BEITRÄGE ZU VERBREITUNG UND ANALYTIK DES SUBTILEN ALTERNATIVEN SPLEIßENS

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Diplom-Biologin
Stefanie Schindler
geboren am 08.09.1981 in Pößneck

Die vorliegende Arbeit wurde in der Zeit von Oktober 2005 bis März 2009 am Leibniz-Institut für Altersforschung – Fritz-Lipmann-Institut in Jena angefertigt.

Gutachter

1. PD Dr. Matthias Platzer
   Leibniz-Institut für Altersforschung – Fitz-Lipmann-Institut, Jena

2. Prof. Dr. Thomas Munder
   Fachhochschule Jena - Fachbereich Medizintechnik und Biotechnologie

3. Prof. Dr. Philip Rosenstiel
   Christian-Albrechts-Universität Kiel - Institut für Klinische Molekularbiologie

Datum der Verteidigung: 30.06.2009

INHALTSVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

AS	alternatives Spleißen
cDNA	„complementary DNA“ (komplementäre DNA)
CE-LIF	“capillary electrophoresis with laser-induced fluorescence “ (Fluoreszenz-basierte Kapillarelektrophorese)
DNA	„deoxyribonucleic acid“ (Desoxyribonukleinsäure)
EST	„expressed sequence tag“ (sequenzbasierter Expressionsmarker)
mRNA	„messenger RNA“ (Boten-RNS)
NCBI	„National Center for Biotechnological Information“ (Nationales Zentrum für Biotechnologische Information, USA)
NMD	„non-sense mediated mRNA decay“
nt	Nukleotid
PAGE-ED	“polyacrylamide gelelectrophoresis with ethidium-bromide densitometry” (Polyacrylamid-Gelelektrophorese mit Ethidiumbromid-basierter Densitometrie)
PCR	„polymerase chain reaction“ (Polymerase-Kettenreaktion)
PSQ	Pyrosequenzierung
RefSeq	Referenzsequenz-Datenbank des NCBI
RNA	„ribonucleic acid“ (Ribonukleinsäure)
RRM	„RNA recognition motif“ (RNA-Erkennungs-Motiv)
RT-PCR	Reverse Transkription gekoppelt mit Polymerase-Kettenreaktion
SNP	„single-nucleotide polymorphism“ (Einzelnukleotid-Polymorphismus)
upstream ORF	„upstream open reading frame“ (stromaufwärts gelegener Leserahmen)
UTR	untranslatierte Region

ABBILDUNGS- UND TABELLENVERZEICHNIS

ZUSAMMENFASSUNG

Das alternative Spleißen (AS) ist ein Hauptakteur der Diversifizierung von Transkriptom und Proteom eines eukaryotischen Organismus und von grundsätzlicher und weitreichender Bedeutung für die Biologie komplexer Organismen. Diese Dissertationsschrift befasst sich mit einem erst kürzlich entdeckten Typ des AS, dem subtilen AS, welcher die Einführung kleiner Variationen im Transkript und in vielen Fällen auch im kodierten Protein bewirkt. Dieses Phänomen wurde wegen seiner subtilen Effekte zunächst nur wenig beachtet, repräsentiert jedoch eine der häufigsten Formen des AS. Die in dieser Arbeit präsentierten neuen Erkenntnisse bilden einen weiteren Schritt zur Erforschung von Verbreitung und Biologie des subtilen AS, sowie zum Verständnis dieser Spleißereignisse im Gesamtkontext der Genregulation in höheren Eukaryoten.

In einer Analyse des *Arabidopsis thaliana*-Genoms konnte eine häufige Präsenz von NAGNAG-Tandem-Motiven als potentielle alternative 3'-Spleißstellen nachgewiesen werden. AS an diesen Tandem-Spleißstellen wurde mithilfe der vorhandenen Transkriptdaten festgestellt, die aufgrund ihres geringen Umfangs mit einer bioinformatischen sequenz-basierten Vorhersagemethode komplettiert wurden. NAGNAG-Tandem-Spleißstellen sind in *A. thaliana* in der Gruppe der SR-Protein-kodierenden Gene überrepräsentiert, die für wichtige Spleißfaktoren kodieren. Auf der Tatsache basierend, dass Spleißvarianten von Spleißfaktoren Auswirkungen auf den Spleißprozess, dessen Regulation und Ergebnis haben können, wurden aus dieser Gruppe Kandidaten für die experimentelle Anlayse des subtilen AS an NAGNAG-Tandem-Spleißstellen unter verschiedenen physiologischen und entwicklungsbiologischen Bedingungen ausgewählt. Bei den als alternativ gesplissen validierten Genen wurde gezeigt, dass sich die Änderungen in den Spleißvarianten-Verhältnissen der verschiedenen Gene ähneln und deshalb weitgehend unabhängig von sequenzspezifischer Spleißregulation zu sein scheinen, sowie eher durch organ- und bedingungsspezifische Änderungen des Spleißosoms vermittelt werden.

Auf der Suche nach subtilen Spleißereignissen im Mensch wurde ein völlig neuer, seltener Typus des subtilen AS entdeckt und systematisch untersucht. Es konnte eine Population von 36 Introns identifiziert werden, die TG-Dinukleotide als alternative 3'-Spleißstellen verwenden und damit den etablierten Regeln des Spleißens widersprechen. TG-3'-Spleißstellen wurden ausschließlich im Kontext einer alternativen AG-Spleißstelle gefunden, von der sie maximal 28 nt entfernt sind. In deren orthologen 3'-Spleißstellen sind TG-Dinukleotid und flankierende intronische Sequenz zwischen Säugetieren auffällig stark konserviert, einige Fälle sogar bis Huhn, Frosch oder Fisch. Interessanterweise steigt die Häufigkeit der Verwendung der TG-Spleißstelle mit der Konservierung von Spleißstelle und

flankierender Intron-Sequenz. Es ist daher naheliegend, dass sehr spezifische *cis*-, möglicherweise auch *trans*-Elemente die Wahl dieser außergewöhnlichen 3'-Spleißstelle vermitteln und deshalb das AS von lediglich 0,01% aller TG/AG-Motive an Intron-Exon-Grenzen ermöglichen.

Da quantitative Information über die Verhältnisse von Spleißvarianten für die Charakterisierung von Spleißereignissen und für die Ermittlung funktioneller Rollen entscheidend ist, wurden zwei Methoden im Zuge meiner Studien für diese Zwecke adaptiert: Die Pyrosequenzierung (PSQ) und die Fluoreszenz-basierte Kapillarelektrophorese (CE-LIF). Beide Methoden bieten den Vorteil, dass Spleißvarianten mit minimalen Längendifferenzen analysiert werden können. In einer systematischen Analyse hinsichtlich Reproduzierbarkeit und Genauigkeit der gemessenen quantitativen Daten, sowie Experimentaufbau, Datenanalyse und Anwendungsspektrum wurden beide Methoden im Vergleich mit der häufig verwendeten Polyacrylamid-Gelelektrophorese in Verbindung mit Ethidiumbromid-vermittelter Densitometrie (PAGE-ED) evaluiert. Außerdem wurde der Einfluss von niedrigen Template-Konzentrationen und Amplikon-Längendifferenzen in der RT-PCR untersucht, die zu zufälligen bzw. systematischen Fehlern führen können. Im Gegensatz zur PAGE-ED erlauben PSQ und CE-LIF genaue Messungen subtiler quantitativer Unterschiede in Hochdurchsatzanalysen. CE-LIF erzielte dabei die höchste Genauigkeit und Reproduzierbarkeit und stellte gleichzeitig die arbeits- und zeiteffizienteste Methode dar.

ABSTRACT

Alternative splicing (AS) is a key player of the diversification of eukaryotic transcriptomes and proteomes and has fundamental importance for the biology of complex organisms. The scope of this thesis is about a recently discovered class of AS – the subtle alternative splicing. Due to its subtle effects on transcripts and encoded proteins this class was originally underestimated and not considered in few studies despite of being a widespread form of AS. The findings of the studies presented herein represent a further step towards an understanding of dispersal and biology of subtle AS and the biological meaning of these splicing events in the context of gene regulation in higher eukaryotes.

In a genome-wide screen in the plant *Arabidopsis thaliana* a frequent occurence of NAGNAG-motifs potentially functioning as alternative 3' splice sites was identified. AS was assessed using available transcript data. Due to the relatively low coverage of the *A. thaliana* genome with transcript data the analysis was complemented with a sequence-based prediction method. NAGNAG tandem splice sites are overrepresented in the *A. thaliana* SR protein-coding genes which encode important splicing factors. Since splice variants of splicing factors may have consequences for alternative splicing and its regulation, alternative NAGNAG splicing of a subset of SR protein-coding genes was experimentally investigated under various conditions. Interestingly, the patterns of changes in splicing ratios are similar for all analyzed genes indicating that the differential effects on NAGNAG AS in the analyzed cases are organ- and condition-specific rather than gene-specific.

Searching for subtle alternative splicing events in human a rare type of subtle AS was discovered and systematically analyzed. A tiny population of 36 introns was identified involving TG dinucleotides functioning as alternative 3' splice sites and therewith violate the established splicing rules. TG 3‘ splice sites were validated in human tissues and were exclusively found in the context of an alternative AG-splice site with a maximum distance of 28 nt. TG 3' splice sites and their flanking intron sequences are substantially conserved among orthologous mammalian genes, in some cases even between human and chicken, frog or fish. Interestingly, the relative frequency of TG splice site usage positively correlates with the intronic sequence conservation. Obviously, *cis*- and/or *trans*-elements mediate the selection of these unusual 3' splice sites in agreement with the finding that only 0.01% of all TG/AG-motifs on intron-exon-boundaries are alternatively spliced.

Quantitative information about splice variants is important to characterize splicing events and to elucidate their functional roles. For that purpose two methods were adapted: the pyrosequencing (PSQ) and the capillary electrophoresis with laser-induced fluorescence (CE-LIF). Both methodes are able to analyze splice variants with small length differences,

which is in particular advantagous for the analysis of subtle AS. In a systematic analysis the techniques were analyzed in terms of accuracy and reproducibility of data obtained as well as assay setup, data analysis and applicability compared to the conventionally used polyacrylamide gelelectrophoresis with ethidium-bromide densitometry (PAGE-ED). Moreover, the influence of template concentrations, amplicons size differences and RT-PCR conditions were analyzed for potential random or systematic errors. In contrast to PAGE-ED, the PSQ and CE-LIF techniques can appropriately determine subtle quantitative differences in high-throughput analyses, whereas the latter turned out to be the most accurate and reproducible as well as the most time- and labor-efficient method.

"ALTERNATIVE SPLICING EMERGES AS AN EVOLUTIONARY WORKSHOP FOR TINKERING WITH THE PROTEIN STRUCTURE AND FUNCTION." [1]

Die vielleicht größte Überraschung des Human-Genom-Projekts war die Entdeckung, dass der Mensch nur über circa 20.000 - 25.000 protein-kodierende Gene verfügt [2-4]. Dies widerspricht zunächst der Anzahl menschlicher Proteine, die mit Schätzungen von über 100.000 die Zahl der Gene erheblich übersteigt [5]. Die Anzahl der Gene eines Organismus scheint also in keinem direkten Verhältnis zu dessen Komplexität zu stehen, wie auch der Vergleich mit dem Fadenwurm *Caenorhabditis elegans* deutlich zeigt: Erstaunlicherweise besitzt dieser mit 20.000 Genen [6] eine vergleichbare Anzahl wie der Mensch, obwohl er ein deutlich weniger komplexer Organismus ist. Kurz nach der Entdeckung von Exons und Introns stellte der Wissenschaftler Walter Gilbert bereits im Jahr 1978 die Frage „Why genes in pieces?" [7]. Gilbert spekulierte, dass die unterschiedliche Verwendung von Exons eines Gens zu funktionell unterschiedlichen Proteinen führen kann, obwohl man bis dato dogmatisch daran festhielt, dass ein Gen in der Regel nur ein Protein kodiert. Mittlerweile sind eine Reihe von Mechanismen bekannt, die die Entstehung verschiedener Proteine aus einem Gen erklären: Ein Hauptakteur der Diversifizierung des Proteoms ist das alternative Spleißen (AS) [8].

Das AS ist von grundsätzlicher und weitreichender Bedeutung für die Biologie höherer Eukaryoten. Die vorliegende Dissertationsschrift befasst sich mit einem erst kürzlich entdeckten Typ des AS – dem subtilen alternativen Spleißen. Dieser Vorgang ermöglicht subtile Änderungen im Transkript und in den meisten Fällen auch im kodierten Protein [9-13]. Die in dieser Arbeit präsentierten Studien liefern Beiträge zur Erforschung von Verbreitung und Häufigkeit des subtilen AS, und greifen regulatorische und mechanistische Aspekte unter funktionellen Gesichtspunkten auf [14-16]. Außerdem werden methodische Strategien zur Identifizierung und Quantifizierung alternativer Spleißereignisse evaluiert.

EINLEITUNG

Ein fundamentaler Unterschied zwischen prokaryotischen und eukaryotischen Genstrukturen ist die Existenz von Introns in Eukaryoten [17]. Die Informationen liegen im eukaryotischen Zellkern also gestückelt vor. Die meisten Gene höherer Eukaryoten setzen sich aus einer bestimmten Anzahl von Exons und Introns zusammen - mit durchschnittlich zehn Exons bzw. neun Introns pro humanem Gen [4]. Diese werden gemeinsam in Form einer primären mRNA (prä-mRNA) transkribiert. Für die Translation dieser gestückelten Information in ein Genprodukt müssen diese Informationen in eine kontinuierliche Folge gebracht werden, indem Teile der prä-mRNA entfernt werden – ein Prozess, der als Spleißen bezeichnet wird. Während des Spleißvorganges entscheidet sich dann, welche Teile der prä-mRNA im reifen Transkript zurückgehalten werden (Exons), bzw. welche in die mRNA nicht integriert und entfernt werden (Introns). Dieser Vorgang wird durch einen der größten molekularen Komplexe in der Zelle ausgeführt – dem Spleißosom. Das Spleißen stellt neben dem 5‘-„Capping“ und der 3‘-Polyadenylierung einen besonderen Vorgang des Prozessierens der prä-mRNA bei Eukaryoten dar. Die reife mRNA gelangt dann vom Kern ins Cytoplasma, wo sie für die Proteinsynthese zur Verfügung steht.

Beim konstitutiven Spleißen werden alle Exons eines Gens in der Reihenfolge ihres genomischen Auftretens aneinander gefügt. Aus einer Transkriptionseinheit entsteht also genau eine mRNA und nach deren Translation ein bestimmtes Polypeptid. Beim AS können aus ein und derselben Transkriptionseinheit mehrere verschiedene reife mRNA-Moleküle und durch deren Translation unterschiedliche Polypeptide gebildet werden [18]. Die Ein-Gen-ein-Protein-Hypothese für Eukaryoten ist somit mittlerweile eher eine Ausnahme, da ein Gen für unterschiedliche Proteine kodieren kann [19], und somit die Anzahl der Proteine die Anzahl der Gene weit überschreitet. Ein eindrucksvolles Beispiel ist *Dscam („*Down syndrome cell adhesion molecule“), ein Gen in der Taufliege *Drosophila melanogaster*, welches das Richtungswachstum von Nervenzellen steuert. Aus diesem einen Gen können durch AS rechnerisch insgesamt 38.016 verschiedene Proteine gebildet werden [20]. Im Gegensatz dazu erscheint die Zahl an Genen in diesem Organismus mit ca. 14.000 vergleichsweise klein [21]. Dies unterstreicht eindrucksvoll, dass die Vielzahl an Proteinen in einem Organismus nicht primär durch die Zahl seiner Gene bestimmt ist, sondern vielmehr durch das AS der prä-mRNA. Die Informationsdichte der DNA wird durch diese Überlagerung also erheblich erhöht. Damit trägt der Prozess des AS in hohem Maße zur Komplexität von Transkriptom und Proteom bei [19,22,23]. Aktuellen Schätzungen zufolge werden 92-94% der humanen Gene alternativ gesplissen [24]. Neben anderen Mechanismen wie alternativer Transkriptionsinitiation, alternativer Polyadenylierung und RNA-Editierung ist dieser post-transkriptionale Prozess das wichtigste Bindeglied zwischen der überraschend

geringen Anzahl an Genen und der wesentlich höheren Anzahl an Transkripten und Proteinen in höheren Eukaryoten [18].

Welche Teile der prä-mRNA beim Spleißvorgang herausgeschnitten werden sollen, markieren Erkennungsstellen an den Nahtstellen von Exons und Introns (Abb. 1). Meist ist der 5'-Terminus eines Introns durch ein GY-Dinukleotid und der 3'-Terminus durch ein AG-Dinukleotid definiert (GY-AG-Regel). Vor letzterer, der sogenannten 3'-Spleißstelle, ist eine Folge von Pyrimidin-Nukleotiden lokalisiert (Polypyrimidin-Trakt) sowie ein „Branchpoint", der sich meist 20-40 nt stromaufwärts befindet.

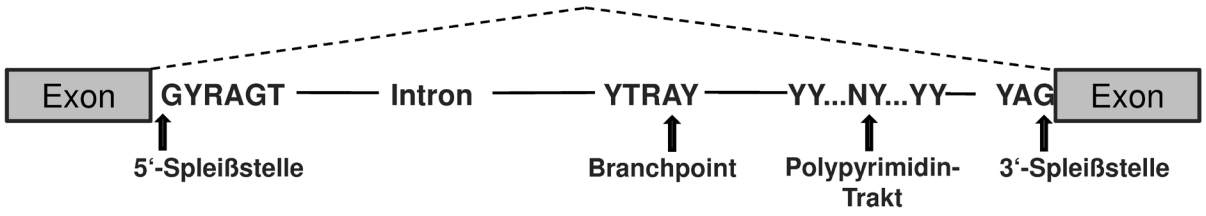


Abbildung 1: Schematische Darstellung der klassischen Spleißsignale eines Introns mit benachbarten Exons. „Y" repräsentiert die Pyrimidine C oder T, ein „R" die Purine A oder G. Nicht alle Nukleotide des Polypyrimidin-Traktes müssen Pyrimidine sein. Die gestrichelte Linie markiert die Stellen, an denen gesplissen wird. In Anlehnung an Hiller *et al.* 2008 [25].

Diese für den Spleißprozess wichtigen Bereiche sind aufgrund ihrer Funktion evolutionär konserviert. Zusätzlich können an bestimmten Sequenz-Motiven (*cis*-Elemente) regulatorische Faktoren binden, die die Auswahl der Spleißstellen beeinflussen können. Diese Motive sind redundant, aber im Gegensatz zu den o.g. obligatorischen Spleißsignalen für den Spleißvorgang nicht essentiell und weisen daher einen niedrigeren Konservierungsgrad auf. Der Prozess des Spleißens gliedert sich in zwei Schritte: Während des ersten Schrittes erfolgt die Definition des intronischen 5'-Terminus und die vorläufige Festlegung des 3'-Terminus, mit nachfolgender Formierung des Intron-Lariats an dem festgelegten Verzweigungspunkt, dem „Branchpoint". Im zweiten Schritt wird die 3'-Spleißstelle endgültig ausgewählt und die Exons miteinander verbunden (Abb. 2).

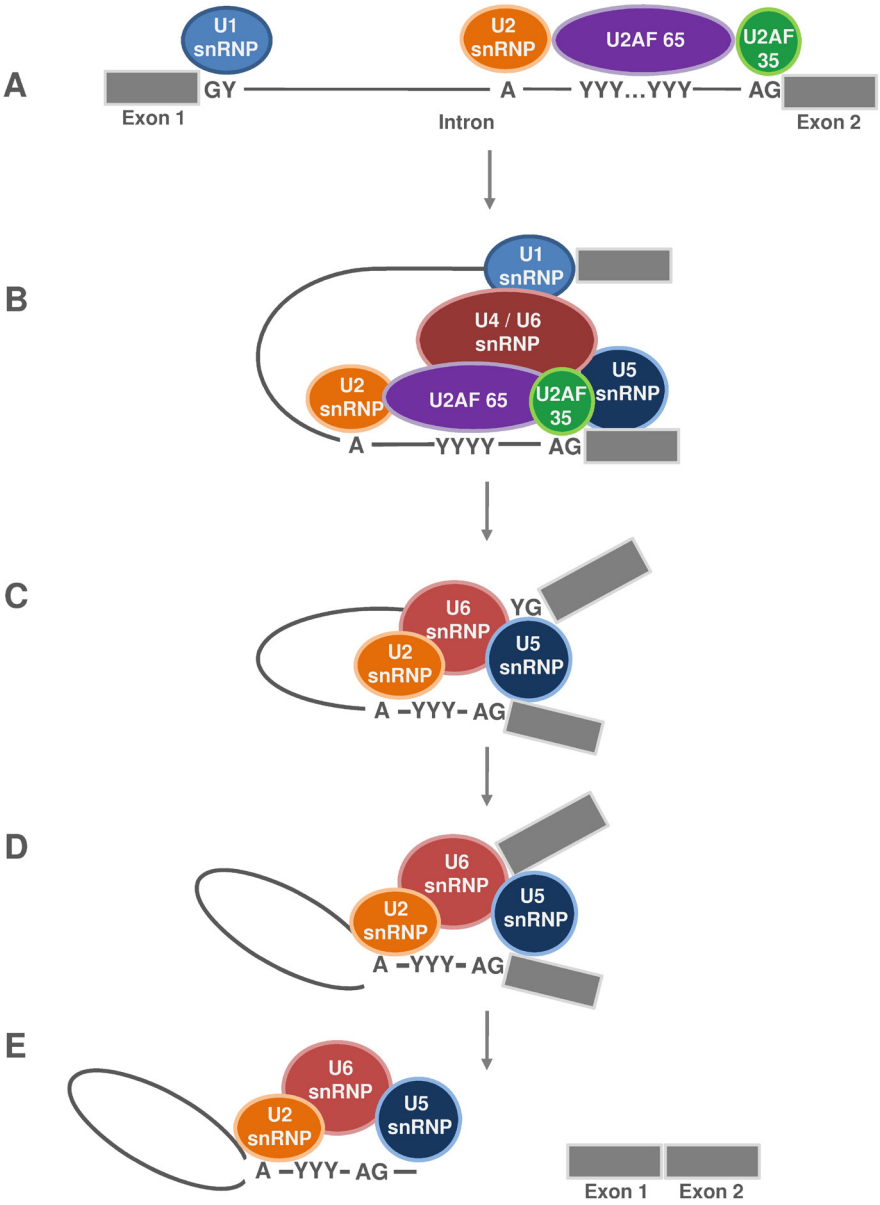


Abbildung 2: Vereinfachte Darstellung des Spleißens eines Introns mit benachbarten Exons. „GY“ bzw. „AG“ repräsentieren 5’- bzw. 3’-Spleißstelle, „A“ den „Branchpoint“, „Y“ die Pyrimidine C oder T des Polypyrimidin-Traktes. Vereinfacht dargestellt, besteht das Spleißosom aus mehr als 150 Proteinen/Proteinpartikeln, wobei die fünf kleinen nukleären Ribonukleoproteinpartikel (snRNPs, „small nuclear ribonucleoprotein particles“) U1, U2, U4, U5, U6 eine entscheidende Rolle spielen. Der Spleißprozess läuft dabei folgendermaßen ab [26]: (A) das U1 snRNP bindet an die 5’-Spleißstelle. Das Protein-Heterodimer U2AF (bestehend aus U2AF35 und U2AF65) bindet an den Polypyrimidin-Trakt und den 3’-Terminus des Introns, dann lagert sich der U2 snRNP am „Branchpoint“ an. (B) Die U4, U5 und U6 snRNPs werden akquiriert, wobei (C) der U6 snRNP den U1 snRNP ersetzt, welcher zusammen mit dem U4 snRNP das Spleißosom verlässt. (D) Die mRNA wird an der 5’-Spleißstelle gespalten, das freie 5’-Intron-Ende klappt zum „Branchpoint“ und formiert die typische Lasso-Struktur. (E) Schließlich wird die mRNA an der 3’-Spleißstelle gespalten, das stromaufwärts und stromabwärts gelegene Exon werden verbunden und das Intron freigegeben. In Anlehnung an http://banon.cshl.edu/cgi-bin/eventbrowser?DB=gk_current&FOCUS_SPECIES=Homo%20sapiens&ID=72163&

Beim AS können mehrere Formen unterschieden werden (Abb. 3): Das Überspringen von Exons, sich gegenseitig ausschließende Exons, das Beibehalten von Introns oder die Benutzung unterschiedlicher 5'- oder 3'-Spleißstellen [27]. Diese Spleißereignisse treten in verschiedenen Phyla in unterschiedlicher Häufigkeit auf. In Vertebraten ist beispielsweise das Überspringen von Exons, in Pflanzen hingegen das Beibehalten von Introns die am häufigsten verwendete Form des AS [28,29].

