# Supplementary Materials.

## 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], Terry Speed[3,5], James Garbe[1], Martha Stampfer[1], Hovig Bayandorian[7], Shannon Dorton[1], Tyson Clark[2], Anthony Schweitzer[2] Andrew Wyrobek[1], Heidi Feiler[1], Paul Spellman[1], John Conboy[1], and Joe 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

* present address Vancouver Prostate Centre, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada

**Supplementary Table 1.** The list of the best 181 alternative splice events in breast cancer – available as a separate text file.

The table contains following columns:

**Gene_name(known_genes=156)** – official gene symbol.

**tr_cl(n=164)** – transcript cluster id as in the Affymetrix Human Junction Array annotation files.

**GroupID(splice_events=181)** – unique splice event ID, compiled of probesets ID the event has been detected with.

**PSIDcount** – number of individual probesets with which the splice events has been detected.

rank_by_NSI –probesets rank based on the variability score$_{i,j,k}$ (see Methods section of the paper).

**percent_of_FIRMA_score_>0.3** – percent of the samples having FiRMA weight score of >0.3

**UCSC_link** – link to the UCSC browser showing custom track with the best splice events reported in this study.

**B_L_diff_in_MeanSI** – difference between means of NSI values of (luminal) subtype and two other subtypes together.

**B_L_p.adjust** – Benjamini-Hochberg adjusted p-value associated with the previous column.

**caludinLOW_L_diff_in_MeanSI** - difference between means of NSI values of claudin-low and luminal subtypes cells.

**claudinLOW_L_p.adjust** - Benjamini-Hochberg adjusted p-value associated with the previous column.

**Subtype_specific** – code for statistically significant differences in splicing based on the 4 previous columns. 0 – no difference; 1 – significant difference according to either of the two comparisons in the previous columns; 2 – significant difference according to both comparisons in the previous columns.

**plasma membrane** – shows whether the gene product is localized to the plasma membrane according to the Gene Ontology.

**Accession_corresponding_to_isoform_specific_toClaudinLow, isoform_uniq_PS_Ids, isoform_uniq_PS__coordinates** - explicitly identify isoforms specific to claudin low cells within genes localized to plasma membrane.

**Supplementary Table 2**. Primer sequences for RT-PCR.

| **Gene** | **Forward Primer** | **Reverse Primer** | **Sizes** |
|---|---|---|---|
| FLNB | TGGAACCTATGACATCTTCTACACA | GACCAGGTCAAAAGGCTTAAATC | 244/172 |
| CLTC | AAACTGCATGGAGGCACAATATC | GGTTGTGTCTCTGTAGCTTGTTCTT | 162/142 |
| FAT | ATCTTGATCCCTGTCTTTCCAAG | TCATAGTTGGGGAACTCTTGTATGT | 156/121 |
| TJP1 | GCCACAACCAATTCATAGAATAGAC | AACATGGTTCTGCCTCATCATTT | 346/106 |
| FER1L3 | TTGGAACAACATATCTACACCTCTC | GCTTCCATAAAGATTCAGGTAACAA | 167/128 |
| UTRN | AGGAAGTTGAAAGCCTTAGAAAGAG | CTTTGAAAATCGAGCATTTATCCAT | 390/207 |
| SLK | AAACAGCAGAAACAGACTATCGAAC | ATGATCTTTTTCAGAGAGCCATCTA | 216/123 |
| MYO6 | AATCCCAACAGCAAGCAGTTCT | CACAAGAAGTATTGATGGTATCACG | 294/225 |
| FAM62B | CTTTCCTCTGCATTGCTGATCTT | ACTGACATCTGGACAACAGGATT | 203/141 |
| DST-1 | TCCAAATGACATAGAAAAAGAGTGG<br>TTTCAGGAATAAAAATGCACAGTAG | TCAGTTTGTCTTCACAGATGACACT | 248/151 |
| DST-2 | TCATGGACCTAAGGACTCGATATAC | ACCTTCGGAAGTTCTTCCTCTACT<br>TTGCTGATATTTGATACCCATTTTT | 427/331 |
| PLEC1 | AGCCATCCAGAACGAGATCAG<br>AGAGAACCAGCTCGGAGGA | AAGTGGGTCAACAAGCACCT | 117/86 |
| KIF21A | GAAATAACCAGTGCTACCCAAAAC | GTTTAAAGGAGCATCCTCATCAGT | 165/144 |
| CLSTN1 (57nt) | ATCCATCCAAGATAGAAACTCAGC | AGACAGTCGATCACCTTCTTATCC | 202/145 |
| CLSTN1 (30nt) | ATAGTCACAGAGAACGACAACACC | GATTTATCCACTACCACTGCATCA | 179/149 |
| ST7 | AGAAAGCCTGGAGAGAGAGAAAC | AAATAATTTTTCTGCTTCAGCAATG | 209/140 |
| MARK3 | CGAGGCTCCACTAATCTCTTTAGTA | CTTCTTTGTTTTCATCTTTTTGCTC | 284/212 |
| PLOD2 | TGAAAGGAACTATTTTGTTCGTGAT | CAGAGGTCATTGTTATAATGGGAAG | 225/162 |
| VPS39 | GTGGGAACCAAACAAGGACAT | CTGCTGAATCTTTTTGGAGAA | 150/117 |
| Fox2 qPCR | GTGTACACAGCCGACCCCTA | CACTTCAGTAGGGGGCAAATC | 112 |

