(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 November 2007 (08.11.2007)

PCT

(10) International Publication Number
WO 2007/126882 A2

(51) International Patent Classification:
C12M 3/00 (2006.01)

(21) International Application Number:
PCT/US2007/007671

(22) International Filing Date: 27 March 2007 (27.03.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/786,550	27 March 2006 (27.03.2006)	US
60/847,278	25 September 2006 (25.09.2006)	US
60/900,679	8 February 2007 (08.02.2007)	US

(71) Applicant *(for all designated States except US)*: **JIVAN BIOLOGICS, INC.** [US/US]; 733 Allston Way, Berkeley, CA 94710 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BINGHAM, Jonathan** [US/US]; 615 Cole Street, #4, San Francisco, CA 94117 (US). **SRINIVASAN, Subha** [US/US]; 20 Corte Patencio, Greenbrae, CA 94904 (US).

(74) Agents: **TODD, Stephen** et al.; Perkins Coie LLP, P.O. Box 2168, Menlo Park, CA 94026 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/126882 A2

(54) Title: ANALYSIS OF SPLICE VARIANT EXPRESSION DATA

(57) Abstract: Indicator molecules, devices, algorithms, software, systems, and methods are described for qualitatively and quantitative measuring the levels of splice variants and determining splicing patterns in a one or more samples.

ANALYSIS OF SPLICE VARIANT EXPRESSION DATA

TECHNICAL FIELD

The compositions, devices, methods, systems, and kits of parts relate to the fields of gene expression and to microarray based methods for measuring gene expression, particularly the expression of splice-variants.

BACKGROUND

Microarray technology has expanded beyond genes to address other applications, including genotyping, comparative genome hybridization, chip-on-chip, genome tiling, and, more recently, alternative RNA splicing. To date, various research groups have conducted: a genome-wide study of alternative splicing in yeast using exon and exon-exon junction probes (Clark, T.A. *et al.* (2002) *Science* **296**:907-910); a human genome-wide investigation of tissue specific splicing using exon-exon junction probes (Johnson, J.M. *et al.* (2003) *Science* **302**:2141-2144); a map of splicing in the brain with junction probes (Ule, J. *et al.* (2005) *Nat Genet.* **37**:844-852); a study of the effects of cancer on splicing in breast cell lines (Li. C, *et al.* (2006) *Cancer Research* **66**:1990-99); and a family-wide genetic study of tissue-specific splicing (Bingham, J.S. *et al.* (2006) *Biochem Biophys Res Commun.*, in press).

RNA splicing is generally agreed to affect a majority of human genes, probably more than 70% (Kan, Z. *et al.* (2001) *Genome Res.* **11**:889-900; Johnson, J.M. *et al.* (2003) *Science* **302**:2141-2144), with important functional implications for protein coding in normal physiology as well as the pathogenesis of disease states. The nuclear machinery performing the splicing reactions comprises five small RNAs and greater than 300 different proteins (Nilsen, T.W. (2003) *Bioessays* **25**:1147-9), exhibiting a level of complexity, in function and composition, that rivals transcriptional and ribosomal systems (Tarn, W.Y. and Steitz, J.A. (1997) *Trends Biochem Sci.* **22**:132-7; Will, C.L. and Luhrmann, R. (1997) *Curr Opin Cell Biol.* **9**:320-8). The reactions executed by this macromolecular complex must be highly coordinated to allow for precise recognition and coupling of exons amongst a sea of introns that average over 3 KB in size (Deutsch, M.R. and Long, N. (1999) Nucleic Acids Res **27**:3219-28). Even

a single nucleotide mistake produced by inappropriate processing can result in a frame shift with potentially catastrophic consequences on translation of the message (Staley, J.P. and Guthrie, C. (1998) *Cell* 1998. **92**:315-26). Perhaps 15% of all human genetic diseases are caused by abolition of essential splice sites or the creation of novel ones (Krawczak, M. *et al.* (1992) *Hum Genet.* **90**:41-54), while other disease-causing mutations are known to impact splicing regulatory elements (Caceres, J.F. and Kornblihtt, A.R. (2002) *Trends Genet.* **18**:186-93). Public databases provide a wealth of information on alternate splicing, notably the Alternative Splicing Annotation Project (ASAP) (Lee, C. *et al.* (2003) *Nucleic Acids Res.* **31**:101-105), Alternative Splicing Database (ASD) (Stamm, S. *et al.* (2006) *Nucleic Acids Res.* **34**:D46-D55), and ECgene (Kim, N.I. *et al.* (2005) *Genome Res.* **15**:566-576).

Data analysis of expression levels is non-trivial. Current methods of splice array data analysis have focused on per-probe analysis using log ratios (Johnson, J.M. *et al.* (2003) *Science* **302**:2141-2144) and on linear equations for combinations of probes that skip or include a given exon, intron or exon extension (Clark, T.A. *et al.* (2002) *Science* **296**:907-910; Fehlbaum. P. *et al.* (2005) *Nucleic Acids Res.* **33**:e47). Algorithms include ASAP, Splicing Index, SPLICE, ASPIRE, and ANOSVA among others.

However, conventional gene expression analysis and transcript expression analysis generally do not consider variations among the expression levels of indicator polynucleotides for the same gene, *i.e.*, they do not, in general, consider the complexities of alternate splicing, which can lead to multiple products of a single gene. Rather, they tend to treat each indicator polynucleotide (*i.e.*, probe) as an independent measurement, or, alternatively, aggregate the expression levels of indicator polynucleotides of a gene as a single measurement of overall transcriptional activity.

Yet, even the expression levels of splice products, considered independently, are not the only quantities of biological interest. The relative expression levels of splice variants of a gene in a single sample, and changes in the relative expression levels of splice variants of a gene between and across samples, can yield meaningful insights into splicing regulation that may have

biological function specific to disease state, tissue, intracellular localization, population, individual, drug treatment, etc.

It would be desirable to improve the accuracy of quantifying gene expression levels from splice variant arrays; be able to measure changes in splicing and to differentiate such changes from changes in overall transcriptional activity; to have algorithms and representations appropriate for quantifying, grouping, ranking, and understanding splicing of genes in samples; to filter and visualize data that provides evidence of changes in splicing; and to be able to synchronize between different types of data views. Finally, it would be desirable to place the results of such algorithms within a data configuration, and to transmit the data configuration from one location to another.

REFERENCES

Each of the following references, as well as references cited in the text, are incorporated by reference in their entirety.

Clark, T.A. *et al.* (2002) *Science* **296**:907-910.

Johnson, J.M. *et al.* (2003) *Science* **302**:2141-2144.

Ule, J. *et al.* (2005) *Nat Genet.* **37**:844-852.

Li. C, *et al.* (2006) *Cancer Research* **66**:1990-99.

Bingham, J.S. *et al.* (2006) *Biochem Biophys Res Commun.* (in press).

Kan, Z. *et al.* (2001) *Genome Res.* **11**:889-900.

Nilsen, T.W. (2003) *Bioessays* **25**:1147-9.

Tarn, W.Y. and Steitz, J.A. (1997) *Trends Biochem Sci.* **22**:132-7.

Will, C.L. and Luhrmann, R. (1997) *Curr Opin Cell Biol.* **9**:320-8.

Deutsch, M.R. and Long, N. (1999) *Nucleic Acids Res* **27**:3219-28.

Staley, J.P. and Guthrie, C. (1998) *Cell* 1998. **92**:315-26.

Krawczak, M. *et al.* (1992) *Hum Genet.* **90**:41-54.

Caceres, J.F. and Kornblihtt, A.R. (2002) *Trends Genet.* **18**:186-93.

Lee, C. *et al.* (2003) *Nucleic Acids Res.* **31**:101-105.

Stamm, S. *et al.* (2006) *Nucleic Acids Res.* **34**:D46-D55.

Kim, N.I. *et al.* (2005) *Genome Res.* **15**:566-576.

Fehlbaum. P. *et al.* (2005) *Nucleic Acids Res.* **33**:e47.

Wang, H. *et al.* (2003) *Bioinformatics* **19**:315-322.

Kim, P. *et al.* (2005) *Nucleic Acids Res.* **33**:D75-79.

Kim, N. *et al.* (2007) *Nucleic Acids Res.***35**:D93-8.

Takeda, J. *et al.* (2007) *Nucleic Acids Res.* **35**:D104-9.

Pospisil, H. *et al.* (2004) *Nucleic Acids Res.* **32**:D70-4.

Zheng, C.L. *et al.* (2005) *RNA* **11**:1767-76.

Holste, D. *et al.* (2006) *Nucleic Acids Res.* **34**:D56-62.

Huang, H.D. *et al.* (2005) *Nucleic Acids Res.* **33**:D80-D85.

Pan, Q. *et al.* (2006) *Genes Dev.* **20**:153-8.

Pan, Q. *et al.* (2004) *Mol Cell.* **16**:929-41.

Sugnet, C.W. *et al.* (2006) *PLoS Comput Biol.* **2**:e4.

Le, K. *et al.* (2004) *Nucleic Acids Res.* **32**:e180.

Srinivasan, K. *et al.* (2005) *Methods* **3**:345-59.

Hu, G.K. *et al.* (2001) *Genome Res.* **11**:1237-1245.

Cline, M.S. *et al.* (2005) *Bioinformatics* **21**:107-15.

Cuperlovic-Culf, M. *et al.* (2006) *Drug Discov Today* **11**:983-90.

Wang, B.B. *et al.* (2006) *Proc Natl Acad Sci U S A.* **103**:7175-7180.

Nuwaysir, E.F. *et al.* (2002) *Genome Res.* **12**:1749-55.

U.S. Patent No. 6,881,571.

U.S. Patent Pub. Nos. 2003-0087261, 2004-0076959, and 2005-0017981.

SUMMARY

The following aspects and embodiments thereof, described and illustrated below are meant to be exemplary and illustrative, not limiting in scope.

In one aspect, a method for determining an expression level for a gene with at least a first splice isoform and a second splice isoform is provided, the method comprising:

(a) obtaining microarray expression level data for a plurality of mutually exclusive indicator polynucleotides for exons, introns, modules, exon-exon junctions, exon-intron junctions, intron-exon junctions or module-module junctions of the gene, and

(b) applying a mathematical algorithm to determine the expression level for the gene.

In some embodiments, the mutually exclusive indicator polynucleotides are non-overlapping. In some embodiments, the mutually exclusive indicator polynucleotides are overlapping.

In some embodiments, at least one mutually exclusive indicator polynucleotide indicates a polynucleotide that is constitutively present in expected splice isoforms of the gene.

In some embodiments, at least one mutually exclusive indicator polynucleotide indicates a polynucleotide that is not constitutively present in expected splice isoforms of the gene.

In some embodiments, the gene expression level is calculated by summing the amount of signal corresponding to the mutually exclusive indicator polynucleotides at each nucleotide base range indicated by the indicator polynucleotides.

In some embodiments, the overall level of gene expression is calculated using the equation:

$$G = \left[\sum_{e} \left(p_e + p_e^c \right) + \sum_{j} \sqrt{\left(p_j + p_{5j}^c \right) * \left(p_j + p_{3j}^c \right)} \right] / P$$

wherein G is the overall gene expression level;

each p_e is separately a signal for an indicator polynucleotide for an exon, intron or module of the gene;

each p_e^c is separately the geometric mean of the signals of indicator polynucleotides complementary to p_e;

each p_j is separately a signal for an indicator polynucleotide for an exon-exon junction, exon-intron junction, intron-exon junction, or module-module junction of the gene, and wherein p_j comprises a 5′ portion and a 3′ portion;

each p_{5j}^c is separately a signal for an indicator polynucleotide that is mutually exclusive with the 5′ portion of p_j;

each p_{3j}^c is separately a signal for an indicator polynucleotide that is mutually exclusive with the 3' portion of p_j;

and P is the total number of included probes.

In some embodiments, a background level of a hybridization signal for a mutually exclusive probe is subtracted from the overall expression level of the probe.

In some embodiments, at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon. In some embodiments, at least one of the plurality of mutually exclusive indicator polynucleotides indicates an intron. In some embodiments, at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon-exon junction. In some embodiments, at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon-intron or intron-exon junction. In some embodiments, at least one of the plurality of mutually exclusive indicator polynucleotides indicates a module-module junction.

In another aspect, software for performing the calculations described above is provided. In some embodiments, the calculation is the overall level of gene expression is calculated using the equation:

$$G = \left[\sum_e \left(p_e + p_e^c\right) + \sum_j \sqrt{\left(p_j + p_{5j}^c\right) * \left(p_j + p_{3j}^c\right)} \right] / P$$

wherein the variables are as defined, above.

In another aspect, a data store comprising expression levels for two or more genes is provided.

In another aspect, a method for identifying alternative splicing of a gene in one or more samples from microarray expression level data for a plurality of indicator polynucleotides for exons, introns, exon-exon junctions, exon-intron junctions, intron-exon junctions or module-module junctions of the gene is provided, the method comprising:

(a) obtaining expression level data for the plurality of indicator polynucleotides in one or more samples;

(b) applying a mathematical algorithm to calculate a value for an alternative

splicing event, wherein the mathematical algorithm involves a gene expression level for the gene; and

(c) identifying the indicator polynucleotides for which the measure exceeds a cutoff value.

In some embodiments, the methods use the calculation described above.

In another aspect, software for performing the calculations described above is provided.

In another aspect, a data store comprising values calculated by a mathematical algorithm are provided. Data stores may contain any of the raw data or data analysis results described above or herein. Splice variant analysis may use data stores in combination with data from a particular experiment.

In a further aspect, a scatter plot of data for a plurality of indicator polynucleotides for one or more genes is provided, wherein the data are the results of applying a mathematical algorithm to the expression levels of the indicator polynucleotides.

In some embodiments, the mathematical algorithm involves a calculation of a gene expression level for each of the one or more genes.

In some embodiments, one or more data points are visually indicated based on results calculated using a mathematical algorithm. In some embodiments, the the visually indicated data points are indicated by a color, hue, saturation, brightness, or transparency level. In some embodiments, the visually indicated data points are indicated by a shape, outline or symbol. In some embodiments, the visually indicated data points are indicated by a label.

In addition to the exemplary aspects and embodiments described above, further aspects and embodiments will become apparent by reference to the drawings and by study of the following descriptions.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a pie chart showing the relative frequency of different splice types in the two cells lines tested.

Figure 2 is a scatter plot comparing the splicing patterns in two cell lines. The left panel shows expression of exon-exon junction, exon-intron junction and exon probes for a genome-wide splice array in MiaPaCa2 vs. HEK293 cell lines. $R^2 = 0.889$. The right panel shows the same data, normalized by the level of gene expression using Equation 2. $R^2 = 0.922$.

Figure 3 shows a series of scatter plots depicting splice variants with a minimum signal of 200 and a "Splice Fold" (*i.e.*, linearized Splice Ratio) score of ≥ 2 (>99.9% confidence for all splice types). Plots on the left (*i.e.*, plots A, C, E, and G) show MCF7 cells vs. CaCO2 cells; plots on the right (*i.e.*, B, D, F, and H) show technical replicates of HEK293 cells. The indicator polynucleotides identified the following types of splice variants: A and B, exon skips and includes; C and D, alternative first and last exons; E and F, alternative donor and acceptor sites; and G and H intron retentions and splices.

Figure 4 is graph showing the distribution of differentially expressed splice types. The splice types and relative amounts of each splice type are indicated. These amounts were determined using the Splicing Index, ASPIRE, and Splice Ratio methods, described.

DETAILED DESCRIPTION

I. Overview

The present indicator molecules, devices and systems comprising indicator molecules, methods for selecting and using indicator molecules, and methods for analyzing splice variant data, allow the detection and quantification of individual splice variants in a biological sample, the quantification of overall gene expression (or transcriptional activity), and the detailed analysis and mapping of essentially all splice variant of one or more genes.

The splice variants may arise from one or more of several different splicing events, some of which are described, below. As used herein, splice variants arise from alternative splicing and/or alternative transcription events that lead to the synthesis of different gene products from the same gene or class of genes. Splice

variants include both polynucleotides and polypeptides encoded by such polynucleotides. Splice variant polynucleotides and polypeptides are collectively referred to as splice isoforms, splice forms, splicesoforms, gene products or isoforms.

In some embodiments, the indicator molecules are polynucleotides that hybridize to the polynucleotides of splice variant RNAs, cRNAs, or cDNAs. Microarrays for use as described herein may be a nucleotide microarray or a protein microarray. A nucleotide microarray may be a spotted oligonucleotide array, an array of *in situ* synthesized oligonucleotides, a bead array, or any other system substantially equivalent to known nucleotide array technology or biochip technology. The microarrays may be two-dimensional or three dimensional. Nucleotide microarrays are described, *e.g.*, in U.S. Patent Nos. 5,202,231, 5,429,807, 5,445,934, 5,525,464, 5,677,195, 5,695,940, 5,700,637, 5,716,785, 5,744,305, 5,800,992, 5,807,522, 5,871,928, 5,891,636, 6,054,270, 6,080,585, 6,110,426, 6,150,095, 6,309,823, 6,379,895, 6,403,957, 6,416,952, 6,480,324, 6,551,784, 6,610,482, 6,770,751, 6,824,866, which are incorporated by reference in their entirety.