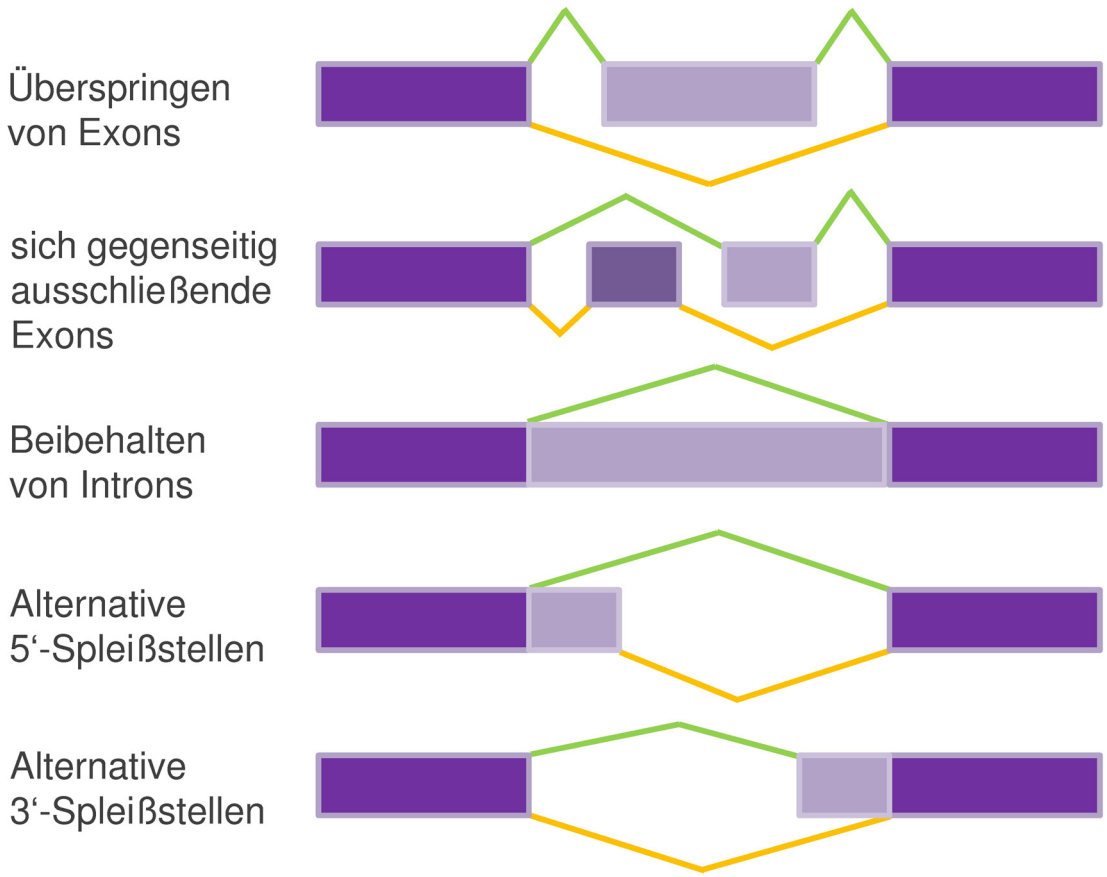


Abbildung 3: Die fünf häufigsten Typen des alternativen Spleißens. Die eine Möglichkeit zu Spleißen ist grün, die andere Alternative gelb gekennzeichnet. In Anlehnung an Cartegni *et al.* 2002 [27].

Die Auswirkungen des AS können strukturelle Änderungen der kodierten Proteine sein (Insertion/Deletion von funktionellen Elementen wie Proteindomänen oder Signalpeptiden), die funktionelle Änderungen in Liganden-Bindungsaffinität, Signalaktivität oder subzellulärer Lokalisation bewirken [19,22,23,30-33]. Darüber hinaus kann die Genexpression über RNA- oder Protein-Halbwertszeit gesteuert werden [22,23,30] oder die globale Enzym- oder Transkriptionsfaktor-Aktivität durch Erzeugung variabler Anteile nicht-funktioneller Isoformen

des aktiven Genproduktes reguliert werden: Durch Auslösen des sog. „non-sense mediated mRNA decay“ (NMD) ist es möglich, die Genexpression post-transkriptional durch die Produktion von Spleißvarianten zu steuern, die durch NMD abgebaut und somit nicht translatiert werden können [34,35]. Dieser regulatorische Prozess wird als RUST („regulated unproductive splicing and translation“) bezeichnet [36,37] und wird beispielsweise bei der Autoregulation von Spleißfaktoren genutzt [38-41]. Außerdem kann durch AS in UTR-Regionen die Stabilität und die translationale Effizienz einer mRNA (untranslatierten Regionen) beeinflusst werden [42].

Die biologische Bedeutung des AS ist sehr weitreichend. Änderungen des Spleißverhaltens durch *cis*- und *trans*-agierende Faktoren können schwerwiegende Folgen haben und das Schicksal einer Zelle in vielerlei Hinsicht bestimmen. Beispielsweise erfordert die Differenzierung neuronaler Zellen in Säugetieren komplexe Veränderungen von Genexpression und des neuronalen Transkriptoms. Daran ist eine Vielzahl spezifischer RNA-bindender Spleißregulatoren beteiligt, die das AS neuronaler Gene regulieren [30], wie z.B. *nPTB*, *Nova*, *Fox-1/2* und *Hu* [43-48]. AS ist auch von therapeutischem Interesse [49]. Treten Spleißdefekte auf bzw. kommt es aufgrund von Mutationen zu einer Veränderungen regulatorischer Prozesse, kann dies zu Krebs und anderen Krankheiten führen [50-53]. Beispielsweise verursacht eine in der menschlichen Population fixierte Punktmutation in einem exonischen spleißregulatorischen Element im Exon 7 des humanen *SMN2*-Gens („survival of motor neuron 2“) in Kombination mit dem Verlust des *SMN1*-Gens spinale Muskelatrophie [50,54]. Im Zuge des AS wird dadurch bei der Mehrheit der Transkripte Exon 7 übersprungen, wodurch ein C-terminal verkürztes, instabiles und vermutlich nicht funktionales Protein gebildet wird, was in der Endkonsequenz zum Krankheitsausbruch durch den Abbau von Motoneuronen in Stammhirn und Rückenmark führt [55].

In den letzten Jahren identifizierten experimentelle und *in silico*-Studien eine weit verbreitete Form des 5’- bzw. 3’-AS, bei denen die alternativen Spleißstellen nur wenige Nukleotide voneinander entfernt sind [9,10,12,56,57]. Diese sogenannten Tandem-Spleiß-stellen generieren 23,7% aller alternativen 5’- sowie 43,7% aller alternativen 3’-Spleiß-ereignisse [9] und sind in Säugetieren, Huhn, Zebrafisch, Taufliege und Pflanzen verbreitet [12,13,28]. Durch deren Verwendung entstehen mRNA, die sich in nur wenigen Nukleotiden voneinander unterscheiden. Im Gegensatz zu den anderen Formen des AS, welche oft beträchtliche Veränderungen der Proteinstruktur bewirken [58], werden durch die Verwendung von Tandem-Spleißstellen auch im kodierten Protein oft nur subtile Änderungen erzeugt. Die Auswirkungen von Tandem-Spleißstellen auf die Proteinstruktur hängen davon

ab, ob die Variationen zu einer Verschiebung des Leserahmens führen (Insertion/Deletion von 2, 4, 5, etc. Nukleotiden) oder ihn erhalten (Insertion/Deletion von 3, 6, 9, etc. Nukleotiden). Die Effekte einer Leserahmen-Verschiebung können schwerwiegend sein, da sie in Proteine mit verändertem C-Terminus oder zum Abbau der mRNA durch NMD resultieren können [9-11]. Leserahmen-erhaltende Variationen können zu sehr unterschiedlichen Effekten auf Proteinebene führen, beispielsweise zu subtilen Insertionen, Deletionen oder Substitution von wenigen Aminosäuren oder zur Einfügung eines Stopp-Codons mit gravierenden Auswirkungen [12,13]. Aus einigen Studien ist bekannt, dass funktionell unterschiedliche Proteine durch diese subtilen alternativen Spleißereignisse generiert werden [59-67]: Durch AS an einer 5'-Tandem-Spleißstelle des *WT1*-Gens („Wilms tumor 1") werden Proteinisoformen mit unterschiedlicher nukleärer Lokalisation und Liganden-Bindungsaffinität gebildet, die die Transkription beeinflussen und letztendlich an der Nieren- und Keimdrüsenentwicklung sowie an der Geschlechtsdetermination beteiligt sind [63,68]. Mutationen, die zur Inaktivierung der proximalen Spleißstelle führen, sind z.B. in die Ausprägung des Frasier-Syndroms involviert [69]. Dieses Beispiel zeigt eindrucksvoll, dass auch subtile alternative Spleißereignisse mit Krankheiten assoziiert sein können.

Das im Jahr 2004 beschriebene AS an NAGNAG-Tandem-Spleißstellen (N=A,C,G,T) beim Menschen ist ein typischer Fall des subtilen AS [12]. Mit einer Distanz von 3 nt zwischen beiden Spleißstellen ist es die häufigste Form aller subtiler alternativer Spleiß-stellen in Pflanzen und Säugetieren [10,28,56,57,70]. In Abhängigkeit davon, welche Spleißstelle benutzt wird, erfolgt entweder eine Insertion bzw. eine Deletion von 3 nt (NAG) in der mRNA (Abb. 4). NAGNAG-Tandem-Spleißstellen sind nicht nur beim Menschen, sondern auch in der Maus [12], Wiederkäuern [71], Huhn [72] und evolutionär weiter entfernten Organismen wie Taufliege [12] und Tomate [73] häufig vertreten. Wegen ihrer subtilen Effekte auf mRNA und Proteinstruktur wurden NAGNAG-Tandem-Spleißstellen oft übersehen oder unterschätzt. Einige Literaturhinweise belegen jedoch eine funktionelle Bedeutung [70,74-79]: Studien aus dem Jahr 1994 zeigen beispielsweise, dass Proteinisoformen des *IGFR1*-Gens („insulin-like growth factor type 1 receptor") mit unterschiedlicher Rezeptoraktivität gebildet werden, die auf einen Austausch von Thr-Gly zu Arg infolge von AS an NAGNAG-Tandem-Spleißstellen zurückzuführen sind [74]. Die auf diese Weise in der Maus modifizierten Proteinisoformen der Transkriptionsfaktor-kodierenden Gene *Pax3* („paired box 3") und *Pax5* („paired box 5") weisen unterschiedliche DNA-Bindungsspezifitäten [78] auf, und die im Menschen durch AS an NAGNAG-Tandem-Spleißstellen entstehenden *DRPLA*-Proteinisoformen („dentatorubral pallidoluysian atrophy") des Gens haben Unterschiede in ihrer subzellulären Lokalisation [70].

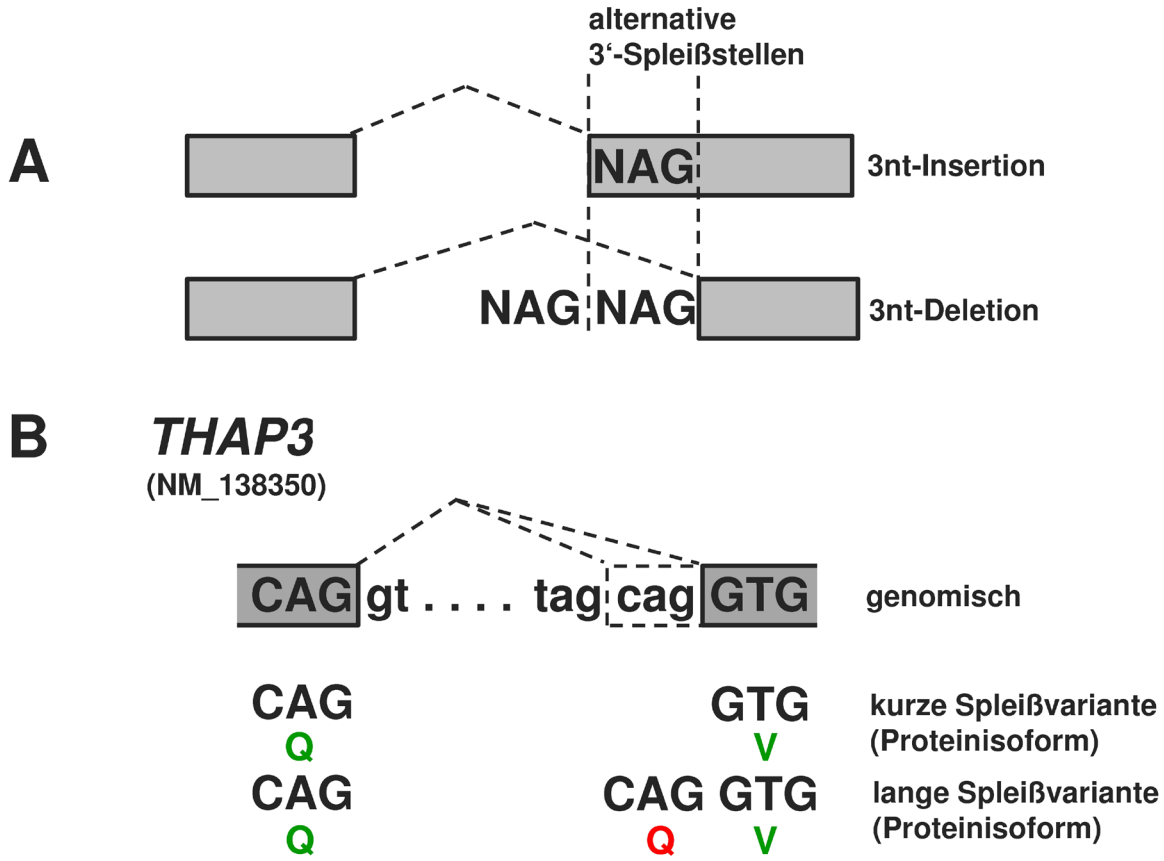


Abbildung 4: Alternatives Spleißen an NAGNAG-Tandem-Spleißstellen. (A) Die Benutzung des intronisch proximal gelegenen AG-Dinukleotids als 3'-Spleißstelle (oben) verursacht eine Insertion von 3 nt in die mRNA. Wird das intronisch distal gelegene AG-Dinukleotid verwendet (unten), verbleiben die 3 nt im intronischen Anteil und werden somit deletiert. Exons sind als graue Rechtecke dargestellt. (B) AS an NAGNAG-Tandem-Spleißstellen im Intron 3 des humanen *THAP3*-Gens („THAP domain containing, apoptosis associated protein 3"). Je nachdem welches AG-Dinukleotid benutzt wird, wird eine um eine Aminosäure verkürzte bzw. verlängerte Proteinisoform gebildet. In Anlehnung an Hiller *et al.*, 2004 [12]

Das subtile AS an NAGNAG-Tandem-Spleißstellen wurde im Jahr 2004 in der Studie von Hiller *et al.* für den Menschen umfassend charakterisiert [12]. NAGNAG-3'-Spleißstellen treten genomisch in ca. 30% der humanen Gene auf und werden in mindestens 12% der Fälle [80] alternativ gesplissen. Gibt es jedoch diese Form des AS auch in anderen, phylogenetisch von den Vertebraten weiter entfernten Organismen? Können die im humanen System erhobenen Befunde auch z.B. in Pflanzen validiert werden? Diese Fragestellung habe ich in meiner Arbeit „Alternative Splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes" [15] aufgegriffen. Das AS in Pflanzen ist bis in die

heutige Zeit nicht in der Breite erforscht wie in Vertebraten oder *D. melanogaster*. Aktuell liegen außerdem nur recht wenige Genome von Pflanzen vollständig sequenziert vor (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html) zu denen das Genom der Modellpflanze *A. thaliana* (Ackerschmalwand) gehört [81]. Basierend auf den Resultaten von Hiller *et al.* (2004) bildete die Analyse des subtilen AS an NAGNAG-Tandem-Spleißstellen in *A. thaliana* somit den ersten Angriffspunkt meiner Studien. Die genomische Präsenz potentieller NAGNAG-Tandem-Spleißstellen warf die grundsätzliche Frage auf, welcher Anteil tatsächlich alternativ verwendet wird. Üblicherweise werden Transkriptdaten herangezogen, um alternative Spleißereignisse zu identifizieren und zu lokalisieren (EST-, RefSeq-, mRNA-Daten). Die Abdeckung des *A. thaliana*-Genoms mit Transkriptdaten ist im Vergleich zu Mensch oder Maus jedoch verhältnismäßig gering. Mensch und Maus verfügen mit 20.000-25.000 protein-kodierenden Genen [4,82] aktuell über ~8.2 bzw. ~4.9 Millionen ESTs, *A. thaliana* derzeitig über lediglich ~1.5 Millionen ESTs bei einer vergleichbaren Zahl protein-kodierender Gene [81] (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Die vorhandenen EST-Daten waren daher für die Charakterisierung von AS an NAGNAG-Tandem-Spleißstellen nicht ausreichend und es bestand die Notwendigkeit, einen alternativen Lösungsweg zu entwickeln. Eine prognostische Methode, die die Wahrscheinlichkeit der alternativen Benutzung potentieller NAGNAG-Tandem-Spleißstellen vorhersagt, sollte den Mangel an Transkript-Abdeckung kompensieren. Darauf basierend, habe ich sowohl das subtile AS an NAGNAG-Tandem-Spleißstellen, als auch die *in silico*-Lösungsstrategie in *A. thaliana* experimentell validiert.

Das Vorhandensein alternativer NAGNAG-Tandem-Spleißstellen wirft eine Frage auf, die bereits von Hiller *et al.* (2004) formuliert wurde – die Frage nach den funktionellen Konsequenzen und der biologischen Relevanz subtiler alternativer Spleißereignisse [12]. In den ersten Studien fiel auf, dass NAGNAG-Tandem-Spleißstellen in humanen Genen überrepräsentiert sind, die für Proteine mit RNA-Erkennungsmotiven (RRM, „RNA recognition motif“) kodieren [12]. Die evolutionär konservierten Serin/Arginin-reichen Proteine (SR-Proteine) besitzen ein solches RRM [83,84]. SR-Proteine sind vielseitig agierende RNA-bindende Proteine, die während des Spleißprozesses eine bedeutende Rolle spielen. Sie binden an spleißregulatorische *cis*-Elemente der prä-mRNA, die exonisch und intronisch lokalisiert sein können (Abb. 5). SR-Proteine unterstützen die Assemblierung des Spleißosoms, vermitteln Erkennung und Wahl der Spleißstellen und können als Spleißfaktoren die Wahl der Spleißstellen beeinflussen [30,85,86].

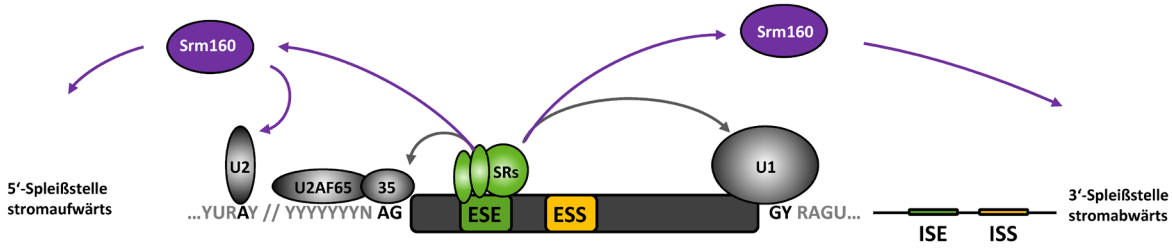


Abbildung 5: Spleißregulatorische Elemente einer prä-mRNA mit Interaktionsmöglichkeiten gebundener SR-Proteine. In den Introns bzw. Exons der prä-mRNA können *cis*-Elemente lokalisiert sein, die wichtig für eine korrekte Identifizierung von Spleißstellen sind und sich von den klassischen Spleißsignalen unterscheiden. Die Bindung von Proteinfaktoren an diese Elemente hat stimulierende (ESE/ISE „exonic/intronic splicing enhancer", in grün gekennzeichnet) oder hemmende Effekte (ESS/ISS „exonic/intronic splicing silencer", in gelb gekennzeichnet) auf das Spleißen. SR-Proteine (grün) binden meist an ESE-Elemente [84,87] und können stromaufwärts und stromabwärts gelegene Spleißstellen beeinflussen. Auch indirekte Interaktionen über den Coaktivator Srm160 (lila) sind möglich [88]. Srm160 interagiert mit dem U2 snRNP und hat stimulierende Effekte [87]. SR-Proteine können weiterhin die Aktivität benachbarter „silencer"-Elemente antagonisieren [89]. In Anlehnung an Cartegini *et al.* 2002 [27].

In *A. thaliana* wurden Beispiele für AS von SR-Proteinen bereits untersucht [90,91]. Es ist außerdem bekannt, dass Spleißvarianten von Spleißfaktoren Auswirkungen auf das AS selbst und dessen Regulation haben können [41]. Durch subtiles AS könnte somit dem Spleißprozess der Pflanze eine zusätzliche Flexibilität verliehen werden, um Transkriptom und Proteom an unterschiedliche Erfordernisse, wie beispielsweise während der Entwicklung oder unter Stress, anpassen zu können. Dies bildete die Motivation, um meine Analyse des AS an NAGNAG-Tandem-Spleißstellen in *A. thaliana* zu vertiefen und gezielt SR-Protein-kodierende Gene und deren Spleißverhalten unter verschiedenen Bedingungen zu untersuchen.

