Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 342 (2006) 197–205
BBRC
www.elsevier.com/locate/ybbrc

Single amino-acid InDel variants generated by alternative tandem splice-donor and -acceptor selection ☆,☆☆

Chun-Hung Lai [a], Ling-Yueh Hu [a,b], Wen-chang Lin [a,c,*]

[a] Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan, ROC
[b] Graduate Institute of Life Sciences, National Defense Medical Center, Taipei 114, Taiwan, ROC
[c] Institute of Bioinformatics, School of Medicine, National Yang-Ming University, Taipei 112, Taiwan, ROC

Received 16 January 2006
Available online 2 February 2006

Abstract

We have investigated putative single amino-acid InDel variants with human ESTs. Examination of the formation process for single amino-acid InDel variants indicates a possible splicing mechanism in addition to the genomic insertion/deletion events as would be expected. The wobble-splicing transcripts were often generated around the intron–exon boundaries by selecting an alternative neighboring splice signal sequence, in particular the tandem agNAG or GTNgt sequence at the splice-acceptor or -donor site, thus creating single amino-acid InDel isoforms. Another category of variants was identified with one altered amino-acid plus one amino-acid InDel, under divergent coding-frame usage. We demonstrate that such minute distance of splice site choice generates an even greater level of transcriptome diversity, and suggest that non-functional synonymous or intronic SNPs could be converted to functionally significant InDel alterations through this process. This subtle alteration in mRNA and protein-coding sequence may elicit a great impact upon human genome and proteome diversity.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Bioinformatics; EST; InDel mutation; Wobble-splicing; Alternative splicing

Completed genome sequences have provided critical information on gene structure and functions in the last decade, especially the human genome project [1]. The identification of certain genes associated with human diseases and the application of such genomic information to the prevention and treatment of such diseases would appear to be the major direction of effort in the genomic research area. Single nucleotide polymorphism (SNP) markers are defined as single base mutations that are found in more than 1% of the human population [2]. Many SNPs could, directly, involve functional regulations, in addition to serving as surrogate markers, if the SNPs occur in the coding or regulatory regions of human genes. Protein variants generated by these non-synonymous cSNPs (coding-region SNPs) are responsible for the majority of the diversity in the characteristics of the human population and the assured culprits of all human diseases associated with inheritance [3,4]. It has been challenging, however, to identify direct and specific disease-causing polymorphisms due to the enormous number of SNPs that have currently been demonstrated to exist, most of which are located in the non-coding region of genes without obvious biological significance.

Our laboratory has been using EST dataset for novel human gene discovery and cSNP identification [5,6]. In order to investigate putative low-frequency but phenotype-linking cSNPs from ESTs, we have modified our comparative gene identification (CGI) program to perform tblastn comparison of human reference-protein sequences with the human EST dataset. Here, we used the human reference-protein sequences as scaffolds to align human ESTs

☆ This paper contains supplementary data that can be found at http://140.109.42.19/wobble.htm.
☆☆ *Abbreviations:* CGI, comparative gene identification; EST, expressed sequence tag; SNP, single nucleotide polymorphism; cSNP, coding-region single nucleotide polymorphism; InDel, insertion and deletion variant.
* Corresponding author. Fax: +886 2 2785 8594.
E-mail address: wenlin@ibms.sinica.edu.tw (W. Lin).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.01.101

and adopted the "good neighborhood" concept as specified in the 2000 work of Altshuler et al. [7]. We defined our potential cSNP by first selecting one single mismatched amino-acid residue surrounded by 10 perfectly aligned amino-acid residues located on either side of the mismatched residue. By this method, not only were non-synonymous cSNPs revealed, but also single amino-acid insertion and deletion mutation variants (InDel) were able to be recognized effectively. In addition to SNP-based point mutations, the less-frequent protein InDel sequence modification can play a significant role in gene evolution and protein function. For example, there are several well-known human diseases including Huntington's disease and spinocerebellar ataxia syndrome [8], which is associated with trinucleotide repeat-based coding sequence expansion of InDel. Single amino-acid InDels might also feature certain functional impacts in terms of cellular functions and regulations [9]. The single amino-acid InDel has not been well characterized in the past. In this report, we have focused our study to discover such single amino-acid InDels from human ESTs.

Materials and methods

InDel discovery in silico from human ESTs. The human reference-protein dataset and EST-human dataset were both obtained from National Center for Biotechnology Information, National Institutes of Health, USA. The reference-protein dataset (released on January, 2003) contained 17,234 different protein sequences, whereas the EST-human dataset (released on November 4, 2002) contained 4,816,479 entries. These datasets were extracted and stored in the MySQL database as well as being appropriately formatted for our Linux-based blast server. The blast server program (version 2.0.4) was obtained from NCBI and established locally. The blast parameters used for this analysis were $e = 10$, $v = 500$, $b = 1000$, and $w = 0$. Java-written computer programs were modified from the CGI program [5] for this study. To investigate only amino-acid sequence-altering SNPs and to avoid EST-associated sequencing quality issues, we defined our potential cSNP by first selecting one single mismatched amino-acid residue surrounded by 10 perfectly aligned residues on either side of the mismatched amino-acid residue following the tblastn searches with whole human reference-protein sequences used as initial search queries. In order to cover the ends of EST sequences, we later used three residues and eight residues on both sides for the first round of cSNP/InDel screening. Selected specific cSNP/InDel regions (21 residues) were then extracted and further examined by another round of tblastn searching with ESTs in order to determine the distribution frequency of individual cSNP/InDel sites. Additional screening filters were established in order to remove duplicated results from ESTs of gene isoforms and paralogous genes. All search and data-mining results were deposited in the MySQL database, and separate data-mining rules were applied to retrieve cSNPs and InDels results from this dataset. The complete list of InDels can be browsed through at http://140.109.42.19/wobble.htm.

