A Statistical Framework for Genome-Wide Discovery of Biomarker Splice Variations with GeneChip Human Exon 1.0 ST Arrays

Ryo Yoshida[*]

yoshidar@ims.u-tokyo.ac.jp

Kazuyuki Numata[*]

numata@ims.u-tokyo.ac.jp

Seiya Imoto[*]

imoto@ims.u-tokyo.ac.jp

Masao Nagasaki

masao@ims.u-tokyo.ac.jp

Atsushi Doi

doi@ims.u-tokyo.ac.jp

Kazuko Ueno

uepi@ims.u-tokyo.ac.jp

Satoru Miyano

miyano@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[*] These authors equally contributed to this work.

Abstract

Alternative splicing is an important regulatory mechanism that generates multiple mRNA transcripts which are transcribed into functionally diverse proteins. According to the current studies, aberrant transcripts due to splicing mutations are known to cause for 15% of genetic diseases. Therefore understanding regulatory mechanism of alternative splicing is essential for identifying potential biomarker for several types of human diseases. Most recently, advent of GeneChip® Human Exon 1.0 ST Array enables us to measure genome-wide expression profiles of over one million exons. With this new microarray platform, analysis of functional gene expressions could be extended to detect not only differentially expressed genes, but also a set of specific-splicing events that are differentially observed between one or more experimental conditions, e.g. tumor or normal control cells. In this study, we address the statistical problems to identify differentially observed splicing variations from exon expression profiles. The proposed method is organized according to the following process: (1) Data preprocessing for removing systematic biases from the probe intensities. (2) Whole transcript analysis with the analysis of variance (ANOVA) to identify a set of loci that cause the alternative splicing-related to a certain disease. We test the proposed statistical approach on exon expression profiles of colorectal carcinoma. The applicability is verified and discussed in relation to the existing biological knowledge. This paper intends to highlight the potential role of statistical analysis of all exon microarray data. Our work is an important first step toward development of more advanced statistical technology. Supplementary information and materials are available from `http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`.

Keywords: human exon microarray data, alternative splicing, analysis of variance, colon cancer, Wnt signaling pathway

1 Introduction

Alternative splicing is known as an important regulatory mechanism in which multiple mRNA transcripts are generated from a single gene. The current studies suggested that $30 - 60\%$ of eukaryotic genes have multiple splice variants [6, 10]. After a gene is transcribed into pre-mRNA, the introns are removed from the transcript isoform and the exons are joined by the spliceosome. During this reaction, a set of exons may be included in one version of mRNA, and skipped in another. The presence of

transcriptional isoforms can alter the protein products which may be associated with the manifestation of cancer and other genetic diseases. For instance, aberrant transcripts due to splicing mutations are known causes for 15% of genetic disease [15]. Therefore, understanding regulatory mechanism of alternative splicing is one challenging task to identify potential biomarkers for several types of human diseases.

Recent advent of human exon microarray technology, e.g. GeneChip® Human Exon 1.0 ST Array, enables us to collect genome-wide expression profiles of over one million human exons. By such a technological innovation, analysis of functional gene regulations could be extended to detect not only changes in the overall gene expression values, but also changes in the splicing variations across different cell classes. In our opinion, the whole transcript analysis with exon expression profiles will plays a key role to clarify mechanism of functional splicing regulations. At the same time, a rapid development of statistical technologies to analyze all exon microarray data must become one of the most significant tasks to be addressed in bioinformatics.

In this study, we target to identify mis-spliced gene loci linkage to a certain disease with the GeneChip® Human Exon 1.0 ST Array. This is an essential step toward understanding the functional regulation of alternative splicing and also the the genome-wide discovery of the potential biomarkers at the exon level. The proposed method is organized according to the following process: (1) Data normalization; In the expression profiles collected by GeneChip exon microarrays, some systematic biases are mixed into the observed expression values. In particular, an upward bias of the probe intensities due to the GC-contents causes a critical issue in the data analysis and removing it from the downstream analysis is necessary. (2) Outlier detection; In this microarray system, each exon is spanned by a small set of probes to quantify the exon specific signal intensity. Typically, the number of probes is ranging from 4 to 20 per an exonic region. If we estimate the exon specific signal based on such a small set of probe intensities, inclusion of the abnormal probe intensities in the downstream analysis often causes indispensable false discoveries of the specific splice variations. (3) Whole transcripts analysis with the analysis of variance (ANOVA); During this process, a set of specific splice variations that are present in a particular cell but absent in the normal control are identified. At this step the method automatically decomposes response of the probe intensities to a target disease into three orthogonal effects, i.e. effect of alternative splicing shared by normal and tumor cells, difference in the overall gene expression level between the different cell types, and effect of specific splice variations.

We tested the proposed approach with application to exon expression profiles of colorectal carcinoma. The dataset was originally distributed by the Affymetrix to the third party developers of exon array data analysis technology. The method was able to detect a wide variety of tumor-specific transcriptional isoforms. Some of them are verified in relation to the existing biological knowledge, for example, some known biomarkers of colon cancer and publicly available database of alternative splicing, such as AltSplice database [3]. As a pioneer work, this paper intends to highlight the potential role of statistical analysis of all exon microarray data and indicates some promising directions toward whole genome study of alternative splicing. Our work is an important first step toward development of more advanced statistical technology. Supplementary information and materials are available from `http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`.

2 Material and Methods

2.1 Probe Design

On the GeneChip® Human Exon 1.0 ST Array, a huge amount of probes, more than 5.5 millions, is tiled to monitor the expression profiles of over one million exons. Along with this new technology, we aim to clarify gene expression program of cells at the exon level, in particular, regulatory mechanism of alternative splicing, e.g. exon skipping, intron retention, mutually exclusive exon usage, alternative

promoter usage, alternative polyadenylation, and so on.

