Studies of Exon Scrambling and Mutually Exclusive Alternative Splicing

by

Rong Kong

B.S., University of Science and Technology of China (2001)

Submitted to the Department of Biology
in Partial Fulfillment of the Requirement for the Degree of
Master of Science in Biology

at the

Massachusetts Institute of Technology

February 2005

Signature of Author ....
Department of Biology
December 31, 2004

Certified by ....
Christopher B. Burge
Associate Professor of Biology
Thesis Supervisor

Accepted by ....
Stephen P. Bell
Professor of Biology
Co-Chair, Department Committee on Graduate Students

Studies of Exon Scrambling and Mutually Exclusive Alternative Splicing

by

Rong Kong

Submitted to the Department of Biology on Dec 31, 2004
in Partial Fulfillment of the Requirement for the Degree of
Master of Science in Biology

ABSTRACT

The goals of this thesis work were to study two special alternative splicing events: exon scrambling at the RNA splicing level and mutually exclusive alternative splicing (MEAS) by computational and experimental methods.

Chapter 1 presents work on the study of exon scrambling, in which exons are spliced at canonical splice sites but joined together in an order different from that predicated by the genomic sequence. The public expressed sequence tag (EST) database was searched for transcripts containing scrambled exons. Stringent criteria were used to exclude genome annotation or assembly artifacts. This search identified 172 human ESTs representing 90 exon scrambling events, which derive from 85 different human genes. In several cases, the scrambled transcripts were validated using an RT-PCR-sequencing protocol, confirming the reproducibility of these unusual events. Exon scrambling of transcripts from the *GLI3* gene, which encodes a transcription factor involved in hedgehog signaling, was also conserved in mouse. Specific gene features, including the presence of long flanking introns were found to be associated with exon scrambling.

Chapter 2 deals with mutually exclusive alternative splicing (MEAS), in which only one of a set of two or more exons in a gene is included in the final transcript. A database with 101 human genes and 25 mouse genes containing mutually exclusive exons (MXE) has been established with GENOA annotation software. Specific sequence features were analyzed. A genome-wide search for a special "tandem MEAS" events was undertaken and 10 such human genes were identified. A fluorescence reporting system was built to study intronic *cis*-elements regulating MEAS.

Thesis Supervisor: Christopher B. Burge
Title: Associate Professor of Biology

ACKNOWLEDGEMENT

I would like to acknowledge the members of the Burge Lab, past and present, for their advice, friendship and support. In particular, I would like to thank Zefeng Wang, Gene Yeo, Dirk Holste, Vivian Tung for their personal and technical assistance.

I would especially like to thank my thesis advisor, Chris Burge, for his advice and support. I have valued his encouragement and help throughout my time in his lab.

I would also like to thank my wife, Ye Gu, for her consistent understanding and support.

TABLE OF CONTENTS

Chapter 1. Studies of Exon Scrambling

Chapter 2. Studies of Mutually Exclusive Alternative Splicing

Chapter 1

Studies of Exon Scrambling

I. Abstract

Exon scrambling is a phenomenon in which exons are spliced at canonical splice sites but joined together in an order different from that predicted by the genomic sequence. In some known cases exon scrambling appears to occur at the RNA level. Although a few examples of exon scrambling have been reported in human genes, this phenomenon has not been systematically studied. Here we undertook a computational search of the public expressed sequence tag (EST) databases for transcripts containing scrambled exons. Stringent criteria were used to exclude genome annotation or assembly artifacts. This search identified 172 human ESTs representing 90 exon scrambling events, which derive from 85 different human genes. In several cases, the scrambled transcripts were validated using an RT-PCR-sequencing protocol, confirming the reproducibility of these unusual events. Exon scrambling of transcripts from the *GLI3* gene, which encodes a transcription factor involved in hedgehog signaling, was also conserved in mouse. Quantification of scrambled *GLI3* transcripts suggested a high frequency of exon scrambling occurs in several tissues. Specific gene features, including the presence of long flanking introns were found to be associated with exon scrambling.

II. Introduction

A typical human mRNA is derived from a much longer primary transcript through the sequential joining of several exons by the nuclear pre-mRNA splicing machinery. Exon scrambling is a phenomenon in which exons are spliced at correct splice sites but joined together in an order different from that predicted by the genomic sequence, e.g., exons are in the order A, B, C, D in genome, but present in a transcript in the order C, D, A, B, or with tandem exact copies of one or more exons which are not duplicated in the genome (Figure 1A). Although the function and mechanisms of exon scrambling are still unclear, it appears to occur during splicing. Therefore, the study of this phenomenon may provide insights into the mechanisms responsible for pairing of exons by the splicing machinery.

Exon scrambling was first discovered in the human tumor suppressor *DCC* (deleted in colorectal carcinoma) gene (Nigro et al. 1991). Subsequently, other mammalian genes were also reported to undergo exon scrambling, including *c-ets-1* (Cocquerelle et al. 1992), *Sry* (Capel et al. 1993), cytochrome P450 2C24 (Zaphiropoulos 1996), putative hypertension-related gene *SA* (Frantz et al. 1999), and the *SNS* voltage-gated sodium channel gene (Akopian et al. 1999), etc. Two mechanisms were proposed (reviewed by Zaphiropoulos 1998): *trans*-splicing and circular RNA splicing. Each can explain some specific examples. However, a large majority of these examples are limited to rats or mice, with very few cases observed in human genes and only one example has been reported to be conserved between primates and rodents (Takahara et al. 2002).

The recent availability of the human genome sequence, together with large numbers of full-length cDNA and EST sequences makes it possible to reliably infer the exon-intron structures of thousands of human genes, and thereby to search for variants or errors in the process of exon joining using EST data. Using a large database of genes with known exon-intron organization, we undertook a computational search of the public expressed sequence tag (EST) databases for exon scrambling events. This search yielded 172 ESTs representing 85 human genes and 90 exon scrambling events. Several of these events were confirmed using an RT-PCR-sequencing protocol in human tissues. Exon scrambling in one of these genes, *GLI3*, was observed to be conserved between human and mouse.

III. Materials and Methods

Data and resources

Genes with known exon-intron organization were obtained by large-scale alignment of cDNAs to the assembled human genome (hg13) using the genome annotation software GENOA (http://genes.mit.edu/genoa; see Yeo et al. 2004), which uses spliced alignment of cDNAs to genomic sequences to infer exon-intron structures of genes. Approximately 5 million human expressed sequence tags (ESTs) were obtained from dbEST (NCBI repository 02202003).

Exon scrambling identification

Exons were obtained from genes in the GENOA database. For each exon, using a 50-bp tag from the 5' terminus of this exon and a 50-bp tag from the 3' terminus of all downstream exons in this gene, a set of all reversed-ordered exon junctions was created (Figure 1A). Exon junctions from directly repeated exons were also created by concatenating 50-bp tags from the 3' and 5' termini of the same exon.

Each such exon junction sequence was searched against the human dbEST database using BLAST 2.0 (Altschul et al. 1997). The possibility of an exon scrambling event was considered when an EST was detected to show significant similarity to the exon-junction sequence ($E < 1e\text{-}40$ with $\geq 90\%$ identity over at least 90 bases of the exon junction sequence) (Figure 1B,C).

A series of checks was then conducted to exclude various types of artifacts (Figure 1C). First, the potential scrambled exons were searched against the genomic sequences to exclude the possibility of unannotated exon duplication or tandem gene duplication in genomic sequences, both of which might produce transcripts similar to those produced by RNA-level exon scrambling. Then exons within the same gene were searched against each other to rule out the possibility of artifacts resulting from sequence similarities between the exons of a gene. ESTs were then aligned to the exons to ensure that they indeed covered the exons producing these scrambled junctions.

RT-PCR and sequencing

DNA polymerase rTth (Applied Biosystems) was used for one-step RT-PCR with gene-specific primers. A two step RT-PCR method was also used where the first-strand cDNA was generated with oligo(dT)$_{20}$ using SuperScript™ III First-Strand Synthesis System (Invitrogen), and subsequently amplified by Taq DNA polymerase (Invitrogen). Human and mouse total RNAs and poly(T)-selected RNAs used in RT-PCR were Premium Total RNAs (Clontech), which were obtained from various healthy tissues. DNA sequencing was conducted using Big Dye v3.1 Terminator Cycle Sequencing Kit with the ABI 3730 capillary DNA sequencer (Applied Biosystems)

RNA quantification

Real-time PCR was conducted using QuantiTect SYBR Green PCR Kit (Qiagen) in the DNA Engine Opticon™ 2 real time PCR system (MJ Research). First-strand cDNAs were synthesized from total RNAs using SuperScript First-Strand Synthesis System (Invitrogen) and subsequently used for real time quantification.

IV. Results

Identification of 85 human genes with evidence of exon-scrambled transcripts

To do a genome-wide survey for exon scrambling events in humans, we searched the human dbEST database with BLAST using concatenated reversed exon-exon junctions including tandem same-exon junctions. After a series of stringent screens to exclude potential sources of error such as cases of tandem gene duplication and unannotated exon duplication in genomic sequences, 172 ESTs spanning scrambled exon junctions were obtained, representing 90 exon scrambling events in 85 human genes (Table 1).

This analysis suggests that exon scrambling is quite rare in the human transcriptome as only 172 ESTs out of about 5 million were detected representing this event, and for most of genes represented there was only one corresponding EST. However, there were some notable exceptions, including 5 genes with two distinct exon scrambling events detected, and 13 exon scrambling events supported by multiple ESTs, including one event (in gene *MRIP2*) that was supported by 13 different ESTs.

