Exon-Level Microarray Analyses Identify Alternative Splicing Programs in Breast Cancer

Anna Lapuk1, Henry Mart1, Lakshmi Jakkula1, Helder Pedro3, Sanchita Bhattacharya1, Elizabeth Purdom2, Zhi Hu1, Ken Simpson2, Lior Pachter3, Steffen Durinck1, Nicholas Wang1, Bahram Parvin1, Gerald Fontenay1, Terence Speed2,7, James Garbe1, Martha Stampfer1, Hovig Bayandorian4, Shannon Dorton1, Tyson A. Clark5, Anthony Schweitzer3, Andrew Wyrobek1, Heidi Feiler1, Paul Spellman1, John Conboy1, and Joe W. Gray1,6

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. 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. Reverse transcription-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. Small interfering RNA 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. Mol Cancer Res; 8(7); 961–74. ©2010 AACR.

Introduction

Breast cancer is a heterogeneous disease that shows 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 (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 thus 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 NH2- or COOH-terminal ends of proteins, whereas alternative pre-mRNA splicing of cassette exons can alter the expression of functionally important internal domains. In fact, recent studies of human transcriptome suggest that >90% of human genes are processed to produce alternative transcript isoforms through one of these mechanisms (5, 6), and it is becoming clear that alternative splicing (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 reverse transcription-PCR (RT-PCR) identified several AS events (ASE) 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 five nonmalignant 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 sequencing using massively parallel sequencing techniques to validate 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.

**Materials and Methods**

**Cell line collection**

Breast cancer cell lines used in this study were obtained from the American Type Culture Collection 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 ref. (10). We analyzed AS using microarrays in 26 breast cancer cell lines and five nonmalignant immortalized human mammary epithelial cells (HMEC) cultured as previously described (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 nonmalignant immortalized human mammary epithelial cells (HMEC) cultured as previously described (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. For RT-PCR validation of predicted ASEs, we expanded the cell line collection to include 48 breast cancer cell lines of basal, luminal, and claudin-low subtypes as well as four normal finite life span 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. (13) have shown that 250MK cells express luminal markers and thus represent the luminal subtype.

**Affymetrix HJAY design and data processing**

Affymetrix HJAY design was designed using content from ExonWalk (C. Sugnet), Ensembl, and RefSeq databases (National Center for Biotechnology Information 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 nonunique locations in the human genome and were likely to give cross-hybridization signal. These were excluded from our analysis. In total, 501,557 of probe sets from 23,546 transcript clusters were retained. Transcript clusters were assigned to known genes using database table refFlat.txt of the University of California at Santa Cruz Genome Browser (http://genome.ucsc.edu/). The HJAY data were preprocessed using the Affymetrix Expression Console Software. Probe set level expression measurements were generated from quantified Affymetrix image files (“.CEL” files) using the RMA algorithm (14). 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 probe sets (PSR) measurements in that cluster.

**Affymetrix microarray profiling**

HJAY profiling of cell lines was done using the GeneChip Whole Transcript Sense Target Labeling Assay kit (Affymetrix). An initial step to remove rRNA was used to minimize background and to increase detection sensitivity and specificity. 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 2 µg of total RNA from each sample, incubated at 70°C for 5 minutes, and then cooled on ice. RiboMinus magnetic beads, prepared by a batch method, were added to the samples and incubated at 37°C for 10 minutes. 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 9.8 µL. Probe synthesis, oligonucleotide array hybridization, and scanning were done according to the standard Affymetrix GeneChip protocol for the Whole Transcript 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 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 (5.5 μg) were hybridized in a volume of 220 μL at 45°C for 16 hours to Affymetrix high-density HJAYs. The arrays were washed and stained with streptavidinphycoerythrin (final concentration, 10 μg/mL). Signal amplification was done 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 Whole Transcript 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 are 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 probe sets in microarray data as follows:

**Data filtering.** Probe set and gene level expression data were filtered to remove noisy data that might contribute to false positives. Briefly, we required the following: (a) expression above the background in at least 25% of samples; (b) probe set exhibiting differential expression; (c) at least three probe sets per transcript cluster exhibiting expression above the background to ensure correct Finding Isoforms using Robust Multichip Analysis (FIRMA) linear model fitting (see below).