**Deep sequencing of cell lines transcriptomes using Illumina technology.**

Library construction and sequencing. 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. mRNA was isolated from the total RNA extracts using Dynal oligo(dT) beads from Invitrogen. The mRNA isolates were then fragmented using Ambion's RNA fragmentation kit. Next, mRNA fragments were converted to double stranded cDNA fragments using random hexamer primers and the SuperScript™ Double-Stranded cDNA Synthesis Kit from Invitrogen. For the next steps, Illumina's genomic DNA sequencing protocol was followed to produce the MCF7 and BT549 sequencing libraries. Fragments of 200bp +/- 25bp were selected in the gel based ligation purification step. Sequencing was performed on Illumina's Genome Analyzer using one flowcell loading lane 1-4 with the BT549 library and lane 5-8 with the MCF7 library. The Illumina analysis pipeline was used for image analysis and base calling to obtain reads of 36nt long. 19,553,572 reads were obtained for the BT549 sample and 18,747,831 for MCF7. ELAND was used to align the reads to the genome.

Mapping of sequencing reads. Each cell line was sequenced using 4 Illumina flow cell lanes producing a total of 19,553,572 and 19,747,831 reads of 36 bp length for the BT549 and MCF7 cell lines respectively. The reads have been trimmed to 32 bp removing last 4 bp of poorer quality and mapped to the entire collection of probeset sequences from Human Junction Array using MAQ software [35]. To ensure adequate mapping we have compiled two reference sets of (i) junction sequences (JUCs) and (ii) exonic sequences (PSRs). The JUC reference set was assembled from sequences of probesets on the array by taking 29 bp of sequence on each side of an exon junction producing sequences of 58 bp length. A small fraction of JUCs (~8 %) were between 32 bp and 58 bp size with at least 16 bp on each side of the junction. For PSR reference set we have used full probesets sequences. The majority of PSR sequences (93%) ranged in size from ~40bp to ~700bp. Mapping was done requiring no more than 2 mismatches and MAQ alignment quality score of 20 for both JUC and PSR sequences, and at least 3 bp span of a splice junction. Supplementary Table 3 summarizes mapping results of Illumina/Solexa data for MCF7 and BT549 cell lines.