RT-PCR tests of EST-supported exon scrambling events

Our computational method relied on the EST sequences, which are known to contain a certain proportion of artifacts. Therefore to check the quality of our data set of EST-supported exon scrambling events, an RT-PCR-sequencing protocol was used to confirm the presence of exon-scrambled transcripts. Five genes were picked from the data set: *GLI3*, a transcription factor involved in the hedgehog pathway, which scrambled across several exons (from exon 8 to 3); mannosidase alpha, which had two predicted scrambling patterns (scrambling from either exon 8 or exon 5 to exon 2); F-Box 7, for which duplication of exon 7 has been previously suggested (Hide et al. 2000), but not experimentally verified; nuclear pore complex interacting protein (NPIP), for which 11 ESTs supported scrambling duplication of exon 2, and Ca^{2+}-promoted Ras GTPase inactivatator (CAPRI), in which scrambling occurred in the 3' terminus (duplication of exon 18).

For each of these genes one RT-PCR primer pair was designed to specifically detect scrambled transcripts and another pair was designed to detect normal transcripts (Figure 2A). With the exception of CAPRI gene, the RT-PCR products in all genes had the predicted sizes and the expected sequence was confirmed by subsequent DNA sequencing (Figure 2B,C).

Therefore, we conclude that a high percentage of predicted scrambling events in our data set is likely to be accurate.

Gene features associated with exon scrambling

The database of EST-supported exon scrambling events was further analyzed to identify gene features that might be associated with exon scrambling. Both the introns immediately upstream and downstream of the scrambled exon junctions were shifted significantly to longer lengths in the CDF plots (Figure 3), suggesting that longer introns are associated with exon scrambling. The increased length of both introns flanking the scrambling events were significant (ANOVA test, $P<2.2e-07$).

Exon scrambling also appeared to favor particular exon positions in the gene. Specifically, 31 out of 90 exon scrambling events joined the 5' ss of a downstream exon to the 3' ss of the second exon of the gene. This bias towards the second exon is consistent with many known examples, such as the *SA*, *COT* and *Sp1* genes (Frantz et al. 1996; Caudevilla et al. 1998; Takahara et al. 2000), and may be related to the fact that the first intron in the gene tends to be longer (Kriventseva and Gelfand 1999).

Among all the 90 exon scrambling events, 38 preserved the normal reading frame of the exons downstream of the scrambled junction. One scrambled transcript has been previously reported to be translated into protein (Caudevilla et al. 1998). Thus a majority of these events might be used to downregulate gene expression, either by producing an mRNA that is a substrate for nonsense-mediated decay (NMD) or producing untranslatable circular RNAs.

Exon scrambling of *GLI3* transcripts is conserved between human and mouse

Exon scrambling of *GLI3* transcripts was analyzed in greater details. Although no EST evidence of exon scrambling was found for the mouse ortholog of *GLI3,* the same design of primers as in the human gene was used to conduct RT-PCR in mouse total RNAs. Interestingly exon-scrambling was observed in several mouse tissues, with the same exons involved in scrambling as in the human gene (Figure 4A). Further analysis on scrambled regions using different RT-PCR primer pairs indicated that the same exon junctions were present in the scrambled transcripts in both human and mouse *GLI3* genes (Figure 4B,C).

Exon scrambling can occur at a high ratio ralative to normal transcript production

In most cases of exon scrambling reported previously, scrambled transcripts occurred at a low level relative to normally spliced transcripts. However, our RT-PCR results indicated that *GLI3* transcripts containing scrambled exon 8-3 junction occurred at a higher abundance than normal transcripts containing the exon 2-3 and 8-9 junctions (Figure 2B,C). To accurately quantify the level of exon-scrambled and unscrambled transcripts, real time PCR was carried out on total RNA samples from different human and mouse tissues (Table 2). The results indicated that although the absolute abundance greatly varied in different tissues, the relative abundance of scrambled *GLI3* transcripts were roughly constant, i.e., a range of ~3-6 times the level of the unscrambled transcript abundance. Despite the variability in *GLI3* expression levels between different batches of tissue total RNAs, this constant relation remained (data not shown). These results showed that the production of scrambled *GLI3* transcript occurred in many tissues of both human and mouse and likely had a constitutive relationship with the normal transcript production.

V. Discussion

The unusual exon reorganization in exon scrambling phenomenon provides another example of how complicated transcript processing can be in mammals. It may also indicate another level of gene expression regulation. In this study, we used a computational method to obtain a list of exon scrambling events in the human transcriptome in conjunction with experimental studies of expression.

The mechanisms that produce exon-scrambled transcripts are still unclear. The requirement for scrambling to occur precisely at splice junctions which was built into our computational method focuses on those exon scrambling events that involve pre-mRNA splicing. Two models have been proposed to explain the phenomenon at the RNA splicing level. The first is circular RNA splicing, in which the splicing machinery joins the 5' ss of the downstream exon to the 3' ss of an upstream exon in the same primary transcript, producing a circular RNA molecule (Nigro et al. 1991). Another possibility is *trans*-splicing. Each of these two models has some supporting evidence in some genes (Cocquerelle et al. 1992; Capel et al. 1993; Zaphiropoulos 1996; Caudevilla et al. 1998; Frantz et al. 1999; Akopian et al. 1999). *Trans*-splicing is a common mechanism for gene maturation in lower eukaryotes, such as trypanosomes and the nematode *C. elegans*, where independently transcribed short sequences called splicing leaders (SL) are spliced to the 5' ends of the transcripts of many genes (Nilsen 2001). This type of *trans*-splicing involves transcripts from two different genes, and therefore is termed "heterotypic *trans*-splicing" to distinguish it from "homotypic *trans*-splicing", in which the two transcripts are from the same gene. Homotypic *trans*-splicing has been proposed to explain some exon scrambling events (Caudevilla et al. 1998). Although no direct observation of natural in vivo *trans*-splicing has been reported in mammals, a spliceosome-mediated RNA *trans*-splicing (SMaRT) method has been used to replace or repair either the 5' or 3' end of a gene using artificial transcript substrates (Mansfield et al. 2003). Several in vitro studies have also indicated that *trans*-splicing can occur under some conditions (Solnick 1985; Konarska et al. 1985). Homotypic *trans*-splicing is the most likely explanation for cases which involve exon repetition.

In some circumstances these two models can be readily distinguished. For example, circular splicing generates products without poly(A) tails, and exon repetition can only be generated by homotypic *trans*-splicing. Although our computational methods do not specifically distinguish between them, we do have evidence for some genes in the dataset. For example, at least 25 exon scrambling events were supported by ESTs covering duplicated exon(s), supporting *trans*-splicing in these genes (e.g., the *NPIP* gene). More direct evidence was obtained for the *GLI3* gene. In this gene the same results were observed when the templates were total RNAs (one step RT-PCR), poly(T)-selected total RNAs (one step RT-PCR), or first strand cDNA synthesized with oligo(dT) primers (two step RT-PCR), respectively (data not shown), which favors the *trans*-splicing model.

As our study has shown, exon scrambling is a very rare event. However, some scrambling events have been reported to generate high abundance of scrambled transcripts. For example, a majority of rat *Sry* transcripts are circular molecules generated by exon scrambling (Capel et al. 1993). Our study gives similar results by accurate real-time RT-PCR quantification of *GLI3* transcripts. The reproducible high rate of scrambled transcripts of some genes as well as other evidence of human/mouse conservation of some exon scrambling events, and the presence of scrambled transcripts in the cytoplasm (Nigro et al. 1991; Capel et al. 1993) indicates that exon scrambling might be an important aspect of the expression regulation of some genes. In particular the scrambled transcripts in *COT* gene has been shown to produce proteins (Caudevilla et al. 1998).

The question of what gene features are required to confer exon scrambling is still unclear. Our statistical analyses showed that the two introns flanking the exon junctions involved in scrambling tend to be significantly longer than other introns in these genes, suggesting that this gene organization facilitates or enables exon scrambling. The bias for the second exon in our dataset might be explained by the intron length bias, since on average first introns are longer than other introns (Kriventseva and Gelfand 1999). Some sequences in the gene might be also required for exon scrambling. Recently, Rigatti et al. (2004) reported that exon scrambling in the rat *SA* gene was allele-specific, suggesting that some sequence elements present only in certain alleles might be required to confer exon scrambling.

The phenomenon of exon scrambling raises questions about the long standing mystery of how exons are paired by the splicing machinery. Exon scrambling is a situation where normal exon pairing is violated. Some of these events could be mistakes of the mechanisms that normally ensure exon pairing. Others might be used to downregulate gene expression through coupling to nonsense-mediated decay (NMD), in cases where premature termination codons are introduced by exon scrambling (Lewis et al. 2003). At least some cases of exon scrambling are almost certainly functional and regulated events because they are well conserved across species. Dissection of the mechanisms of these events may shed light on gene products involved in exon pairing. Therefore, the 85 genes we identified provide a resource for future studies of cis, *trans*- and circular RNA splicing.

VI. References

Akopian A.N., Okuse K., Souslova V., England S., Ogata N., and Wood J.N. 1999. *Trans*-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Lett.* 445(1):177-82.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-402

Capel B., Swain A., Nicolis S., Hacker A., Walter M., Koopman P., Goodfellow P., and Lovell-Badge R. 1993. Circular transcripts of the testis-determining gene *Sry* in adult mouse testis. *Cell.* 73(5):1019-30.

Cocquerelle C., Daubersies P., Majerus M.A., Kerckaert .JP., and Bailleul B. 1992. Splicing with inverted order of exons occurs proximal to large introns. *EMBO J.* 11(3):1095-8.

Caudevilla C., Serra D., Miliar A., Codony C., Asins G, Bach M., and Hegardt F.G 1998. Natural *trans*-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci U S A.* 95(21):12185-90.

Frantz S.A., Thiara A.S., Lodwick D., Ng L.L., Eperon I.C., and Samani NJ. 1999. Exon repetition in mRNA. Pr*oc Natl Acad Sci USA.* 96(10):5400-5.