**Splicing index.** We calculated a splicing index (SI) using filtered probe set level expression data according to the formula:

\[
SI_{i,j,k} = \frac{c_{i,j,k}}{\bar{g}_{j,k}}
\]

in which \(c_{i,j,k}\) is the expression level of the \(i\) probe set, in experiment \(j\), within the \(k\)th transcript cluster. \(\bar{g}_{j,k}\) is the transcript cluster expression level estimate of the \(j\) experiment and \(k\) transcript cluster calculated as the mean of expression of its probe sets 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 probe sets were unlikely to represent clear AS cases. To avoid prioritization of such clusters in the downstream selection process, we converted the SI to a normalized SI (NSI) data according to the formula:

\[
NSI_{i,j,k} = \frac{SI_{i,j,k}}{N_k}
\]

in which \(N_k\) 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:

\[
N_{i,k} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (c_{i,j,k} - \mu_{i,k})^2}
\]

in which \(N_{i,k}\) is the SD of expression of a probe set \(i\) within a transcript cluster \(k\), \(m\) is the number of samples, \(\mu_{i,k}\) is the mean of \(c_{i,j,k}\).

**Selection of highly variable probe sets.** NSI data were used to select probe sets with the highest variability among cell lines due to AS. We computed a variability score for a probe set \(i\), from transcript cluster \(k\) using SDs of NSI and expression of probe sets across samples according to the formula:

\[
\sqrt{\frac{1}{m} \sum_{i=1}^{m} (NSI_{i,j,k} - \bar{NSI}_{j,k})^2} * \sqrt{\frac{1}{m} \sum_{i=1}^{m} (e_{i,j,k} - \mu_{j,k})^2}
\]

in which \(m\) is the number of samples, \(\bar{NSI}_{j,k}\) and \(\mu_{j,k}\) are the means of NSI and \(c_{i,j,k}\) respectively. We selected probe sets with the highest 1% of variability scores. This conservative selection strategy likely misses many valid ASEs; however, it increases the probably that AS calls are valid. In total, 2,783 probe sets from 1,760 transcript clusters passed this cutoff.

**Selection of statistically robust AS probe sets using FIRMA.** We used a minor modification of the program FIRMA (15) to assess the robustness of the highly variable probe sets 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 (perfect match, PM) values:

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

in which \(c_i\) 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 \(\log_2(PM_{ij})\) is the log (base2) of the background-corrected, normalized perfect match signal for probe \(j\) on chip \(i\).

The model is fitted using iteratively reweighted least squares (16), 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 probe sets within a transcript—likely due to differential splicing. Instead of using residuals as described in the Purdom et al. (15), 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 cutoff of \(w_{ij} \leq 0.7\) for an indication of AS taking place at a given probe set in a given sample. To generate FIRMA weight data for preselected highly variable probe sets, we ran the FIRMA algorithm using R function rlm on a subset of 78,050 probe sets. These were all unique probe sets from the 1,760 candidate transcript clusters described above.

**Selection of the best AS candidate events.** The 2,783 most variable probe sets were further filtered to select the most reliable alternatively spliced events within the breast cancer cell lines. Of 2,783 probe sets, we retained those that had the following: (a) FIRMA weights \(w_{ij}\) of \(\leq 0.7\) in at least 10% of cell lines, (b) average probe set expression of \(\log_2 \geq 5.9\), and (c) belonged to a transcript cluster mapped unambiguously to a known gene. We grouped
the remaining probe sets based on the location within the genome. Probe sets located within 150 bp from each other were considered to describe the same ASE. This process yielded 392 probe sets belonging to 181 alternative splice events of ≥2 probe sets 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 cutoff was log₂ ≥ 3). We measured the variability of expression for every transcript cluster on the array by calculating the coefficient of variation (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 gene ontology term enrichment analysis

