Single-Nucleotide Polymorphisms in NAGNAG Acceptors Are Highly Predictive for Variations of Alternative Splicing

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

[1]Institute of Computer Science, Chair for Bioinformatics, Friedrich-Schiller-University Jena, and [2]Genome Analysis, Leibniz Institute for Age Research–Fritz Lipmann Institute, Jena, Germany; and [3]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany

Aberrant or modified splicing patterns of genes are causative for many human diseases. Therefore, the identification of genetic variations that cause changes in the splicing pattern of a gene is important. Elsewhere, we described the widespread occurrence of alternative splicing at NAGNAG acceptors. Here, we report a genomewide screen for single-nucleotide polymorphisms (SNPs) that affect such tandem acceptors. From 121 SNPs identified, we extracted 64 SNPs that most likely affect alternative NAGNAG splicing. We demonstrate that the NAGNAG motif is necessary and sufficient for this type of alternative splicing. The evolutionarily young NAGNAG alleles, as determined by the comparison with the chimpanzee genome, exhibit the same biases toward intron phase 1 and single–amino acid insertion/deletions that were already observed for all human NAGNAG acceptors. Since 28% of the NAGNAG SNPs occur in known disease genes, they represent preferable candidates for a more-detailed functional analysis, especially since the splice relevance for some of the coding SNPs is overlooked. Against the background of a general lack of methods for identifying splice-relevant SNPs, the presented approach is highly effective in the prediction of polymorphisms that are causal for variations in alternative splicing.

SNPs, as the most abundant form of genetic variation, contribute significantly to phenotypic individuality and disease susceptibility. SNPs are mostly biallelic and are therefore easy to assay once they are described. Given their abundance in the human genome (~1 SNP every 300 bp [Ke et al. 2004]) and their ease of high-throughput typing, SNPs progressively replace microsatellites as first-choice genetic markers in association and linkage studies.

Much interest focuses on SNPs that are located in coding regions, since those SNPs may alter the protein sequence. However, SNPs can also influence splicing, which usually has a greater effect on the resulting protein than does the alteration of a single codon. Recently, splicing mutations have been suspected to be the most frequent cause of hereditary diseases (Lopez-Bigas et al. 2005). Accordingly, an increasing number of SNPs have been described that cause diseases by a change or disruption of the normal splicing pattern (for review, see Cartegni et al. [2002] and Garcia-Blanco et al. [2004]). These splice-relevant SNPs affect donor and acceptor splice sites, branch points, exonic as well as intronic splicing enhancers and silencers or alter important mRNA secondary structures. For example, the G allele of the silent coding SNP *rs17612648* in the *PTPRC* gene that is associated with multiple sclerosis destroys an exonic splicing silencer and abolishes the skipping of exon 4 (Lynch and Weiss 2001), and the SNP *rs2076530* in *BTLN2* that is associated with sarcoidosis leads to the activation of a cryptic donor site and a cryptic donor splice site 4 nt upstream (Valentonyte et al. 2005). Since the impact of SNPs on splicing is hard to predict *in silico* and is difficult to analyze experimentally, silent or intronic SNPs that may cause a phenotype or a disease by changing splicing patterns are often not investigated (Pagani and Baralle 2004). Thus, novel approaches are urgently needed to identify splice-relevant SNPs.

Recently, we reported the widespread occurrence of subtle alternative splice events that insert or delete the sequence NAG (N denotes A, C, G, or T) in mRNA (Hiller et al. 2004). This happens if both AG alleles of a NAGNAG acceptor can be chosen by the spliceosome. We termed the upstream acceptor in this tandem motif the "E acceptor" and the downstream one the "I acceptor." The products that arise from the use of E and I acceptors are called "E and I transcripts and proteins," respectively. The consequences of NAG insertion/deletions (indels) in mRNAs for the respective protein sequences are highly diverse and comprise eight different single–amino acid (aa) indel events, the exchange of a dipeptide and an unrelated aa, or the creation/destruction of a stop codon. Tandem acceptors are conserved

Received September 9, 2005; accepted for publication November 22, 2005; electronically published December 22, 2005.

Address for correspondence and reprints: Dr. Matthias Platzer, Genome Analysis, Fritz Lipmann Institute, Beutenbergstrasse 11, 07745 Jena, Germany. E-mail: mplatzer@fli-leibniz.de

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2006;78:291–302. 0002-9297/2006/7802-0011$15.00

Table 1

Correlation between Acceptor Genotypes and the Appearance of E and I Transcripts

dbSNP ID	Gene Symbol	Observations for Genotype: Homozygous NAGNAG, No. of Probands	Homozygous NAGNAG, cDNA Transcripts	Heterozygous, No. of Probands	Heterozygous, cDNA Transcripts	Homozygous Non-NAGNAG, No. of Probands	Homozygous Non-NAGNAG, cDNA Transcript
rs2245425	*TOR1AIP1*[a]	3	E+I	6	E+I	2	I
rs2275992	*ZFP91*[a]	1	E+I	7	E+I	4	E
rs1558876	*KIAA1001*	0	…	6	E+I	6	E
rs2290647	*KIAA1533*	0	…	4	E+I	8	E
rs4590242	*GABRR1*	11	E+I	1	E+I	0	…
rs1152522	*C14orf105*	0	…	0	…	12	I

Note.—E+I indicates presence of both E and I transcripts; E indicates only E transcripts; I indicates only I transcripts.

[a] See also figure 2.

between human and mouse, and the use of E or I acceptors can be controlled in a tissue-specific manner. Our results concerning the frequency and tissue specificity were confirmed by others (Tadokoro et al. 2005). Furthermore, E/I protein isoforms have functional differences (Condorelli et al. 1994; Tadokoro et al. 2005), and the SNP *rs1650232* within a NAGNAG acceptor is associated with respiratory-distress syndrome (Karinch et al. 1997).

Since NAGNAG acceptors occur in ~30% of human genes, we were interested in finding SNPs that may affect this type of alternative splicing. By scanning the SNP annotation of the human reference sequence, we identified those SNPs and provide experimental evidence of respective variations in the alternative splicing patterns. In addition, we introduce a classification for NAGNAG acceptors, with respect to their splicing plausibility, to bring forward a highly effective approach for predicting splice-relevant SNPs.

Methods

Identification of SNPs Affecting NAGNAG Acceptors

We downloaded the human genome assembly from the UCSC Genome Browser (UCSC Human Genome Browser, hg17, May 2004) as well as from RefSeq (refGene.txt.gz, January 12, 2005) and SNP annotations (snp.txt.gz, January 9, 2005). From the transcripts, we extracted a list of unique genomic positions of acceptor sites. We used the genomic position of the acceptors to select those SNPs that overlap the first 3 nt of an exon or the last 6 nt of an intron. Then, we evaluated whether one of both AG alleles or one of the two Ns in the NAGNAG pattern is polymorphic. SNPs are the only type of polymorphisms that were considered.

To check whether a tandem acceptor is EST confirmed, we used BLAST with a search string of 30 nt from the upstream exon and 30 nt from the downstream exon—taking the non-annotated acceptor into account—against the human fraction of the dbEST database (December 2004) and against the mRNA sequences downloaded from GenBank (December 2004). At most, one mismatch or one gap was allowed.

Comparison with the Chimpanzee Genome

We downloaded the chimpanzee genome working draft assembly from UCSC Genome Browser (UCSC Chimpanzee Genome Browser, panTro1, November 2003). We compared human polymorphic sites with the chimpanzee sequence, using BLAST, with 101-nt queries consisting of one of the SNP alleles, as well as 50 nt upstream and 50 nt downstream. Only hits with at least 95% identity and no other mismatch in the $-5\ldots+5$ context of the SNP were considered.