In other embodiments, the indicator molecules are polypeptides, such as antibodies, that recognize splice variant polypeptides or fragments encoded by the polynucleotides. Polypeptide (including antibody) arrays are described, *e.g.*, U.S. Patent Nos. 6,777,239, 6,696,620, 6,689,568, 6,448,387, 5,143, and 5,081,584, which are incorporated by reference in their entirety.

In some cases, combinations of exon-exon junction indicator moecules or exon indicator molecules are used to detect simple and well-defined splicing such as the exon skip event described (with limitations, as discussed, herein). In other cases, indictor molecules may indicate exon, introns, exon-exon junctions, exon-intron junction, intron-exon junctions, modules, module-module junctions, or other splice variant features, some of which are described herein. Splice variants of a gene may also result from other splicing events, which are often described by their products, including intron retention events, alternative 3' exons, 3' extensions, alternative 5' exons, 5' extensions, micro-exons, etc.

For each type of splicing event, the present compositions and methods can further be used to determine a minimal (*i.e.*, restricted and sufficient) set of

indicator polynucleotides that are capable of differentiating among different splice variants. Exemplary sets of indicator polynucleotides for use in detecting various splice variants are described. In some embodiments, the present compositions and methods provide a minimal set of indicator polynucleotides for detecting several splice events involving of a gene or group of genes.

Because differential gene expression and differential splicing can be independently controlled, a gene may appear to have the same expression level in two different samples even though the dominant splice isoforms present in each sample differ. Conversely, a gene may have different expression levels in two samples without there being change in the dominant splice variant. To distinguish between gene expression and regulated splicing, it is insufficient to look at differential expression of individual splice variants, which may not reflect the relevant population o splice variants in a sample. Multiple probes for the same gene must be compared determine the relative composition of splice variants in splice variants in a sample.

The present indicator molecules, devices, algorithms, software, data stores, data representations, systems, and methods permit the detailed analysis of splice variant data, providing a far more complete picture of differential gene expression than possible using convention technology.

II. Definitions

Unless otherwise specified, all scientific terms and expressions are as used in the art. Singular terms can include the plural, such that "a," "an," and, "the" can intend both singular and plural.

The following terms are defined for clarity. Other terms are defined in the text. Terms not defined should be given their ordinary meaning as used in the art.

A "portion" of an indicator molecule corresponds to a sequence in a splice variant characterized by a structural or functional significance, such as the polynucleotide sequence of an exon, intron, junction, or module sequence, or the polypeptides encoded by these polynucleotides. For example, where a polynucleotide hybridized across a junction of, *e.g.*, exon 1 and exon 2, the polynucleotide has a portion corresponding to exon 1 and a portion corresponding to exon 3. The meaning of the term "portion" is typically apparent from context.

"Length balancing" means exactly or approximately balancing the length of different portions of an indicator polynucleotide, or different indicator polynucleotides, such that the difference between the length of the first and second portion of an indicator polynucleotide, or first and second indicator polynucleotides, is zero (*i.e.*, they are the same length), at most one base, at most two bases, at most three bases, at most four bases, at most five bases, at most six bases, at most eight bases, or at most ten bases, or even at most twelve bases. A portion of an indicator polynucleotide corresponds to a particular sequence in a splice variant, such as an exon, intron, junction, or module.

"Temperature balancing" means exactly or approximately balancing the melting temperature of different portions of an indicator polynucleotide, or different indicator polynucleotides, such that that the difference between the melting temperatures of the first and second portions of an indicator polynucleotide, or the first and second indicator polynucleotides, is at most 1°C, at most 2°C, at most 3°C, at most 4°C, at most 5°C, at most 6°C, or at most 8°C.

"Tiling a junction" (*e.g.*, exon-exon, exon-intron, intron-exon, exon-exon-exon, intron-exon-exon, exon-exon-intron) means using two, three, four, five, six or more indicator polynucleotides that are shifted across a junction by an increment of 1 base, 2 bases, 3 bases, 4 bases, or 5 or more bases.

A "compound junction indicator polynucleotide," an "indicator polynucleotide for a compound junction," or an "indicator polynucleotide that indicates a compound junction," are equivalent terms relating to an indicator polynucleotide having three or more portions, each portion indicating an exon. Such indicator polynucleotides are useful for detecting very short exons (*e.g.*, micro-exons). Exemplary compound indicator polynucleotides are exon-exon-exon junction indicator polynucleotides, intron-exon-exon junction indicator polynucleotides, exon-exon-intron junction indicator polynucleotides, and exon-exon-exon-exon junction indicator polynucleotides. An exon-exon-exon junction indicator polynucleotide indicates a junction between a first exon and a second exon, and a junction between the second exon and a third exon. There may be one or more intervening (skipped) exons between the first exon and the second exon, or between the second exon and the third exon. An intron-exon-exon junction indicator polynucleotide indicates a junction between an intron and a first exon and

a junction between the first exon and a second exon. There may be one or more intervening (skipped) exons between the first exon and the second exon.

The term "polynucleotide" includes DNA, RNA, and nucleic acid derivatives having modified backbone linkages such as phosphorothioates, methylphosphonates, benzylphosphonates, etc. Such polynucleotides may also have modified sugar groups, such as 2'-methoxyethyl sugars or 2'-methoxyethoxy sugars, or modified bases such as 5-methyl cytosine, 2'-deoxyuracil, or 7-deaza-2'-deoxyguanosine. As used herein, polynucleotides include their complements where appropriate. For example, where a pair of indicator polynucleotides are desribed for use in detecting a splice variant RNA using cDNAs immobilized in a microarray, and also described for use in a PCR reaction to amplify a product from the mRNA, one skilled in the art will appreciate that one of the indicator polynucleotides used in the PCR reaction would ordinarily be the complement of that used in the microarray, so as to produce a PCR product.

"Complementary" describes the relationship between two single-stranded nucleic acid sequences that hybridize through base-pairing. For example, 5'-AGT-3' pairs with its complement, 3'-TCA-5'.

"Splice variants" refer to gene products resulting from different transcriptional events and/or splicing events involving the same gene or set of genes, or derivatives, thereof. The gene products may be mRNA, rRNA, amplified antisense RNA (aRNA), cDNA, dsRNA, DNA, polypeptides, polypeptide fragments, etc. Polynucleotides may be modified as described, herein. Polypeptides include antibodies that recognize polypeptide forms of splice variants, *i.e.*, polypeptides translated from splice variant RNAs. Different forms of the same gene may be called "isoforms." As used herein, a "spliceoform," a "splice variant," and "isoform," and a "splice event" may be used interchangeably.

A "micro-exon" is a form of an exon that is smaller than a conventional exon. Microexons are typically less than 35 bases, less than 30 bases, less than 25 bases, less than 20 bases, less than 15 bases, less than 12 bases, less than 10 bases, or even less than 8 bases. A micro-exon may encode only a small number of amino acid residues (*e.g.*, 1, 2, 3, 4, 5, up to 8, up to 10, or up to 12 residues). While the terms exons and micro-exons are use to aid in the description of the indicator molecules and methods, one skilled in the art will

recognize that exon length varies in nature and that use of the terms "exon" or "micro-exon" is not intended as limiting.

As used herein, a "sample" is a source of at least one splice variant for detection using the present indicator molecules and methods. The sample may be a fluid sample, such as an aqueous sample, or a dry sample that can be suspended in an aqueous reagent, typically water, Tris-acetate, or similar for nucleic acids, and PBS or one of a number of buffers for polypeptides. The sample may be derived from a biopsy sample or from a body fluid, including but not limited to blood, plasma, mucous, sputum, semen, vaginal fluid, spinal fluid, throat or nasal washings, anal washings, amniotic fluid, urine, etc. The samples may be from cell culture or *in vitro* reactions. Sample may include RNase inhibitors or protease inhibitors, as appropriate, to protect splice variants from degredation.

As used herein, "a minimal set of indicator molecules," or a similar term, as used with respect to detecting several splice events involving of a gene or group of genes, is the minimum number of polynucleotides or polypeptides required to detect a predetermined or estimated number and type of splice variants in a sample. A set of indicator polynucleotides may contain 100 or more, 1,000 or more, 10,000 or more, 100,000 or more, or 1,000,000 or more indicator polynucleotides. A set of indicator polypeptides may contain 100 or more, 1,000 or more, 10,000 or more, 100,000 or more, or 1,000,000 or more indicator polypeptides. The set of indicator molecules may be for detecting 10 or more, 100 or more, 1,000 or more, 10,000 or more, 100,000 or more, or 1,000,000 or more splice events; 100 or more, 1,000 or more, 10,000 or more, 100,000 or more, or 1,000,000 or more splice variants; or 10 or more, 100 or more, 1,000 or more, 10,000 or more, or 20,000 or more genes of one or more organisms.

As used herein, "detecting" a splice variant refers to identifying a splice variant via binding of a tagged or labeled indicator molecule, such as a polynucleotide (*e.g.*, via hybridization) or an antibody (*e.g.*, via protein-protein interactions). The presence, and optionally, the amount of a particular splice variant can be determined by detecting the binding of such indicator molecules, *e.g.*, in an array.

An exon, exon-exon junction, module, exon-intron junction, intron-exon junction or module-module junction is a "constitutive" polynucleotide if all known and/or expected splice variants comprise that exon, exon-exon junction, module, exon-intron junction, intron-exon junction or module-module junction. Otherwise, it is "non-constitutive".

A first indicator polynucleotide is considered "mutually exclusive" with a second indicator polynucleotide if the first indicator polynucleotide indicates a first exon, module or junction of a first splice variant and the second indicates a second exon, module or junction of a second splice variant, wherein the first splice variant does not comprise the second exon, module or junction and the second splice variant does not comprise the first exon, module or junction.

A pair of mutually exclusive indicator polynucleotides comprises a first indicator polynucleotide and a second indicator polynucleotide. The pair is "overlapping" if the first indicator polynucleotide comprises a polynucleotide comprised by the second indicator polynucleotide. The pair is "non-overlapping" if the first does not comprise a polynucleotide comprised by the second. This will be clear to one of ordinary skill in the arts from the following examples.

As used herein, "essentially all splice variants" refers to all splice variants predicted based on known exons, as described herein, excluding spurious or unexpected splice events resulting from rare or non-canonical splice events, or unique biological events unlikely to be repeated with statistical regularity. Essentially all splice variants may be greater than 90% of all splice variants, greater than 95% of all splice variants, greater than 98% of all splice variants, greater than 99% of all splice variants, greater than 99.5% of all splice variants, or even greater than 99.8% of all splice variants.

As used herein, a "mean" indicates a mean, geometric mean, trimmed mean or trimmed geometric mean.

III. Indicator molecules for detecting splice variants

In one aspect the invention provides a set of indicator molecules for identifying known or predicted splice events. The hybridization of indicator molecules in the set to splice variants present in a sample may be mutually exclusive, as illustrated by the following example.

A. Indicator molecules for identifying known or predicted splice events

Mutually-exclusive indicator molecules for identifying known or predicted splice events are exemplified below:

1. In one example, where a gene comprises at least a first exon, an intron, and a second exon, and the second exon has a short form and a long form that differ at the 5' end, the gene may have a first splice variant that comprises the first exon and the long form of the second exon, and a second splice variant that comprises the first exon and the short form of the second exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the exon-exon junction between the first exon and the long form of the second exon is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the first exon and the short form of the second exon.

b. An indicator polynucleotide for the module at the 5' end of the second exon that is part of the long form but not the short form is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the first exon and the short form of the second exon.

c. An indicator polynucleotide for the intron-exon junction between the intron and the short form of the second exon is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the first exon and the short form of the second exon.

2. In another example, where a gene comprises at least a first exon, an intron, and a second exon, and the first exon has a short form and a long form that differ at the 3' end, the gene may have a first splice variant that comprises the long form of the first exon and the second exon., and a second splice variant that comprises the short form of the first exon and the second exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the exon-exon junction between the long form of the first exon and the second exon is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the short form of the first exon and the second exon.

b. An indicator polynucleotide for the module at the 3' end of the first exon that is part of the long form but not the short form is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the short form of the first exon and the second exon.

c. An indicator polynucleotide for the exon-intron junction between the short form of the first exon and the intron is mutually exclusive (overlapping) with an indicator polynucleotide for the exon-exon junction between the short form of the first exon and the second exon.

3. In another example, where a gene comprises at least a first exon, a second exon and a third exon, a first splice variant may comprise the first exon, the second exon and the third exon. A second splice variant may comprise the first exon and the third exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the junction between the first exon and the second exon is mutually exclusive (overlapping) with an indicator polynucleotide for the junction between the first exon and the third exon.

b. An indicator polynucleotide for the junction between the second exon and the third exon is mutually exclusive with an indicator polynucleotide for the junction between the first exon and the third exon.

c. An indicator polynucleotide for the junction between the first exon and the third exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the second exon.

4. In another example, where a gene comprises at least a first exon, a second exon, a third exon and a fourth exon, a first splice variant may comprise the first exon and the fourth exon. A second splice variant may comprise the first exon, the second exon, the third exon and the fourth exon. In this example, an indicator polynucleotide for the junction between the first exon and the fourth exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the junction between the second exon and the third exon.

5. In another example, where a gene comprises at least a first exon, a second exon, a third exon and a fourth exon, a first splice variant may comprise the first exon and the third exon. A second splice variant may comprise the

second exon and the third exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the junction between the first exon and the third exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the junction between the second exon and the fourth exon.

b. An indicator polynucleotide for the second exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the third exon if the gene comprises no splice variant that comprises both the second exon and the third exon.

6. In another example, where a gene comprises at least a first exon, a second exon and a third exon, a first splice variant may comprise the first exon and the third exon and starts with the first exon. A second splice variant may comprise the second exon and the third exon and starts with the second exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the junction between the first exon and the third exon is mutually exclusive (overlapping) with an indicator polynucleotide for the junction between the second exon and the third exon.

b. An indicator polynucleotide for the first exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the second exon if the gene comprises no splice variant that comprises both the first exon and the second exon.

7. In another example, where a gene comprises at least a first exon, a second exon and a third exon, a first splice variant may comprise the first exon and the third exon and ends with the third exon. A second splice variant may comprise the first exon and the second exon and ends with the second exon. In this example, several indicator molecules are useful for detecting and distinguishing these splice variants in a sample.

a. An indicator polynucleotide for the junction between the first exon and the third exon is mutually exclusive (overlapping) with an indicator polynucleotide for the junction between the first exon and the second exon.

b. An indicator polynucleotide for the second exon is mutually exclusive (non-overlapping) with an indicator polynucleotide for the third exon if the gene

comprises no splice variant that comprises both the first exon and the second exon.

B. Combinations of indicator molecules for detecting types of splicing events

The present compositions and methods include combinations of indicator polynucleotides for detecting and distinguishing particular types of splice event, such as exon skips, trimmed exons, intron retentions, etc.

To detect exon skips, one compares the exon-exon junction probe that skips an exon (*e.g.*, the junction between exons 1 and 3) with the exon-exon junction and/or exon probes that include the exon (*e.g.*, the junction between 1 and 2, exon2, and the junction between 2 and 3). When using splice signals, the data can be normalized against the gene expression level, which acts as a control. In the case of junction-only arrays, the same algorithm works, merely omitting the exon measurement.

To detect retained introns, one compares the exon-exon junction probe that skips an intron (*e.g.*, the junction between exons 1 and 2) with the exon-intron junction and/or intron probes that include the intron (*e.g.*, the junction between exon 1 and intron 1, intron 1, and the junction between intron 1 and exon 2). When using splice signals, the data can be normalized against the gene expression level. In the case of junction-only arrays, the same algorithm works, merely omitting the intron probe.

Short and long forms of exons (*e.g.*, resulting from 3' exon trims or extensions) are common in nature, and their analysis depends upon the exon-exon junction probe for the long form exon (*i.e.*, the junction between long exon 1 and exon 2), the exon-exon junction probe for the short form exon (*i.e.*, the junction between short exon 1 and exon 2), and perhaps the extended piece of the long exon. When using splice signals, the data can be normalized against the gene expression level, which acts as a control.

Alternative 5' exons may occur because of different promoters and may or may not include alternate splicing after the initial exon. If the long form contains all of the short form exons, the two forms may be regarded as either transcript variants or splice variants. We found more alternate first exons than alternate last exons, both in sequence data mining and in the microarray results.

An alternative 3' exon may also result from alternative splicing. Additionally or alternatively, a 3' untranslated region (UTR) may terminate with an earlier poly(A) tract or polyadenylation signal. In the former case, the exon can be identified using the same probe combination as an exon skip, omitting the final exon-exon junction (since the splice product ends rather than joining the downstream exon). If two 3' exons differ only in the length of the UTR, they may more properly be considered transcript variants than splice variants.

Complex splicing events—such as a multi-exon skip plus a truncated exon—may not be readily detectable or distinguishable using indicator molecules designed to detect the splice variants described above. These more complicated splice variants can be detected by determining the fold-change of an indicator molecule signal, *e.g.*, in an array device. In effect, this embodiment compares the signal detected using each individual indicator molecule to the overall level of gene expression. Because this method considers individual probes, it is statistically less robust.

Transcription level differences in gene products due to alternative promoters or UTR length also affect RNA stability and protein coding. Such events/splice variants can be detected using fold change of splice signals as described above. This type of data analysis can also lead to the discovery of novel splice isoforms. For example, an exon-exon junction that is differentially expressed without a corresponding change in either exon may suggest a previously unknown exon insertion, intron read-through or truncated or extended exon.

