Genomics & Informatics Vol. 4(1) 33-39, March 2006

Deducing Isoform Abundance from Exon Junction Microarray

Pora Kim[1], S. June Oh[2] and Sanghyuk Lee[3]*

[1]Bioinformatics Team, IT-BT group, Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea, [2]Department of Pharmacology & Pharmacogenomics Research Center, College of Medicine, Inje University, Busan 614-735, Korea, [3]Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea

Abstract

Alternative splicing (AS) is an important mechanism of producing transcriptome diversity and microarray techniques are being used increasingly to monitor the splice variants. There exist three types of microarrays interrogating AS events-junction, exon, and tiling arrays. Junction probes have the advantage of monitoring the splice site directly. Johnson *et al.*, performed a genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays (Science 302:2141-2144, 2003), which monitored splicing at every known exon-exon junctions for more than 10,000 multi-exon human genes in 52 tissues and cell lines. Here, we describe an algorithm to deduce the relative concentration of isoforms from the junction array data. Non-negative Matrix Factorization (NMF) is applied to obtain the transcript structure inferred from the expression data. Then we choose the transcript models consistent with the ECgene model of alternative splicing which is based on mRNA and EST alignment. The probe-transcript matrix is constructed using the NMF-consistent ECgene transcripts, and the isoform abundance is deduced from the non-negative least squares (NNLS) fitting of experimental data. Our method can be easily extended to other types of microarrays with exon or junction probes.

Keywords: Alternative Splicing, Junction Microarray, Non-negative Matrix Factorization (NMF), Non-Negative Least Squares (NNLS)

Introduction

Microarrays are increasingly used to study alternative splicing recently. The major advantage is that splicing pattern of numerous exons can be monitored simultaneously (Shoemaker *et al.*, 2001). It is the most promising technique that can show the transcript variation due to alternative splicing in various tissues, developmental, and pathological conditions (Xu *et al.*, 2002; Hu *et al.*, 2002). Proper interpretation can give information on transcript structure and expression pattern of genes in an efficient way. However, there is no standard in manufacturing arrays and in interpreting the data yet.

To monitor alternative splicing events with microarray technique, one needs special probe design that can report whether specific exons or splice sites are present in the transcript or not. Arrays of this type are called 'splice arrays' and oligonucleotides are typically used as probe sequences. Splice arrays can be further classified into three sub-groups - junction, exon, and tiling array as shown in Fig. 1.

Exon probe sequences are part of a single exon, thereby manifesting expression of the corresponding exon. Multiple probes are necessary to identify alternatively spliced exons using exon probes. Its main disadvantage would be that exact splice sites cannot be obtained. On the other hand, junction probes consist of concatenation of two neighboring exons. Presence of each splice site is directly examined and any variation in the splice site is reflected in the probe intensity. However, the sequence composition is fairly limited since they cannot be far from the splice site. Hybridization efficiency might be different for each probe, which is a difficult problem to overcome in analyzing experimental result. Recent trend is to use both the junction and exon probes for reliable prediction of splice site variation (Pan *et al.*, 2004; Fehlbaum *et al.*, 2005; Ule *et al.*, 2005).

Tiling array is collection of probes reflecting the genomic map. Original purpose of the tiling array was

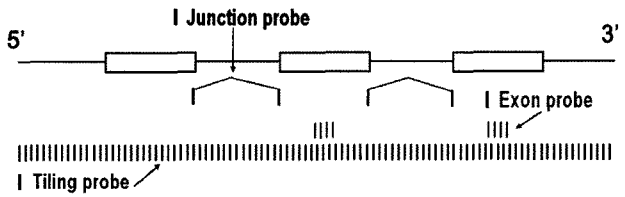

Fig. 1. Various types of splice arrays. Design principles for junction, exon and tiling probes are illustrated.