Null Model for Gain of NAGNAG Acceptors

Briefly described, we determined the ancestral allele variant for 2,439 SNPs that overlap an acceptor in the 9-nt context by comparing the genomic sequence context with the chimpanzee genome. In addition, we selected a set of 8,082 acceptor sites not affected by known SNPs. Then, the 2,439 SNPs were randomly assigned to one of those acceptors, given that the ancestral allele variant is present at the respective position. This position was replaced by the nonancestral allele, and we evaluated and counted the possible impact on a NAGNAG acceptor. More details are given in appendix A.

Experimental Verification of Alternative Splicing at Polymorphic NAGNAG Acceptors

Genomic DNA and cDNA from 12 whites were kindly provided by Gerd Birkenmeier (Leipzig) and were purified from whole blood by standard methods. First-strand cDNA was derived from oligo-dT primed reverse transcription.

For determination of the respective genotypes, ~20 ng of genomic DNA was used to PCR amplify the regions of the respective SNP through use of Ready-To-Go PCR beads (Amersham). PCR conditions were 1 cycle of denaturation at 95°C for 30 s; followed by 38 cycles of denaturing at 92°C for 30 s, annealing at 59°C for 30 s, and extension at 72°C for 60 s; and 1 cycle of final extension at 72°C for 5 min. PCR products were purified by precipitation and were sequenced with the same primers used for PCR amplification by the dye terminator method by use of BigDye v3.1 (Applied Biosystems). To identify E and I transcripts, cDNA from the genotyped individuals was amplified using the same PCR conditions with transcript-specific primers.

For amplification of genomic DNA and subsequent sequenc-

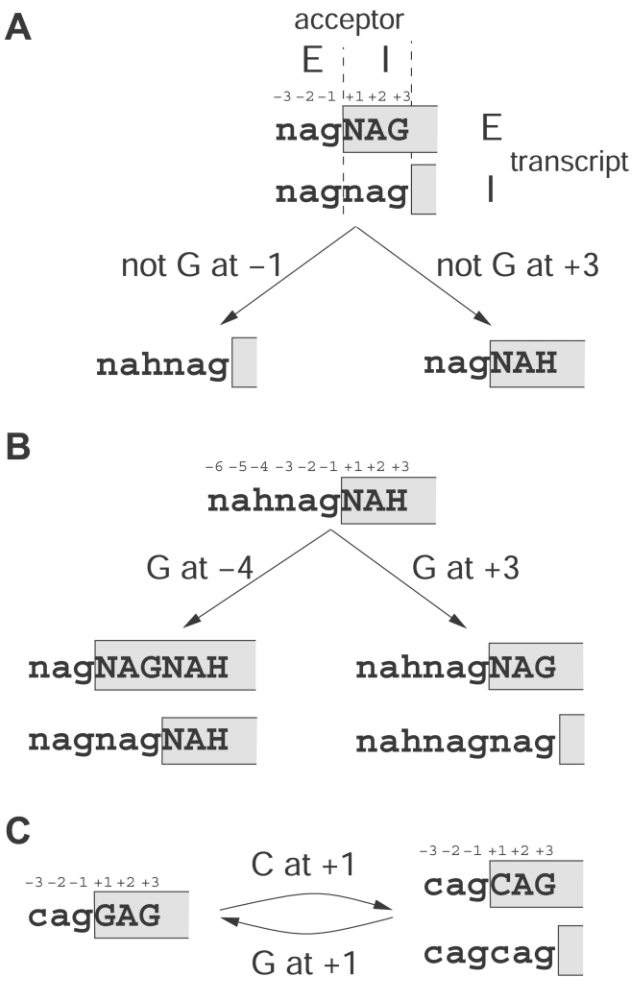

Figure 1 Schematic illustration of how SNPs affect splicing at NAGNAG acceptors. *A*, SNP alleles at position −2, −1, +2, or +3 of a NAGNAG acceptor destroy this motif by affecting the E (*left*) or I (*right*) acceptor, thus preventing alternative splicing. *B*, SNP alleles at intron positions −5 and −4 can create a novel E acceptor (*left*) and, at exon positions +2 and +3, a novel I acceptor (*right*), thus yielding a NAGNAG motif. Acceptors at these alleles may allow alternative splicing, as indicated by the two transcripts (*E transcript above; I transcript below*). *C*, SNP alleles at position −3 or +1 of a NAGNAG acceptor can convert a plausible NAGNAG that allows alternative splicing (*left*) to an implausible one that allows only the expression of one transcript (*right*), or vice versa. Positions refer to a standard intron-exon boundary. H denotes A, C, or T; upper- and lowercase letters indicate exonic and intronic nucleotides, respectively; exonic nucleotides are boxed.

ing of the resulting amplicons that correspond to SNPs listed in table 1, we used primers 5′-CAGCTACGGTTTGCTGAGAA-3′ and 5′-ACAGAGGGGACAGGGAGATT-3′ for genotyping *rs2245425*, 5′-GATTTTCCTGGAGGAGAGGG-3′ and 5′-CAAGTTCAAAGCAAGCCTCC-3′ for *rs1558876*, 5′-AGGAGGCGTGCTATCTGGTA-3′ and 5′-GTAGGAAGCCCTGGAGGAAG-3′ for *rs2290647*, 5′-GCCATTGAGTTGTCATCACC-3′ and 5′-ACCCATTAGCTTGGCAACAG-3′ for *rs2275992*, 5′-AAGAATGGCGTCCATTTCAC-3′ and 5′-TTTCTGATCCTTGGTGAGGG-3′ for *rs4590242*, and 5′-CCTTCAACCTCAATGACGAAA-3′ and 5′-CACAAAGGACTTGTCAGGGA-3′ for *rs1152522*. RT-PCR for transcript amplification was done with primers 5′-GAAAGCGCGTACTACCTTCG-3′ and 5′-AATCCCTGGATCTGGCCTTA-3′ for *TOR1AIP1*, 5′-AGGCTACAACCACCCTCCTT-3′ and 5′-ACTTCCCCCTTGACGAGTTT-3′ for *KIAA1001*, 5′-AGAGGAGGACAAGGAGGAGC-3′ and 5′-GAACAGCGTCTGTGTCTCCA-3′ for *KIAA1533*, 5′-GGACATCTGTTTCTCGCCAT-3′ and 5′-ATCCTTCCATCTCACAACGG-3′ for *ZFP91* (GenBank accession number NM_170768), 5′-TCTTTCTTTTGTGGTGGGGA-3′ and 5′-TGTCAGGGACCCAGATCTTC-3′ for *GABRR1*, and 5′-TGCAGGACCAGAATAAAGCC-3′ and 5′-TATGGTCCCTTGGACTTTGC-3′ for *C14orf105*. For *ZFP91* and *TOR1AIP1*, the amplicons obtained by RT-PCR from individuals with each of the possible genotypes were cloned into PCR2.1-TOPO (Invitrogen) and were propagated in *Escherichia coli* TOP10 cells, respectively. Plasmids were isolated from several isolated clones, and their inserts were sequenced using plasmid primers. SNPs exhibiting nonancestral plausible NAGNAGs without EST evidence were selected by high frequencies of the minor alleles *rs1638152* (*DTX2*), *rs5248* (*CMA1*), and *rs17105087* (*SLC25A21*). Genomic primers used were for *DTX2* (5′-TTTCCTCCTGGCAGCTTAGA-3′ and 5′-GCTGGGAGATGAAACCAAAG-3′), *CMA1* (5′-GGCTCCAAGGGTGACTGTTA-3′ and 5′-CCCCACTTTCCCGTTTAACT-3′), and *SCL25A21* (5′-AACTCCATGTCGTCCCAAAG-3′ and 5′-CAAAATCGTTTGTTCTTTGCC-3′). Transcript-specific primers were used for *DTX2* (5′-CAGGCATGACGAGTGTTCTG-3′ and 5′-CACAGCTAGGGACCCGAT-3′) and *CMA1* (5′-CCCTGCTGCTCTTTCTCTTG-3′ and 5′-ACACACCTGTTCTTCCCCAG-3′).