We used the DAVID Functional Annotation tool (17, 18) and the Ingenuity Pathway Analysis software (Ingenuity Systems, http://www.ingenuity.com) to perform enrichment analysis for 156 best AS genes and 224 top differentially expressed genes. Gene ontology (GO) enrichment was also done with DAVID with a Benjamini Hochberg–adjusted P value cutoff 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 the Ingenuity Pathway Analysis 6.0 database and used for comparative pathway enrichment analysis. Human Genome U133A array data were used as a reference to rank top statistically significant overrepresented canonical signaling and metabolic pathways in both sets. Fisher’s exact test was applied to examine the statistical overrepresentation of pathways, using a threshold of Benjamini Hochberg adjusted P value of ≤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 10-cm dishes under the standard conditions for each cell type, using the RNeasy Mini kit (QIAGEN, Inc.). Cell lysis was done in 600 μL RLT Buffer with mechanical shearing; the RNA was recovered according to the manufacturer’s instructions.

RT-PCR validation

Primer design. Primers were designed using the Primer3 software (19) 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 toward a common constitutive exon primer. Validation was done using a panel of 48 cell lines and 7 primary breast tumors. For primers sequences see Supplementary Materials.

Reverse transcription-PCR. Three microgram of total RNA from each cell line and 1 μg from each primary tumor was used to produce cDNA primed with random hexamers in a 20-μL volume using the SuperScript III First Strand Synthesis kit (Invitrogen) with standard protocols. An aliquot of the resulting cDNA (0.5 μL) was used in each 10-μL PCR 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 (4 μL) were separated on 5% polyacrylamide gels using a Dual Triple Wide Mini-Vertical Electrophoresis System (CBS Scientific) and were imaged.

Fox2 knockdown experiments. Breast cancer cell lines at 55% to 65% confluency were transfected with Fox2-specific small interfering RNA (siRNA; Fox2 On-TARGETplus SMARTpool L-020616-01) or control siRNA (On-TARGETplus Non-Targeting Pool; D-001810-10-05) from Dharmacon. Transfection was done with 25 nmol/L siRNA using Dharmafect I Transfection Reagent according to the manufacturer’s instructions (Dharmacon). After 48 hours, 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 AS 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. Reads (19,553,572) of 32-bp length were obtained for the BT549 sample, and 18,747,831 reads were obtained for MCF7. ELAND was used to align the reads to the HJAY probe set sequences, which were used as a reference. For mapping procedure details, see Supplementary Data. Probe set expression levels were derived from read count data (number of reads mapped to the reference sequence)
by normalizing for the reference sequence length. Because the minimal reference sequence length was 32 bp, we defined a normalized read count (NRC) data as:

$$NRC = \frac{RC}{32} \cdot \frac{L}{C}$$

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

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

Further, expression estimates using both the 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, probe set–wise SI score differences have been calculated for the pair of cell lines for each platform. A probe set that had a SI score difference at least one SD away from the mean of SI differences of all probe sets in that gene was called AS.

We tested the overall agreement of the two platform expression profiles and the ability to measure differential expression. Between platforms, correlations of expression values and log₂ ratios were comparable with those published before for microarray platform comparisons (Spearman correlation of 0.5-0.7; ref. 20) and for microarray sequencing platform comparisons (Spearman correlation of 0.73-0.75; ref. 21). 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 ($\chi^2$ test, $\chi^2 = 489.5$; degrees of freedom = 1; $P < 2.2e-16$). 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 nonmalignant immortalized breast cell lines and 26 breast
FIGURE 2. Unsupervised hierarchical clustering of AS predictions. Three hundred ninety-two probe set 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 probe set for the exclusion event (JUC0700130578_st) and relatively lower expression of probe sets 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.
cancer cell lines including 13 classified as luminal, 6 classified as basal (1, 2), and 7 classified as claudin-low (3). The five 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 Materials and Methods). We used a computational pipeline designed to detect hallmarks of probe set expression associated with alternative pre-mRNA splicing events among the cell lines. The pipeline used two techniques, SI (22) and FIRMA (15). 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 probe set within a transcript in a statistically robust manner so that it identified those probe sets (exons) whose expression pattern across samples did not follow overall transcript expression pattern across the same samples. We included probe set (exon) level and transcript level filtering to remove probe sets that exhibited high background, high level cross hybridization, and/or weak hybridization (for details, see Materials and Methods) to reduce the false-positive rate in AS detection. Finally, we added postprocessing to select the top 1% of probe sets to identify recurrent AS patterns among the cell lines. We performed hierarchical clustering of the 392 AS probe sets 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). Nonmalignant immortalized mammary epithelial lines clustered together with basal subtype cell lines. Application of the standard Student’s t test (Benjamini-Hochberg-adjusted \( P < 0.05 \)) to NSI data for the 392 AS probe sets identified 74 (of 181) splicing events that correlated with a breast cancer subtype. As expected, reciprocal behavior was observed for probe sets representing exon inclusion and exclusion events (Fig. 2, cluster D).

