



SpliceArray™ Products

Advancing healthcare through RNA splicing

High Resolution Analysis of the Human Transcriptome: Detection of Extensive Alternative Splicing Independent of Transcriptional Activity

W. Zhou, H. Jordan, M. Brenner, S. Johnson, D. Wu,
D. Pallares, P. Fehlbaum, F. Rouet, L. Bracco,
C. Soucaille, R. Einstein

High Resolution Analysis of the Human Transcriptome: Detection of Extensive Alternative Splicing Independent of Transcriptional Activity

W. Zhou*, H. Jordan*, M. Brenner*, S. Johnson*, D. Wu*, D. Pallares#, P. Fehlbaum#, F. Rouet#, L. Bracco#, C. Soucaille#, R. Einstein*[†]
*ExonHit Therapeutics Inc, Gaithersburg, MD, USA; #ExonHit Therapeutics SA, Paris, France

Abstract

Commercially available microarrays have been used in many settings to generate expression profiles for a variety of applications, including target selection for disease detection, classification, profiling for pharmacogenomic response to therapeutics, and potential disease staging. However, commercial microarrays do not resolve a large portion of the transcriptome, as most of these transcripts are produced by alternative splicing. The inconsistency between genes and transcripts as explained through alternative splicing, is a major mechanism for driving proteomic diversity through transcript heterogeneity. Recent advances in the design of expression arrays and bioinformatics analysis techniques for the identification of alternative transcripts have produced a unique microarray design that allows the detection of alternatively spliced events within the human genome through the use of exon body and exon junction probes to provide a direct measure of each transcript, through simple calculations derived from expression data. Over 138,000 putative events were identified with direct evidence of publicly available cDNA sequences, with novel exons as the most abundant type of splice event detected. A custom array was manufactured to detect splicing events in the human genome and the performance was measured against standards recently published (MAQC Project) and the array monitored over 400,000 potential splicing events (known and predicted). The array was shown to be highly quantitative through sample titration for probe, gene and splice event level analysis. The array highly correlated with the Affymetrix HG-U133 Plus 2.0 array on the gene level, and provided more extensive coverage of each gene. Almost 60% of genes demonstrating differential expression of greater than 3 fold also contained extensive splicing alterations. Also, almost 10% of genes having constant overall expression values contained evidence of transcript diversity when examined in detail. Two examples illustrate the types of events identified: Lim domain 7 showed no differential expression, but demonstrated an exon skip event, while Erythrocyte membrane protein band 4.1 –like 3 was differentially expressed 11 fold and contained a skipped exon isoform. A novel design for the detection of specific alternatively spliced transcripts is described and shown to be highly reproducible, quantitative and provides high resolution for transcript analysis. Significant changes were also detected independent of transcriptional activity, indicating that the controls for transcript generation and transcription are distinct, and require novel tools in order to detect changes in specific transcript quantity. This array design will provide researchers with the ability to identify and quantify specific changes not only at the gene level, but also at the transcript level.

Introduction

A large portion of the diversity within the transcriptome is generated by alternative splicing, which in some cases, can produce thousands of transcripts from a single gene or locus (1,2). This has important implications in biology where extensive alterations in transcripts resulting from alternative splicing produce structurally different products and impacts the function of genes in biology, disease (3,4), and also in processes such as evolution (5,6). The fine granularity of the transcriptome has not been determined with clarity, and new tools are required in order to begin to identify with certainty the actual content (or parts list) of the transcriptome.

Several outstanding attempts have been made to generate databases that contain alternatively spliced isoforms, and these databases provide a glimpse of the potential diversity that exists within the transcriptome (see 7, for list). However, there are many different conditions in biology that have not been investigated, and applicable information on the transcripts that are expressed under these defined conditions does not exist. The ability to detect transcript diversity through commercial means has not been straight forward. Many alterations in expression of spliced isoforms have been difficult to detect due to probe design and the difficulty in establishing robust analytical methods. SpliceArray™ microarrays were developed to experimentally define the composite of transcripts that are present within biological samples and have the ability to determine subtle differential changes in gene expression for different alternatively spliced isoforms. A novel approach to probe design and analysis is presented here for the efficient detection of these alterations on a genome wide scale. The performance of this microarray was assessed, guided by the recent publications from the MicroArray Quality Control consortium.

The MicroArray Quality Control (MAQC) project is a FDA sponsored consortium consisting of over 140 members from academia, government, pharma & biotech, and was founded to address concerns of the microarray community regarding reproducibility of expression profiling experiments. The purpose of the MAQC project is to generate quality control tools for the microarray community in order to avoid procedural failures and to develop guidelines for microarray data analysis by providing the public with large reference datasets along with readily accessible reference RNA samples. The project set out to define performance parameters for different microarray platforms that are used in pharmacogenomic and toxicogenomic studies, with the hope of applying the results towards advancing personalized medicine. The study found, that overall,

the platforms perform similarly (8) and were validated with alternative quantitative gene expression platforms (9). However, all the platforms tested contain a similar bias in that probes were designed to monitor the overall level of the gene, and do not give any expression information toward the isoform diversity produced from each gene through alternative splicing. We provide here an analysis of a new microarray design in accordance with the approach outlined by the MAQC Consortium and demonstrate that these arrays are highly reproducible, quantify transcripts, and are sensitive in detecting subtle changes in transcript ratios.