The GeneChip® Human Exon 1.0 ST array employs a comprehensive probe design strategy and supports most exonic regions for both well-annotated human genes and abundant novel transcripts. A total of over one million exonic regions is registered in this microarray system. The probe sequences are designed based on two kinds of genomic sources, i.e. cDNA-based content which includes the human RefSeq mRNAs, GenBank and ESTs from dbEST, and the gene structure sequences which are predicted by GENSCAN, TWINSCAN, Ensemble and so on. The majority of the probe sets are composed of four perfect match (PM) probes of length 25 bp, whereas the number of probes for about 10 percent of the exon probe sets is limited in less than four due to the length of probe selection region and sequence constraints. With this microarray platform, no mismatch (MM) probes are available to perform data normalization, for example, background correction of the monitored probe intensities. In stead of the MM probes, we can infer and remove the existing systematic biases based on the observed intensities of the background probe probes (BGP) which are pre-designed by the Affymetrix. The BGPs are composed of the genomic and the antigenomic ones. The genomic BGPs were selected from a research prototype human exon array design based on NCBI build 31. The antigenomic background probe sequences are derived based on reference sequences that are not found in the human (NCBI build 34), mouse (NCBI build 32), or rat (HGSC build 3.1) genomes. For more details about the BGP design, see the Affymetrix's data sheet `http://www.affymetrix.com/support/technical/datasheets/exon_arraydesign_datasheet.pdf`.

In the GeneChip® Human Exon 1.0 ST array, the transcript cluster ID is assigned to each locus on entire human genome. Furthermore, in order to quantify exon specific signal intensity, each exon is spanned by a number of probe set where each probe set contains about four probes. After a hybridization of the target cells, the expression value of each exon is predicted based on the corresponding probe intensities. According to the observed pattern of the probe intensities in a particular locus, one can predict the splicing isoform (see Figure 1).
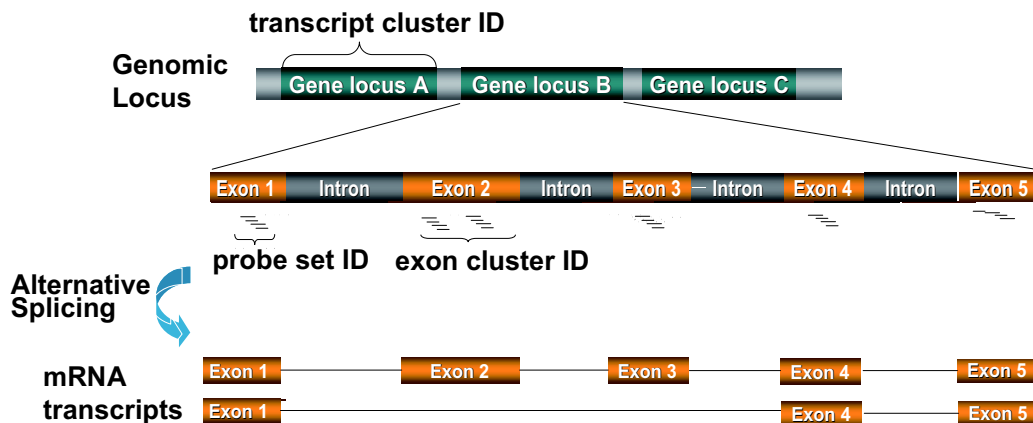

Figure 1: GeneChip® Human Exon 1.0 ST microarray system. In this microarray, a locus ID which is referred to as the transcript cluster ID is assigned to each locus on the human genome. To estimate the exon specific-expression values, each exon is spanned by a number of probe set where each set typically contains four probes. The exon cluster ID and the probe set ID are also assigned to each exonic region and the probe set. After in situ hybridization of the target cells, the expression value of each exon is estimated based on the observed probe intensities. According to the observed pattern of the probe intensities in a particular locus, one can predict what splice variations occurs in the cells.

2.2 Tissue Samples

We discuss the potential applicability of the GeneChip® Human Exon ST 1.0 Array along with analysis of early access single-stranded WTA colon cancer dataset which is now publicly available from `http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx`. This dataset is originally generated for third party software developers. Total RNA obtained from colon cancer tumors is compared against their corresponding adjacent normal tissue. There are 10 colon cancer tumor/normal pairs that were isolated from 10 different individuals. Therefore, a total of 20 samples is prepared with no technical replicates. For more details of the sample information, see Supplementary Table S1 of `http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`.

2.3 Data preprocessing

One of the most critical issues in design of the statistical analysis is the way GC content-specific biases are handled. What is shown in Supplementary Figure S1 (`http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`) is boxplots of the logged-probe intensities of 3-2T.CEL (tumor cells of patient No 3) across the GC bins which are ranging from 3 to 25. From these plots, we observed the following facts: (1) The median intensities increase exponentially as the corresponding GC bins tend to be large. (2) The range of the two quartile points become large as the GC contents increase. Such a upward bias due to the high GC contents is caused by the higher affinity of GC-rich probes than AT-rich probes and should be removed from the downstream data analysis to avoid crucial false discoveries.

The Affymetrix's white paper [1] suggests the background estimation by using median of the logarithmic intensities of BGPs with same GC-content, i.e. $\log(\text{Raw Intensity}) - \log(\text{BGP})$, based on the controlled experiments. Following this guide, we also applied a background correction method that uses the antigenomic BGPs to estimate the GC-content-dependent background intensities. The left and the right panel in Supplementary Figure S2 (`http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`) shows the upward trend of the antigenomic BGP intensities across the GC-contents and the background corrected intensities, respectively. For the GC-rich probes, for example, more than 20 bin, the upward bias of the background corrected probe intensities still remained as shown in the right panel in Supplementary Figure S2 (`http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`). Notice that for the median values of the antigenomic BGPs, the upward trend is not clear for the GC-rich probes, e.g. across 23 to more (the left panel in Supplementary Figure S2). This implies that ability of the antigenomic probe with high GC contents is less accurate in the estimation of GC-content-dependent bias and also the background intensities. Such an inaccurate estimation is possibly associated with either small sample size of antigenomic BGPs for the high GC bins and the extremely high affinity of GC rich probes which may lead to a large amount of the cross-hybridizations. For example, the number of antigenomic BGPs with GC-content 24 is 268 which is equal to only 27% of that with GC-content 14.

