tumors (3,10) and 5 non-malignant breast cell lines. We assessed AS by computational analysis of exon level expression profiles measured using the Affymetrix Human Junction Array (HJAY) technology (11,12) and we applied RT-PCR and deep RNA sequencing for validation. These studies identified 156 AS genes including ~40% for which splicing differed between the luminal, basal and claudin-low transcriptional subtypes. Analysis of the genomic context of alternatively spliced exons suggested Fox1/Fox2 family proteins as regulators of AS. Together these observations suggest the existence of a subtype specific AS program in breast cancer that may be exploited therapeutically and diagnostically.

## Methods

### Cell lines collection

Breast cancer cell lines used in this study were obtained from the ATCC or from collections developed in the laboratories of Drs. Steve Ethier and Adi Gazdar and have been carefully controlled for quality and identity as described in (10). We analyzed alternative splicing using microarrays in 26 breast cancer cell lines and 5 non-malignant immortalized human

mammary epithelial cells (HMEC) cultured as described previously (10,13). The 26 breast cancer cell lines were comprised of 13 having transcription level profiles similar to luminal breast cancers, 6 classified as basal and 7 classified as claudin-low. Five cell lines in our collection (184A1, 184B5, MCF10A, MCF10F, and MCF12A) represented non-malignant immortalized cell lines derived from abnormal but not cancerous tissues.

For RT-PCR validation of predicted AS events we expanded the cell line collection to include 48 breast cancer lines of basal, luminal and claudin-low subtypes as well as 4 normal finite lifespan HMEC strains. The latter included 184D, 48RT, and 240LB strains, which were obtained from reduction mammoplasty tissues and were shown to have mixed, predominantly basal phenotypes, and the 250MK derived from aspirated milk fluids. Garbe *et al* have demonstrated that 250MK cells express luminal markers and thus represent the luminal subtype (14).

### Affymetrix Human Junction Array design and data processing

Genome-wide, exon-level expression and alternative splicing were analyzed using Affymetrix GeneChip® Human Junction Arrays (HJAY, a noncommercial format in collaboration with Affymetrix (Santa Clara, California)). The HJAY array platform was designed using content from ExonWalk (C. Sugnet), Ensembl, and RefSeq databases (NCBI build 36). It interrogates ~315,000 human transcripts from ~35,000 genes and contains ~260,000 junction (JUC) and ~315,000 exonic (PSR) probe sets. A fraction of probe sets had non-unique locations in the human genome and were likely to give cross-hybridization signal. These were excluded from our analysis. In total 501,557 of probesets from 23,546 transcript clusters were retained. Transcript clusters were assigned to known genes using database table refFlat.txt of the UCSC Genome Browser (http://genome.ucsc.edu/). The HJAY data was pre-processed using Affymetrix Expression Console™ Software. Probe set level expression measurements were generated from quantified Affymetrix image files (".CEL" files) using the RMA algorithm (15). The cell lines in the collection were analyzed simultaneously creating a data matrix of probe sets log2 expression values in each cell line. Transcript level expression levels were generated by averaging exonic probesets (PSRs) measurements in that cluster.

### Affymetrix Microarray Profiling

HJAY profiling of cell lines was performed using the GeneChip® Whole Transcript (WT) Sense Target Labeling Assay Kit (Affymetrix). An initial step to remove ribosomal RNA was used to minimize background and to increase detection sensitivity and specificity. Ribosomal RNA subtraction was conducted using a protocol that was modified by Affymetrix for the RiboMinus Transcriptome Kit (Invitrogen). Diluted poly-A RNA controls and RiboMinus probe (in a betaine-containing hybridization buffer) were added to 2ug of total RNA from each sample, incubated at 70C for 5min and then cooled on ice. RiboMinus magnetic beads, prepared by a batch method, were added to the samples and incubated at 37C for 10min. The beads containing the rRNA were isolated using a magnetic separator and the supernatant was transferred to a fresh tube. The beads were washed, separated, and the supernatant was added to the tube. IVT cRNA cleanup columns (Affymetrix) were used to concentrate the subtracted RNA to a volume of 9.8ul. Probe synthesis, oligonucleotide array hybridization and scanning were performed according to the standard Affymetrix GeneChip® protocol for the WT Sense Target Labeling Assay with Control Reagents (rev. 2). Double-stranded cDNA was synthesized with random hexamers tagged with a T7 promoter sequence and used as a template in the presence of T7 RNA polymerase to produce cRNA. In the second cycle of cDNA synthesis, random hexamers were used to reverse transcribe cRNA from the first cycle, producing single-stranded DNA (ssDNA) in the sense orientation. The ssDNA was fragmented by the uracil DNA

glycosylase and apurinic/apyrimidinic endonuclease 1 which recognizes the dUTP incorporated in the ssDNA during the second-cycle, first-strand reverse transcription reaction, and breaks the DNA strand. The fragmented ssDNA was labeled with terminal deoxynucleotidyl transferase and a DNA labeling reagent that is covalently linked to biotin. The fragmented, biotinylated ssDNA probes (5.5ug) were hybridized in a volume of 220ul at 45°C for 16 hours to Affymetrix high density Human Junction Arrays. The arrays were washed and stained with streptavidinphycoerythrin (SAPE, final concentration 10 μg/ml). Signal amplification was performed using a biotinylated anti-streptavidin antibody. The arrays were scanned on an Affymetrix GeneChip® Scanner 3000 7G scanner with an autoloader, according to the Affymetrix GeneChip® WT Sense Target Labeling Assay protocol for the GeneChip® Exon 1.0 ST array. Scanned images were inspected for the presence of obvious defects (artifacts or scratches) on the array; none were detected. The raw and processed expression and splicing data is available at the ArrayExpress data repository with accession number E-MTAB-183.

### Detection of differential splicing using microarray data

We developed an analysis pipeline to detect alternately spliced probesets in microarray data as follows:

**Data filtering—**Probeset and gene level expression data were filtered to remove noisy data that might contribute to false positives. Briefly, we required: (a) expression above the background in at least 25% of samples; (b) probeset exhibiting differential expression; (c) at least 3 probe sets per transcript cluster exhibiting expression above the background to ensure correct FIRMA linear model fitting (see below).

**Splicing Index—**We calculated a Splicing Index (SI) using filtered probeset level expression data according to the formula:

$$SI_{i,j,k} = \frac{e_{i,j,k}}{g_{j,k}}$$

where $e_{i,j,k}$ is the expression level of the $i$ probeset, in experiment $j$, within the $k^{th}$ transcript cluster. $g_{j,k}$ is the transcript cluster expression level estimate of the $j$ experiment and $k$ transcript cluster calculated as the mean of expression of its probesets in a given experiment (sample). Transcript clusters differed substantially in the level of SI variation of their probe sets. Clusters with high variation in many probesets were unlikely to represent clear alternative splicing cases. To avoid prioritization of such clusters in the downstream selection process we converted the splicing index to a Normalized Splicing Index (NSI) data according to the formula:

$$NSI_{i,j,k} = \frac{SI_{i,j,k}}{\tilde{N}_k}$$

where $\tilde{N}_k$ is a measure of a transcriptional noise within a given transcript cluster within a sample set. It is estimated as a median of the following values:

$$N_{i,k} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(e_{i,j,k} - \mu_{j,k}\right)^2}$$

where $N_{i,k}$ is the SD of expression of a probeset *i* within a transcript cluster *k*, *m* is the number of samples, $\mu_{j,k}$ is the mean of $e_{i,j,k}$.

**Selection of highly variable probesets—**NSI data was used to select probesets with the highest variability among cell lines due to alternative splicing. We computed a variability score for a probeset *i*, from transcript cluster *k* using standard deviations of NSI and expression of probesets across samples according to the formula:

$$variability\ score_{i,k} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(NSI_{i,j,k} - \overline{NSI_{j,k}}\right)^2} * \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(e_{i,j,k} - \mu_{j,k}\right)^2}$$

where m is the number of samples, $\overline{NSI_{i,j}}$ and $\mu_{i,j}$ are the means of $NSI_{i,j,k}$ and $e_{i,j,k}$ respectively. We selected probesets with the highest 1% of variability scores. This conservative selection strategy likely misses many valid AS events however, it increases the probably that AS calls are valid. In total 2783 probesets from 1760 transcript clusters passed this cutoff.

**Selection of statistically robust AS probesets using FIRMA—**We used a minor modification of the program FIRMA (16), to assess robustness of the highly variable probesets selected as described above. FIRMA tests the consistency of expression pattern of all probes within a transcript cluster within sample set. For each gene FIRMA fits the following additive model, to background corrected and normalized log2(PM) values:

$$\log_2(PM_{ij}) = c_i + p_j + e_{ij}$$

where $c_i$ is the chip effect (expression level) for chip i, $p_j$ is the probe effect for probe $_j$, $e_{ij}$ is a random error and log2($PM_{ij}$) is the log (base2) of the background–corrected, normalized perfect match (PM) signal for probe *j* on chip *i*.