Results

Sequence data was obtained from the NCBI GenBank database and used in an analysis to determine splicing alterations. A set of potential protein coding genes were selected for a total representation of 20,649 genes. From this group, 19,066 genes or 92.3% were found to contain evidence of potential alternative splicing within the selected sequence collection. As previously reported (12,13), the most common alteration found was an exon skip or novel exon event. The least common event was a retained intron and the distribution of the different event types identified is listed in Table 1. Over 71,000 exons were identified to be involved in some form of alternative splicing, and represents over 35% of exons found in the gene set. Interestingly, this indicates that the majority of exons are constitutive and are consistent throughout the expression repertoire for each locus.

Samples were selected in accordance with the MAQC project, using Statagene's Human Universal Reference RNA (Sample A) and Ambion's Human Brain RNA (Sample B). In addition, two titration samples were generated consisting of 1:3 and 3:1 ratios of sample A to sample B, respectively (Figure 1), which provide a means to validate the performance of the array through the assessment of a titration response. As with the

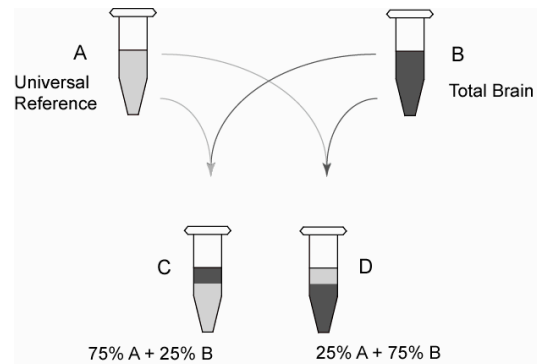


Figure 1. MAQC sample set up. Two reference samples were selected based on the MAQC study; Stratagene Universal RNA (sample A) and Ambion brain RNA (sample B) were used to generate two dependent titration samples where sample C contained 75% sample A and 25% sample B, and sample D contained 25% sample A and 75% sample B.

MAQC project (Shippy et al, 2006), we assumed a linear response for each probe across the four titration samples. All four samples were analyzed on the Human GWSA in triplicate. The frequency distribution of the expression values for all 12 arrays showed no anomalies (Figure S1A) and displayed a normal distribution of signal as expected for the platform. The reproducibility of the arrays was found to be highly concordant and produced low coefficients of variation for each sample analyzed (Table 2). The mean CV% were found to be slightly higher than the median values (5.5 versus 4.23, respectively), indicating a slight bias towards higher values. On a gross level, a principle component analysis (Figure S1B) shows close grouping of the replicates.

The titration samples were analyzed at the probe set, and gene level to define the quantitative nature of the expression data generated by the array. There were 527,574 probe sets identified in which expression was higher in the universal sample (A) than in the brain sample (B), (A>B) and 472,085 probe sets where B>A. Based on the fold change between samples A and B, probes were assessed for the titration response of expression values such that where the probe showed a fold change of A > B, that it also demonstrated that [sample A > sample C > sample D > sample B]. Where probes were found to be B > A, calculations were made to identify probes where B > D > C > A. Exon body probes demonstrated the ability to titrate slightly more efficiently than junction probes for both groups and were more efficient in detecting the proper response (Figure 2).

The probe configuration was used to generate a total gene level expression value that was also compared back to more conventional, 3' biased microarrays, such as the U133 Plus 2.0 GeneChip. The gene level

Event Type	Number of events
Total putative alternative splicing events with cDNA supporting evidence	138,636
novel exon	46,352
novel exons	9,413
exon skipped	31,163
exons skipped	11,905
alternative splice donor (ASD)	10,281
alternative splice acceptor (ASA)	12,606
intron retention	5,999
novel intron	10,917
predictions of single exon skips with no supporting cDNA evidence	142,697
introns with structural probes for prediction of ASA, ASD and intron retentions	176,000

Table 1. Distribution of splice events from an analysis of the human genome. Results from the human genome analysis for splicing events from 20,649 genes is tabulated according to the type of splice event. (The complete set of information is available online at portal.splicearray.com.)

Samples	CV Median (%)	CV Mead (%)
A	4.45	5.81
B	4.2	5.56
C	3.99	5.16
D	4.33	5.6

Table 2. Coefficients of variation for the Human Genome Wide SpliceArray™. All four samples were analyzed in triplicate and the CV% were calculated for each probe set, and the median and mean values for each sample is indicated in the table.

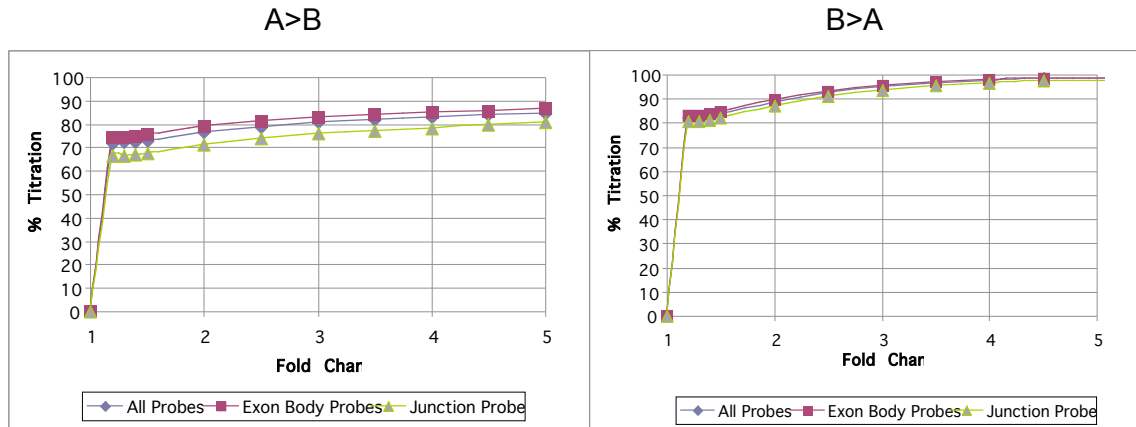


Figure 2. Titration analysis of probe set response. Probes were selected for analysis based on the expression results detecting different transcripts. 527,574 probes were identified where expression was higher in sample A than sample B (A>B, left graph) and 472,085 probes where B>A (right graph). Probes were binned based on the fold change between samples A and B, and assessed for correct titration of expression values. Probes were filtered statistically and those with p values <0.001 were analyzed. Probe types were separated according to exon body and junction probes.