Such an unreliable GC-rich probe intensities largely affects the downstream data analysis and sometimes cause a large number of false discoveries, particularly, when we analyze a locus which is composed of a small number of exonic regions. That is, the estimated exon expression value is usually biased by presence of outlier probe intensities due to inclusion of such an unreliable probes because in this microarray the number of probes which interrogate an exonic region is fairly small. Therefore, in this study, to achieve a conservative scheme to discovery of specific-splicing events, we decided to exclude the GC-rich probes more than 21 from the following whole transcript analysis.
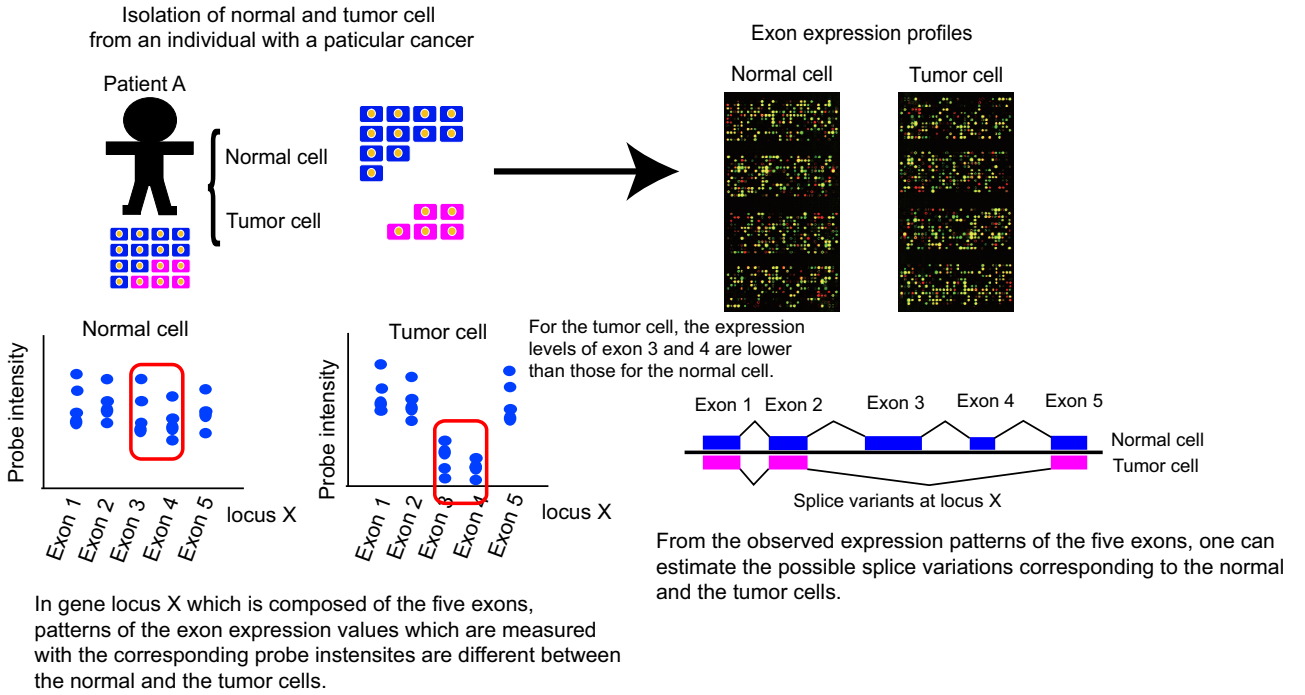
Figure 2: Schematic for the identification of the tumor-specific biomarker splice variations from exon expression profiles of normal and tumor cells which are isolated from one individual. With the observed intensities of the tiled probes, one can estimate the expression values of whole human exons. Based on the expression patterns of exons in a particular locus, we predict the transcriptional isoforms which are generated during the alternative splicing. Problem of detecting the specific-splicing events amounts to finding loci that exhibit differentially expressed exons between normal and tumor cells.

2.4 Whole transcript analysis with analysis of variance

Suppose that a pair of normal and tumor cells which are isolated from an individual have been profiled by the two microarray experiments. Here, a target gene locus is assumed to be composed of m exons. Let x_{ijk} be a background corrected probe intensity which corresponds to the ith exon ($i = 1, \cdots, m$) and the kth probes ($k = 1, \cdots, n_i$) for the normal $j = 1$ and tumor cell $j = 2$, respectively. Then, the task to be addressed is to deduce what splicing process functions in a locus from the observed probe intensities x_{ijk}, $i = 1, \cdots, m$, $j = 1, 2$ and $k = 1, \cdots, n_i$.

With regard to detection of spliced gene loci specific to tumor cell, the problem amount to finding a set of loci that exhibit the differentially expressed exonic regions across the cell types (Figure 2). To this end we apply the classical ANOVA method which is based on a simple fixed effect model as follows:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

The μ represents an overall mean of the probe intensities common the entire probes. The parameters α_i, $i = 1, \cdots, m$ respond to changes in the baseline intensities for the m exonic regions. The estimated α_is are expected to take one or more different values respond to presence of splice variations shared between the both tissue type. The parameters β_j, $j = 1, 2$, correspond to differences in the overall means between normal and tumor cells. The estimated β_js take different values respond to changes in the gene expression level across the cell types. The parameter γ_{ij}s represent interaction effect for

each combination of the m exons and two cell categories. It is likely that at least one values in the estimated γ_{ij}s is different with each other if the alternative splicing is present in a particular cell, but absent in the other. The estimation of the interaction parameters captures the effect of tumor-specific alternative splicing. The mechanism of the mixed linear model is summarized in Figure 3.