The model is fitted using iteratively reweighted least squares (IRLS) (17), as implemented in the R function rlm. rlm returns parameter estimates, weights and residuals at convergence. The weights and residuals can be used for detection of probe sets that behave inconsistently with other probesets within a transcript– likely due to differential splicing. Instead of using residuals as described in the Purdom et al (16) we used weights that ranged from 0 (strong evidence of AS) to 1 (no AS). Based on the careful observation of the data, we set an arbitrary cut off of $w_{i,j} \leq 0.7$ for an indication of alternative splicing taking place at a given probe set in a given sample. To generate FIRMA weights data for pre-selected highly variable probesets, we ran the FIRMA algorithm using R function rlm on a subset of 78,050 probe sets. These were all unique probesets from the 1760 candidate transcript clusters described above.

**Selection of the best AS candidate events—**The 2783 most variable probesets were further filtered to select most reliable alternatively spliced events within the breast cancer cell lines. Out of 2783 probe sets, we retained those that had: (a) FIRMA weights $w_{i,j}$ of ≤0.7 in at least 10% of cell lines; (b) average probeset expression of $\log_2 \geq 5.9$; and (c) belonged to a transcript cluster mapped unambiguously to a known gene. We grouped the remaining probesets based on the location within the genome. Probesets located within 150bp from each other, were considered to describe the same AS event. This process yielded 392 probe sets belonging to 181 alternative splice events of ≥2 probesets within 156 known genes.

### Selection of differentially expressed transcript clusters using microarray data

We used the transcript cluster level expression data to identify genes that were differentially expressed among the breast cancer cell lines. Clusters with consistently low expression in all cell lines were excluded from the analysis (expression value cut off was $\log_2$=3). We measured the variability of expression for every transcript cluster on the array by calculating the CV of expression among cell lines. We ranked transcript clusters according to these values and selected those with the highest CVs. This yielded 224 known genes with CV ≥ 0.3.

### Pathway, protein networks and GO terms enrichment analysis

We used the DAVID Functional Annotation tool (18,19) and Ingenuity Pathway Analysis (IPA) software (Ingenuity®Systems, www.ingenuity.com) to perform enrichment analysis for 156 best AS genes and 224 top differentially expressed genes. GO enrichment was also performed with DAVID with a Benjamini Hochberg adjusted p-value cut off of ≤0.01. The Ingenuity knowledge base includes an extensive library of well characterized signaling and metabolic pathways and was used for pathway and network enrichment analysis. Of the 156 AS candidate and 224 differentially expressed genes, 140 and 222, respectively were well annotated in IPA 6.0 database and used for comparative pathway enrichment analysis. Human Genome U133A array data was used as a reference to rank top statistically significant over-represented canonical signaling and metabolic pathways in both sets. Fisher's exact t-test was applied to examine the statistical over representation of pathways, using a threshold of Benjamini Hochberg adjusted p-value ≤ 0.05. Further, the genes of interest were overlaid onto the global molecular network developed from information contained in Ingenuity's Knowledge Base to identify networks that were significantly enriched. These networks are algorithmically generated based on their connectivity. These networks were analyzed further to identify the major nodes (genes in each network with the highest number of interactions with other genes) and the functions associated with the genes in the network. The Functional Analysis of a network identified the biological functions that were most significant to the molecules in the network. The network molecules associated with biological functions in Ingenuity's Knowledge Base were considered for the analysis. Right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function assigned to that network is due to chance alone.

### Nucleic Acid Isolation

Total RNA was extracted from cells grown in 10cm dishes under the standard conditions for each cell type, using the RNeasy Mini Kit (QIAGEN Inc., Valencia, CA). Cell lysis was performed in 600uL RLT Buffer with mechanical shearing; the RNA was recovered according to manufacturer's instructions.

### RT-PCR validation

**Primer design—**Primers were designed using Primer3 software (20) within constitutive exons immediately upstream and downstream of predicted alternatively spliced exons. In instances of alternative terminal exons, a unique primer was designed in each terminal exon and amplified towards a common constitutive exon primer. Validation was performed using a panel of 48 cell lines and 7 primary breast tumors. For primers sequences see Suppl. Materials.

**RT-PCR—**3ug of total RNA from each cell line and 1ug from each primary tumor was used to produce cDNA primed with random hexamers in a 20uL volume using the SuperScript III First Strand Synthesis Kit (Invitrogen) with standard protocols. An aliquot of the resulting cDNA (0.5ul) was used in each 10uL PCR reaction using specific primers. PCR was carried

out with the same conditions for 35 cycles: 35 seconds at 94°C, 35 seconds at 55°C, and 60 seconds at 72°C. The PCR products (4ul) were separated on 5% polyacrylamide gels using a Dual Triple Wide Mini-Vertical Electrophoresis System (CBS Scientific) and imaged.

**Fox2 knock down experiments—**Breast cancer cell lines at 55–65% confluency were transfected with Fox2-specific siRNA (Fox2 On-TARGETplus SMARTpool L-020616-01) or control siRNA (On-TARGETplus Non-Targeting Pool; D-001810-10-05) from Dharmacon. Transfection was performed with 25nM siRNA using Dharmafect I Transfection Reagent according to the manufacturer's instructions (Dharmacon). After 48hrs RNA was isolated as described above. Equal amounts of RNA from control and knockdown cell were analyzed by semiquantitative RT-PCR to evaluate changes in alternative splicing of candidate Fox2-regulated exons.

## Deep sequencing of cell lines transcriptomes using Illumina technology

Total RNA was extracted from MCF7 and BT549 cells using Qiagen's RNeasy Mini Kit. An early version of Illumina's mRNA-Seq protocol was used to prepare the sequencing libraries. For details, see Supplementary Data. 19,553,572 reads of 32bp length were obtained for the BT549 sample and 18,747,831 for MCF7. ELAND was used to align the reads to the HJAY probeset sequences, which were used as a reference. For mapping procedure details see Supplementary Data. Probesets expression level were derived from read count data (number of reads mapped to the reference sequence) by normalizing for the reference sequence length. Since the minimal reference sequence length was 32 bp we defined a normalized read count (NRC) data as:

$$NRC = RC^{*}32/L$$

where RC is the raw read count data and L is the length of reference sequence in bp. NRC data was filtered to remove unreliable low coverage probesets from further analyses. We required, that a reference sequence (a probeset from HJAY array) had the sum of NRC between two cell lines of ≥ 1, which roughly corresponded to 2 mapped reads per average 58 bp JUC probe. Next, NRC data was log2 transformed to make it comparable with the microarray log2 scale expression summaries:

$$Expression = \log_2(NRC+1).$$

Further, expression estimates using both Illumina technology and Affymetrix microarrays were used to generate AS calls. First, log2 expression values for each of the two cell lines have been transformed to SI values. Second, probeset-wise SI score differences have been calculated for the pair of cell lines for each platform. A probeset that had a SI score difference at least one standard deviation away from the mean of SI differences of all probesets in that gene was called AS.

We tested the overall agreement of the two platforms expression profiles and ability to measure differential expression. Between platforms correlations of expression values and $\log_2$ ratios were comparable to those published before for microarray platform comparisons (Spearman correlation of 0.5–0.7 (21) and for microarray-sequencing platforms comparisons (Spearman correlation of 0.73–0.75 (22). Second, we determined if the best 181 splice events could be validated with Illumina technology. This comparison revealed that 59% of well expressed genes were detected with sequencing technology and the set of the best AS candidates was significantly enriched with differential splicing signal from both platforms (Chi-squared test, X2=489.5, df = 1, p-value < 2.2e-16). For details see Supplementary Data.

## Results

### Detection of alternative pre-mRNA splicing in breast cancer cell lines

We measured exon level transcriptional profiles for 5 non-malignant immortalized breast cell lines and 26 breast cancer cell lines including 13 classified as luminal, 6 classified as basal (1,2) and 7 classified as claudin-low (3). The 5 nonmalignant immortalized lines were classified as basal. Exon level expression was assessed using the Affymetrix HJAY platform (11,12) that measured expression levels for ~315,000 known exons and ~260,000 exon junctions (for array design details see Methods). We used a computational pipeline designed to detect hallmarks of probeset expression associated with alternative pre-mRNA splicing events among the cell lines. The pipeline utilized two techniques, Splicing Index (SI)(23) and Finding Isoforms using Robust Multichip Analysis (FIRMA)(16). The SI provided a quantitative measure of differential exon level expression along each gene independent of the transcript level expression. FIRMA assessed differential splicing status of every probeset within a transcript in a statistically robust manner so that it identified those probesets (exons) whose expression pattern across samples did not follow overall transcript expression pattern across the same samples. We included probeset (exon) level and transcript level filtering to remove probesets that exhibited high background, high level cross hybridization and/or weak hybridization (for details see Methods) in order to reduce the false positive rate in AS detection. Finally, we added post-processing to select the top 1% of probesets (exons) showing the most prominent differential splicing among the cell lines that mapped to annotated genes and that exhibited significant AS in at least 10% of the cell lines as measured by FIRMA. This analysis identified 181 splice events (supported by 392 probesets) and implicated 156 known genes as alternately spliced among the cell lines (Supplementary Table 1). Figure 1 shows an example of alternative splicing predicted in the *SLK* (*STE20*-like kinase) gene, as represented by the normalized splicing index (NSI) for exons distributed across the gene. The alternatively spliced exon in the center of the gene is preferentially skipped in cell lines of the claudin-low subtype (depicted by red lines) as indicated by the low splicing index for probesets interrogating the alternative exon itself and its junction with the downstream exon (upper panel), and the relatively higher splicing index for the junction probe interrogating the exon skipping event (lower panel).