expression values were calculated from the array data using the probes common to most isoforms of the gene (probes F and T, Figure S2). These probes are common to many forms of the gene, and as such represent all common features for all transcripts at a locus. For genes that were found to be significantly differentially expressed between samples A and B, it can be seen that the majority illustrate the proper response. On average, 87% of genes with greater than a 2.0 fold change between samples A and B demonstrated correct titration of all four samples (Figure 3, Table 3), indicating the array quantifies gene expression at the gene level in an accurate manner. As expression values generated on different platform parameters, labeling methods, probe sequences, etc., we compared the relative expression difference between samples A and B. We map genes from the U133 Plus 2.0 GeneChip back to the Human Genome Wide SpliceArray™ and a comparison of the annotations from each array provided 11,089 genes in common and 7,147 genes expressed after removing genes exhibiting low expression. The probe sets were

highly correlated between the two platforms with a linear correlation of 0.85 (Figure 4, Table 3). When considering probe sets that showed a 2 fold change between samples, the coefficient rose to 0.90, and increased as the fold change for the probe set increased, indicating that the Human Genome Wide SpliceArray™ provides highly concordant quantification of gene expression, when compared to the U133 Plus 2.0 GeneChip.

The Human Genome Wide SpliceArray™ has the ability to detect alternatively spliced genes due to the extensive design of the probes on the array. The set of genes identified as expressed above background were separated into three groups: genes that are not differentially expressed between samples A and B (fold changes within the range of $-1.2 < x < 1.2$), genes with a greater than 3-fold change ($x < -3$ or $x > 3$); and genes differentially expressed between 1.2 and 3-fold (this gene set was not investigated). In order to determine the extent of differential splicing occurring for the first two gene categories, a splicing ratio was calculated that consisted of the long form probe (B) divided by the

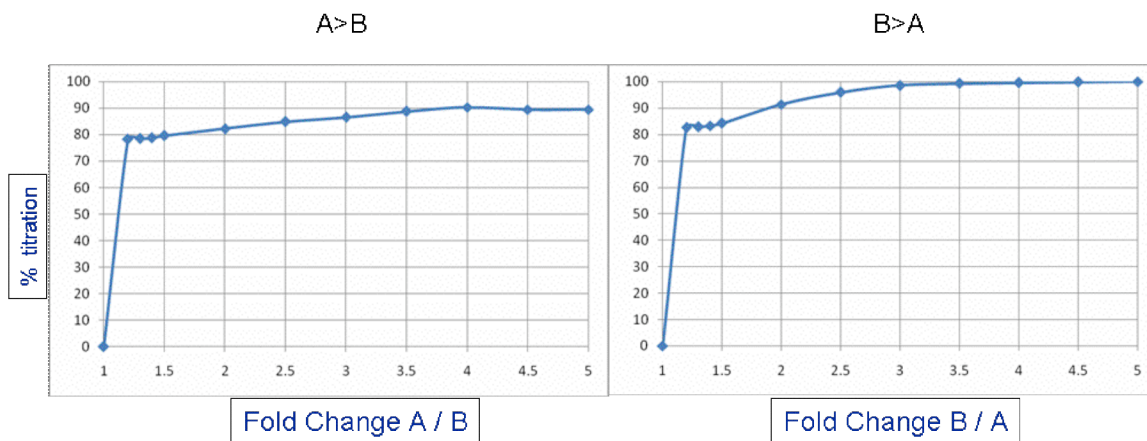


Figure 3. Titration response based on gene level analysis. To calculate the gene level expression values, the median expression value was calculated from probes common to both isoforms (the F and T probes). The fold change for each gene was determined between samples A and B, binned and titration was assessed as described above. The left graph indicates genes which were more highly expressed in sample A than sample B (A > B), and the right graph displays genes which were more highly expressed in sample B versus sample A (B > A).

Probe Set (FC)	Correlation
All Probes	0.85
1.5 FC	0.90
2.0 FC	0.92
3.0 FC	0.933

Table 3. Correlation of gene level analysis between the U133 plus 2.0 array and the Human GWSA™. Gene level analysis was calculated for the Human GWSA and the U133 Plus 2.0 array. As intensity values were significantly different due to different formats, we compared the fold change between the universal (sample A) and brain (sample B) between the two designs. The correlation between the two platforms was higher when genes that were found to be differentially expressed were considered distinctly, as would be expected where genes showing no differential expression would exhibit a higher variability.

short form probe (E), called the B/E ratio (see Figure S2 for probe design), and events that had 3-fold change or greater in this ratio between samples A and B were identified. Greater than 72% of the significant events titrated all four samples correctly demonstrating the extensive quantification available on the array (not shown).