RT-PCR and sequence validation of cSNP and wobble-spliced transcripts. All cDNA was prepared from human gastric cancer-cell lines as indicated [10]. Briefly, reverse transcription was carried out using 25 μg total RNA, oligo$(dT)_{15}$V, and SuperScript™ II reverse transcriptase (Invitrogen; Carlsbad, CA, USA). The quality of reverse transcription products was monitored by agarose-gel electrophoresis after PCR with GAPDH-specific primers. The PCR primer pairs were used to amplify the target sequences. The gene-specific PCRs were conducted at 94 °C for a period of 5 min, generally 35 cycles of 94 °C/20 s, 58 °C/30 s, and 72 °C/30 s, and the final extension phase at 72 °C for 10 min using a PCR thermocycler and Takara *Taq* polymerase (Takara; Shiga, Japan). The final PCR products were cleaned and subjected to restriction endonuclease digestion to distinguish between transcript isoforms as indicated. A part of the amplified fragments were subcloned into a pGEM-T Easy cloning vector provided by Promega (Madison, WI, USA). Following the cloning procedure, several clones were randomly selected, plasmids were extracted, and their sequence was determined by an autosequencer in our Institute's core facility.

Results and discussion

Single amino-acid InDel identification

ESTs have long been used for gene expression profiles and splicing analysis [11–15]. In an initial study featuring 17,234 human reference-protein sequences and 4,816,479 human ESTs, a total of 217,473 potential non-synonymous cSNPs and 1477 single amino-acid InDel residues were identified (Supplementary material about the original blast search results and lists of reference gene NP number and gene symbol can be accessed at http://140.109.42.19/wobble.htm). Among the 1477 InDel sites initially discovered, 271 redundant loci were removed for they are the products of alignments with repetitive amino-acid strings as query sequences, each of these sites should be counted once instead of twice or more. Four hundred and four InDels featuring AAA sequence-encoding lysine residues (365 sites) and TTT sequence-encoding phenylalanine residues (39 sites) deriving from some particular cDNA libraries were noted and held for further analysis. A closer inspection of all of these transcripts revealed the complex nature of them. These ESTs often carry multiple homo-AA/TT dinucleotide or homo-AAAA/TTTT tetranucleotide deletion besides the homo-AAA/TTT trinucleotide deletion on a same transcript (Supplement Figure 1), hence not to be the subject of our study here and are excluded from the statistical analysis.

Although these InDels usually presented at a relatively low frequency in EST numbers, it may be that they are functionally significant even with only one amino-acid alteration if such an alteration occurred at a critical structure or if it arose at a variety of catalytic sites [16,17]. Herein, single amino-acid InDels were further analyzed in detail for the particular origin of insertion or deletion at the DNA level. Fourteen InDel sites could not be assigned due to duplicated sequence at the ends of adjacent exons and incomplete human genomic sequence at that region (Table 1). It is apparent that 297 InDels did occur in the middle region of exons and, as such, may represent a bona fide three base-pair InDel event at the genomic DNA level. By using restriction enzymes recognizing the specific InDel regions, we demonstrated that cells with different genotypes generated specific types of mRNA transcript (Fig. 1A). One cell line (COLO 205) did express both alleles of NM_015925 (NP_057009) LISCH7 gene.

However, for most of the other InDel cases, the putative InDel sites proved to be located right at intron–exon boundaries following manual inspections. A nucleotide sequence distribution tendency was observed toward 'GT'

Table 1
Nucleotide codon sequence summary on the 802 InDel amino-acid residues identified

InDel a.a. codon	Numbers of InDel located at			
	5′-Splice donor site	3′-Splice acceptor site	Exonic position	Not determined[b]
AAA			14	
AAC, ACA, CAA		0, 0, 2	16, 2, 3	
AAG, AGA, GAA	1, 0, 0	40, 7, 56	19, 1, 11	0, 0, 1
AAT, ATA, TAA[a]			13, 2, —	1, 0, —
ACC, CCA, CAC	1, 0, 0	1, 0, 0	5, 4, 5	
ACG, CGA, GAC			1, 0, 3	
ACT, CTA, TAC			5, 0, 3	1, 0, 0
AGC, GCA, CAG	0, 2, 0	43, 122, 137	10, 6, 11	
AGG, GGA, GAG	0, 0, 1	0, 8, 18	2, 7, 9	
AGT, GTA, TAG[a]	0, 8, —	8, 12, —	5, 7, —	
ATC, TCA, CAT		0, 0, 1	12, 1, 2	
ATG, TGA[a], GAT		2, —, 0	5, —, 16	1
ATT, TTA, TAT			1, 2, 4	
CCC			3	
CCG, CGC, GCC			1, 2, 1	
CCT, CTC, TCC			2, 1, 6	
CGG, GGC, GCG	1, 0, 0	1, 0, 1	2, 1, 2	
CGT, GTC, TCG			3, 3, 1	
CTG, TGC, GCT		2, 0, 0	10, 5, 7	
CTT, TTC, TCT			4, 5, 0	
GGG		1	5	1
GGT, GTG, TGG	5, 7, 0	0, 1, 1	0, 9, 2	0, 9, 0
GTT, TTG, TGT			4, 3, 2	
TTT		1	6	
Subtotal	26	465	297	14