Estimation of α_i, β_j and γ_{ij} amounts to decomposing the probe responses into three orthogonal effects, i.e. exon effect, overall gene effect and effect of specific-splicing event. To assess the significance of each effect with the observed intensities, we proceed to the following statistical testing:

Test 1 (Exon effect);
$$H_0 : \alpha_i = \alpha_j \text{ for any } i \neq j$$
$$H_1 : \alpha_i \neq \alpha_j \text{ for at least one pair of } \{i \neq j\}$$
Test 2 (Overall gene effect);
$$H_0 : \beta_1 = \beta_2$$
$$H_1 : \beta_1 \neq \beta_2$$
Test 3 (Effect of Specific alternative splicing);
$$H_0 : \gamma_{ij} = \gamma_{hk} \text{ for any } \{i, j\} \neq \{h, k\}$$
$$H_1 : \gamma_{ij} \neq \gamma_{hk} \text{ for at least one pair of } \{i, j\} \text{ and } \{h, k\}$$

Rejection of the null hypothesis that all interaction terms are equal suggests that any splicing events associated with cancer manifestation are likely to occur during the regulation process at the locus. After repeating the ANOVA for the entire loci which are featured by the human exon microarray, one can assess significance for presence of specific-splicing event for the over 300,000 transcripts clusters and then assign the p-values of the null hypothesis in Test 3 for each locus.
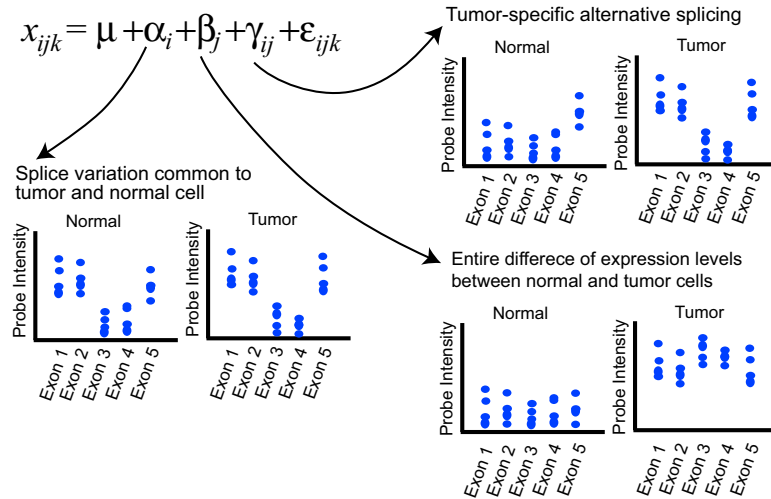

Figure 3: Mechanism of the mixed linear model. The model parameters α_i, β_j and γ_{ij} respond to alternative splicing common to normal and tumor cells, overall difference in the gene expression level and specific-splicing variations. For example, at least one parameter in α_i, $i = 1, \cdots, m$ or in β_j, $j = 1, 2$ takes a value different from the others, respond to the alternative splicing shared between both normal and tumor or overall difference in the gene expression levels, respectively. In addition, at least one γ_{ij}, $i = 1, \cdots, m$ and $j = 1, 2$ is different with the others if a splice variation is present in a cell, but absent in the other.

3 Results

3.1 Detection of Colon Cancer-Specific Splice Variations

We first demonstrate the potential of the proposed method with the application to exon expression profiles of an individual (patient No2, i.e. 3-2T.CEL and 4-2N.CEL, in Supplementary Table S1). The results of ANOVA, which include the computed p-values for the significance of alternative splicing shared between normal and tumor cells, entire difference of gene expression levels and tumor-specific splice variations, are available from the supplementary website `http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`. The overall probes (more than five millions) tiled on the GeneChip® Human Exon 1.0 ST Array are used in this analysis. Among total 300,000 transcript clusters, the 3016 and 1470 gene loci were identified as to generate the tumor-specific splice variants by choosing 5% or 1% significant level, respectively. Note that because we repeatedly use F-test for the transcript clusters, the method involves multiple comparisons, and the p-values should be interpreted accordingly. In order to correct the biased false positive rates, instead we possibly choose a somewhat more stringent p-value threshold. Hence, as a guide, we compute the q-value [12] that is an estimate of the false discovery rate in the identified significant loci. The technical details and the computed q-values are available from the supplementary information site `http://bonsai.ims.u-tokyo.ac.jp/~yoshidar/IBSB2006_ExonArray.htm`.

Table 1 shows a part of the significant genes which attained the smallest 20 p-values for the interaction parameters γ_{ij}. The most relevant gene was LAMA3 which codes homo sapiens lamini alpha 3 (RefSeq ID: NM198129). Laminins are basement membrane components thought to mediate the attachment, migration and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components. The protein encoded by this gene is the alpha-3 chain of laminin 5, which is a complex glycoprotein composed of three subunits (alpha, beta, and gamma). Laminin 5 is thought to be involved in cell adhesion, signal transduction and differentiation of keratinocytes.

The alternatively spliced transcriptional variants encoding the multiple isoforms have been identified. For example, in AltSplice database [3] (European Bioinformatics Institute), the five splice variants of LAMA3 are registered (AltSplice-Human: Entry ENSG00000053747). These splice variants, i.e. sp1, sp2, sp3, sp4 and sp5, are displayed in the left panel of Figure 4 with the observed patterns of the probe intensities over the normal and tumor cells in the right panel. The patterns of the observed probe intensities suggest that a splice variation is specific to tumor cell, particularly, in tumor cell, the exonic regions around in chr18:19705029-19786890(+) are highly expressed than those in normal control cell. In the left panel of Figure 4, we also show a prediction of the splicing form specific to the tumor cell where the significantly expressed exonic regions are identified by repeatedly applying the t-test with the null hypothesis that average value of the probe intensities in an exonic region is identical to zero. This observation implies that the observed expression pattern of LAMA3 in the tumor cell approximately corresponds to the known transcriptional variant sp3 or sp4.