It will be apparent that the examples of indicator molecules may be combined in various ways to detect and quantitate a wide range of alternative splicing events and alternative splice forms of a gene. Accordingly, the present compositions and methods apply to a plurality of indicator polynucleotides for detecting exons, introns, modules, exon-exon junction, exon-intron junctions, intron-exon junctions, or module-module junctions of a gene, a plurality of genes, such as a set of genes for a gene/protein family, a pathway, or an organism's entire genome.

In some embodiments, the present compositions, devices, systems, and methods use one or more mathematical algorithms, in combination with data

obtained using the indicator molecules in a microarray, to calculate the expression levels of one or more splice variants in a sample. In some embodiments, a mathematical algorithm is applied to splice variant expression levels determined using a plurality of indicator polynucleotides that indicate non-constitutive polynucleotides in splice variants of a gene. In another embodiment, a mathematical algorithm is applied to expression levels determined using a plurality of indicator polynucleotides that indicate mutually exclusive (overlapping) polynucleotides in splice variants of a gene. In another embodiment, a mathematical algorithm is applied to expression levels determined using a plurality of indicator polynucleotides that indicate mutually exclusive (non-overlapping) polynucleotides in splice variants of a gene. In another embodiment, a mathematical algorithm is applied to expression levels determined using indicator polynucleotides that indicate constitutive, non-constitutive, or mutually exclusive (overlapping or non-overlapping) polynucleotides in splice variants.

In some embodiments, the compositions, devices, systems and methods determine an overall level of gene expression in a sample by summing (arithmetically combing) the expression levels determined using mutually exclusive indicator polynucleotides. In another embodiments, background signals from control array experiments (*e.g.*, not using a sample) are subtracted from the overall expression levels detected using mutually exclusive indicator polynucleotides. In a particular example, the present compositions, devices, systems and methods determine an overall gene expression level by averaging (including log averaging, taking the geometric mean of, or otherwise applying a mathematical algorithm to) one or more of the following expression levels:

(1) an expression level determined using an indicator polynucleotide that identifies the first exon;

(2) the sum of expression levels determined using mutually exclusive indicator polynucleotides;

(3) the sum of expression levels determined using the mutually exclusive indicator polynucleotides in III.A.1.b., above;

(4) the sum of expression levels determined using the mutually exclusive indicator polynucleotides in III.A.1.c.; and

(5) an expression level determined using an indicator polynucleotide that detects the second exon.

In some examples, a mathematical algorithm is applied to the expression levels detected according to (1) and (2), above, or (2) only, or (2) and (3), etc. In related embodiments, the algorithm sums background-subtracted expression levels in (2) and (3). In some embodiment, the algorithm applies a mathematical function of the background readings, such as a sum or log average or geometric mean, to the background subtracted expression levels determined using the indicator polynucleotides.

In another embodiment, a mathematical algorithm is applied to expression levels determined using a plurality of sets of mutually exclusive indicator polynucleotides for a single gene. In another embodiment, a mathematical algorithm is applies to expression levels determined using a plurality of sets of mutually exclusive indicator polynucleotides for identifying splice variants of a plurality of genes, such a family of evolutionarily related genes, or suspected related genes. The genes may be from a single species or multiple species. Some genes may be associated with one or more disease states while other genes are associated with a normal state.

It will be apparent that numerous combinations of indicator molecules and mathematical analysis, thereof, can be performed using the various examples of constitutive, non-constitutive, mutually exclusive overlapping, or mutually exclusive non-overlapping indicator molecules described. Exemplary equations are provided herein and in the Examples.

IV. Calculating Percent Composition of Splice Variants in a Sample

Where a gene produces multiple expected splice variants, it may be desirable to determine the relative quantity for each splice variant in a sample, independent of the overall level of gene expression. The relative amount of each splice variant in a sample is referred to as the "splice composition."

In one example, where a gene produces a first splice variant and a second splice variant, a first splice variant may have an expression level of 600 and a second splice variant may have an expression level of 1,200 (arbitrary units). In a second sample, the first splice variant may have an expression level of 6,000 and

the second splice variant may have an expression level of 12,000. The sum of the expression levels of the first and second splice variants in the first sample is 1800 and in the second sample is 18,000. The expression level data show that the gene is upregulated in the second sample. The absolute expression level of the first splice variant is ten times greater in the second sample than in the first sample. Likewise, the absolute expression level of the second splice variant is ten times greater in the second sample than in the first sample.

However, the relative proportions of the two splice variants are the same in the first sample and second samples. In the first sample, the first splice variant has an expression level equal to one half of the expression level of the second splice variant. In the second sample, the first splice variant also has an expression level equal to one half of the expression level of the second splice variant. Therefore, although the overall transcriptional level of the gene is different in the first and second samples, the relative composition of the splice variants within the samples is the same. One conclusion to be drawn from such data is that there seems to be no change in the splicing pattern or the levels of alternative or differential splicing between the samples. Only upregulation of total gene expression (at least of the particular splice variants for which indicator polynucleotides were provided) in the second sample is evident.

In another example, a gene has a first splice variant and a second splice variant, as before. In a first sample, the level of expression of the first splice variant is 600 and the level of expression of the second splice variant is 1,200. In a second sample, the level of expression of the first splice variant is 1200 and the level of expression of the second splice variant is 600. The sum of the expression levels of the first and second splice variants is 1,800 in the first sample and also 1,800 in the second sample. In this example, the overall level of gene expression level is the same in both samples.

However, the absolute level of expression of the first splice variant is two times greater in the second sample than in the first sample. Similarly, the absolute level of expression of the second splice variant is two times greater in the first sample than in the second sample. Therefore, the relative proportions of the two splice variants are different in the first and second samples. In the first sample, the first splice variant has an expression level equal to one-half the expression

level of the second splice variant. In the second sample, the first splice variant has an expression level equal to twice the expression level of the second splice variant. Although the overall transcriptional level of the gene appears to be the same in the first and second samples, the relative composition of the splice variants in the samples is different. Thus, in this example, there is evidence of a change in the splicing pattern between the samples.

In some embodiments, splice variant composition in a sample is described by the percent-composition of each splice variant in a sample. For example, data obtained using indicator molecules may indicate that the splice variant composition in a sample is 35% of a first splice variant and 65% of a second splice variant.

In some embodiments, splice variants that are not expressed above a predetermined threshold level are ignored or excluded from calculations for determining absolute or relative levels of splice variant expression. For example, if in a sample a particular splice variant has an expression level of 100, and the minimum threshold for detection is set at 150, the expression of the particular splice variant is ignored or set at zero.

In some embodiments, percent splice variant composition calculations are based on the use of indicator molecules that are mutually exclusive in their detection of splice variants in a sample. For example, a splice variant composition analysis may determines that a sample comprises 35% of a first splice variant based on a first indicator polynucleotide, and 65% of a second splice variant, based on a second indicator polynucleotide that is mutually exclusive with the first indicator polynucleotide.

In further embodiments, the above-described splice variant composition data is an input for further data analysis algorithms, such as a RMA (robust multichip average) calculation. In another embodiment, percent composition data is used as input to a clustering algorithm, a heat map, or a grouping algorithm. In further embodiments, percent composition data is used to generate in a visual representation of the expression of splice variants of a gene, as described herein.

V. Measuring and Calculating Different Patterns of Splice Variants

Exemplary algorithms for measuring and calculating different patterns of splice variants are described, below. These algorithms may be combined or

modified, optionally with other algorithms described herein or used in the art. All algorithms may be executed using appropriate computer software running in a suitable environment. Stored data or analyses may be combined with data from one or more microarray experiments. The results of data analysis may be represented as further described, herein.

A. Regulated splicing value, splice log ratio, and splice-fold ration

A difference in the composition of splice variants in two or more samples reflects a difference in regulation of gene splicing, producing in different patterns (or relative levels) of splice variants. Such differences in splicing patterns may be expressed in terms of fold-change in the expression levels of a particular splice variant, *e.g.*, relative to another splice variant or relative to total gene expression. A p-value, z-score, or expect(ation) value can be computed from fold-change data using classical or Bayesian statistical methods. The following examples exemplify indicator molecules, arrays and methods for determining splice variant expression patterns.

In a first example, a gene has a first splice variant and a second splice variant. In a first sample, the first splice variant has an expression level of 600 and the second splice variant has an expression level of 1,200. In a second sample, the first splice variant has an expression level of 6,000 and the second splice variant has an expression level of 12,000. The sum of the expression levels of the first and second splice variants is 1,800 in the first sample and 18,000 in the second sample. As before, the overall level of gene expression is upregulated in the second sample. The absolute expression level of the first splice variant is ten times greater in the second sample than in the first sample. Likewise, the absolute expression level of the second splice variant is ten times greater in the second sample than in the first sample.

The relative proportions/ratios of the two splice variants is the same in the first and second samples, *i.e.*, the first splice variant has an expression level equal to one-half the expression level of the second splice variant. Although the overall level of transcriptional of the gene is different in the first and second samples, the composition of splice variants in each samples is the same. Thus, there is no difference in the "splicing pattern," and no apparent change in splicing regulation,

the use of alternative or differential splicing pathways, or the use of some other natural or acquired cellular splicing mechanism, which would change the relative proportions/ratios of different polynucleotide splice variants produced by a gene or group of genes.

In another example, a gene has a first splice variant and a second splice variant. In a first sample, the first splice variant has an expression level of 600 and the second splice variant has an expression level of 1,200. In a second sample, the first splice variant has an expression level of 1,200 and the second splice variant has an expression level of 600. The sum of the expression levels of the first and second splice variants is 1,800 in both the first sample and the second sample, *i.e.*, the overall levels of gene expression are the same. However, the absolute expression levels of the first splice variant is two times greater in the second sample than in the first sample. The absolute expression level of the second splice variant is two times greater in the first sample than in the second sample. Therefore, the relative proportions of the two splice variants are different in the first sample compared to the second sample. Although the overall transcriptional levels of the gene is the same in the first and second samples, the relative composition of the splice variants are different. There is apparently a change in splicing regulation, the use of alternative or differential splicing pathways, or the use of some other natural or acquired cellular splicing mechanism, which would change the relative proportions/ratios of different polynucleotide splice variants produced by a gene or group of genes.

The relative amounts of splice variants in the samples (*i.e.*, the splicing pattern) can be calculated by applying a mathematical algorithm to the expression data. The devices and system described herein are best implemented on a microprocessor-based device, such as a computer. Any number of calculation may be performed using appropriate software. Microarrays and scanners for microarrays are known in the art and exemplified, herein.

In one embodiment, a measure of splicing regulated is calculated (or computed) using expression data obtained for two splice variants of a gene using the equation:

$$R = F(E1,1 , E1,2, E2,1, E2,2)$$

where R is a measure of alternative splicing, F is a mathematical algorithm (or "function"), E1,1 is an expression level for a first splice variant a first sample as determined using a first indicator polynucleotide, E1,2 is an expression level for a first splice variant a second sample as determined using a second indicator polynucleotide, E2,1 is an expression level for a second splice variant a first sample as determined using a first indicator polynucleotide, and E2,2 is an expression level for a second splice variant a second sample as determined using a second indicator polynucleotide.

A mathematical algorithm using log ratios can be used to produce a "splice log ratio." For example, where E1,1 = 600; E1,2 = 1,200; E2,1 = 6,000; and E2,2 = 12,000:

$$R = \log(E1,1 / E1,2) - \log(E2,1 / E2,2)$$
$$R = \log(0.5) - \log(0.5) = \log(0.5/0.5) = \log(1) = 0.$$

Here, R equals zero, indicating that no change in regulated splicing is taking place between the first sample and the second sample.

However, in another where E1,1 = 600; E1,2 = 1,200; E2,1 = 1,200; and E2,2 = 600:

$$R = \log(600/1200) - \log(1200/600)$$
$$R = \log(0.5) - \log(2) = \log(0.25) = -0.602$$

Here, R is not equal to zero, indicating a difference in regulated splicing between the first sample and the second sample.

In other embodiments, a mathematical algorithm calculates a fold-change (instead of a log ratio). The calculation can be called a "splice-fold" or "splice-fold change."

Expression levels of splice variants detected using indicator molecules may be normalized for overall expression levels prior to use in calculating splicing patterns. Normalization includes using background subtraction, Loess normalization, mutation controls, swap-controls, and other routine background subtraction measures (see, below). In some embodiments, splice variants that are

not expressed above a preset threshold level are ignored (*i.e.*, assigned a value of zero) when calculating the relative levels of splice variants.

A plurality of data relating to regulated splicing may be converted into z-scores, p-values, or another measure of statistical significance, *e.g.*, by assuming a normal distribution, t-distribution, or another probability distribution for the data obtained using the indicator molecules. Statistics for use with microarray data are described herein and in the documents identified herein.

An example will be used to illustrate splice variant. In this example, a gene produces three polynucleotide splice modules, a first module, a second module and a third module. The three modules may be exons. The first module may be an exon, the second module an intron, and the third module an exon. The first module may be an exon, the second module an extension of an exon, and the third module an exon. The gene may have a first splice isoform containing the first module, the second module and the third module; and a second splice isoform containing the first module and the third module. Suitable indicator polynucleotides for detecting these splice variants hybridize to sequence corresponding to the first module (M1); the junction between the first module and the second module (J1-2); the second module (M2); the junction between the second module and the third module (J2-3); the junction between the first module and the third module (J1-3); and the third module (M3).

A nucleotide array may include a number of indicator polynucleotides sufficient for detecting the three junctions and the three modules. For example, the nucleotide array may include indicator polynucleotides for the three junctions; indicator polynucletide for the junction between the first module and the second module and the indicator polynucleotide for the junction between the first module and the third module; indicator polynucleotide for the junction between the first module and the third module and the indicator polynucleotide for the junction between the second module and the third module; or a combination or subset of the six types of indicator polynucleotides. The nucleotide array may comprise multiple indicator polynucleotides for a junction between a first module and a second module or multiple indicator polynucleotides for a module, and so forth.

In the same example, the following expression levels were determined using the indicated indicator polynucleotides:

M1 = 300

M2 = 200

M3 = 300

J1-2 = 200

J1-3 = 100

J2-3 = 200

The gene expression level, G, is calculated using the equations:

$$G = \sum I_i = (\prod B_j)^{1/n} \quad (1)$$

$$B_{exon} = p + p^c \quad (2)$$

$$B_{junc} = \sqrt{((p_5 + p_5^c) * (p_3 + p_3^c))} \quad (3)$$

where G is the gene expression level; each I is a splice isoform; each B_j is the sum of expression levels of all splice isoforms at a base range targeted by one or more probes; n is the number of probes for the gene; B_{exon} covers the case of an exon or intron probe, with p equal to a probe signal and p^c equal to the sum of signals of probes exclusive with (or complementary to) p; and B_{junc} addresses the case of an exon-exon or exon-intron junction probe, with p_5 being the 5′ portion of the junction and p_3 being the 3′ portion, each portion having its own complement.

p5 is the 5' portion of an exon-exon or exon-intron junction probe p. p3 is the 3' portion of an exon-exon or exon-intron junction probe p. p^c is the "complement", meaning the set of probes that are complementary to p. A complementary probe p^c is mutually exclusive with a probe p, meaning that the same splice isoform cannot contain both the polynucleotide indicated by p and the polynucleotide indicated by p^c. In this manner, $p5^c$ is the complement of the 5' portion of p and $p3^c$ is the complement of the 3' portion of p.

For example, for a first splice isoform containing exon-exon junction J1-3 and a second splice isoform contains exon-exon junction J2-3, p5 is the signal for the portion of J1-3 for exon 1; $p5^c$ is the complementary signal from the second splice isoform, J2-3. For the second splice isoform, p5 is the signal for the portion of J2-3 for exon 2; $p5^c$ is the complementary signal from J1-2. The two splice isoforms have mutually exclusive splicing patterns, so each is a complement of the other. In this example, the two probes share the same 3' portion, so in both cases,

p3 is the signal for the portion that detects exon 3. For the first splice isoform, J2-3 is the complement. For the second splice isoform, J1-3 is the complement.

Returning to the example using the above-identified expression levels for the indicator polynucleotides,

$$G = (B_{M1} * B_{M2} * B_{M3} * B_{J1-2} * B_{J1-3} * B_{J2-3})^{1/6}$$

$$= ((M1 + 0)$$
$$* (M2 + J1\text{-}3)$$
$$* (M3 + 0)$$
$$* \sqrt{((J1\text{-}2 + J1\text{-}3) * (J1\text{-}2 + J1\text{-}3))}$$
$$* \sqrt{((J1\text{-}3 + J1\text{-}2)*(J1\text{-}3 + J2\text{-}3))}$$
$$* \sqrt{((J2\text{-}3 + J1\text{-}3) * (J2\text{-}3 + J1\text{-}3))})^{1/6}$$
$$= (300$$
$$* (200 + 100)$$
$$* 300$$
$$* \sqrt{((200 + 100) * (200 + 100))}$$
$$* \sqrt{((100 + 200) * (100 + 200))}$$
$$* \sqrt{((200 + 100) * (200 + 100))})^{1/6}$$
$$= (300^6)^{1/6} = 300.$$

Background levels may be taken into account and even subtracted from expression level data when solving the equation. Alternatively, background can be subtracted from the 'complementary' indicator polynucleotides in the equation. For example, where the background level for each splice variant detected using an indicator polynucleotide is 100:

$$G = (B_{M1} * B_{M2} * B_{M3} * B_{J1-2} * B_{J1-3} * B_{J2-3})^{1/6}$$

$$= ((M1 + 0)$$
$$* (M2 + J1\text{-}3 - k)$$
$$* (M3 + 0)$$
$$* \sqrt{((J1\text{-}2 + J1\text{-}3 - k) * (J1\text{-}2 + J1\text{-}3 - k))}$$
$$* \sqrt{((J1\text{-}3 + J1\text{-}2- k)*(J1\text{-}3 + J2\text{-}3 - k))}$$
$$* \sqrt{((J2\text{-}3 + J1\text{-}3 - k) * (J2\text{-}3 + J1\text{-}3 - k))})^{1/6}$$

$= (300$
$* (200 + 100 - 100)$
$* 300$
$* \sqrt{((200 + 100 - 100) * (200 + 100 - 100))}$
$* \sqrt{((100 + 200 - 100) * (100 + 200 - 100))}$
$* \sqrt{((200 + 100 - 100) * (200 + 100 - 100))})^{1/6}$
$= (300 * 200 * 300 * \sqrt{(200*200)} * \sqrt{(200*200)} * \sqrt{(200*200)})^{1/6}$
$= (300^2 * 200^4)^{1/6}$

Other methods of background subtraction may be used, as known in the art.