Results

SNPs in NAGNAG Acceptors Influence Alternative Splicing

We extracted from the UCSC Human Genome Browser (hg17, May 2004) all annotated SNPs that are located within the last 6 nt of an intron or within the first 3 nt of an exon, given intron-exon boundaries from RefSeq transcripts. From these SNPs, we selected those that affect a NAGNAG acceptor. With respect to the human reference genome sequence, the alternative SNP allele can create or destroy a NAGNAG acceptor by affecting one of both AG alleles (fig. 1*A* and 1*B*). Since the nucleotide upstream of any acceptor AG is usually C or T (Stamm et al. 2000) and a change at this position is likely to alter alternative splicing at a tandem acceptor, we also considered SNPs at the N positions in an existing tandem (fig. 1*C*). We found a total of 137 NAGNAG-affecting SNPs (table 2). Aware of the uncertainty about the true nature of SNPs in segmental duplications (Fredman et al. 2004; Taudien et al. 2004), we excluded seven (5%) of the variations from further analysis. Our precaution was justified by genotyping SNP *rs1638152* in 12 whites; we consistently found both alleles and both transcripts

Table 2

SNPs That Affect NAGNAG Acceptors

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

(*DTX2* [GenBank accession numbers DQ082728 and DQ082730]), which is a strong indication for paralogous sequence variants and/or multisite variations (combinatorial $P = .0003$). Since dbSNP entries sometimes are the result of sequencing errors, we manually examined the trace data (if available) and excluded a further nine SNPs (7%). Thus, we considered a total of 121 bona fide SNPs affecting NAGNAG acceptors.

Searching dbEST (December 2004), we obtained confirmation for alternative splicing at 16% (19 of 121) of these tandem acceptors. However, this percentage must be considered a lower bound. In addition to the general limitations of an EST-based evaluation of alternative splicing (insufficient EST coverage, especially for tandem acceptors that are spliced in a tissue-specific manner), the allele frequencies of the NAGNAG alleles and populational biases in EST sampling introduce further constrictions. Noteworthy, 18 (95%) of the 19 confirmed tandem acceptors match the consensus HAGHAG (H denotes A, C, or T). Thus, 26% of the 68 polymorphic HAGHAGs are EST confirmed, whereas only 1.9% of the 53 acceptors carrying G at one or both variable positions of the NAGNAG motif are EST supported. This is in line with our previous genomewide analysis, in which 31% of the HAGHAGs and only 1.7% of the remaining NAGNAGs were found to be experimentally confirmed (see table 1 of Hiller et al. [2004]). On the basis of these differences in the degree of confirmation by mRNA and EST data, we propose to subdivide all tandem acceptors into "plausible" (HAGHAG) and "implausible" (GAGHAG, HAGGAG, or GAGGAG) acceptors. Further support for this classification comes from the genomewide observation that all plausible NAGNAGs have the same bias toward intron phase 1, as described elsewhere (Hiller et al. 2004) for experimentally confirmed NAGNAGs, whereas the introns with implausible tandem acceptors are not biased toward phase 1 (table 3).

Accordingly, 68 (56%) of the 121 SNPs affect a plausible NAGNAG. However, four of those convert a plausible into another plausible NAGNAG, which has presumably no drastic consequence for NAGNAG splicing, even though we cannot exclude the possibility of changes in the ratio of E to I transcripts or of changes in tissue specificity. Thus, we consider the remaining 64 (53%) SNPs as relevant for NAGNAG splicing (table 4).

Cases of SNPs that comprise NAGNAG-acceptor and non-NAGNAG–acceptor alleles represent knockout experiments made by nature. We took this opportunity to investigate the assumed correlation between NAGNAG-acceptor genotypes and the appearance of E and I transcripts. Such a study seemed reasonable, since, so far, it has been performed in artificial splicing systems only (Tadokoro et al. 2005). We selected six SNPs with a heterozygosity of >0.2 that affect EST-confirmed HAGHAG acceptors for genotyping and detection of transcript forms. In two cases, we did not find either genotypes with at least one NAGNAG allele or genotypes that are homozygous for the non-NAGNAG allele. In the remaining four cases, we consistently observed E and I transcripts in cells with at least one HAGHAG allele, whereas cells that do not have a HAGHAG acceptor allele produced only one transcript (table 1). This strict correlation between NAGNAG alleles and alternative splicing is illustrated for *ZFP91* and *TOR1AIP1* in figure 2. These results confirm that NAGNAG motifs are necessary for this type of alternative splicing.

Next, we asked whether NAGNAG motifs created by the nonancestral SNP alleles are also sufficient for alternative splicing. With regard to the human reference sequence, in 36 (56%) of 64 cases, a novel NAGNAG is created; in 18 (28%), a known NAGNAG is destroyed by affecting an AG; and in 10 (16%), the N positions are changed. Since the appearance of a SNP allele in the current human genome build is rather random and does not reflect either the relative allele frequency in a defined population or its evolutionary history, the best reference for the question of gain versus loss of NAGNAG acceptors is the UCSC Chimpanzee Genome Browser (panTro1, November 2003). When the sequence context of the 64 plausible NAGNAG-affecting SNPs is compared, for 61 (95%), the orthologous chimpanzee nucleotide is identical to one of both human alleles, which we therefore consider the ancestral one (Watanabe et al. 2004). In 43 cases, the plausible NAGNAG is gained (nonancestral), and, in 18 cases, it is lost (ancestral). Consistent with our assumption that novel plausible NAGNAGs are very likely functional, we found EST evidence of alternative splicing in 16% (7 of 43) (table 4). To pro-

Table 3

Phase Distribution of Human Introns and NAGNAG Acceptors

	No. (%) of Introns by Phase		
Intron Class	0	1	2
Confirmed NAGNAGs[a,b]	349 (39.8)	379 (43.2)	150 (17.0)
Plausible NAGNAGs[b]	1,111 (42.5)	1,099 (42.0)	405 (15.5)
Implausible NAGNAGs[b]	2,568 (54.5)	1,466 (31.1)	677 (14.4)
All introns[c]	(46)	(33)	(21)

Note.—Only NAGNAGs that are located upstream of a coding exon are considered.

[a] EST/mRNA confirmed.

[b] From table 1 of Hiller et al. (2004).

[c] Genomewide frequencies (Long and Deutsch 1999).