Expression estimates. Probesets expression level has been 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 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. Whole genome NRC data has been filtered to remove unreliable low coverage probesets from further analyses. We have required a probeset to have 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 has been log2 transformed to make it comparable with the microarray log2 scale expression summaries:

$$\text{Expression} = \log2\ (NRC +1).$$

Determination of Alternative Splicing (AS calls)

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

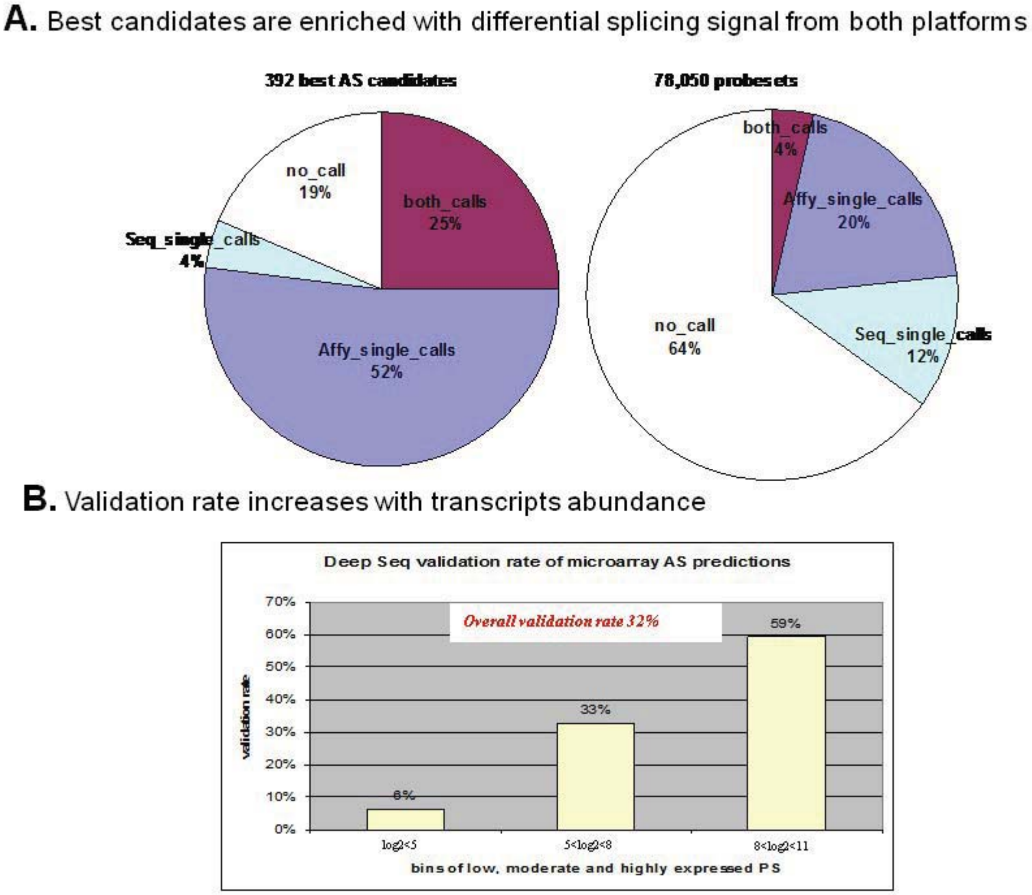
Comparison of splicing measured by microarrays and sequencing technology.

High throughput sequencing was used as an independent approach for detection of alternative splicing in breast cancer cell lines. We obtained ~ 19 million 32 basepair Illumina sequencing reads of the transcriptomes for two different subtypes of breast cancer, including one basal B (BT549) and one luminal (MCF7) line. Specifically we compared (1) the best 152 genes/392 probesets (termed best set) and (2) a strongly variable set (variable set) of 78,050 probesets from 1760 transcript clusters from the selection process (see "Selection of highly variable probe sets" section of the Methods). To test the overall agreement between expression profiles measured by the HJAY and Illumina platforms we compared estimated expression level (absolute signal) and relative expression (BT549 vs MCF7) for the 78K probesets. The relative expression of BT549 vs MCF7 (log2 ratios) of illumina vs HJAY was 0.78.