We recognized the potential InDel by selecting one single amino-acid residue inserted or deleted surrounded by perfectly aligned 10 residues on each side. 17,234 human reference-proteins and 4,816,479 human ESTs were used for tblastn searches as noted in Materials and methods. Insertion/deletion of a single amino-acid residue in the original human reference-protein sequences was summarized and their coding nucleotide sequences are listed in this table according to their genomic position related to the exon–intron boundaries. The three stop codons were not considered in the search algorithm. We observed 297 InDels occurred in the middle of exons, while most of InDels did occur at the splicing junctions (491 sites).

[a] Stop codons.

[b] 14 InDel sites cannot be determined due to various issues: duplicated sequences at the ends of adjacent exons, incomplete genomic sequence information, etc.

or 'AG' at the InDel-coding sites (Table 1), which led us to examine the precise InDel location in the genomic sequence. A total of 491 of 802 InDels belonged to this intron–exon boundary location type, and almost all of these InDels were located at the splice-acceptor site (465 out of 491 cases). Most of these splice sites were typical GT-AG pairs, as demonstrated by the high degree of GT or AG representation. Only 24 out of 491 sites contained the non-canonical splicing pairs. We observed a total of only 26 InDels located at the splice-donor site and the remaining 14 cases were not able to be assigned to either the splice-donor or -acceptor site due to the highly homology of intron–exon junction sequences at both ends (Table 1). As shown in Figs. 1B and C, intron–exon boundary InDels were validated by RT-PCR in several cell lines. These InDel variants were then validated by DNA and mRNA sequencing-verification procedures to confirm the origins of InDel sequences. Although no three base-pair genomic DNA insertion/deletion was observed in these samples, one could still observe the co-existence of typical and InDel RT-PCR isoforms for almost all cell lines examined, as long as this gene had been expressed (Figs. 1B and C).

The observation of the presence of tandem GTNgt and agNAG at intron–exon junctions in combination with our additional experimental results below indicated that these InDel variants might be generated mostly from a wobble-splicing mechanism (Supplement Figure 2). Due to the relatively close proximity of splice-donor or -acceptor signal sequences (GTNgt or agNAG), a possible slippage of three base-pairs at splicing junctions could lead to the generation of such single amino-acid InDel coding transcripts. Such spliceosome slips around the splice junction were first described in mouse transcriptome by Zavolan and 161 sites were found with possible 3 nt alternative splice at 3′-splice-acceptor sites [18,19].

Wobble-splicing observation at adjacent splicing junctions in a human gene

Alternative mRNA splicing would appear to be a complex mechanism that occurs in order to generate mRNA