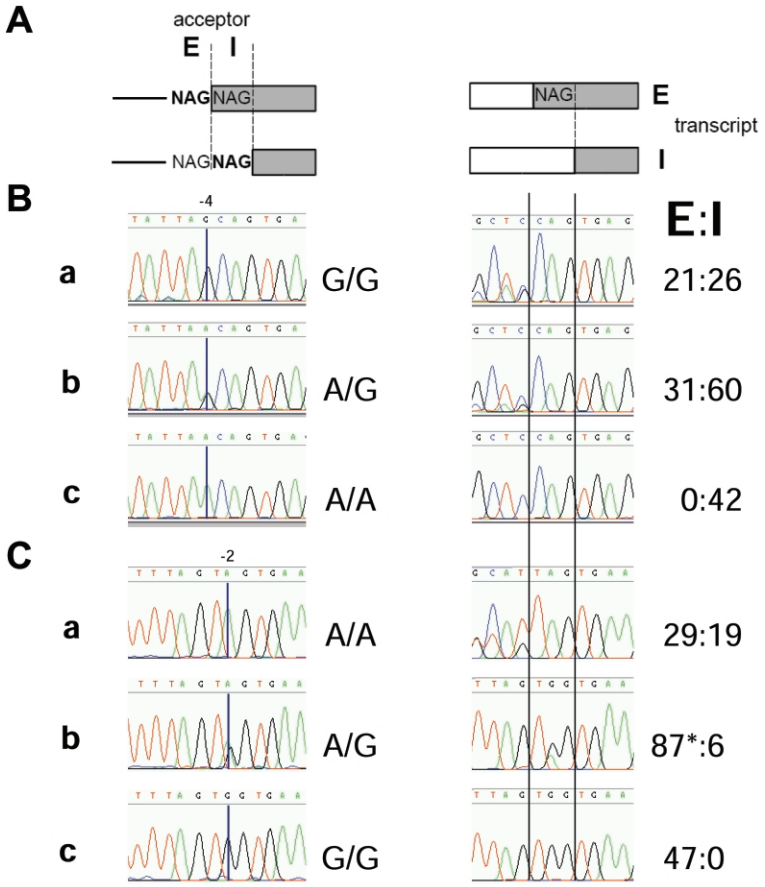

Figure 2 SNPs that affect plausible NAGNAG acceptors as knockout experiments made by nature. *A,* Schematic representation of the nomenclature of NAGNAG acceptors (*left*) and transcripts (*right*). *B,* SNP *rs2245425* affecting the E acceptor of *TOR1AIP1* exon 3 leads to the exclusive expression of the I transcript from the A allele (NAGNAG position −4; for numbering scheme, refer to fig. 1). *C,* SNP *rs2275992* affecting the I acceptor of *ZFP91* exon 5 leads to the exclusive expression of the E transcript from the G allele (position −2). Homozygous NAGNAG allele (*a*), heterozygous (*b*), and homozygous non-NAGNAG allele (*c*) are shown as genomic with genotypes (*left*); cDNA with E:I transcript ratio determined by counting subcloned and sequenced RT-PCR fragments (*right*). The asterisk (*) denotes E transcripts that can be assigned to the SNP alleles in the I acceptor (A = 15; G = 72).

vide further experimental support that respective SNP alleles enable alternative NAGNAG splicing, we selected two nonancestral plausible NAGNAGs without EST evidence. As expected, in leukocytes of individuals heterozygous or homozygous for the respective tandem allele of *rs5248,* we observed the expression of E and I transcripts (GenBank accession numbers DQ082727 and DQ082729) in the ratios 4:14 and 11:7, respectively (table 4). In the case of *rs17105087,* we were unable to identify the nonancestral allele in our white population sample. By analyzing the human-chimpanzee genomic sequence context of the eight confirmed nonancestral NAGNAG alleles, we found three cases in which both genomes are identical in a long range (*rs2287800* [−140/+123 identical nucleotides], *rs3765018* [−130/+95 nt], and *rs2290647* [−105/+70 nt]). Since most splice enhancers function only within a distance of <100 nt from the affected splice site (Schaal and Maniatis 1999), these findings suggest that NAGNAG motifs are sufficient for alternative splicing in the context of a previously non-NAGNAG acceptor.

Evolutionary Aspects of SNPs in NAGNAG Acceptors

At first glance, surprisingly, the large majority (43 [70%] of 61) of the plausible NAGNAGs are created (35 novel tandem AG alleles and 8 conversions of implausible into plausible), whereas only 18 are destroyed (16 AG destructions and 2 conversions of plausible into implausible). Therefore, we questioned whether there is a trend toward gain-of-NAGNAG acceptors in the human lineage. To test this, we used a null model that maps SNPs to randomly chosen acceptors (see appendix A) and found nearly the same relation for gain and loss of

Table 4

SNPs Affecting Plausible NAGNAG Acceptors

dbSNP ID	Gene Symbol	RefSeq ID	Exon	Nucleotide Pattern[a]	mRNA/EST[b]	Intron Phase[c]	Protein Impact[d]	Chimpanzee[e]	Human versus Chimpanzee[f]
rs2307130	*AGL*	NM_000644	2	CTCT***A***G|AAG→CTCT***G***G|AAG		…	5′ UTR	TAGAAG	Loss
rs12944821	*AP1GBP1*	NM_007247	3	TTTCAG|***C***AG→TTTCAG|***G***AG	12:5	1	Delete G	CAGCAG	Loss
rs363209	*APPBP1*	NM_003905	7	AA***A***CAG|CAC→AA***G***CAG|CAC		2	Insert S	AAACAG	Gain
rs2287800	*AQP8*	NM_001169	2	C***G***GCAG|ATA→C***A***GCAG|ATA	1:14	0	Insert Q	CGGCAG	Gain
rs13228988	*AUTS2*	NM_015570	8	CGATA***G***|CAG→CGATA***A***|CAG	4:2	2	Delete S	TAGCAG	Loss
rs8176139	*BRCA1*[h]	NM_007304	8	GTT***T***AG|CAG→GTT***G***AG|CAG	26:7	0	Delete Q	TAGCAG	Loss
rs1152522	*C14orf105*	NM_018168	4	TCAT***A***G|CAG→TCAT***G***G|CAG	3:6	0	Delete Q	TAGCAG	Loss
rs1044833	*C1QDC1*[h]	NM_001002259	18	AAACA***G***|CAG→AAACA***T***|CAG		1	Delete A	CAGCAG	Loss
rs11567804	*C3AR1*	NM_004054	2	TTGCA***G***|AAG→TTGCA***A***|AAG		…	5′ UTR	CAGAAG	Loss
rs11660370	*CABLES1*	NM_138375	3	TTTCAG|A***T***G→TTTCAG|A***A***G		2	Delete R	CAGATG	Gain
rs1804783	*CACNA1A*	NM_023035	39	TTGCAG|G̲AG→TTGCAG|***T***AG		1	Delete V	CAGGAG	Gain
rs5248	*CMA1*	NM_001836	3	CA***A***CAG|GTC→CA***G***CAG|GTC	15:21[g]	2	Insert S	CAACAG	Gain
rs3014960	*COG3*	NM_031431	14	ATACAG|CA***G***→ATACAG|CA***A***		0	Delete Q	CAGCAA	Gain
rs10914468	*COL16A1*	NM_001856	5	CTCCAG|A***A***G→CTCCAG|A***C***G		2	Delete R	CAGACG	Gain
rs2425068	*CPNE1*	NM_152927	16	CCCCAG|CA***A***→CCCCAG|CA***G***	79:7	0	Delete Q	CAGCAA	Gain
rs9463545	*CRISP1*	NM_001131	3	TAAC***A***G|AAG→TAAC***C***G|AAG		0	Delete K	CAGAAG	Loss
rs11597439	*CUEDC2*	NM_024040	2	CTTCAG|AA***G***→CTTCAG|AA***C***	49:2	…	5′ UTR	CAGAAG	Loss
rs3020724	*CYP17A1*	NM_000102	8	C***T***GCAG|AGC→C***A***GCAG|AGC		1	Insert A	CTGCAG	Gain
rs3025420	*DBH*	NM_000787	11	CA***C***CAG|GTT→CA***G***CAG|GTT		2	Insert S	CACCAG	Gain
rs12760076	*DMAP1*	NM_019100	8	TTGCAG|A***T***G→TTGCAG|A***A***G		0	Delete K	CAGATG	Gain
rs11661706	*EPB41L3*	NM_012307	12	CTGCAG|A***G***G→CTGCAG|A***A***G		1	Delete E	CAGAGG	Gain
rs2271959	*ETV4*	NM_001986	3	T***C***GCAG|AAA→T***A***GCAG|AAA	3:17	0	Insert Q	TCGCAG	Gain
rs13251099	*FLJ36980*	NM_182598	2	AA***A***TAG|GTC→AA***G***TAG|GTC		…	5′ UTR	AAATAG	Gain
rs3765018	*FLJ46354*	NM_198547	23	TCCCAG|AA***G***→TCCCAG|AA***A***	11:1	…	3′ UTR	CAGAAA	Gain
rs6285	*GABRB1*	NM_000812	3	C***G***GCAG|GGC→C***A***GCAG|GGC		1	Insert A	NA	NA
rs4590242	*GABRR1*	NM_002042	2	T***G***GTAG|GCC→T***A***GTAG|GCC	2:2	2	Insert S	TAGTAG	Loss
rs2409496	*GART*	NM_175085	6	AATCAG|***G***AG→AATCAG|***C***AG		0	Delete Q	CAGGAG	Gain
rs751517	*GGA1*	NM_013365	10	TTCCAG|C***G***G→TTCCAG|C***A***G		1	Delete A	NA	NA
rs2010657	GGT1[h]	NM_013421	2	CCCCAG|C***G***G→CCCCAG|C***A***G		…	5′ UTR	CAGCGG	Gain
rs9644946	*GOLGA1*	NM_002077	8	AA***A***TAG|GAG→AA***G***TAG|GAG		0	Insert stop	AAATAG	Gain
rs2243187	*IL19*	NM_153758	5	TCACA***G***|CAG→TCACA***A***|CAG		0	Delete Q	CAGCAG	Loss
rs2290609	*IL5RA*	NM_000564	5	CA***A***CAG|TTT→CA***G***CAG|TTT		1	I versus TV	CAACAG	Gain
rs2297988	*KIAA0690*	NM_015179	33	***A***AGCAG|AAA→***G***AGCAG|AAA		0	Insert Q	GAGCAG	Gain