In further embodiments the level of gene expression is calculated from the levels of hybridization in the sample obtained using the junction indicator polynucleotides:

$G = (B_{J1-2} * B_{J1-3} * B_{J2-3})^{1/3}$
$= (\sqrt{((J1-2 + J1-3) * (J1-2 + J1-3))}$
$* \sqrt{((J1-3 + J1-2) * (J1-3 + J2-3))}$
$* \sqrt{((J2-3 + J1-3) * (J2-3 + J1-3))})^{1/3}$
$= \sqrt{((200 + 100) * (200 + 100))}$
$* \sqrt{((100 + 200) * (100 + 200))}$
$* \sqrt{((200 + 100) * (200 + 100))})^{1/3}$
$= (300^3)^{1/3} = 300$

Background levels may be subtracted as before.

B. Splicing Index

In another embodiment, data are analyzed using a "Splicing Index, which can generally be defined as:

$$\text{Splicing Index} = R_{short} - R_{long} \quad (5)$$

$$R_{short} = \log(J_s/J_t) - [\sum \log(D_{is}/D_{it})] / m \quad (6)$$

$$R_{long} = [\sum \log (K_{is} / K_{it})] / n - [\sum \log(D_{is}/D_{it})] / m \quad (7)$$

where J is a junction signal for a short splice form (an exon skip, later first exon, earlier last exon, intron splice or truncated exon from alternative

donor or acceptor), D_i are signals for m exons that flank the alternative splice site (m=1 for an alternative first or last exon, m=2 for all other splice events); K_i are signals for n long splice form junctions or exons; and s and t are samples or computational pools of samples using an equation such as mean, geometric mean, median, or geometric median, perhaps omitting outliers.

As an example of the generalized Splicing Index, consider a single-exon skip event with long splice form probes for J1-2, E2 and J2-3 which include the exon, and a short splice form probe for J1-3 which skips the exon. Probes for the flanking E1 and E3 may be included as well. Suppose the signals in sample s are as follows:

J1-2 = E2 = J2-3 = 200

J1-3 = 400

E1 = E3 = 600

and in sample t they are

J1-2 = E2 = J2-3 = 400

J1-3 = 200.

E1 = E3 = 600.

Hence,

R_{short} = log 400/200 – (log 600/600 + log 600/600) / 2.

= log 2

R_{long} = (log 200/400 + log 200/400 + log 200/400) / 3 – (log 600/600 + log 600/600) / 2.

= log (1/2)

Splicing Index = log 2 – log (1/2)

In one embodiment, the generalized Splicing Index omits the signal for the skipped exon. It should be noted that the term for the flanking exons cancels out. For example, consider the above example of an exon skip, but without J1-2. This example addresses the case of an alternative first exon. The gene either begins with J1-3 or J2-3. The generalized Splicing Index

can be applied as before, except that the term for J1-2 will be omitted. Consider the above example of an exon skip, but without J2-3. This example without J2-3 addresses the case of an alternative last exon. Either the gene ends with J1-3 or with J1-2. The generalized Splicing Index can be applied as before, except that the term for J2-3 will be omitted.

In an embodiment, the generalized Splicing Index combines long form probes using a function other than the average log ratio. It applies a function of the form F(log K_{is} / K_{it}), where K, i, s and t are defined as above. F might be the mean, minimum, maximum, median, or other function of the values of K_i. In an embodiment, F takes the minimum unless two (log (K_{is} / K_{it})) pairs have opposite signs, in which case F gives the value of zero (0).

The splicing analysis method computes a score for splicing changes using a generalized ASPIRE algorithm:

$$|ASPIRE| = k * \min (|R_{short}|, |R_{long}|); \; R_{short} * R_{long} < 0 \quad (8)$$

$$R_{short} = \log [J_s/J_t] \quad (9)$$

$$R_{long} = \min (\log [K_{is} / K_{it}]) \quad (10)$$

where the variables have the same meanings as for the generalized Splicing Index, and k is a constant, such as 2. If any (log [K_{is} / K_{it}]) pair has opposite signs, the data analysis method sets R_{long} = 0.

For example, take the example data from the generalized Splicing Index for an exon skip. Let k = 2. Then:

$$|ASPIRE| = 2 * \min (|\log 2|, |\log ½|) = 2 * \log 2.$$

In this case, the generalized ASPIRE value and the generalized Splicing Index value are the same. Now suppose R_{short} = log 4.

$$|ASPIRE| = 2 * \min (|\log 4|, |\log ½|) = 2 * |\log ½|.$$

In one embodiment, the splicing analysis method sets the sign of the ASPIRE value based on R_{short}. In that case, for these two examples, the ASPIRE value is positive. In another embodiment, the splicing analysis method sets the sign of the ASPIRE value based on R_{long}. In that case, for these two examples, the ASPIRE value is negative.

As with the generalized Splicing Index, the generalized ASPIRE equation can address all types of splicing events.

In an embodiment, the generalized ASPIRE algorithm combines long form probes using a function other than the minimum log ratio. It applies a function of the form F(log K_{is} / K_{it}), where K, i, s and t are defined as above. F might be the mean, minimum, maximum, median, or other function of the values of K_i.

Another method computes a score for splicing changes using a Splice Ratio equation that combines elements of the generalized Splicing Index and generalized ASPIRE. It identifies splicing changes that occur in opposite directions, but relative to the global or local gene expression level G rather than in absolute terms, as is the case with ASPIRE:

$$|\text{Splice Ratio}| = 2 * \min(|R'_{short}|, |R'_{long}|); \; R'_{short} * R'_{long} < 0 \qquad (11)$$

where R' is the ratio computed after normalizing all probes by the gene expression level using equation 4. In an embodiment, the splicing analysis method uses a global gene expression level to normalize probes. In another embodiment, the data analysis method uses a local gene expression level to normalize probes.

In one embodiment, a mathematical algorithm determines a splicing score for a splice event with more than two alternatives. In some embodiment the algorithm is the generalized Splicing Index, generalized ASPIRE algorithm, or Splice Ratio

In one example, a gene has three spliceoforms based on four possible exons. The spliceoforms are S1-3-4, S1-4 and S1-2-3-4. An exon probes targets each of the four exons (E1, E2, E3 and E4), and an exon-exon junction probes target each of the five distinct exon-exon junctions (J1-3, J3-4, J1-4, J1-2 and J2-3). In one embodiment, a mathematical algorithm is applied to determine a splicing score for each spliceoform.

For example, where J1-2, E2 and J2-3 all have signals of 200 in a first sample and 400 in a second sample, J1-3 has a signal of 100 in the first and second samples, and J1-4 has a signal of 300 in the first sample and 600 in the second samples, the, a mathematical algorithm to is applied to the hybridization signals. For example, the algorithm could assign a score for S1-2-3-4 = log (200 / 400) – log ((100 + 300) / (100 + 600)). In one embodiment, the signals of each long form probe and treats the sum of signals as if it were a single alternative splice form. In other words, R_{long} = the log of the sum of signals for each long

splice form in the first sample minus the log of the sum of signals for each long splice form in the second sample.

$$R_{long} = F (\log [L_{js} / L_{jt}]) \quad (12)$$

where F is a function such as the minimum or mean and each L_j is a long splice form in a sample s or t. L is further defined as

$$L = G (K_{is}) \quad (13)$$

where G is a function such as the sum or geometric mean and each K_i is a probe detecting the long splice form i. In one embodiment, this value for R_{long} is substituted in the equations for the generalized Splicing Index, generalized ASPIRE or Splice Ratio equations, as herein.

A splicing score may be converted to a fraction or a percentage. For example, suppose the score is -2 in $\log_2$, *i.e.*, the fold change is -4. Therefore the signal for the short splice form is ¼ the signal of the long splice form in a first sample relative to a second sample. Hence, the fraction is 0.2 for the short splice form and 0.8 for the long splice form, and the percentages are 20% and 80% respectively. In one embodiment, a splicing score for a single sample is converted to a fraction or a percentage. For example, replace the variables for the second sample, sample t, with one ('1') in an equation above for generalized Splicing Index, generalized ASPIRE or Splice Ratio. A splicing score can be straightforwardly calculated for the first sample, sample s. For each spliceoform, a percentage or fraction can be determined. For example, consider the case of an alternative acceptor site. A short splice form might have a signal of 400 and a long splice form might have three probes (a module probe, an exon-exon junction probe and a module-module junction probe) having a geometric mean signal of 600. The Splicing Index could be calculated using equation 1 as log 400 – log 600: The resulting value can be converted to a fraction using the equations.

$$F + F_c = 1 \quad (14)$$

$$\log F / F_c = \log (400 / 600) \quad (15)$$

where F is a fraction of gene expression deriving from a first spliceoform and F_c is a fraction of gene expression deriving from a second spliceoform or two or more alternative spliceoforms mutually exclusive with the first spliceoform. Multiplying F by 100 gives a percentage.

A splicing score for two spliceoforms may be converted to a fraction or a percentage. The above examples address this case. A splicing score for three or more spliceoforms in one sample may be converted to a fraction or a percentage. Equation 12 provides a way to compute a splicing score using several of the algorithms above. Equations 14 and 15 provide a way to compute a fraction or a percentage for each spliceoform. A splicing score for two spliceoforms in two samples may be converted to a fraction or a percentage. A splicing score for two spliceoforms in more than two samples, or more than two spliceoforms in two samples, or more than two spliceoforms in more than two samples, may be converted to a fraction or a percentage.

A mathematical algorithm may be applied to determine a splicing score for an indicated polynucleotide such as a single exon, intron, exon-exon junction, exon-intron junction, module or module-module junction (as opposed to a splice event or a spliceoform). In some embodiments, a generalized Splicing Index, generalized ASPIRE or Splice Ratio algorithm is applied to determine a splicing score for an indicated polynucleotide.

Where multiple indicator molecules detect a splice form (and one detects an exon-exon junction, an exon, an exon-intron junction, a module, a module-module junction and an intron), the splicing analysis method applies a mathematical equation to determine a signal for the multiple probes. For example, the probe signal intensities might be averaged, log averaged, a median value might be used, outliers might be discarded, the values might be first normalized using a z-score, etc. The resulting value derived from multiple probes can be substituted for J (a short form junction) or K (long form junctions, exons, introns or modules) in the equations above). For a long splice form, each indicated polynucleotide can be treated individually. This gives a splicing score for each probe. Conceptually, the change involves treating each indicated polynucleotide as a "short form" in the equations above, and the mutually exclusive splice events or spliceoforms as the "long form".

An example will make this clear. The example refers to the three spliceoforms described above: S1-3-4, S1-2-3-4, and S1-4. A splicing score can be calculated for J1-2. Suppose the signal measured for J1-2 is 300. This indicated polynucleotide (sometimes, for simplicity's sake, referred to as a probe,

even though in fact this 'probe' may comprise multiple probes) can be treated as a 'short form' in the equations above such as Equation 1. The corresponding 'long forms' relative to this indicated polynucleotide are J1-3 and J1-4. Similarly, a splicing score can be calculated for E2. Again the 'long forms' relative to the indicated polynucleotide are J1-3 and J1-4. It is interesting to note that the splicing scores for J1-2, E2 and J1-3 may be different, although in theory, if the gene model comprises only three spliceoforms, the scores should be the same. Differences may arise because of an unexpected additional spliceoform (a surprising result) or because of experimental error. Application of an equation above can determine a splicing score, a fraction or a percent composition for each indicated polynucleotide.

A score for splicing changes is used to filter data for one or more splice events, and splice events pass the filter if the score satisfies some filtering criterion, F(score). In one embodiment, the splicing analysis method accepts a score if the absolute value exceeds a constant value. For example, suppose there are ten exon skip events detected by probes on a microarray, and a Splicing Index score is calculated for each. The splicing analysis method may filter out any events that have a score with an absolute value less than 1.0 in log base two (ie., a splice fold change of less than two). Suppose that two of the exon skip events pass the filtering criterion.

Other algorithms such as ASAP or genASAP may be used. In another embodiment, the splicing analysis method calculates a score using a machine learning algorithm, a genetic algorithm, a neural network, a simulated annealing algorithm, or another algorithm that may occur to one of skill in the arts.

In some embodiments, data is filtered by the type of splice event. For example, data from a high-density microarray may be processed to output all data for probes that detect alternative donor sites. In one embodiment, data is filtered to identify data for a single splice event type. In another embodiment, data is filtered to identify multiple splice event types.

In one embodiment, data is filtered based on whether a signal is up-regulated or down-regulated compared to another sample. Multiple criteria may be used.

C. Normalization

Relative splice variant compositions may be normalized for overall gene expression levels, such that both relative and absolute levels of splice variants are determined. For example, where the level of a first splice variant in a sample is 3,500 based on data obtained using a first indicator polynucleotide, and the level of a second splice variant in a sample is 6,500 based on data obtained using a second indicator polynucleotide that is mutually exclusive (overlapping or non-overlapping) with the first indicator polynucleotide, the splice variant composition data are normalized for the overall expression levels, such that the levels of the first and second splice variants account for 100% of total gene expression.

In other embodiments, a mathematical equation is applied to normalize expression levels for a splice variant of a gene. The equation may be of the form:

$$S = p\,M/G \qquad (4)$$

wherein S is the gene-normalized signal of a probe in a sample, p is the probe signal, G is the gene expression level, and M is a scalar, such as the geometric mean of G and the geometric mean of gene expression levels in the sample and at least one other sample.

M's value may be a function of the gene expression level relative to the probe in one or more samples (the "local gene expression level"); a function of the local gene expression level in the sample; a function of the local gene expression level in two or more samples; or the average of the local gene expression level in two or more samples. (*e.g.*, if the local gene expression levels in two samples are 500 and 1000, M's value is the average, which is 750.) M may be equal to the geometric mean of the local gene expression levels in two or more samples; the median of the local gene expression levels in two or more samples; the geometric median of the local gene expression levels in two or more samples; or an arbitrary function of the local gene expression levels in two or more samples. For example, M might discard outliers and take a function, such as the geometric mean, of the remaining values.

G may be calculated using the equation above or by combining proximal probes relative to the probe for which p is the signal, such as the flanking exon signals for an exon-exon junction probe for an exon skip, or the one flanking exon signal for an alternative first or last exon, or the flanking exon signals for a probe

for a retained intron, or the flanking exon signals for a probe for an alternative acceptor or donor site.

Examples for calculation of G as a "global gene expression level" appear above. For "local gene expression levels," examples follow. Where a short splice form probe (an exon-exon junction probe for a skipped exon, a spliced intron, or an alternative donor or acceptor site that truncates an exon) has two flanking exons with probe signals of 200 and 250, the local transcription level G could be calculated as F (exon1, exon2). F may takes the average of the two exon signals, so in the example, the value is average (200, 250) = 225. In other embodiments, F takes the median; the geometric mean = square root (200 * 250); or the geometric median. In another embodiment, it performs an arbitrary function on the flanking exon probes.

Where the probe is an exon-exon junction probe or an exon probe for an alternative first exon, F (common_exon) computes a signal value for the first nearest common exon between the splice forms. For example, where a gene has a first spliceoform that contains J1-3 and a second spliceoform that contains J2-3, probes for these two junctions can be normalized using the transcriptional level G computed as F(common_exon) = F(E3). In one embodiment, F is the identity function. In another embodiment, F is an arbitrary function.

In an illustrative example, a nucleotide array comprises a first indicator polynucleotide (p_1) and a second indicator polynucleotide (p_2) for the same gene. In an array experiment, the first indicator polynucleotide indicates an expression level of 400 in a first sample and 600 in a second sample. The second indicator polynucleotide indicates an expression level of 800 in the first sample and 1,200 in the second sample. The gene has an overall expression level of 500 in the first sample and 1,000 in the second sample. Applying the above equation:

$$S_1 = 400 * \sqrt{(500 * 1000)} / 500$$

$$S_2 = 600 * \sqrt{(500 * 1000)} / 500$$

And in the second sample:

$$S_1 = 800 * \sqrt{(500 * 1000)} / 1000$$

$$S_2 = 1200 * \sqrt{(500 * 1000)} / 1000$$

Note that S_1 is equal in both samples, and S_2 is equal in both samples. The normalization adjusts the values based on gene expression level. The remaining

differences between the two samples may result from different splicing patterns. In this case, there are no remaining differences between the two samples, hence there may not be a difference in splicing.