Hide W.A., Babenko V.N., van Heusden P.A., Seoighe C., and Kelso J.F. 2001. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *G*enome Res. 11(11):1848-53.

Takahara T., Kanazu S.I., Yanagisawa S. and Akanuma H. 2000. Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic *trans*-splicing. *J Biol Chem.* 275(48):38067-72.

Takahara T., Kasahara D., Mori D., Yanagisawa S., Akanuma H. 2002. The *trans*-spliced variants of Sp1 mRNA in rat. *Biochem Biophys Res Commun.* 298(1):156-62.

Konarska M.M., Padgett R.A., and Sharp PA. 1985. *Trans*-splicing of mRNA precursors in vitro. *Cell.* 42(1):165-71.

Kriventseva E.V. and Gelfand M.S. 1999. Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn.* 17(2):281-8.

Lewis B.P., Green R.E. and Brenner S.E. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A.* 7;100(1):189-92

Mansfield S.G., Clark R.H., Puttaraju M., Kole J., Cohn J.A., Mitchell L.G. and Garcia-Blanco M.A. 2003. 5' exon replacement and repair by spliceosome-mediated RNA *trans*-splicing. *RNA.* 9(10):1290-7

Nigro J.M., Cho K.R., Fearon E.R., Kern S.E., Ruppert J.M., Oliner J.D., Kinzler K.W., and Vogelstein B. 1991. Scrambled exons. *Cell.* 64(3):607-13.

Nilsen T.W. 2001. Evolutionary origin of SL-addition *trans*-splicing: still an enigma. *Trends Genet.* 17(12):678-80.

Rigatti R., Jia J.H., Samani N.J., and Eperon I.C. 2004. Exon repetition: a major pathway for processing mRNA of some genes is allele-specific. *Nucleic Acids Res.* 32(2):441-6

Solnick D. 1986. Does *trans*-splicing in vitro require base pairing between RNAs? *Cell.* 44(2):211.

Yeo G., Holste D., Kreiman G., and Burge C.B. 2004. Variation in alternative splicing across human tissues. *Genome Biol.* 5(10):R74.

Zaphiropoulos P.G. 1996. Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. *Proc Natl Acad Sci USA.* 93(13):6536-41.

Zaphiropoulos P.G. 1998. Mechanisms of pre-mRNA splicing: classical versus non-classical pathways. *Histol Histopathol.* 13:585-9

Chapter 2

Studies of Mutually Exclusive Alternative Splicing

I. Abstract

Mutually exclusive alternative splicing (MEAS) is a process in which only one of a set of two or more exons in a gene is included in the final transcript. In most cases how MEAS is regulated is still largely unknown. Here, a database comprised of 101 human genes and 25 mouse genes that contain mutually exclusive exons (MXE) has been built with the GENOA annotation software. This database was analyzed for specific sequence features that might be associated with MEAS. A special "tandem MEAS" pattern was found, in which two MXEs are next to each other without any sequence between them. A genome-wide search for tandem MEAS was done and 11 such human genes were identified. A fluorescence reporting system was also built to study intronic *cis*- regulatory elements for MEAS.

II. Introduction

Mutually exclusive alternative splicing (MEAS) is a process in which only one of a set of two or more exons in a gene is included in the final transcript. It is an unusually complicated splicing pattern because it involves coordination of multiple exons (Black 2003). In most cases how MEAS is regulated is still largely unknown. MEAS may also plays an important role in gene evolution, since it provides the potential to modulate protein functions simply by swapping the mutually exclusive exons (MXE) without disrupting protein size or structure (Letunic et al. 2002; Kondrashov and Koonin 2002). What is particularly interesting is that MEAS can provide multitude of distinct mRNA/protein isoforms. One of the most striking examples is the *Drosophila DSCAM* gene, which can potentially produce 38,016 different mRNAs and proteins (Schmucker et al. 2000; Celotto and Graveley 2001).

An important unanswered question about MEAS is: why are the exons mutually exclusive? Two mechanisms have been uncovered to date. The first one is that the branch point of the downstream MXE is located too close to the 5' splice site of the upstream exon for the spliceosome to form; therefore the 3' splice sites of the two MXEs in a transcript cannot both be used.

Examples include the alpha-tropomyosin and alpha-actinin genes (Smith and Nadal-Ginard 1989; Southby et al., 1999). Another mechanism is that pairs of MEAS exons are flanked by splice sites of incompatible types. Because each of the two types of spliceosomes (U2-type and U12-type) requires distinct sequences at the 5' and 3' end of introns, 'chimerical" introns cannot be processed (Hall et al. 1994; Tarn and Steitz 1996; Burge et al. 1998). Therefore, splicing has to be mutually exclusive there. This mechanism has been proposed for *JNK1* and *p38* (Letunic et al, 2002; Katz and Burge, unpublished data). However, these two mechanisms cannot apply to genes like *DSCAM*, which contains arrays of more than two consecutive MXEs. Regulation of so many MXEs is likely to be more complicated. Therefore, knowing more MEAS genes will be important for us to understand the MEAS. Some novel experimental designing, such as a convenient MEAS reporting system, will also be of great help.

Using the genome annotation software GENOA we have built a database of mutually exclusive exons in several organisms, including human, mouse, and *Drosophila melanogaster*. This database was analyzed for sequence features that are associated with mutually exclusive alternative splicing. In particular we are interested in the genes with multiple MXEs, since these complicated examples might provide direct evidence for MEAS regulation.

Studies on our MXE database have identified some of such examples. A novel fluorescence reporting system was also designed for further investigations of intronic *cis*-regulatory elements.

III. Methods and Results

Build-up of MXE database

Genes with known exon-intron organization were obtained by large-scale alignment of cDNAs to the assembled human genome (hg13) using the genome annotation software GENOA (Yeo et al. 2004; see http://genes.mit.edu/genoa). MXEs were identified on the basis that there is at least one aligned transcript that includes each of a set of adjacent exons, but none of these transcripts includes more than one of these exons, or excludes all of these exons (Figure 5). As the result a MXE database containing 101 human genes and 35 mouse genes has been built (Table 3).

Analyses of MXE database

To investigate the possible characteristic features of MEAS, we further analyzed the MXE database. Because many MXEs have generally been thought to evolve by exon duplication, we first compared sequences between MXE pairs in our database using BLAST2.0 (Altschul et al. 1997). The result showed that 35 out of 101 MXE pairs are highly similar to each other at the amino acid level, suggesting these MXEs are very likely to originate from very recent exon duplication events.

Based on this assumption the MXEs in MEAS genes were divided into two categories: recently duplicated MXEs, and other MXEs, according to the sequence similarity between the MXEs in the gene. They appear to have different properties in many ways. For example, MEAS was conserved in orthologous human/mouse genes for most of the genes in the first category, whereas it is only conserved in very few genes in the second category.

We also checked the reading frames of the transcripts from MEAS, because in order for MEAS not to be detrimental, the reading frame should be preserved. As expected, in more than half of these genes alternation of MXEs does not introduce frame shift, and therefore preserves their potential to generate functional proteins.

These features may help us understand the mechanisms of MEAS. For example, we can therefore focus our studies on (1) the conserved sequences elements, which might be functionally important and therefore conserved between organisms; (2) the sequence variations between MXEs, which might control the choice of MXEs, or (3) the introns flanking MXEs.

A novel splicing pattern we called 'tandem MEAS' has been identified.

As we discussed before, the short distance between two MXEs can force the splicing in a mutually exclusive way. Therefore, it is interesting to know the minimal distance between two MXEs. Surprisingly, it was observed that in the gene *Rbp 7*, there is no sequence between two MXEs. RT-PCR has confirmed that either exon can be included in the final transcript (Figure 6). We named this new pattern "tandem MEAS" because the two MXEs are located in tandem positions.

Subsequently we conducted a genome-wide survey in humans for this tandem MEAS phenomenon. Steps to identify tandem MEAS genes have been illustrated in Figure 7. We checked the regions immediately after exons ending at AG for the potential hidden "tandem MXE". Junctions made by concatenate 50 bp of the 3' end of the preceding exon and first 50 bp of this region were searched against human EST database. The EST hits satisfying the criteria of E<1e-40, >90% identity and >90 bp coverage were reported as potential regions containing hidden MXEs. Junctions made of 50 bp immediately before any GT in this region and the first 50 bp of the next exon were searched against human EST database again. If ESTs confirmed the presence of this junction, the 3' of this potential hidden exon was determined. At the final step, this whole hidden exon sequence was searched against human EST database to confirm its presence. Similar searching was done in the regions immediately before exons starting at GT.

As the result 11 candidate genes were identified (Table 4). This small number, although possibly limited by the incomplete EST data, indicated that this splicing pattern is not common. However, this phenomenon provides a special example to show how MEAS can be possibly accomplished.

Identification of human genes containing multiple MXEs

One striking phenomenon of MEAS is its potential for generating hundreds of distinct mRNA or proteins if they have one or more groups of multiple MXEs, which has been reported in Drosophila *DSCAM* gene (Schmucker et al. 2000).

From above analysis we concluded that at least a proportion of MXEs can be identified by sequence similarity at the amino acid level. Based on this conclusion, a set of MEAS genes containing multiple MXEs in human were identified. Using the MXEs from MEAS database, these exons were searched against the introns flanking these exons, including the region between them for the third 'hidden' MXE. As the result four MEAS genes in our database were identified as candidate genes containing multiple MXEs.

One of them is the L-3-phophoserine phosphatase gene (Figure 8), in which two known MXEs have the same lengths (35bp). The newly identified potential MXE located between them also has the same length. Due to the high sequence similarity of the "hidden" exon to the other two MXEs, it is easy to be missed, which may explain the reason why it has not been annotated.

