nature genetics

Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity

Michael Hiller[1,4], Klaus Huse[2,4], Karol Szafranski[2], Niels Jahn[2], Jochen Hampe[3], Stefan Schreiber[3], Rolf Backofen[1] & Matthias Platzer[2]

Splice acceptors with the genomic NAGNAG motif may cause NAG insertion-deletions in transcripts, occur in 30% of human genes and are functional in at least 5% of human genes. We found five significant biases indicating that their distribution is nonrandom and that they are evolutionarily conserved and tissue-specific. Because of their subtle effects on mRNA and protein structures, these splice acceptors are often overlooked or underestimated, but they may have a great impact on biology and disease.

Alternative splicing is a main source of transcriptome and proteome diversity and is therefore relevant to disease and therapy[1]. A scan of 20,213 human mRNAs from the RefSeq division of GenBank found that 5% (8,105 of 152,288) of the splice acceptors contained a NAGNAG motif (N stands for A, C, G or T). In these structures, the upstream AG is called the E acceptor, which gives rise to the E transcript, because part of the tandem will be exonic; in comparison, the whole tandem is intronic for the I acceptor (**Fig. 1a**). Of these tandem acceptors, 627 and 152 belong to introns that are exclusively located in the 5′ and 3′ untranslated regions, respectively. We focused on the 7,326 NAGNAG acceptors that are situated upstream of an exon annotated as part of the protein coding sequence (**Table 1**). Alternative splicing for 40 NAGNAGs was indicated in the two respective RefSeq entries. Searching dbEST yielded support for 791 tandem acceptors. After eliminating redundancies, we identified 878 experimentally confirmed tandem acceptors. Thus, of 20,213 RefSeq transcripts, 30% (6,004) contain at least one observed NAGNAG acceptor and 5% (1,054) contain at least one confirmed NAGNAG acceptor in the coding sequence (**Supplementary Methods** and **Supplementary Table 1** online).

Tandem acceptors are biased towards intron phase 1 (40% phase 0, 43% phase 1 and 17% phase 2; **Supplementary Table 2** online), which

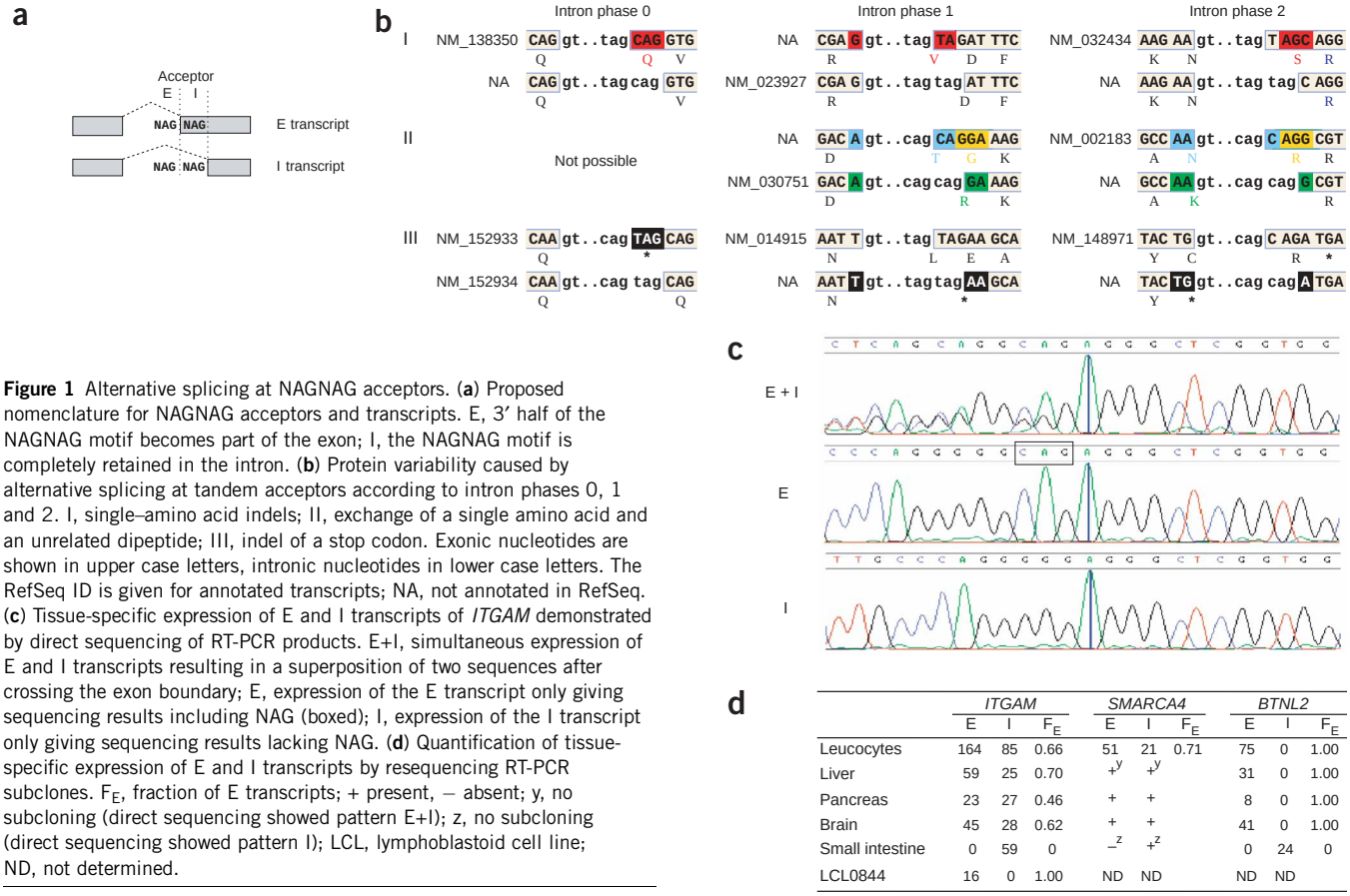
Table 1 NAGNAG acceptors in human RefSeq coding sequence