Die Analyse des subtilen AS wirft auch in bioinformatischer Hinsicht Fragen auf. Zur Identifizierung und Lokalisierung alternativer Spleißprozesse werden Transkriptdaten unterschiedlicher Art herangezogen. Die in der NCBI „Reference Sequence Database" (RefSeq) annotierte Transkriptionseinheit eines Genes wird überwiegend nur durch eine mRNA maximaler Länge repräsentiert und weist somit in den meisten Fällen keine auf AS hinweisende Redundanz auf (http://www.ncbi.nlm.nih.gov/RefSeq). RefSeq-Einträge werden anhand von Standards manuell generiert und werden qualitätskontrolliert. In der Sequenzdatenbank GenBank werden dagegen für jedes Gen alle verfügbaren mRNA-Sequenzen gespeichert, die durch Wissenschaftler weltweit erhoben und eingereicht wurden (http://www.ncbi.nlm.nih.gov/GenBank). Hier verantwortet der einzelne Wissenschaftler die Validität der Daten. Daneben gibt es weitere Transkriptdaten, die durch Hochdurchsatz-

cDNA-Sequenzierprojekte generiert werden. Diese sogenannten ESTs („expressed sequence tags“) repräsentieren in der Mehrheit nur Abschnitte von mRNA und werden ohne Qualitätskontrolle als Rohdaten direkt in Datenbanken abgelegt. Im Gegensatz zu RefSeq-weisen EST-Daten eine hohe Redundanz auf. Da die Qualität der ESTs schwankt, besteht die Gefahr des „technischen Rauschens“, indem Variationen durch Sequenzierfehler vorgetäuscht werden. Um diese Gefahr zu minimieren, wurden in vielen Studien Filterkriterien angewendet, die Trankripte mit minimalen Variationen herausfiltern und unberücksichtigt lassen. Somit waren diese Ansätze nicht geeignet, mithilfe von EST-Datensätzen subtile alternative Spleißereignisse umfassend und verlässlich zu identifizieren [56,92-96]. Wie kann man jedoch technisches Rauschen von subtilen natürlichen Ereignissen unterscheiden? Welche Informationen können durch stringente Filterkriterien gewonnen bzw. verloren gehen? In der Arbeit „Violating the splicing rules: TG dinucleotides function as alternative 3’ splice sites in U2-dependent introns“ [16] wird ein Lösungsweg aufgezeigt.

Aus Studien ungewöhnlicher Spleißsignale war bekannt, dass eine seltene Population von Introns existiert, die auf die terminalen Dinukleotide AT-AC enden und 0,7-0,9% aller Introns in Vertebraten ausmachen [97]. Diese werden von einem speziellen Spleißosom, dem U12-Spleißosom, prozessiert. Die Erkennung der 3’-Spleißstelle ist beim U12-Spleißsosom eher unspezifisch, sodass verschiedene zusätzliche Varianten von Intron-Termini existieren [98,99]. Im Gegensatz dazu ist die Wahl der 3’-Spleißstelle des bei weitem häufigeren Spleißosoms (U2-Spleißsosom), von dem die Mehrzahl aller Introns gesplissen wird, hoch spezifisch. Daher existieren neben den Introns mit GY-AG-Termini [97,99,100] auch nur sehr seltene Ausnahmen (GA-AG- und [101] AT-AC-Intron-Termini [99,102,103]). Unter diesen Ausnahmen stellen die AT-AC-Termini bis dato die einzigen belegten Nicht-AG-Spleißstellen dar, die in U2-gesplissenen Introns vorkommen. Hinweise aus einigen *in silico* Studien ließen jedoch das Vorhandensein weiterer ungewöhnlicher Spleißstellen vermuten [104,105]. Eine im Jahr 2000 durchgeführte systematische Analyse von Spleiß-stellen in Säugetier-Genomen konnte keine gesicherten Beweise erbringen [100]. Es gab jedoch vereinzelte Literaturhinweise, dass im humanen G-Protein-Gen *GNAS* („adenylate cyclase-stimulating G alpha protein“) [106,107], im *DLG4*-Gen („discs, large homolog 4“) des „presynaptic density“ Proteins 95 [108] und im Dopamin-Rezeptor-D2-Gen *DRD2* („dopamine receptor 2“) [109] in unmittelbarer Nachbarschaft eines AG-Dinukleotids überraschenderweise auch ein TG als alternative 3’-Spleißstelle verwendet wird (Abb. 6). Dieser Befund wurde bis dato nicht im Detail untersucht und unterstrich damit die Notwendigkeit genomweiter Ansätze, um in den Transkriptom-Daten des Humangenoms

gezielt zu suchen. In den nachfolgenden Analysen beziehe ich mich ausschließlich auf U2-abhängige Introns.

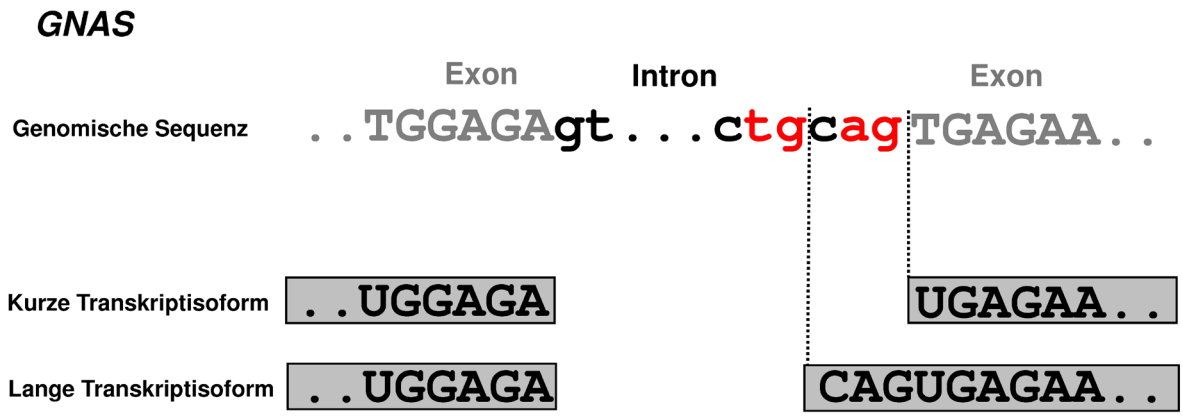


Abbildung 6: Alternatives Spleißen am Intron 3 im humanen *GNAS*-Gen. *GNAS* kodiert für die primäre stimulatorische Komponente der Adenylatcyclase, das GTP-bindende Protein $G\alpha_s$. Im Intron 3 wird nicht nur eine AG-3'-Spleißstelle, sondern auch eine alternative TG-3'-Spleißstelle verwendet. Bemerkenswerterweise zeigt das homologe Gen in *D. melanogaster* das gleiche ungewöhnliche Spleißereignis in einem anderen Intron [106].

Es erschien also möglich, dass noch eine weitere, ungewöhnliche Form des subtilen AS an 3'-Spleißstellen existiert. Dies würde der etablierten GY-AG-Regel widersprechen und bildete somit die Motivation der Arbeit „Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns" [16]. An dieser Stelle ergeben sich weitere Fragen: Was genau verursacht die Verletzung der Spleißregel? Gibt es Auffälligkeiten, die auf einen besonderen Mechanismus hindeuten? Damit steht auch diese Arbeit in enger Beziehung zu der grundsätzlichen Frage: Kann subtiles AS von biologischer Bedeutung sein oder wird es lediglich durch eine Ungenauigkeit des Spleißosoms verursacht [9]?

Der experimentelle Nachweis und die Quantifizierung von Spleißvarianten bildeten ein wichtiges Element meiner Studien des subtilen AS. Experimentell erhobene quantitative Daten sind entscheidend, um Spleißereignisse zu charakterisieren und deren funktionelle Rollen zu eruieren. Im Zuge dessen habe ich zwei für die Quantifizierung von Spleiß-varianten mit kleinen Längenunterschieden bis dato selten eingesetzte Methoden für diesen Zweck adaptiert. Die PSQ, eine bioluminometrische Sequenziertechnik [110], wurde ursprünglich für Genotypisierungen entwickelt [111-114]. Da sie neben qualitativer auch

quantitative Sequenzinformation bietet, kann sie für die Detektion und Quantifizierung von Spleißvarianten angepasst werden. Mittels CE-LIF ist es möglich, DNA-Fragmente mit einem Größenunterschied von nur einem Nukleotid (nt) präzise aufzutrennen und in einem Elektropherogramm zu visualisieren [115,116]. Anhand der Flächen der spezifischen Fluoreszenz-Peaks können Populationen von Nukleinsäuremolekülen quantifiziert werden. In meinem kürzlich eingereichten Manuskript „Comparison of methods for splice variant quantification“ [14] habe ich meine Arbeiten zu methodischen Aspekten der Spleiß-variantenquantifizierung zusammengefasst und systematisch verglichen. Als Referenz-Methode habe ich die weit verbreitete PAGE-ED herangezogen [70,117]. Es sollte untersucht werden, welche Vorzüge, Nachteile und Limitierungen die drei Methoden aufweisen und welche Aspekte bei der Konzipierung von Experimenten beachtet werden müssen. Da niedrige Template-Konzentrationen infolge von niedriger exprimierten Genen bzw. die Bedingungen der RT-PCR zu verzerrten Resultaten führen können, habe ich auch diese Einflussfaktoren getestet.

PUBLIKATIONEN UND MANUSKRIPTE

STEFANIE SCHINDLER
KAROL SZAFRANSKI
MICHAEL HILLER
GUL SHAD ALI
SAIPRASAD G. PALUSA
ROLF BACKOFEN
MATTHIAS PLATZER
ANIREDDY S.N. REDDY

Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes

BMC Genomics 2008, 9:159.

NAGNAG-Tandem-Spleißstellen stellen die häufigste Formen aller subtilen alternativen 3'-Spleißstellen in Säugetieren dar. Eine häufige Präsenz konnte auch in der Pflanze *Arabidopsis thaliana* nachgewiesen werden. AS an diesen Tandem-Spleißstellen wurde mithilfe der vorhandenen Transkriptdaten und einer sequenzbasierten Vorhersagemethode festgestellt. Interessanterweise sind diese Tandem-Spleißstellen in den SR-Protein-Genen überrepräsentiert, die für wichtige Spleißfaktoren kodieren. Da Spleißvarianten von Spleißfaktoren Auswirkungen auf den Spleißprozess, dessen Regulation und Ergebnis haben können, wurden aus dieser Gruppe Kandidaten für eine experimentelle Analyse unter verschiedenen Bedingungen ausgewählt. Diese zeigte, dass die differentiellen Effekte eher durch organ- und bedingungspezifische Änderungen des Spleißosoms als durch genspezifische Spleißregulation verursacht werden.

BMC Genomics

BioMed Central

Research article

Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes

Stefanie Schindler*[1], Karol Szafranski[1], Michael Hiller[2], Gul Shad Ali[3], Saiprasad G Palusa[3], Rolf Backofen[2], Matthias Platzer[1] and Anireddy SN Reddy[3]

Address: [1]Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany, [2]Institute of Computer Science, Bioinformatics Group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany and [3]Department of Biology and Program in Molecular Plant Biology, Colorado State University, Fort Collins, CO, USA

Email: Stefanie Schindler* - sschindl@fli-leibniz.de; Karol Szafranski - szafrans@fli-leibniz.de; Michael Hiller - hiller@informatik.uni-freiburg.de; Gul Shad Ali - gsali@lamar.colostate.edu; Saiprasad G Palusa - saigoud@lamar.colostate.edu; Rolf Backofen - backofen@informatik.uni-freiburg.de; Matthias Platzer - mplatzer@fli-leibniz.de; Anireddy SN Reddy - reddy@colostate.edu

* Corresponding author

Published: 10 April 2008

BMC Genomics 2008, **9**:159 doi:10.1186/1471-2164-9-159

Received: 20 September 2007
Accepted: 10 April 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/159

Abstract

Background: Several recent studies indicate that alternative splicing in Arabidopsis and other plants is a common mechanism for post-transcriptional modulation of gene expression. However, few analyses have been done so far to elucidate the functional relevance of alternative splicing in higher plants. Representing a frequent and universal subtle alternative splicing event among eukaryotes, alternative splicing at NAGNAG acceptors contributes to transcriptome diversity and therefore, proteome plasticity. Alternatively spliced NAGNAG acceptors are overrepresented in genes coding for proteins with RNA-recognition motifs (RRMs). As SR proteins, a family of RRM-containing important splicing factors, are known to be extensively alternatively spliced in Arabidopsis, we analyzed alternative splicing at NAGNAG acceptors in SR and SR-related genes.

Results: In a comprehensive analysis of the *Arabidopsis thaliana* genome, we identified 6,772 introns that exhibit a NAGNAG acceptor motif. Alternative splicing at these acceptors was assessed using available EST data, complemented by a sequence-based prediction method. Of the 36 identified introns within 30 SR and SR-related protein-coding genes that have a NAGNAG acceptor, we selected 15 candidates for an experimental analysis of alternative splicing under several conditions. We provide experimental evidence for 8 of these candidates being alternatively spliced. Quantifying the ratio of NAGNAG-derived splice variants under several conditions, we found organ-specific splicing ratios in adult plants and changes in seedlings of different ages. Splicing ratio changes were observed in response to heat shock and most strikingly, cold shock. Interestingly, the patterns of differential splicing ratios are similar for all analyzed genes.

Conclusion: NAGNAG acceptors frequently occur in the Arabidopsis genome and are particularly prevalent in SR and SR-related protein-coding genes. A lack of extensive EST coverage can be compensated by using the proposed sequence-based method to predict alternative splicing at these acceptors. Our findings indicate that the differential effects on NAGNAG alternative splicing in SR and SR-related genes are organ- and condition-specific rather than gene-specific.

Background

Alternative splicing is an important mechanism for regulating gene expression at the post-transcriptional level and contributes to proteome complexity [1-3]. This widespread process comprises various mechanisms such as exon skipping, mutually exclusive exons, intron retention, or the usage of alternative 5' or 3' splice sites [4]. Alternative splicing has been extensively studied in mammals but less in plants. Recent evidence indicates more than 60% of the genes in the human genome alternatively spliced [5] compared to about 20–30% in plants [6,7], based on EST/cDNA data. A hallmark of plant introns is their relatively short length (~150 vs. ~740 nt in humans, on average) [8] and Uridine-richness [9]. Furthermore, plant introns exhibit a weaker polypyrimidine tract than mammals [2,9]. Datasets of spliced alignments from the TIGR [6,10] and RIKEN [11] databases of full-length cDNAs and ESTs provide useful annotated versions of the Arabidopsis genome sequence for the detection of various alternative splice events. Based on the final TIGR annotation release, a total of 26,207 genes are annotated in Arabidopsis [12]. Although splicing machinery is generally conserved between plants and animals [2,9], plants exhibit a much higher fraction of retained introns (more than 40% of the events) compared to ~10% reported for humans [5-8].

The fidelity of intron excision from a pre-mRNA relies on the precise recognition of exonic and intronic sequence signals and the complex interplay of different spliceosomal RNAs and proteins. Among these, SR proteins direct splice site selection by recognizing splice sites and splicing regulatory sequences (enhancers and silencers), thereby facilitating spliceosome assembly [13,14]. SR proteins are important factors for constitutive and alternative splicing. This evolutionary conserved protein family contains structurally related proteins that possess one or two RNA-recognition motifs (RRM) at the N-terminus and a C-terminal arginine/serine-rich (RS) domain [15,16]. A recent genome-wide survey on Arabidopsis splicing-related genes revealed variations in SR proteins and hnRNP proteins between plants and mammals, suggesting plant-specific differences in splicing-regulation mechanisms [17]. The *A. thaliana* genome encodes 19 SR proteins, almost twice as many as in humans [18,19]. They can be subdivided into seven families [20]. Whereas SF2/ASF, 9G8 and SC35 are orthologues between plants and metazoa, the RS, RS2Z, SCL and SR45 subfamilies seem to be plant-specific. Most of the SR protein genes are themselves alternatively spliced to a great extent in Arabidopsis [20,21]. Fifteen of the 19 genes coding for SR proteins in Arabidopsis undergo alternative splicing and produce at least 95 transcripts [21]. In some cases, it was shown that alternative splicing correlates with the intron length [22]. Splicing patterns of Arabidopsis SR protein genes are under tight spatio-temporal control, leading to a different abundance of splice variants in different tissues and at developmental stages [21,23-26]. Several plant SR proteins have been shown to regulate the splicing of their own transcripts and transcripts of other SR genes [25,27-29]. Environmental conditions can also modulate the splicing pattern of a gene, as shown by the temperature dependent alternative acceptor selection of *SR1B*/*SR1* in Arabidopsis [30]. Furthermore, stresses such as exposure to cold, heavy metals or anaerobiosis, affect the efficiency or patterns of splicing [21,31-33], but the mechanisms by which some types of stress influence alternative splicing in plants are largely unknown.

In plants and mammals, the most frequent distance between alternative acceptors is 3 nt [6,34]. Such tandem acceptors have been termed NAGNAG acceptors based on the existence of a NAGNAG acceptor motif (N = A, C, G, T) [35,36]. In the NAGNAG motif, the upstream acceptor is termed the E-acceptor (since the downstream NAG becomes exonic upon splicing at this site) and the downstream one the I-acceptor (since the whole tandem becomes intronic) [36]. Alternative splicing at NAGNAG acceptors is widespread in many species [37,38] and also in plants [6,39] with *Caenorhabditis elegans* [36] being the only exception known so far. The selection of either AG in the splicing process results in the insertion/deletion (indel) of the I-acceptor NAG in mRNAs. This leads to diverse effects at the protein level with the majority of the events involving the indel of a single amino acid. A fraction of these events is estimated to be under purifying selection, suggesting an evolutionary conserved function [40]. Interestingly, it was demonstrated that the distribution of NAGNAG acceptors is highly similar between mammals and plants, for example, polar amino acid residues were found to be predominantly affected in both kingdoms [36,41].

An example for functional alternative splicing in Arabidopsis is a TAGCAG acceptor affecting the RNA-binding domain of the U11–35K protein that results in different binding affinity for SR proteins and the U11 snRNA *in vitro* [42]. In contrast, both splice variants derived from a CAGCAG acceptor in the tomato prosystemin gene are active signaling components of the wound response pathway, without detectable functional differences [39].

Previously, we found that human genes coding for RNA binding proteins including many splicing factors are preferentially equipped with NAGNAG acceptors [36]. Here, we observed a similar overrepresentation of NAGNAG acceptor motifs in Arabidopsis. This agrees with a very recent study, where NAGNAG alternative splicing was also found to be accumulated in genes for RNA-binding proteins in Arabidopsis [41].

Since splice variants of splicing factors may have consequences for alternative splicing and its regulation [43], we investigated alternative NAGNAG splicing at SR and SR-related genes. We determined the splicing ratios of 15 NAGNAG acceptors in splicing factors for several plant tissues, seedlings of different ages and in response to cold and heat stresses. We detected organ-specific variations and differences between the developmental stages. Cold stress was found to induce the most remarkable changes in the splicing ratios.

Results

NAGNAG acceptors are frequent in the Arabidopsis genome

A comprehensive list of introns was constructed from the annotated Arabidopsis genome sequence based on RIKEN [11] and TIGR [6,10] cDNA sequences. Out of 112,934 intron-exon boundaries (taken from 26,207 annotated protein-coding genes), 6,772 showed a NAGNAG motif within 5,381 genes (Additional table 1). Thus, 6% of all introns and 21% of all annotated genes in Arabidopsis harbor a genomic NAGNAG acceptor motif. For comparison, in human, 5% of introns and 30% of genes harbor such a motif [36]. We categorized all Arabidopsis cases according to their EST coverage (Additional table 2). In 229 cases (3%), no EST support exists for either of the possible acceptor sites. In 1,899 cases (28%), a single EST supports either acceptor. Out of the remaining 4,644 cases with minimally required EST coverage (two or higher, 69%), 242 cases (5%) have supporting ESTs for both acceptor sites. Naturally, EST-based evidence for alternatively spliced NAGNAGs depends on their isoform frequencies and the EST coverage, which is low in Arabidopsis compared to other species such as human or mouse. For example, if a minor isoform occurs with 10% frequency, at least 29 ESTs are necessary to reach a probability of 95% that it will be found (binomial test). Hence, in many cases, native alternative splicing remains undetected, and certainly more NAGNAG sites than those indicated by the current transcript data are expected to be alternatively spliced.

In order to overcome this limitation of EST coverage we established a sequence-based prediction method for alternative splicing at NAGNAG acceptors. There is evidence that a narrow context of flanking nucleotides captures most of the information relevant for prediction of the splice variant ratio [44,45]. Conservatively, we chose a heptameric context NAGNAGN, comprising the two acceptor AG dinucleotides and three additional variable positions, and divided all NAGNAG cases into 64 heptamer classes. The EST counts within each of the classes were pooled, and the resulting splicing variant ratio (fraction of E-transcripts) was considered representative for all cases of that heptamer class. For example, the average frequency of the E-isoforms of 55 observed CAGCAGA acceptors, based on 227 pooled ESTs, is 48%, and this was taken as the predicted frequency for any CAGCAGA acceptor motif (Additional Table 3). The validity of the heptamer-based approach is corroborated by the finding that maximum-likelihood estimators mostly agree between models for Arabidopsis and human. The high level of agreement is explained by the basic finding that the splicing ratios follow the basic rules of sequence preferences seen for isolated 3' splice sites: position -3 with $C \geq T > A > G$, position +1 with $G \geq A > T > C$ (data not shown).

Applying this method to the 2,128 cases with insufficient EST coverage (less than two ESTs), 482 (23%) are predicted to have a minor transcript frequency of at least 10%. Using this conservative threshold for isoform abundance gives a lower-bound estimate of the fraction of alternatively spliced NAGNAG sites. Applying this prediction method to all NAGNAG cases in Arabidopsis, 14% are predicted to be alternatively spliced with a minor transcript frequency of at least 10%, 21% with ≥ 5%, and 33% with ≥ 2%, respectively. Interestingly, as EST coverage increases, NAGNAG acceptors are less often predicted to be alternatively spliced (<2 ESTs: 23%, 2–5 ESTs: 11%, >5 ESTs: 8%). These results indicate that the occurrence of alternatively spliced NAGNAG acceptors is negatively correlated with the transcript levels of the genes.

Many SR and SR-related protein transcripts contain NAGNAG acceptors

For identification of SR and SR-related genes we searched for characteristic protein signatures in the gene products associated with NAGNAG acceptors [46]. Of all Arabidopsis proteins, 84 proteins had RRM domains and are rich for R/S dipeptides. Of these 84, 19 were previously identified as SR proteins [18,19], leaving 65 SR-related proteins. The intersection with NAGNAG cases gave 36 introns in 30 genes (Table 1). Thus, 36% of SR and SR-related protein-coding genes exhibit NAGNAG acceptors (7 out of 19 SR, 23 out of 65 SR-related). This is significantly higher than the average frequency of NAGNAG-containing genes (21%), even if we account for a higher fraction of multi-exon genes and a slightly higher number of introns in the SR/SR-related gene family ($P = 0.068$, permutation test). This finding is very similar to human where alternatively spliced NAGNAG motifs were found to be enriched in RRM-containing proteins [36].

SR33/*SCL33* is the only case which exhibits EST support for alternative NAGNAG splicing (Table 1). Intriguingly, in 14 cases, the sequence-based prediction argues for the usage of both acceptor sites with a predicted minor transcript frequency of 2%. This permissive 2% threshold was applied in narrowing the list of experimental candidates in order to retain those which have a substantial chance to

Table 1: NAGNAG acceptors in Arabidopsis SR and SR-related protein-coding genes. In summary, 36 NAGNAG-containing introns occur in 30 genes. Genes are classified into SR and SR-related protein coding genes. Splicing ratios are given as absolute EST counts ("#"). Column 'heptamer motif' specifies the heptamer sequence of the NAGNAG acceptor sites used for the sequence-based prediction; here, "|" marks the annotated acceptor. Predicted E-transcript proportions are listed in column 'E-transcript predicted'. Gene names are grey shaded if they contain two NAGNAG acceptors.