*Corresponding author: E-mail sanghyuk@ewha.ac.kr
Tel +82-23-277-2888, Fax +82-23-277-2384
Accepted 2 March 2006

interrogating whether a specific part of the genome is transcribed or not. Since it is unbiased mapping of transcription, it can be used to monitor alternative splicing. The main problems are cost and resolution. Researchers at the Affymetrix Inc. reported tiling array experiment for chromosomes 21 and 22 at 35 nt. resolution as early as in 2002 (Kampa *et al.*, 2004). Their recent update includes 10 chromosomes at 5 nt. resolution (Cheng *et al.*, 2005). Even though it cannot be applied for general purpose due to high cost, the experimental data can be a reference point for transcription with additional tissue data.

Johnson et al. performed a junction array experiment that covers ~10,000 human RefSeq genes in 52 tissues and cell lines (Johnson *et al.*, 2003). Each probe sequence consists of 36 nucleotides from two adjacent exons (last 18-mer in the preceding exon and the first 18-mer in the following exon) and a linker sequence. Every exon-exon junction within RefSeq is taken into consideration in designing the probe sequences. Resulting hybridization reflects the splice-site variation owing to AS. For example, exon skipping of the 8-th exon produces reduced probe intensity for two adjacent probes at donor and acceptor splice sites of the exon 8.

Even though Johnson *et al.*, demonstrated successfully that microarray could be used to study alternative splicing at genome-wide level, their analysis is fairly limited in several aspects. The most important drawback is that their interpretation lacks any quantitative prediction of isoform concentration. They intentionally avoided building any transcript models from the experimental data and just examined if any probe intensity is below expectation, which is a good indication of alternative splicing event.

Predicting transcript structure and abundance for each isoform simultaneously would be the most important, but truly demanding task in the presence of experimental noise. A research group at the Affymetrix Inc. and the University of California at Santa Cruz developed a method to determine the relative abundance of known splice variants through a maximum likelihood estimation framework (Wang *et al.*, 2003). Their analysis assumes that splice variants and their gene structure present in the sample are known. No attempt to detect any novel splice variants was made.

Here, we describe an algorithm to deduce the relative concentration of isoforms from the junction array data. First, Non-negative Matrix Factorization (NMF; Lee and Seung, 1999) is applied to obtain the transcript structure inferred from the expression data. Then we choose the transcript models consistent with the ECgene model of alternative splicing since the transcript models predicted from NMF are often unreliable. ECgene transcripts have mRNA or EST sequences as a supporting evidence. The probe-transcript matrix is constructed using the resulting ECgene transcripts and the isoform abundance is deduced from the non-negative least squares (NNLS) fitting of experimental data. Our method takes the non-negative nature of concentration into account explicitly, which gives dramatic difference from other decomposition-based methods. It can be easily extended to other types of microarrays with exon or junction probes (e.g. Affymetrix's exon chips).

Methods

Datasets

Johnson *et al.*,'s junction-array data (GSE740) were downloaded from the GEO website (Gene Express Omnibus, http://www.ncbi.nlm.nih.gov/geo). It contains more than 10,000 multi-exon human genes in 52 tissues and cell lines. Oligonucleotide probes were placed at every exon-exon junction in each transcript. The data covers 105,398 probes, 10,274 genes, 11,138 GenBank accession and 1,646 genomic contigs.

Every probe sequence should be mapped onto the splice junctions in the ECgene transcripts. We compared the gene symbols, GenBank accession numbers, and explicit sequences used in GSE740 with the ECgene annotations. The resulting gene symbol, GenBank accession, ECgene ID, transcript ID, exon number, and probe sequences were stored in the MySQL database. Using the latest ECgene version 1.2, we were able to map 95,945 probes out of 105,389 probes in GSE740. Some RefSeqs could not be aligned against the genome with good quality.

One subtle point is worth mentioning. Johnson *et al.*, made probes for each RefSeq. In other words, if a gene had more than two RefSeqs representing splice variants, junction probe sequences can be redundant. We searched all redundant probe sequences and took the average value in the subsequent steps.