Construction of a fluorescence MEAS reporting system

As a subtype of alternative splicing MEAS is precisely regulated in different developmental and differentiation stages by various *cis*- and *trans*-acting regulatory elements (Lopez 1998; Smith et al 2000; Grabowski et al. 2001; Graveley, 2001; Maniatis and Tasic 2002; Black, 2003). In the known examples, most *cis*-regulations come from introns. To study intronic *cis*-regulatory elements for MEAS, we designed a fluorescence MEAS reporting system based on the fact that CFP and YFP fluorescence proteins are almost identical except a few amino acids (Figure 9).

The CFP/YFP cDNA sequences were divided into three pieces. The first and third pieces are identical in CFP and YFP and all sequence variations are only within the second piece. The second pieces from both genes as well as the first and last pieces were designed as a construct with four exons by inserting intronic sequences among them. By including the restriction sites within the intronic sequences, we can test the functions of intronic sequence elements by monitoring the fluorescence of produced proteins using the cell cytometry method.

IV. Conclusions

To understand more about alternative splicing, especially mutually exclusive alternative splicing, we used GENOA software and built a database containing 101 human genes and 37 mouse genes that undergo MEAS. Our statistical analyses on this database showed that most of these genes tend to preserve natural reading frames during MXE alternation. Many of MXE pairs in these genes showed high sequence similarity at the amino acid level and were conserved between human and mouse.

Based on the property of sequence similarity we identified MEAS genes containing multiple MXEs, such as the L-3-phophoserine phosphatase gene. Studies on these genes may be helpful to answer one of the most interesting questions about MEAS: how are so many exons precisely regulated in genes like *DSCAM*.

An unknown "tandem MEAS" phenomenon was also identified. Although only very few such examples were found by our genome-wide search, this special organization may help us understand how MEAS is regulated in the special circumstances where two MXEs are very close to each other.

We also built a fluorescence reporting system based on the CFP/YFP fluorescence proteins. This system can be used to study intronic *cis*-regulatory elements for MEAS.

V. References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402

Black D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem.* 72:291-336.

Burge C.B., Padgett R.A., Sharp P.A. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell.* 2(6):773-85.

Celotto A.M., Graveley B.R. 2001. Alternative splicing of the Drosophila *DSCAM* pre-mRNA is both temporally and spatially regulated. *Genetics.* 159(2):599-608.

Grabowski P.J. and Black D.L. 2001. Alternative RNA splicing in the nervous system. *Prog. Neurobiol.* 65:289–308.

Graveley, B.R. 2001 Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17:100-107.

Hall S.L. and Padgett R.A. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splices sites. *J. Mol. Biol.* 239:357-365.

Kondrashov F.A., Koonin E.V. 2001. Origin of alternative splicing by tandem exon duplication. *Hum Mol Genet.* 10(23):2661-9.

Letunic I., Copley R.R., Bork P. 2002. Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet.* 11(13):1561-7.

López A.L. 1998. Alternative splicing of pre mRNA: developmental consequences and mechanisms of regulation. *Ann. Rev. Genet.* 32:279–305.

Maniatis T. and Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature.* 418:236–243.

Schmucker D., Clemens J., Shu J., Worby C., Xiao J., Muda M., Dixon J., and Zipursky L. 2000. *Cell.* 101:671-684

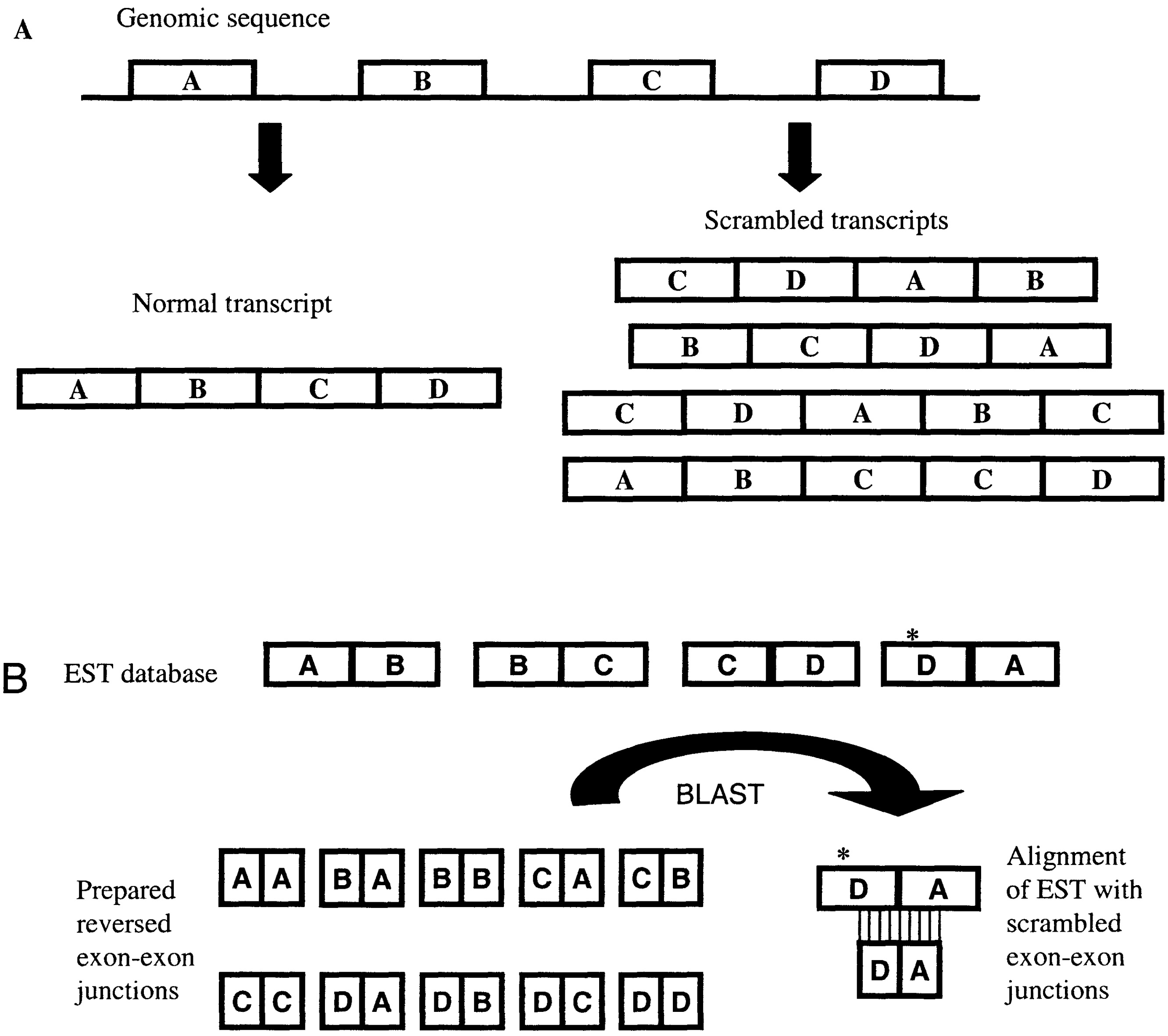
Smith C.W. and Nadal-Ginard B. 1989. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell.* 56(5):749-58.

Southby J., Gooding C., Smith C.W. 1999. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. *Mol Cell Biol.* 19(4):2699-711.

Tarn W.Y. and Steitz J.A. 1996. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell.* 84:801-811.

Yeo G., Holste D., Kreiman G., and Burge C.B. 2004. Variation in alternative splicing across human tissues. *Genome Biol.* 5(10):R74.

Figure 1


A
Genomic sequence
A
B
C
D
Normal transcript
A
B
C
D
Scrambled transcripts
C D A B
B C D A
C D A B C
A B C C D
B
EST database
A B
B C
C D
*
D A
BLAST
Prepared reversed exon-exon junctions
A A
B A
B B
C A
C B
C C
D A
D B
D C
D D
*
D A
D A
Alignment of EST with scrambled exon-exon junctions