We performed hierarchical clustering of the 392 AS probesets to identify recurrent AS patterns among the cell lines. Figure 2 shows three distinct clusters that are mostly concordant with the basal, luminal, and claudin-low subtypes defined previously using hierarchical clustering according to overall gene transcription level (3,4). Non-malignant immortalized mammary epithelial lines clustered together with basal subtype cell lines. Application of the standard Students t-test (Benjamini-Hochberg adjusted p-value $<0.05$) to NSI data for the 392 AS probe sets identified 74 (out of 181) splicing events that correlated with a breast cancer subtype. As expected, reciprocal behavior was observed for probesets representing exon inclusion and exclusion events (Figure 2, cluster D).

We tested the validity of predicted AS using deep sequencing and RT-PCR. Deep sequencing of one claudin-low (BT549) and one luminal (MCF7) cell line was performed to obtain ~19 million 32bp sequence reads from each cell line. Analysis of these data validated ~60% of HJAY microarray-predicted splicing differences among the highly expressed genes (See methods and Supplementary Data). However, the cost of sequencing needed to validate AS in transcripts of moderate or low abundance was prohibitive so we used RT-PCR to test 12 subtype-specific AS predictions across an expanded panel of 7 basal, 14 claudin-low, 18 luminal, 5 non-malignant immortalized breast cell lines and 4 normal finite life span HMEC strains. The AS events selected for verification included cassette exons, alternative 5' and 3' ends, and tandem cassette exons. Table 1 and Figure 3 show that 11 of the 12 AS predictions exhibited splicing differences among the cell lines. In addition to these, 8 other predicted AS

events were used in Fox2 knock-down experiments and 7 of them showed alternative splicing in a smaller panel of 12 cell lines representing the same cancer subtypes (Fig 5B, additional AS events *CLSTN1-57nt*, *CLSTN1-30nt*, *KIF21A*, *PLOD2*, *ST7*, *MARK3*, *ENAH* and *VPS39*). Thus the validation rate of our predictions was ~90% (total of 18 out of 20), which supported the robustness of our computational detection strategy for prediction of AS using HJAY profiles. In general, the claudin-low cell lines consistently showed a different splicing pattern than the basal and luminal subtypes due to differences in the regulated alternative splicing of internal cassette exons.

## Mechanism of Breast Cancer Subtype-specific Splicing

A substantial literature supports the concept that differentiated normal cells execute cell type-specific alternative splicing programs to tailor the structure and function of encoded proteins to the needs of individual cells. One particularly striking observation in this regard is the increased frequency of the binding site, UGCAUG, for the Fox1/Fox2 class of splicing factors, in the introns adjacent to tissue-specific alternative exons in muscle, brain, and erythroid cells compared to introns adjacent to constitutively spliced exons or non tissue-specific alternative exons (11,24–28). The recurring association of this Fox binding site (UGCAUG) with tissue-specific alternative exons suggests an important role in regulation of alternative splicing. These reports motivated our investigation of the possibility that Fox1/Fox2 class splicing factors also influence breast cancer subtype specific splicing.

The importance of Fox1/Fox2 regulation of subtype specific splicing in breast cancer is supported by the plot of the expression levels of Fox2 in Figure 2 that shows that its expression is significantly elevated in basal and claudin-low subtypes compared to luminal subtype cells. We further explored the role of Fox2 class splicing factors in breast cancer subtype specific splicing by assessing the presence of the consensus Fox binding site, UGCAUG, in 22 internal alternative exons for which at least two probesets supported differential alternative splicing between luminal vs. basal and/or claudin-low cells (Supplementary Table 1). Remarkably, Figure 4 shows that 19/22 of these exons possessed one or more UGCAUG binding site in the intron sequences within 400nt of the differentially expressed exons. This frequency greatly exceeded that expected for a random hexamer, which should occur only once every 4kb, or in one out of five such cases. In addition, some exons lacking these more proximal UGCAUG motifs possessed Fox binding sites more distally. In two cases (*ENAH*, *ST7*), clusters of predicted Fox2 binding sites were located in the long downstream intron between 1.4–1.8kb from the regulated exon. While these sites are more distant than most described splicing enhancers, there is precedent for functional Fox sites >1kb from the regulated exon (29).

The functional significance of these associations is further supported by the observation that the Fox binding sites are highly conserved in evolution (Figure 4). In 18/22 cases, orthologs of genes (including the three with distal Fox sites) displaying subtype specific splicing encoded conserved Fox binding sites in the same relative intronic regions. Phylogenetic conservation of intronic UGCAUG motifs in some cases extended not only within mammalian genomes, but also to other vertebrate orders including avian (chicken), amphibian (frog), reptilian (lizard), and fish (zebrafish) species. This result strongly supports the functional importance of Fox binding sites near subtype-associated exons.

To further explore whether Fox2 regulates breast cancer subtype-specific splicing, we asked whether siRNA knockdown of Fox2 altered the splicing efficiency of putative target exons by evaluating the effect on AS of treating a non-malignant immortalized mammary epithelial cell line and cell lines representing the three breast cancer subtypes with siRNAs against Fox2 or an irrelevant siRNA. The Fox2 siRNA consistently reduced expression by 75–80% as determined by qRT-PCR (Figure 5A). Splicing changes induced by Fox2

knockdown were then examined for fourteen subtype-specific alternative exons in four representative cell lines from each of the luminal, basal, and claudin low subtypes (Figure 5B). This analysis showed (a) that the majority of subtype-specific exons predicted to be Fox2-regulated did indeed exhibit differences in splicing efficiency when Fox2 was knocked down, and (b) that splicing responses between subtypes was often distinct, while responses within each subtype were generally very consistent. In most cases (11 of 14 exons), lower Fox2 mRNA expression correlated with reduction in splicing efficiency of the target exons. The relative effects of Fox2 knockdown were large for some exons, such that exon inclusion was almost eliminated (e.g., *CLSTN1*, 57nt). In other cases, effects were modest but reproducible among cell lines representing the same subtype (e.g., *ST7*). Inclusion of one exon (in *FAM62B*) was modestly increased under the same conditions, while two exons (in *FER1L3* and *VPS39*) were little affected. These observations indicate that Fox2 functions predominantly as a splicing enhancer in these cells. Control siRNA did not affect any of these splicing events. We conclude that Fox2 activity plays a major role in regulating a set of breast cancer subtype-specific alternative splicing events, as predicted by the associated consensus binding sites.

Fox2 enhancer activity was detected most frequently in claudin low cells, with eight exons exhibiting marked Fox2-dependence in exon inclusion levels (*CLSTN1*-57 and -30, *KIF21A, PLOD1, ST7, FAT, TJP1, MARK3*). Evidence for Fox2 enhancer effects on the same set of exons was often but not always observed in basal cells even though Fox2 is relatively highly expressed in both in claudin low and basal subtypes (Figure 2). Interestingly, a few exons that were predominantly skipped in claudin low cells paradoxically exhibited Fox2-dependent splicing in luminal cells even though this subtype has lower Fox2 expression. This feature was most evident for *ENAH*, but was also reproducibly observed as a minor effect in *CLTC* and *SLK*. These results indicate that Fox2 cannot be the sole determinant of subtype-specific splicing. Presumably, subtype-specific differences in Fox2-dependent splicing efficiency reflect combinatorial effects of multiple splicing regulators with antagonistic or synergistic activities, each with its own subtype-specific activity profile, together with exon-specific constellations of binding sites for these factors.

### Important pathways and networks affected by alternative splicing

We analyzed 156 alternatively spliced genes detected in breast cancer cell lines using the DAVID functional annotation tool (http://david.abcc.ncifcrf.gov) (30,31) and Ingenuity software (http://www.ingenuity.com) and compared these to 224 genes that were differentially expressed among the cell lines. In general, AS involved different genes than differential expression since only 2 genes were in common between these lists. Analyses of AS genes showed preferential enrichment of biological processes related to cytoskeleton and actin. This held true for the 156 gene list and a subset of 63 genes showing subtype specific AS. The Ingenuity pathway enrichment analysis implicated AS in aspects of signaling involving Axon guidance, Ephrin receptor, Integrin and Tight Junctions (Figure 6, Table 2). The top three protein networks that were highly enriched with AS genes had major nodes involving *MYC*, *Actin* and *EGFR* genes. The main functions associated with the merged network were cytoskeleton organization, biogenesis and cell signaling (Figure 2 in Supplementary Materials). On the other hand, Figure 6 shows that 224 differentially expressed but not alternately spliced genes predominantly influenced aspects of metabolism.

## Discussion

Alternative splicing of pre-mRNAs is now well established as a mechanism for increasing protein functional diversity during normal development. In addition, alternative splicing events (ASEs) have been implicated specifically in breast and ovarian cancer genesis and progression by PCR screening for cancer associated known ASEs listed in a RefSeq

database of known isoforms (28). These studies identified over 200 breast cancer specific ASEs that were spliced differently in *ER*+ tumors vs. normal tissue (32). Moreover, the splicing factor, Fox2, has been associated with normal tissue specific splicing (24) and has been implicated as an important splicing factor in breast and ovarian cancer development (32). We extended these findings here by interrogating alternative splicing in a collection of breast cancer cell lines that exhibit transcriptional programs found in breast cancers classified as luminal (*ER*+), basal (*ER*−) and claudin-low (*ER*− with stem cell like features) using splice-sensitive Affymetrix HJAY microarrays. Our analysis implicated 156 genes as alternately spliced including 63 whose splicing patterns were associated with the luminal, basal and claudin-low subtypes.