Gene	Name	SR	SR-related	Intron	# ESTs E-transcript	# ESTs I-transcript	Heptamer motif	E-transcript predicted (%)
At5g52040	*RS41*	*x*		2	5	17	AAG|CAG,G	7
At4g31580	*RSZ22*	*x*		3	15	0	TAG|GAG,C	99
At3g13570	*SCL30a*	*x*		3	13	0	TAG|GAG,G	99
At1g55310	*SR33/SCL33*	*x*		3	0	8	CAG,CAG|A	48
At1g16610	*SR45*	*x*		7	0	12	CAG,CAG|G	48
At1g16610	*SR45*	*x*		9	0	12	AAG,CAG|G	7
At1g23860	*SRZ21*	*x*		3	13	0	CAG|GAG,A	100
At2g24590	*SRZ22a*	*x*		3	8	0	CAG|AAG,A	98
At1g07350			*x*	2	21	0	TAG|GAG,A	100
At1g22910			*x*	1	7	0	TAG|GAG,C	99
At1g53650	*CID8*		*x*	4	7	0	CAG|AAG,G	97
At1g60000	*cp29*		*x*	1	6	0	CAG|GAG,T	100
At1g60900	*U2AF65*		*x*	2	3	0	CAG|GAG,A	100
At1g60900	*U2AF65*		*x*	4	0	2	TAG,CAG|G	17
At1g76940			*x*	1	2	0	TAG|AAG,G	83
At2g24350			*x*	4	0	2	AAG,CAG|T	7
At2g24350			*x*	1	5	0	CAG|GAG,T	100
At2g35410	*cp33*		*x*	3	16	0	TAG|GAG,T	100
At2g37220	*cp29*		*x*	2	22	0	TAG|GAG,T	100
At2g43370	*U11–35K*		*x*	4	0	6	TAG,CAG|G	17
At2g43370	*U11–35K*		*x*	5	5	0	CAG|GAG,T	100
At3g23830	*GR-RBP4/GRP4*		*x*	3	15	0	CAG|GAG,T	100
At3g26420	*ATRZ-1A*		*x*	3	5	0	TAG|AAG,G	83
At3g51950			*x*	4	6	0	CAG|GAG,G	100
At3g54230			x	5	0	3	GAG,CAG|G	1
At3g54230			*x*	6	0	3	TAG,CAG|C	25
At4g35785			*x*	4	0	0	CAG,AAG|T	98
At4g36960			*x*	8	5	0	TAG|GAG,A	100
At5g02530			*x*	2	6	0	TAG|GAG,A	100
At5g03580	*PABP*		*x*	1	0	1	GAG,TAG|G	0
At5g09880			*x*	1	5	0	CAG|AAG,A	98
At5g44200	*CBP20*		*x*	1	4	0	CAG|GAG,A	100
At5g44200	*CBP20*		*x*	7	6	0	CAG|GAG,G	100
At5g47320	*RPS19*		*x*	3	8	0	TAG|GAG,T	100
At5g53180	*PTB*		*x*	7	0	7	GAG,TAG|G	0
At5g59950			*x*	1	0	11	TAG,CAG|A	7

be alternatively spliced. We selected 15 SR and SR-related protein-coding genes for experimental analysis, including *SR33/SCL* as a positive control (Table 2).

Experimental evidence for NAGNAG isoforms in SR and SR-related protein genes

For experimental detection of splice variants, cDNA from different adult plant organs (root, leaf, stem, inflorescence) and from callus and seedlings of different ages (3d, 5d, 10d, 15d) was sampled to cover a broad spectrum of transcript sources. Three independent RT-PCRs were performed per cDNA sample and gene, and splice variants were separated by capillary electrophoresis and subsequently quantified based on fluorescence intensity. We considered a NAGNAG candidate as alternatively spliced if the measurements indicated an average minor transcript frequency of at least two times the standard deviation. In a conservative approach, we evaluated the averages of the plant samples, in order to avoid extreme values from single samples that could cause false positives.

Eight cases of SR and SR-related protein-coding genes were found to be alternatively spliced at their NAGNAG acceptor sites (53% of 15 tested, 22% of NAGNAGs in this family; Table 2, Additional Table 4). In addition, the alternative splicing patterns of *RS41* and *SR33* were independently confirmed by Sanger sequencing of at least 100 clones (data not shown). Using the same approach, alter-

Table 2: Experimental candidates with corresponding E-transcript proportions based on EST and experimental data. EST ratios are given as absolute EST counts. Column 'predicted' lists the E-transcript proportions obtained from the sequence-based prediction. Grey shaded values mark the cases where NAGNAG alternative splicing was validated (avg [minor isoform in organs, seedlings] > 2× avg [error]). At4g35785 is lacking EST data. Therefore, an EST-based E-transcript frequency cannot be shown and is indicated by '-'.

		E-transcript (%)	
Gene	**EST**	**Predicted**	**Experiments**
RS41	23	7	16.7 ± 0.6
SR33/SCL33	0	48	29.9 ± 0.7
SR45i7	0	48	0.6 ± 0.5
SR45i9	0	7	0.4 ± 0.6
SRZ22a	100	98	99.5 ± 0.6
CID8	100	97	99.1 ± 0.5
At1g76940	100	81	98.8 ± 0.9
At2g24350	0	7	3.2 ± 1.5
ATRZ-1A	100	83	98.9 ± 1.4
At3g54230	0	25	2.3 ± 0.5
At4g35785	-*	98	97.4 ± 0.3
At5g09880	100	98	99.0 ± 0.7
At5g59950	0	7	2.8 ± 0.4
U1 1–35K	0	17	8.5 ± 0.9
U2AF65	0	17	5.8 ± 0.6

native NAGNAG splicing could not be detected in *SR45i9* and *SRZ22a*, consistent with the quantitative capillary electrophoresis results. It is noteworthy that the relatively high frequency of alternative splicing in *SR33/SCL33* was not indicated by the eight ESTs that exist for that transcript region.

In the tested cases, the E-acceptor was found to represent the minor acceptor in nearly all cases, and the prediction for alternative splicing was more often accurate for the major-I subclass compared to major-E subclass. This is consistent with the global case distribution evident from EST data, which divides into 13% constitutive I, 17% alternative major-I, 13% alternative major-E, 57% constitutive E cases (based on the 2%-abundance threshold).

Genome-wide, the sequence-based prediction method suggested that 33% of NAGNAGs are alternatively spliced, producing minor isoforms with at least 2% frequency. For the 15 SR/SR-related cases that fulfill this prediction criterion, 53% were actually validated by our experiments. The prediction accuracy is positively correlated with the predicted minor transcript frequency. For example, cases predicted to have 2–5% minor transcripts are validated with a rate of 25% whereas cases predicted to have more than 5% minor transcripts are validated with a rate of 64% (Table 2). Consequently, a threshold of 2% seems to fully capture the fraction of likely alternatively spliced NAGNAG acceptors, as was intended for the selection of experimental candidates. Unfortunately, independent measures for the fraction of non-SR protein genes that undergo alternative NAGNAG splicing do not exist. However, based on the prediction results, we expect that SR/SR-related protein genes have a slightly higher propensity for alternative splicing (42% versus 33%).

Organ-specific alternative splicing of NAGNAG acceptors and differential splicing ratios during development

Splicing patterns of Arabidopsis SR protein genes are under tight spatio-temporal control, leading to a different abundance of splice variants in different tissues and at developmental stages [21,23-26]. Thus, we considered the occurrence of possible differences in the splice variant distribution in various plant organs (root, leaf, stem and inflorescence) and in callus. Based on the prior results, we similarly tested the cDNAs from those candidates, where the NAGNAG alternative splicing was successfully validated. In four cases (Figure 1, Table 3), a significant organ-specificity was observed (ANOVA, Table 3). Interestingly, inflorescence tissue shows reduced splicing of the minor acceptor (mostly E-acceptor) in nearly all cases. This trend is also observed for the NAGNAG cases that do not show significant organ-specific splicing.

Next, we asked if the splicing ratios at the NAGNAG acceptors exhibit developmental variations. To this end, we analyzed cDNA derived from seedlings at the ages of 3d, 5d, 10d and 15d (Table 4, Figure 1). In four analyzed cases, statistically significant splicing ratio changes could be detected (ANOVA, Table 4). Comparing the values of the 3d, 5d, 10d and 15d probes, our results show a general trend towards increased minor acceptor usage with seedling development.

NAGNAG splicing ratios under heat and cold shock

Finally, we examined whether temperature stresses can modulate the NAGNAG splicing pattern, as was previously illustrated by temperature-controlled splicing ratios of *SRp34/SR1* and other SR transcripts [21,30]. Seedlings were kept in hot or cold conditions and compared to an untreated control. Rather slight splicing ratio changes could be observed in the heat-shocked probes (Figure 1, Table 5). However, this difference was statistically significant in four cases (ANOVA, Table 5). In contrast, more obvious splicing ratio changes could be detected under cold shock (Figure 1). In six cases, a significant rise in minor acceptor usage was observed (ANOVA, Table 5), that clearly increased with the duration of treatment.

Discussion

SR proteins are important regulatory splicing factors and facilitate the correct interplay of components of the splicing machinery. Alternative splicing of SR protein genes is

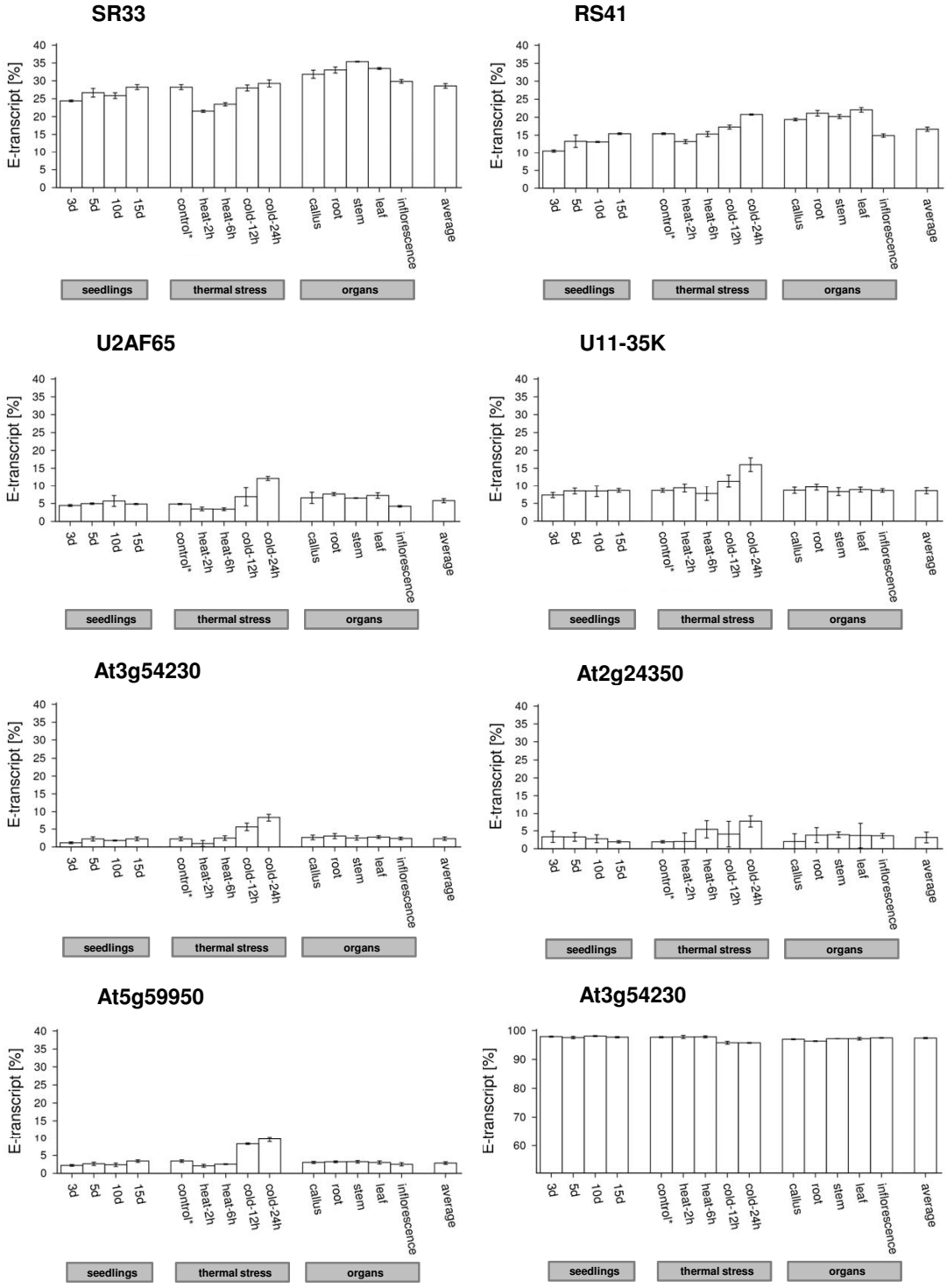


Figure 1

E-transcript proportions in SR and SR-related protein-coding genes under several conditions. E-transcript proportions among different seedling ages, under heat and cold shock and in different organs are depicted. The 15d old seedlings also serve as the control for the thermal stress treatments, indicated by '*'.

Table 3: E-transcript proportions in plant organs. The E-transcript frequencies in plant organs (root, leaf, stem, inflorescence) and callus are illustrated. Values were obtained from three independent experiments. Column 'ANOVA' displays the p-values, '+' $p \leq 0.05$ and '++' $p \leq 0.01$.

	E-transcript(%)					
Gene	**Callus**	**Root**	**Stem**	**Leaf**	**Inflor**	**ANOVA**
RS41	19.4 ± 0.4	21.1 ± 0.8	20.2 ± 0.6	22.0 ± 0.7	14.9 ± 0.5	++
SR33/SCL33	31.8 ± 1.2	33.1 ± 0.9	35.4 ± 0.2	33.5 ± 0.3	29.8 ± 0.6	++
At2g24350	2.0 ± 2.2	3.8 ± 2.1	3.9 ± 0.8	3.7 ± 3.5	3.6 ± 0.7	
At3g54230	2.7 ± 0.7	3.0 ± 0.8	2.5 ± 0.7	2.8 ± 0.4	2.4 ± 0.4	
At4g35785	97.0 ± 0.2	96.3 ± 0.2	97.2 ± 0.1	97.2 ± 0.5	97.5 ± 0.2	+
At5g59950	3.0 ± 0.3	3.2 ± 0.3	3.2 ± 0.4	3.0 ± 0.5	2.5 ± 0.5	
U11–35K	8.7 ± 0.9	9.6 ± 1.0	8.3 ± 1.2	8.8 ± 0.7	8.6 ± 0.5	
U2AF65	6.5 ± 1.6	7.6 ± 0.5	6.5 ± 0.1	7.2 ± 0.8	4.2 ± 0.3	+

able to confer a spatial flexibility to the architecture of the spliceosome and thus may influence the splicing process and its outcome. Subtle changes in the protein composition induced by alternative splicing at NAGNAG acceptors could contribute to this flexibility as previously suggested [36]. Here we explored the degree of alternative splicing at NAGNAG acceptors in Arabidopsis in general and of SR and SR-related protein-coding genes in detail. In a genome-wide *in silico* screening we identified 6,772 introns that exhibit a NAGNAG acceptor motif. Out of this group, we identified 36 introns within 30 SR and SR-related protein-coding genes. Intriguingly, NAGNAG acceptor motifs are more frequent in Arabidopsis SR and SR-related protein-coding genes (36%) than on average (21%). This is equivalent to the situation in human [36].

EST and mRNA data are the main resources to identify and locate alternative splicing of a gene. The total number of ESTs for a respective gene correlates with the diagnostic power of ESTs. Due to the relatively low EST coverage of the Arabidopsis genome, the EST data alone is not sufficient for a comprehensive characterization of the alternative splicing of NAGNAG motifs. For example, guided by the sequence-based prediction we could experimentally show an E-transcript frequency of 30% for *SR33/SCL33* despite initially lacking EST evidence for alternative splicing. This illustrates that the limitations of a low EST coverage can be at least in part circumvented with an appropriate prediction method. Currently, the EST data provides evidence for alternative splicing of 5% (242 cases) of NAGNAG acceptors, which represents the lower bound for genome-wide estimates. On the other hand, our sequence-based prediction method suggested 33% of genes produce minor isoforms with at least 2% frequency. But these predictions were found to be too optimistic, with only 53% of the cases actually giving detectable amounts of NAGNAG isoforms. Extrapolating these results to genome scale, 17% of Arabidopsis NAGNAGs are likely to be alternatively spliced. All prediction work neglects possible differences between tissues or developmental stages. In fact, our results for SR and SR-related protein genes indicate that organ- and development-specific, as well as stress-induced differences exist. However, the mechanisms that underlie tissue-specific regulation of alternative splicing are not yet understood and are not predictable by any current method.

Table 4: E-transcript proportions among different developmental stages. The E-transcript frequencies in seedlings of different ages (3d, 5d, 10d, 15d) are presented. Values were obtained from three independent experiments. Column 'ANOVA' displays the p-values, '+' $p \leq 0.05$ and '++' $p \leq 0.01$.

	E-transcript(%)				
Gene	**3d**	**5d**	**10d**	**15d**	**ANOVA**
RS41	10.6 ± 0.4	13.3 ± 1.8	13.1 ± 0.2	15.4 ± 0.3	+
SR33/SCL33	24.4 ± 0.3	26.7 ± 1.3	25.9 ± 0.8	28.2 ± 0.7	+
At2g24350	3.3 ± 1.6	3.3 ± 1.3	2.8 ± 1.1	1.9 ± 0.4	
At3g54230	1.2 ± 0.3	2.2 ± 0.6	1.8 ± 0.2	2.3 ± 0.6	+
At4g35785	97.9 ± 0.2	97.6 ± 0.5	98.1 ± 0.2	97.7 ± 0.3	
At5g59950	2.2 ± 0.3	2.6 ± 0.5	2.4 ± 0.5	3.4 ± 0.4	+
U11–35K	7.3 ± 0.8	8.5 ± 0.8	8.4 ± 1.6	8.6 ± 0.6	
U2AF65	4.4 ± 0.3	4.9 ± 0.3	5.7 ± 1.6	4.8 ± 0.2	

Table 5: E-transcript proportions under heat and cold shock. The E-transcript frequencies of seedlings kept in hot and cold conditions (2 h vs. 6 h and 12 h vs. 24 h, respectively) compared to untreated seedlings. Values were obtained from three independent experiments. Column 'ANOVA' displays the p-values, '+' $p \leq 0.05$ and '++' $p \leq 0.01$.

	E-transcript(%)					ANOVA	
Gene	Untreated control	Heat shock 2 h	Heat shock 6 h	Cold shock 12 h	Cold shock 24 h	Heat shock	Cold shock
RS41	15.4 ± 0.3	13.2 ± 0.6	15.3 ± 0.8	17.3 ± 0.6	20.7 ± 0.2	+	++
SR33/SCL33	28.2 ± 0.7	21.5 ± 0.4	23.5 ± 0.5	28.0 ± 0.9	29.3 ± 1.0	++	
At2g24350	1.9 ± 0.4	2.0 ± 2.4	5.4 ± 2.5	4.1 ± 3.6	7.7 ± 1.7		
At3g54230	2.3 ± 0.6	1.0 ± 0.9	2.5 ± 0.7	5.6 ± 1.1	8.2 ± 1.1		+
At4g35785	97.7 ± 0.3	97.8 ± 0.6	97.8 ± 0.4	95.8 ± 0.6	95.7 ± 0.2		+
At5g59950	3.4 ± 0.4	2.1 ± 0.5	2.5 ± 0.2	8.2 ± 0.3	9.6 ± 0.8	+	++
U11–35K	8.6 ± 0.6	9.4 ± 1.2	7.8 ± 2.0	11.3 ± 1.8	16.0 ± 2.0		+
U2AF65	4.8 ± 0.2	3.5 ± 0.6	3.4 ± 0.4	6.9 ± 2.6	12.1 ± 0.6	+	+

We found a negative correlation of the occurrence of alternatively spliced NAGNAG acceptors with the transcript levels of the genes. Though this finding needs further validation, it suggests that genes with high transcript abundance are not representative for the transcriptome. This would have profound consequences for studies extrapolating from highly expressed genes to the remaining transcriptome.

The effects of splicing factors are often dependent on their concentration, localization and phosphorylation, resulting in gradual changes of the alternative splicing pattern of certain transcripts [2,16]. Thus, splicing ratio changes, leading to differential abundance of splicing factor isoforms, could enhance the flexibility of the spliceosome composition and the splicing process itself. Hence, we asked whether this is the case for the genes shown to have an alternatively spliced NAGNAG. We experimentally tested several organs, developmental stages and environmental influences. Interestingly, significant organ-specific differences of splicing ratios were detected in four cases. Most notably, inflorescence showed reduced splicing of the minor acceptor in nearly all experimental candidates. A very similar effect was seen in early developmental stages (3d compared to later stages). A common reason may be that stem cells, enriched in both these samples, disfavor minor acceptor usage. This should be further tested in future experiments. Finally, the most pronounced effect on the splicing ratio was seen after cold shock, consistent with previous observations [21,47].

For the analyzed gene family, it seemed reasonable to ask for the impact of NAGNAG splicing on the RRM domain. We found that none of the eight NAGNAG acceptors in SR proteins do affect the RRM. In contrast, 12 of the 23 SR-related proteins have a NAGNAG acceptor located in the RRM domain. Previously, functional differences due to a NAGNAG acceptor in the RRM were observed for the U11–35K protein [42] and, more generally, NAGNAG alternative splicing in RRM-containing proteins was suggested to have an impact on the tertiary structures [41]. Also the usage of the E-acceptor site results in a protein with one additional serine in *SR33/SCL33* and *RS41*. Serine residues in SR proteins are the targets of phosphorylation, and numerous studies have shown that the phosphorylation status of SR proteins is critical for their splicing activity as well as subcellular localization [2,14,16,27].

Most notably, the pattern of differential splicing ratios is similar for all analyzed genes. Thus, the differential effects on NAGNAG alternative splicing seem to be organ- and condition-specific rather than gene-specific. This favors the hypothesis that differential splicing of NAGNAG acceptors is mostly independent of sequence-specific splicing regulators, and is rather mediated by (subtle) organ- and condition-specific differences of the spliceosomal core composition. Intuitively, such lack of tight regulation seems to argue against a functional relevance of splice variants, as was suggested earlier [44]. However, several tandem splice sites with clear functional implications exhibit constant splicing ratios. Vice versa, it was shown that alternative splicing events producing variable splicing ratios do not always imply a function [48]. Surely, the functional relevance of the alternative splice events analyzed in this study remains to be evaluated.

Conclusion

We demonstrated, that NAGNAG acceptors frequently occur in the Arabidopsis genome and are particularly prevalent in SR and SR-related protein-coding genes. Insufficient EST coverage can be compensated using the sequence-based method to predict alternative splicing of NAGNAG acceptors. The observed differential effects on NAGNAG alternative splicing appear to be organ- and condition-specific rather than gene-specific. In particular, inflorescence and early seedling stages consistently show reduced levels of the minor transcript isoforms.