rs1558876	*KIAA1001*	NM_014960	6	TTTCAG|CA***C***→TTTCAG|CA***G***	10:2	2	Delete S	CAGCAG	Loss
rs3746373	*KIAA1510*	NM_020882	6	CCCCAG|C***C***G→CCCCAG|C***A***G		1	Delete A	CAGCCG	Gain
rs2290647	*KIAA1533*	NM_020895	11	CTCCAG|C***G***G→CTCCAG|C***A***G	34:35	1	Delete A	CAGCGG	Gain
rs3738833	*LSM10*	NM_032881	2	CCACAG|CA***A***→CCACAG|CA***G***		…	5′ UTR	CAGCAA	Gain
rs479984	*MGC35555*	NM_178565	5	TACTA***G***|AAG→TACTA***A***|AAG		1	Delete E	TAGAAG	Loss
rs3751353	*MGC48915*	NM_178540	4	ATTTAG|***G***AG→ATTTAG|***C***AG		1	Delete A	TAGGAG	Gain
rs11042902	*MRVI1*	NM_006069	2	AACC***G***G|CAG→AACC***A***G|CAG	2:2	2	NR versus K	CAGCAG	Loss
rs2298847	*MT1G*	NM_005950	2	T***A***GCAG|GTG→T***T***GCAG|GTG	59:14	1	Insert A	TTGCAG	Gain
rs2273431	*NID2*	NM_007361	10	ATGCAG|A***G***G→ATGCAG|A***A***G		1	Delete E	CAGAGG	Gain
rs12974798	*NTE*	NM_006702	35	TCGCAG|***G***AG→TCGCAG|***A***AG		0	Delete K	CAGGAG	Gain
rs17173698	*PAPSS2*	NM_004670	2	TTATAG|***G***AG→TTATAG|***A***AG		0	Delete K	TAGGAG	Gain
rs3842776	*PARVG*	NM_022141	4	TTCCAG|***G***AG→TTCCAG|***C***AG		1	Delete A	CAGGAG	Gain
rs1438073	*PDE1A*	NM_001003683	3	AA***A***TAG|ACT→AA***G***TAG|ACT		2	Insert R	AAATAG	Gain
rs3816280	*PPP4R1*	NM_005134	5	CA***A***TAG|AAC→CA***G***TAG|AAC		1	Insert V	CAATAG	Gain
rs879022	*REGL*	NM_006508	3	GGACAG|***G***AG→GGACAG|***A***AG		1	Delete E	CAGGAG	Gain
rs1127307	*RGS19IP1*	NM_202494	6	CAATAG|C***G***G→CAATAG|C***A***G	87:8	1	Delete A	NA	NA
rs16960071	*SEMA6D*	NM_020858	16	A***T***GAAG|GCT→A***A***GAAG|GCT		2	Insert R	ATGAAG	Gain
rs11553436	*SERHL*	NM_170694	11	CTCCAG|C***G***G→CTCCAG|C***A***G		…	3′ UTR	CAGCGG	Gain
rs2243603	*SIRPB1*	NM_006065	5	TTCCAG|AA***G***→TTCCAG|AA***C***		1	Delete E	CAGAAC	Gain
rs17105087	*SLC25A21*	NM_030631	7	CTGCAG|CA***A***→CTGCAG|CA***G***		0	Delete Q	CAGCAA	Gain
rs2521612	*SLC4A1*	NM_000342	17	C***C***GTAG|GCT→C***A***GTAG|GCT		2	Insert R	CCGTAG	Gain
rs9621415	*SLC5A4*	NM_014227	9	C***G***GCAG|GTC→C***A***GCAG|GTC		0	Insert Q	CGGCAG	Gain
rs9606756	*TCN2*	NM_000355	2	TCTAAG|AA***A***→TCTAAG|AA***G***	62:2	1	Delete E	AAGAAA	Gain
rs11466221	TGFA[h]	NM_003236	2	CA***A***CAG|GTA→CA***G***CAG|GTA		1	Insert A	CAACAG	Gain
rs2245425	*TOR1AIP1*[h]	NM_015602	3	TA***G***CAG|TGA→TA***A***CAG|TGA	13:41	1	Insert A	TAGCAG	Loss
rs1071716	*TPM2*	NM_213674	6	CCCCAG|C***C***G→CCCCAG|C***A***G		2	Delete S	CAGTCG	Gain
rs4434604	*TRIM55*	NM_033058	8	TACCAG|A***A***G→TACCAG|A***G***G		1	Delete E	CAGAAG	Loss
rs7862221	*TSC1*	NM_000368	14	CTTCAG|A***A***G→CTTCAG|A***G***G		1	Delete E	CAGAAG	Loss
rs11574323	*WRN*	NM_000553	23	GGGTAG|AA***T***→GGGTAG|AA***G***		2	QS versus H	TAGAAT	Gain
rs2275992	ZFP91[h]	NM_170768	5	TTTTAG|T***A***G→TTTTAG|T***G***G	31:7	2	Delete S	TAGTAG	Loss
rs200925	*ZNF248*	NM_021045	5	TA***A***CAG|GGT→TA***G***CAG|GGT		1	Insert A	TAACAG	Gain

NOTE.—NA = sequence context not available in panTro1.

[a] 9-nt acceptor context; reference genome sequence is on the left, and SNP allele is on the right. Polymorphic position is shown in bold italics.

[b] Number of mRNA and ESTs that match the E:I transcripts (shown only if both transcripts are EST confirmed).

[c] Phase of the intron or 5′/3′ UTR.

[d] Impact of alternative NAGNAG splicing on the protein sequence.