Algorithmic details

Problem setup: The experimental data can be represented as an *expression matrix* E whose element E_{pt} is the hybridization intensity of probe p for tissue t. The expression value for each probe depends on several factors including isoform abundance, isoform structure and probe affinity. The relation can be written as

$$E_{pt} = \sum M_{pi} X_{it} \quad (1)$$

E_{pt} : probe intensity matrix of probe p for tissue type t
M_{pi} : probe-transcript matrix of probe p for transcript

model (isoform) *i*

X_{it} : transcript abundance matrix of transcript model *i* in tissue type *t*

X_{it} is the abundance (i.e. concentration) of isoform *i* in tissue *t*. The matrix **M** is the *probe-transcript matrix* whose element M_{pi} represents the presence of probe sequence *p* in isoform *i*. M_{pi} is 1 if the probe sequence is part of mRNA sequence for isoform *i*. In matrix representation, Eq. (1) can be rewritten as

$$\boldsymbol{E = MX} \qquad (2)$$

Fig. 2 shows an example of matrix setup for a case with three transcript models.

The goal is to deduce the abundance of each isoform (X) from the expression data (E). This is not an easy task since (i) we do not have the full catalog of transcripts (i.e. splice variants), (ii) all parameters in Eq. (2) are non-negative. Without the first problem, the NNLS (non-negative least squares) fitting method or the maximum likelihood optimization (Wang *et al.*, 2003) can be used to obtain the transcript abundance in various tissues. They take the non-negative nature of abundance explicitly.

Our method, named as the "isoAbundance" algorithm, combines NMF and NNLS procedures to obtain reliable transcript models and their relative concentration in each tissue. Fig. 3 shows a brief flowchart of the algorithm.

Data Pre-processing: Before any data processing, we filtered bad or non-informative probes. Probes whose mean natural log intensities across 52 tissues were less than 0.65 or larger than 11 were flagged as dark and saturated, respectively. Probes with constant intensity were also removed since they are likely due to cross-hybridization.

Selection of putative transcripts using NMF: Non-negative Matrix Factorization was applied to solve Eq. (2). Original version of NMF was to decompose a matrix $\mathbf{V} \approx \mathbf{WH}$ where **V**, **W**, **H** are non-negative matrices (Lee and Seung, 1999). This is exactly the same situation as in Eq. (2). Non-negativity constraint makes NMF distinguished from other methods such as the principal component analysis (PCA) or NMF allows a parts- based representation because only additive combinations are allowed. NMF applied to decompose the expression matrix $\mathbf{E} \approx \mathbf{MX}$, allows the extraction of hidden localized patterns such as alternative splicing of exons. Applying NMF to the expression matrix of dimension (*p*×*t*), we obtain two non-negative factors. Matrix **W** has size (*p*×*i*), which each of the *i* columns defining the presence of probes in the isoform *i*. Matrix **H** has size (*i*×*t*), representing the expression level of isoform *i* in the tissue *t*. Therefore **W**