		Confirmed							
		mRNA			EST			mRNA and EST	
Motif	Observed	E	I	E+I	E	I	E+I	E+I	
AAGAAG	54	50	5	1	33	15	15	20*	37%
AAGCAG	164	46	122	4	72	112	62	69	42%
AAGGAG	199	199	0	0	164	4	4	4	2%
AAGTAG	32	7	27	2	12	21	11	15	47%
CAGAAG	888	868	25	5	727	96	92	104	12%
CAGCAG	720	509	233	22	500	366	302	343	48%
CAGGAG	2,882	2,874	9	1	2,468	39	38	41	1%
CAGTAG	96	71	26	1	65	35	26	28	29%
GAGAAG	9	2	7	0	1	7	1	1	11%
GAGCAG	227	6	221	0	10	190	9	10	4%
GAGGAG	15	10	5	0	8	4	2	2	13%
GAGTAG	45	0	45	0	2	37	2	2	4%
TAGAAG	366	357	10	1	312	55	55	59	16%
TAGCAG	258	201	60	3	203	147	136	142	55%
TAGGAG	1,334	1,332	2	0	1,153	17	17	18	1%
TAGTAG	37	27	10	0	28	21	19	20	54%
Sum	7,326	6,559	807	40	5,758	1,166	791	878	13%

*Acceptors were also considered confirmed if one acceptor has an EST hit and the other is annotated in RefSeq.
Observed NAGNAG motifs and E and I acceptors confirmed by mRNA (from RefSeq) or EST hits are shown. The percentages indicate the portion of observed NAGNAG acceptors that is confirmed. NAGGAG and GAGNAG preferably function as E and I acceptors, respectively (**Supplementary Note** online). YAGYAG is the most efficient tandem acceptor (**Supplementary Note** online).

[1]Institute of Computer Science, Friedrich-Schiller-University Jena, Chair for Bioinformatics, Ernst-Abbe-Platz 2, 07743 Jena, Germany. [2]Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstr. 11, 07745 Jena, Germany. [3]Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstr. 12, 24105 Kiel, Germany. [4]These authors contributed equally to this work. Correspondence should be addressed to M.P. (mplatzer@imb-jena.de).

Published online 31 October 2004; doi:10.1038/ng1469



	ITGAM			SMARCA4			BTNL2		
	E	I	F_E	E	I	F_E	E	I	F_E
Leucocytes	164	85	0.66	51	21	0.71	75	0	1.00
Liver	59	25	0.70	$+^y$	$+^y$		31	0	1.00
Pancreas	23	27	0.46	+	+		8	0	1.00
Brain	45	28	0.62	+	+		41	0	1.00
Small intestine	0	59	0	$-^z$	$+^z$		0	24	0
LCL0844	16	0	1.00	ND	ND		ND	ND	