In a second example an array includes a first indicator polynucleotide (p3) and a second indicator polynucleotide (p4) for the same gene. The first indicator polynucleotide indicates an expression level of 400 in a first sample and 600 in a second sample. The second indicator polynucleotide has indicates an expression level of 1,200 in the first sample and 800 in the second sample. The gene has an overall expression level of 500 in the first sample and 1,000 in the second sample.

Applying the above equation, for the data obtained using the first sample:

$S_3 = 400 * \sqrt{(500 * 1000)} / 500 =$

$S_4 = 600 * \sqrt{(500 * 1000)} / 500 =$

And for the second sample:

$S_3 = 1200 * \sqrt{(500 * 1000)} / 1000$

$S_4 = 800 * \sqrt{(500 * 1000)} / 1000$

Note that now S_3 in the second sample is larger than S_3 in the first sample, while S_4 in the second sample is smaller than S_4 in the first sample. The difference may suggest a difference in splicing between the two samples.

In some embodiments, levels of expression obtained using indicator molecules are normalized using a gene expression level value derived from more than two samples. Thus, instead of normalizing the two samples to each other (*e.g.*, with the value of $\sqrt{(500 * 1000)}$ in the present example), normalization is to another value of m, such as the geometric mean of gene expression levels in each of a plurality of samples, or a gene expression level from a second sample.

Normalization can be applied to data obtained using any number of indicator molecules.

D. Statistical analysis

Data from one or more microarray experiments and/or from one or more data stores may be analyzed to determine statistical significance. In one embodiment, the calculation is standard deviation. For example, where Splicing Indexes are calculated for each exon skip event and there are 1,000 exon skip events, there will be 1000 Splicing Index scores. The standard deviation of these

1000 scores may be calculated, or a standard error of data, and/or a confidence level. For example, the cutoff of the 10 scores (out of 1000) with the highest absolute values could be used as the 99% confidence level. For the present purposes, the standard deviation, standard error and confidence interval will all be referred to collectively as "confidence intervals." Confidence intervals can be calculated for a single splice event type, for example, the 1000 exon skip events mentioned above, and/or for splice events that involve the same number of probe measurements. For example, where a microarray contains 4 probes per exon skip event (3 exon-exon junction probes plus one exon probe) and the microarray also contains 4 probes per intron retention event (1 exon-exon junction probe, 2 exon-intron junction probes, and 1 intron probe), statistical calculations based separately on the two splice events might reasonably be predicted to yield similar results, *i.e.*, the same confidence interval, standard deviation or standard error could be used to analyze both splice event types. Similarly, alternative donor sites and alternative acceptor sites might be detected by the same number of probes, perhaps 2 exon-exon junction probes for each (one probe for the short form, one for the long form), and donor and acceptor sites, and so forth.

A statistical confidence interval may be used as a filter. For example, alternative first and last exons have a 99% confidence interval of +/- 0.10 for the Splice Ratio. When comparing data for two samples, a potential alternative first exon may have a Splice Ratio score of 0.15. This value lies outside of the confidence interval, which is centered at zero. Hence, the alternative first exon event passes the filtering criterion and is statistically significant with $p < 0.01$. Additional criteria might be applied, of course, and the event would have to pass these additional criteria as well. A confidence interval may be used as a filtering criterion for a single splice event type or two or more splice event types, *e.g.*, alternative first and alternative last exons. In one embodiment, a confidence interval is used as a filtering criterion for all splice data. A confidence interval may be calculated for gene expression, *e.g.*, as described above. Gene expression data derived from a splice variant microarray is first filtered using a confidence interval(s), as described.

Other methods of statistical analysis are known in the art.

VI. Data Representation, Storage, and Synchronization

In some aspects, the present compositions, devices, software, systems, and methods present splice variant data in one or more data configurations or representations. Such representation are particularly useful for presenting data obtained using more several indicator molecules, as in an array. Many of the representation formats can be combined, providing the user with an integrated system for splice variant analysis.

In one embodiment, the data representation is in the form of a table including a column of splice variant identifications (IDs) or symbols, and a column containing the levels of each splice variant obtained using a mathematical algorithm applied to data obtained using indicator molecules. The table may include a header row with names for the columns and/or additional columns; a plurality of columns containing splice variant levels derived using various mathematical algorithms; and/or columns containing splice variant expression levels obtained from multiple samples. The rows of the table may contain a splice variant ID or symbol, the level of expression in a first sample, and the level of expression in a second sample.

In another embodiment, the data representation includes a second table containing splice variant expression data obtained using one or more particular exon-exon junction indicator polynucleotides, exon-intron junction indicator polynucleotides, intron-exon junction indicator polynucleotides, exon indicator polynucleotides, module indicator polynucleotides, or module-module junction indicator polynucleotides. The second table may comprise at least a first column of identifiers for the indicator polynucleotide (or the splice variants they detect), and a second column containing splice variant expression level data obtained using the indicator polynucleotides. The second table may include a header row with names for the columns. and/or additional columns. It may include columns for mean, median, mode, min or max expression levels of pixels from a scanner. It may include columns containing the levels of splice variants in each of multiple samples. A row of the table may contain an indicator polynucleotide or splice variant ID, an expression level for splice variant in a first sample, an expression level for a splice variant in a second sample, etc.

In another embodiment, the data representation is in the form of a single table containing splice variant IDs, the levels of splice variants in one or more samples, etc. An exemplary table is a tab-delimited file, a spreadsheet file for Microsoft Excel or OpenOffice, a database table, or a csv file.

In another embodiment, the data representation is in the form of a XML document containing splice variant IDs or symbols and splice variant levels calculated using a mathematical algorithm applied to data obtained using indicator molecules as described. The XML document may comprise additional information (or elements). For example, it may include splice variant expression levels calculated using various mathematical algorithms. The XML document may include splice variant expression data using multiple samples, such that an element of the comprises a splice variant ID or gene symbol, an expression level in a first sample, a expression level in a second sample, etc.

In some embodiments, the data representation includes a second XML document containing expression level data obtained using exon-exon junction indicator polynucleotides, exon-intron junction indicator polynucleotides, intron-exon junction indicator polynucleotides, exon indicator polynucleotides, module indicator polynucleotides, or module-module junction indicator polynucleotides. The second XML document may also contain indicator polynucleotide or splice variant identifiers and the expression levels of splice variants in a sample. The second XML document may include additional information. For example, it may include mean, median, mode, min or max expression level data based human and/or computer-readable microarray data.

The second XML document may include expression levels obtained from multiple samples, such that an element of the XML file includes an indicator polynucleotide or splice variant ID, an expression level in a first sample, an expression level in a second sample, etc.

In one embodiment, the data configuration is in the form of a single XML document containing all the information described. The XML document may include expression level data for a plurality of samples, and may further include annotation information.

In another embodiment, the data representation is in the form of an electronic file containing splice variant or indicator polynucleotide IDs or symbols

and splice variant expression level data obtained from a mathematical algorithm, as above. The electronic file may include additional information, for example, splice variant expression levels obtained using various mathematical algorithms, splice variant expression levels for multiple samples, etc., such that an element of the file includes a splice variant or indicator molecule ID or symbol, a splice variant expression level in a first sample, a splice variant expression level in a second sample, etc.

In some embodiments, the data representation includes a second electronic file containing splice variant expression data obtained using exon-exon junction indicator polynucleotides, exon-intron junction indicator polynucleotides, intron-exon junction indicator polynucleotides, exon indicator polynucleotides, module indicator polynucleotides, or module-module junction indicator polynucleotides. The second electronic file may include indicator polynucleotide or splice variant identifiers and expression level data obtained using these indicator polynucleotides. The second electronic file may include additional information, for example, mean, median, mode, min or max expression levels based on microarray data. It may include splice variant expression level data from multiple samples, such that an element includes an indicator polynucleotide or splice variant ID, an expression level of a splice variant in a first sample, an expression level of a splice variant in a second sample, etc.

In one embodiment, the data representation is in the form of a single electronic file containing the above-described data. Exemplary electronic files are binary files and serialized data objects.

Data representations may include expression data obtained for the splice variants of at least one gene, 20 or more genes, 100 or more genes, 1,000 or more genes, or even 100,000 or more genes. Data representation may include splice variant expression data obtained using one or more indicator polynucleotides, 20 or more indicator polynucleotides, 100 or more indicator polynucleotides, 1,000 or more indicator polynucleotides, or even 100,000 or more indicator polynucleotides. Data representation may include data for at least one sample or for multiple samples, such as 2 or more, 5 or more, 10 or more, 20 or more, or even 100 or more samples.

The splice variant data and/or representations of the data may be transmitted from one location or entity to another, *e.g.*, by email, email attachment, FTP, HTTP, a socket connection, UDP, 802.11a/b/g, wireless transmission, or by another means or protocol for electronic data transfer. In other embodiments, the splice variant data are stored on a portable storage device such as print material, flash memory, or a CD, DVD or removable hard drive, and then transmitted from one location or entity to another. Exemplary transmission methods are domestic (including U.S.) mail, international mail, FedEx, UPS, DHL, courier, private delivery, or interpersonal exchange.

More advanced data representations for mapping and calculating splice patterns may include the results of calculations, such as log splice ratios, splice fold changes, z-scores, or p-values. The data representation may be in the form of a table, such as a spreadsheet, tab-delimited file, csv file, database table, etc. In another embodiment, the data representation is in the form of an XML document. In another embodiment, the data representation is in the form of an electronic file, such as a serialized object or binary file.

Data representations for mapping and calculating splice patterns may include data obtained using at least one, 20 or more, 100 or more, 1,000 or more, or even 100,000 or more indicator polynucleotide. A regulated splicing data representation may include data obtained using at least one sample pair. Alternatively, the data may be from multiple sample pairs, such as 2 or more pairs, 5 or more pairs, 10 or more pairs, 20 or more pairs, or even 100 or more pairs. Such data may be stored and transmitted as above.

Exemplary data representation formats are shown in the Figures and Tables. For example, Table 3 (below, referring to Example 5) identifies the number of instances of each data type (gene or splice event type), such as alternatively spliced short spliceoforms, or alternative donor sites (short form or long form). It shows the number of instances of each splice event that pass the filtering criterion defined in the columns Min Fold, Min Signal and Evidence. For example, the MCF7 cell line contained 38 exon include events that passed the filtering criterion of a minimum fold change (linearized Splice Ratio) of 1.4 and minimum signal intensity for each probe of 300. The changes were observed in all sample comparisons between the sample group labeled MCF7 and the sample

group labeled MCF10A, *e.g.*, if there were only one sample or replicate for each cell line, there would be one pairwise comparison. If there were replicates, there would be two samples compared to two other samples.

In other embodiments, a report for the comparison of two sample groups is created. In another embodiment, a report for a multi-sample comparison, comparing each sample to all others or to the average of all others, is created. In another embodiment, a report for spliceoforms of each splice event type present or absent in one or more samples is created. For example, there may be 58 exon skip/include events where the skip form is present in a sample and 21 events where the include form is present in the sample.

In other embodiments, a profile of splicing scores for splice events in two or more samples is created. For example, a profile may comprise a vector of log ratio values for each sample, *e.g.*, (-0.01, 0, 0.006, -0.15, 1.21). The first element in the vector is the score for a first sample vs. a second sample; the second element is the score of a second sample vs. a third sample. As another example, the scores may be present/absent scores for samples taken independently. Given two such profiles, the splicing analysis method computes a distance measure. In one embodiment, a Pearson correlation coefficient between two splicing profiles is creates. In another embodiment, a Euclidean distance between two splicing profiles is created. In another embodiment, splicing profiles are converted into bit strings and a Hamming distance is computed. For example, where a bit-conversion assigns two bits to each sample and assigns a bit string of 00 for splicing scores less than plus or minus the standard error or confidence interval; a bit string of 01 is assigned to positive splicing scores greater than the standard error or outside of the confidence interval, and a bit string of 10 is assigned to negative splicing scores less than minus the standard error or outside of the confidence interval. The Hamming distance between two such bit strings is equal to the number of bits that differ. For example, the Hamming distance between 00 10 01 and 00 10 01 is 2, since the two middle bits differ. Other algorithms may be used, as known in the art.

A matrix of splicing profiles may be created, *e.g.*, each row contains the splicing scores for a splice event, and each column contains the splicing scores for

a sample or sample comparison. Alternatively, the matrix could be arranged with splicing events in columns and samples in rows.

In some embodiments, a mathematical algorithm is applied to profiles or matrices of splicing scores. In one embodiment, principle component analysis is performed. In another embodiment, a greedy clustering algorithm is used to iclusters the profiles. In another embodiment, self-organizing maps are used to cluster data. In another embodiment, a hierarchical clustering algorithm is used. In another embodiment k-means are used. Other algorithms are known in the art.

In some embodiments, data visualization is used to represent data that pass a filtering test, such as a filtering test in a scatter plot in two or more dimensions. Scatter plots are described and depicted herein, and are an aspect of the present systems and methods. Scatter plots enable a scientist to visually determine the extent of alternative splicing in a multi-sample comparison. In an embodiment, the splicing visualization method visually indicates data (in a scatter plot or spreadsheet or computer file) that passes the filtering criterion by using a different color from other data points. In one embodiment, the plots visually indicate different data by changing the size of the symbol used to indicate the data (for example by using a larger circle or larger square), by changing the symbol shape (for example by displaying highlighted data with squares and non-highlighted data with circles), by outlining, by making blink, or by giving an animation effect to certain data. In other embodiments data may be visualized by adjusting its opacity, saturation, hue, brightness, transparency, or one or more other visual attributes. In other embodiments data is visualized by displaying a label, a popup, a tooltip, or a text message. In another embodiment data is visualized by using a different font: a different font face, font style, font decoration, font size, or other attribute of the type face.

In some embodiments, the data representation visually differentiates between short form and long form probe data. For example, where the data contains one exon-exon junction probe for a spliced intron and two exon-intron junctions for a retained intron, a representation differentiates between the exon-exon junction probe's data and the two exon-intron junction probes' data. In another embodiment, short form and long form data is displayed in different colors, *e.g.*, red for short form data and orange for long form data.

In another embodiment, the score for splicing data is visualized by adjusting the hue. For example, a positive score could be red and a negative score green in a spreadsheet or scatter plot. In another embodiment, it visually indicates the score using the color saturation. For example, larger positive scores might be brighter red, larger negative scores brighter green, and scores close to zero nearly black. In another embodiment, transparency is used to highlight certain data. For example, scores close to zero might be nearly transparent, whereas scores with large absolute values might be more opaque. In another embodiment, it visually indicates the score using the symbol size. For example, larger points in a scatter plot may indicate scores with larger absolute values. The data analysis method may visually indicate the score using any of the methods suggested above, or using another method that may occur to one of skill in the arts.

Data relating to one splice event type may be represented one way and data relating to another splice event type may be represented another way, *e.g.*, different shaped data points or different colors.

A spreadsheet, scatter plot, line plot, or other image with visually indicated data (for splice events, for data that passes a filter, for up- or down-regulation, for a splicing score, etc) is an aspect of the present compositions, devices, systems, and methods. Data in a spreadsheet or scatter plot may be used in combination with a gene model viewer (a "splice graph"). In one embodiment, the user indicates/selects data in a scatter plot and the corresponding part of the gene model is visually represented. For example, the user may click on a spreadsheet row containing data for an exon-exon junction probe, and the gene model viewer may open, if it is not already open, and that particular exon-exon junction will be visually represented. Similarly, the user may move a mouse over a point in a scatter plot, and the relevant part of the gene model will be visually indicated, *e.g.*, by highlighting. The visualization of the selected data may employ any of the methods already mentioned, or other known in the art. For example, the visual indication might involve changing the color of a portion of the gene model, or underline, outline, label with text, or display an icon near the portion of the gene model.

Splicing analysis may be linked to software resources such as gene ontology tools, alternative splicing databases, sequence databases, genome

browser, pathway software, chemistry database, gene model viewer, molecular biology/proteomics tools, etc.

For example, suppose an exon-exon junction probe is annotated with ACCESSION_ID and GENE_SYMBOL and has genomic coordinates on chromosome 6 of 5000 to 5020 for the first exonic portion and of 6000 to 6020 for the second exonic portion. In a software application, the probe annotation might be displayed in a spreadsheet, or represented as a point in a scatter plot or line plot, or in a region of a gene model viewer. The user might indicate the probe annotation, representation or region by moving a cursor or navigating using the keyboard. The indicated probe might then be highlighted, or a tooltip or popup window or context menu might appear. The user then gives a specified cue; such as a mouse cue (left click, right click, center click, mouse wheel, mouse drag, hover) or keyboard cue (key press) or input on a touch screen, or a combination of these or other methods. Afterward, the software application launches an external tool. For example, a hyperlink might open with a URL to a web-based tool. In one embodiment, a probe is linked to a genome browser displaying information related to the genomic or chromosomal region targeted by the probe. For example, the University of California Santa Cruz genome database and web browser might open, displaying a base range that includes the probe's genomic locations. The browser might display genomically aligned sequences within that base range. As another example, the genome browser displays the nucleotide sequence of the probe. As another example, the genome browser displays the nucleotide sequence of the genomic region to which the sequence with the ACCESSION_ID aligns. For example, suppose the probe detects a spliceoform indicated by accession ABC12345 and the sequence with that accession has been aligned to chromosome 6 with a given set of coordinates. The genome browser displays the nucleotide sequence of that coordinate set.