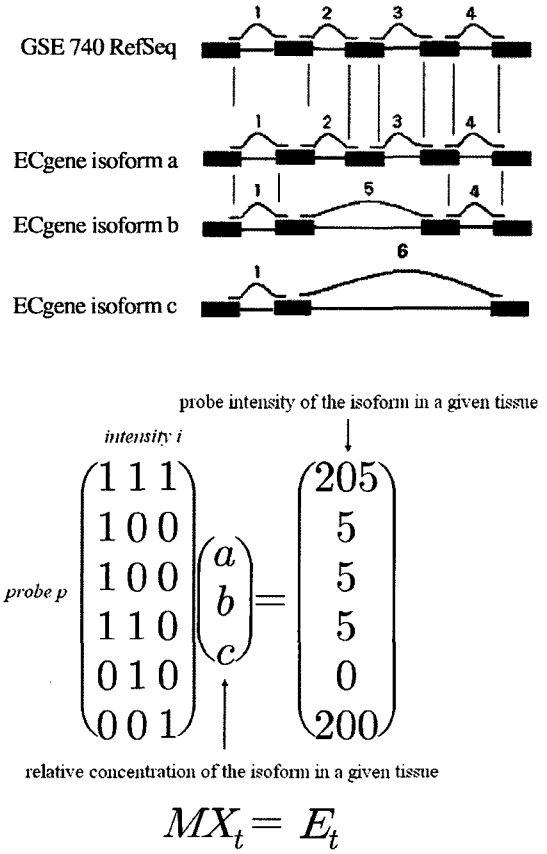

Fig. 2. Example of matrix setup for a gene with three transcript models. (a) transcript structure and probe positions (b) matrix representation. Concentration and expression values are given as vectors for this case of a specific tissue. They become multi-dimensional matrices for the experiment with various tissues. Solving this linear equation for this hypothetical case gives a = 5, b = 0, c = 200.

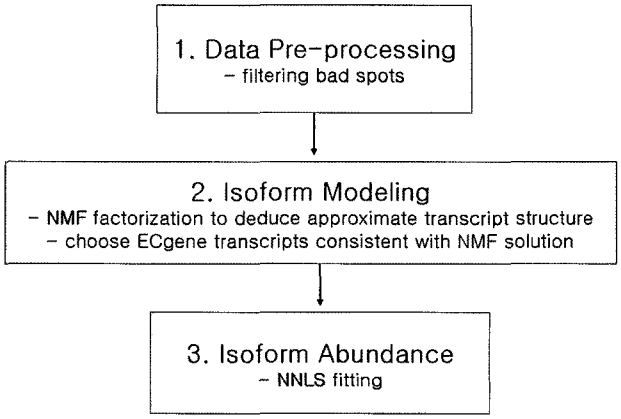

Fig. 3. Flow chart of the algorithm.

and **H** are equivalent to the probe-transcript matrix **M** and the transcript abundance matrix **X**, respectively.

However, we find that NMF method has many drawbacks to apply for the junction array data. Iterative optimization is not always robust and may fail occasionally. It is also sensitive to the experimental noise. NMF solution to Eq. (2) may not be unique - i.e. random seeding may give different combinations of **W** and **H**. Furthermore, many transcript models inferred from NMF factorization have unrealistic structure. In an effort to take these problems into consideration, we decided to keep transcripts whose exon structure is consistent with any ECgene models. Since ECgene is based on genomic alignment of mRNA and EST sequences, this ensures that each isoform model in subsequent sections has supporting evidence of clones. We simply examined the absent probes from the NMF-derived probe-transcript matrix to test the identity of NMF and ECgene models.

NNLS optimization to obtain isoform abundance: The probe-transcript matrix is constructed using the resulting ECgene transcripts and the isoform abundance is deduced from the non-negative least squares (NNLS) fitting of experimental data. NNLS method solves "**Ax=b**" in a *least squares sense*, under constraint "*vector x has non-negative elements*". **A** is equivalent to the probe-transcript matrix. Two vectors (**x** and **b**) are isoform abundance and its expression intensity in a specific tissue, essentially being a column of **X** and **E** matrices, respectively. The solution corresponds to the relative abundance of isoforms.

A subtle problem of weighting arises in NNLS fitting. Since most exons are constitutive (not showing AS event), there exist only limited number of probes with information on AS events. It is desirable to increase the weight of those informative (AS-related) probes. For simplicity, we substituted all constitutive probes with a hypothetical probe whose expression was just the average value. This allows that NNLS algorithm sensitively picks up the variation in isoform abundance.

Results and Discussion

We applied our method to three genes (OCRL1, HMGCR, and APP) whose RT-PCR and sequencing results were reported by Johnson *et al.*, For these case studies, we used the ECgene transcripts whose gene structure is the same with those in the Johnson *et al.*'s paper.

OCRL1 gene is known to have two isoforms as shown in Fig. 4A. Long isoform (NM_000276) is expressed in retina and fetal brain, whereas a short form (NM_001587) is observed in kidney that lacks a 24-nt. exon. They confirmed the expectation from junction array data with RT-PCR experiment and sequencing. Our result agrees well with the experimental result. The intersection of gene model from the NMF and the ECgene transcripts found 2 isoforms. Variant #2 and #3 correspond to NM_000276 and NM_001587, respectively. Table 1 shows the summary of our algorithm. Long isoform is expressed in all three tissues, whereas the short one is pre-dominant form in kidney and fetal brain.