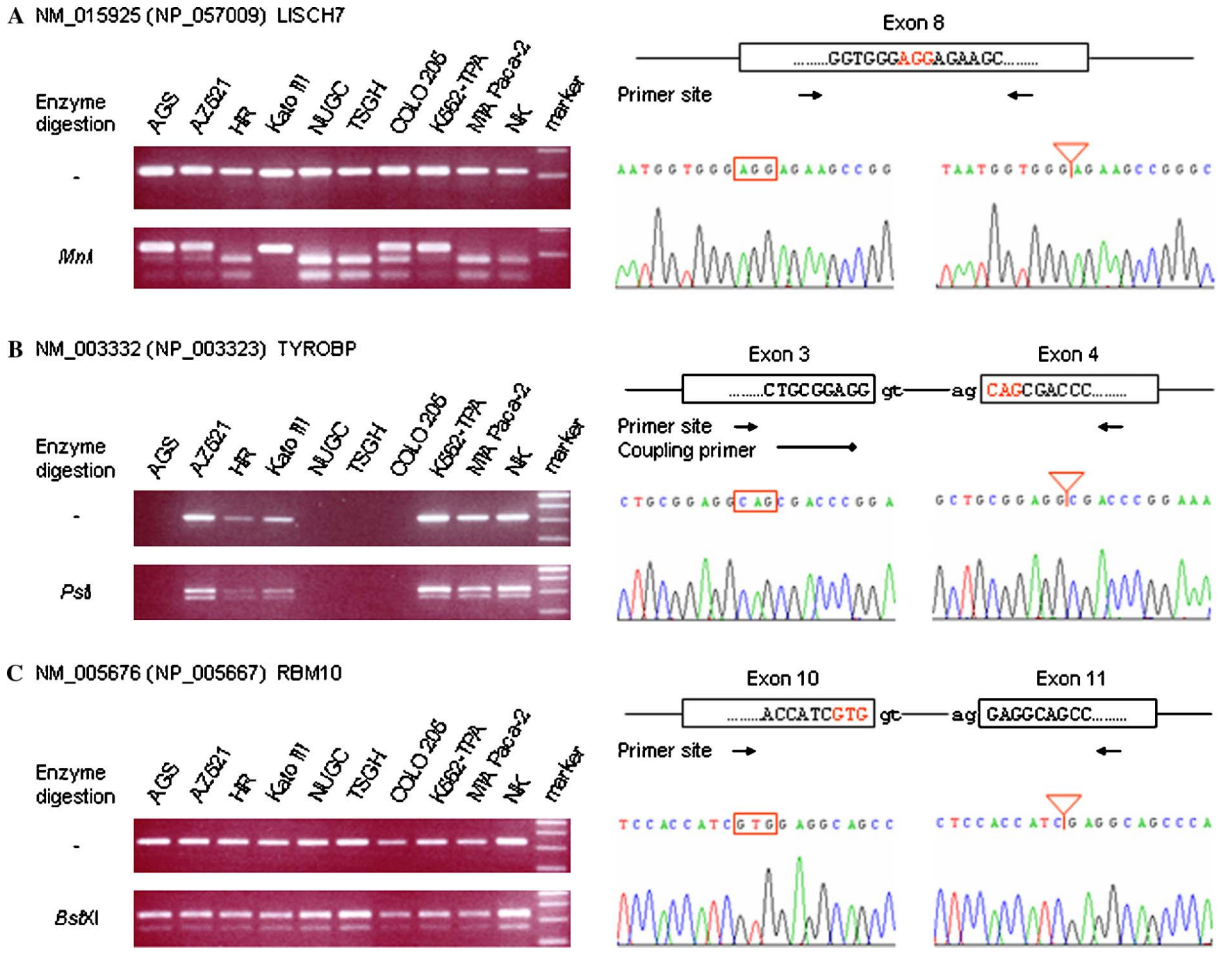
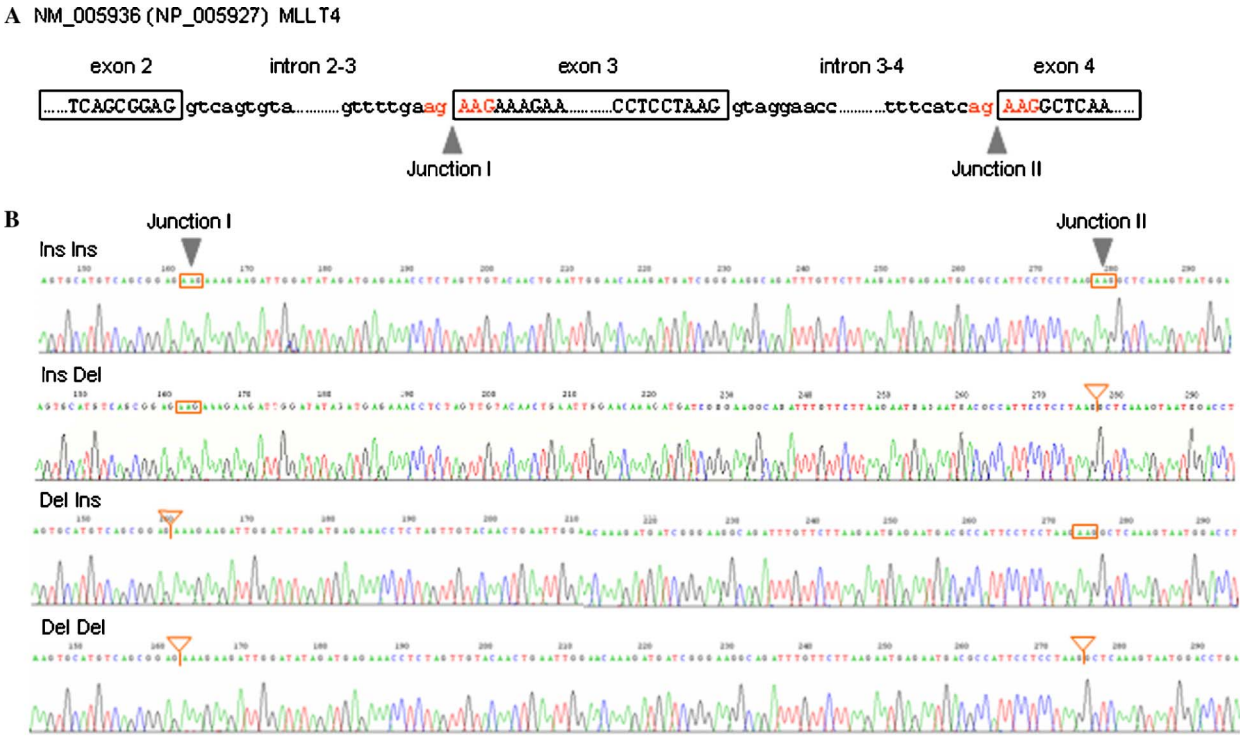


Fig. 1. Three types of InDel variants were discovered and classified according to their genomic locations. (A) A three base-pair (AGG) insertion or deletion arising in the middle of exon 8 of LISCH7 gene (NM_015925/NP_057009). RT-PCR products were screened by *Mnl*I restriction-enzyme digestion to confirm the InDel transcripts. (B) A three base-pair (CAG) insertion or deletion arising at the beginning of exon 4 of TYROBP (NM_003332/NP_003323). The formation of an agCAG tandem splice-acceptor site was noted, with intronic sequences marked in lowercase while exonic sequences are marked in uppercase. RT-PCR products were screened by *Pst*I digestion together with a third restriction site-modifying coupling primer to generate the *Pst*I cutting site. (C) A three base-pair (GTG) insertion or deletion arising at the end of exon 10 of RBM10 (NM_005676/NP_005667). RT-PCR products were screened by *Bst*XI digestion. The formation of a GTGgt tandem splice donor site can be noted.