We next focused on LGR5 (GPR49) which codes leucine-rich repeat-containing G protein-coupled receptor 5. Figure 5 displays three splice variants registered in the AltSplice database and the observed probe intensities which predict the splice variation specific to the tumor cell and the form of the splice variant as shown in the left panel. LGR5 is a member of the glycoprotein hormone receptor subfamily, which includes the thyroid-stimulating hormone receptor (TSHR), follicle-stimulating hormone receptor (FSHR), and luteinizing hormone receptor (LHR). According to [16], overexpression of LGR5 was frequently observed in HCC (hepatocellular carcinoma) with mutation in β-catenin exon 3 (14 of 16 cases, 87.5%). Moreover, introduction of mutant β-catenin into mouse hepatocytes in culture caused up-regulation of the LGR5 mouse homologue. Due the observed facts, they concluded that LGR5 is a target gene activated by Wnt-signaling. Wnt signaling is known to trigger the destabilization of free cytoplasmic β-catenin. In addition, the β-catenin is involved in both cadherin-mediated cell-cell adhesion. Recently it has been established that aberrant activation of β-catenin contributes to the onset

of a variety of tumors, particularly, colorectal carcinoma [5, 11]. In sequel we discuss the identified genes with the significant specific splice variations in relation to the Wnt signaling pathways.

Gene ID	Gene symbol	*p*-value (*q*-value)	Description
3781794	LAMA3	$2.10 \times 10^{-49}(3.24 \times 10^{-45})$	Homo sapiens laminin, alpha 3 (LAMA3), transcript variant 1 and 2, mRNA
3584443	SNRPNH	$5.38 \times 10^{-34}(4.99 \times 10^{-30})$	Homo sapiens small nuclear ribonucleoprotein polypeptide N (SNRPN), transcript variant 2, mRNA
3422144	LGR5	$3.75 \times 10^{-29}(2.32 \times 10^{-25})$	Homo sapiens leucine-rich repeat-containing G protein-coupled receptor 5 (LGR5), mRNA
3988165	SLC6A14	$1.15 \times 10^{-27}(5.32 \times 10^{-24})$	Homo sapiens solute carrier family 6 (amino acid transporter), member 14 (SLC6A14), mRNA.
2611848	SLC6A6	$4.04 \times 10^{-27}(1.50 \times 10^{-23})$	Homo sapiens solute carrier family 6 (neurotransmitter transporter, taurine), member 6 (SLC6A6), mRNA
3514879	THSD1	$6.34 \times 10^{-27}(1.96 \times 10^{-23})$	Homo sapiens thrombospondin, type I, domain containing 1 (THSD1), transcript variant 2, mRNA.
3786868	SLC14A1	$2.41 \times 10^{-26}(6.40 \times 10^{-23})$	Homo sapiens solute carrier family 14 (urea transporter), member 1 (Kidd blood group) (SLC14A1), mRNA
3494137	LMO7	$1.28 \times 10^{-25}(2.97 \times 10^{-22})$	Homo sapiens LIM domain 7 (LMO7), mRNA.
2985781	THBS2	$2.29 \times 10^{-23}(4.73 \times 10^{-20})$	Homo sapiens thrombospondin 2 (THBS2), mRNA
3604147	KIAA1199	$3.87 \times 10^{-21}(7.18 \times 10^{-18})$	Homo sapiens KIAA1199 (KIAA1199), mRNA
3911217	TMEPAI	$1.06 \times 10^{-20}(1.78 \times 10^{-17})$	Homo sapiens transmembrane, prostate androgen induced RNA (TMEPAI), transcript variant 1,2,3,4, mRNA
3299585	LIPA	$3.35 \times 10^{-20}(5.19 \times 10^{-17})$	Homo sapiens lipase A, lysosomal acid, cholesterol esterase (Wolman disease) (LIPA), mRNA
3508330	HSPH1	$1.69 \times 10^{-19}(2.41 \times 10^{-16})$	Homo sapiens heat shock 105kDa/110kDa protein 1 (HSPH1), mRNA
2809245	ITGA2	$2.21 \times 10^{-19}(2.94 \times 10^{-16})$	Homo sapiens integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor) (ITGA2), mRNA
3581637		$1.67 \times 10^{-18}(2.07 \times 10^{-15})$	
2818517	CSPG2	$9.20 \times 10^{-17}(1.07 \times 10^{-13})$	Homo sapiens chondroitin sulfate proteoglycan 2 (versican) (CSPG2), mRNA
2382781	SPR9,EPHX1	$5.06 \times 10^{-14}(5.53 \times 10^{-11})$	Homo sapiens signal recognition particle 9kDa (SRP9), mRNA.
			Homo sapiens epoxide hydrolase 1, microsomal (xenobiotic) (EPHX1), mRNA.
2409104	SLC2A1	$5.09 \times 10^{-14}(5.25 \times 10^{-11})$	Homo sapiens solute carrier family 2 (facilitated glucose transporter), member 1 (SLC2A1), mRNA
3462816	PHLDA1	$5.86 \times 10^{-14}(5.72 \times 10^{-11})$	Homo sapiens pleckstrin homology-like domain, family A, member 1 (PHLDA1), mRNA.
3867629		$2.62 \times 10^{-13}(2.43 \times 10^{-10})$	cdna:Genscan chromosome:NCBI35:19:54242724:54253836:-1
3479935	FAM58A	$3.04 \times 10^{-13}(2.69 \times 10^{-10})$	Homo sapiens family with sequence similarity 58, member A (FAM58A), mRNA.