[e] Chimpanzee sequence orthologous to human NAGNAG.

[f] Gain = plausible NAGNAG in one of the human alleles and no or implausible NAGNAG in chimpanzee; loss = no or implausible NAGNAG in one of the human alleles and plausible NAGNAG in chimpanzee.

[g] Experimentally confirmed within the present study.

[h] Experimentally confirmed elsewhere (Tadokoro et al. 2005).

plausible NAGNAG acceptors. Thus, the high number of nonancestral plausible NAGNAGs is presumably a consequence of the fact that NAGNAG motifs represent only 5% of all human acceptors (Hiller et al. 2004). In consequence, in recent primate genomes, a constant bias seems to exist toward the accumulation of NAGNAG acceptors, which leads to an increased complexity of the transcriptome and proteome, antagonized by purifying selection. The question of whether the currently observed NAGNAG fraction among human acceptors represents the saturation level has to be addressed by further comparative genomewide analyses.

Furthermore, we observed striking differences in the numbers of SNPs that affect the AG of the E or I acceptor in ancestral plausible and implausible NAGNAGs, respectively. For the 16 ancestral plausible HAGHAGs, the E acceptor is affected in 11 cases and the I acceptor in 5. In contrast, for 22 implausible HAGGAGs (one ancestral GAGGAG and two GAGHAGs were omitted), we found 5 and 17 cases, respectively (Fisher's exact test $P = .00766$). Interestingly, we observed the same trend by comparing all 138 human NAGNAGs that are not conserved in the chimpanzee genome (one GAGGAG and seven GAGHAGs were omitted). The I acceptors of 79 HAGHAGs are affected in 56% (44), whereas the GAG of 59 HAGGAGs is affected in 83% (49) (Fisher's exact test $P = .0009$). Implausible GAGGAG and GAGHAG motifs were not considered, since the number of cases is too small.

Since tandem acceptors are nonrandomly distributed in the human genome, with a bias toward intron phase 1 and toward single-aa indels in phase 1 and 2, we questioned whether the nonancestral plausible NAGNAGs are also biased. Indeed, these NAGNAGs show the same bias toward intron phase 1, and they also have a strong tendency to result in single-aa indels (table 5). Thus, the process of establishing SNPs that are relevant for alternative NAGNAG splicing in the human population seems to be a nonrandom process that is subject to the same evolutionary forces as the maintenance of the tandem acceptors themselves.

Discussion

Since splicing variations are coming more and more into the research focus of human molecular genetics (Lopez-Bigas et al. 2005), novel approaches are needed to identify splice-relevant SNPs. By data mining the SNP annotation of the UCSC Human Genome Browser, we identified 121 variations that may affect alternative splicing by creation, destruction, or changing of NAGNAG acceptors. To improve the specificity of our prediction, we classified NAGNAG acceptors into "plausible" (HAGHAG) and "implausible" (GAGHAG, HAGGAG, or GAGGAG) ones. This subdivision of the tandem acceptors, primarily based on the degree of confirmation by mRNA and EST data, is further supported by (1) the fact that GAG acceptors are very rare (Stamm et al. 2000), (2) our genomewide observation that only plausible and not implausible NAGNAGs have the same bias toward intron phase 1 as experimentally confirmed NAGNAGs (Hiller et al. 2004), and (3) the observed differences in the numbers of SNPs that affect the AG of the E or I acceptor in ancestral plausible and implausible NAGNAGs, respectively. The last indicates that the selection pressure to maintain the E acceptor for HAGGAGs is higher than the pressure to preserve the coding sequence, since destruction of the HAG acceptor will leave a GAG that is unlikely to act as an acceptor site. In contrast, for plausible HAGHAGs, destruction of either AG is much less deleterious, since the other will still function as an acceptor. Thus, the identified 64 SNPs in plausible NAGNAGs are highly predictive of variations in alternative splicing. Nevertheless, it represents an experimental and bioinformatic challenge for future research to elucidate what makes the rare cases of confirmed implausible NAGNAG acceptors.

Although it seems obvious that the disruption of a plausible NAGNAG acceptor abolishes the formation of alternative transcripts, SNPs in these motifs provide us with unique knockout experiments by nature to confirm this hypothesis experimentally. Analyzing the expression of E and I transcripts in cells with at least one HAGHAG allele or without HAGHAG alleles, we have shown that

Table 5

Intron Phase Distribution and Single aa Events of Nonancestral Plausible NAGNAG Acceptors

Acceptor	No. (%) of Introns by Phase			No. (%) of Single-aa Events, Phases 1 and 2
	0	1	2	
Nonancestral NAGNAG alleles[a]	12 (31.6)	16 (42.1)	10 (26.3)	24 (92.3)
Nonpolymorphic confirmed NAGNAGs[b]	349 (39.8)	379 (43.2)	150 (17.0)	487 (92.1)

Note.—Only NAGNAGs that are located upstream of a coding exon are considered.

[a] Plausible polymorphic NAGNAGs for which the chimpanzee acceptor has no NAGNAG.

[b] EST/mRNA-confirmed NAGNAGs (Hiller et al. 2004).

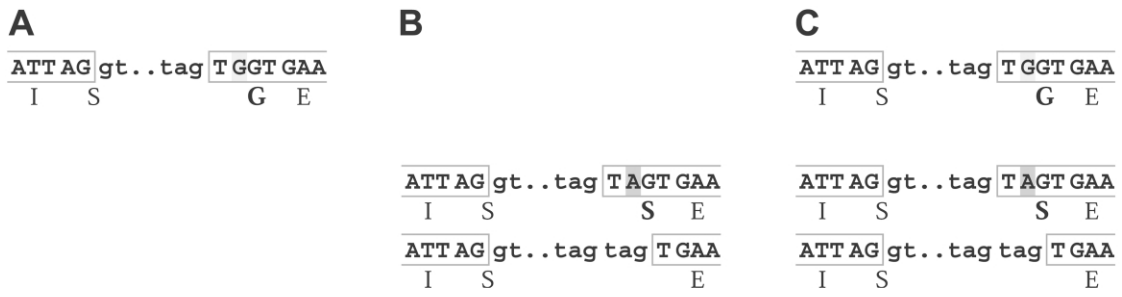

Figure 3 SNP affecting the I acceptor and the aa sequence of the E protein (*rs2275992* in *ZFP91*). Homozygosity of the G allele without a NAGNAG results in the expression of one protein (*A*), homozygosity of the A allele with the NAGNAG results in two (*B*), and heterozygosity results in three isoforms (*C*). All three transcripts are confirmed by at least four ESTs/mRNAs. The two allele variants are highlighted in light and dark gray. Amino acids are shown below the second codon position. Upper- and lowercase letters indicate exonic and intronic nucleotides, respectively. Exons are boxed.

the NAGNAG motif is necessary for this type of alternative splicing. In a subsequent analysis, we asked whether NAGNAG motifs created by the nonancestral SNP alleles allow alternative splicing. Usually, the introduction of an AG anywhere in the pre-mRNA does not create a functional acceptor site, since a polypyrimidine tract upstream and possibly enhancer sequences are required for recognition by the spliceosome. However, we suppose that the creation of a second AG 3 bases up or downstream of an existing acceptor is very likely to result in a functional tandem acceptor, since the splice-relevant sequence context is already present.

Referring to the chimpanzee genome as the reference for ancestral SNP alleles, we found EST and RT-PCR evidence that novel plausible NAGNAGs are most likely functional. This implies that a change of a normal acceptor to a plausible NAGNAG acceptor by a single mutation is sufficient to enable alternative splicing. Although the mechanism of NAGNAG splicing is not understood in detail, our findings argue against a general involvement of signals other than the NAGNAG itself. Thus, we conclude that SNPs in plausible NAGNAGs have an influence on NAGNAG splicing, regardless of whether the NAGNAG is ancestral. However, additional signals might be necessary for regulation of alternative splicing at tandem receptors.