We tested the validity of predicted AS using deep sequencing and RT-PCR. Deep sequencing of one claudin-low

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene</th>
<th>Specificity</th>
<th>AS type</th>
<th>No. of PS</th>
<th>PCR</th>
<th>Affy_call</th>
<th>Illu_call</th>
</tr>
</thead>
<tbody>
<tr>
<td>117</td>
<td>DST-1</td>
<td>cancer</td>
<td>Cassette exon</td>
<td>3</td>
<td>AS</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>DST-1</td>
<td>cancer</td>
<td>Alt start</td>
<td>3</td>
<td>No AS</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>DST-2</td>
<td>cancer</td>
<td>Alt start</td>
<td>2</td>
<td>AS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>146</td>
<td>DST-2</td>
<td>cancer</td>
<td>Alt stop</td>
<td>3</td>
<td>AS</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>107</td>
<td>FAM62B</td>
<td>CL</td>
<td>Cassette exon</td>
<td>4</td>
<td>AS</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>58</td>
<td>FAT</td>
<td>CL</td>
<td>Cassette exon</td>
<td>3</td>
<td>Weak AS</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>135</td>
<td>FER1L3</td>
<td>CL, LUM</td>
<td>Cassette exon</td>
<td>3</td>
<td>AS</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>FLNB</td>
<td>CL</td>
<td>Cassette exon</td>
<td>4</td>
<td>AS</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>MYO6</td>
<td>LUM, BA</td>
<td>Tandem cassette exons</td>
<td>6</td>
<td>AS</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>SLK</td>
<td>CL, cancer</td>
<td>Novel alt exon</td>
<td>3</td>
<td>AS</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>170</td>
<td>UTRN</td>
<td>NV</td>
<td>Novel alt exon</td>
<td>4</td>
<td>No AS</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviation: NV, not validated.

NOTE: 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 with the reference) is most frequent. The ratio of alternative isoforms in the nonmalignant mammary gland cell lines were taken as a reference. The AS type column refers to the type of ASE being tested. Right columns show the result of assessing AS in the cell lines MCF7 (luminal) and BT549 (claudin-low) using Illumina sequencing. The No. of PS column shows the number of probe sets in microarray and sequencing data describing the same ASE validated using RT-PCR. The last three columns summarize RT-PCR, microarray, and sequencing data on AS 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.
One luminal (MCF7) cell line was done to obtain \( \sim 19 \) million 32-bp sequence reads from each cell line. Analysis of these data validated \( \sim 60\% \) of HJAY microarray–predicted splicing differences among the highly expressed genes (see Materials and 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 nonmalignant immortalized breast cell lines and 4 normal finite life span HMEC strains. The ASEs selected for verification included cassette exons, alternative 5’ and 3’ ends, and tandem cassette exons. Table 1 and Fig. 3 show that 11 of the 12 AS predictions exhibited splicing differences among the cell lines. In addition to these, eight other predicted ASEs were used in Fox2 knockdown experiments and seven of them showed AS in a smaller panel of 12 cell lines representing the same cancer subtypes (Fig. 5B, additional ASEs CLSTN1-57nt, CLSTN1-30nt, KIF21A, PLOD2, ST7, MARK3, ENAH, and VPS39). Thus, the validation rate of our predictions was \( \sim 90\% \) (total of 18 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 AS 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 AS 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 with introns adjacent to constitutively spliced exons or nontissue-specific alternative exons (11, 23-27). The recurring association of this Fox binding site (UGCAUG) with tissue-specific alternative exons suggests an important role in regulation of AS. 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 Fig. 2 that shows that its expression is significantly elevated in basal and claudin-low subtypes compared with luminal subtype cells. We further explored the role of Fox2 class splicing factors in breast cancer by examining the expression levels of several Fox2 target genes across the same cancer subtypes (Fig. 5B). These results suggest that Fox2 may play a role in regulating subtype-specific splicing in breast cancer.
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 probe sets supported differential AS between luminal versus basal and/or claudin-low cells (Supplementary Table S1). Remarkably, Fig. 4 shows that 19 of 22 of these exons possessed one or more UGCAUG binding site in the intron sequences within 400 nt of the differentially expressed exons. This frequency greatly exceeded that expected for a random hexamer, which should occur only once every 4 kb, or in one of five such cases. In addition, some exons lacking these more proximal UGCAUG motifs possessed Fox binding sites more distally. In two cases (ENAH and ST7), clusters of predicted Fox2 binding sites were located in the long downstream intron between 1.4 to 1.8 kb from the regulated exon. Although these sites are more distant than most described splicing enhancers, there is precedent for functional Fox sites >1 kb from the regulated exon (28).