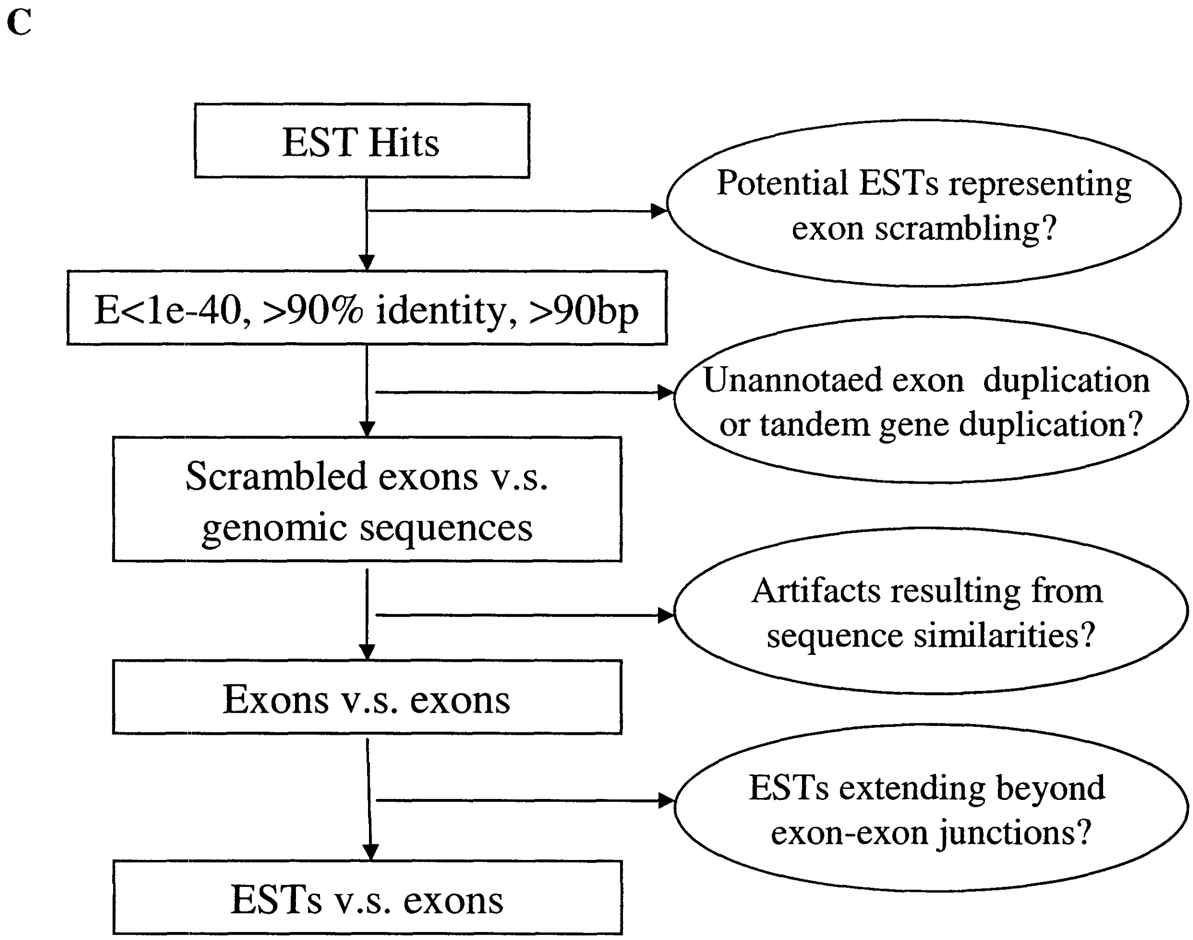

C
EST Hits
Potential ESTs representing exon scrambling?
E<1e-40, >90% identity, >90bp
Unannotaed exon duplication or tandem gene duplication?
Scrambled exons v.s. genomic sequences
Artifacts resulting from sequence similarities?
Exons v.s. exons
ESTs extending beyond exon-exon junctions?
ESTs v.s. exons

Figure 1. Identification of ESTs representing exon scrambled events. (A) In exon scrambling exons are spliced at correct sites but joined in a unusual order. A,B,C,D represent exons. (B) The strategy to identify exon scrambling is to search exon-exon junctions against EST database for ESTs representing exon scrambling events. Exon D was spliced to exon A in the figure, which was represented by a EST covering this abnormal D-A junction. (C) Several steps were taken to exclude various artifcacts.

Figure 2

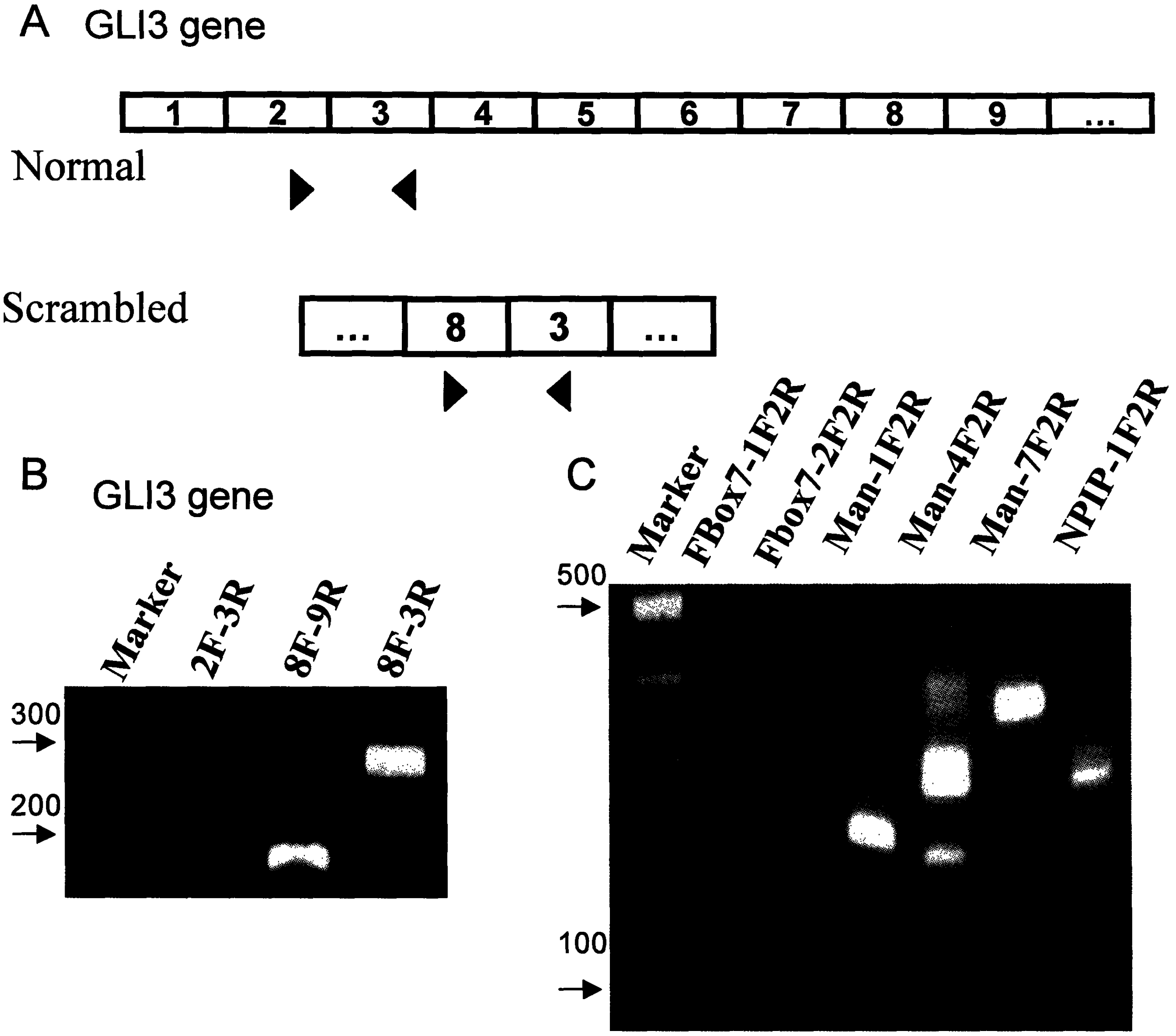


Figure 2. RT-PCR confirmation of exon scrambling. (A) GLI3 gene was illustrated. Primers were designed to specifically detect scrambled transcripts. (B) A product of predicted size was produced with primer set 8F-3R. Primer sets 2F-3R and 8F-9R were also used to confirm the normal junctions. (C) More genes were tested. Exon scrambling events in *F-box 7, Mannosidase,* and *NPIP* genes were validated.

Figure 3

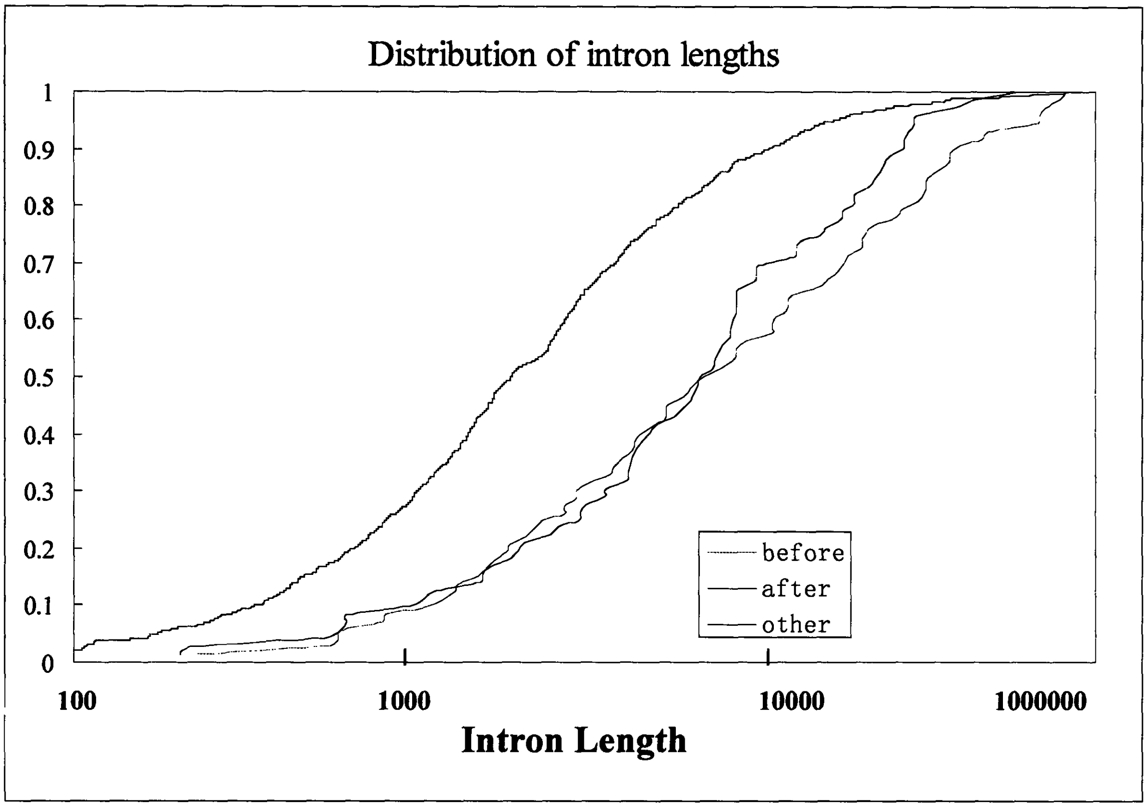


Figure 3. Cumulative distribution function of intron length in genes within our dataset. Both the introns immediately upstream and downstream of the scrambled exon junctions were shifted significantly to longer lengths compared with other introns. The increased length of both introns flanking the scrambling events were significant (ANOVA test, P<2.2e-07).