Table 1: Identified genes with statistical evidence of the specific splice variations. The 20 genes listed here were the most significant genes in which the *p*-values and the *q*-values are denoted in the third column. The first column denotes the transcript cluster IDs to which the GeneChip Human 1.0 ST array specifies. The gene descriptions correspond to RefSeq annotations.
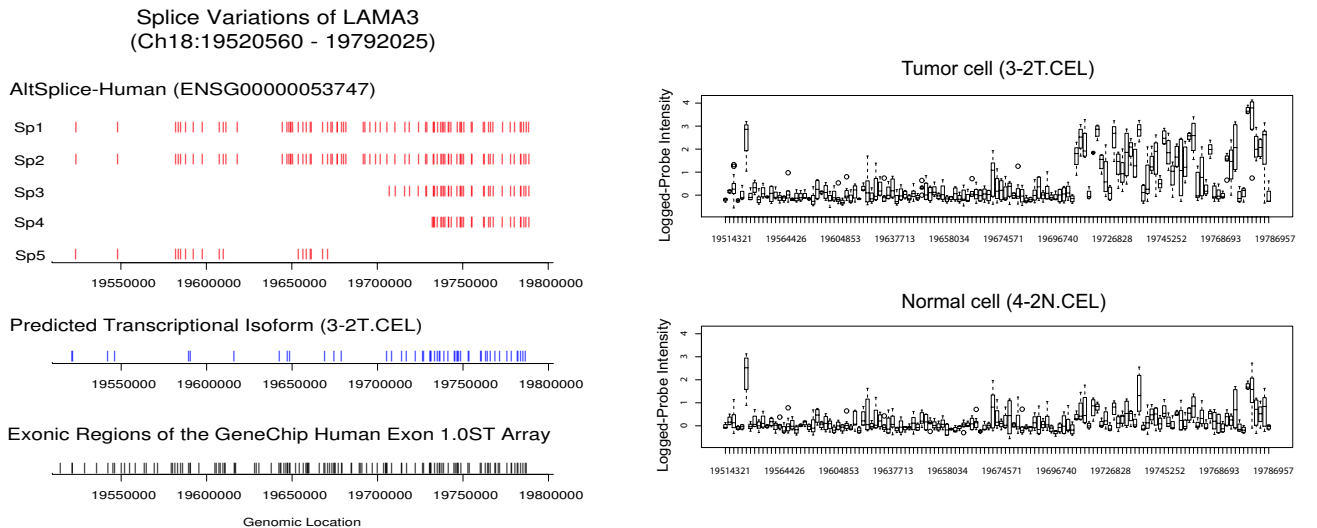

Figure 4: The splice variations of LAMA3 (left) and patterns of the observed probe intensities across normal and tumor cells (right). In the left panel, the five splice variants registered at the current AltSplice database, i.e. sp1, sp2, sp3, sp4 and sp5, are displayed. In addition to these, the predicted splicing form from the observed probe intensities and the exonic regions which are designed by this exon array system are also shown in the left panel.
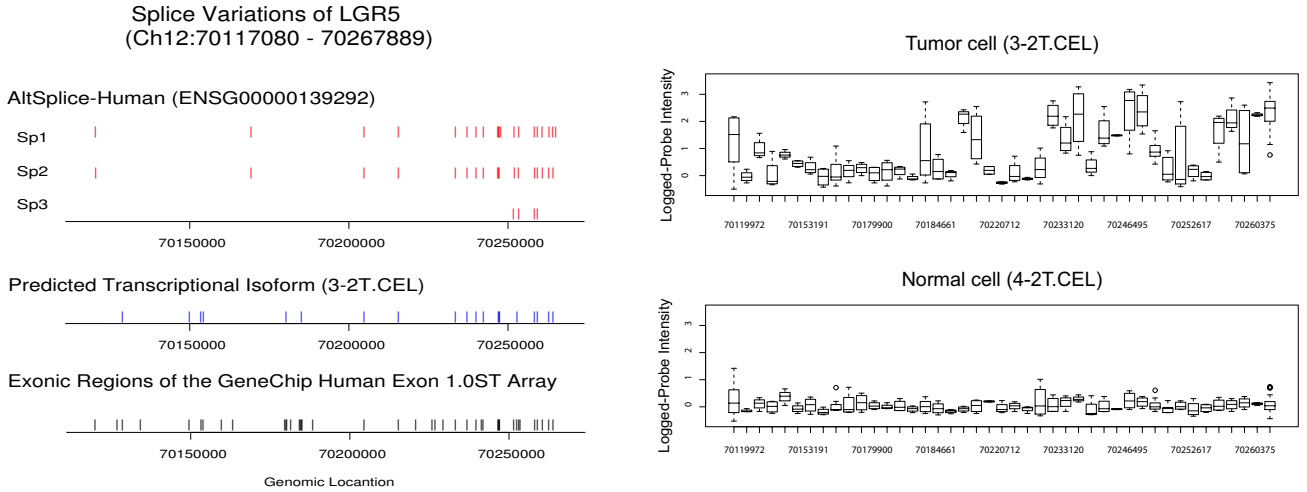
Figure 5: The splice variations of LGR5 (left) and patterns of the observed probe intensities across normal and tumor cells (right). In the left panel, the three splice variants registered at the current AltSplice database, i.e. sp1, sp2 and sp3, are displayed. In addition to these, the predicted splicing form from the observed probe intensities and the exonic regions which are designed by this exon array system are also shown in the left panel.