The functional significance of these associations is further supported by the observation that the Fox binding sites are highly conserved in evolution (Fig. 4). In 18 of 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 nonmalignant 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% to 80% as determined by qRT-PCR (Fig. 5A). Splicing changes induced by Fox2 knockdown were then examined for 14 subtype-specific alternative exons in four representative cell lines from each of the luminal, basal, and claudin-low subtypes (Fig. 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, whereas 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, 57 nt). In other cases,

**FIGURE 4.** Frequent association of Fox sites with breast cancer subtype–specific alternative exons. Diagrams represent the genomic regions spanning 1 kb upstream to 1 kb downstream of the regulated exons. Filled ovals, locations of evolutionarily conserved occurrences of the Fox splicing regulatory motif UGCAUG. Empty ovals, nonconserved sites present only in the human genome. The highest frequency of Fox binding sites occurs within 400 nt (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 one site 1.5 kb upstream.
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, whereas 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 ASEs, 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, and MARK3). Evidence for Fox2 enhancer effects on the same set of exons was often but not always observed in basal cells, although Fox2 is relatively highly expressed in claudin-low and basal subtypes (Fig. 2). Interestingly, a few exons that were predominantly skipped in claudin-low cells paradoxically exhibited Fox2-dependent splicing in luminal cells, although 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.

**FIGURE 5.** Knockdown of Fox2 alters splicing in breast cancer cells. A, level of Fox2 siRNA knockdown achieved in four breast-derived cell lines from nonmalignant tissue and three different malignant breast cancer subtypes. Irrelevant siRNA and mock transfection serve as controls. 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 MDA MB 157; samples 9–12). Splicing in each cell line was compared after treatment with a control siRNA (first lane of each pair), versus treatment with the Fox2 siRNA (second lane of each pair). Left, identity of genes containing the 14 subtype-specific exons. Upper PCR bands, exon inclusion; lower bands, exon-skipping products. For VPS39, only the inclusion product was detected. For MARK3, the amplified region spans two alternative exons of 27 nt and 45 nt, and can generate four products. The 27 nt exon is subtype specific (skipped in luminal but partially included in basal and claudin-low) and Fox2 dependent (abundance of the +27+45 nt and +27−45 nt bands is greatly reduced in claudin-low cells after Fox2 knockdown).
Important pathways and networks affected by AS

We analyzed 156 alternatively spliced genes detected in breast cancer cell lines using the DAVID functional annotation tool (http://david.abcc.ncifcrf.gov; refs. 29, 30) and the 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 because only two 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 (Table 2; Fig. 6). 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 (Fig. 2 in Supplementary Materials). On the other hand, Fig. 6 shows that 224 differentially expressed but not alternately spliced genes predominantly influenced aspects of metabolism.