Methods

Screening for NAGNAG acceptor tandems

The annotated genome sequence of *Arabidopsis thaliana* was obtained from GenBank, to serve as a data basis for the locations of intron-exon boundaries and their sequence. Boundaries with the sequence NAGNAG| or NAG|NAG (where "|" indicates the annotated boundary) were sampled. Redundancies due to annotation of multiple transcript isoforms were filtered. Potential splice variants derived from the genomic NAGNAG patterns were detected and quantified by a WU-BLASTN search of 60-nt sequence windows around the resulting exon-exon junctions against all Arabidopsis ESTs from TIGR and RIKEN databases [49,50], using parameters W = 13 N = -8 nogap S = 180 hspmax = 1. BLAST matches were considered valid if perfect sequence identity was found in a 12-nt window around the exon-exon junctions [51].

Prediction of splicing ratios

All NAGNAG-containing introns with supporting EST data for E- and I-acceptor were divided into subsets of 64 motif classes, according to the heptameric motif NAGNAGN. Maximum-likelihood estimators for E-to-I transcript ratios were calculated by combining the EST counts per class. In order to prevent a bias caused by cases with an extremely high EST coverage, counts were limited to a maximum of 10 per isoform per NAGNAG site, and eventually downscaled.

Identification of SR/SR-related protein genes

The complete set of non-redundant Arabidopsis proteins was screened for existing RNA-recognition motifs, and its derivatives, using Pfam HMM definitions (PF00076, PF04059, PF08777) and hmmsearch (HMMer package, [52]) applying recommended cutoff parameters. Additionally, the relative content of RS or SR dipeptides of each gene product was determined. A significance threshold >0.016 for R/S-richness was applied, corresponding to the transition point of a two-exponential case distribution. A subset of 84 proteins had both significant RRM profile hits and R/S-rich sequence. Of these 84, 19 are identified as SR proteins *sensu strictu* [18,19].

Plant material and stress treatments

A. thaliana ecotype Columbia seedlings were grown on Murashige and Skoog (MS) medium at 22°C with 16 h/8 h light/dark cycle and harvested after 3d, 5d, 10d and 15d. Callus tissue was generated from roots of one-week-old seedlings by transferring them onto a callus induction medium (1× Gamborg's B5 medium, 2% glucose, 0.5 g/l MES (pH 5.7), 0.8% agar, 0.5 mg/l 2,4-D [2,4-Dichlorphenoxyacetate] and 0.005 mg/l kinetin). Callus tissue was collected and frozen in liquid nitrogen for RNA extraction. Heat and cold stress treatments were done with 15d-old seedlings. Seedlings were grown for 15-days and exposed to heat (38°C) for two and six hours or cold (4°C) for 12 and 24 hours, the untreated control seedlings were kept at 22°C for the corresponding time period.

RT-PCR and splice-variant analysis by quantitative capillary electrophoresis

RNA from plant tissues, seedlings or callus was isolated using RNeasy Plant Mini Kit (Qiagen) and quantified spectrophotometrically at 260 nm. RNA was treated with DNAseI and used to synthesize first strand cDNA with oligo (dT) primer using SuperScriptII (Invitrogen). For validation of splice variants, three independent RT-PCRs for each candidate were performed with cDNA from different organs, developmental seedling stages and stress treatments to yield amplicons covering the respective exon-exon junction. Reactions were set up with BioMix Red (Bioline, Randolph, USA) and 10 pmol primer in 50 μl total volume, according to the manufacturer's instructions. Each forward primer was labelled with 6-carboxyfluorescein (FAM) for subsequent analysis on a capillary sequencer. The thermocycle protocol was 1 min 30 sec initial denaturation at 94°C, followed by 35 cycles of 50 sec denaturation at 94°C, 45 sec annealing at 55–59°C, 1 min extension at 72°C, and a final 1 h extension step at 72°C. The following gene-specific primers were used: *RS41* reverse GCTGGCGGCGAACGAGA, *RS41* forward GAGAAGGGAAAGCAGGAGTC, *SR33* forward GCTGCTGATGCAAAACATC, *SR33* reverse CTCCCATCATATCGCTCTTC, *SRZ22a* forward CGTGGTGGTTCTGATTTGAAG, *SRZ22a* reverse GATCTAGCACGAGGGCTGTAA, *SR45i9* forward GTCGCTCTCGTTCAAGTTCC, *SR45i9* reverse TTTACGAGGTGGAGGTGGTG, *SR45i7* forward AGGCCGTTCTCCATCTTCTC, *SR45i7* reverse CCTTCTGGGACTTGGTGAAC, *At2g24350* forward CTGCGCTCTGTCATTGTTTC, *At2g24350* reverse ACATGAGGCTCCGTTTCTTG, *At1g53650* forward AGTTCTTCGCTTTGCGTTTG, *At1g53650* reverse GCAGGCAGACTGAAAGAAGG, *At5g09880* forward GGAAGAGAAGGAACCCGAAG, *At5g09880* reverse CCATTGGAACTGACATCACG, *At5g59950* forward TGGATGGAAAACCCATGAAG, *At5g59950* reverse ACCACCTCGTTGTTGACCTC, *At1g76940* forward ACATCATCCTCCTGGTGGTC, *At1g76940* reverse CCACCTTCTCCTGATTGCAC, *At2g43370* forward GGAGCTTCAC GAGGATATGG, *At2g43370* reverse CTCAGGCGGAAGCT GAATAC, *At4g35785* forward ATCTCCTTCACCCCGAAA AG, *At4g35785* reverse CAAGACGCAACCTTTCCTTC, *At1g60900* forward GCGCCTCCTGATATGTTAGC, *At1g60900* reverse AGGCCACCAACATAGACTCG, *At3g54230* forward GGGTCCTTTGCATCATGTTC, *At3g54230* reverse ACATCCGCTGAAGGAGAATC. The PCR products were appropriately diluted (1/20 to 1/50) and 1 ul was supplemented with 10 ul formamide (Roth, Karlsruhe) and 0.5 ul of GeneScan 500 LIZ (Applied Biosystems). The mixture was than separated on an ABI 3730 capillary

sequencer and analyzed with the GeneMapper 4.0 software. The E-transcript proportion (%) was calculated as follows: peak area for the E-isoform/(E-isoform+I-isoform) × 100.

Splice-variant analysis by clone-counting

For validation of splice variants (*RS41*, *SR33*, *SRZ22a*, *SR45*), RT-PCR was performed with cDNA from root, leaf, stem and inflorescence to yield amplicons covering the respective exon-exon junction. The following gene-specific primers were used: *RS41* forward 5'-AAG AGG AGG GAA AGC AGGAG-3' and reverse 5'-GCG ATT TCG AAT GGA GTC AT-3'; *SRZ22a* forward 5'-GCA AGA ATG GAT GGA GGG TA-3' and reverse 5'-CCA CGA GGA GAA GGA CTA CG-3'; *SR33/SCL33* forward 5'-AGG GTT TGG GTT CGT TCA AT-3' and reverse 5'-CTC CGT GAC CGA GAT CTA CC-3'; *SR45* forward 5'-CAC CTC CAA GGA GAC TAC GC-3' and reverse 5'-CAG TGG CCT CTT AGG ACT GC-3'. PCR products were gel purified using the QIAquick Gel Extraction Kit and the isolated fragments were cloned into pCR2.1-TOPO (Invitrogen) according to the supplier's recommendations. 25 clones per gene and plant sample were selected and Sanger-sequenced using M13 standard reverse primer (20-mer). Sequence analysis was performed using SPIDEY [53].

Authors' contributions

SS performed the capillary electrophoresis experiments, analyzed and interpreted data and wrote the initial manuscript. KS conceived and designed experiments, performed the statistical analyses, interpreted data and contributed to the manuscript. MH performed the genome-wide screening. GSA and SGP performed experiments, analyzed and interpreted data. ASNR, RB and MP as principal investigators conceived the experiments. All authors contributed to the final manuscript.

Additional material

Additional file 1

All NAGNAG acceptor cases identified within the Arabidopsis genome. *The absolute numbers of ESTs supporting the E- or the I-acceptor ("ESTs_E" and "ESTs_I", respectively) and the sequence-based prediction of E-transcript frequency ("expected_E") are given.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-159-S1.xls]

Additional file 2

Counts of Arabidopsis NAGNAG cases depending on local EST coverage. *The number of cases at least reaching a given coverage is presented.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-159-S2.xls]

Additional file 3

Heptamer motif classification. *Heptameric motif classes are presented with corresponding maximum-likelihood estimators for E-to-I-transcript ratios, used for sequence-based prediction.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-159-S3.xls]

Additional file 4

Comprehensive list of experimental values from all analyzed SR and SR-related protein-coding genes. *Column 'avg' lists the averaged E-transcript proportions derived from three independent experiments per probe with corresponding standard deviation in column 'sd'. The values derived from the 3d, 5d, 10d, 15d old seedlings, callus and organs were averaged (column 'avg organs-seedlings') as well as the corresponding standard deviations ('avg error') to gain appropriate values for a comparison with the sequence-based predictions (see column 'predicted'). Grey shaded values mark the cases where NAGNAG alternative splicing was validated [(avg error × 2) < (avg organs-seedlings)].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-159-S4.xls]

Acknowledgements

This work was supported by grants from the German Ministry of Education and Research (0313652D) and the Deutsche Forschungsgemeinschaft (SFB604-02) to M.P as well as the Department of Energy (DE-FG02-04ER15556) to A.S.N.R.

References

1. Lareau LF, Green RE, Bhatnagar RS, Brenner SE: **The evolving roles of alternative splicing.** *Curr Opin Struct Biol* 2004, **14(3):**273-282.
2. Reddy AS, N : **Alternative splicing of pre-messenger RNAs in plants in the genomic era.** *Annu Rev Plant Biol* 2007, **58:**267-294.
3. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing.** *Gene* 2005, **344:**1-20.
4. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 2002, **3(4):**285-298.
5. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes.** *Nucleic Acids Res* 2007, **35(1):**125-131.
6. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR: **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis.** *BMC Genomics* 2006, **7:**327.
7. Wang BB, Brendel V: **Genomewide comparative analysis of alternative splicing in plants.** *Proc Natl Acad Sci USA* 2006, **103(18):**7175-7180.
8. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R: **Intron retention is a major phenomenon in alternative splicing in Arabidopsis.** *Plant J* 2004, **39(6):**877-885.
9. Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W: **Pre-mRNA splicing in higher plants.** *Trends Plant Sci* 2000, **5(4):**160-167.
10. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res* 2003, **31(19):**5654-5666.
11. Sakurai T, Satou M, Akiyama K, Iida K, Seki M, Kuromori T, Ito T, Konagaya A, Toyoda T, Shinozaki K: **RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome.** *Nucleic Acids Res* 2005:D647-650.
12. Moskal WA Jr, Wu HC, Underwood BA, Wang W, Town CD, Xiao Y: **Experimental validation of novel genes predicted in the**

un-annotated regions of the Arabidopsis genome. *BMC Genomics* 2007, **8:**18.
13. Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem* 2003, **72:**291-336.
14. Manley JL, Tacke R: **SR proteins and splicing control.** *Genes Dev* 1996, **10:**1569-1579.
15. Fu XD: **The superfamily of arginine/serine-rich splicing factors.** *RNA* 1995, **1(7):**663-680.
16. Graveley BR: **Sorting out the complexity of SR protein functions.** *RNA* 2000, **6(9):**1197-1211.
17. Wang BB, Brendel V: **The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing.** *Genome Biol* 2004, **5(12):**R102.
18. Lorkovic ZJ, Barta A: **Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana.** *Nucleic Acids Res* 2002, **30(3):**623-635.
19. Reddy AS: **Plant serine/arginine-rich proteins and their role in pre-mRNA splicing.** *Trends Plant Sci* 2004, **9(11):**541-547.
20. Kalyna M, Barta A: **A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions?** *Biochem Soc Trans* 2004, **32(Pt 4):**561-564.
21. Palusa SG, Ali GS, Redddy ASN: **Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins and its regulation by hormones and stresses.** *Plant J* 2007, **49:**1091-1107.
22. Kalyna M, Lopato S, Voronin V, Barta A: **Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins.** *Nucleic Acids Res* 2006, **34(16):**4395-4405.
23. Golovkin M, Reddy AS: **An SC35-like protein and a novel serine/arginine-rich protein interact with Arabidopsis U1–70K protein.** *J Biol Chem* 1999, **274(51):**36428-36438.
24. Lazar G, Schaal T, Maniatis T, Goodman HM: **Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF.** *Proc Natl Acad Sci USA* 1995, **92(17):**7672-7676.
25. Lopato S, Kalyna M, Dorner S, Kobayashi R, Krainer AR, Barta A: **atSRp30, one of two SF2/ASF-like proteins from Arabidopsis thaliana, regulates splicing of specific plant genes.** *Genes Dev* 1999, **13(8):**987-1001.
26. Lopato S, Waigmann E, Barta A: **Characterization of a novel arginine/serine-rich splicing factor in Arabidopsis.** *Plant Cell* 1996, **8(12):**2255-2264.
27. Ali GS, Reddy AS: **ATP, phosphorylation and transcription regulate the mobility of plant splicing factors.** *J Cell Sci* 2006, **119(Pt 17):**3527-3538.
28. Isshiki M, Tsumoto A, Shimamoto K: **The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA.** *Plant Cell* 2006, **18(1):**146-158.
29. Kalyna M, Lopato S, Barta A: **Ectopic expression of atRSZ33 reveals its function in splicing and causes pleiotropic changes in development.** *Mol Biol Cell* 2003, **14(9):**3565-3577.
30. Lazar G, Goodman HM: **The Arabidopsis splicing factor SR1 is regulated by alternative splicing.** *Plant Mol Biol* 2000, **42(4):**571-581.
31. Luehrsen KR, Taha S, Walbot V: **Nuclear pre-mRNA processing in higher plants.** *Prog Nucleic Acid Res Mol Biol* 1994, **47:**149-193.
32. Reddy AS, N : **Nuclear Pre-mRNA Splicing in Plants.** *Crit Rev, Plant Sci* 2001, **20:**523-571.
33. Simpson GG, Filipowicz W: **Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery.** *Plant Mol Biol* 1996, **32(1–2):**1-41.
34. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site.** *RNA* 2006, **12(12):**2047-2056.
35. Akerman M, Mandel-Gutfreund Y: **Alternative splicing regulation at tandem 3' splice sites.** *Nucleic Acids Res* 2006, **34(1):**23-31.
36. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet* 2004, **36(12):**1255-1257.
37. Ferranti P, Lilla S, Chianese L, Addeo F: **Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein.** *J Protein Chem* 1999, **18(5):**595-602.
38. Rogina B, Upholt WB: **The chicken homeobox gene Hoxd-11 encodes two alternatively spliced RNA species.** *Biochem Mol Biol Int* 1995, **35(4):**825-831.
39. Li L, Howe GA: **Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway.** *Plant Mol Biol* 2001, **46(4):**409-419.
40. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection.** *RNA* 2008 in press.
41. Iida K, Shionyu M, Suso Y: **Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals.** *Mol Biol Evol* 2008, **25(4):**709-718.
42. Lorkovic ZJ, Lehner R, Forstner C, Barta A: **Evolutionary conservation of minor U12-type spliceosome between plants and humans.** *RNA* 2005, **11(7):**1095-1107.
43. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements.** *Nature* 2007, **446(7138):**926-929.
44. Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A simple physical model predicts small exon length variations.** *PLoS Genet* 2006, **2(4):**e45.
45. Tsai KW, Tarn WY, Lin WC: **Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3' tandem splice site selection.** *Mol Cell Biol* 2007, **27(16):**5835-5848.
46. Birney E, Kumar S, Krainer AR: **Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors.** *Nucleic Acids Res* 1993, **21(25):**5803-5816.
47. Fung RW, Wang CY, Smith DL, Gross KC, Tao Y, Tian M: **Characterization of alternative oxidase (AOX) gene expression in response to methyl salicylate and methyl jasmonate pretreatment and low temperature in tomatoes.** *J Plant Physiol* 2006, **163(10):**1049-1060.
48. Hiller M, Platzer M: **Widespread and subtle: alternative splicing at short-distance tandem sites.** *Trends Genet* 2008 in press.
49. **RIKEN Arabidopsis Full-Length Clone Databse** [ftp://pfgweb.gsc.riken.jp/rafl/sequence/]
50. **The TIGR *Arabidopsis thaliana* Database** [http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=arab]
51. Thanaraj TA: **A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures.** *Nucleic Acids Res* 1999, **27(13):**2627-2637.
52. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.
53. Wheelan SJ, Church DM, Ostell JM: **Spidey: a tool for mRNA-to-genomic alignments.** *Genome Res* 2001, **11(11):**1952-1957.

STEFANIE SCHINDLER*
KAROL SZAFRANSKI*
STEFAN TAUDIEN
MICHAEL HILLER
KLAUS HUSE
NIELS JAHN
STEFAN SCHREIBER
ROLF BACKOFEN
MATTHIAS PLATZER

Violating the splicing rules: TG dinucleotides function as alternative 3‘ splice sites in U2-dependent introns

Genome Biology 2007, 8:R154.

TG/AG-Tandem-Spleißstellen wurden als eine neue, jedoch seltene Form des subtilen AS in einer kleinen Population von Introns im Menschen nachgewiesen und systematisch untersucht. Diese Spleißstellen verletzen die etablierte GY-AG-Spleißregel, da ein TG-Dinukleotid als 3‘-Spleißstelle verwendet wird. Das AS an diesen Tandem-Spleißstellen wurden experimentell in humanen Geweben validiert. TG-Spleißstellen treten ausschließlich mit einer zusätzlichen AG-Spleißstelle auf, von der sie maximal 28 nt entfernt sind. In Übereinstimmung mit einer überdurchschnittlich starken Konservierung von Spleißstelle und Sequenzkontext deutet dies auf einen speziellen Mechanismus bzw. auf die Beteiligung von *cis*- und/oder *trans*-Elementen hin.

*Beide Autoren lieferten gleichwertige Beiträge.

Research

Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns

Karol Szafranski[¤*], Stefanie Schindler[¤*], Stefan Taudien[*], Michael Hiller[†], Klaus Huse[*], Niels Jahn[*], Stefan Schreiber[‡], Rolf Backofen[†] and Matthias Platzer[*]

Addresses: [*]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstr., 07745 Jena, Germany. [†]Institute of Computer Science, Bioinformatics Group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee, 79110 Freiburg, Germany. [‡]Institute of Clinical Molecular Biology, Christian Albrechts University Kiel, Schittenhelmstr., 24105 Kiel, Germany.

¤ These authors contributed equally to this work.

Correspondence: Karol Szafranski. Email: szafrans@fli-leibniz.de

Published: 1 August 2007

Genome ***Biology*** 2007, **8:**R154 (doi:10.1186/gb-2007-8-8-r154)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/8/R154

Received: 8 March 2007
Revised: 14 June 2007
Accepted: 1 August 2007

Abstract

Background: Despite some degeneracy of sequence signals that govern splicing of eukaryotic pre-mRNAs, it is an accepted rule that U2-dependent introns exhibit the 3' terminal dinucleotide AG. Intrigued by anecdotal evidence for functional non-AG 3' splice sites, we carried out a human genome-wide screen.

Results: We identified TG dinucleotides functioning as alternative 3' splice sites in 36 human genes. The TG-derived splice variants were experimentally validated with a success rate of 92%. Interestingly, ratios of alternative splice variants are tissue-specific for several introns. TG splice sites and their flanking intron sequences are substantially conserved between orthologous vertebrate genes, even between human and frog, indicating functional relevance. Remarkably, TG splice sites are exclusively found as alternative 3' splice sites, never as the sole 3' splice site for an intron, and we observed a distance constraint for TG-AG splice site tandems.

Conclusion: Since TGs splice sites are exclusively found as alternative 3' splice sites, the U2 spliceosome apparently accomplishes perfect specificity for 3' AGs at an early splicing step, but may choose 3' TGs during later steps. Given the tiny fraction of TG 3' splice sites compared to the vast amount of non-viable TGs, *cis*-acting sequence signals must significantly contribute to splice site definition. Thus, we consider TG-AG 3' splice site tandems as promising subjects for studies on the mechanisms of 3' splice site selection.

Background

Intervening sequences (introns), primary transcript regions that are removed during mRNA maturation, are an outstanding feature of eukaryotic gene structure. Introns are excised through two transesterification reactions involving the collaboration of five different small nuclear ribonucleoprotein

particles and additional proteins that associate to form the spliceosome. Rearrangements of the spliceosome and, consequently, splicing catalysis is driven by the sequential action of ATP-dependent helicases [1,2]. The assembly of the early spliceosomal complex relies on sequence-specific contacts between the intron terminal regions and the spliceosome subunits U1, U2, and U2AF [1,3,4]. From accumulating intron sequence data, it was noted that invariant dinucleotides must represent important signals for the definition of intron termini, the so-called GT-AG rule (for simplicity, we use the nucleotide symbol T to denote thymidine in DNA as well as uridine in RNA sequences). With respect to the role of intron termini in the transesterification reactions, the 5' GT site and the 3' AG site were named donor and acceptor splice sites, respectively.

Early work on unusual splice signals revealed introns with the terminal dinucleotides AT-AC [5], and these were later shown to be processed by an independent splicing pathway, the U12 spliceosome. The U12 spliceosome recognizes highly specific donor site and branchpoint motifs [6] while recognition of 3' splice sites is rather unspecific. As a result, there are several variants of intron termini besides the prominent combinations GT-AG and AT-AC [7,8].

Among U2-dependent introns, the most frequent exception to the GT-AG rule are GC-AG intron termini, which comprise 0.7-0.9% of vertebrate introns [5,8,9]. Other rare exceptions are GA-AG intron termini in the FGFR gene family [10] and AT-AC termini, mostly found in introns of the SCN gene family [8,11,12]. While the latter cases are the only reported non-AG 3' splice sites, results from *in silico* studies have repeatedly suggested that other unusual 3' splice sites occur in U2-type introns [13,14]. An in-depth, systematic screening effort could not reveal significant evidence for additional unusual intron 3' termini above the noise level brought by annotation errors [9]. However, it was noted that a few exceptional U2-spliceosomal introns exist that involve unusual 3' splice sites in scenarios of alternative splice site choice. For example, intron 3 of the human guanine nucleotide binding protein gene *GNAS* is spliced at either TG or AG in the 3' intron sequence CTGCAG [15,16]. Remarkably, the homologous *Drosophila* gene shows the same unusual splicing pattern for another intron [17]. Moreover, unusual TG splice acceptors appear to be involved in alternative splicing of the human gene for presynaptic density protein 95, *DLG4* [18], and the human dopamine D2 receptor gene, *DRD2* [19].

We have previously reported a widespread type of alternative splicing mediated by the tandem splice acceptor motif NAGNAG [20]. From the analysis of single-nucleotide polymorphisms (SNPs) we concluded that a NAGNAG motif is necessary and sufficient to explain three-nucleotide variant splicing at intron-exon boundaries [21]. In contrast, alternative splicing of an intron 3' terminus in the *GNAS* gene appears to occur independently of a NAGNAG motif. Furthermore, it has been suggested that unusual splice sites could be selectively involved in alternative splicing [5,9], although this was never examined in detail. Here, we report a systematic screening of the human transcriptome that identified 36 introns with *bona fide* TG 3' splice sites. These TG splice sites are exclusively found as alternative 3' splice sites, each associated with a canonical AG 3' splice site. The evolutionary conservation of these introns and their alternative splicing patterns indicate physiological relevance and point to the requirement for *cis*-regulatory sequence elements to promote usage of TG 3' splice sites.