Figure 1 Alternative splicing at NAGNAG acceptors. (**a**) Proposed nomenclature for NAGNAG acceptors and transcripts. E, 3′ half of the NAGNAG motif becomes part of the exon; I, the NAGNAG motif is completely retained in the intron. (**b**) Protein variability caused by alternative splicing at tandem acceptors according to intron phases 0, 1 and 2. I, single–amino acid indels; II, exchange of a single amino acid and an unrelated dipeptide; III, indel of a stop codon. Exonic nucleotides are shown in upper case letters, intronic nucleotides in lower case letters. The RefSeq ID is given for annotated transcripts; NA, not annotated in RefSeq. (**c**) Tissue-specific expression of E and I transcripts of *ITGAM* demonstrated by direct sequencing of RT-PCR products. E+I, simultaneous expression of E and I transcripts resulting in a superposition of two sequences after crossing the exon boundary; E, expression of the E transcript only giving sequencing results including NAG (boxed); I, expression of the I transcript only giving sequencing results lacking NAG. (**d**) Quantification of tissue-specific expression of E and I transcripts by resequencing RT-PCR subclones. F_E, fraction of E transcripts; + present, − absent; y, no subcloning (direct sequencing showed pattern E+I); z, no subcloning (direct sequencing showed pattern I); LCL, lymphoblastoid cell line; ND, not determined.

is significantly different ($P < 0.0001$ by χ^2 test) from all human introns (46% phase 0, 33% phase 1 and 21% phase 2)[2]. The intron phase determines the outcome of a NAG insertion-deletion (indel) and includes single–amino acid indels (Gln, Glu and Lys in phase 0; Ala, Glu, Gly and Val in phase 1; and Arg and Ser in phase 2), exchange of single amino acid and an unrelated dipeptide or the creation or destruction of a stop codon (**Fig. 1b**). Ten confirmed NAGNAGs create or destroy a stop codon (**Supplementary Note** online). Notably, confirmed NAGNAGs in phase 1 and phase 2 are significantly enriched in single–amino acid indels ($P = 0.0009$ and $P = 0.0007$, respectively, by Fisher's exact test; **Supplementary Note** online). We suppose that single–amino acid indels are functionally more compatible than an exchange of a single amino acid for two unrelated amino acids. Nevertheless, an indel of a charged amino acid (Arg, Glu or Lys) might be an important event for a protein. The ten amino acids on each side of confirmed Glu indels were significantly enriched in similarly charged amino acids ($P = 0.0046$ by *t*-test). For Arg and Lys we observed the same trend (**Supplementary Note** online). These findings further support our view that tandem acceptors evolved to introduce subtle protein changes.

Tandem acceptors are not restricted to the human genome and occur in ruminants[3], chicken[4] and tomato[5], among other organisms. An investigation of the RefSeq and EST databases for *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans* showed that NAGNAG acceptors are frequent in other genomes, too (**Supplementary Table 3** online). The number of RefSeqs and ESTs (expressed sequence tags) is comparable in *C. elegans* and *D. melanogaster*, but *C. elegans* has a relatively low fraction of confirmed tandem acceptors. This may reflect unique features of the splicing machinery in *C. elegans*[6], whose introns typically lack both branch point and polypyrimidine tract consensus sequences, preventing an extensive utilization of tandem acceptors in this species. We further determined whether purifying selection acts on NAGNAG acceptors. We found that 73% of orthologous NAGNAG acceptor pairs from human and mouse were conserved (**Supplementary Table 4** and **Supplementary Note** online). This high conservation rate is consistent with the observation that NAGNAG acceptors represent nearly half of all human-mouse conserved alternative 3′ splice sites[7].

Alternative splicing is often controlled in a tissue-specific or developmental stage–specific manner[8]. Tissue or cell specificity of alternative splicing at tandem acceptors would be an indication that this process is biologically important and not the result of imprecision in the splice machinery. To address this question, we showed *in silico* that 5 of 15 tandem acceptors with the highest numbers of EST hits in dbEST are tissue-specific and showed experimentally that NAGNAG acceptors of *ITGAM*, *SMARCA4* and *BTNL2* are tissue-specific (**Fig. 1c,d** and **Supplementary Note** online).

Polar residue hot spots have been observed at protein-protein binding sites[9]. The ten amino acids on each side of confirmed NAGNAG exon junctions are significantly more polar than those on each side of non-NAGNAG junctions ($P < 0.0001$ by *t*-test). The database of interacting proteins and 'RNA recognition' Pfam members are enriched in proteins encoded by transcripts with confirmed tandem acceptors ($P < 0.0001$ by χ^2 test and $P = 0.025$ by binomial test; **Supplementary Table 5** and **Supplementary Note** online). Accordingly, several genes involved in splicing (*e.g.*, *PRPF3*, *PRPF8*, *U2AF1*

and *U2AF2*) are equipped with tandem acceptors that are conserved between human, mouse and rat. Moreover, tandem acceptors raise interesting questions about the 3′ splice site AG selection[10]. Their alternative recognition requires some flexibility of the interactions of branch point, polypyrimidine tract and splice AG with splice factors. This flexibility of the splice machinery may be enhanced by isoforms of its protein components, such as *U2AF1* and its interacting partner *U2AF2*. Tandem acceptor–derived isoforms of *U2AF* subunits might promote flexibility in the spatial architecture of the spliceosome, with functional consequences for the splicing process and even for splicing at NAGNAG acceptors themselves.