Extensive alternative splicing activity was identified in two important categories; genes that show high level of differential expression at the gene level (greater than 3 fold change), and genes where no differential change was detected, but splicing alterations were identified (Table 4). 1,844 genes were found to be differentially expressed greater than 3 fold at the gene level, and 13,323 potential splicing events were identified within

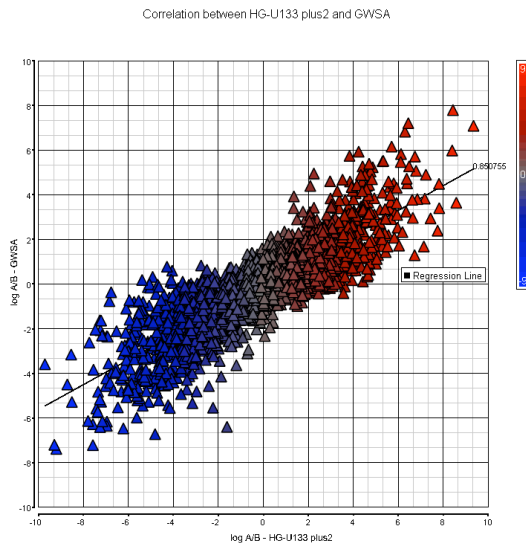


Figure 4. Gene level comparison of the U133 plus 2 array with the Human GWSA. Gene annotations from the Human GWSA™ were mapped against the U133 Plus 2.0 array and after removing the low expressed genes, 7147 genes remained for comparative analysis. Fold changes for each gene were calculated between samples A and B for each platform and plotted against each other. Correlation coefficients (found in table 3) were calculated based on fold change gene sets. A direct comparison of the intensity values was prevented due to the significant differences between the platforms (5 micron versus 11 micron feature size, different labeling technology, and probe sequences) and prevented a direct comparison of the intensity values. Therefore, an assessment of the ability to detect changes between samples A and B was performed to assess the performance of each platform.

Fold Change Set	Total Genes	Total Events	B/E Ratio >3 FC
> 3 or < -3 FC	1844	13323	7883
-1.2 < X < 1.2	4434	32917	3021

Table 4. Event analysis of genes grouped by overall gene expression. Genes were classified into two groups: genes that were not differentially expressed between samples A and B (fold changes within the range of $-1.2 < x < 1.2$), and genes with a greater than 3-fold change ($x < -3$ or $x > 3$). In order to determine the differential splicing occurring for each gene, the ratio of the long form probe (B) and the short form probe (E) was calculated, and events were selected where the ratio was differentially expressed between samples A and B by greater than 3-fold. On average, greater than 71% of the significant events titrated all four samples correctly (not shown).

these genes. Of these potential events, 7,883 or 59% of these events had B/E ratios that were differentially expressed by greater than 3 fold indicating a change in the isoforms for these genes. In addition, 4,434 genes displayed highly similar expression levels between samples A and B (less than 1.2 fold change between the samples), and contained 32,917 potential splicing events. 9% or 3021 from these events displayed a 3 fold expression change in the B/E ratio. Significant splicing changes were detected in groups of genes regardless of their overall gene expression characteristic and demonstrate that transcription and alternative splicing are independent events.

In order to validate the array analysis and the analytical methods, a total of 169 events were selected from both categories. First, events were selected from genes which did not show a significant difference in expression between the universal RNA and the human brain RNA samples (median of F and T probe values ≤ 1.2 or ≥ 1.2), but did give a B/E ratio of >3 , indicating that changes in splicing were occurring. 81% of the events chosen from this list were validated by RT-PCR, and the results showed that the isoform ratio was altered in both samples relative to the expression levels from the GWSA data. Of the validated events, 56% were novel exon(s), 22% were exon skips, and 22% were intron retention events. One such event was contained in the LIM domain 7 gene which was found to be expressed at similar levels in samples A and B; however, there was an 8 fold change in the B/E ratio for event 4008.10.2, an exon skip event of exon 28 (Figure 5A). PCR validation demonstrated that the reference or long form was expressed only in sample A (universal RNA) while both forms are expressed at approximately equal amounts in sample B (brain, Figure 5B). The exon skip event is an out of frame deletion that alters the translation frame and results in a loss of 10 amino acids in length with the formation of an early stop codon with a loss of the C-terminal LIM domain (Figure 5C).

Events were also selected from the list of genes which did show significant gene level differences in expression and also showed a change in the B/E ratio of ≥ 3 . 98% of these events were validated by RT-PCR, showing altered isoform ratios in both samples. 60% were exon skip events, 40% were novel exons. An example of an exon skip event is illustrated by the erythrocyte membrane protein band 4.1-like 3 (EMPB41L3) which was found to be more highly expressed in brain than in the universal RNA sample. Event 23136.004.4, an exon skip event of exon 17 (Figure 6A) was also found to have a change in alternative splicing indicated by an 11-fold upregulation

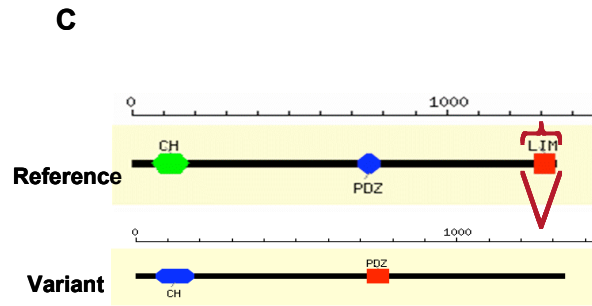
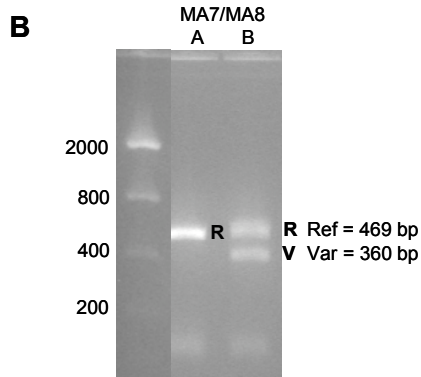
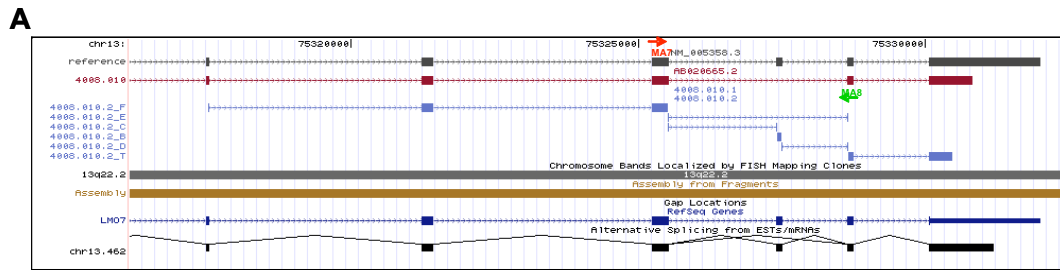