diversity as well as proteomic complexity within the human genome [20]. We believe such wobble-splicing mechanism exists within many genes and that this phenomenon has not been systematically explored until the analysis of the increasing contents of biological information databases [21,22]. To further demonstrate the existence of wobble-splicing within human genes, we selected a human MLLT4 gene NM_005936 (NP_005927) with two adjacent wobble-splicing sites at exon 3 and exon 4 (junctions I and II as depicted in Fig. 2A). Subsequent to RT-PCR amplification of such junctions I and II from a single cell line, we isolated 95 clones and screened them by restriction-enzyme patterns as well as sequence verification (Fig. 2B). Four different types of clones were observed with most of the resultant transcripts containing the AAG insertion within at least one junction (87 out of 95). To the best of our knowledge, this is the first experimental data to report that the wobble-splicing event operates independently at adjacent exons within a particular gene, and the identification of all four types of splicing isoforms in a single cell line reinforces the concept of a wobble-splicing event.

The regulation of such subtle splicing junction site selection may, in fact, change the distribution of different InDel variants and may thus feature possible functional or disease associations. Splicing-regulating elements, however, still clearly warrant further research. Only few splicing-related enhancers appear to have been investigated at the whole exon inclusion/exclusion level [23] and the precise elements regulating the wobble-splicing process yet remain to be clearly elucidated [24]. Recent reports published in 2002 and 2004 by Fairbrother et al. and Wang et al. [25–27] have revealed the existence of exonic-splicing enhancer sequences and also the relative importance of serine-arginine-rich (SR) protein-recognition events. The relationship between the location of such wobble-splicing



C

PCR clones from AGS cell line and classified by restriction pattern (total=95 colonies)

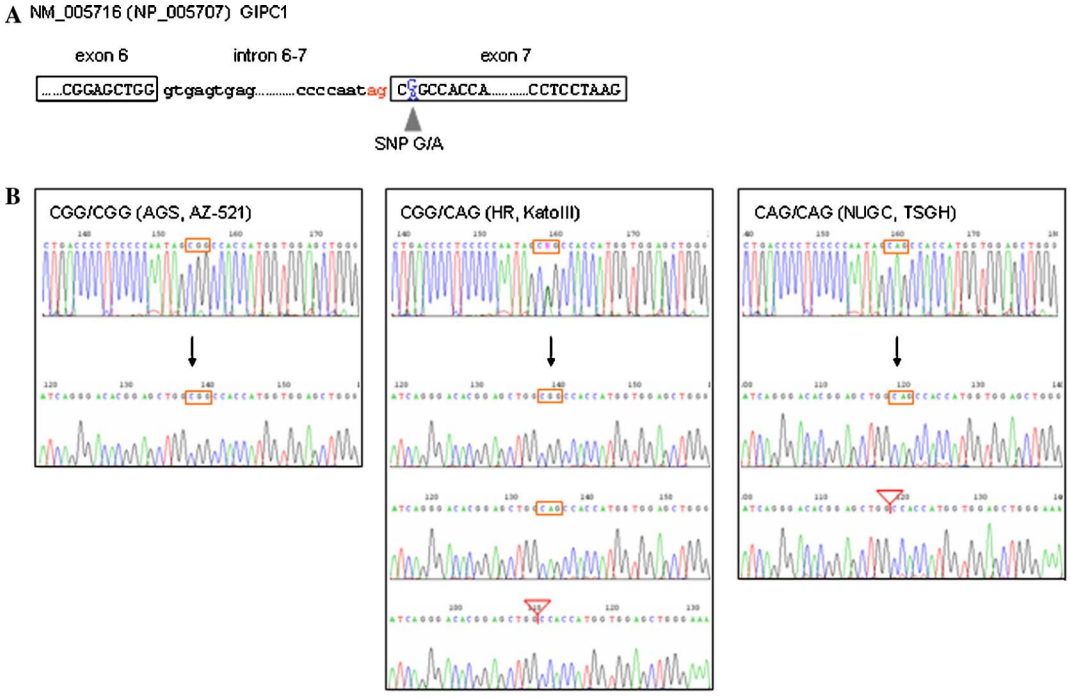
Junction		
I	II	Number
AAG	AAG	36
AAG	--	25
--	AAG	26
--	--	8

Fig. 2. Wobble-splicing transcript verification within the MLLT4 (NM_005936/NP_005927) gene. (A) Junction I and II sequences of exon 3 and exon 4 in the MLLT4 gene. Two adjacent possible wobble-splicing sites at exon 3 and exon 4 with agAAG wobble-splicing sequences can be noted by arrowheads. (B) Four types of wobble-splicing transcripts and their nucleotide sequence-representative chromatograms: type I (Ins Ins) with both junctions I and II containing AAG sequences in red colored closed boxes; type II (Ins Del) and type III (Del Ins) with only one junction containing AAG; type IV (Del Del) transcript without any AAG sequences. (C) Numbers of PCR clones identified for each type of wobble-splicing variant were listed. Following RT-PCR amplification of the junction I and II regions from a human gastric-cancer cell line (AGS), 95 clones were isolated randomly and screened by restriction-enzyme digestion as well as by sequencing validation.

and its neighboring sequences clearly needs to be carefully examined.