In contrast to alternative splicing in general, which often severely affects the protein structure[11], tandem acceptors provide a mechanism to create subtle changes. These changes may nonetheless be of functional relevance by changing local hydrophobicity and charge, varying the distances between relevant sites in proteins or changing recognition sequences for post-translational modifications. In the case of *IGF1R*, the two protein isoforms resulting from the change from Thr-Gly to Arg have different signaling activities and receptor-mediated internalization[12]. Moreover, one NAGNAG polymorphism is suspected to cause an abnormal phenotype[13]. The subtle effects of alternative splicing at tandem acceptors on multiple proteins simultaneously might be of particular importance in the pathogenesis of complex diseases. For instance, four of the six genes in which single-gene mutations are known to cause obesity contain NAGNAG acceptors (**Supplementary Note** online).

NAGNAG acceptors can increase proteome plasticity markedly, because many proteins may be affected simultaneously. Furthermore, 68 RefSeqs have more than one confirmed tandem acceptor that increases the number of possible protein variants (**Supplementary Table 6** online). Alternative splicing for the *D. melanogaster* gene *Dscam* results in 38,016 possible splice forms; this provides a potential mechanism for generating a unique cell identity[14]. Whether tandem acceptors are an alternative mechanism to individualize cells remains to be elucidated.

To summarize, we found *in silico* and experimental evidence that tandem acceptors may result in increased proteome diversity in a wide range of species, including mammals and fruit flies. Because of its minor effects on mRNA and protein structures, this phenomenon is frequently overlooked or underestimated in laboratory PCR experiments as well as in systematic genome-wide *in silico* analyses[7,15] and annotation efforts (*e.g.*, RefSeq). Therefore, we encourage further research focused on the prevalence, regulation and mechanism of alternative splicing at tandem acceptors; its functional effects on the affected proteins; and its impact on biology in general and human complex diseases in particular.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank I. Görlich and M.-L. Schmidt for technical assistance. This work was supported by grants from the German Ministry of Education and Research to S.S. and to M.P. as well as from the Deutsche Forschungsgemeinschaft to M.P.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 6 September; accepted 6 October 2004
Published online at http://www.nature.com/naturegenetics/

1. Garcia-Blanco, M.A., Baraniak, A.P. & Lasda, E.L. *Nat. Biotechnol.* **22**, 535–546 (2004).
2. Long, M. & Deutsch, M. *Mol. Biol. Evol.* **16**, 1528–1534 (1999).
3. Ferranti, P., Lilla, S., Chianese, L. & Addeo, F. *J. Protein Chem.* **18**, 595–602 (1999).
4. Rogina, B. & Upholt, W.B. *Biochem. Mol. Biol. Int.* **35**, 825–831 (1995).
5. Li, L. & Howe, G.A. *Plant Mol. Biol.* **46**, 409–419 (2001).
6. Zhang, H.B. & Blumenthal, T. *RNA* **2**, 380–388 (1996).
7. Sugnet, C.W., Kent, W.J., Ares, M. Jr. & Haussler, D. *Pac. Symp. Biocomput.* **2004**, 66–77 (2004).
8. Stamm, S. *et al. DNA Cell Biol.* **19**, 739–756 (2000).
9. Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. *Proc. Natl. Acad. Sci. USA* **100**, 5772–5777 (2003).
10. Chen, S., Anderson, K. & Moore, M.J. *Proc. Natl. Acad. Sci. USA* **97**, 593–598 (2000).
11. Resch, A. *et al. J. Proteome Res.* **3**, 76–83 (2004).
12. Condorelli, G., Bueno, R. & Smith, R.J. *J. Biol. Chem.* **269**, 8510–8516 (1994).
13. Karinch, A.M., deMello, D.E. & Floros, J. *Biochem. J.* **321**, 39–47 (1997).
14. Neves, G., Zucker, J., Daly, M. & Chess, A. *Nat. Genet.* **36**, 240–246 (2004).
15. Zavolan, M. *et al. Genome Res.* **13**, 1290–1300 (2003).