Results

Prior considerations

We used an *in silico* approach based on expressed sequence tags (ESTs) to identify unusual 3' splice sites that are found in pairs of 3' splice variants. ESTs, as first-pass results from high-throughput cDNA sequencing projects, are clearly prone to errors. Therefore, we assumed that single ESTs are insufficient to indicate genuine subtle splice variants since technical artifacts contribute to false positives. We considered a variant as sufficiently evident if it is supported by at least two independent ESTs, and expect the EST variant ratio to serve as an approximation for the natural ratio of splice variants. An additional threshold was applied to the relative abundance of splice variants, since our experimental approach, that is, sequencing of 100 individual RT-PCR clones, had a detection limit. Using a random binomial distribution to model the occurrence of splice variants in the RT-PCR clones, we calculated a diagnostic power of 95% (β error 5%) if a splice variant occurs with at least 3% frequency. It is important to note that we have not inferred anything for the cases that failed the threshold criteria. It is possible that they actually represent natural splice variants; however, the evidence for such cases is weak and the experimental approach did not provide sufficient sensitivity for validation.

TG dinucleotides function as non-canonical alternative 3' splice sites

We initially aimed to identify unusual 3' splice sites that are found in pairs of 3' splice variants that differ by 3 nucleotides (nt), such as in the *GNAS* intron 3 splice site tandem [15,16]. Identification of 3 nt splice variant pairs (Δ3SVPs) was based on 3' splice sites as indicated by spliced alignments of human ESTs [22]. After a reduction of false positives performed by a series of filtering steps (Figure 1), we identified 65 'unusual' Δ3SVPs that were supported by high-quality local EST alignments. Of these, 20 meet the requirements that the minor splice variant is supported by at least two ESTs and 3% of the matching ESTs (see considerations below). However, after close inspection and re-sequencing, we identified 6 of the 20 unusual Δ3SVPs as false positives (Additional data file 1), explained by: 3 nt deletion variants due to sequencing errors; mouse ESTs erroneously attributed to human; or alignment artifacts. Another six Δ3SVPs can be explained by SNPs,

Table 1

Unusual TG splice acceptors identified in the human transcriptome

	Intron		3' Splice site pair		ESTs for unusual 3' splice sites	
Gene	No.	Length	Distance	Motif	Fraction	No.
GNAS	3	7843	3	CTG,CAG|	0.15-0.62*†	282
PCGF2	1	224	3	AAG|ATG,	0.50	4
CNBP	3	168	3	TTGTTG,AAG|	0.25	257
	3	168	6	TTG,TTGAAG|	0.01	10
FBXO17‡	3	1999	3	CAG|ATG,	0.14	4
C21orf63	3	9975	3	CAG|ATG,	0.09	2
BRUNOL4	6	1147	3	CAG|CTG,	0.07	2
PCID2	2	2162	3	CAG|ATG,	0.04	7
TNNT2	1	4354	3	TTG,GAG|	0.04	2
CACNA1A	9	2532	3	TTGTTG,GAG|	?†	-
	9	2532	6	TTG,TTGGAG|	0.17†	-
GPBP1	9	1377	4	CAG|GATG,	0.03	2
KIAA0494	1	1459	5	TTG,AGCAG|	0.09	2
OSBPL8	2	36530	6	CTG,TTGTAG|	0.11	2
SAP30	1	1892	6	CTG,TTTCAG|	0.04	2
DRD2	6	1485	6	CTG,GTGCAG|	0.02†	5†
SUV420H2	5	134	7	CTG,GCTCCAG|	0.20	3
SSRP1	1	489	7	TTG,AATTCAG|	0.20	16
FREQ	2	16849	7	CTG,CCTCCAG|	0.04	2
IL21	3	2753	8	TTG,ATTTCTAG|	0.13	2
RYK	7	3107	9	TTG,GCTCCTTAG|	0.77	27
DLG4	5	131	9	CTG,GAGTTGCAG|	0.62	8
SMARCA4	29	6174	9	TTG,ACCCTGAAG|	0.41	34
FBXL10	15	177	9	TTG,GCCTACAAG|	0.21	3
HNRPR	7	2839	9	TTG,GTTTAACAG|	0.13	15
RRAD	1	214	9	CTG,ATCCCCTAG|	0.06	2
TGM1	6	454	10	CTG,TCCTGGGCAG|	0.13	2
ALAS1	11	1599	11	CTG,TTTCTCCTCAG|	0.04	5
ARS2	18	182	12	TTG,TACTCCCCCCAG|	0.74	75
PCBP2	7	1337	12	CTG,ACTCTCTCCCAG|	0.43	169
PTPN11	10	4269	12	TTG,GCTCTACTCCAG|	0.33	3
MSH5	6	164	12	CTG,ATCCCCTCCCAG|	0.25	5
SYTL2	9	1259	13	TTG,CCCTCCTGAGTAG|	0.09	3
TOMM40	1	95	16	CTG,ACCTCTCCCCTAGCAG|	0.07	2
MARK3	3	20478	17	TTG,TTTGTTTTTTTTTTTAG|	0.07	3
BAT3	6	832	18	CTG,ACTCTCCCCTACCTTCAG|	0.01	1
SH3D19	6	838	21	TTG,GTTTTGTTTTGGTCTCGTCAG|	0.07	1
LOC346653	1	3097	27	CTG,ACCCATGTACCTGAGGCTGATTTCCAG|	0.60	3
ACAD9	10	253	28	TTG,TTTCTTGTGTTTTTTCTGAACACTCCAG|	0.09	4

Entries in bold have RefSeq transcripts supporting the unusual TG acceptor site. Each TG splice variant is supported by at least two ESTs and at least 3% of all covering ESTs, except for some RefSeq-supported cases, *CACNA1A* [24,35], *DRD2* [19] and *BAT3*. In the 'Motif' column, a vertical line (|) indicates a canonical splice site, and a comma (,) marks the TG splice site. Splice ratios are given as absolute EST counts (No.) as well as the fraction of TG splice variants. A question mark indicates that an explicit fraction is not given in the referenced article, although the authors performed quantitative experiments. *EST ratio depends on the exon junction; the upstream exon 3 may be skipped. †Splice variants were previously quantified by others: *GNAS* [16,26], *CACNA1A* (splice ratio cited from [24,35]), *DRD2* (splice ratio cited from [19]). ‡Alternative splicing at *FBXO17* intron 3 was not experimentally reproducible in this study.

where the SNP allele corresponding to a NAGNAG splice site motif is not displayed by the human reference genome sequence [21]. Strikingly, all the remaining eight Δ3SVPs suggest that TG dinucleotides function as alternative 3' splice sites (Table 1).

Since all the unusual alternative Δ3-nt splice acceptors identified display TG dinucleotides, we investigated their occurrence in a wider scope. Analogously to the screen for Δ3SVPs (Figure 1), we performed a search for alternative TG splice acceptors at larger distances, up to 36 nt from from the canonical splice site. The same filter procedures were applied, and close inspection did not reveal obvious artifacts or

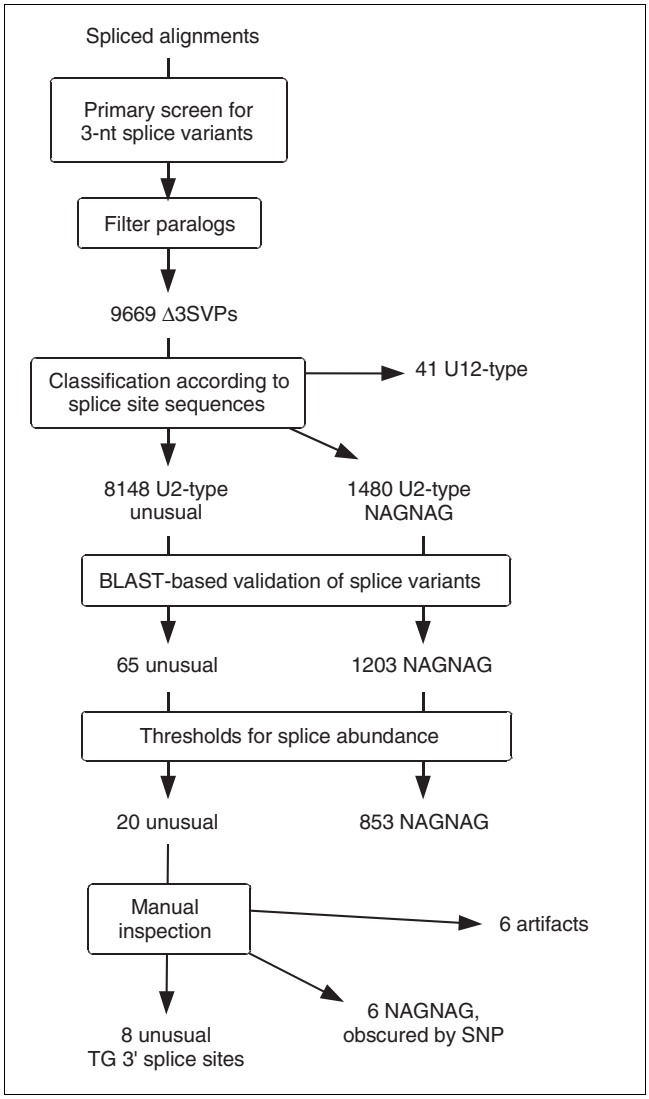


Figure 1
Screening procedure for unusual 3' splice sites found in pairs of 3' splice variants that differ by 3 nt (Δ3SVPs). Processing of AG-AG tandem cases ('NAGNAG', parallel branch on the right) was performed as a comparison to unusual 3' splice site tandems.

explanatory SNP alleles. We identified 26 additional EST-supported splice variant pairs that suggest alternative TG splice acceptors functioning in U2-dependent introns (Table 1).

We sought to screen for unusual splice acceptors using an independent approach in order to cross-validate our initial findings and to make a link to previous studies that were primarily based on curated transcript data [5,9]. An analysis of RefSeq-to-genome alignments identified 122 putative introns with terminal TG dinucleotides (120 unique genomic sites) out of 228,925 total introns (171,605 unique genomic sites). Of these, 39 introns have a canonical GT donor dinucleotide (Additional data file 1). A previous study, performing a similar screening approach for unusual splice sites using curated transcript data, showed high enrichment of annotation artifacts [9]. Therefore, we checked the identified TG acceptor cases thoroughly. In fact, cases failed this quality check for several reasons: known SNPs masking existing canonical AG splice sites; RefSeqs lacking transcript evidence; and misleading RefSeq-to-genome alignments. Since the overall false positive rate seemed very high, we additionally required independent transcript entries (mRNA or EST) to support the unusual splice site. In summary, 9 of the 39 RefSeqs showed robust support for unusual TG acceptor sites (Table 1, entries in bold), and 6 of these overlap with cases obtained from the EST screening approach while the others are exclusively identified by the RefSeq-based approach (*SH3D19*, *BAT3*, *CACNA1A*). Intriguingly, these three EST-independent RefSeq-supported cases all comprise 'alternative' splice sites, although this was not a screening criterion. Taking into account that about 1% of all introns have alternative 3' termini [23], this strongly indicates that TG splice acceptors are functionally linked to nearby AG splice sites and cannot function in a constitutive manner (P = 0.000001, binomial test). Altogether, the two screens identified 37 introns with 39 alternative TG splice acceptors (Table 1).

Negation of genome sequence errors and polymorphisms

Since six putatively unusual 3' splice sites can be explained by SNP-affected NAGNAG acceptors (which were filtered; Figure 1), we asked whether undiscovered SNPs, or even inaccuracies in the available human genome sequence, may explain some of the remaining candidates. The genomic sequence of the splice site regions of *GNAS* and *CACNA1A* had been experimentally verified by others [16,24]. For 10 other genes (listed in Table 2), we analyzed PCR products obtained from genomic DNA, pooled from 100 individuals, for sequence variations. The re-sequenced genomic regions were in perfect agreement with the available genome sequence, negating the possibility that unusual splice sites are trivial sequencing errors (data not shown). Moreover, we identified no SNP alleles that confer explanatory AG splice sites on any of the observed unusual splice variants, demonstrating that the TG splice sites are real and genetically invariant.

Validation and quantification of splice variants

To verify the existence of TG-derived splice variants, we performed RT-PCR experiments designed to yield amplicons that cover the exon-exon junctions under consideration. Cloning of the PCR products and sequencing allowed us to detect splice variants. Subclassification and counting of clones gave measures of splice variant ratios. This way, the alternative splicing pattern was reproduced and quantified for 11 out of 12 analyzed cases (Table 2). Generally, the splice ratios obtained from clone counting agree well with the EST data. The observed deviations can be explained by significant fluctuations depending on the analyzed tissue (*C21orf63*, *BRUNOL4*, and *CNBP* in Table 2). The splice variant valida-

Table 2

Validation and quantification of alternative splice variants

Splice junction		Tissue	Fraction of TG splice	Method
GNAS	Intron 3, exon junction 3-4	Leukocytes	0.14	n = 115
PCGF2	Intron 1	Placenta	0.99	n = 89
CNBP	Intron 3, indel AAG	Leukocytes	0.52	n = 69
		Brain	0.38	n = 58
		Placenta	0.33	n = 70
	Intron 3, indel TTGAAG	Leukocytes	0.01	n = 69
FBXO17	Intron 3	Leukocytes	-	n = 96
		Liver	-	n = 151
C21orf63	Intron 3	Leukocytes	0.09	n = 110
		Brain	-	Direct sequencing
BRUNOL4	Intron 6	Lung	-	n = 90
		Brain	0.19	n = 92
PCID2	Intron 2	Leukocytes	0.02	n = 142
TNNT2	Intron 1	Heart	0.03	n = 91
CACNA1A	Intron 9, indel GAG	Brain	0.85	n = 90
	Intron 9, indel TTGGAG	Brain	0.03	n = 90
ARS2	Intron 18	Leukocytes	0.55	n = 37
PCBP2	Intron 7	Leukocytes	0.54	n = 47
LOC346653	Intron 1	Testis	0.50	Direct sequencing

In the 'Methods' column, n represents the number of subclones sequenced.

tion failed for *FBXO17*, a gene for which 4 out of 29 ESTs had suggested a TG-derived splice variant. All supporting ESTs originated from the same EST library, NIH_MGC_100, derived from a hepatocellular carcinoma. A peculiarity of the source material, either the NIH_MGC_100 cell line or the single-individual liver sample used for our RT-PCR experiments, may be the reason for the inconsistent results concerning this putative splice variant. This example illustrates that at least two ESTs from independent sources are required to indicate a natural splice variant with high reliability. Overall, the success rate of the validation experiments was high (92%), and extrapolated to the 25 non-tested cases, about 2 false positives are expected.

Tissue-specificity of splice ratios

According to the results of the PCR cloning approach, *BRUNOL4* displayed remarkable tissue-specific splice ratios. The TG-derived splice variant was not detected in lung cDNA whereas the same variant constituted 20% of brain *BRUNOL4* transcripts (Table 2, Figure 2a). So we asked if splice ratios of TG-derived and AG-derived variants generally show tissue-specific differences. We analyzed the splice variant ratios more extensively in other genes using pyrosequencing, a method that allows accurate and cost-effective quantification of mixtures of polymorphic DNA populations [25,26]. *ARS2* and *CNBP*, both having a ubiquitous expression profile, show tissue-dependent fluctuations in splice ratios (55-65% and 20-40%, respectively; Figure 2). While these differences are numerically significant in each of these genes ($\alpha = 0.01$, ANOVA), their biological relevance is debatable. We conclude that splice ratios of TG-AG tandems are tissue-specific for particular introns but are rather stable for others.

Evolutionary conservation of introns with TG 3' splice sites

Since splicing at TG sites occurs in a very small number of introns, one might argue that these represent 'accidental' events attributable to spliceosome dysfunction. To address this question, we first analyzed the conservation of splice sites in homologous introns as an indication of alternative splice variants being under purifying selection. Out of 36 introns with 3' TG splice sites (37 minus the false-positive *FBXO17* intron), 26 (72%) are conserved between the human and mouse genomes. In 14 of these cases (39% of the total), mouse ESTs indicate homologous TG-derived splice variants. For comparison, this rate is three- to four-fold higher than that of alternative exons found in both human and mouse [27-29]. In some cases, EST evidence for orthologous TG-derived splice variants even exists for distantly related species, such as chicken (*CNBP*, *BRUNOL4*, *RYK*), and frog or fish (*BRUNOL4*, *RYK*, *FBXL10*). An outstanding example of conserved intron sequence and homologous splice variants is intron7 of the *RYK* gene (Figure 3). The ratio of 3' splice variants is remarkably similar between human and chicken, as can be inferred from the available EST data (EST ratios of 24:6 and 5:1 for human and chicken, respectively). In general, it should be noted that homologous splice variants may

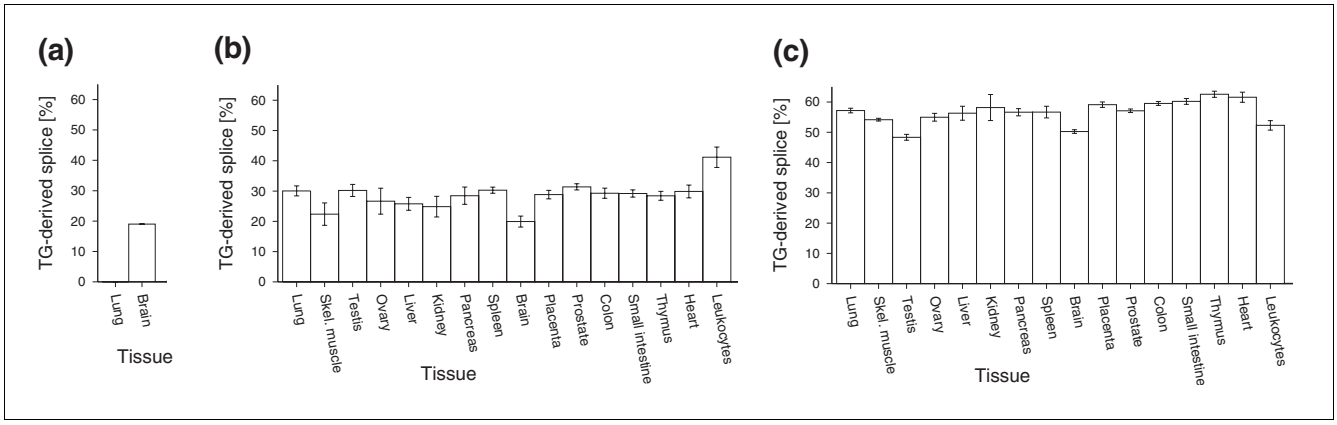


Figure 2
Tissue-specific fractions of TG-derived splice variants. **(a)** *BRUNOL4* (values are as shown in Table 2); **(b)** *CNBP*; and **(c)** *ARS2*. Pyrosequencing assays (for (b,c)) were performed multiple times for each sample (two to four times). Error bars depict the standard deviation of individual measurements.

remain undetected due to the limited depth of EST coverage [23,27].

Independently, we analyzed intron sequence conservation as an indication of the functional relevance of alternative splicing [27-29]. A data set of human-mouse orthologous intron-exon boundaries was used to determine the degree of conservation within a 50 nt intron sequence upstream of the splice acceptor, or acceptor tandem. Intronic flanks of TG splice sites show an average sequence similarity of 74%, whereas flanks of AG splice sites within canonical (AG-only) introns are 65% similar on average ($P << 0.00001$, permutation test). A plot of flanking sequence conservation against the abundance of the TG-derived splice variant (Figure 4) shows that

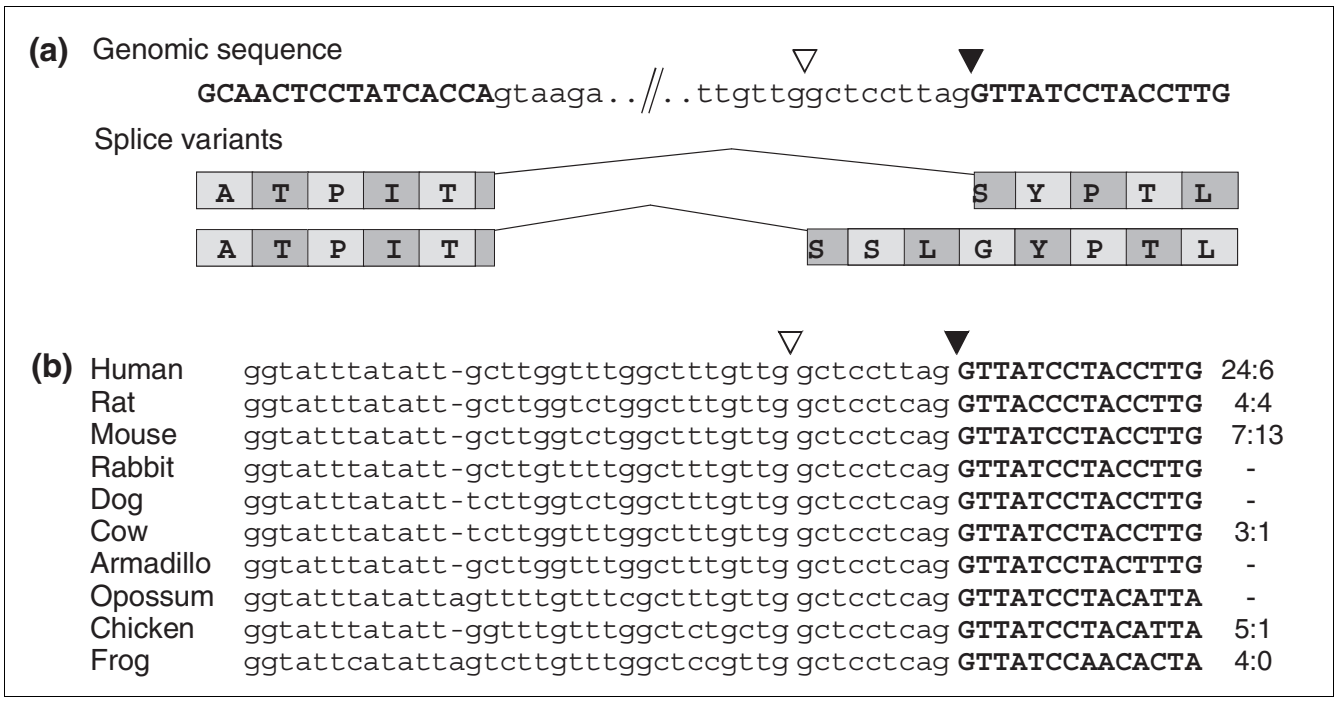


Figure 3
Conservation of the TG splice site found in intron7 of the *RYK* gene from human to chicken. **(a)** Human genomic sequence and derived splice variants. Canonical (filled triangle) and TG 3' splice site (open triangle) are marked. **(b)** Alignment of orthologous exon-intron boundary regions from several vertebrate genomes, splice sites highlighted as in (a). Numbers on the right display the ratios of species-specific ESTs for the TG and AG splice sites, respectively.