3.2 Splice Variations in Wnt Signaling Pathways

There are several excellent reviews on Wnt signaling and colon cancer [5, 11]. Adenomatous polyposis coli (APC), glycogen synthase kinase GSK-3β, Axin, and the transcriptional cofactor β-catenin play a central role in this pathway. β-catenin is known to code cadherin-associated protein, beta1 (CTNNB1) and to be stabilized and translocates to the nucleus to bind members of the T cell factor (Tcf)/lymphoidenhancing factor (LEF) family of transcription factors and induce target gene expression. The downstream targets of the canonical Wnt signaling pathway comprise several genes important for cellular proliferation underscoring the importance of Wnt signaling in the development of cancers, e.g. c-myc, c-Jun, c-Fos, CLDN1, cyclin D1 (CCND1), MMP3, and so on. Figure 6 shows summary of the Wnt signaling pathways. The identified genes with evidence of the specific splice variations are depicted by red. For example. the TRANSPATH [4] suggests that Wnt signaling target genes, c-Jun and c-Fos, are known to regulate MET which had a specific splice variation with the significant p-value 1.34×10^{-6} and codes met proto-oncogene (hepatocyte growth factor receptor). Either significant genes CLDN1 and the CDH11 (cadherin) encode the transmembrane-spanning proteins to produce cell adhesion molecules. In this pathway, much more significant splice variations were observed, for example, c-myc MMP3, MMP12, CDCA7, MAT2A, ETS2. Moreover, at the gene level, APC (adenomatosis polyposis coli) and β-catenin were judged to be differentially expressed genes between normal and tumor cell with the p-values of the overall gene effect β_j, 3.03×10^{-46} and 5.72×10^{-20} ,respectively. In the Wnt pathways, the APC protein normally binds to β-catenin in the cytoplasm. This biding leads to a rapid degradation of free β-catenin. On the other hand, inactivation of the APC gene triggers a reduced degradation of β-catenin. This results in an aberrant accumulation of beta-catenin in nucleus and the accumulated beta-catenin binds to a transcription factor TCF/LEF acting Wnt target genes. In this experiments, we observed that the expression of APC was present in the normal, but absent in the tumor. On the contrary, β-catenin was highly expressed specifically in the tumor cell. This observations are consistent to the above mentioned gene regulatory mechanism.
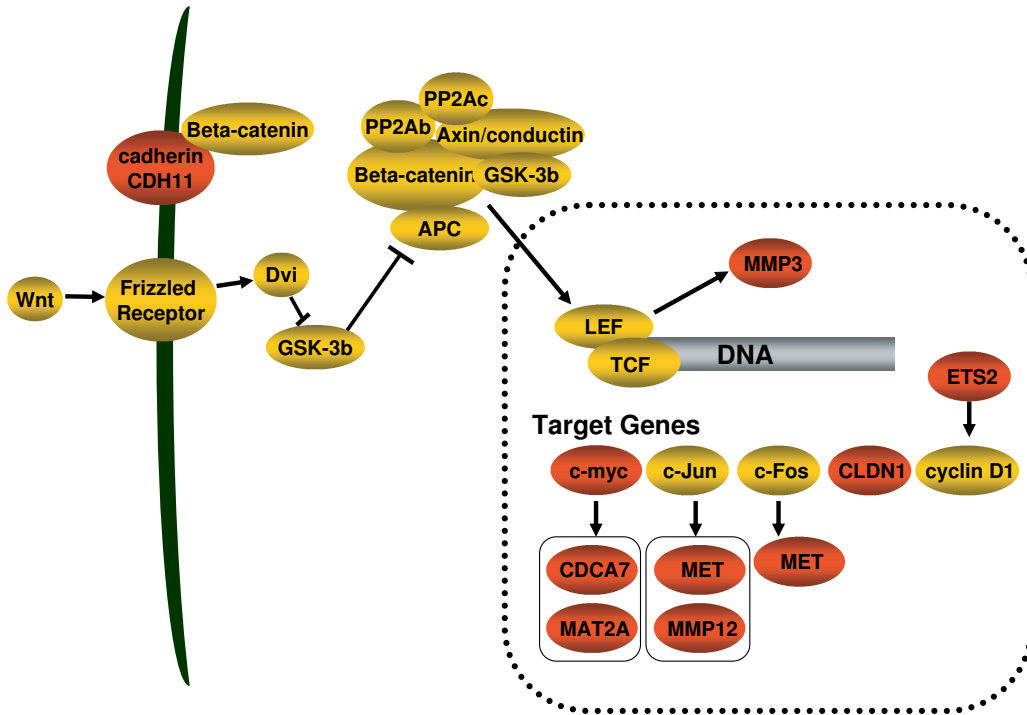
Figure 6: Summary of specific splice variations in the Wnt signaling pathways. The downstream targets of the canonical Wnt signaling pathway comprise several genes important for cellular proliferation underscoring the importance of Wnt signaling in the development of cancers, e.g. c-myc, c-Jun, c-Fos, CLDN1, cyclin D1 (CCND1), MMP3, and so on. The identified genes with evidence of the specific splice variations are depicted by red.

3.3 Discussion

Advent of GeneChip Human Exon ST Array will open up a way to whole genome analysis of functional regulation of alternative splicing. We discussed the potential of this microarray platform along with the statistical analysis to discover splicing variations which possibly cause or are caused by colon cancer manifestation. To our knowledge, this work is first as whole-genome statistical analysis of exon expression profiles with this new data production. The ANOVA method automatically identified more than 3000 loci with evidence of the tumor related splice variations. We then elucidated link between some identified splice variations and the existing biological knowledge. Some observed splicing patterns captured by the proposed method were highly consistent to the transcriptional isoforms which are registered in the AltSplice database. Moreover we also performed a pathway level analysis in order to relate the splicing variations to the aberrant accumulation of free β-catenin in the Wnt signaling pathways which are known to affect the manifestation of colorectal carcinoma. These results are sufficient to highlight the potential power of the statistical analysis of exon expression profiles.