Johnson *et al.*, also showed that the junction array data could predict novel isoforms for the case of HMGCR gene. Only one isoform was known at that point. Their junction array data clearly indicated alternative splicing of the 13-th exon as indicated in Fig. 4B. Their RT-PCR result for this gene showed two bands in most of the 44 tissues. Our analysis predicted that two isoforms were present in 46 tissues out of 52 tissue (data not shown). Furthermore, NMF and ECgene suggest another novel transcript with different first exon (isoform #1). The reality of this isoform should be

Case study on the APP gene is rather complicated. Johnson *et al.*, found that it had three isoforms shown in Fig. 4C. Two isoforms (NM_000484 and X06989) are present in most nonneuronal samples, e.g., melanoma and lung carcinoma. Exon 8 of NM_000484 is frequently missing in all brain tissues (X06989). Additional exon 7 of NM_000484 is skipped in fetal brain tissue (Y00264). Our analysis in Table 2 seems substantially different from their result. NMF and ECgene suggested a novel isoform which is exactly the same with NM_000484 except the first exon. This can happen if the probe intensity for the junction between the first and second

Table 1. The relative abundance of gene OCRL1 in various tissues

	Relative abundance (variant's conc./total conc.)		
variant \ tissue	Fetal brain	Kidney	Retina
#2 (NM_000276)	20.6	490.3	1583.8
#3 (NM_001587)	147.4	1031.7	0.0

Table 2. The relative abundance of isoforms in APP gene

	Relative abundance (variant's conc./total conc.)					
variant \ tissue	Brain	Brain amygdala	Cerebellum	Fetal brain	Lung carcinoma	Melanoma
#84 (New isoform)	0.0	0.0	4669.3	660.0	15489.7	8604.6
#5 (NM_000484)	118.6	384.3	0.0	0.0	0.0	0.0
#6 (X06989)	195.1	341.2	0.0	430.2	4448.2	308.2
#7 (Y00264)	56.4	0.0	9061.6	478.0	6707.2	8387.0