In order to understand the role of splicing in global regulation of cellular processes, we performed a pathway, GO terms and protein network enrichment analysis for 156 genes with strongest evidence of alternate splicing across breast cancer cell lines. We observed a significant enrichment for pathways involving axonal guidance, integrin signaling, tight junction signaling, Ephrin receptor signaling, and actin cytoskeleton (Figure 6, Table 2) and GO terms mostly related to cytoskeleton and actin. The involvement of the cytoskeleton in morphology and motility suggests the possibility that alternative splicing plays a significant role in determining phenotypic differences between these breast cancer subtypes (10,33). Interestingly, we found that genes and the biological processes influenced by alternative splicing were different from those influenced by expression regulation. In general, alternative splicing influenced aspects of cell surface protein mediated signaling that affected morphology and motility while differential gene expression seemed to influence metabolism and signaling controlling cell proliferation. This is in line with earlier observations that splicing and transcription regulation mechanisms function in parallel to mediate cellular processes (34,35). The importance of AS in protein function regulation is supported by the theoretical protein structure analysis of Wang et al showing that ~90% of AS regions are located within regions of “loop” secondary structures on the surface of proteins and thus likely mediate protein-protein and protein-ligand interactions (36) In addition, Hughes and Friedman showed that AS genes tend to interact with other AS genes in genetic and protein interaction networks (37). Taken together, our data and the published literature suggest that AS plays an important role in the formation and regulation of protein-protein interactions involved in cell motility and morphology(38).

Expression of Fox2 is associated with subtypes showing up-regulation of expression in claudin-low and basal cells and down regulation in luminal cells. This observation suggests that Fox2 is an important regulator of subtype specific splicing differences between luminal and basal/claudin-low subtypes. Evidence for this includes (a) moderate association between increased Fox2 expression and internal exon cassette inclusion in the non-luminal subtypes (Pearson’s correlation of 0.5), (b) evolutionary conservation of intronic Fox2 consensus binding sites and (c) reduced inclusion of target exons in non-luminal subtype cells after treatment with a siRNA against Fox2. Notably, even the *ST7* alternative exon having a cluster of three distal intronic Fox binding sites ~1.8kb downstream exhibited reduced splicing efficiency when Fox2 expression was reduced by siRNA knockdown. However, it is clear that Fox2 is not the only regulator of subtype specific splicing events, since there was a wide range of Fox2-dependence in splicing efficiency among the tested exons, and a few were insensitive to changes in Fox2 expression. Individual exons in the the subtype-specific splicing programs are likely regulated by a combination of factors with antagonistic and/or synergistic activities that can fine tune the distributions of spliced forms for each subtype. In particular, the recently described epithelial splicing regulatory proteins ESRP1 and ESRP2(39,40) are good candidates for contributing to subtype-specific splicing programs.

A comparison of the 156 alternately spliced genes revealed by our study with the 247 genes listed by Venables *et al* as alternately spliced in breast cancer shows only 10 that are common to both studies. This is not surprising since the experimental designs and sample set composition used in our studies differed from those used by Venables *et al*. For example, Venables *et al* used a sensitive PCR based approach able to detect alternately spliced isoforms present at a level of 10% of the total amount of all transcripts while we used a microarray approach that is relatively insensitive to the presence of low abundance transcripts. In addition, we focused on alternative splicing differences between cell lines derived from both *ER*+ positive and *ER*-negative (*ER*−) breast tumors while Venables *et al* focused on the detection of tumor specific splice events that differed between *ER*+ tumors and normal tissues. As a result, the *ER*+ breast cancer specific splice isoform markers found by Venables and splice specific markers of breast cancer subtypes discovered in our study provide a more comprehensive picture of the role of AS in breast cancer pathophysiology. A combination of the approaches seems appropriate for future splicing studies.

Finally, information about subtype specific alternative splicing of cell surface proteins in breast cancer may have important translational applications. In this study, these included cell surface proteins encoded by the genes *CD47, CLTC, DST, FAM62B, FAT, FER1L3, FLNB, MET, PLEC1, PPFIBP1 and PTPRK*. Molecular assays for specific protein isoforms for these genes may increase the sensitivity and specificity of anatomic and blood based detection of specific breast cancer subtypes. Subtype specific cell surface protein isoforms also are attractive candidate therapeutic targets since agents that specifically attack a cancer specific protein isoforms may have reduced reactivity with other protein isoforms that are expressed in otherwise rate limiting normal tissues. Genes showing strong claudin-low specific alternative splicing including *FLNB* and *FAT* are interesting as targets for stem cell specific therapies considering that the claudin low cells carry many molecular features associated with stem cell function (41). Targeting this breast cancer subtype is highly important given the growing evidence that such cancers may be particularly resistant to conventional therapies. (42). Supplementary Table 1 describes alternative exon usage in alternatively spliced cell surface proteins to guide efforts to develop subtype specific markers and therapies.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

**Grant support:** This work was supported by the Director, Office of Science, Office of Biological & Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, USAMRMC BC 061995, and NIH grants CA58207, CA112970, and CA 126477 (JWG); by NIH grant HL045182 (JGC); and by the FCT SFRH / BD 33203 2007 (HP).

## References

1. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001; 98:10869–10874. [PubMed: 11553815]
2. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747–752. [PubMed: 10963602]
3. Hennessy BT, Gonzalez-Angulo AM, Stemke-Hale K, et al. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. Cancer Res. 2009; 69:4116–4124. [PubMed: 19435916]

4. Herschkowitz JI, Simin K, Weigman VJ, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007; 8:R76. [PubMed: 17493263]
5. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]
6. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40:1413–1415. [PubMed: 18978789]
7. Venables JP. Aberrant and alternative splicing in cancer. Cancer Res. 2004; 64:7647–7654. [PubMed: 15520162]
8. Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A. 2008; 105:20179–20184. [PubMed: 19088194]
9. Venables JP, Klinck R, Bramard A, et al. Identification of alternative splicing markers for breast cancer. Cancer Res. 2008; 68:9525–9531. [PubMed: 19010929]
10. Neve RM, Chin K, Fridlyand J, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006; 10:515–527. [PubMed: 17157791]
11. Yamamoto ML, Clark TA, Gee SL, et al. Alternative pre-mRNA splicing switches modulate gene expression in late erythropoiesis. Blood. 2009; 113:3363–3370. [PubMed: 19196664]
12. Lin L, Liu S, Brockway H, et al. Using high-density exon arrays to profile gene expression in closely related species. Nucleic Acids Res. 2009
13. Garbe JC, Bhattacharya S, Merchant B, et al. Molecular Distinctions between Stasis and Telomere Attrition Senescence Barriers Shown by Long-term Culture of Normal Human Mammary Epithelial Cells. Cancer Res. 2009
14. Garbe JC, Bhattacharya S, Merchant B, et al. Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. Cancer Res. 2009; 69:7557–7568. [PubMed: 19773443]
15. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). 2003; 4:249–264.
16. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics (Oxford, England). 2008; 24:1707–1714.
17. Marazzi, A. Algorithms, Routines and S Functions for Robust Statistics. Wadsworth & Brooks/ Cole; 1993.
18. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]
19. Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3. [PubMed: 12734009]
20. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000; 132:365–386. [PubMed: 10547847]
21. Yauk CL, Berndt ML, Williams A, Douglas GR. Comprehensive comparison of six microarray technologies. Nucleic Acids Res. 2004; 32:e124. [PubMed: 15333675]
22. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18:1509–1517. [PubMed: 18550803]
23. Clark TA, Schweitzer AC, Chen TX, et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol. 2007; 8:R64. [PubMed: 17456239]
24. Das D, Clark TA, Schweitzer A, et al. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic Acids Res. 2007; 35:4845–4857. [PubMed: 17626050]
25. Minovitsky S, Gee SL, Schokrpur S, Dubchak I, Conboy JG. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic Acids Res. 2005; 33:714–724. [PubMed: 15691898]

26. Sugnet CW, Srinivasan K, Clark TA, et al. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol. 2006; 2:e4. [PubMed: 16424921]
27. Zhang C, Zhang Z, Castle J, et al. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. Genes Dev. 2008; 22:2550–2563. [PubMed: 18794351]
28. Castle JC, Zhang C, Shah JK, et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet. 2008; 40:1416–1425. [PubMed: 18978788]
29. Guo N, Kawamoto S. An intronic downstream enhancer promotes 3' splice site usage of a neural cell-specific exon. J Biol Chem. 2000; 275:33641–33649. [PubMed: 10931847]
30. Huang da W, Sherman BT, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID gene ID conversion tool. Bioinformation. 2008; 2:428–430. [PubMed: 18841237]
31. Huang da W, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007; 8:R183. [PubMed: 17784955]
32. Venables JP, Klinck R, Koh C, et al. Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol. 2009
33. Kenny PA, Lee GY, Myers CA, et al. The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. Mol Oncol. 2007; 1:84–96. [PubMed: 18516279]
34. Pan Q, Shai O, Misquitta C, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell. 2004; 16:929–941. [PubMed: 15610736]
35. Le K, Mitsouras K, Roy M, et al. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. Nucleic Acids Res. 2004; 32:e180. [PubMed: 15598820]
36. Wang P, Yan B, Guo JT, Hicks C, Xu Y. Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc Natl Acad Sci U S A. 2005; 102:18920–18925. [PubMed: 16354838]
37. Hughes AL, Friedman R. Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm Caenorhabditis elegans. Genetica. 2008; 134:181–186. [PubMed: 18026854]
38. Blencowe BJ. Alternative splicing: new insights from global analyses. Cell. 2006; 126:37–47. [PubMed: 16839875]
39. Warzecha CC, Shen S, Xing Y, Carstens RP. The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. RNA Biol. 2009; 6:546–562. [PubMed: 19829082]
40. Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Carstens RP. ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. Mol Cell. 2009; 33:591–601. [PubMed: 19285943]
41. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol.
42. Kakarala M, Wicha MS. Implications of the cancer stem-cell hypothesis for breast cancer prevention and therapy. J Clin Oncol. 2008; 26:2813–2820. [PubMed: 18539959]