Figure 5. Lim Domain 7 Splicing alteration. Lim domain 7 protein (LMO7) was selected from a set of genes that displayed little difference at the gene level between samples A and B, but had event ratios that were significantly higher than 3 fold between these samples, indicating a splicing alteration. A) Using the UCSC browser, the event monitored is illustrated along with the variant used to identify the event, the probe target regions, and the PCR primers used to validate the event. B) RT-PCR results for the event validation using the primers indicated in A. C) Protein domain analysis was done using CDART (30) and shows the C terminal Lim domain is missing from the splice variant.

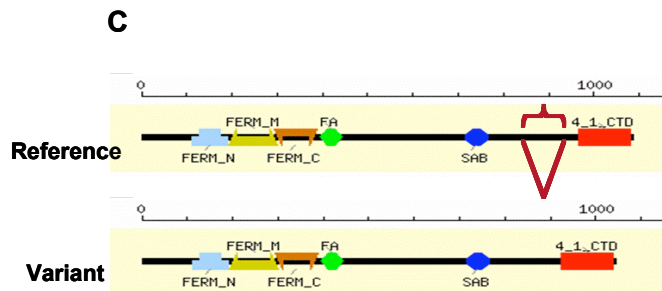
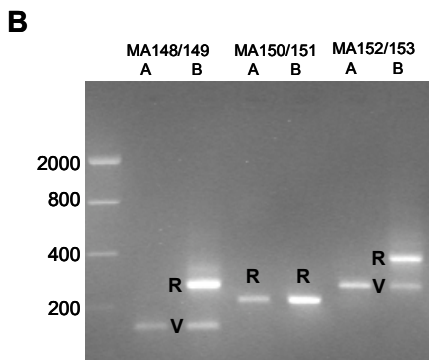
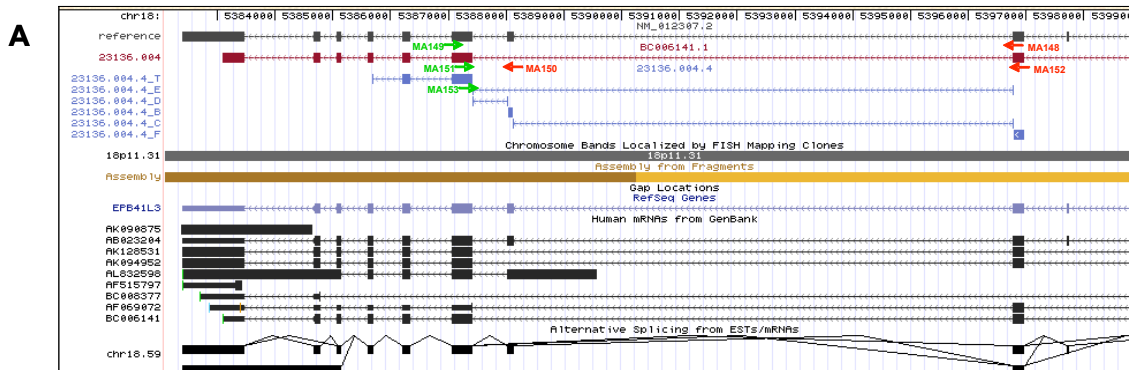


Figure 6. Erythrocyte membrane protein band 4.1-like 3 (EMP41L3) Splicing alteration. EMP41L3 was selected from a set of genes that displayed a large difference at the gene level between samples A and B, and, in addition, had event ratios that were significantly higher than 3 fold between these samples, indicating a splicing alteration. A) Using the UCSC browser, the event monitored is illustrated along with the variant used to identify the event, the probe target regions, and the several pairs of PCR primers used to validate the event. B) RT-PCR results for the event validation using the primers indicated in A. C) Protein domain analysis was done using CDART (30) and shows a portion of the C terminal region is removed and may affect the function of the protein.

of the B/E ratio in brain. RT-PCR analysis clearly indicates that brain RNA contains a significantly higher amount of the long form (the reference), while the short form (the exon skip event) is equal or potentially higher in the universal RNA sample (Figure 6B). The exon skip event results in an in-frame deletion, leading to the loss of 41 amino acids in the C-terminal portion of the protein, between the spectrin-actin domain (SAB), and the 4.1 C-terminal domain (Figure 6C).