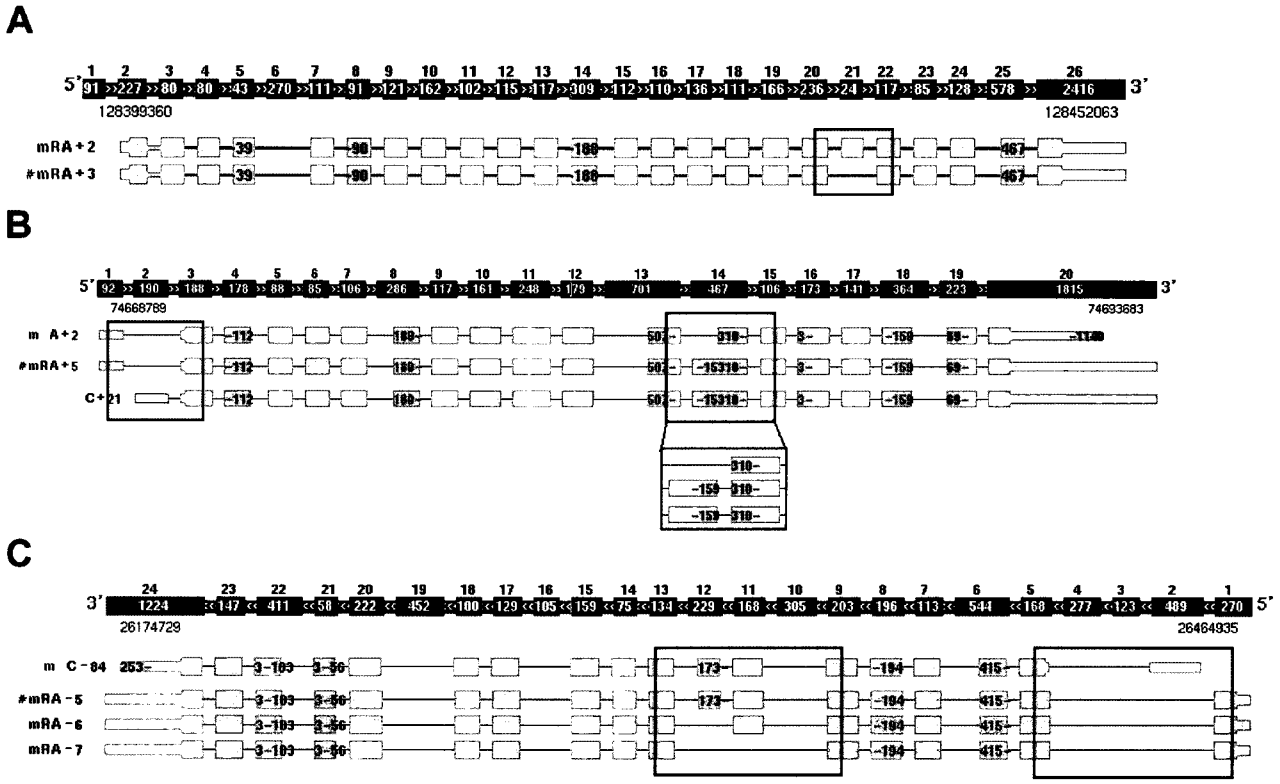

Fig. 4. Transcript structure used in the case studies. **A** OCRL1 gene. The variant #2 and #3 correspond to the RefSeq NM_000276 and NM_001587, respectively. **B** HMGCR gene. The variant #5 is the RefSeq NM_000859. The variant #2 is an unknown variant whose 13th exon (indicated as boxes) is being skipped. This exon skipping is obvious in the lower magnified view. **C** APP gene. It has three known splice variants (#5, #6, #7) whose accession numbers are NM_000484, X06989 and Y00264, respectively. The third variant (#1 of HMGCR gene) in figure B and the first variant (#84 of APP gene) in figure C are novel isoforms that both NMF and ECgene models predicted.

exons of NM_000484 is low compared to other constitutive probes. It remains to be seen if this new isoform is real or it should be regarded as equivalent to the RefSeq NM_000484. Results for melanoma and lung carcinoma do not agree well with the RT-PCR result.

The APP case clearly indicates that our algorithm requires some improvements. There are many factors to take into consideration. Different hybridization efficiency of junction probes is the most serious problem. Conventional probe sequences are designed to have similar melting temperatures. Since junction probes have limited freedom of sequence selection, their hybridization efficiency may be vastly different. Method to estimate hybridization efficiency for each probe is essential for successful interpretation of junction array data. Observed probe intensities for constitutive exons could give some idea on different hybridization efficiency of each probe. Matrix **W** from NMF calculation should reflect the hybridization efficiency too in ideal situation. One can calculate the melting temperature from the probe sequence and may be able to estimate the hybridization efficiency theoretically. One of these methods or an entirely different approach is imperative to overcome the problem of different hybridization efficiency in junction or exon array data.

Another critical step is the proper selection of isoforms. Almost all methods analyzing splice array data predict alternative splicing at the exon level. For example, they can predict a specific event of exon skipping fairly accurately. However, concurrent events at several exons should be correlated to predict gene expression at the isoform level. It would be quite challenging since both AS and different hybridization affects the probe intensities simultaneously. No method is advanced enough to deduce the transcript structure from the microarray data yet. An iterative procedure that learns the hybridization affinity and AS events alternately may be developed to solve this problem.

The problem of isoform selection can be alleviated as

the catalog of splice variants becomes more complete. RefSeq and Ensembl are the two main resources for gene structure. Databases focused on alternative splicing comprise the ASD, AceView and ECgene. Adjusting weights for constitutive and alternative exons might be helpful to emphasize the effect of AS in the NNLS fitting step.