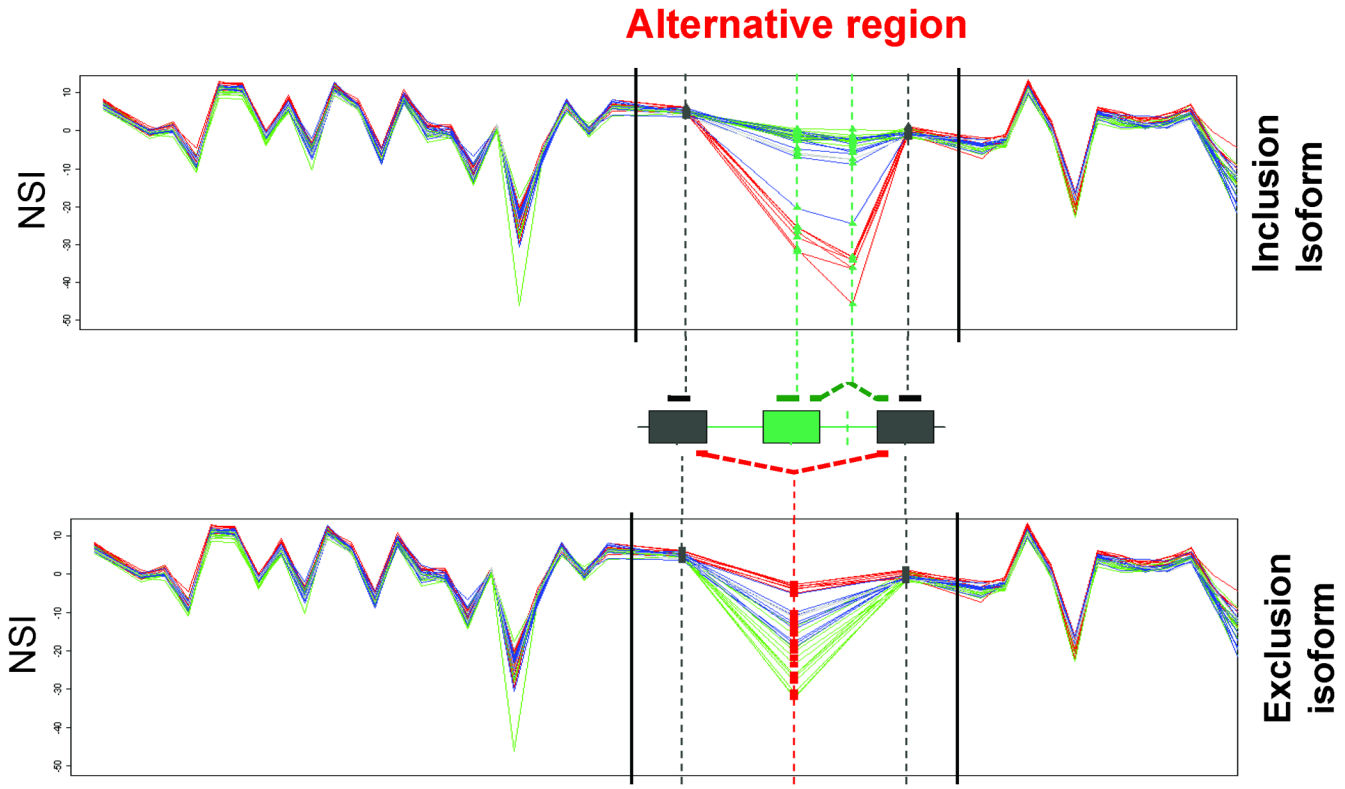



**Figure 1. Alternative Splicing for *SLK***

The probeset NSI data for two isoforms are plotted separately as functions of distance along the genome. Data for the inclusion isoform is shown above the exon-intron cartoon and data for the exclusion isoform is shown below it. The AS region between the vertical black lines involves a single alternative exon. The AS region is expanded to make AS events more clear. Probesets interrogating exonic regions and respective junctions are shown with vertical dotted lines. Subtypes of cell lines are color coded so that claudin-low results are red, basal results are blue, luminal results are green and HMEC results are grey. NSI values show that the exclusion isoform is prominent in claudin-low cell lines and diminished in the luminal and basal subtypes; conversely, the inclusion isoform is diminished in claudin-low and prominent in two other subtypes.

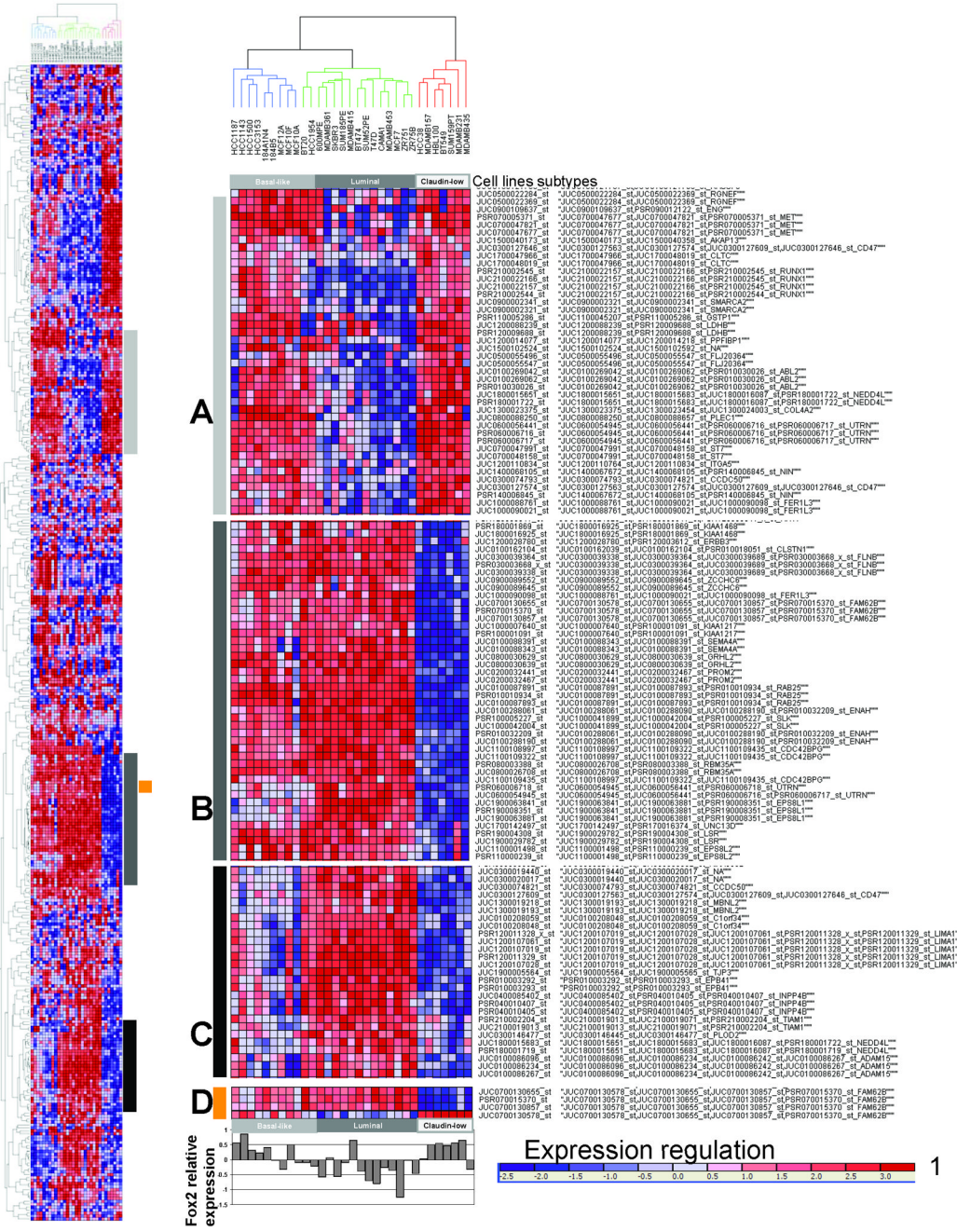



**Figure 2. Unsupervised hierarchical clustering of AS predictions**

392 probeset NSI data from 156 genes have been clustered. Three major gene clusters A, B and C expanded on the right show distinct subtype specific AS patterns. Cluster D at the bottom shows the *FAM62B* gene used for PCR validation. Preferential exclusion of an alternative exon in claudin-low cells is indicated by strong expression of the probeset for the exclusion event (JUC0700130578_st) and relatively lower expression of probesets targeting alternative exon inclusion event. Relative expression of the Fox2 splicing factor measured by HJAY profiling (mean normalized) is shown at the bottom of the figure.