Most interestingly, 23% (15 of 64) of SNPs in plausible NAGNAGs are translationally nonsilent and, thus, introduce a novel dimension of variability on the protein level by changing the I acceptor *and* the aa sequence of the E protein. Whereas homozygotes express either one or two isoforms, heterozygosity results in three different proteins (fig. 3). As listed in the Human Gene Mutation Database, the aa change can be dramatic—for example, as from Glu to the oppositely charged Lys in *PAPSS2* (*rs17173698*), which leads to a decrease in immuno-

Table 6

Human Disease Genes with SNPs Affecting Plausible NAGNAG Acceptors

dbSNP ID	Gene Symbol	RefSeq ID	Disease	MIM Number(s)	PubMed ID(s)
rs3020724	*CYP17A1*	NM_000102	Adrenal hyperplasia, congenital	#202110, *609300	4303304
rs12042060	*FIBL-6*	NM_031935	Age-related macular degeneration	#603075, *608548	14570714
rs2243187	*IL19*	NM_153758	Asthma	*605687	15557163
rs8176139	*BRCA1*	NM_007304	Breast cancer	*113705, #114480	9167459
rs11567804	*C3AR1*	NM_004054	Bronchial asthma	*605246	15278436
rs3025420	*DBH*	NM_000787	Congenital dopamine-beta-hydroxylase deficiency	#223360, *609312	14991826
rs2409496	*GART*	NM_175085	Down syndrome	*138440	9328467
rs1804783	*CACNA1A*	NM_023035	Episodic ataxia-2, familial hemiplegic migraine, spinocerebellar ataxia-6, idiopathic generalized epilepsy	#183086, #141500, #108500, *601011	8988170, 8898206, 9302278
rs2010657	*GGT1*	NM_013421	Glutathionuria	+231950	238530, 7623451
rs2307130	*AGL*	NM_000644	Glycogen storage disease type III	+232400	9032647, 10925384
rs1833783	*FTL*	NM_000146	Hyperferritinemia-cataract syndrome	#600886, *134790	7493028, 12199804
rs11661706	*EPB41L3*	NM_012307	Meningioma, lung cancer	*605331	10888600, 9892180
rs2275992	*ZFP91*	NM_170768	Acute myeloid leukemia	#601626	12738986,
rs1071716	*TPM2*	NM_213674	Nemaline myopathy-4, distal arthrogryposis 1	#609285, #108120, *190990	11738357, 12592607
rs2521612	*SLC4A1*	NM_000342	Renal tubular acidosis, ovalocytosis, spherocytosis	#179800, 166900, +109270	9600966, 1737855, 9973643
rs9644946	*GOLGA1*	NM_002077	Sjogren syndrome	270150, *602502	9324025
rs17173698	*PAPSS2*	NM_004670	Spondyloepimetaphyseal dysplasia	*603005	9714015
rs9606756	*TCN2*	NM_000355	Transcobalamin II deficiency	+275350	14632784
rs7862221	*TSC1*	NM_000368	Tuberous sclerosis	#191100, *605284	12773162, 14551205
rs11574323	*WRN*	NM_000553	Werner syndrome	#277700, *604611	9012406, 8968742

reactive protein (Xu et al. 2002). However, the third isoform of the protein generated by alternative NAGNAG splicing had not been taken into consideration. Moreover, it is conceivable that some of the SNPs in NAGNAG acceptors that allow the formation of three protein isoforms in heterozygotes may confer a heterozygous advantage.

Alternative splicing at tandem acceptors can result in the gain/loss of a premature stop codon in the mRNA. Among SNPs affecting plausible NAGNAGs, the G allele of SNP *rs9644946* changes the acceptor context of *GOLGA1* exon 8 from AAATAG to AAGTAG. Since intron 7 resides in phase 0, an inframe TAG insertion would be the consequence if the novel E acceptor is used. Interestingly, the gene codes for an autoantigen associated with Sjogren syndrome (MIM 270150). Since the E acceptor is preferred in alternative NAGNAG splicing (Hiller et al. 2004), the novel AAG acceptor is likely to be functional. The resulting E transcript is a candidate for nonsense-mediated mRNA decay (Maquat 2004). Thus, the AAGTAG allele would result in a lower protein expression. Alternatively, it is possible that the mRNA containing the premature stop codon escapes degradation, and the truncated protein may exhibit autoantigenic properties. It remains to be elucidated in populations with a sufficiently high allele frequency (e.g., 0.099 in the PERLEGEN panel that contains 24 samples of Chinese descent), regardless of whether alternative splicing at the AAGTAG acceptor contributes to the disease.

A second example of potential disease relevance is the SNP *rs363209,* the G allele of which creates a novel plausible AAGCAG acceptor of intron 6 in *APPBP1* (GenBank accession number NM_003905). The APP-BP1 protein binds to the carboxyl-terminal region of the amyloid precursor protein (APP) and interacts with the ubiquitin-activating enzyme E1C (UBE1C [homolog to yeast Uba3]) in the process of neddylation (Walden et al. 2003). APP plays a central role in Alzheimer disease and Down syndrome. Dysfunction of the APP-BP1 interaction with APP has been suggested to be one cause of Alzheimer disease (Chen 2004). The protein-protein interactions of the APP-BP1 E and I isoforms may be different and modulate the respective processes. It should be mentioned that the *UBE1C* gene (GenBank accession number NM_003968) itself contains a tandem acceptor (CAGAAG in front of exon 11). This may further increase the flexibility of the neddylation process by all four combinations of the E/I protein isoforms from two genes each.

The disease relevance of a NAGNAG SNP is demonstrated for the *ABCA4* gene (Maugeri et al. 1999). Maugeri et al. (1999) describe a NAGNAG mutation (2588G→C, changing the acceptor site TAGGAG→TAGCAG) that has a much higher frequency in patients with Stargardt disease 1 (STGD1 [MIM 248200]) and that is assumed to be a mild mutation that causes STGD1 in combination with a severe *ABCA4* mutation. By experimental analysis of the splice patterns of two patients with STGD1 who carry the mutation and one control individual, they found that only the alleles with the TAGCAG produce two splice forms. Our study exactly predicts this mutation outcome.

In general, most of the SNPs that are described in the present study—in particular, these in plausible NAGNAGs—affect the E:I transcript ratio, depending on the cell's genotype. SNP alleles with a destroyed E acceptor cause the exclusive expression of the I transcript. Alleles that destroy an I acceptor result in an exclusive expression of the longer E transcript. SNPs that comprise a plausible and an implausible NAGNAG allele will seriously hamper or disable splicing at the GAG acceptor. It has already been shown that a change in the ratio of alternative splice forms can cause diseases. For example, the change in the ratio of the alternative *MAPT* transcripts containing three or four microtubule-binding repeats may be causal for frontotemporal dementia (MIM 600274) (Spillantini et al. 1998). Another example is the *WT1* gene, in which alternative donor usage results in two protein isoforms that differ in 3 aa (+KTS/−KTS isoforms) and function (Englert et al. 1995). The altered ratio of +KTS/−KTS leads to Frasier syndrome (MIM 136680) (Barbaux et al. 1997). This situation is similar to that of NAGNAG acceptors, since E/I protein isoforms are observed that have functional differences (Condorelli et al. 1994; Tadokoro et al. 2005).