For validation both microarray and Illumina/Solexa data for sets 1 and 2 have been processed to generate differential alternative splicing calls (AS calls) using SI measures. The summary of AS calls within two platforms is in Supplementary Table 4. Within the best set out of 302 AS calls made for the Affymetrix platform, 98 (32%) have been also made by Illumina platform. In comparison, the variation set rate was 15.6%. Overall, within the best set, 25% of probesets had AS calls in both platforms. Within the variable set only 4% were called from both platforms (Supplementary Figure 1). Thus, the best 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). Notably, even though the 78K set included differentially spliced probesets (the set originates from step 1 of the selection process and 4% of probesets exhibited differential splicing with both platforms), their removal did not significantly improve correlations. This observation supports our assumption that 78K is an unbiased representative sampling of the whole genome data.

**Supplementary figure 1.** Illumina sequencing validation of the AS microarray predictions


A. Best candidates are enriched with differential splicing signal from both platforms

B. Validation rate increases with transcripts abundance

A) Pie charts summarize detection of AS in two cell lines, MCF7 and BT549, using the best validation set of 392 probesets and representative (reference) set of 78,050 probesets. Affymetrix calls, Illumina calls and Both calls refer to the number of probesets, which showed differential signals between the two cell lines in a given platform or in two platforms simultaneously. Single calls refer to the number of probesets, which showed differential signal in a given platform, but not in the other. No call refers to absence of differential signal in both platforms. B) A list of best 392 probesets have been stratified according to average expression level of corresponding transcripts and sets of transcripts from each bin have been validated using Illumina data.

**Supplementary Table 3.** Summary of Illumina/Solexa reads mapping onto the collection of probe sets from HJAY array.

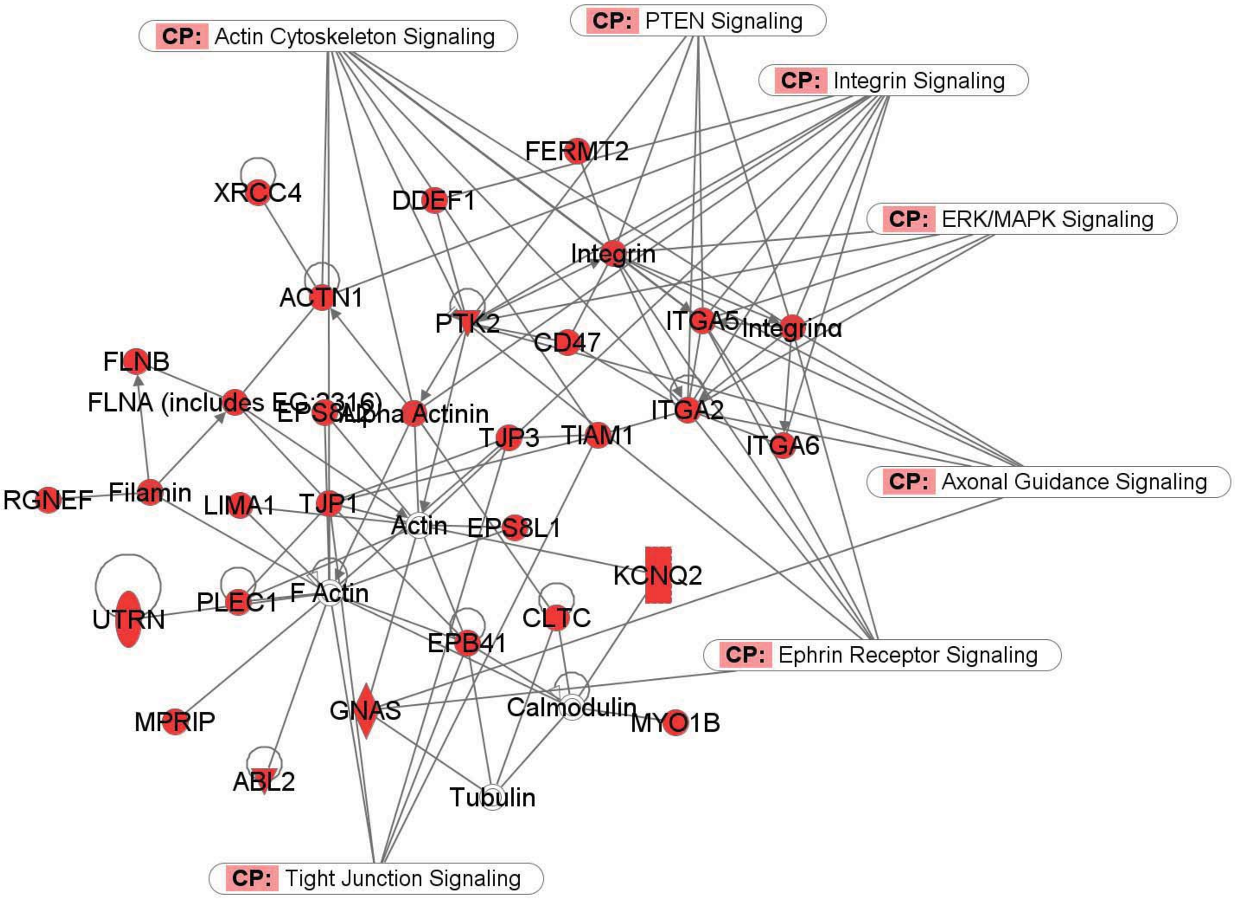
| | BT549 | MCF7 |
|---|---|---|
| Number of JUC probe sets | 252,741 | |
| Number of PSR probe sets | 307,435 | |
| Total number of reads | 19,553,572 | 19,747,831 |
| Reads that map to JUC | 1,593,103 | 1,533,075 |
| Reads that map to PSR | 10,239,925 | 8,872,672 |
| Number of JUC probes with at least one mapped read | 101,290 | 103,180 |
| Number of PSR probes with at least one mapped read | 143,536 | 145,746 |