Discussion

Gene expression analysis has provided an important resource for investigators to identify genes involved in different biological processes, and has been used to generate profiles where the expression level of a predefined set of genes can help to identify and predict a variety of pathological states and prognoses (14,15). However, many of these studies have ignored the large diversity of transcripts that are generated from each locus by alternative splicing, and have treated each gene as a single entity. As such, singular gene expression measurement can be misleading and may not define splice variant transcripts and their corresponding translated isoforms that bear etiological relevance. This large diversity of transcripts now renders the dogma of one gene, one protein as an invalid concept, and it is quickly becoming clear that the vast majority of genes, potentially over 90% in our analysis, undergo alternative splicing, producing many different proteins from a single gene or locus that can be heavily influenced by pathological states (16,17). Studies have been performed where probes have been reordered into new probe sets to define transcripts more accurately (18-20). The attempt to identify different transcripts in these studies suffers from the same flaw, that the original design of the microarray was not focused on detecting alternative transcripts *per se*, but to provide an accurate measurement of the overall expression of each gene, not transcript. The difficulty in quantifying different transcripts with this design is due to the high similarity of the transcript sequences. As such, by using small oligonucleotide probes, it becomes an extremely difficult task to identify expression of distinct transcripts. Sets of probes can be designed to detect differences in expression of a specific splice event. There is one report (21) where probes were designed and used to predict the transcript expression level with success, however, knowledge of all potential transcripts is one requirement for this method that is not available on a genome-wide scale. Although one can assume some composition for the transcriptome, there is sufficient ignorance of the complete collection of transcripts and isoform regulation to render this approach unreasonable.

The Human Genome Wide SpliceArray™ was designed to detect splice events in all potential coding frames. Each gene was analyzed for alternatively spliced events based on several criteria; evidenced through expressed sequences available from GenBank, or through the prediction of exon skip events. Probes were designed for each identified event, whether evidenced or predicted, and the arrays were shown to be highly reproducible, suggesting that the reproducibility is more a function of the platform and labeling procedures than of the probe design.

The probes and probe sets demonstrated similar performance on this platform as compared to, and reported on these platforms for the MAQC project (10). Interestingly, a similar response was found when the data was analyzed at the gene or event level. Analysis of the array data at the gene level proved highly concordant with data produced in the MAQC project on the U133 Plus 2.0 array. Even though the platforms were slightly different (the U133 Plus 2.0 array contains 11 micron features while the SpliceArray™ contains 5 micron features), the genes showed similar levels of expression, and importantly, similar quantified changes between the universal and brain RNA samples. The titration clearly showed a high concordance between these similar platforms with probe sets that were designed for different applications. In addition, the SpliceArray™ also provides splice event information available from the same experiment.

Several groups have simultaneously produced microarrays to monitor splice events, and interestingly have used almost the exact same design (11,22,23). Many methods have been developed to determine the extent of alternative splicing and identify the expression level for each transcript within a gene. We used a simple approach to identify cases where the ratio of the sequence inclusive and exclusive forms changed dramatically, by calculating the ratio of probes that were specific for the inclusive or long form (i.e. probe B) versus probes specific for the exclusive or short form (i.e. probe E). This approach was shown to provide very powerful results when assessed against PCR validation of the events. Although the approach does suffer from the inability to determine actual transcript levels, it is a simple and applicable approach that provides strong filtering methods for real positives, unlike other approaches which require an extensive programming and array fitting, and are not easily applicable to most investigators (23).

RT-PCR validation was performed based on the analysis of the ratio of long to short form. This type of assessment is an important aspect to determining the extent of alternative splicing as in many cases the ratio of isoforms will determine the biological response (24-27). By assessing the statistical power of the ratio, we were able to validate a high rate of events; overall, 81% of statistically selected events were validated by RT-PCR. The high rate of validation was found regardless of the overall transcriptional changes identified. Interestingly, almost 60% of the events identified to have gene level, transcriptional alterations of 3 fold or more were found to have an event fold change of 3 fold or more. This indicates that not only is there a change in overall gene expression, but the transcripts produced are different as generated by alternative splicing. Surprisingly, genes that were found to not change in their expression levels represented a rich source of alternative spliced transcripts. Almost 10% of these genes contained evidence of splicing alterations by changes in the ratio of the different isoforms. This is an important finding as it brings a point that genes which have been ignored because of equivocal expression between samples actually have important changes that occur in producing different isoforms. Importantly, this suggests that there is a clear separation between general transcription and alternative splicing. This

confirms earlier evidence (28) that illustrates transcriptional activity is independent of splicing, even though the two processes function simultaneously. These findings should encourage investigators to look more closely at the genes they discard for important clues in mechanisms of biology.

Alternative splicing affects not only the structure of the mRNA but ultimately the structure of the protein produced. Many exons encode protein domains that can be removed by an exon skip event (29) and can lead to the production of a transcript lacking functional domains, potentially producing dominant negative proteins, constitutively active isoforms, creating soluble homologues, or regulate the overall activity of the protein by reducing the overall expression of the protein. The Lim domain 7 (LMO7) was identified as one example where there was no overall change in gene expression but significant change in splicing was identified. Lim domains are protein-protein interaction domains and found in many key regulators of development. In addition, LMO7 contains a calponin homology domains (CH), an actin binding domain found in cytoskeletal and signal transduction proteins; and PDZ domains, which bind protein and lipid, and are modified by phosphorylation (30). As LMO7 mutants display retinal, muscular, and growth retardation, different isoforms will impact the function of this protein and potentially be altered in disease states. Multiple alternative splice variants have been described (31). The second example was differentially expressed at the gene level as well as at the transcript level. The erythrocyte membrane protein band 4.1-like 3 (EMPB41L3) is highly expressed in brain, and loss of

heterozygosity is found in 60% of meningiomas and represents an early event in tumorigenesis, suggesting that these proteins are critical growth regulators in meningioma pathogenesis (32). Normally expressed at high levels in brain, with lower levels in kidney, intestine, and testis, the function of the variant is unknown, yet contains some well described protein domains, including Ferm domains found in many cytoskeletal proteins, the N domain found in ubiquitin-like structural domain, and the C domain found in tyrosine phosphatases (30). Different isoforms certainly will affect the function of these different domains and potentially impact the function and may play a role pathological states.