In one embodiment, splicing data is linked to a resource using a hyperlink; a resource using a menu item in the menu bar or in a context menu; a resource using a toolbar button; a resource using a keyboard shortcut; a resource using a mouse cue; or a resource using a mouse cue and a keyboard shortcut. In some embodiments splicing data is linked to a resource using another input method or

cue. A variety of methods may be employed to connect splicing data to software and database resources.

In some embodiments, the present compositions and methods are used to generate data that is stored on a medium. In related embodiments, the storage medium contains control data relating to polynucleotide controls. The storage medium may contain data obtained from a study using one or more indicator polynucleotides, optionally along with appropriate controls. The storage medium may contain data obtained from a study using one or more arrays containing indicator polynucleotides, optionally along with appropriate controls.

The storage medium, for example, can contain expression level data for one or more indicator polynucleotide, such as swap controls or balanced mutation controls, or other controls. The storage medium may take the form of, for example, a computer hard drive, a memory device, a compact disc (CD), a digital video disk or digital versatile disk (DVD), computer cache, paper, magnetic tape, or any medium capable of temporary or permanent storage of a data set, data representation, or a computer file. The computer file can be, for example, a tab-delimited or csv file, a nucleotide sequence file, a spreadsheet file, a database table, an XML document, an HTML table, or the like. The computer file may be binary, ASCII, encrypted, encoded, etc. The data representation may comprise a portion of a file (*e.g.*, part of a database) or it may span multiple files (*e.g.*, a normalized database comprising multiple data tables stored in separate locations). The data representation may comprise print material, such as a computer printout or lab notebook. The data may be used for data analysis purposes, for archival purposes, etc.

In one embodiment, the data contained on the storage medium may comprise an identifier for an indicator polynucleotide; an identifier for an indicator polynucleotide control; a numerical value, such as an expression level, for the indicator polynucleotide; a numerical value, such as an expression level, for the indicator polynucleotide control; a sequence for the indicator polynucleotide; and/or a sequence for the indicator polynucleotide control. The data may include other information such as a public database identifier; an exon index, intron index, pair of exon indexes or an exon index and an intron index; a background level from an array experiment, a mean or median value from a fluorescent scanner; a

number of pixels; a flag for sequence quality or presence of homologues; or other information that may occur to one of ordinary skill in the art. The data set may comprise any combination of these types of information. The data set may comprise such information for a plurality of pairs. The representation may comprise such data for 100 or more pairs, or 1,000 or more pairs, or 10,000 or more pairs, or 100,000 or more pairs, or 1,000,000 or more pairs.
The data may be transmitted via a networking protocol such as UDP, TCP-IP, FTP, SMTP or HTTP. Alternatively, the data may be transmitted via domestic mail, international mail, courier, etc. .

The data may be transferred in any of the ways mentioned in the patents and patent applications incorporated by reference, or in any other way that may occur to one skilled in the art.

It will be appreciated that the methods described above, and the analysis described below, can be accomplished using a software program written to conduct the described methods or analysis. The software is stored on a suitable storage medium, such as those already mentioned above.

Data storage, data transmission, and software applications are aspects of the present invention. Data stores may be created using any of the raw data, calculations, or analyses provided by the present compositions, methods, and systems.

VII. Comparison of Gene Expression in Different Cells

In experiments performed in support of the present compositions, devices, software, systems and methods, an exemplary splice variant polynucleotide microarray was used detect and quantify splice variants in samples obtained from two cell lines with expected differences in splicing patterns.

A. MiaPaCa2 vs. HEK293 cells

In a first experiment, splicing patterns were analyzed in MiaPaCa2 pancreatic cancer cells and HEK293 embryonic kidney cells (ATCC Manassas, VA, USA). These experiments are described in Examples 1-4.

Table 1, below, summarizes the differential splicing and transcriptional observed in MiaPaCa2 and HEK293 cell lines with a 99.9% significance interval

calculated using the empirical distribution from replicate data as a control. In the first three rows of Table 1, data obtained using each indicator molecule is considered separately. In subsequent rows, data obtained using multiple indicator molecules is combined to detect the indicated types of splicing events. Table 2 shows normalization of the data by overall gene expression

The arrays were able to detect differences in the expression of short-form vs. long-form splice variants with exon skips, intron retentions, exon trims and extensions, alternative first and last exons, and other features described in the Examples.

B. MCF7 vs. CaCO2 cells

In another experiment performed in support of the present compositions, devices, software, systems and methods, splicing patterns were compared in MCF7 breast cancer cells and CaCO2 colon carcinoma cells (Example 5). Figures 3A-3H show scatter plots obtained using indicator polynucleotides (in a microarray) corresponding to particular polynucleotide sequences of differentially expressed splice variants with a minimum signal of 200 and a "Splice Fold" (linearized Splice Ratio) score ≥ 2 (a >99.9% confidence interval for all splice types). Plots on the left show the splice variants present in MCF7 cells vs. CaCO2 cells. Plots on the right show technical replicates from HEK293 cells. The indicator polynucleotides used in the microarrays detect exon skip events (panels A and B); alternative first and last exons (C and D), intron retentions (E and F); and alternative acceptor and donor sites (G and H). Different gene isoforms are clearly present in the different cell types.

The data were analyzed using a database of alternative splicing in human (*i.e.*, the SpliceExpress Human Spliceome database, or SEHS), or the Splicing Index, ASPIRE, and Splice Ratio methods described herein. A comparison of the results is shown in Figure 4. A summary of some of the results is shown in Table 3.

The results of the experiments described in (A) and (B) show the qualitative and quantitative differences between the splicing patterns in different cells lines

The following example is provided to illustrate the methods. Additional embodiments will apparent to one skilled in the art without departing from the scope of the invention.

EXAMPLES

Example 1 Microarray Design

A human genome-wide microarray was designed using Build 35 of the NCBI genome assembly and by aligning sequences present in multiple data bases (RefSeq, mRNAs, ESTs and Alternative Splicing Database (ASD)) for comparison of splice isoforms against the normal loci. Splice isoforms were filtered based on the number of sequences providing evidence of each splice site, sequence quality, and intron acceptor and donor sites. After filtering redundant splice isoforms, 98,382 isoforms contained unique genomic coordinates. They included 279,438 non-redundant exon-exon or exon-intron junctions and 301,475 distinct exons or portions of exons potentially subject to alternate transcription or splicing. Of the exon-exon and exon-intron junctions, 134,212 targeted alternately spliced sites.

To measure RNA splicing, 148,603 indicator polynucleotides (oligonucleotide probes) were designed as described herein. The set of indicator polynucleotides included probes for alternately spliced exon-exon and exon-intron junctions as well as probes for single-exon genes and exons or exon-exon junctions not known to be alternately spliced. The probes had an average length of 38 nucleotides. Variability in melting temperatures was minimized by varying probe lengths between 34 and 42 nucleotides and isothermally balancing junction probes across the splice sites. Sense strand oligonucleotides with a 10-base polyT linker at the 3' end were synthesized by NimbleGen Systems (Madison, WI) using photodeposition chemistry and digital light projection.

To produce data for analysis and algorithm development, we chose two human cell lines with extensive differences expected in both gene expression and alternative splicing: (i) a pancreatic carcinoma cell line MiaPaCa2 and (ii) embryonic kidney cell line HEK 293. Both cell lines were obtained from the American Tissue Culture Collection (Manassas, VA) and cultured according to ATCC specifications. The cell lines were grown in 10 cm dishes to 70-80% confluency, harvested with cell dissociation solution, and lysed in Trizol at $1x10^7$

cells/ml (InVitrogen). RNA was purified from each Trizol suspension using the Qiagen RNEasy Isolation Kit (Qiagen Inc., Valencia, CA) and the integrity and concentration established by the Agilent (Santa Clara, CA) 2100 Bioanalyzer and spectrophotometry. Aliquots of total RNA were used to produce cDNA using the Invitrogen (Carlsbad, CA) Superscript™ Double Stranded cDNA Synthesis Kit, following the standard protocol with the exception that after second strand cDNA synthesis was stopped by the addition of 0.5 M EDTA but before the phenol-chloroform-isoamyl alcohol step, 1 µl of 10 mg/ml RNase A solution was added to the tubes, which were incubated at 37°C for 10 minutes and rehydrated to a final concentration of 250 ng/µl.

Direct one-color labeling was performed using 1 µg of cDNA. Briefly, 40 µl of Cy3 labeled random primers (Trilink Biotechnologies, San Diego, CA) were mixed with 1 ug of cDNA in 80 µl of water, heated to 98° for 5 minutes, and chilled in an ice water bath. An aliquot of 10 µl 50X dNTP mix was added to each tube followed by 8 µl of water and 2 µl of high concentration NEB (Sunnyvale, CA) Klenow fragment (40 U/µl). Reactions were incubated at 37°C for 2 hours, precipitated with isopropanol, and evaporated to dryness using a SpeedVac concentrator/lyophylizer on low heat for 5 min. Prior to hybridization, samples were resuspended in 20 µl water. Dye incorporation and concentration were determined by spectrophotometry.

For hybridization, 13 µg of Cy3-labeled sample was added to 40 µl of 2X Nimblegen hybridization buffer, and heated at 95°C for 5 min. Samples were loaded onto the microarrays through the MAUI Mixer sample ports and hybridized for 16-20 hrs. at 42°C in a MAUI Hybridization System (Biomicro, Salt Lake City, UT). The arrays were washed three times, dried in an Array-Go Round, and scanned with an Axon (Sunnyvale, CA) scanner. Replicate data were generated for each cell line and the relevant features were extracted.

Example 2 Raw Data Analysis

Raw data from Example 1 were matched with probe annotations. Data analysis was performed in R and Java after Loess normalizing by local mean. Gene expression was calculated by taking the geometric mean of the expression

level at each base range (or 'splice module') along the gene model that was targeted by at least one probe:

$$G = \sum I = (\prod B)^{1/n} \quad (1)$$

$$B_{exon} = p + p^c \quad (2)$$

$$B_{junc} = \sqrt{((p_5 + p_5^c) * (p_3 + p_3^c))} \quad (3)$$

where G is the gene expression level; each I is a splice isoform; each B is the sum of expression levels of all splice isoforms at a base range targeted by one or more probes; n is the number of probes for the gene; B_{exon} covers the case of an exon or intron probe, with p equal to a probe signal and p^c equal to the sum of signals of probes exclusive with (or complementary to) p; and B_{junc} addresses the case of an exon-exon or exon-intron junction probe, with p_5 being the 5' portion of the junction and p_3 being the 3' portion, each portion having its own complement.

To identify regulated splicing, we normalized per-probe expression data by the gene expression level from Equation 1. The equation

$$S = p\ M/G \quad (4)$$

performs the normalization, wherein S is the normalized 'splice signal' of a probe in a sample, p is the probe signal, G is the gene expression level, and M is the geometric mean of G and the geometric mean of gene expression levels in the sample and at least one other sample. Then, to compute regulated splicing, we used the equation:

$$R = R_{short} - R_{long} \quad (5)$$

$$R_{short} = \log J_1/J_2 \quad (6)$$

$$R_{long} = \text{ave}(\log K_1 / K_2) \quad (7)$$

$$R_{long} = d_1 * \min |\log K_1 / K_2| \quad (8)$$

where R is the splice ratio for the splice event, R_{short} is the log ratio of the short splice form (an exon skip, intron excision or truncated exon) in two samples, J_1 is the signal in the first sample of the junction probe spanning the alternately spliced region, and J_2 is the signal in the second sample; R_{long} is the log ratio of the long splice form (an exon inclusion, intron retention or extended exon) in two samples, K_1 equal to the signal in the first sample of a long form probe (a junction or an alternately spliced module), K_2 equal to the signal in the second sample, and d_1 the sign of K_1; if two K_1/K_2 pairs had opposite signs, we set $R_{long} = 0$. As a further constraint to filter out cases where one splice isoform was not expressed in

either sample, we required the additional condition that splice ratios change in opposite directions ($R_{short} * R_{long} < 0$), and then used a more stringent value:

$$|R| = 2 * \min (|R_{short}|, |R_{long}|) \quad (9)$$

We computed G, S and R for both replicate pairs as well as for control pairs. To filter out false positives, we took the minimum value of R in the two replicates, or zero if the log ratio (for gene expression) or splice ratio (for regulated splicing) disagreed in direction. P-values were calculated using the empirical distribution from the control pairs. For gene expression and for each category of splice event—skipped middle exon, retained intron, 5′ trim, etc.—p-values were calculated separately to ensure sensitivity to any differences resulting from particular probe combinations.

Example 3. Results of Data Analysis

Introduction

Based on the raw data analysis in Example 2, both gene expression levels and RNA splicing patterns differed considerably in the two cell lines tested (Table 1). More exon trims occurred in MiaPaCa2 cells, with exon skips and spliced introns being more common in HEK293 cells. The relative frequency of observed splice events with 99.9% confidence was greatest for exon skips and intron retentions, together forming a majority (Figure 1). Alternative first exons were observed in both the sequence databases and the splice array results with greater frequency than alternative last exons. Trims or extensions at the 5′ end of exons were more common than at the 3′ end. However, the ratio of observed to expected splice events was approximately constant for all events, suggesting that the proportions/ratios depended on the splice events the array was designed to detect.

Table 1. Splicing and Transcriptional Changes

	Points	Events	Observed	Expected	MiaPaCa2	HEK293
Genes	18169	18169	3030	18	1268	1755
Probes	148603	148603	18044	149	9247	8712
Alt-spliced	134212	134212	12073	134	6070	5817
Exon Skips	56342	24943	488	25	217	271
Middle Exon	27602	9361	230	9	114	116
First Exon	11367	5856	140	6	66	74
Last Exon	6768	3455	50	3	23	27
Exon Trims	21297	11108	243	11	146	96
5' Trim	11683	5849	111	6	72	39
3' Trim	9268	4639	89	5	50	39
Intron spliced	31103	10638	194	11	76	118

Determing splicing patterns

To identify splicing patterns, probe expression data was normalized to gene expression levels (computed using Equation 1). The log-linear translation slides the probes for a gene perpendicularly to the diagonal until the gene expression level lies exactly along it (Figure 2). Non-expressed probes cause the spread near the origin. Additional normalization by local variance could help mitigate the effect, though one should of course be wary of probes with expression levels close to background.

After normalizing by gene expression, points farther from the diagonal may indicate differential splicing, independent of changes in gene expression level. The specific probes that lie away from the diagonal may target alternately spliced regions of a gene. If a gene contains multiple such points, the evidence is more persuasive. Using splice signal levels, fold change ('splice fold') and log ratio ('splice ratio') can be computed using standard spreadsheet software or a gene expression analysis package. Splice signals have the benefit that they retain an absolute expression level, making it easy to visually distinguish between large splice fold changes in high expressed versus low expressed probes. Splice signals

allow a scientist to simultaneously view all splice fold changes for a two category experiment.

Normalization by gene expression reduced the variance by 30.75% for the sample comparison, having, as expected, little effect on the controls (Figure 2, Table 2). The difference in variance can be attributed to gene expression differences between the samples, the remaining variation resulting from splicing, error in the estimate of gene expression, and noise in probe expression levels. For highly expressed probes, the reduction was greater for multiple reasons: highly expressed probes affect gene expression calculation more; many splice forms have low expression; and normalization leaves the diffuse scatter already mentioned. By comparison, experimental error from replicate experiments accounted for 12.41% as much variance, leaving potentially 56.83% of the variance due to alternate splicing as well as residual error in the gene expression calculation. Put another way, alternate splicing may have accounted for up to 1.85 (56.83/30.75) times as much of the variance as gene expression.

Table 2. Effects of Normalizing Probes by Gene Expression

	Probes		Normalized		Change in σ^2
	R^2	σ^2	R^2	σ^2	
MiaPaCa2 vs HEK293 (1)	0.889	0.0762	0.922	0.0523	-30.97%
MiaPaCa2 vs HEK293 (2)	0.885	0.0790	0.919	0.0549	-30.54%
MiaPaCa2 vs MiaPaCa2	0.982	0.0122	0.982	0.0119	-3.05%
HEK293 vs HEK293 (1)	0.985	0.0103	0.985	0.0102	-0.98%
HEK293 vs HEK293 (2)	0.985	0.0105	0.985	0.0106	+0.59%

Example 4. Further data analysis

Conventional methods of splice variant analysis have been based on the difference of log ratios (see, *e.g.*, Equation 5 *et seq.* in Example 2), *i.e.*, the ratio of expression of one splice variant in two samples minus the ratio of a second splice variant in the same two samples. The relationship can also be described as a splice fold change or a change in percent composition of the splice forms. In such

cases, probes typically detect a short splice form having an exon-exon junction that eliminates an intron, exon, or exon extension present in a long splice form. The long form is detected by a first exon-exon or exon-intron junction probe, an exon probe, and a second exon-exon or exon-intron junction probe.

A limitation of such convention methods is that it considers only two splice forms, a short and a long form. In cases where there are more splice forms, such analysis will overlook other splice variants.

In one embodiment of the present compositions, devices, systems, and methods the minimum log ratio (Equation 8 in Example 2) was used rather than Equation 7, allowing the screening for only those alternative splice events where all measurements (based on indicator molecules) were in agreement.