In order to analyze the custom-made exon microarray data, a variety of computational approaches has been proposed with the successful applications. For example, Wang *et al.* [14] presented a structural modeling to quantify the relative abundance of known splice variants with the custom designed Affymetrix exon microarrays for 21 well-documented genes. Unfortunately, applicability of this approach is highly limited because the users are required to pre-specify forms of the transcriptional isoforms. Cline *et al.* [6] recently used classical analysis of variance (ANOVA) to identify tissue-specific splice variants from the observed exon expression profiles of mouse genes. Use of ANOVA is a promising approach to detect specific splice patterns that are differentially observed between one or more exon expression profiles. Indeed, most existing studies which intended to analyze exon expression profiles have designed their data analysis methods based on the ANOVA, for example, [2, 15]. In this

paper, following this direction, we also exploit the statistical method within the framework of ANOVA. However, it should be stressed that, in analysis of human diseases, it is necessity to handle issue of individual specificity, e.g. sex, age, SNPs and so on, which have some impact on regulatory mechanism of alternative splicing. For example, the current studies of oncogenesis reported that colonic tumors have a female predominance, indeed, some genes are known to significantly over-express in specimens from male compared with female colon cancer patients [8]. Moreover, another non-specific factors, e.g. age and degree of tumor differentiation is likely to influence splicing variation in any way.

Our ultimate goal is to discover "universal biomarkers", i.e. mis-spliced gene loci common to all individuals with a particular disease. One intuitive direction toward the discovery of universal biomarker is to identify loci to which the sufficiently small p-values are assigned across the entire patients. As an illustration, we applied the method of ANOVA to LGR5 (TCID:3422144) and TDGF1 (TCID:2620937) for each of the 10 colon cancer patients. Then, the computed p-values for the specific splicing variations of the 10 individuals were given by $(2.904 \times 10^{-2}, 3.751 \times 10^{-29}, 3.044 \times 10^{-7}, 1.161 \times 10^{-3}, 0.9957, 4.095 \times 10^{-22}, 1.0242 \times -23, 1.344 \times 10-17, 1.874 \times -11, 0.999)$ and $(0.9873, 0.9326, 5.419 \times 10^{-4}, 1.656 \times 10^{-5}, 0.6622, 8.0286 \times 10^{-6}, 0.9836, 2.2921 \times 10^{-3}, 3.5919 \times 10^{-3}, 0.84572)$ for LGR5 and TDGH1, respectively. Whereas for the LGR5 the assigned p-values were small across the most individuals, those of the TDGH1 vary over the individuals, particularly, more than 50% scores are assigned to the patient no1, no2, no5, no7 and no10, respectively. Indeed, as shown in Supplementary Figure S3 and S4, the observed splice patterns are almost same over the entire experiments for LGR5, but differ from each other for TDGH. The later pattern of probe intensity is possibly affected by the non-specific splicing factors. We conclude that removal of such non-specific splicing factors is a key for success in selecting a set of biomarkers for several types of human diseases. One intuitive direction to solve such a problem is to identify a set of genes which have small p-values across all of the collected patients. Following this, we will exploit a statistical technique in the future work.

References

[1] Affymetrix's white paper: Exon array background correction v1.0, `http://www.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf`

[2] Affymetrix's white paper: Alternative transcript analysis methods for exon arrays v1.1, `http://www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf`

[3] AltSplice, `http://www.ebi.ac.uk/asd/`

[4] TRANSPATH, `http://www.transpath.com/`

[5] Bienz M, Clevers H: Linking colorectal cancer to Wnt signaling. Cell 2000, 103:311-320

[6] M.S. Cline, J. Blume, S. Cawley, T.A. Clark, J.-S. Hu, G. Lu, N. Salomonis, H. Wang, and A. Williams, ANOSVA: A statistical method for detecting splice variation from expression data, *Bioinformatics*, **21**:i107–i115, 2005.

[7] Huang J, Papadopoulos N, McKinley AJ, Farrington SM, Curtis LJ, Wyllie AH, Zheng S, Willson JK, Markowitz SD, Morin P, Kinzler KW, Vogelstein B, Dunlop MG: APC mutations in colorectal tumors with mismatch repair deficiency. Proc Natl Acad Sci USA 1996, 93:9049-9054

[8] Jean-Pierre J. Issa, Nita Ahuja, Minoru Toyota, Mary P. Bronner and Teresa A. Brentnall, Accelerated Age-related CpG Island Methylation in Ulcerative Colitis *Cancer Research*, 2001, 61, 3573–3577

[9] Korinek V, Barker N, Morin PJ, van Wichen D, de Weger R, Kinzler KW, Vogelstein B, Clevers H: Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. Science 1997, 275:1784-1787

[10] Keith Le, Katherine Mitsouras, Meenakshi Roy, Qi Wang, Qiang Xu, Stanley F. Nelson and Christopher Lee, Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data, *Nucleic Acids Research*, 2004, Vol. 32, No. 22, e180.

[11] Polakis P: Wnt signaling and cancer. Genes Dev 2000, 14:1837- 1851

[12] Storey JD. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64: 479-498.

[13] T. A. Thanaraj, Stefan Stamm, Francis Clark, Jean-Jack Riethoven, Vincent Le Texier and Juha Muilu ASD: the Alternative Splicing Database, *Nucleic Acids Research*, 2004, Vol. 32, Database issue, D64–D69.

[14] H. Wang, E. Hubbell, J.-S. Hu, G. Mei, M. Cline, G. Lu, T. Clark, Michael A. Siani-Rose, M. Ares, D.C. Kulp, and D. Haussler, Gene structure-based splice variant deconvolution using a microarray platform, *Bioinformatics*, **19**:i315–i322, 2005.

[15] Yuanyuan Xiao, Yee H. Yang, Todd A. Burckin, Lily Shiue, Grant A. Hartzog, Mark R. Segal, Analysis of a Splice Array Experiment Elucidates Roles of Chromatin Elongation Factor Spt4-5 in Splicing, *PLOS Computational Biology*, Volume 1, Issue 4, e39, 0276–0288, 2005.

[16] Yoshiya Yamamoto, Michiie Sakamoto, Gen Fujii, Hitomi Tsuiji, Kengo Kanetaka, Masahiro Asaka, and Setsuo Hirohashi, Overexpression of Orphan G-Protein-Coupled Receptor, Gpr49, in Human Hepatocellular Carcinomas With -Catenin Mutations *HEPATOLOGY*, Vol. 37, No. 3, 2003, 528-533.