Discussion

AS of pre-mRNAs is now well established as a mechanism for increasing protein functional diversity during normal development. In addition, 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 (27). These studies

### Table 2. Component gene lists from enriched pathways from Fig. 6

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Component genes from Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axonal guidance signaling</td>
<td>PTK2, SLIT3, PRKACB, GNAS, ADAM15, ITGA2, ITGA5, LIMK2, RASSF5, SEMA4A</td>
</tr>
<tr>
<td>Integrin signaling</td>
<td>PTK2, DDEF1, ITGA2, ITGA6, ITGA5, ACTN1</td>
</tr>
<tr>
<td>Actin cytoskeleton signaling</td>
<td>PTK2, TIAM1, ITGA2, ITGA5, LIMK2, ACTN1</td>
</tr>
<tr>
<td>Tight junction signaling</td>
<td>PRKACB, EPB41, TIAM1, TJP3, TJP1</td>
</tr>
<tr>
<td>PTEN signaling</td>
<td>PTK2, ITGA2, ITGA5</td>
</tr>
<tr>
<td>Androgen and estrogen metabolism</td>
<td>STS</td>
</tr>
<tr>
<td>NF-κB signaling</td>
<td>PRKACB</td>
</tr>
<tr>
<td>Ephrin receptor signaling</td>
<td>PTK2, GNAS, GRIN2D, ITGA2, ITGA5, LIMK2</td>
</tr>
<tr>
<td>ERK/MAPK signaling</td>
<td>PTK2, PRKACB, ITGA2, ITGA5</td>
</tr>
<tr>
<td>Xenobiotic metabolism signaling</td>
<td>GSTM1, GSTP1, ALDH3A2, SULT1A1</td>
</tr>
<tr>
<td>Metabolism of xenobiotics by cytochrome P450</td>
<td>GSTM1, GSTP1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Component genes from Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axonal guidance signaling</td>
<td>SEMA3A, SEMA3D, PIK3C2G, AKT3, EPHA3, BMP5, ITGA4</td>
</tr>
<tr>
<td>Ephrin receptor signaling</td>
<td>PIK3C2G, AKT3, EPHA3, ITGA4</td>
</tr>
<tr>
<td>Integrin signaling</td>
<td>PIK3C2G, AKT3, ITGB6, ITGA4</td>
</tr>
<tr>
<td>Actin cytoskeleton signaling</td>
<td>FGFR2, PIK3C2G, ITGA4, MSN</td>
</tr>
<tr>
<td>PTEN signaling</td>
<td>GHR, AKT3, ITGA4</td>
</tr>
<tr>
<td>ERK/MAPK signaling</td>
<td>PLA2G4A, PIK3C2G, ELF5, ITGA4</td>
</tr>
<tr>
<td>Xenobiotic metabolism signaling</td>
<td>IL1A, CAMK4, ALDH1A1, UGT2B4, PIK3C2G, IL1B, SULT1E1, SULT1B1, UGT2B28</td>
</tr>
<tr>
<td>Metabolism of xenobiotics by cytochrome p450</td>
<td>AKR1C1, UGT2B4, DHRS2, CYP4F11, UGT2B28</td>
</tr>
<tr>
<td>Androgen and estrogen metabolism</td>
<td>AKR1C1, UGT2B4, SULT1E1, HSD17B2, UGT2B28</td>
</tr>
<tr>
<td>NF-κB signaling</td>
<td>IL1B, IL1A, GHR, PIK3C2G, IL1B, AKT3</td>
</tr>
<tr>
<td>Wnt/β-catenin signaling</td>
<td>CDH2, MMP7, GJA1, SFRP2, AKT3, SFRP1, DKK1</td>
</tr>
<tr>
<td>p53 signaling</td>
<td>CCND2, PIK3C2G, AKT3, SERPINE5</td>
</tr>
<tr>
<td>FXR/RXR activation</td>
<td>IL1A, IL1A, UGT2B4, IL1B, AKT3</td>
</tr>
<tr>
<td>Tyrosine metabolism</td>
<td>TYRP1, DHRS2 (includes EG:10202), TYR, DCT</td>
</tr>
<tr>
<td>Retinol metabolism</td>
<td>ALDH1A1, UGT2B4, UGT2B28</td>
</tr>
<tr>
<td>Pentose and glucuronate interconversions</td>
<td>UCHL1, UGT2B4, UGT2B28</td>
</tr>
</tbody>
</table>
identified over 200 breast cancer-specific ASEs that were spliced differently in \( ER^+ \) tumors versus normal tissue (31). Moreover, the splicing factor, Fox2, has been associated with normal tissue-specific splicing (23) and has been implicated as an important splicing factor in breast and ovarian cancer development (31). We extended these findings here by interrogating AS 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.