Figure 4

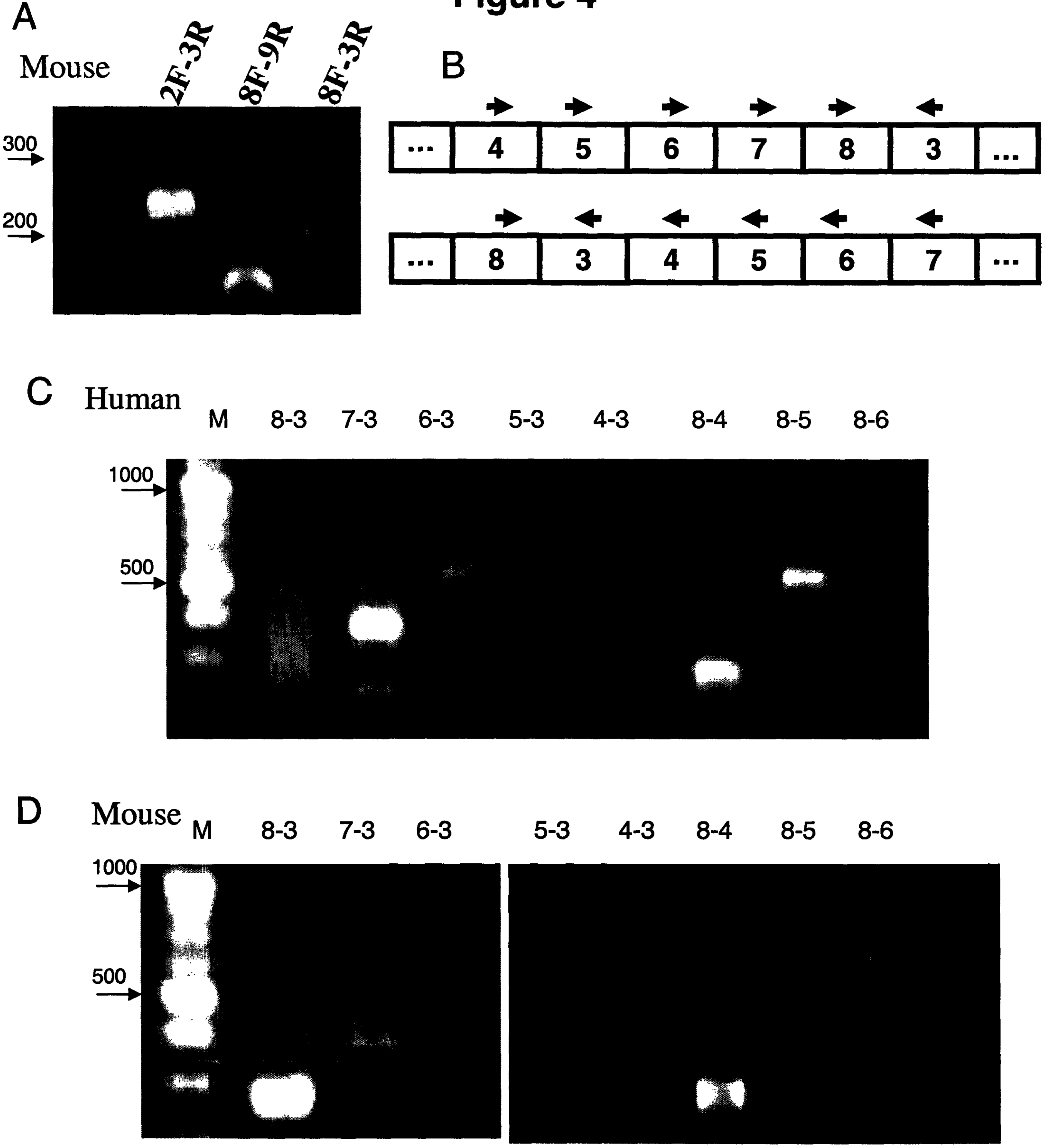


Figure 4. Exon scrambling of *GLI3* gene was conserved between human and mouse. (A) Using the same primer designing exon scrambling of *GLI3* gene in mouse was also verified. (B) The designing of primer sets used to exclude the possibility of RT-PCR artifacts. (C) RT-PCR in human tissues. (D) RT-PCR in mouse tissues.

Figure 5

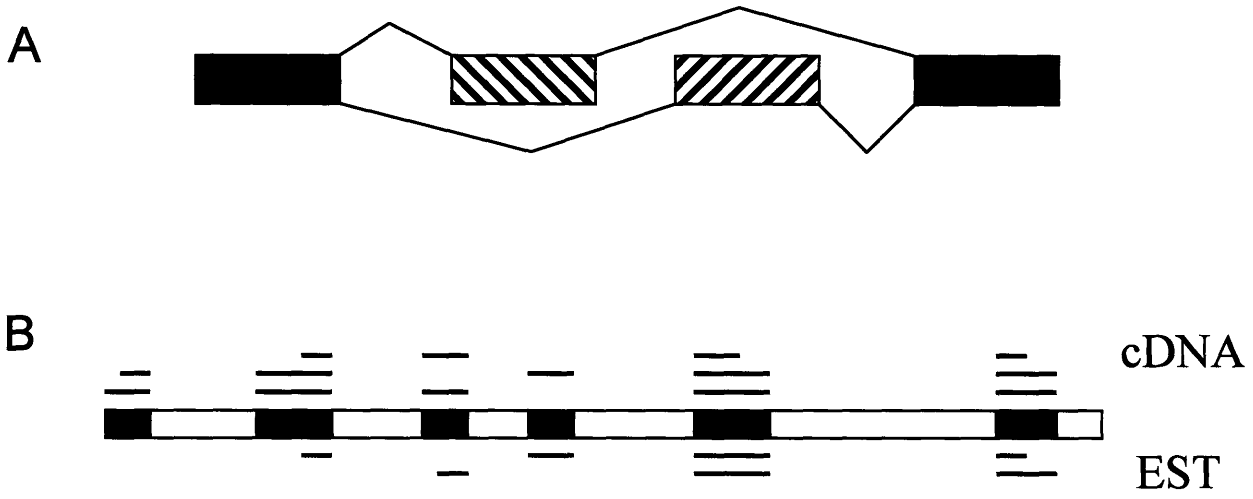


Figure 5. Schematic illustration of mutually exclusive alternative splicing (MEAS). (A) In MEAS only one of a set of two or more exons in a gene is included in the final transcript. (B) GENOA definition of MEAS. Mutually exclusive exons (MXE) were identified on the basis that there is at least one aligned transcript (cDNA/EST) that includes each of a set of adjacent exons, but none of these transcripts includes more than one of these exons, or excludes all of these exons.

Figure 6

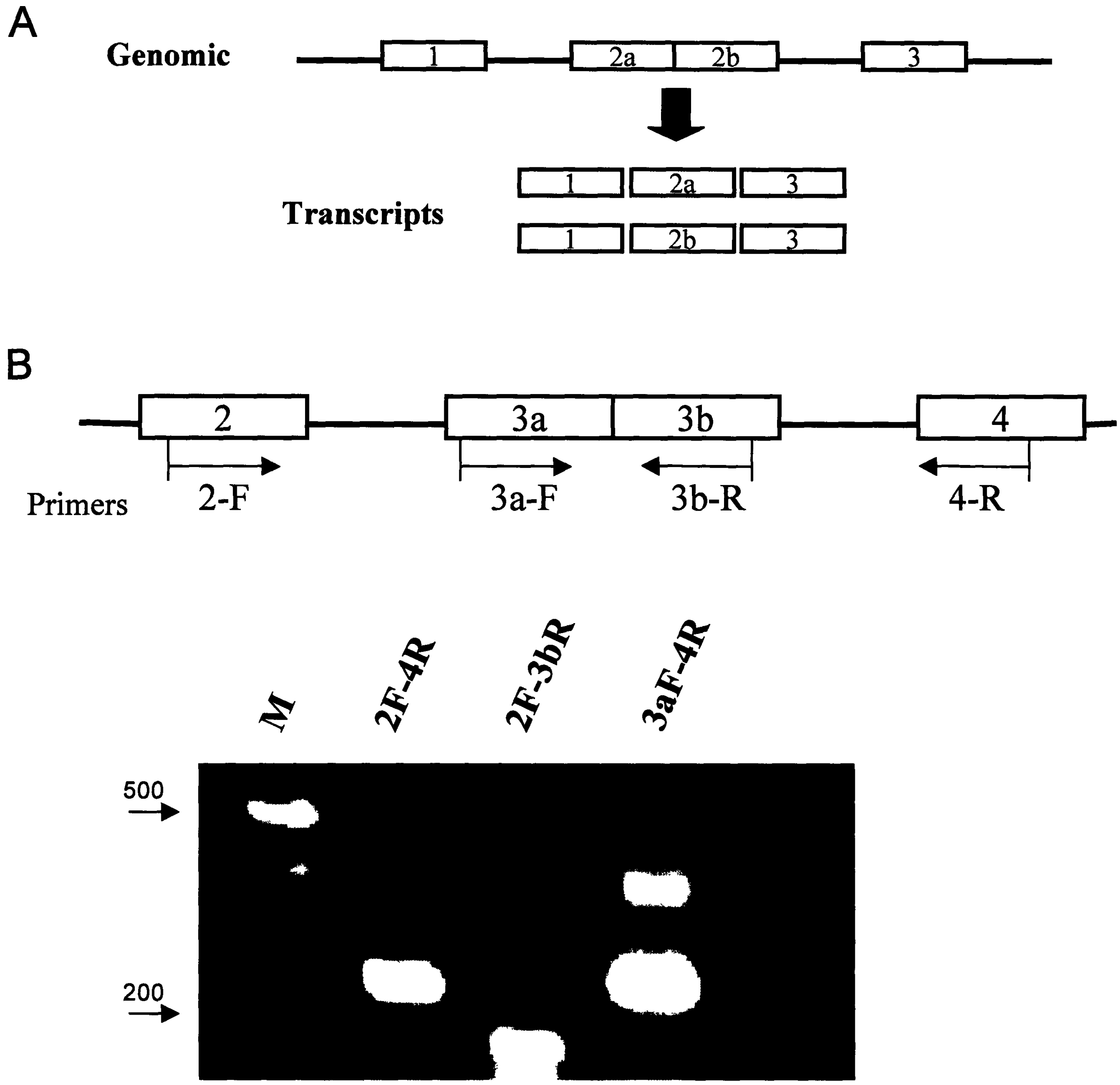


Figure 6. Tandem mutually exclusive alternative splicing. (A) In tandem MEAS two MXE are next to each other without any sequence in between. (B) In *Rbp 7* gene tandem MEAS was studied by RT-PCR. Both putative exon 3a and 3b can appear in the RT-PCR products.