New definitions are being proposed in order to correctly talk about expression studies and this is important so that we all speak about the same entities. Genes are no longer considered a single entity and are now considered more of chromosomal regions of transcriptional activity (33). Transcripts can be thought of as specific structures produced from each genomic location, and as such, have a many-to-one relationship with genes. The identification of expressed transcripts is at the heart of expression analysis and these examples, as well as recently identified novel spliced isoforms demonstrating diverse activity (4,17) illustrate the importance of correct determination of the expression of each transcript. Tools such as arrays specifically designed to detect the differences in transcripts will allow researchers to more fully explore the transcriptome under different physiological and pathological conditions.

References:

1. Lander et al., International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15;409(6822):860-921.
2. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*. 2003;72:291-336.
3. Heinzen EL, Yoon W, Weale ME, Sen A, Wood NW, Burke JR, Welsh-Bohmer KA, Hulette CM, Sisodiya SM, Goldstein DB. Alternative ion channel splicing in mesial temporal lobe epilepsy and Alzheimer's disease. *Genome Biol*. 2007;8(3):R32.
4. Einstein R, Jordan H, Zhou W, Brenner M, Moses EG, Liggett SB. Alternative splicing of the G protein-coupled receptor superfamily in human airway smooth muscle diversifies the complement of receptors. *Proc Natl Acad Sci U S A*. 2008 Mar 24; [Epub ahead of print]
5. Boue S, Letunic I, Bork P. Alternative splicing and evolution. *Bioessays*. 2003 Nov;25(11):1031-4.
6. Calarco JA, Xing Y, Cáceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, Blencowe BJ. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev*. 2007 Nov 15;21(22):2963-75.
7. Birzele F, Küffner R, Meier F, Oefinger F, Potthast C, Zimmer R. ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res*. 2008 Jan;36(Database issue):D63-8.
8. MAQC Consortium, Shi L et al., The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006 Sep;24(9):1151-61.
9. Canales RD, Luo Y, Willey JC, Austermler B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY, Ma Y, Maqsoodi B, Papallo A, Peters EH, Poulter K, Ruppel PL, Samaha RR, Shi L, Yang W, Zhang L, Goodsaid FM. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol*. 2006 Sep;24(9):1115-22.
10. Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E, Sun YA, Willey JC, Thierry-Mieg J, Thierry-Mieg D, Setterquist RA, Wilson M, Lucas AB, Novorodovskaya N, Papallo A, Turpaz Y, Baker SC, Warrington JA, Shi L, Herman D. Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat Biotechnol*. 2006 Sep;24(9):1123-31.
11. Fehlbaum P, Guihal C, Bracco L, Cochet O. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res*. 2005 Mar 10;33(5):e47.
12. Resch A, Xing Y, Alekseyenko A, Modrek B, Lee C. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res*. 2004 Feb 24;32(4):1261-9.
13. Hide WA, Babenko VN, van Heusden PA, Seoighe C, Kelso JF. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res*. 2001 Nov;11(11):1848-53.
14. He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002 Dec 19;347(25):1999-2009.
15. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, Robertson JF, Aparicio S, Ellis IO, Brenton JD, Caldas C. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*. 2007 Mar 1;26(10):1507-16.
16. Kozyrev SV, Abelson AK, Wojcik J, Zaghlool A, Linga Reddy MV, Sanchez E, Gunnarsson I, Svenungsson E, Sturfelt G, Jönsen A, Truedsson L, Pons-Estel BA, Witte T, D'Alfonso S, Barrizzone N, Danielli MG, Gutierrez C, Suarez A, Junker P, Lauststrup H, González-Escribano MF, Martin J, Abderrahim H, Alarcón-Riquelme ME. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet*. 2008 Feb;40(2):211-6.

17. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*. 2008 Mar 13;452(7184):230-3.
18. Hu GK, Madore SJ, Moldover B, Jatkoa T, Balaban D, Thomas J, Wang Y. Predicting splice variant from DNA chip expression data. *Genome Res*. 2001 Jul;11(7):1237-45.
19. Fan W, Khalid N, Hallahan AR, Olson JM, Zhao LP. A statistical method for predicting splice variants between two groups of samples using GeneChip expression array data. *Theor Biol Med Model*. 2006 Apr 7;3:19.
20. Lu J, Lee JC, Salit ML, Cam MC. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*. 2007 Mar 29;8:108.
21. Wang H, Hubbell E, Hu JS, Mei G, Cline M, Lu G, Clark T, Siani-Rose MA, Ares M, Kulp DC, Hausssler D. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*. 2003;19 Suppl 1:i315-22.
22. Le K, Mitsouras K, Roy M, Wang Q, Xu Q, Nelson SF, Lee C. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*. 2004 Dec 14;32(22):e180.
23. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ, Blencowe BJ. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*. 2004 Dec 22;16(6):929-41.
24. Lord KA, Creasy CL, King AG, King C, Burns BM, Lee JC, Dillon SB. REDK, a novel human regulatory erythroid kinase. *Blood*. 2000 May 1;95(9):2838-46.
25. Meng J, Tsai-Morris CH, Dufau ML. Human prolactin receptor variants in breast cancer: low ratio of short forms to the long-form human prolactin receptor associated with mammary carcinoma. *Cancer Res*. 2004 Aug 15;64(16):5677-82.
26. Frey UH, Nüchel H, Dobrev D, Manthey I, Sandalcioglu IE, Eisenhardt A, Worm K, Hauner H, Siffert W. Quantification of G protein Gaalphas subunit splice variants in different human tissues and cells using pyrosequencing. *Gene Expr*. 2005;12(2):69-81.
27. Diernfellner AC, Schafmeier T, Meroz MW, Brunner M. Molecular mechanism of temperature sensing by the circadian clock of *Neurospora crassa*. *Genes Dev*. 2005 Sep 1;19(17):1968-73. Epub 2005 Aug 17.
28. Ip JY, Tong A, Pan Q, Topp JD, Blencowe BJ, Lynch KW. Global analysis of alternative splicing during T-cell activation. *RNA*. 2007 Apr;13(4):563-72.
29. Xing Y, Lee CJ. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet*. 2005 Sep;1(3):e34.
30. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. *Genome Res*. 2002 Oct;12(10):1619-23.
31. Putilina T, Jaworski C, Gentleman S, McDonald B, Kadiri M, Wong P. Analysis of a human cDNA containing a tissue-specific alternatively spliced LIM domain. *Biochem Biophys Res Commun*. 1998 Nov 18;252(2):433-9.
32. Tran YK, Bögler O, Gorse KM, Wieland I, Green MR, Newsham IF. A novel member of the NF2/ERM/4.1 superfamily with growth suppressing properties in lung cancer. *Cancer Res*. 1999 Jan 1;59(1):35-43.
33. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbil JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. What is a gene, post-ENCODE? History and updated definition. *Genome Res*. 2007 Jun;17(6):669-81.