InDel variants generated by SNP conversion through wobble-splicing

One profound implication of wobble-splicing and its relationship to cSNPs is that the latter may feature a more observable impact upon protein diversity and human diseases [28]. We previously identified a SNP at a splicing-acceptor site (ag to ac) resulting in an in-frame insertion of 36 bp intronic sequences [29]. Now, with the wobble-splicing mechanism herein, it remains a possibility that a new type of functional SNP located at splicing junctions could be converted from a synonymous cSNP or intronic SNP to a functionally significant coding single amino-acid InDel. Simply conducting genomic DNA genotyping would, however, not be able to distinguish between the potential different wobble-splicing transcripts for a given individual. Thus, our finding would appear to be beneficial in that it highlights a mechanism to link previously non-functional SNPs to potential InDels for specific diseases and gene-association studies. As illustrated in Fig. 3, the GIPC1 gene NM_005716 (NP_005707) carries a synonymous SNP (dbSNP: rs1127307) at codon 284 (GCG and GCA) and this particular SNP is located at the intron–exon junction at 5′ of exon 7 (agC[G/A]G). We have observed possible InDel wobble-splicing transcripts in this junction. Sequence validation of this junction was performed by RT-PCR and genomic DNA

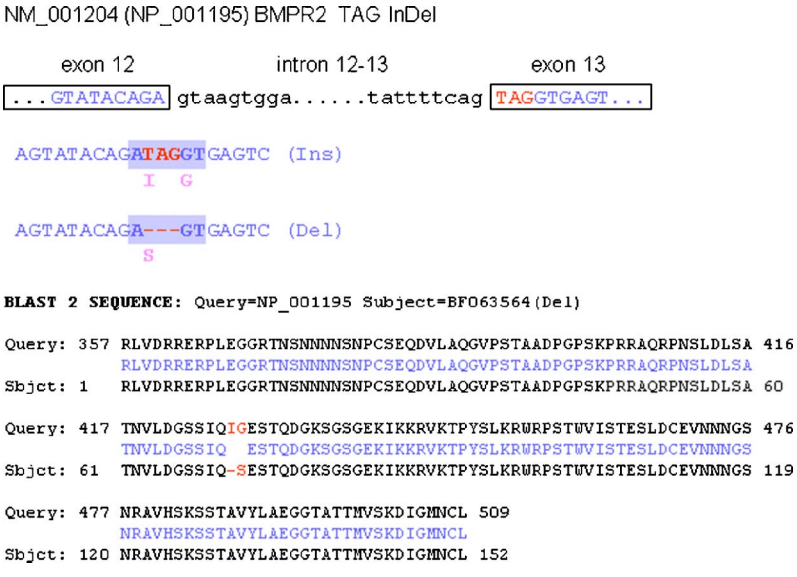


C

PCR clones identified by restriction pattern of three representative cell lines

		Putative transcript		
		CGG	CAG	Del*
Genotype:	CGG/CGG (AGS)	376	0	0
	CGG/CAG (HR)	184	190	5
	CAG/CAG (NUGC)	0	185	3

Fig. 3. Identification of genomic sequence and wobble-splicing transcripts of GIPC1 (NM_005716/NP_005707). (A) Wobble-splicing junction of exon 6–7 of GIPC1 gene with a synonymous G/A SNP (dbSNP: rs1127307) site for codon 284 (GC**G**/GC**A**). (B) Following analysis of human gastric-cancer cell lines, two cell lines (AGS and AZ521) feature CGG/CGG genotypes in their genome sequence as shown in the top panel. The only transcript generated was a CGG transcript. Two cell lines (HR and KatoIII) feature the CGG/CAG heterozygous genotype in their genomic sequence. Three transcripts can be generated including the wobble-splicing variant with one amino-acid coding sequence deletion as depicted in the bottom panel. Two cell lines (NUGC and SC-M1) feature the CAG/CAG homozygous genotype. Two different transcripts have been generated including one wobble-splicing variant. (C) Numbers of PCR clones screened by restriction-enzyme digestion. *Del indicates the wobble-splicing transcript detected at a very low frequency in the CAG genotype.



PCR using six gastric cancer cell lines [10]. In Fig. 3, three different genomic genotypes were discovered following genomic DNA PCR and direct sequencing (agCGG/agCGG and agCAG/agCAG homozygous cells as well as agCGG/agCAG heterozygous genotypes). Only cells carrying the agCAG genotype would appear to be able to generate the apparent wobble-spliced isoforms, with a low frequency less than 2% of total clones (Figs. 3B and C). In total, three additional SNP records were found located in our InDel sites by dbSNP cross-examination (rs2425068, rs1558876, and rs11509437, Supplement Figures 3A and B). These sequences increased the complexity of splicing site determination in different scenarios (Supplement Figure 3).

Genomic scanning of GTNGT and AGNAG junction sites

In order to learn more about the distribution of potential wobble-splicing variants within the human genome, we used the human genome assembly sequences for exploration purposes and systematically searched, in particular, for agNAG or GTNgt intron–exon boundary sequence patterns. Among 322,390 intron–exon boundaries examined, 7983 agNAG sites and 3215 GTNgt possible wobble-splicing sites were observed to be located at their respective intron–exon boundaries (Supplement Table 1). A part of these putative sites were sequence-confirmed majorly only at agNAG sites in previous reports [21,22]. In addition to single amino-acid coding InDels, such three base-pair wobble-spliced transcripts can generate dipeptide variants with one altered amino-acid plus one amino-acid InDel under different codon usage, so as to thus create greater disease-implicating mutations within certain human proteins (Fig. 4).