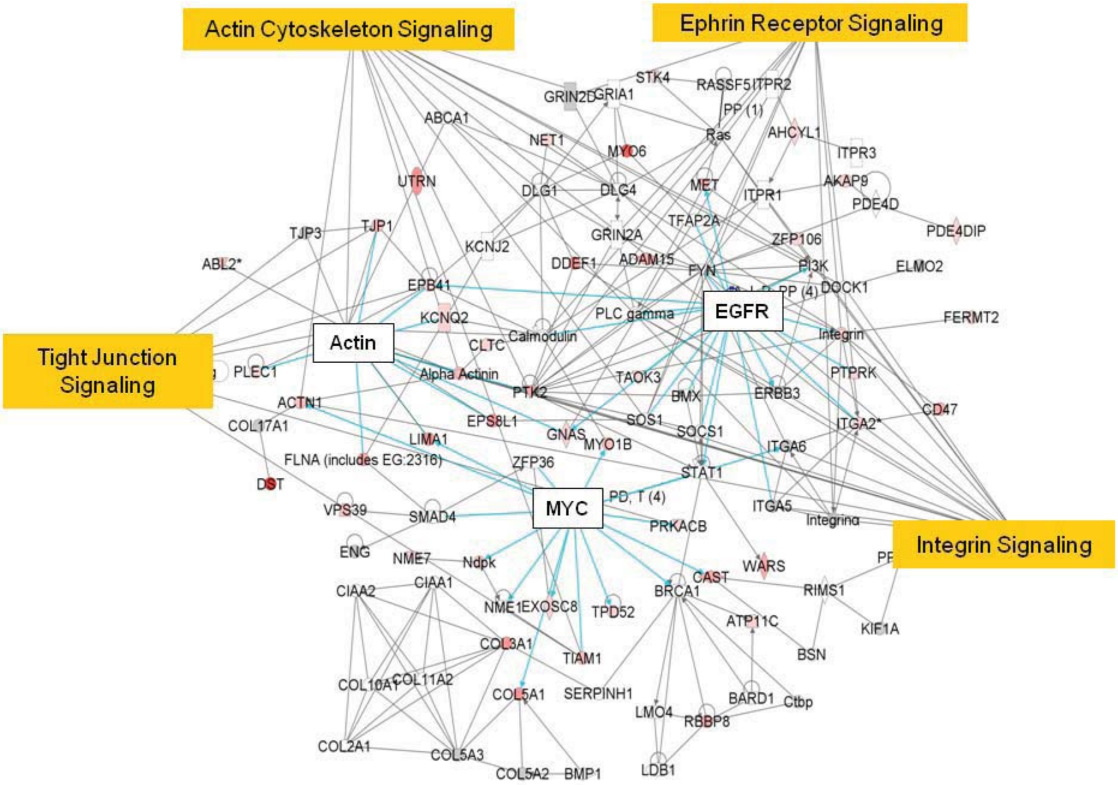
**Supplementary Table 4.** Summary of AS detection with microarray and sequencing platforms**.**