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 (Table 2; Fig. 6) and GO terms mostly related to cytoskeleton and actin. The involvement of the cytoskeleton in morphology and motility suggests the possibility that AS plays a significant role in determining phenotypic differences between these breast cancer subtypes (10, 32). Interestingly, we found that genes and the biological processes influenced by AS were different from those influenced by expression regulation. In general, AS influenced aspects of cell surface protein-mediated signaling that affected morphology and motility, whereas 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 (33, 34). The importance of AS in protein function regulation is supported by the theoretical protein structure analysis of Wang et al. (35).
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. In addition, Hughes and Friedman (36) showed that AS genes tend to interact with other AS genes in genetic and protein interaction networks. 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 (37).

Expression of Fox2 is associated with subtypes showing the upregulation of expression in claudin-low and basal cells, and downregulation 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 nonluminal subtypes (Pearson’s correlation of 0.5), (b) evolutionary conservation of intronic Fox2 consensus binding sites, and (c) reduced inclusion of target exons in nonluminal subtype cells after treatment with a siRNA against Fox2. Notably, even the 5′UTR alternative exon having a cluster of three distal intronic Fox binding sites ~1.8 kb 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 because 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 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 (38, 39) 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. (31) as alternately spliced in breast cancer shows only 10 that are common to both studies. This is not surprising because 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, whereas we used a microarray approach that is relatively insensitive to the presence of low-abundance transcripts. In addition, we focused on AS differences between cell lines derived from both ER+ and ER− breast tumors, whereas Venables et al. focused on the detection of tumor-specific splice events that differed between ER+ tumors and normal tissues. Consequently, 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 AS 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 because 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 AS 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 (40). Targeting this breast cancer subtype is highly important given the growing evidence that such cancers may be particularly resistant to conventional therapies (41). Supplementary Table S1 describes alternative exon usage in alternatively spliced cell surface proteins to guide efforts to develop subtype-specific markers and therapies.

Disclosure of Potential Conflicts of Interest

J.W. Gray received early access to microarrays from Affymetrix.

Grant Support

Director, Office of Science, Office of Biological & Environmental Research, of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. USAMRMC W81XWH-07-1-0663 and NIH grants CA58207, CA112970, and CA 126477 (J.G. Conboy) by NIH grant HL045182 (J.W. Gray); and by the USAMRMC W81XWH-07-1-0663 and NIH grants CA58207, CA112970, and CA 126477 (J.G. Conboy) by NIH grant HL045182 (J.W. Gray); and by the FCT SFRH/BD 33203 2007 (H. Pedro).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 12/07/2009; revised 05/21/2010; accepted 06/01/2010; published OnlineFirst 07/06/2010.

References

3. Hennessy BT, Gonzalez-Angulo AM, Stemke-Hale K, et al. Charac-

Molecular Cancer Research

Exon-Level Microarray Analyses Identify Alternative Splicing Programs in Breast Cancer

Anna Lapuk, Henry Marr, Lakshmi Jakkula, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1541-7786.MCR-09-0528

Supplementary Material
Access the most recent supplemental material at:
http://mcr.aacrjournals.org/content/suppl/2010/09/10/1541-7786.MCR-09-0528.DC1

Cited articles
This article cites 39 articles, 19 of which you can access for free at:
http://mcr.aacrjournals.org/content/8/7/961.full.html#ref-list-1

Citing articles
This article has been cited by 22 HighWire-hosted articles. Access the articles at:
/content/8/7/961.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.