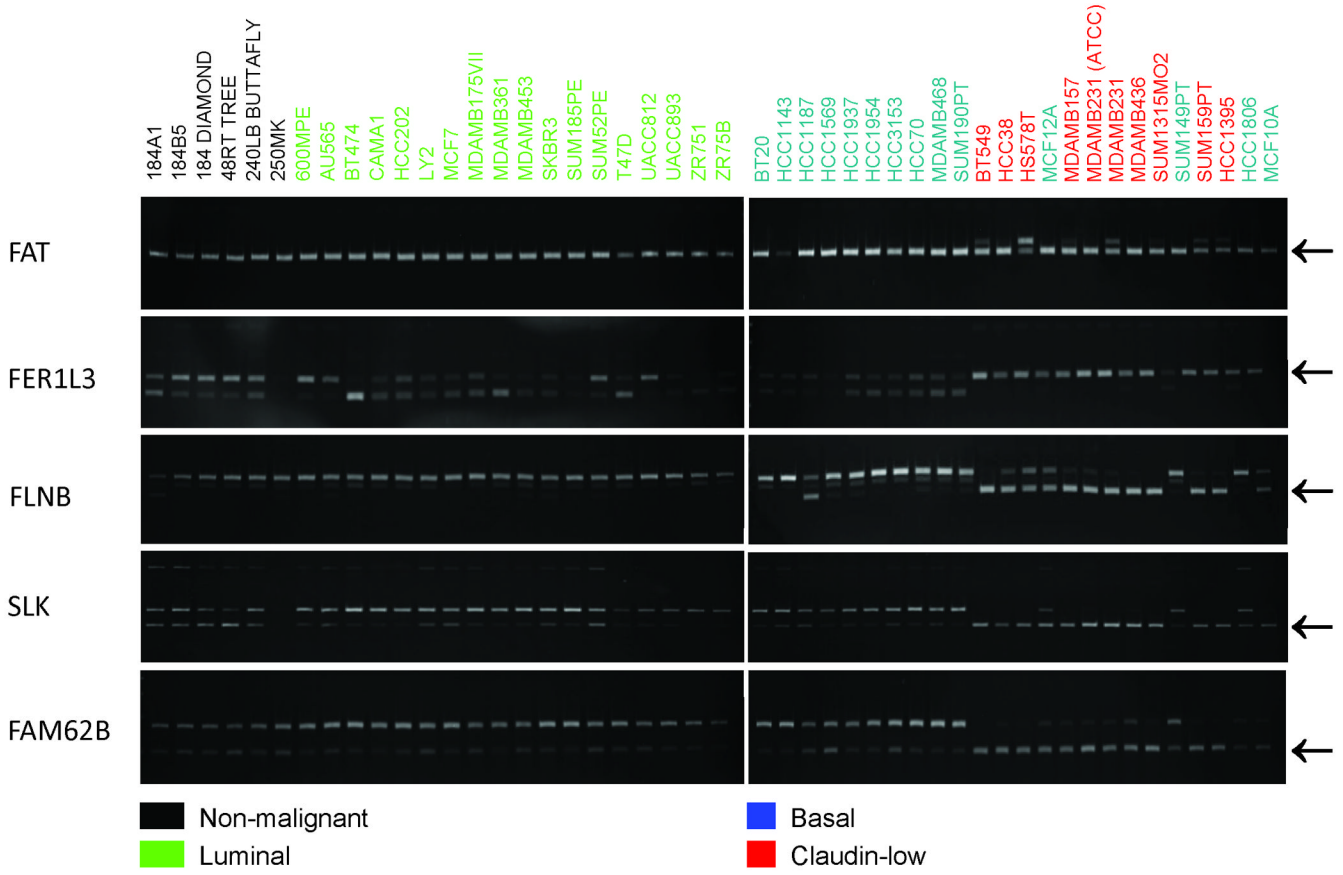



**Figure 3. PCR validation of breast cancer subtype-specific alternative splicing**

Splicing differences predicted by exon junction microarrays for 5 genes shown in the figure were examined by RT-PCR analysis in an extensive panel of breast cancer cell lines. RNA from each cell line was amplified using primers in the flanking constitutive exons and the products were analyzed using gel electrophoresis. The upper PCR band in each analysis corresponds to an exon inclusion isoform while the lower band represents an exon exclusion isoform. Results for these genes show that most claudin-low cell lines were spliced differently than the other subtypes. Arrow indicates the position of the claudin-low-enriched PCR product.

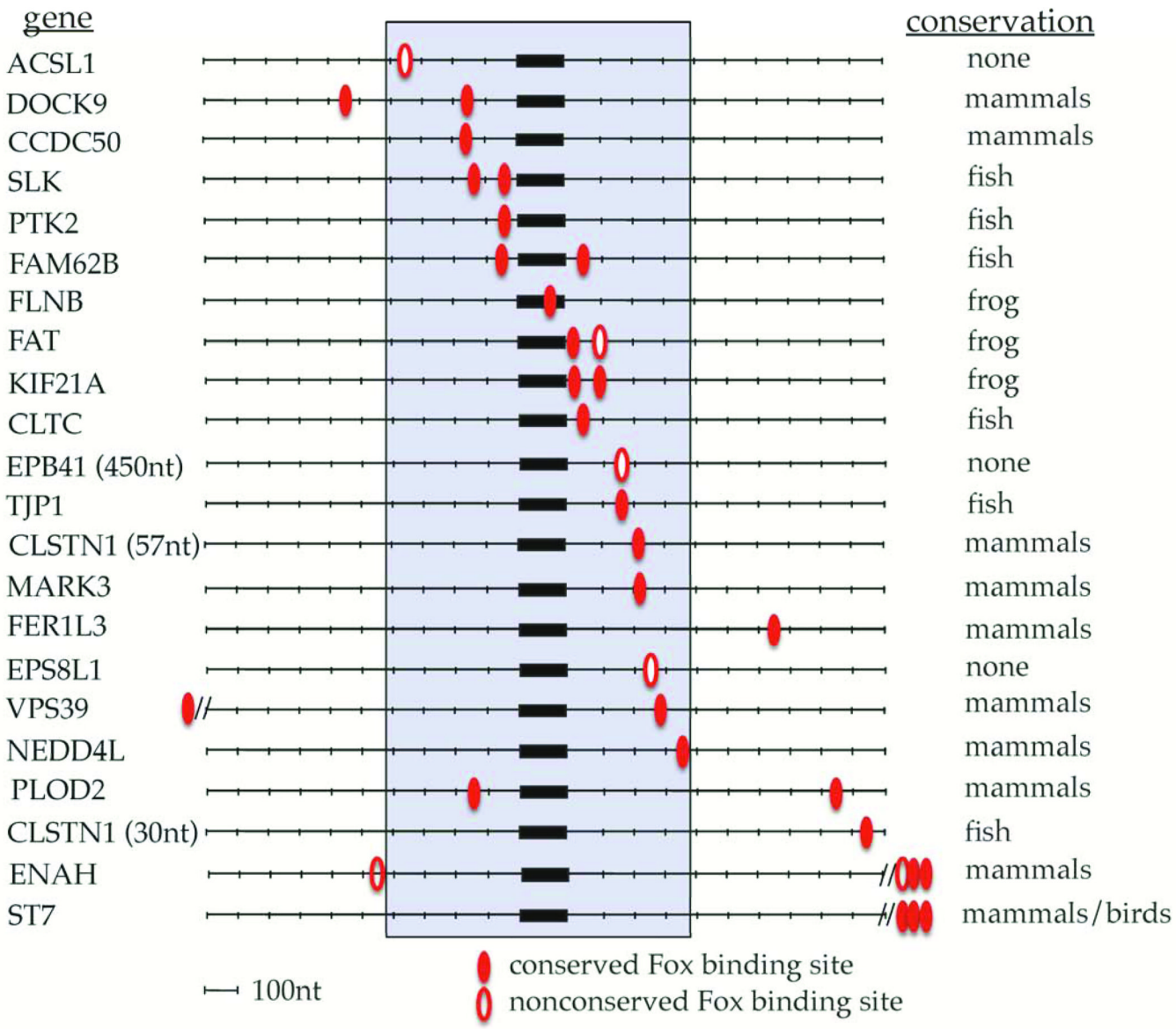
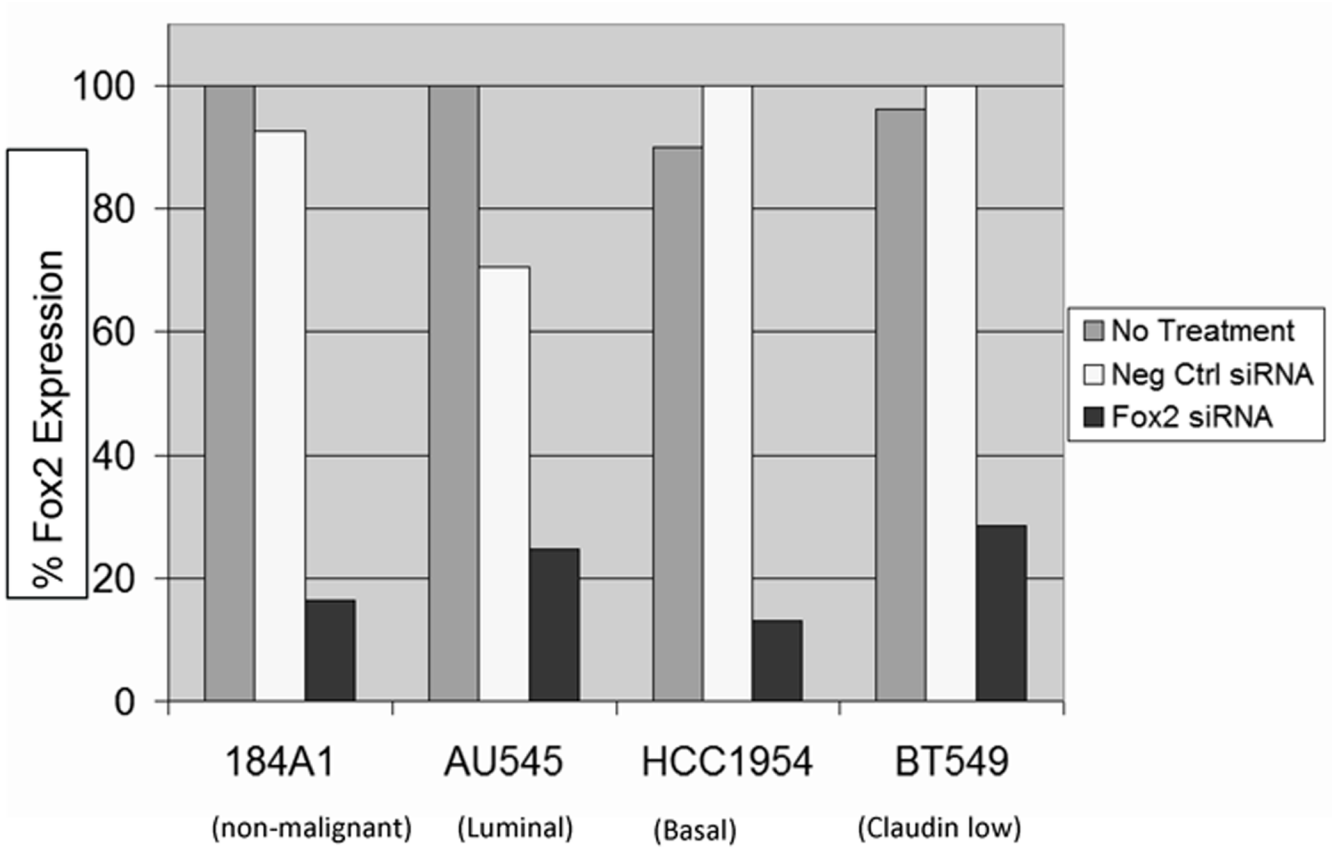