| AS calls | validation set | reference set |
|---|---|---|
| Affymetrix calls | 302 | 18149 |
| Illumina_calls | 115 | 12045 |
| both_calls | 98 | 2832 |
| Affymetrix_single_calls | 204 | 15317 |
| Illuminal_single_calls | 17 | 9213 |
| no_call | 73 | 50688 |
| total | 392 | 78050 |
| validation rate | 32.45% | 15.60% |

The table summarizes detection of AS in two cell lines, MCF7 and BT549, using the best validation set of 392 probesets and representative (reference) set of 78,050 probesets. Affymetrix calls, Illumina calls and Both calls refer to the number of probesets, which showed differential signals between the two cell lines in a given platform or in two platforms simultaneously. Single calls refer to the number of probesets, which showed differential signal in a given platform, but not in the other. No call refers to absence of differential signal in both platforms.

**Supplementary Figure 2.** Networks and Pathways associated with splice variants in breast cancer cell lines.


CP: Actin Cytoskeleton Signaling
CP: PTEN Signaling
CP: Integrin Signaling
CP: ERK/MAPK Signaling
CP: Axonal Guidance Signaling
CP: Ephrin Receptor Signaling
CP: Tight Junction Signaling
FERMT2
XRCC4
DDEF1
Integrin
ACTN1
PTK2
CD47
ITGA5
Integrinα
FLNB
FLNA (includes EG:2316)
EPS8
Alpha Actinin
ITGA2
TJP3
TIAM1
ITGA6
RGNEF
Filamin
LIMA1
TJP1
Actin
EPS8L1
KCNQ2
UTRN
PLEC1
F Actin
CLTC
EPB41
Calmodulin
MYO1B
MPRIP
GNAS
ABL2
Tubulin



Top network highly saturated with AS genes is shown at the top and the merged network (composed of three top networks) is shown below. A network is a graphical representation of the molecular relationships between molecules. Molecules are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least 1 reference from the literature, from a textbook, or from canonical information stored in the Ingenuity Pathways Knowledge Base. The intensity of the node color indicates the degree of up- (red) or down- (green) regulation. Nodes are displayed using various shapes that represent the functional class of the gene product. Edges are displayed with various labels that describe the nature of the relationship between the nodes (e.g., P for phosphorylation, T for transcription). For the shown merged network MYC, Actin, EGFR are the major nodes. The main functions associated with the merged network are cytoskeleton organization and biogenesis and cell signaling. There was also a significant enrichment for a number of pathways such as Integrin Signaling, Tight Junction Signaling, Ephrin Receptor Signaling, Actin Cytoskeleton Signaling and Focal adhesion. Genes involved with differential splicing are coded with grey, pink and red, depending on the frequency of differential splicing (as defined by FIRMA).