Altogether, 28% (18 of 64) of the plausible NAGNAG SNPs occur in known disease genes (table 6). Thus, they are preferable candidates for more-detailed functional analysis and association studies to link alternative splicing with diseases. Currently, there are no general methods that allow the prediction of splice-relevant SNPs. Focusing on SNPs that affect NAGNAG acceptors, we present a highly effective approach for the identification of SNPs that result in variations in alternative splicing patterns.

Acknowledgments

The skillful technical assistance of Ivonne Görlich is gratefully acknowledged. This work was supported by German Ministry of Education and Research grants 01GS0426 (to S.S.) and 01GR0105 and 0312704E (to M.P.) as well as Deutsche Forschungsgemeinschaft grant SFB604-02 (to M.P.).

Appendix A

Randomization Null Model for NAGNAG SNPs

To assess whether there is a preference for creating plausible NAGNAGs, we used a simulation that assigns a new acceptor to the 2.896 SNPs that overlap an acceptor in the 9-nt context and evaluates a possible NAGNAG-relevant outcome. For the 2,896 SNPs, we blasted the 101-nt genomic context (50 nt upstream and 50 nt downstream of the SNP) against the chimpanzee genome to determine the ancestral allele variant. We kept alignments with at least 95% identity and no mismatches in a ± 5-nt context around the SNP position. This yielded a total of 2,439 SNPs. Then, we blasted the 103-nt contexts (50 nt up- and downstream of the acceptor NAG) of 10,000 human acceptor sites (excluding the acceptors that are overlapped by a known SNP) against the chimpanzee genome and kept 8,082 for which we found an alignment (95% identity and no mismatch ± 10 nt around the NAG). Then, we assigned a new acceptor (randomly chosen from the 8,082) to a given SNP. We chose an acceptor with the ancestral allele variant at the respective position (e.g., if a SNP changes a C→G at position 4 of the 9-nt context, the new acceptor must also have a C at position 4). Since a methylated C in a CG context frequently mutates to a T, we assigned a new acceptor with the same sequence context at this position if the SNP represents a C→T mutation in a CG context (or a G→A mutation in a GC context on the opposite strand). This assures that context-dependent mutations are simulated in the same context. If a new acceptor is assigned to a SNP, we evaluated the possible impact on a NAGNAG acceptor. For each of the 2,439 SNPs, we successively assigned 10 randomly chosen acceptors (avoiding duplicate assignments).

The whole procedure was repeated 10 times, with different starts of the random-number generator. We calculated the following statistics from the 10 runs: (1) minimum and maximum percentage of creation versus destruction of a plausible NAGNAG, (2) minimum and maximum percentage of changes from a plausible to an implausible NAGNAG versus changes from an implausible to a plausible NAGNAG, and (3) minimum and maximum percentage of "gain of plausible NAGNAG" versus "loss of plausible NAGNAG." "Gain of plausible NAGNAG" is the sum of created, plausible NAGNAGs and changes from implausible to plausible. "Loss of plausible NAGNAG" is the sum of destroyed, plausible NAGNAGs and changes from plausible to implausible. These values were compared with the observed values by Fisher's exact test. For (1), we obtained P values between .52 and .75, for (2), P values between .72 and 1, and, for (3), P values between .66 and .88. Thus, the observed bias toward "gain of plausible NAGNAG" is comparable to the expectation.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

dbEST, ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/est_human.gz (for the human portion of dbEST)

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for the human mRNA download, *ZFP91* [accession number NM_170768], *DTX2* [accession numbers DQ082728 and DQ082730], *CMA1* [accession numbers DQ082727 and DQ082729]), *APPBP1* [accession number NM_003905], and *UBE1C* [accession number NM_003968])

Human Gene Mutation Database, http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for Sjogren syndrome, STGD1, frontotemporal dementia, and Frasier syndrome)

UCSC Chimpanzee Genome Browser, http://hgdownload.cse.ucsc.edu/goldenPath/PANTro1/bigZips/ (for source download panTro1 [November 2003])

UCSC Human Genome Browser, http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/ (for source download hg17)

References

Barbaux S, Niaudet P, Gubler MC, Grunfeld JP, Jaubert F, Kuttenn F, Fekete CN, Souleyreau-Therville N, Thibaud E, Fellous M, McElreavey K (1997) Donor splice-site mutations in WT1 are responsible for Frasier syndrome. Nat Genet 17:467–470

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Chen YZ (2004) APP induces neuronal apoptosis through APP-BP1-mediated downregulation of β-catenin. Apoptosis 9:415–422

Condorelli G, Bueno R, Smith RJ (1994) Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics. J Biol Chem 269:8510–8516

Englert C, Vidal M, Maheswaran S, Ge Y, Ezzell RM, Isselbacher KJ, Haber DA (1995) Truncated WT1 mutants alter the subnuclear localization of the wild-type protein. Proc Natl Acad Sci USA 92: 11960–11964

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861–866

Garcia-Blanco MA, Baraniak AP, Lasda EL (2004) Alternative splicing in disease and therapy. Nat Biotechnol 22:535–546

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet 36:1255–1257

Karinch AM, deMello DE, Floros J (1997) Effect of genotype on the levels of surfactant protein A mRNA and on the SP-A2 splice variants in adult humans. Biochem J 321:39–47

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Long M, Deutsch M (1999) Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol Biol Evol 16:1528–1534

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R (2005) Are

splicing mutations the most frequent cause of hereditary disease? FEBS Lett 579:1900–1903

Lynch KW, Weiss A (2001) A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. J Biol Chem 276:24341–24347

Maquat LE (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5:89–99

Maugeri A, van Driel MA, van de Pol DJR, Klevering BJ, van Haren FJJ, Tijmes N, Bergen AAB, Rohrschneider K, Blankenagel A, Pinckers AJLG, Dahl N, Brunner HG, Deutman AF, Hoyng CB, Cremers FPM (1999) The 2588G→C mutation in the *ABCR* gene is a mild frequent founder mutation in the Western European population and allows the classification of *ABCR* mutations in patients with Stargardt disease. Am J Hum Genet 64:1024–1035

Pagani F, Baralle FE (2004) Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5:389–396

Schaal TD, Maniatis T (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol Cell Biol 19:261–273

Spillantini MG, Murrell JR, Goedert M, Farlow MR, Klug A, Ghetti B (1998) Mutation in the tau gene in familial multiple system tauopathy with presenile dementia. Proc Natl Acad Sci USA 95:7737–7741

Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ (2000) An alternative-exon database and its statistical analysis. DNA Cell Biol 19:739–756

Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, Toyoda M, Ozaki M, Ono M, Miki N, Miyashita T, Yamada M (2005) 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 50:382–394

Taudien S, Galgoczy P, Huse K, Reichwald K, Schilhabel M, Szafranski K, Shimizu A, Asakawa S, Frankish A, Loncarevic IF, Shimizu N, Siddiqui R, Platzer M (2004) Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. BMC Genomics 5:92

Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, Franke A, Haesler R, Koch A, Lengauer T, Seegert D, Reiling N, Ehlers S, Schwinger E, Platzer M, Krawczak M, Muller-Quernheim J, Schurmann M, Schreiber S (2005) Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat Genet 37:357–364

Walden H, Podgorski MS, Schulman BA (2003) Insights into the ubiquitin transfer cascade from the structure of the activating enzyme for NEDD8. Nature 422:330–334

Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, et al (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature 429:382–388

Xu ZH, Freimuth RR, Eckloff B, Wieben E, Weinshilboum RM (2002) Human 3′-phosphoadenosine 5′-phosphosulfate synthetase 2 (PAPSS2) pharmacogenetics: gene resequencing, genetic polymorphisms and functional characterization of variant allozymes. Pharmacogenetics 12:11–21