Figure 7

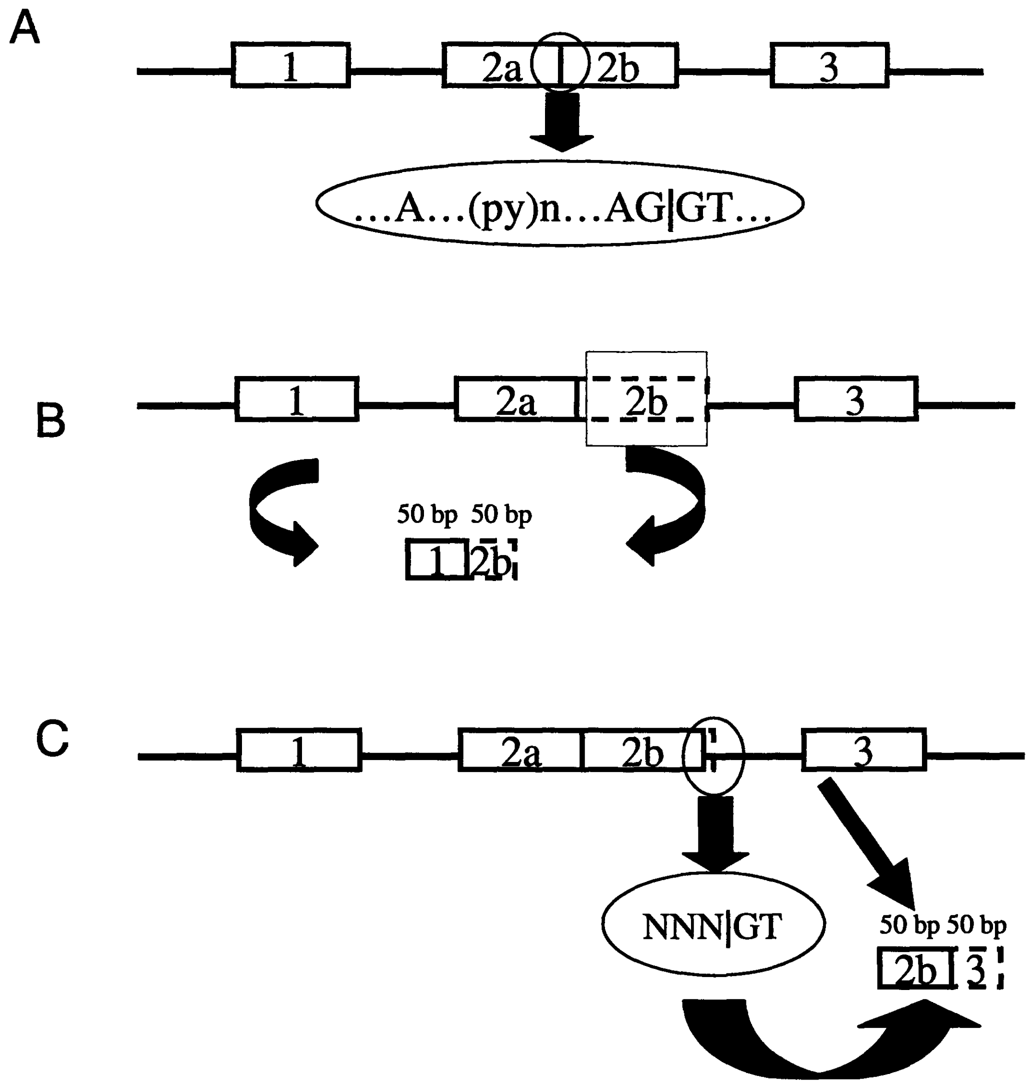


Figure 7. Three steps were taken to identify tandem MEAS. (A) Find all the exons (illustrated as exon 2a in the figure) ending at AG. (B) Make junctions by concatenating 50 bp of the 3' end of exon 2a and first 50 bp of the downstream region and search against human EST database. (C) Make junctions by combining 50 bp immediately before any GT in this region and the first 50 bp of exon 3 were searched against human EST database again. If ESTs confirmed the presence of this junction, the 3' of this potential this hidden exon 2b was determined.

Figure 8

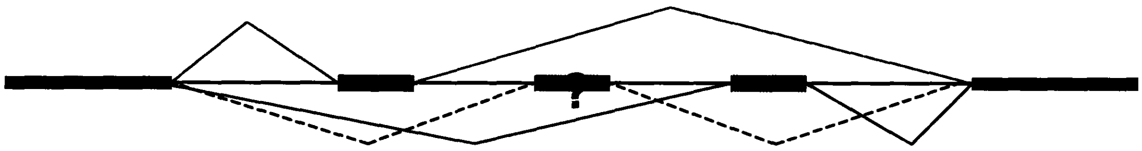


Figure 8. One example of genes containing multiple MXEs: L-3-phophoserine phosphatase gene. Two known MXEs of this gene (red and blue) were searched against the introns flanking these exons, including the region between them. The hidden exon (green) was found base on its sequence similarity to the two known MXEs.

Figure 9

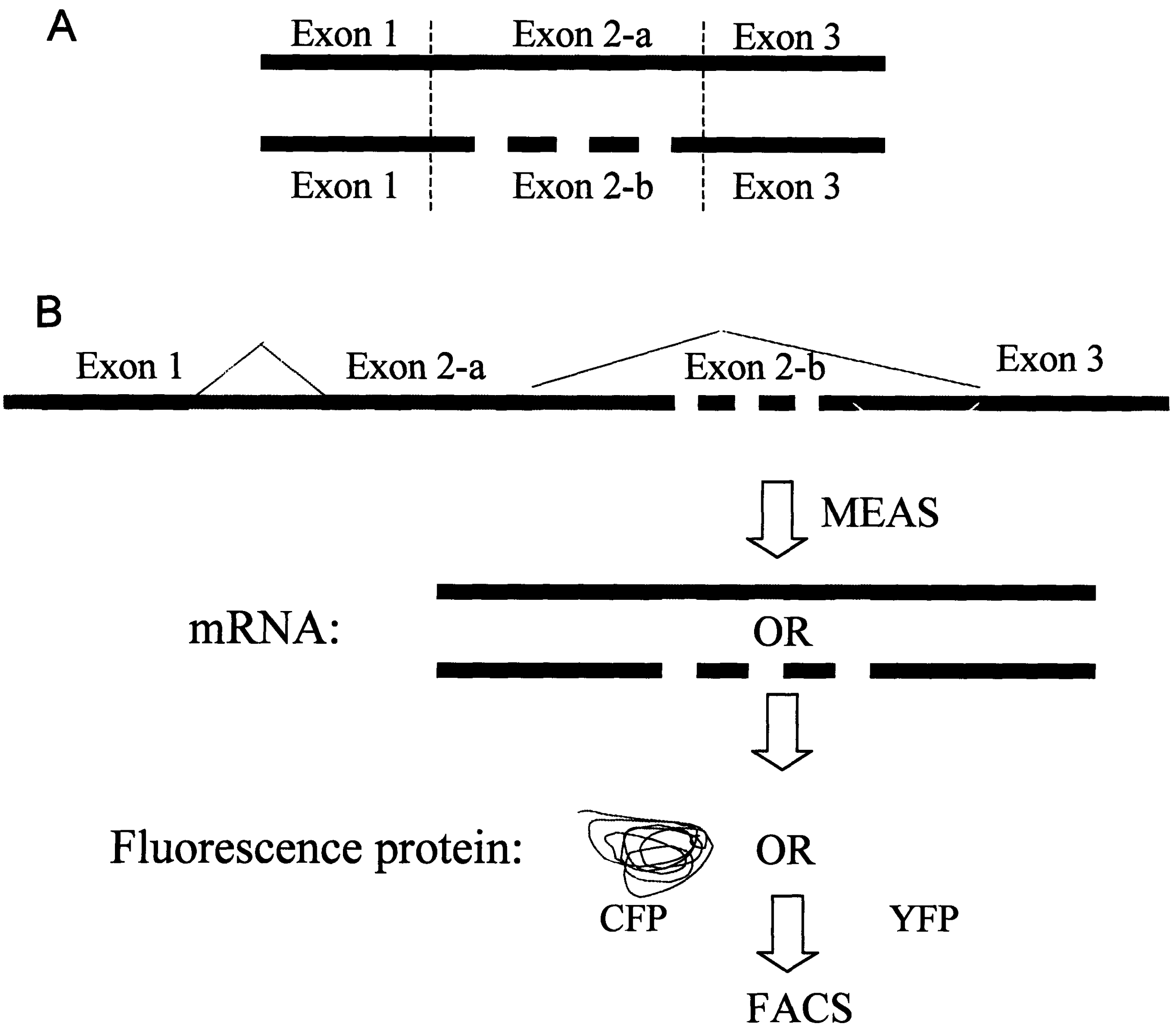


Figure 9. Designing of fluorescence reporting system for MEAS. (A) CFP and YFP proteins are almost identical except a few amino acids. (B) The construct was made by chopping the CFP/YFP genes into three pieces and using the middle parts which contain the sequence variation as MXEs. MEAS can be monitored by the fluorescence of the produced protein using FACS.