Recently, several labs are trying to identify AS events using the so-called splice array that contains junction, exon and intron intron probes for a single splice event (Pan *et al.*, 2004; Fehlbaum *et al.*, 2005; Ule *et al.*, 2005). Notably, Shai *et al.*, developed GenASAP that predicted the levels of AS and the hybridization profiles simultaneously using Bayesian learning in an unsupervised probability model (Shai *et al.*, 2005). Even though it is limited to detect exon skipping events only, its ability to learn the hybridization affinity may be applied to augment the NMF method.

Conclusion

We developed an algorithm that predicts the isoform abundance from the junction array data. Unlike other algorithms dealing with individual AS event at the exon level, our method uses the NMF algorithm to obtain the transcript models consistent with the ECgene prediction. Then the relative isoform abundance is deduced from the NNLS fitting of expression data.

Recent development of human exon chips from the Affymetrix Inc. provides an exciting opportunity to examine transcription at the exon level on the genome-wide scale. It contains ~1.4 million probes that comprising ~4 probes/exon and ~40 probes/gene on average. Methods to analyze such type of data are urgently needed.

Our method can be easily extended to other types of microarrays with exon and junction probes. It is just a matter of defining the probe-transcript matrix according to the probe position in the transcript structure. However, it should be pointed out that the catalog of splice variants is not complete even with today's vast amount of mRNA and EST sequences in the GenBank. Methods that take the differential hybridization efficiency and identification of novel splice variants into account simultaneously should be developed in the near future.

Acknowledgements

This work was supported by the Ministry of Science and Technology of Korea through the Bioinformatics Research Program (Grant No. 2005-00201).

References

Bracco, L. and Kearsey, J. (2003). The relevance of alternative RNA splicing to pharmacogenomics. *TRENDS in Biotech.* 21, 346-352.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol.* 18, 630-634.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D.K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D.S., and Gingeras, T.R. (2005). Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science* 308, 1149-1154.

Fehlbaum, P., Guihal, C., Bracco, L., and Cochet, O. (2005). A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res.* 33, *e*47.

Hu,G.K., Madore,S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. (2001). Predicting splice variant from DNA chip expression data. *Genome Res.* 11, 1237-1245.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., and *et al.*, (2004). Integrative annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLOS Bio.* 2, 856-875.

Johnson, J.M., Castel, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141-2144.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H., and Gingeras, T.R. (2004). Novel RNAs Identified From an In-Depth Analysis of the Transcriptome of Human Chromosomes 21 and 22. *Genome Res.* 14, 331-342.

Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y., and Lee, S. (2005). ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.* 33, D75-D79.

Lawson, C.L. and Hanson, R.J. (1974). Solving Least Squares Problem, Prentice-Hall, Engelwood Cliffs, N.J.

Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788-791.

Modrek, B., Resch, A., Grasso C., and Lee, C. (2001).

Genome-wide detection of alternative splicing in expressed sequences of human gene. *Nucleic Acids Res.* 29, 2850-2859.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T., Morris, Q.D., Frey, B.J., and Blencowe, B.J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* 16, 929-941.

Shai. O., Morris. Q.D., Blencowe, B.J., and Frey, B.J. (2005). Inferring global leels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics* 22, 606-613.

Shoemaker,D.D., Schadt,E.E., Armour, C.D., He, Y.D., Garrett-Engele, P. *et al.* (2001). Experimental annotation of the human genome using microarray technology. *Nature* 409. 922-927.

Strausberg, L.R., Feingold, E.A., Klausner, R.D., and Collins, F.S. (1999). The mammalian gene collection. *Science* 286, 455-457.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J. Cline, M. Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B.R., Kane, D., Weinstein, J.N., Blume, J., and Darnell, R.B. (2005). Nova regulates brain-specific splicing to shape the synapse. *Nature Genet.* 37, 844-852.

Wang, H., Hubbell, E., Hu, J., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M.A., Ares, M., Kulp, D.C., and Haussler, D. (2003). Gene structure-based splice variant deconvolution using microarray platform. *Bioinformatics* 19, i315-i322.

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome, *Nucleic Acids Res.* 30, 3754-3766.