The recent tandem splicing acceptor studies completed in 2004 by Hiller et al. and 2005 by Tadokoro et al. [21,22] demonstrated the genome-wide distribution of NAGNAG sequences at human splice-acceptor sites, identified NAG-based single amino-acid InDel variants, and demonstrated that some with tissue-specific expression pattern. This NAGNAG-based tandem repeat would also appear to be the major wobble-splicing site observed in our study, and our results also implicated the additional GTNGT splice-donor site as well as the non-canonical GT-AG splicing signatures as sites that can be used to generate InDels in addition to the bona fide genomic DNA InDel events (297 sites). Based upon the distribution of canonical and non-canonical splice sites as previously examined for mammalian genes [30], it is evident that there exists GT-AG bias distribution at the intron–exon junction and a higher frequency of junction-based GTNGT and AGNAG tandem signals among wobble-splicing variants than is the case for non-GT-AG junctions. It would be interesting to further explore the adjacent nucleotide distribution frequency in order to learn more about the splice-site choice and selection process. We have found a tendency toward AGAAG and AGGAG in the splice-acceptor sites in comparing the AGNAG (nucleotide after first AG) and AGNAGN (nucleotide after the second AG) sequence usages (Supplement Figure 4).

The exposition of the possible wobble-splicing mechanism indicates the likelihood of an even higher degree of diversity of human gene transcripts than was previously thought to be the case [31]. Such a finding should be considered in the context of genomic annotation to provide a comprehensive gene-organization procedure, such as precise exon-boundary assignment. The presence of such a single amino-acid InDel has, previously, been noted to be contained within the human AF-6 gene (MLLT4, NM_005936) as an alternative spliced transcript [32], similar splicing transcripts also having been found for the mouse EST dataset in an orthologous gene (data not shown). Therefore, wobble-splicing transcripts have been observed not only within ESTs from many different human cDNA libraries, but also within libraries corresponding to different mouse developmental stages [15]. In most cases previously, these tandem spliced isoforms were often overlooked, since they appeared to represent only a minor sub-population of the total gene-transcript population, which need to be carefully examined with restriction enzyme and PCR-based enrichment procedures (Supplement Figure 5).

Multiple wobble spliced isoforms by adjacent non-tandem splicing site selection

We also observed that the wobble-splicing that occurred separately, as three and nine base-pair deletion EST transcripts, which might thus be classified as the cryptic splicing site choice in addition to the three base-pair wobble-splicing mechanism (Supplement Figure 5). Based upon our observations on the generation of the micro InDels around the splice junction, many factors should be considered as the annotation to be precise in syncopation before further elucidation, such as the genomic DNA alteration events by SNP or InDel, sequence identity around both splicing donor and acceptor region, and their interplay

Fig. 4. Additional dipeptide variant detected with one amino-acid InDel and one altered amino-acid in BMPR2 gene (NM_001204/NP_001195). Following the scanning of specific sequences of the human genome and intron–exon boundaries, a bioinformatic tool was developed in order to detect such wobble-splicing variants. dbEST_human was employed to validate the existence of these variants. In total, 51 out of 1184 predictions were confirmed with EST matches, as illustrated here. The intron–exon junctions can be seen at the top; the coding-sequence alterations (IG to S) were listed with the removal, by wobble-splicing, of a TAG sequence in exon 13. A total of 14 ESTs were matched with IG coding sequences and three ESTs revealing the deletion and S-coding changes. The detailed BLAST alignment is listed in the bottom panel.

with canonical and non-canonical splice signatures. We thus hypothesize that the wobble-splicing phenomenon discussed herein could be generated by a mis-splicing event in addition to alternative splice-site selections. We believe that most of these wobble-splicing transcripts are generated by the mis-splicing event. Since the mRNA-surveillance mechanism [33] might not be sufficiently effective to eliminate these in-frame wobbled mis-splicing variants, these in-framed spliced transcripts may be able to be processed and exported as translation templates. mRNA surveillance, also known as nonsense-mediated mRNA decay (NMD), is an mRNA quality-control mechanism that degrades abnormal mRNAs such as mis-spliced mRNA transcripts [34]. By recognizing mRNAs containing premature termination codons, NMD eliminates the production of the encoded truncated protein coded by mis-spliced transcripts [34], presumably because the truncated protein could function to the detriment of cells.

In conclusion, we have identified over 800 possible single amino-acid InDel variants in human proteome and demonstrated a wobble-splicing mechanism that appears to generate this new type of functional InDel in transcriptome in addition to genomic DNA three base-pair InDels. This wobble-splicing event appears to occur at the boundaries of intron–exon splicing junctions by a process of selecting alternative adjacent splicing signal sequences. It is also possible that this process could convert non-functional nucleotide polymorphisms to functional InDel protein variants in order to generate a higher complexity of translated proteins. This would have great implication in genome diversity and disease associations.

Acknowledgments

The authors thank Jian-Yuan Chiu and Gu-Liang Wang for their excellent technical assistance in computer programming. We are grateful to the critical suggestions of Drs. Lloyd A. Culp and Dan Robinson in the preparation of this manuscript. This work was supported, in part, by Academia Sinica research project grants and National Science Council grants (94-2311-B-001-033, 93-2311-B-001-057, and 92-2311-B-001-097) that had been awarded to Wen-chang Lin.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.01.101.