Another limitation of conventional methods arises when one splice variant is not expressed in a sample being compared. Whenever the over transcription level of the gene changes, the log ratio of the unexpressed splice variant relative to the other splice variant will also change. Consideration of the absolute expression level, or log average of the probe intensities, can help to screen out such false positives. However, on larger data sets, this approach may be inadequate, especially when an array detects many rare splice isoforms. In this situation, the log ratios for the rare splice variants will change anytime the overall gene expression level changes, even without the rare forms being expressed. To address this case, we filtered data by requiring that two splice isoforms be regulated in opposite directions: one up and one down, or vice versa (Equation 9). In the ordinary scatter plot, the changes are absolute and relative to gene expression levels (Figure 2). This 'opposite' stipulation, specified as a minimum fold change for up- and down-regulation, seems to effectively filter out many non-expressed splice forms.

By considering multiple probes together for each splice event, filtering changes using a stringent minimum log ratio, and requiring splice forms to change in opposite direction relative to gene expression, the sensitivity of detection is greater than by looking at changes in individual probes alone.

Levels of gene expression were also calculated by taking splicing into account (using Equation 1) and plotted against gene expression data calculated by naively averaging all probes, including those specific to rare splice forms. The

results suggest that splice variant arrays may lead to higher estimates of gene expression levels, and that the levels may differ from results derived from conventional gene arrays. The difference between the two estimates may depend on the proportion of indicator molecules that target alternately spliced regions, since constitutive probes should behave the same under both methods (not shown).

Some additional observations are in order, related to background levels, alternative first and last exons, and micro-exons. When summing probe signal levels, background readings from non-expressed probes have the potential to falsely inflate the expression level. Therefore, background-subtracted levels can be used. When a gene has known splice variants with alternative first or last exons, the signal of the short form should be included in the complements of the long form base ranges in Equation 2 and 3. The short form signal can often be determined from a probe near the terminus of the short form sequence, provided that it uniquely identifies the short form. In the case of micro-exons, some junction probes may in fact span three or more exons. Equation 3 can readily be generalized to the cube root of the products of the sums of each of the three base ranges in an exon-exon-exon junction.

The reliability of gene expression measurements from splice variant microarrays using Equation 1 will depend upon the underlying microarray platform as well as on the set of expected splice products—so long as many splice forms are inferred only from partial ESTs, there will be guesswork involved. However, as more splice variants are discovered and sequenced, the accuracy of arrays will improve. By predicting exons and exon-exon junctions, all possible splice isoforms can in theory be anticipated and included in a model. However, each predicted probe will bring its own experimental error to the calculation, so there may be a tradeoff between completeness and noise.

Example 5. Analysis of spice variants in MCF7 cells and CaCO2

Figures 3A-3H show scatter plots obtained using indicator polynucleotides corresponding to particular polynucleotide sequences of differentially expressed splice variants with a minimum signal of 200 and a "Splice Fold" (linearized Splice

Ratio) score ≥ 2 (a >99.9% confidence interval for all splice types). The indicator polynucleotides were used in a microarray.

Plots on the left show the splice variants present in MCF7 cells vs. CaCO2 cells. Plots on the right show technical replicates from HEK293 cells. The indicator polynucleotides used in the microarrays detect exon skip events (panels A and B); alternative first and last exons (C and D), intron retentions (E and F); and alternative acceptor and donor sites (G and H). Different gene isoforms are clearly present in the different cell types.

The data were analyzed using a database of alternative splicing in human (*i.e.*, the SpliceExpress Human Spliceome database, or SEHS), or the Splicing Index, ASPIRE, and Splice Ratio methods described herein. A comparison of the results is shown in Table 3 and Figure 4.

Table 3: Summary of analysis of splice types

Data Type	Instances	MCF7	MCF10A	Min Fold	Min Signal	Evidence
Genes Up	2679	247	313	2	300	All
Genes Down	2679	313	247	2	300	All
Alt Spliced: Short	16247	426	456	1.4	300	All
Alt Spliced: Long	16247	456	426	1.4	300	All
Exon Skips	2217	30	38	1.4	300	All
Exon Includes	2217	38	30	1.4	300	All
Alt First Exons: 3'	1155	27	27	1.4	300	All
Alt First Exons: 5'	1155	27	27	1.4	300	All
Alt Last Exons: 5'	530	23	13	1.4	300	All
Alt Last Exons: 3'	530	13	23	1.4	300	All
Alt 5' Donor: Short	712	24	33	1.4	300	All
Alt 5' Donor: Long	712	33	24	1.4	300	All
Alt 3' Acceptor: Short	1071	58	27	1.4	300	All
Alt 3' Acceptor: Long	1071	27	58	1.4	300	All
Intron Splices	708	14	7	1.4	300	All
Intron Retentions	708	7	14	1.4	300	All

Table 3 identifies the number of 'instances' of each data type (gene or splice event type), such as alternatively spliced short spliceoforms, or alternative donor sites (short form or long form). It shows the number of instances of each

splice event that pass the filtering criterion defined in the columns Min Fold, Min Signal and Evidence. For example, the MCF7 cell line contained 38 exon include events that passed the filtering criterion of a minimum fold change (linearized Splice Ratio) of 1.4 and minimum signal intensity for each probe of 300. The changes were observed in all sample comparisons between the sample group labeled MCF7 and the sample group labeled MCF10A, *e.g.*, if there were only one sample or replicate for each cell line, there would be one pairwise comparison. If there were replicates, there would be two samples compared to two other samples.

CLAIMS

What is claimed is:

1. A method for determining an expression level for a gene having at least a first splice isoform and a second splice isoform, the method comprising:

(a) obtaining microarray expression level data for a plurality of mutually exclusive indicator polynucleotides for exons, introns, modules, exon-exon junctions, exon-intron junctions, intron-exon junctions or module-module junctions of the gene, and

(b) applying a mathematical algorithm to determine the expression level for the gene.

2. The method of claim 1, wherein the mutually exclusive indicator polynucleotides are non-overlapping.

3. The method of claim 1, wherein the mutually exclusive indicator polynucleotides are overlapping.

4. The method of any of claims 1-3, wherein at least one mutually exclusive indicator polynucleotide indicates a polynucleotide that is constitutively present in expected splice isoforms of the gene.

5. The method of any of claims 1-3, wherein at least one mutually exclusive indicator polynucleotide indicates a polynucleotide that is not constitutively present in expected splice isoforms of the gene.

6. The method of any of claims 1-5, wherein the gene expression level is calculated by summing the amount of signal corresponding to the mutually exclusive indicator polynucleotides at each nucleotide base range indicated by the indicator polynucleotides.

7. The method of any of claims 1-6, wherein the overall level of gene expression is calculated using the equation:

$$G = \left[\sum_{e} \left(p_e + p_e^c \right) + \sum_{j} \sqrt{\left(p_j + p_{5j}^c \right) * \left(p_j + p_{3j}^c \right)} \right] / P$$

wherein G is the gene expression level;

each p_e is separately a signal for an indicator polynucleotide for an exon, intron or module of the gene;

each p_e^c is separately the geometric mean of the signals of indicator polynucleotides complementary to p_e;

each p_j is separately a signal for an indicator polynucleotide for an exon-exon junction, exon-intron junction, intron-exon junction, or module-module junction of the gene, and wherein p_j comprises a 5′ portion and a 3′ portion;

each p_{5j}^c is separately a signal for an indicator polynucleotide that is mutually exclusive with the 5′ portion of p_j;

each p_{3j}^c is separately a signal for an indicator polynucleotide that is mutually exclusive with the 3′ portion of p_j; and

P is the total number of included probes.

8. The method of any of claims 1-7, wherein a background level of a hybridization signal for a mutually exclusive probe is subtracted from the overall expression level.

9. The method of any of claims 1-8, wherein at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon.

10. The method of any of claims 1-8, wherein at least one of the plurality of mutually exclusive indicator polynucleotides indicates an intron.

11. The method of any of claims 1-10 wherein at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon-exon junction.

12. The method of any of claims 1-11 wherein at least one of the plurality of mutually exclusive indicator polynucleotides indicates an exon-intron or intron-exon junction.

13. The method of any of claims 1-11 wherein at least one of the plurality of mutually exclusive indicator polynucleotides indicates a module-module junction.

14. Software for performing the calculations of claim 1.

15. Software for performing the calculations of claim 7.

16. A data store comprising expression levels for two or more genes, the expression levels being calculated using any of the methods of any of claims 1-13.

17. A method for identifying alternative splicing of a gene in one or more samples from microarray expression level data for a plurality of indicator polynucleotides for exons, introns, exon-exon junctions, exon-intron junctions, intron-exon junctions or module-module junctions of the gene, the method comprising:

(a) obtaining expression level data for the plurality of indicator polynucleotides in one or more samples;

(b) applying a mathematical algorithm to calculate a value for an alternative splicing event, wherein the mathematical algorithm involves a gene expression level for the gene; and

(c) identifying the indicator polynucleotides for which the measure exceeds a cutoff value.

18. The method of claim 17 wherein the gene expression level is calculated using the method of claim 1-13.

19. Software for performing the calculations in claim 18.

20. A data store comprising the values calculated by the mathematical algorithm of claim 18.

21. A scatter plot of data for a plurality of indicator polynucleotides for one or more genes, wherein the data are the results of applying a mathematical algorithm to the expression levels of the indicator polynucleotides.

22. The scatter plot of claim 21 wherein the mathematical algorithm involves a calculation of a gene expression level for each of the one or more genes.

23. The scatter plot of claim 21 wherein one or more data points are visually indicated based on results calculated using a mathematical algorithm

24. The scatter plot of claim 23 wherein the results are calculated using the method of claim 22.

25. The scatter plot of claim 29 wherein the visually indicated data points are indicated by a color, hue, saturation, brightness, or transparency level

26. The scatter plot of claim 29 wherein the visually indicated data points are indicated by a shape, outline or symbol.

27. The scatter plot of claim 29 wherein the visually indicated data points are indicated by a label.

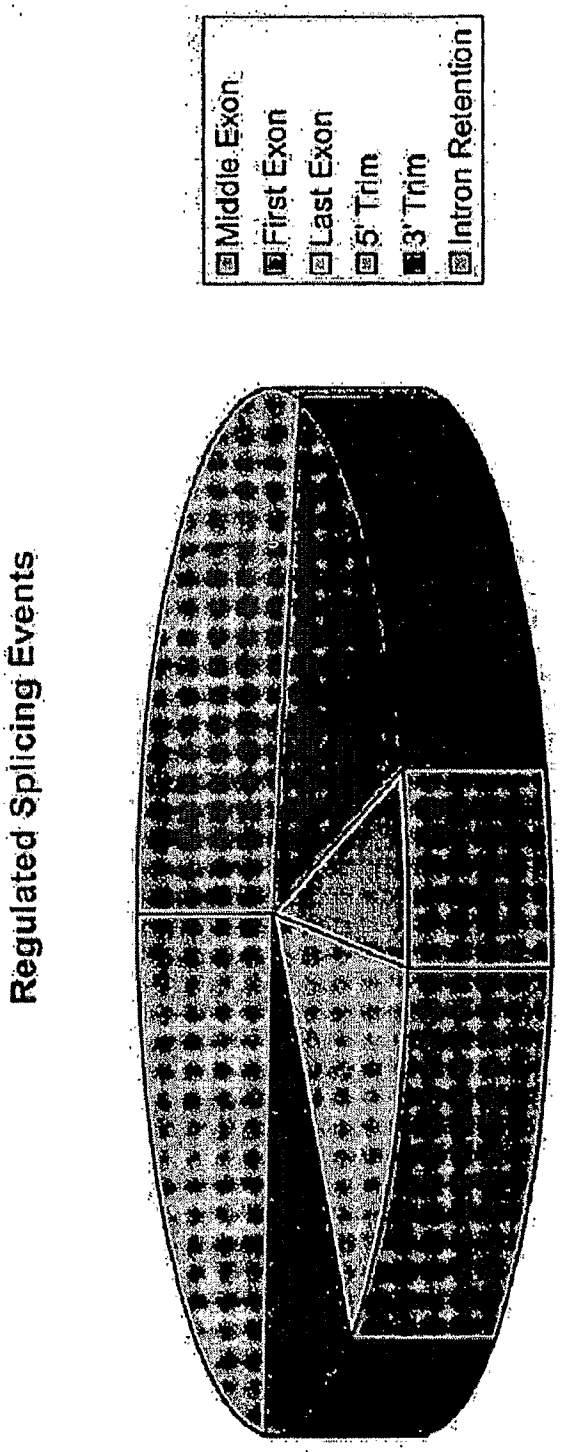
Middle Exon
First Exon
Last Exon
5' Trim
3' Trim
Intron Retention
Regulated Splicing Events

Figure 1

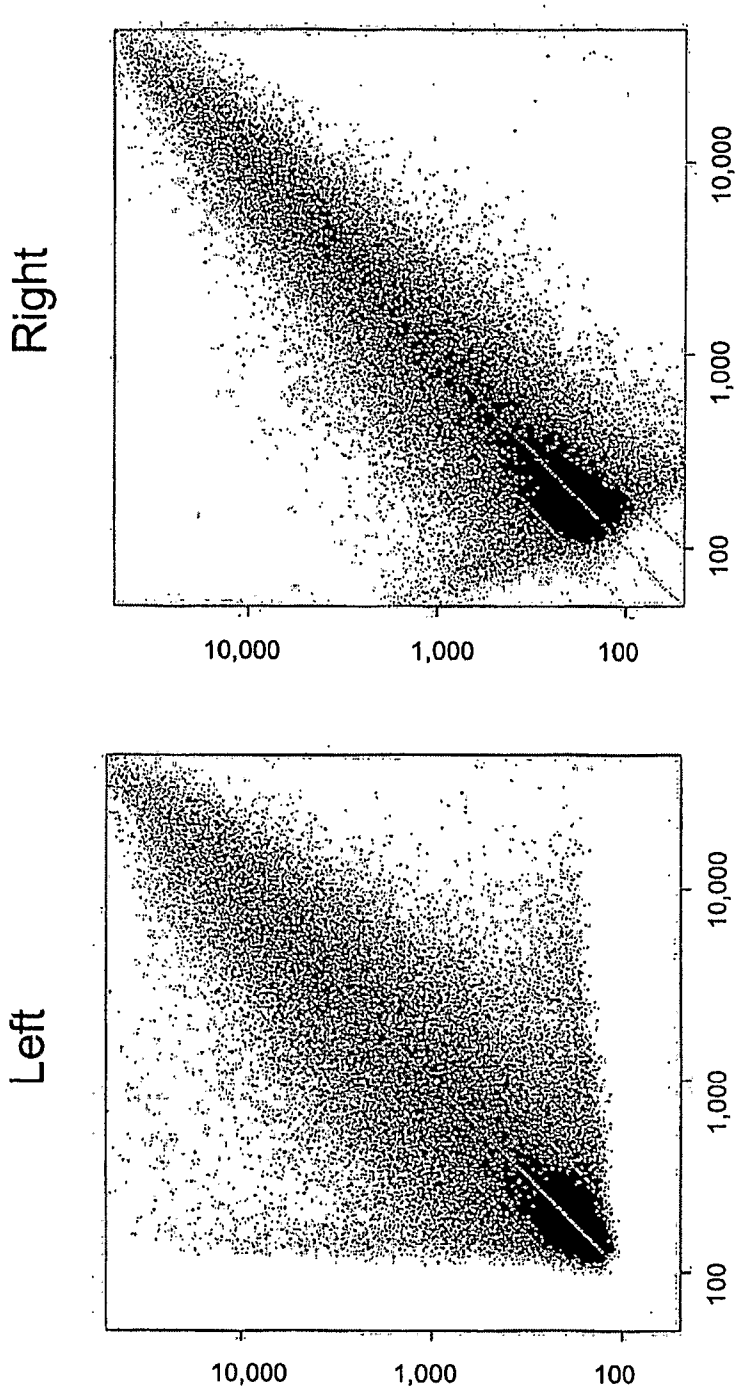
Right
10,000
1,000
100
10,000
1,000
100
Left
10,000
1,000
100
10,000
1,000
100