**Figure 4. Frequent association of Fox sites with breast cancer subtype-specific alternative exons**

Diagrams represent the genomic regions spanning 1kb upstream to 1kb downstream of the regulated exons. Locations of evolutionarily conserved occurrences of the Fox splicing regulatory motif UGCAUG are indicated by filled ovals. Non-conserved sites present only in the human genome are indicated by empty ovals. The highest frequency of Fox binding sites occurs within 400nt (shaded) proximal to the intronic region however a few distal sites are also evolutionarily conserved. For example, *ENAH* encodes two sites ~1.9 kb downstream, *ST7* encodes three sites ~1.8 kb downstream (one of which is conserved in birds) and *VPS39* encodes 1 site 1.5 kb upstream.



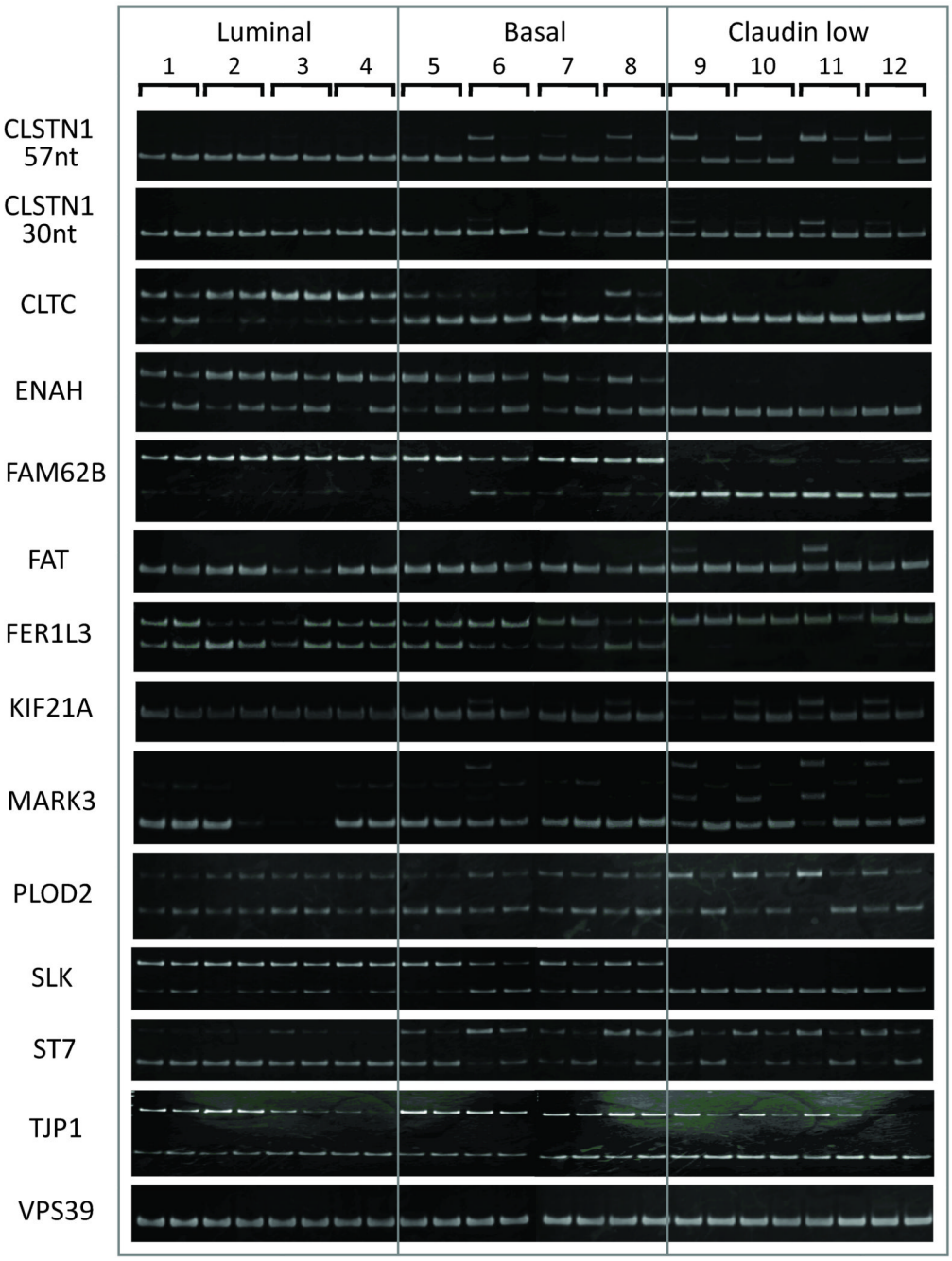



**Figure 5. Knockdown of Fox2 alters splicing in breast cancer cells**

Panel A. Level of Fox2 siRNA knockdown achieved in four breast derived cell lines from non-malignant tissue and three different malignant breast cancer subtypes. Irrelevant siRNA and mock transfection serve as controls. Panel B. RT-PCR analysis of changes in splicing efficiency induced by Fox2 knockdown in four luminal cell lines (LY2, 361, T47D, and ZR75B; samples 1–4, respectively), four basal cell lines (BT20, HCC1569, HCC1954, and HCC3153; samples 5–8) and four claudin low cell lines (BT549, HCC38, HS578T, and 157; samples 9–12). Splicing in each cell line was compared after treatment with a control siRNA (first lane of each pair), vs treatment with the Fox2 siRNA (second lane of each pair). Identity of genes containing the 14 subtype-specific exons is given at the left. Upper PCR

bands represent exon inclusion, while lower bands represent exon skipping products. For *VPS39*, only the inclusion product was detected. For *MARK3*, the amplified region spans two alternative exons of 27nt and 45nt and can generate four products. The 27nt exon is subtype-specific (skipped in luminal but partially included in basal and claudin low) and Fox2-dependent (abundance of the +27+45nt and +27−45nt bands is greatly reduced in claudin low cells after Fox2 knockdown).