References

[1] J.D. Watson, The human genome project: past, present, and future, Science 248 (1990) 44–49.
[2] R. Sachidanandam, D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers, D.R. Bentley, P.Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, D. Altshuler, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928–933.
[3] K.M. Weiss, In search of human variation, Genome Res. 8 (1998) 691–697.
[4] D. Altshuler, J.N. Hirschhorn, M. Klannemark, C.M. Lindgren, M.C. Vohl, J. Nemesh, C.R. Lane, S.F. Schaffner, S. Bolk, C. Brewer, T. Tuomi, D. Gaudet, T.J. Hudson, M. Daly, L. Groop, E.S. Lander, The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes, Nat. Genet. 26 (2000) 76–80.
[5] C.H. Lai, C.Y. Chou, L.Y. Ch'ang, C.S. Liu, W. Lin, Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics, Genome Res. 10 (2000) 703–713.
[6] L.Y. Hu, C.H. Lai, J.Y. Chiu, W.C. Lin, Functional variants/non-synonymous cSNP discovery by the CGI bioinformatic tool, J. Genet. Mol. Biol. 16 (2005) 56–62.
[7] D. Altshuler, V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, E.S. Lander, An SNP map of the human genome generated by reduced representation shotgun sequencing, Nature 407 (2000) 513–516.
[8] C.M. Everett, N.W. Wood, Trinucleotide repeats and neurodegenerative disease, Brain 127 (2004) 2385–2405.
[9] G. Condorelli, R. Bueno, R.J. Smith, Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics, J. Biol. Chem. 269 (1994) 8510–8516.
[10] H.W. Kao, H.C. Chen, C.W. Wu, W.C. Lin, Tyrosine-kinase expression profiles in human gastric cancer cell lines and their modulations with retinoic acids, Br. J. Cancer 88 (2003) 1058–1064.
[11] Z. Kan, E.C. Rouchka, W.R. Gish, D.J. States, Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, Genome Res. 11 (2001) 889–900.
[12] B. Modrek, C. Lee, A genomic view of alternative splicing, Nat. Genet. 30 (2002) 13–19.
[13] N. Volfovsky, B.J. Haas, S.L. Salzberg, Computational discovery of internal micro-exons, Genome Res. 13 (2003) 1216–1221.
[14] T.A. Thanaraj, A.J. Robinson, Prediction of exact boundaries of exons, Brief Bioinform. 1 (2000) 343–356.
[15] T.A. Thanaraj, S. Stamm, Prediction and statistical analysis of alternatively spliced exons, Prog. Mol. Subcell Biol. 31 (2003) 1–31.
[16] M.S. Taylor, C.P. Ponting, R.R. Copley, Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes, Genome Res. 14 (2004) 555–566.
[17] F.S. Collins, Cystic fibrosis: molecular biology and therapeutic implications, Science 256 (1992) 774–779.
[18] M. Zavolan, S. Kondo, C. Schonbach, J. Adachi, D.A. Hume, Y. Hayashizaki, T. Gaasterland, Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome, Genome Res. 13 (2003) 1290–1300.
[19] M. Zavolan, E. van Nimwegen, T. Gaasterland, Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome, Genome Res. 12 (2002) 1377–1385.
[20] B.R. Graveley, Alternative splicing: increasing diversity in the proteomic world, Trends Genet. 17 (2001) 100–107.
[21] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, M. Platzer, Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity, Nat. Genet. (2004).

[22] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, M. Ozaki, M. Ono, N. Miki, T. Miyashita, M. Yamada, Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products, J. Hum. Genet. 50 (2005) 382–394.
[23] H. Miyaso, M. Okumura, S. Kondo, S. Higashide, H. Miyajima, K. Imaizumi, An intronic splicing enhancer element in survival motor neuron (SMN) pre-mRNA, J. Biol. Chem. 278 (2003) 15825–15831.
[24] C.W. Smith, J. Valcarcel, Alternative pre-mRNA splicing: the logic of combinatorial control, Trends Biochem. Sci. 25 (2000) 381–388.
[25] W.G. Fairbrother, G.W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P.A. Sharp, C.B. Burge, RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons, Nucleic Acids Res. 32 (2004) W187–W190.
[26] Z. Wang, M.E. Rolish, G. Yeo, V. Tung, M. Mawson, C.B. Burge, Systematic identification and analysis of exonic splicing silencers, Cell 119 (2004) 831–845.
[27] W.G. Fairbrother, R.F. Yeh, P.A. Sharp, C.B. Burge, Predictive identification of exonic splicing enhancers in human genes, Science 297 (2002) 1007–1013.
[28] D.L. Black, Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology, Cell 103 (2000) 367–370.
[29] C.H. Lai, J.Y. Chiu, W. Lin, Identification of the human crooked neck gene by comparative gene identification, Biochim. Biophys. Acta 1517 (2001) 449–454.
[30] M. Burset, I.A. Seledtsov, V.V. Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes, Nucleic Acids Res. 28 (2000) 4364–4375.
[31] P.A. Sharp, RNA splicing and genes, JAMA 260 (1988) 3035–3041.
[32] S. Saito, M. Matsushima, S. Shirahama, T. Minaguchi, Y. Kanamori, M. Minami, Y. Nakamura, Complete genomic structure DNA polymorphisms, and alternative splicing of the human AF-6 gene, DNA Res. 5 (1998) 115–120.
[33] M.J. Moore, Nuclear RNA turnover, Cell 108 (2002) 431–434.
[34] L.E. Maquat, Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics, Nat. Rev. Mol. Cell Biol. 5 (2004) 89–99.