Figure 2

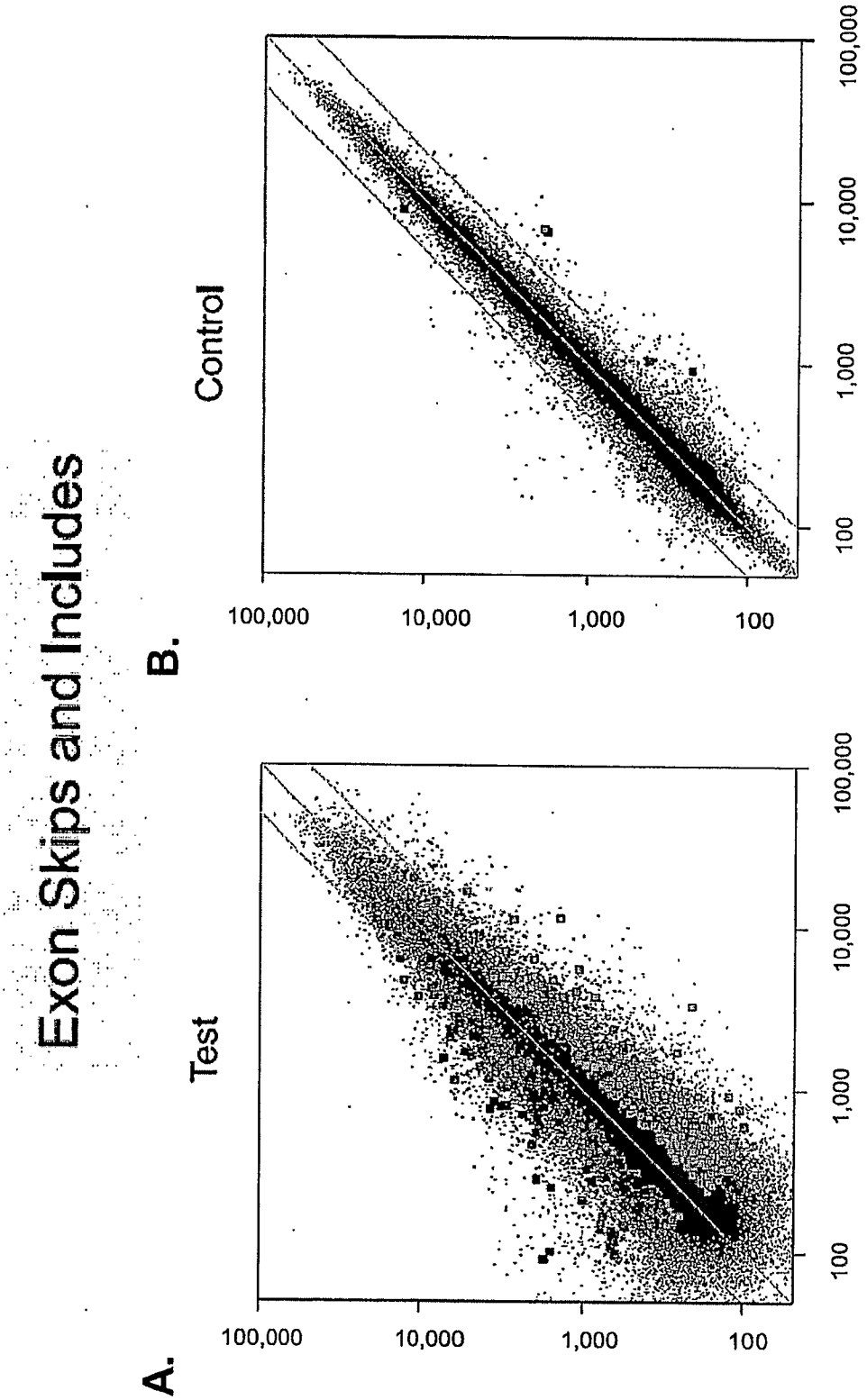
Exon Skips and Includes
A.
Test
100,000
10,000
1,000
100
B.
Control
100,000
10,000
1,000
100

Figure 3

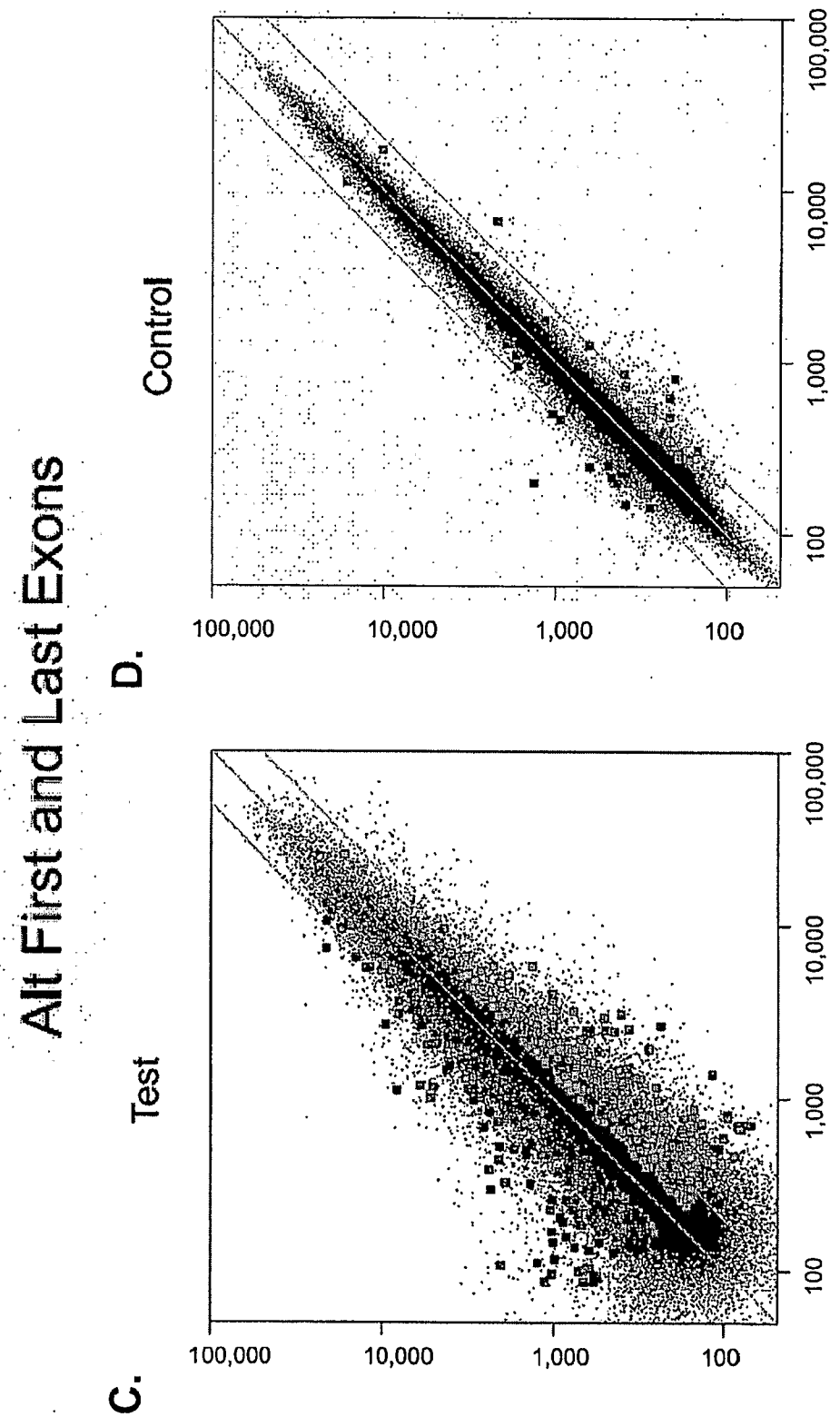

Figure 3 (cont.)

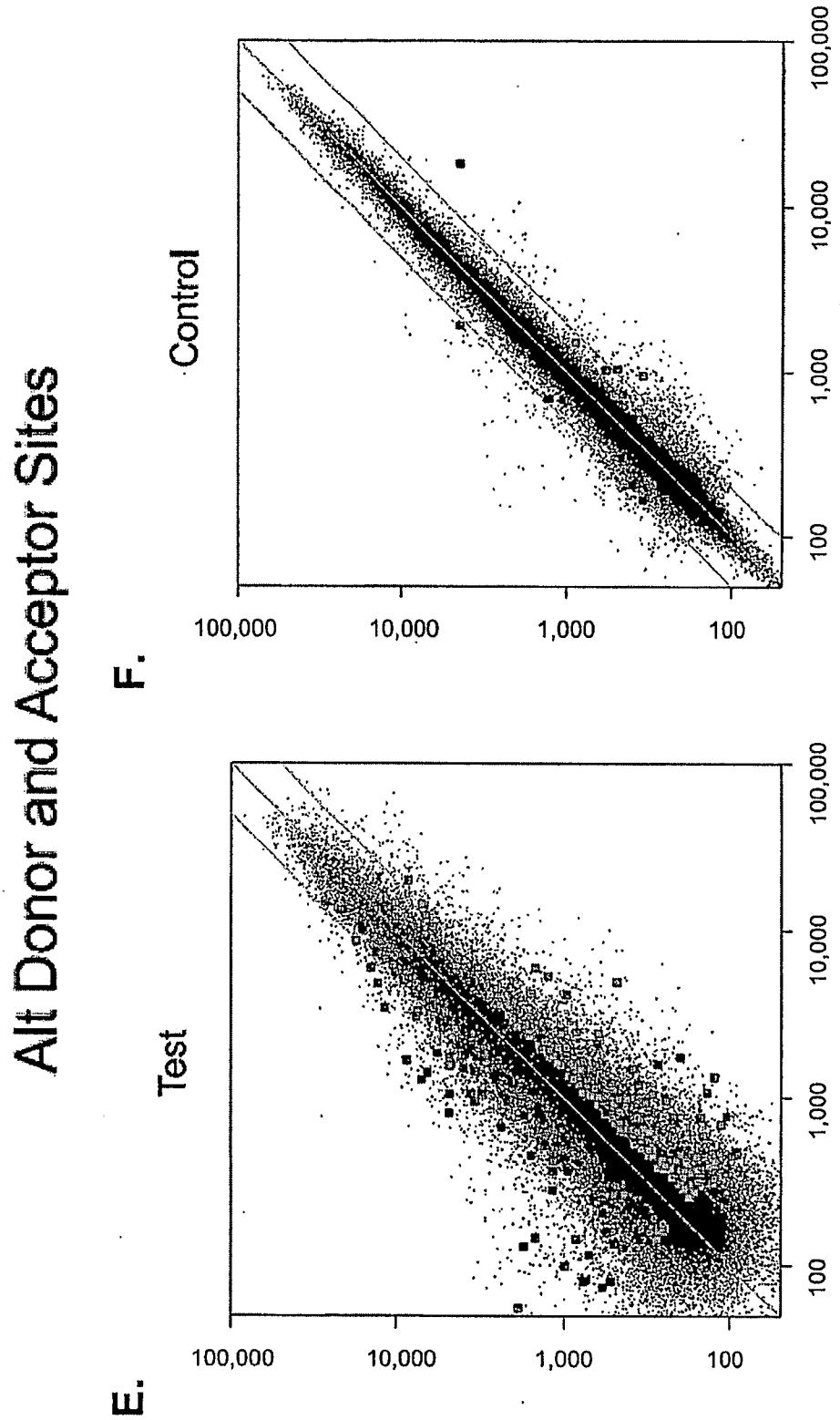

Figure 3 (cont.)

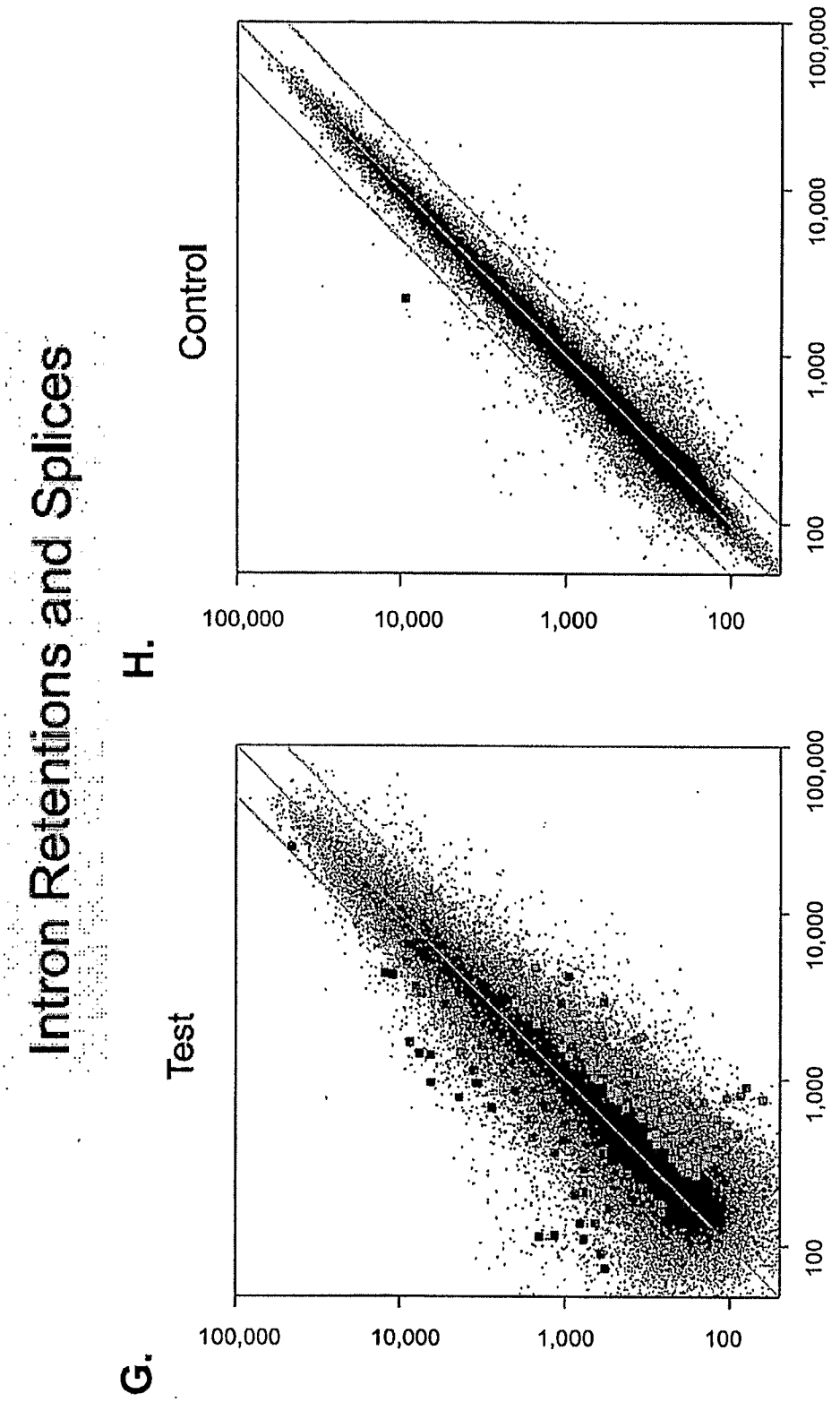

Figure 3 (cont.)

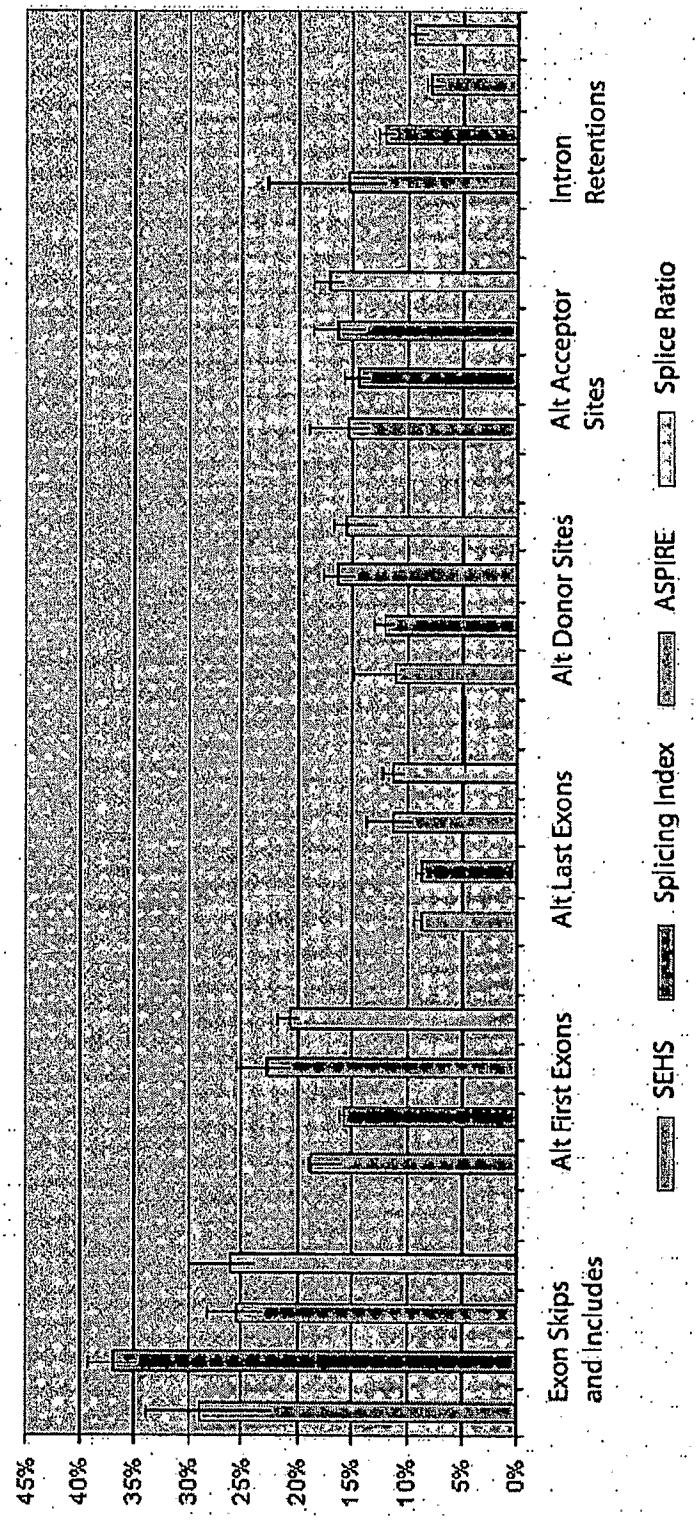
Distribution of Differentially Expressed Splice Types
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Exon Skips and Includes
Alt First Exons
Alt Last Exons
Alt Donor Sites
Alt Acceptor Sites
Intron Retentions
SEHS
Splicing Index
ASPIRE
Splice Ratio

Figure 4