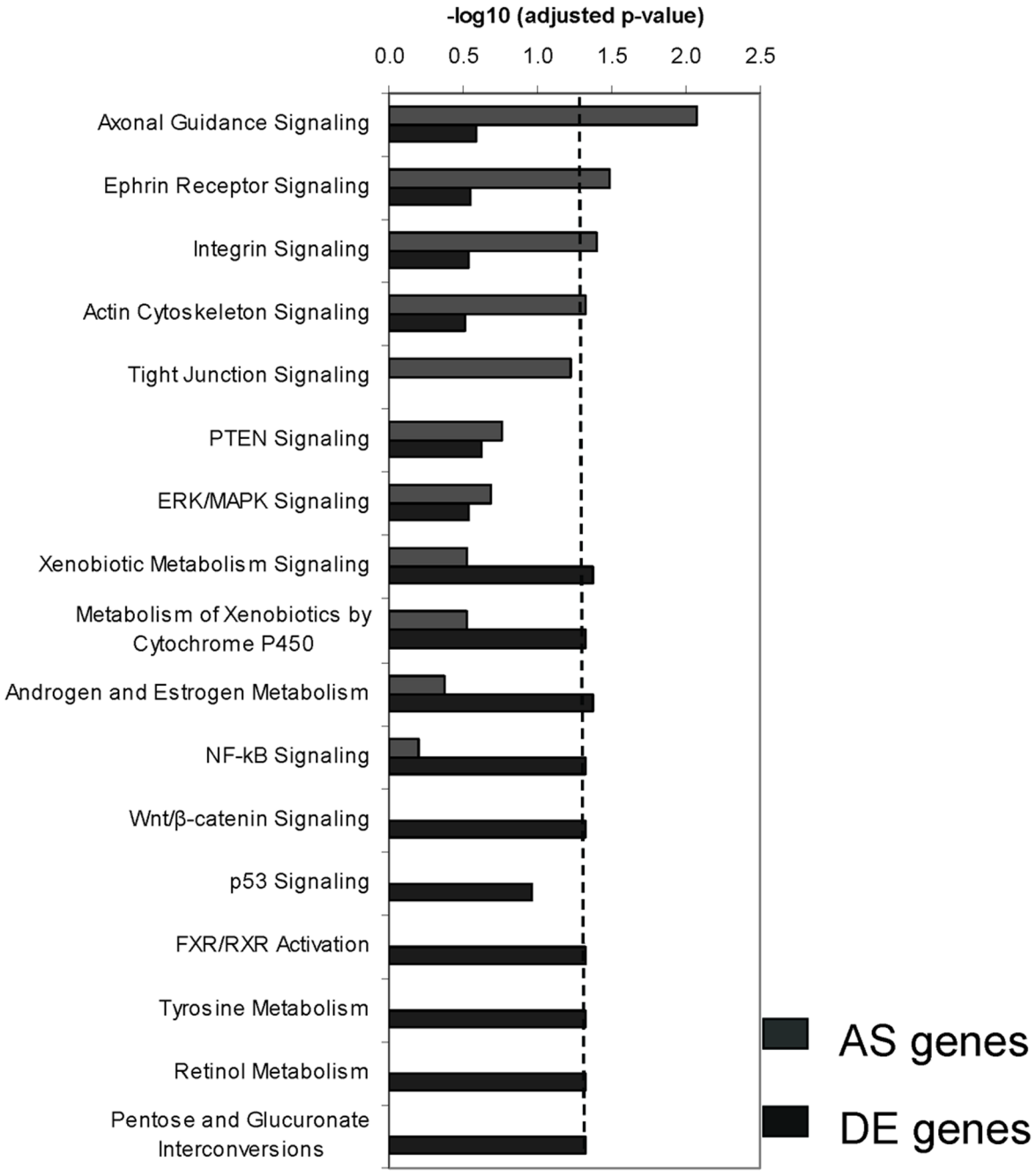



**Figure 6. Comparative Pathway enrichment for AS and DE genes**

One hundred forty (out of 156) AS genes and 222 (out of 224) DE genes were analyzed in Ingenuity IPA 6.0 for pathway enrichment. Pathways enriched with AS genes are shown in grey and with DE genes in black. A p-value cut off of 0.05 is shown with a vertical dotted line.

**Table 1**

**Validation of AS predictions**

Left columns show the result of assessing AS in a panel of 48 cell lines using RT-PCR. Rank is based on the ordered NSI score data across all probe sets on the HJAY array. Specificity refers to the subgroup in which aberrant splicing (compared to the reference) is most frequent (NV is not validated). The ratio of alternative isoforms in the non-malignant mammary gland cell lines were taken as a reference. The AS type column refers to the type of alternative splicing event being tested. Right columns show the result of assessing AS in the cell lines MCF7 (luminal) and BT549 (claudin-low) using Illumina sequencing. The #_of_PS column shows the number of probe sets in microarray and sequencing data describing the same AS event validated using RT-PCR. The last 3 columns summarize RT-PCR, microarray and sequencing data on alternative splicing between two cell lines - MCF7 and BT549. For microarray and Illumina platforms the number of probe sets exhibiting differential signal between the two cell lines is shown.

| PCR validation in cell line panel | | | | Illumina validation in MCF7/BT549 | | | |
|---|---|---|---|---|---|---|---|
| **rank** | **gene** | **specificity** | **AS type** | **#_of_PS** | **PCR** | **Affy_call** | **Illu_call** |
| 117 | **TJP1** | cancer | cassette exon | 3 | **AS** | 0 | 2 |
| 34 | **DST-1** | cancer | alt start | 3 | **no AS** | 0 | 0 |
| 43, 1223* | **PLEC1** | cancer | alt start | 2 | **AS** | 1 | 2 |
| 146 | **DST-2** | cancer | alt stop | 3 | **AS** | 3 | 1 |
| 107 | **FAM62B** | CL | cassette exon | 4 | **AS** | 4 | 4 |
| 58 | **FAT** | CL | cassette exon | 3 | **weak AS** | 3 | 0 |
| 135 | **FER1L3** | CL, LUM | cassette exon | 3 | **AS** | 3 | 2 |
| 1 | **FLNB** | CL | cassette exon | 4 | **AS** | 4 | 3 |
| 6 | **MYO6** | LUM, BA | tandem cassette exons | 6 | **AS** | 3 | 1 |
| 35 | **SLK** | CL, cancer | cassette exon | 3 | **AS** | 3 | 3 |
| 40 | **CLTC** | CL | novel alt exon | 2 | **no AS** | 2 | 0 |
| 170 | **UTRN** | NV | novel alt exon | 4 | **no AS** | 4 | 0 |

## Table 2

Component gene lists from enriched pathways from Figure 6.

| | **Pathway genes from 140 AS gene list** |
|---|---|
| **Pathway Name** | **Component genes from Pathway** |
| Axonal Guidance Signaling | PTK2, SLIT3, PRKACB, GNAS, ADAM15, ITGA2, ITGA5, LIMK2, RASSF5, SEMA4A |
| Integrin Signaling | PTK2, DDEF1, ITGA2, ITGA6, ITGA5, ACTN1 |
| Actin Cytoskeleton Signaling | PTK2, TIAM1, ITGA2, ITGA5, LIMK2, ACTN1 |
| Tight Junction Signaling | PRKACB, EPB41, TIAM1, TJP3, TJP1 |
| PTEN Signaling | PTK2, ITGA2, ITGA5 |
| Androgen and Estrogen Metabolism | STS |
| NF-kB Signaling | PRKACB |
| Ephrin Receptor Signaling | PTK2, GNAS, GRIN2D, ITGA2, ITGA5, LIMK2 |
| ERK/MAPK Signaling | PTK2, PRKACB, ITGA2, ITGA5 |
| Xenobiotic Metabolism Signaling | GSTM1, GSTP1, ALDH3A2, SULT1A1 |
| Metabolism of Xenobiotics by Cytochrome P450 | GSTM1, GSTP1 |
| | **Pathway genes from 222 DE gene list** |
| **Pathway Name** | **Component genes from Pathway** |
| Axonal Guidance Signaling | SEMA3A, SEMA3D, PIK3C2G, AKT3, EPHA3, BMP5, ITGA4 |
| Ephrin Receptor Signaling | PIK3C2G, AKT3, EPHA3, ITGA4 |
| Integrin Signaling | PIK3C2G, AKT3, ITGB6, ITGA4 |
| Actin Cytoskeleton Signaling | FGF2, PIK3C2G, ITGA4, MSN |
| PTEN Signaling | GHR, AKT3, ITGA4 |
| ERK/MAPK Signaling | PLA2G4A, PIK3C2G, ELF5, ITGA4 |
| Xenobiotic Metabolism Signaling | IL1A, CAMK4, ALDH1A1, UGT2B4, PIK3C2G, IL1B, SULT1E1, SULT1B1, UGT2B28 |
| Metabolism of Xenobiotics by Cytochrome P450 | AKR1C1, UGT2B4, DHRS2, CYP4F11, UGT2B28 |
| Androgen and Estrogen Metabolism | AKR1C1, UGT2B4, SULT1E1, HSD17B2, UGT2B28 |
| NF-kB Signaling | IL18, IL1A, GHR, PIK3C2G, IL1B, AKT3 |
| Wnt/β-catenin Signaling | CDH2, MMP7, GJA1, SFRP2, AKT3, SFRP1, DKK1 |
| p53 Signaling | CCND2, PIK3C2G, AKT3, SERPINB5 |
| FXR/RXR Activation | IL18, IL1A, UGT2B4, IL1B, AKT3 |
| Tyrosine Metabolism | TYRP1, DHRS2 (includes EG:10202), TYR, DCT |
| Retinol Metabolism | ALDH1A1, UGT2B4, UGT2B28 |
| Pentose and Glucuronate Interconversions | UCHL1, UGT2B4, UGT2B28 |