Supplemental Figures

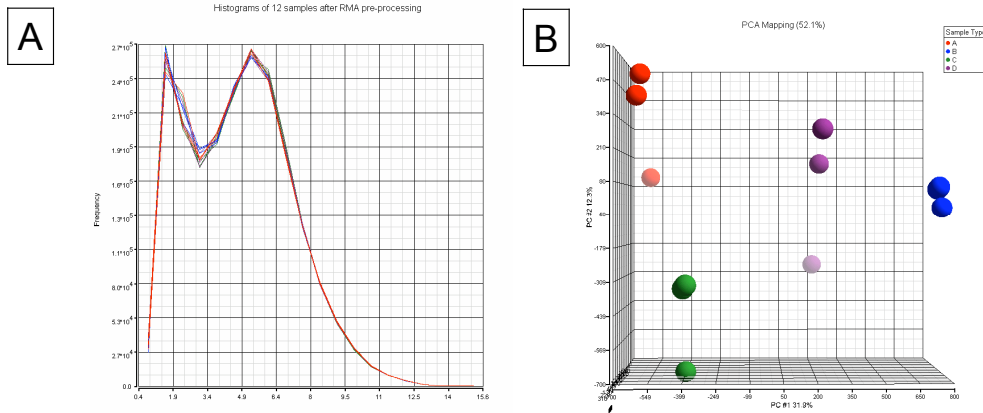


Figure S1. Analysis of the Human GWSA. The four samples described in the manuscript were analyzed on the SpliceArray™. A) A frequency distribution is shown after normalization of the entire data set. B) A Principle Component Analysis (PCA) was performed on the normalized distribution and displays the correct orientation of the samples (sample A on the left, followed by samples C, D, and B moving to the right), as well as good reproducibility among the replicates on the array.

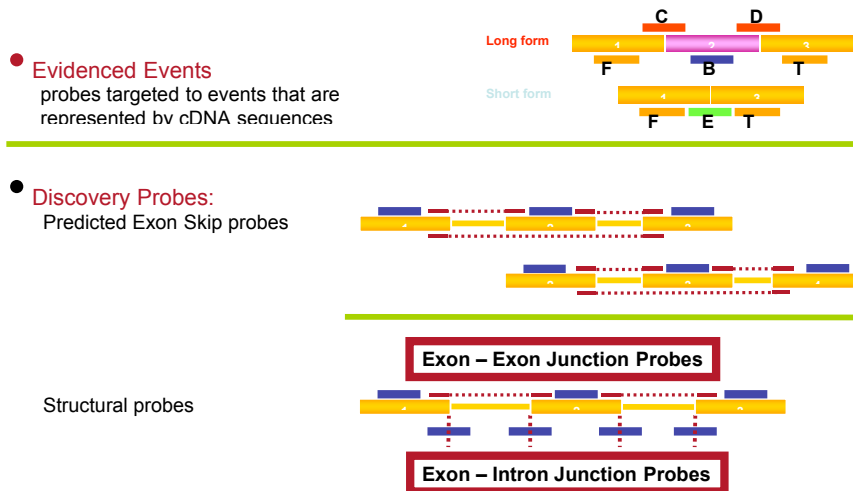


Figure S2. Probe configuration for the Human Genome Wide SpliceArray™. Potential splice events were identified after investigation of expressed sequence from the public databases. These “evidenced events” were selected and probes were designed to detect both the short, exclusive form and the long, inclusive form. F and T probes were common to both forms while B, C, and D probes are exclusive to the long form, while the E probe is specific only for the short form. Discovery probes come in two types: predicted exon skips and structural probes. If an exon was not detected to be skipped from the sequence analysis, then the exon was predicted to be skipped and probes were designed to detect each prediction. Structural probes are designed against the exon-intron structure of the gene, such that each intron contains three probes; an exon-exon junction probe, an exon-intron and intron-exon probe to monitor exon extension events.



ExonHit Therapeutics, Inc.
217 Perry Parkway, Building #5
Gaithersburg, MD 20877 USA
Toll free US and Canada 1-888-811-7070
International +1-240-404-0333
E-mail splicearray@exonhit.com
www.exonhit.com



Agilent
Certified
Services Provider
Microarray-Based
Genomic Analysis