Table 1

ENSEMBL ID	Gene Name	Exon Junction	EST COUNT
ENSG00000002746	NEDD4-like ubiquitin-protein ligase 1	23_10	1
ENSG00000005483	Myeloid/lymphoid or mixed-lineage leukemia 5	5_3	1
ENSG00000008177	Mitogen-activated kinase kinase kinase 5	7_2	1
ENSG00000008952	Human Sec62 homolog	7_2	1
ENSG00000011260	WD-repeat protein CGI-48	4_2	1
ENSG00000041357	Proteasome subunit alpha type 4	6_4	1
ENSG00000042832	Thyroglobulin	23_23	1
ENSG00000049541	Human replication factor C, 40-kDa subunit	7_5	1
ENSG00000067369	Tumor suppressor p53-binding protein 1	11_10	2
ENSG00000072736	T cell transcription factor NFAT4	9_4	2
ENSG00000073921	Phosphatidylinositol-binding clathrin assembly protein	12_2	1
ENSG00000074416	Human lysophospholipase homolog	3_3	1
ENSG00000074416	Monoglyceride lipase	3_3	1
ENSG00000075151	Eukaryotic translation initiation factor 4 gamma 3	4_1	1
ENSG00000081148	Interphotoreceptor matrix proteoglycan 2	11_3	1
ENSG00000089006	Sorting nexin 5	11_2	1
ENSG00000090905	Trinucleotide repeat containing 6	5_2	2
ENSG00000100060	Manic fringe precursor	7_7	1
ENSG00000100225	F-box only protein 7	7_7	1
ENSG00000100664	Eukaryotic translation initiation factor 5	9_4	1
ENSG00000101040	Protein kinase C binding protein 1	14_13	1
ENSG00000101773	CtBP interacting protein CtIP	12_10	1
ENSG00000103194	Ubiquitin carboxyl-terminal hydrolase 10	3_2	1

ENSEMBL ID	Gene Name	Exon Junction	EST COUNT
ENSG00000105808	Ca2+-promoted Ras inactivator	18_18	1
ENSG00000105821	M-phase phosphoprotein 11	10_8	1
ENSG00000106571	Zinc finger protein GLI3	8_3	1
ENSG00000107368	Transducin-like enhancer protein	6_5	1
ENSG00000109458	GRB2-associated binding protein 1	2_2	1
ENSG00000109920	Formin binding protein 4	12_5	1
ENSG00000112699	GDP-mannose 4,6 dehydratase	2_2	1
ENSG00000114346	Epithelial cell transforming sequence 2 oncogene	16_6	1
ENSG00000114416	Fragile X mental retardation syndrome related protein 1	14_14	1
ENSG00000115310	Neurite outgrowth inhibitor	5_4	1
ENSG00000115919	L-kynurenine hydrolase	9_4	1
ENSG00000117523	HBxAg transactivated protein 2	14_9	1
ENSG00000117713	SWI-SNF complex protein p270	4_2	1
ENSG00000122512	PMS1 protein homolog 2	4_1	3
ENSG00000123965	Postmeiotic segregation increased 2-like 5	5_2// 6_2	6//1
ENSG00000124177	Chromodomain-helicase-DNA-binding protein 6	4_3	1
ENSG00000124795	Death Kinase	8_2	1
ENSG00000126858	Ras homolog gene family	4_3	1
ENSG00000128487	Sperm antigen HCMOGT-1	2_2	1
ENSG00000135093	Ubiquitin specific protease 30	?	1
ENSG00000141720	Phosphatidylinositol-4-phosphate 5-kinase	7_2	1
ENSG00000143842	SRY-like DNA binding protein	6_4	1
ENSG00000145216	FIP1-like 1	12_9	1
ENSG00000147010	SH3-domain kinase binding protein 1	6_5	2

ENSEMBL ID	Gene Name	Exon Junction	EST COUNT
ENSG00000147044	Peripheral plasma membrane protein CASK	8_6	1
ENSG00000147649	LYRIC/3D3 protein	11_6	1
ENSG00000151883	Poly (ADP-ribose) polymerase 8	9_3	1
ENSG00000158158	Cyclin M4	5_2	1
ENSG00000158636	EMSY protein	8_7	1
ENSG00000159256	Zinc finger CW-type coiled-coil domain protein 3	7_5	1
ENSG00000160271	Ral guanine nucleotide dissociation stimulator	2_2	1
ENSG00000162959	Protein C2orf4	6_4	1
ENSG00000164253	WD repeat domain 41	8_4	2
ENSG00000164769	Aspartyl(asparaginyl)beta-hydroxylase	3_2	1
ENSG00000170603	Nuclear pore complex interacting protein NPIP	2_2	11
ENSG00000170776	LBC oncogene	19_13	1
ENSG00000177425	Apoptosis response protein 4	3_3	1
ENSG00000198162	Mannosidase alpha class 1A member 2	5_2// 8_2	2//1

Table 1. Partial list of EST-supported exon scrambling events. Their ENSEMBL IDs, gene names, EST-supported scrambled exon junctions, and the numbers of supporting ESTs were shown in the table.

Table 2

	RNA amount (10^{-11} mg)		
Tissue	2-3	8-9	8-3
Heart	1.46±0.03	4.20±1.23	15.6±1.43
Brain	4.66±0.16	11.2±2.34	31.6±4.34
Fetal Brain	4.21±1.1	4.50±0.35	11.2±2.39
Plancenta	9.57±2.05	7.00±1.28	39.6±6.78
Lung	7.05±0.05	8.80±0.86	42.4±3.52
Skeletal muscle	3.05±1.23	3.30±0.59	14.2±2.06

Table 2. Amount of transcripts of GLI3 gene in various tissues was quantified by real-time RT-PCR. The amount of transcripts containing exon junctions 2-3, 8-9 and 8-3 were quantified, respectively. The amount of 8-3 junctions indicates the level of exon scrambling.

Table 3

Human MEAS Gene	Mouse MEAS gene
FIBROBLAST GROWTH FACTOR RECEPTOR 1	**FIBROBLAST GROWTH FACTOR RECEPTOR 1**
ALPHA-TROPOMYOSIN	**ALPHA-TROPOMYOSIN**
VOLTAGE-DEPENDENT CALCIUM CHANNEL ALPHA-1A	**VOLTAGE-DEPENDENT CALCIUM CHANNEL ALPHA-1A**
ALPHA-ACTININ	**ALPHA-ACTININ**
ACYL-COENZYME A OXIDASE 1	**ACYL-COENZYME A OXIDASE 1**
FIBROBLAST GROWTH FACTOR RECEPTOR 2	GTP-BINDING PROTEIN SAR1A
MAPK8/JNK1	MITOGEN-ACTIVATED PROTEIN KINASE 9 (JNK2)
DNA-DIRECTED RNA POLYMERASE II 19 KDA POLYPEPTIDE	ATAXIN 2-BINDING PROTEIN
TROPONIN T BETA	MITOGEN-ACTIVATED PROTEIN KINASE 14 (P38)
PYRUVATE KINASE M1	ANTITHROMBIN-III
DYNAMIN 2	SYNAPTOSOMAL-ASSOCIATED PROTEIN 25 (SNAP-25)
ARF GTPASE-ACTIVATING PROTEIN GIT2	ALPHA CASEIN
FAD SYNTHETASE	RAS-RELATED PROTEIN RAB-6A
MITOGEN-ACTIVATED PROTEIN KINASE P38ALPHA	DNA-DIRECTED RNA POLYMERASE II 33 KDA POLYPEPTIDE
CYTOCHROME P450 3A5	ELONGATION FACTOR 1-DELTA (EF-1-DELTA)
ANNEXIN A8	TRANSCRIPTIONAL REGULATOR ERG (FRAGMENT)
N-GLYCANASE 1	BETA-TROPOMYOSIN
TRANSCRIPTION FACTOR (P38 INTERACTING PROTEIN)	STAR-RELATED LIPID TRANSFER PROTEIN 4 (STARD4)
6-PHOSPHOFRUCTOKINASE, TYPE C	X TRANSPORTER PROTEIN 2
NUCLEAR FACTOR 1A	TRAF AND TNF RECEPTOR ASSOCIATED PROTEIN

Table 3. Partial list of genes in the GENOA MXE database, which contains 101 human genes and 35 mouse genes. Left column lists human MEAS genes, and right column lists mouse MEAS genes. MEAS pattern was conserved between human and mouse in genes in red color (bold).

Table 4

Gene name
Small nuclear ribonucleoprotein polypeptide N
KIAA0618 protein
Splicing factor, arginine/serine-rich 7
Myelin associated glycoprotein
RNA polymerase II seventh subunit 7
Interferon regulatory factor 3
Zinc finger DHHC domain containing 19
cDNA DKFZp434
Retinoblastoma binding protein 7
CDNA clone IMAGE: 4821863
12 BAC RP11-512M8

Table 4. Tandem MEAS in 11 genes were supported by ESTs.