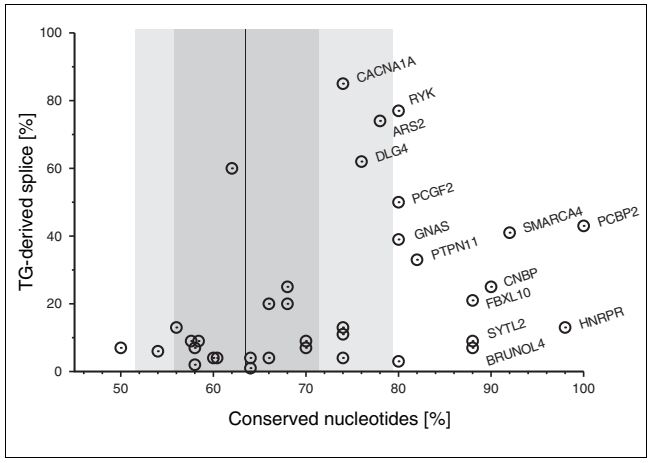


Figure 4
Intron flank conservation of TG-AG splice acceptor tandems. Orthologous human/mouse intron-exon boundaries involving TG splice sites are displayed in a two-dimensional plot according to two properties: horizontal axis = sequence identity of 50 nt sequence upstream of both splice sites; vertical axis = relative abundance of the TG-derived splice variant, as reflected by the fraction of TG-spliced ESTs (except for *CACN1A1*, where the data are taken from Table 2). Data points are labeled with the gene symbol if the conservation score and/or the fraction of TG-derived splice variant are significantly high. Conservation properties of canonical introns are indicated by shaded intervals: black line = median; dark gray = 66% percentile; light gray = 90% percentile.

these two measures are positively correlated. Introns that give rise to less than 10% TG-derived splice variants have an average human-mouse intron sequence identity of 64%, indistinguishable from canonical introns. In contrast, introns with a TG-derived splice variant making up more than 10% of the transcripts show an average sequence identity as high as 80%. This parallels a previous finding that the abundance of splice variants correlates with sequence conservation of alternative exons [27]. Consistently, high intron sequence conservation is strongly correlated with conservation of the TG splice site (13 of the 14 cases with gene labels in Figure 4). Our results indicate that splicing at TG acceptors may arise from neutral evolution, presumably showing low splicing efficiency. However, efficiently spliced TG 3' splice sites seem to evolve and to be maintained by evolutionary selection.

Structural and sequence characteristics of TG 3' splice sites

We analyzed the context properties of TG 3' splice sites in order to find an explanation for these rare exceptions to the GT-AG rule. With regard to the gene structure context, TG-splicing introns are indistinguishable from canonical introns in several respects: length of the affected as well as the downstream introns, and length of the upstream and downstream exons (data not shown). TG 3' splice sites are significantly often found in the first intron of the gene; 8 of 36 (22%) TG-AG introns compared to 11% of other introns (P = 0.02, Fisher's Exact). This bias is also found for AG-AG splice site tandems and is certainly due to neutral evolution of introns located in the 5' untranslated transcript region.

TG 3' splice sites were exclusively found within a context of alternative splice site choice (Table 1). This is clearly significant for results from the RefSeq-based screening procedure, which are unbiased with respect to constitutive or alternative transcripts. Taking into account that about 1% of all introns have alternative 3' termini [23], it strongly indicates that TG 3' splice sites are functionally linked to nearby AG splice sites and cannot function in a constitutive manner. This conclusion is further supported by studies of human AG>TG 3' splice site mutants [30-32], which always resulted in activation of neighboring AG splice sites, but not splicing at TG. Furthermore, we observed that TG-AG tandems display a splice site distance restriction with a limit of 28 nt, which is not seen for AG-AG tandems (Table 1, Additional data file 1). Thus, splicing of a 3' TG does not depend only on an additional AG splice site, this dependency also seems to pose a constraint on TG-AG splice site distance. The observed distance limit corresponds well with the distance between the branch site and the 3' splice site, which is typically 20-40 nt [33].

Splice site strength was scored using the maximum entropy method 'maxent' of Yeo and Burge [34]. The 5' splice site scores are indistinguishable between TG-AG introns and canonical introns. On the other hand, the AG 3' splice sites in TG-AG tandems score significantly lower than canonical introns (6.1 ± 3.5 versus 8.5 ± 2.8, respectively; P = 0.00001, Student's t-test). The 3' splice site score remains significantly small if low-scoring outliers are excluded from the analysis (6.7). Interestingly, the sequence context of TG 3' splice sites is very similar to canonical AG splice sites in that it shows a preference for pyrimidines at position -3 and preference for purines at position +1 (Additional data file 1). However, TG 3' splice sites changed to AG yield an average score of 6.6, again significantly lower than that of canonical AG splice sites (P < 0.00001, Student's t-test). This disfavors the simple explanation that TG and AG splice sites compete for the same recognizing factors and the neighboring nucleotide composition (that is, the feature scored by maxent) alone acts to direct splice site choice towards TGs. Finally, it remains questionable if maxent, trained on canonical AG 3' splice sites, has any predictive power for the functionality of TG splice sites.

The fraction of TGs functioning as splice acceptors is extremely small, about 0.01% of candidate motifs. Thus, *cis*-regulatory elements must play a crucial role in the definition of TG splice acceptors. From the ratio of functional/non-functional TG-AG tandems, in comparison with AG-AG tandems, we estimate that at least 6 nt of *cis*-regulatory sequence information is required to promote splice site usage of 3' TGs (Additional data file 1). This is in agreement with 5-20 unchanged nucleotides in excess over the average intron mutation rate, found in about half of the TG-splicing introns,

that is, those considered to be subject to purifying selection for splicing of 3' TGs. However, we failed to identify specific regulatory motifs (data not shown). This may be due to the dispersed arrangement of *cis*-regulatory elements, or the contextual cooperation of diverse elements. Due to the relatively small sample size for TG 3' splice sites, available methods for motif discovery have limited detection power.

Discussion

Previous studies provided incidental evidence for unusual 3' terminal dinucleotides in U2-dependent introns, particularly TG dinucleotides that are used as alternative 3' splice sites. Few directed efforts have been made so far to verify such instances and to elucidate underlying mechanisms and consequences [16,35]. Here, we report 36 human U2-spliceosomal introns with TG dinucleotides functioning as 3' splice sites, identified by thoroughly filtered EST-to-genome alignments. The high accuracy of the EST-based screening approach was validated by RT-PCR with a success rate of 92%. Though it might seem paradoxical, the analysis of EST data gave superior results compared to an analysis of curated data, that is, RefSeq transcripts. We found that the abundance of EST data allows the application of statistical methods for obtaining valid results whereas curated data sets, which are typically devoid of redundancy, may contain errors that are rarely captured by filtering criteria, consistent with the findings of others [36]. In practice, we found that two independent ESTs are strong evidence for a natural splice variant. Given this rather permissive threshold [9,37], we expect that the established screening protocol achieves high sensitivity.

Since our screening procedure is EST-based, certainly more unusual 3' splice sites remain undiscovered in transcript regions that lack sufficient EST coverage. Moreover, there are indications that even other unusual dinucleotides, apart from TG, may function as alternative 3' splice sites. For example, others reported an AT 3' splice site in the mammalian *DGCR2* gene [8], a CG 3' splice site in the *Drosophila per* gene [38], and we found that a TG splice acceptor in human *CNBP* intron 3 is replaced by a viable GG in the chicken ortholog (results not shown). The occurrence of a TG splice acceptor in the *Drosophila gnas* gene suggests that they occur throughout metazoan organisms.

Other studies have questioned the extent to which alternative splicing is functionally relevant [27-29]. Since TG splice acceptors are extremely rare compared to AG acceptors, one might think that these cases reflect a fuzziness of the splicing reaction. However, multiple findings support the idea that TG splice sites are activated by directed mechanisms and that the resulting splice variants fulfill functional roles: first, several TG splice acceptors are used with a high frequency or can even be the preferred splice site, which excludes splicing errors as a plausible explanation (Table 1, Figure 4); second, TG splice acceptors and their adjacent intron sequence are remarkably conserved between orthologous mammalian genes (Figure 4); third, tissue-specific splice patterns are observed for *GNAS* [16,26] as well as *BRUNOL4* (this study; Figure 2), suggestive of specific regulatory processes; and fourth, the TG splice site-mediated protein isoform of the mammalian calcium channel subunit α_{1A} (*CACNA1A*) has been shown to result in significant differences in neuronal excitability [35].

Thinking of splice site evolution as a process of functional engineering, we might ask about the functional options that distinguish TG-AG splice acceptor tandems from AG-AG tandems. During analysis of orthologs of human TG splice acceptors, we did not identify any case of orthologous AG splice sites, suggesting that TG and AG splice site dinucleotides are functionally non-equivalent. The inserted/deleted nucleotide sequence differs only if TG is positioned downstream of the tandem splice site. Apart from the possible impact on the protein sequence, an NAGATG tandem acceptor allows insertion of a start codon. For example, this seems to be realized in intron 1 of human *PCGF2*, where the observed splice variants differ by the presence of an upstream open reading frame. Preliminary results indicate that this ATG insertion has an effect on the translation efficiency of the mRNA (results not shown). It is also worth noting that the *Drosophila gnas* gene has a TG splice acceptor, like the human gene, but it is located in a non-homologous intron [17]. Given the overall low frequency of TG 3' splice sites (0.02%), this example of convergent evolution indicates a functional benefit of the unusual splice site, independent of its impact on protein sequence. It is tempting to speculate that splicing of TG splice acceptors, rather than providing a pathway for alternative transcripts or protein isoforms, may play a role as a regulatory bottleneck for maturation of the transcript, as was suggested for U12-type introns [39].

Considering functional classes, a significant fraction of TG-spliced genes represent regulators of chromatin structure (*PCGF2*, *GPBP1*, *SAP30*, *SUV420H2*, *SSRP1*, *SMARCA4*) as well as splicing factors and translational modulators (*CNBP*, *BRUNOL4*, *HNRPR*, *PCBP2*). Interestingly, two of the affected RNA-binding proteins are reported to bind DNA as well [40,41]. Together, these enrichments suggest a regulatory cross-talk between transcription on the one hand, and splicing, mRNA maintenance, and translation on the other. Together with another subgroup associated with receptor-mediated signal transduction (*GNAS*, *DRD2*, *FREQ*, *IL21*, *RYK*, *DLG4*, *RRAD*, *PTPN11*, *SYTL2*, *MARK3*, *SH3D19*), most of the genes' functions may be circumscribed with 'information processing', a term that was introduced to describe the functional characteristics of U12-dependent introns [6]. However, as a statistical analysis of Gene Ontology functional classification terms does not reveal any significant over- or under-representation (results not shown),

further work is required to determine the relevance of these findings.

TG-AG splice acceptor tandems illustrate the flexibility as well as the specificity of splice site selection by the U2-type spliceosome. The spliceosome is flexible enough to choose TG dinucleotides as splice acceptors. Despite this flexibility, a TG splice site depends on a neighboring AG splice acceptor, since constitutive TG splice acceptors are not found, and TG-AG acceptor tandems show a distance constraint. We assume that an AG splice acceptor, within the typical context of a branch-point motif and polypyrimidine tract, is essentially required for intron definition to promote splicing stepI *in vivo*. Consistent with this, a recent report showed that the essential splicing factor U2AF^{35} in cooperation with other factors mediates the spliceosome's specificity for AG 3' intron termini during splicing stepI [42]. Assuming that splicing stepI does not ultimately define the 3' splice site, we hypothesize that definite splice site choice takes place during reaction stepII, allowing TG dinucleotides to function as 3' splice sites. Since U2AF dissociates from the spliceosomal complex after stepI [43,44], other factors may influence splice site choice at a later step. Two different modes of 3' splice site selection after splicing stepI have been suggested for AG-AG splice site tandems. First, a second 3' AG may be chosen as the site of exon ligation during splicing stepII if it is located a few nucleotides downstream of the first-step AG, defined by U2AF binding [45]. This rather unspecific mechanism is the likely explanation for the high propensity of small-distance AG-AG tandems to result in alternative splicing, and may also be relevant for TG-AG acceptor tandems, which are found overrepresented at a 3-nt distance compared to larger distances (Figure S1 in Additional data file 1). Another mechanism is exemplified by intron 2 of the *Drosophila sxl* gene [46] as well as intron 1 of the β-globin mutant β^{110} [47,48]. Here, the downstream AG is essential for splicing while the dispensable upstream AG may be chosen in splicing stepII, even as the preferred splice site. The splicing factor SPF45 was shown to bind to the upstream AG dinucleotide during splicing stepII, promoting splice site choice [46]. It remains to be tested if SPF45 or other factors contribute to TG splice site choice.

Given the extremely low ratio of viable versus non-viable TG-AG tandems at intron-exon boundaries, contextual sequence signals must contribute to TG splice site definition and influence splice site choice. In agreement, half of the TG splice acceptors are associated with outstandingly high intron sequence conservation. Notably, the alternative TG splice acceptor of *GNAS* intron 3 has been shown to be flanked by three putative exonic splice enhancer motifs (specific for SF2/ASF, SC35, and SRp40), and TG splice site choice has been experimentally shown to be modulated by the ratios of SF2/ASF and hnRNPA1 [16]. We could not identify specific sequence motifs associated with TG splice sites (results not shown). Due to the relatively small sample size for TG 3' splice sites, available methods for motif discovery have limited detection power, especially if *cis*-regulatory elements are highly dispersed, or if diverse elements cooperate in a contextual manner. Presumably, each individual TG-AG tandem recruits a characteristic ensemble of splice regulators to facilitate unusual splice site choice. Thus, the compilation of TG splice sites could serve as a rich source of splicing-relevant contextual sequence signals to be examined in future experimental studies.

Materials and methods

Screening for non-canonical 3' splice sites

From the UCSC Genome Browser site [22] we obtained spliced alignments of human ESTs (file all_est.txt, released 2005-07-14) and of human RefSeq transcripts (refGene.txt, 2005-07-23) [49], as well as a compilation of all human EST sequences (est.fa, 2005-11-26). First, we sampled EST-supported 3' splice sites to identify 3-nt splice variant pairs (Δ3SVPs). In parallel, we identified ESTs that were mapped to multiple genome locations, indicative of paralogous gene loci including pseudogenes. We discarded those Δ3SVPs whose EST support for the minor splice variant did not exceed the number of these ambiguously mapped ESTs. Furthermore, we retained only those Δ3SVPs that have at least one splice site corresponding to a RefSeq transcript, according to the RefSeq-to-genome alignment. Then, we separated cases that involve the dinucleotide AG at both 3' splice sites, that is, NAGNAG tandem splice acceptors, as well as U12-dependent introns, identified by their characteristic donor site and branch-point motifs [6]. The remaining Δ3SVPs were considered 'unusual' since they comprise at least one non-AG splice acceptor in a U2-spliceosomal intron. The splice variants of these Δ3SVPs were validated and quantified by a WU-BLASTN search of 60-nt sequence windows around the resulting exon-exon junctions against all human ESTs, using parameters W = 13, N = -8, nogap S = 180, hspmax = 1. BLAST matches were considered valid if perfect sequence identity was found in a 12-nt window around the exon-exon junctions [37]. Finally, Δ3SVPs were considered highly reliable if the minor 3' splice site was found in at least two ESTs and was used in at least 3% of the covering ESTs. A screen for splice variant pairs for distances of 4-36 nt was performed analogously, restricting the search to tandems of TG-AG splice sites.

PCR and RT-PCR

For validation of splice variants, nested PCR was performed using 1 ng cDNA templates from the Human Multiple Tissue cDNA PanelsI and II (Clontech, Mountain View, CA, USA). For a given gene, suitable tissues were determined from expression data obtained from the Stanford SOURCE database [50]. However, pooled leukocyte cDNA from 200 individuals was preferably chosen in order to obtain comparable results. Verification of the genomic sequence and an analysis of potential polymorphisms were done by nested PCR using 200 ng of pooled genomic DNA from 100 Caucasian

individuals (Roche, Mannheim, Germany) as template. Primers were obtained from Metabion (Martinsried, Germany) (Additional data file 1). Reactions were set up with PuReTaq Ready-To-Go PCR beads (GE Healthcare, Munich, Germany) and 10 pmol primer in 25 μl total volume, according to the manufacturer's instructions. A typical thermocycle protocol was 3 minutes initial denaturation at 94°C, followed by 25 cycles of 1 minute denaturation at 94°C, 1 minute annealing at 53-55°C, 1 minute extension at 72°C, and a final 10 minute extension step at 72°C. In the second round of nested PCR, 1 μl of the first-round product was amplified for 30 cycles. For cloning, PCR products were separated on agarose, DNA was extracted applying the Millipore (Billerica, MA, USA) Montage Gel Extraction kit, followed by ethanol precipitation. Isolated fragments were cloned in pCR2.1-TOPO (Invitrogen, Karlsruhe, Germany), and cloned DNA was Sanger sequenced using M13 standard reverse primer (17-mer).

Splice variant quantification by pyrosequencing

Templates for pyrosequencing were generated using universal biotinylated primers [51]. RT-PCR amplicons of the exon-exon junctions were ligated into pCR2.1-TOPO (Invitrogen) according to the supplier's recommendations and subsequently re-amplified with all four possible combinations of 5'-biotinylated M13 standard primers (17-mers) and unlabeled insert-specific primers (Additional data file 1). The latter also served to prime the pyrosequencing reaction. Biotin-labeling of DNA, single strand preparation and sequencing were performed as described [51].

Orthologous intron-exon boundaries

A data set of orthologous intron-exon boundaries was constructed automatically to obtain sufficient data (especially reference data) to test for evolutionary constraints on intron flanking sequence. Sets of human (data as described for the splice site screen) and mouse transcript annotations (UCSC genome assembly mm7, RefSeq-to-genome alignment 2006-05-21) were processed as described earlier (supplementary methods in [20]). For 97,107 unambiguous orthologous pairs (57% of unique human intron-exon boundaries, including 23 of 36 TG-AG splice site tandems), 100 nt flanking intron sequences were aligned using CLUSTALW [52], using the optimized parameter -gapopen = 2.2. The degree of conservation was determined for 50 nt of the human intron sequence upstream of the splice site (tandem), giving a score of 1 for an identical aligned nucleotide in mouse, a score of 0 for a mismatch, and a penalty of -1 for inserted mouse sequence. Since a histogram of sequence conservation in canonical introns showed a non-normal distribution, statistical testing was performed using a permutation test. Intron samples of given size were simulated by random drawings from the intron data set, and the average sequence identity was calculated, repeating the sampling procedure 10,000 times.

Where automated processing failed (13 of 36 TG-AG splice site tandems), orthologous intron-exon boundaries were retrieved using the UCSC genome browser. These cases were not used for the statistical analysis since these represent a likely biased subset with regard to sequence conservation, and an appropriate large data set for comparison is not available.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a Word file containing a description of the estimation of the amount of *cis*-regulatory sequence context, three supplementary tables and two supplementary figures. Supplementary Table 1 lists the putative unusual splice sites evident from EST-to-genome alignments that failed the quality checks. Supplementary Table 2 provides data about the comprehensive analysis of putative 3' TG splice sites suggested by spliced alignments of RefSeq transcripts. Supplementary Table 3 contains all primer sequences. Supplementary Figure S1 shows the distance-dependent occurrence of TG-AG and AG-AG splice acceptor tandems. Supplementary Figure S2 shows a LOGO representation of the TG 3' splice site sequence context.

Acknowledgements

We thank M-L Schmidt and I Görlich for expert technical assistance, F Liu and Z-G Han for providing clone material, members of the RefSeq Division staff of the National Center for Biotechnology Information for helpful discussions, many members of the FLI and two anonymous referees for critical reading of the manuscript and helpful suggestions. This work was supported by grants from the German Ministry of Education and Research to SS (01GS0426) and MP (01GR0504, 0313652D) as well as from the Deutsche Forschungsgemeinschaft (SFB604-02) to MP.

References

1. Burge CB, Tuschl TH, Sharp PA: **Splicing precursors to mRNAs by the spliceosomes.** In *The RNA World* 2nd edition. Edited by: Gesteland RF, Cech T, Atkins JF. Plainview, NY: Cold Spring Harbor Laboratory Press; 1999:525-560.
2. Konarska MM, Query CC: **Insights into the mechanisms of splicing: more lessons from the ribosome.** *Genes Dev* 2005, **19:**2255-2260.
3. Reed R: **Mechanisms of fidelity in pre-mRNA splicing.** *Curr Opin Cell Biol* 2000, **12:**340-345.
4. Moore MJ: **Intron recognition comes of AGe.** *Nat Struct Biol* 2000, **7:**14-16.
5. Jackson IJ: **A reappraisal of non-consensus mRNA splice sites.** *Nucleic Acids Res* 1991, **19:**3795-3798.
6. Burge CB, Padgett RA, Sharp PA: **Evolutionary fates and origins of U12-type introns.** *Mol Cell* 1998, **2:**773-785.
7. Levine A, Durbin R: **A computational scan for U12-dependent introns in the human genome sequence.** *Nucleic Acids Res* 2001, **29:**4006-4013.
8. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R: **Comprehensive splice-site analysis using comparative genomics.** *Nucleic Acids Res* 2006, **34:**3955-3967.
9. Burset M, Seledtsov IA, Solovyev VV: **Analysis of canonical and non-canonical splice sites in mammalian genomes.** *Nucleic Acids Res* 2000, **28:**4364-4375.
10. Brackenridge S, Wilkie AOM, Screaton GR: **Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes.** *EMBO J* 2003, **22:**1620-1631.
11. Wu Q, Krainer AR: **Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs.** *RNA* 1997, **3:**586-601.
12. Dietrich RC, Incorvaia R, Padgett RA: **Terminal intron**

dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell* 1997, **1:**151-160.
13. Chong A, Zhang G, Bajic VB: **Information for the Coordinates of Exons (ICE): a human splice sites database.** *Genomics* 2004, **84:**762-766.
14. van Nimwegen E, Paul N, Sheridan R, Zavolan M: **SPA: a probabilistic algorithm for spliced alignment.** *PLoS Genet* 2006, **2:**e24.
15. Kozasa T, Itoh H, Tsukamoto T, Kaziro Y: **Isolation and characterization of the human Gs alpha gene.** *Proc Natl Acad Sci USA* 1988, **85:**2081-2085.
16. Pollard AJ, Krainer AR, Robson SC, Europe-Finner GN: **Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice site.** *J Biol Chem* 2002, **277:**15241-15251.
17. Quan F, Forte MA: **Two forms of *Drosophila melanogaster* Gs alpha are produced by alternate splicing involving an unusual splice site.** *Mol Cell Biol* 1990, **10:**910-917.
18. Stathakis DG, Udar N, Sandgren O, Andreasson S, Bryant PJ, Small K, Forsman-Semb K: **Genomic organization of human DLG4, the gene encoding postsynaptic density 95.** *J Neurochem* 1999, **73:**2250-2265.
19. Seeman P, Nam D, Ulpian C, Liu IS, Tallerico T: **New dopamine receptor, D2(Longer), with unique TG splice site, in human brain.** *Brain Res Mol Brain Res* 2000, **76:**132-141.
20. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet* 2004, **36:**1255-1257.
21. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing.** *Am J Hum Genet* 2006, **78:**291-302.
22. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, *et al.*: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34:**D590-D598.
23. Sugnet CW, Kent WJ, Ares MJ, Haussler D: **Transcriptome and genome conservation of alternative splicing events in humans and mice.** *Pac Symp Biocomput* 2004:66-77.
24. Soong TW, DeMaria CD, Alvania RS, Zweifel LS, Liang MC, Mittman S, Agnew WS, Yue DT: **Systematic identification of splice variants in human P/Q-type channel alpha1(2.1) subunits: implications for current density and Ca2+dependent inactivation.** *J Neurosci* 2002, **22:**10142-10152.
25. Neve B, Froguel P, Corset L, Vaillant E, Vatin V, Boutin P: **Rapid SNP allele frequency determination in genomic DNA pools by pyrosequencing.** *Biotechniques* 2002, **32:**1138-1142.
26. Frey UH, Nuckel H, Dobrev D, Manthey I, Sandalcioglu IE, Eisenhardt A, Worm K, Hauner H, Siffert W: **Quantification of G protein Gaalphas subunit splice variants in different human tissues and cells using pyrosequencing.** *Gene Expr* 2005, **12:**69-81.
27. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs.** *Genome Res* 2002, **12:**1837-1845.
28. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB: **Identification and analysis of alternative splicing events conserved in human and mouse.** *Proc Natl Acad Sci USA* 2005, **102:**2850-2855.
29. Sorek R, Dror G, Shamir R: **Assessing the number of ancestral alternatively spliced exons in the human genome.** *BMC Genomics* 2006, **7:**273.
30. Kanno H, Fujii H, Wei DC, Chan LC, Hirono A, Tsukimoto I, Miwa S: **Frame shift mutation, exon skipping, and a two-codon deletion caused by splice site mutations account for pyruvate kinase deficiency.** *Blood* 1997, **89:**4213-4218.
31. Thongnoppakhun A, Rungroj N, Wilairat P, Vareesangthip K, Sirinavin C, Yenchitsomanus PT: **A novel splice-acceptor site mutation (IVS13-2A>T) of polycystic kidney disease 1 (PKD1) gene resulting in an RNA processing defect with a 74-nucleotide deletion in exon 14 of the mRNA transcript.** *Hum Mutat* 2000, **15:**115.
32. Yamada H, Shinmura K, Tsuneyoshi T, Sugimura H: **Effect of splice-site polymorphisms of the TMPRSS4, NPHP4 and ORCTL4 genes on their mRNA expression.** *J Genet* 2005, **84:**131-136.
33. Kol G, Lev-Maor G, Ast G: **Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation.** *Hum Mol Genet* 2005, **14:**1559-1568.
34. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.** *J Comput Biol* 2004, **11:**377-394.
35. Bourinet E, Soong TW, Sutton K, Slaymaker S, Mathews E, Monteil A, Zamponi GW, Nargeot J, Snutch TP: **Splicing of alpha 1A subunit gene generates phenotypic variants of P- and Q-type calcium channels.** *Nat Neurosci* 1999, **2:**407-415.
36. Furey TS, Diekhans M, Lu Y, Graves TA, Oddy L, Randall-Maher J, Hillier LW, Wilson RK, Haussler D: **Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing.** *Genome Res* 2004, **14:**2034-2040.
37. Thanaraj TA: **A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures.** *Nucleic Acids Res* 1999, **27:**2627-2637.
38. Citri Y, Colot HV, Jacquier AC, Yu Q, Hall JC, Baltimore D, Rosbash M: **A family of unusually spliced biologically active transcripts encoded by a *Drosophila* clock gene.** *Nature* 1987, **326:**42-47.
39. Patel AA, McCarthy M, Steitz JA: **The splicing of U12-type introns can be a rate-limiting step in gene expression.** *EMBO J* 2002, **21:**3804-3815.
40. Kim SS, Pandey KK, Choi HS, Kim SY, Law PY, Wei LN, Loh HH: **Poly(C) binding protein family is a transcription factor in mu-opioid receptor gene expression.** *Mol Pharmacol* 2005, **68:**729-736.
41. Michelotti EF, Tomonaga T, Krutzsch H, Levens D: **Cellular nucleic acid binding protein regulates the CT element of the human c-myc protooncogene.** *J Biol Chem* 1995, **270:**9494-9499.
42. Soares LMM, Zanier K, Mackereth C, Sattler M, Valcarcel J: **Intron removal requires proofreading of U2AF/3' splice site recognition by DEK.** *Science* 2006, **312:**1961-1965.
43. Bennett M, Michaud S, Kingston J, Reed R: **Protein components specifically associated with prespliceosome and spliceosome complexes.** *Genes Dev* 1992, **6:**1986-2000.
44. Chiara MD, Palandjian L, Feld Kramer R, Reed R: **Evidence that U5 snRNP recognizes the 3' splice site for catalytic step II in mammals.** *EMBO J* 1997, **16:**4746-4759.
45. Chua K, Reed R: **An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing.** *Mol Cell Biol* 2001, **21:**1509-1514.
46. Lallena MJ, Chalmers KJ, Llamazares S, Lamond AI, Valcarcel J: **Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45.** *Cell* 2002, **109:**285-296.
47. Krainer AR, Reed R, Maniatis T: **Mechanisms of human-globin-pre-mRNA splicing.** In *Genetic Chemistry: the Molecular Basis of Heredity* Edited by: Berg P. Houston, TX: The Robert A Welch Foundation; 1985:353-382. [Conferences on Chemical Research, vol.XXIX]
48. Zhuang Y, Weiner AM: **The conserved dinucleotide AG of the 3' splice site may be recognized twice during *in vitro* splicing of mammalian mRNA precursors.** *Gene* 1990, **90:**263-269.
49. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-D504.
50. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
51. Groth M, Huse K, Reichwald K, Taudien S, Hampe J, Rosenstiel P, Birkenmeier G, Schreiber S, Platzer M: **Method for preparing single-stranded DNA templates for pyrosequencing using vector ligation and universal biotinylated primers.** *Anal Biochem* 2006, **356:**194-201.
52. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31:**3497-3500.

STEFANIE SCHINDLER
MONIKA HEINER
MATTHIAS PLATZER
KAROL SZAFRANSKI

Comparison of methods for splice variant quantification

Biotechniques. 2009 (zur Begutachtung eingereicht).

Quantitative Information über die Verhältnisse von Spleißvarianten ist für die Charakterisierung von Spleißereignissen und für die Ermittlung funktioneller Rollen entscheidend. Die Methoden der Pyrosequenzierung und der Fluoreszenz-basierten Kapillarelektrophorese wurden für diesen Zweck adaptiert, da sie in der Lage sind, Spleißvarianten mit kleinen Längenunterschieden zu analysieren. Beide Methoden wurden hinsichtlich Reproduzierbarkeit und Genauigkeit der gemessenen Daten, sowie Experimentaufbau, Datenanalyse und Anwendungsspektrum systematisch analysiert und mit der verbreiteten Methode der Polyacrylamid-Gelelektrophorese mit Ethidiumbromid-basierter Densitometrie verglichen. Außerdem wurde der Einfluss von niedrigen Template-Konzentrationen und von PCR-Bedingungen analysiert, die zu verzerrten Ergebnissen führen können.

Comparison of methods for quantification of subtle splice variants

Stefanie Schindler[1*], Monika Heiner[2], Matthias Platzer[1], and Karol Szafranski[1]

[1] Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

[2] Institute of Biochemistry, Justus-Liebig-University of Giessen Heinrich-Buff-Ring 58, 35392 Giessen, Germany

* Corresponding author.

Tel: +49 3641 656254. Fax: +49 3641 656255.

Email address: sschindl@fli-leibniz.de (S. Schindler).

Keywords

alternative splicing; splice variant quantification; RT-PCR; pyrosequencing; capillary electrophoresis; polyacrylamide gelelectrophoresis

Abstract

Alternative splicing is capable of generating multiple mRNA variants from a single gene and is hence a key mediator of molecular diversity generated at the transcript level. Consequently, delivering quantitative information on the fractions of splice variants is essential for the understanding of their functional and biological roles. Here we compare techniques for subtle splice variant quantification that are able to resolve length differences as small as one nucleotide: pyrosequencing (PSQ), capillary electrophoresis with laser-induced fluorescence (CE-LIF), and polyacrylamide gelelectrophoresis with ethidium-bromide densitometry (PAGE-ED). We give comprehensive descriptions of assay designs and calibration procedures and present an evaluation of these methods in terms of accuracy, reproducibility and applicability. We also examined the impact of template concentrations and RT-PCR conditions potentially leading to biased results which were observed for PCR amplicons with larger size differences. As proof of concept, we determine the splice ratios of variants differing by 3 and 12 nt in several human tissues. We demonstrate that CE-LIF is the most precise and also the most labor- and time-efficient method.

Introduction

Alternative splicing (AS) is widespread among eukaryotes and generates multiple mRNAs from many genes, contributing an essential part of transcriptome complexity (1-3). The major consequences of AS are changes in structure and function of encoded proteins as well as changes in regulation of gene expression (1,2,4). Thus, the biological impact of AS is dramatically high, as for instance alterations of the splicing patterns by *cis*- or *trans*-acting factors can lead to profound consequences. It is meanwhile broadly proven that AS determines cell fate in numerous contexts, such as sexual differentiation in *Drosophila*, apoptosis in mammals, and aberrant regulation of AS has been implicated in human

diseases (5-7). To understand the impact of AS on the complexity of eukaryotic transcriptomes, its *cis*- and *trans*-mediated regulation, and the functional relevance of particular splice variants relies on quantitative data. Hence, accurate and robust techniques for quantification of transcript variants are important for monitoring changes in AS patterns and ratios.

AS at 5' or 3' splice sites located just a few nucleotides apart is a widespread type of AS in many species. It results in subtle changes in the transcripts and often in the encoded proteins. Several of these splice events contribute to the repertoire of functionally different proteins (8-12). Consequently, a technique for quantification of such splice variants has to meet the demand to resolve even small length differences as small as 3 nt. Numerous methods are commonly used for the determination of the relative amounts of splice variants. Northern Blotting, as the classical method for quantitative analysis of RNA (13,14), relies on denaturing gel electrophoresis and hybridization with a labeled probe. The method is time-consuming, requires relatively large amounts of RNA samples (about 20 µg), and different isoforms are resolved only if their size difference is significant. The ribonuclease protection assay, an extension of the Northern Blot technique, is able to transform subtle sequence differences into differential fragment patterns (14-16). Like Northern Blotting, it requires relatively large amounts of RNA samples, and hardly detects rare transcripts. Real-time quantitative PCR can be applied for quantification of splice variants (17,18), but it is unspecific for the differentiation of small-difference isoforms (19). Transcriptome-wide monitoring of AS with microarrays provides semi-quantitative information, and the selection of a specialized probe set has to be taken into account (20,21). Furthermore, MALDI-TOF MS involving primer extension has been applied for high-throughput monitoring of AS (22).

Here we compare three high-resolution techniques for quantification of splice variants with the capability to detect length differences as small as one nucleotide: polyacrylamide gelelectrophoresis with ethidium-bromide densitometry (PAGE-ED), pyrosequencing (PSQ),

and capillary electrophoresis with laser-induced fluorescence (CE-LIF). In contrast to PAGE-ED, PSQ and CE-LIF are well-suited for processing large numbers of samples in micro-titer plate formats. As PAGE-ED has been extensively used for separation of PCR amplicons and subsequent semi-quantitative measurements of isoform ratios by analytical gel densitometry (23,24), we used this method as a benchmark reference for the evaluation of the high-throughput PSQ and CE-LIF applications. PCR amplicons are separated by gel electrophoresis and visualized by staining or autoradiography, and the digitized gel image is then used for quantification by densitometry using appropriate software.

PSQ is a bioluminometric, real-time DNA sequencing technique for analyzing short DNA sequences. It has been widely used for mutation screening and genotyping, species typing, and estimation of SNP allele frequencies in DNA pools (25-28). Based on a sequencing-by-synthesis principle, this technique yields both qualitative and quantitative sequence information by enzymatically monitoring the incorporation of nucleotides into a primer-template hybrid (29). The order in which the nucleotides are added in the course of the sequencing assay can individually be defined in a custom dispensation order. In the case of quantitative applications special efforts are needed to suppress background signals due to primer dimers and/or secondary structure formation (30). The technique can be adapted for the quantification of splice variants via appropriate sequencing primers and specific dispensation orders. The splice ratios can be calculated based on the variant-specific nucleotide incorporation. Previously, this technology was used by Frey *et al.* for a quantitative analysis of G protein stimulatory α subunit gene (*GNAS*) splice variants (19).

The CE-LIF is based on PCR amplification with fluorescently labeled primers and PCR product separation on a capillary sequencer. Those sequencers can precisely separate DNA fragments that differ in size by only 1 nt (31,32). In a population of nucleic acid molecules the components can be quantified based on the respective fluorescence peak area in the

electropherogram. This powerful technology has been used to analyze RT-PCR products of individual genes (33,34) and splice variants (35).

The three techniques are evaluated and compared in terms of assay design, accuracy, reproducibility and applicability. Since all techniques are coupled to a PCR amplification, we initially performed tests aiming at optimal PCR conditions under consideration of rare transcripts and demonstrating potential pitfalls of RT-PCR. As test cases for validation we used splice variants from two genes with size differences of 3, 12 and 195 nt. In a series of experiments, using known proportions of cloned splice variants, we examined the three quantification techniques with respect to accuracy. The results demonstrate that differences in accuracy exist, especially around extreme values of splice ratios (0 or 100%). As a proof of concept we quantify RT-PCR products obtained from cDNA from human tissues using the three techniques.

Materials and Methods

Plasmids used as PCR templates

Our experimental test cases are the genes *CNBP* and *ARS2*. The splice variants of *CNBP* originating from insertion or deletion of 3 nt at the 3'-terminus of intron 3 are termed Δ3+ and Δ3-, respectively. AS at the 3'-terminus of *ARS2* intron 18 results in insertion or deletion of 12 nt, termed Δ12+ and Δ12-, respectively. Additional *ARS2* splice variants were used arising from retention of intron 18, termed Δ195+ and Δ195-. Partial cDNA clones of splice variants from *ARS2* (exon 18-20, framed by primer sequences 5'-TAACAACTTCCTCACTGATGC and 5'-CTCGACCAGCACCATAC), and *CNBP* (exon 3-4, framed by the primer sequences 5'-GGTGAGTCTGGTCATCTTG and 5'-GCAGTCCTTGGCAATGTG) were obtained as described previously (11). The clones were verified by Sanger sequencing. For PCR, plasmids were linearized by restriction with *BglII*,

purified by ethanol precipitation, resuspend in TE buffer (10 mM TrisHCl pH 8.0, 1 mM EDTA), and stock solutions of 1 ng/µl in TE buffer were made.

Serial dilutions of a 1:1 mixture of linearized ∆3+ and ∆3- plasmids were prepared, containing a total of 100/250/500/1,000/2,500 molecules in 1 µl TE buffer. Molecule numbers were calculated based on the average molecular weight of a basepair (660g/mol), the size of plasmids, and the stock concentration of 1 ng/µl. In the case of ∆3-, for example, the plasmid size is 3,990 bp, and the molecule count in the 1 µl stock solution is calculated as ~$2.3x10^8$. To obtain 100 template molecules in 1 µl, the stock solution was diluted in four steps to 1:$2.3x10^6$ in TE buffer.

Nested PCR conditions

Reactions were started with 1,000 DNA template molecules, or 0.2 ng poly(A)+ cDNA, BioMix Red (Bioline, Randolph, USA) and 10 pmol primer in 25 µl total volume, according to the manufacturer's instructions. A nested PCR approach was applied. In general, the PCR reaction started with a 2 min initial denaturation at 94°C, followed by cycles of 45 sec denaturation at 94°C, 50 sec annealing at 53°C, 1 min extension at 72°C, and a final 2 min extension step at 72°C. In the first round of PCR, M13 standard primers (20-mer) and 20 cycles were used. In the second PCR, 1 µl of the first-round product was amplified with insert-specific primers for a variable number of cycles.

Statistical analyses

The measurements of splice variant fractions were based on at least three independent experiments, for each sample and method. Statistical measures such as average (avg), standard error of the mean (SEM), and coefficient of determination (R^2) were calculated based on standard formula as implemented in Microsoft Excel.

PAGE-ED analysis

The second round of nested-PCR was carried out with unlabelled reverse primers. The products were separated by electrophoresis using native polyacrylamide gels. The Δ3 amplicons were separated on 12% native gels (12% acrylamide/bis-acrylamide 19:1, 400 µl 10% APS, 40 µl TEMED, in 50 ml 1x TBE) for 6 h at 300 V, and theΔ12 amplicons were separated on 8% native gels (8% acrylamide/bis-acrylamide 29:1, 400 µl 10% APS, 40 µl TEMED, in 50 ml 1x TPE) for 17 h at 60 V.

The separated products were visualized by ethidium-bromide staining. Images were sampled by a CCD camera, and the integrated optical density of detected bands was measured by the GeneTools gel image analysis software (Syngene). The splice ratios were calculated from an average of at least three independent experiments.

PSQ analysis

The 2nd round of nested PCR was carried out with 5'-biotinylated primers together with an unlabeled primer which also served to prime the PSQ reaction: *CNBP* (Δ3+/Δ3-) forward 5'-GATCTTCAGGAGGAT with dispensation order A G C T A G C T G, reverse 5'-CCGCAGTTATAGCA with dispensation order A G C T A C A T; *ARS2* (Δ12+/Δ12-) forward 5'-CACCTGGCCCCGC with dispensation order T C A G A T A C A T C A G, reverse 5'-GTCCTGGGGTCAAAC with dispensation order A C T G A G C T A T C T G. Biotin-labeled PCR products were immobilized on Streptavidin Sepharose (GE Healthcare Lifesciences) by mixing 20 µl of the PCR product with 6 µl bead suspension, 10 µl water, and 40 µl Binding Buffer (Biotage), followed by shaking at room temperature for 30 min. The samples were sequentially washed with 70% ethanol and then 0.5 M NaOH, using the PyroMark Vacuum Prep Tool (Biotage). Immobilized ssDNA was then washed with Washing Buffer for 10 s, transferred to 40 µl Annealing Buffer plus 6 µl pyrosequencing primer (10 pmol/µl in water), and heated at 81.5°C for 10 min to allow for hybridization of the pyrosequencing primer. After equilibration to room temperature, the sequencing reaction was performed using the Pyro

Gold Reagent Kit (Biotage) on a PSQ 96 MA apparatus (Biotage), according to the manufacturer's instructions.

For the analysis step, the following quality-control criteria were used: (i) low relative error between multiple isoform-specific peaks: SEM (isoform-specific values) / avg (reference values) ≤ 0.2, for each of the isoforms; (ii) the sum of isoform-specific signals should be close to the of reference signals: 1 - ∑ avg (isoform-specific values) / avg (reference values) ≤ 0.2; (iii) control nucleotides in the dispensation scheme produce low noise signals: x (control nucleotide value) / avg (reference values) ≤ 0.04, for all control nucleotides.

CE-LIF analysis

The second round of nested-PCR was carried out with 5'-6-carboxyfluorescein (FAM)-labelled reverse primers and a final elongation time of 30 minutes. The FAM-labelled PCR products were appropriately diluted (1/20 to 1/50) and 1µl was supplemented with 10 µl formamide (Roth, Karlsruhe) and 0.5 µl of GeneScan 500 LIZ (Applied Biosystems). The mixture was denatured at 95°C for 5 min, and cooled on ice. The denatured products were then separated on an ABI 3730 capillary sequencer and analyzed with the GeneMapper 4.0 software. The proportion of the long isoform (%) was calculated as follows: peak area for the long isoform/(long isoform+short isoform) x 100.

RT-PCR on human tissue samples

cDNA from the Human Multiple Tissue Panels I and II (Clontech) was amplified by nested PCR with the following gene-specific primers: First round primers *ARS2* (Δ12) 5'-GCAGAGAAAATTGAGGAAGTG and 5'-CCTCGGAAGGCATCATAG, *CNBP* (Δ3) 5'-TCGCTGTGGTGAGTCTG and 5'-GCTCTCGCTCTCTCTTG; second round primers as described in section "Plasmids used as PCR templates". The subsequent analyses via PAGE-ED, PSQ and CE-LIF were done as itemized.

Results

Effect of different PCR conditions

In order to avoid false results due to PCR biases or stochastics as a consequence of low transcript numbers, we tested the robustness of our nested PCR approach with respect to template concentration and number of PCR cycles in the 2nd round.

We first examined the stability of measurements with varying template amounts and particularly checked low numbers of template molecules in the PCR reaction. This is important for the analysis of genes with low expression levels resulting in not abundant transcript isoforms and bears the danger of stochastic effects leading to higher error rates in measurements of splice variant ratios. For that purpose, serial dilutions of a 1:1 mixture of linearized plasmids containing splice variants that differ in 3 nt (Δ3+ and Δ3-) were prepared, with a total of 100, 250, 500, 1,000, and 2,500 molecules. In all cases, an average Δ3+ fraction of ~50% was measured (Supplementary fig. 1), and even with the smallest template amount the standard error was less than 4%. However, the variation clearly increased with decreasing number of template molecules. We chose 1,000 molecules for further analyses since this appeared to allow for measurements with a resolution of 1% or even higher.

In order to assess the impact of human cDNA as carrier, we also checked the usage of *S. cerevisiae* RNA and herring sperm DNA, showing minor effects on the measurements and accuracy (data not shown).

Next, we tested whether the number of PCR cycles in the 2nd round of nested-PCR influences the obtained isoform ratios with respect to PCR biases expected, e.g. towards the shorter amplicon. Defined mixtures of cloned splice variants in the range from 5 to 95% fraction of the longer isoform were made and amplified in the 2nd PCR with 15, 20, 25, 30, 35, and 40 cycles. Overall, over the range of 20 to 30 cycles stable results were obtained (fig.

1A, B) and no PCR bias could be detected for the Δ3 and Δ12 cases. Equivalently, this was done with the cloned Δ195 splice variants as an example for PCR amplicons with extremely varying sizes, e.g. originating from exon skipping/inclusion or intron retention (36). Notably, a drastic PCR bias was observed for long-isoform rates >20%, even with 15 PCR cycles in the 2nd PCR (fig. 1C). Consequently, the Δ195 test case was excluded from further experiments, and for the Δ3 and Δ12 cases 25 cycles in 2nd PCR were chosen.

Assay setups

The setup of the PAGE-ED assay for the test cases was straightforward as the established nested PCR primers could directly be used for amplification. The amplicons were separated by PAGE and densitometrically quantified (fig. 2B).

Similarly, for CE-LIF, merely one of the established primers in the 2nd round of nested PCR had to be synthesized with a FAM-label. The labeled PCR products were then separated on the ABI sequencer and quantified using the GeneMapper software. Representative electropherograms are shown in figure 4. Notably, the addition of an adenosine overhang, a typical characteristic of the *Taq* DNA Polymerase, was observed. The separated fragments were shifted in size by one nucleotide (fig. 3A; Δ3: 87 nt and 90 nt; Δ12: 239 nt and 251 nt). As a control, we used *Pfu* DNA Polymerase and obtained the fragments with their expected sizes (fig. 3B; Δ3: 86 nt and 89 nt; Δ12: 238 nt and 250 nt). Typically, we observed “stutter” products with both polymerases. However, the quantification of splice ratios was not affected by inclusion of the “stutter”-peak data into the calculation (data not shown).

For PSQ, appropriate sequencing primers had to be established, first. The primers were set in close proximity to the region of interest, as the signal intensity decreases with sequencing length and thus the experimental accuracy drops with distance. Second, the order in which the nucleotides are dispensed into the reaction should be defined for both sequencing directions, with consideration of the following aspects:

- Specific nucleotide incorporations have to be defined, which are incorporated in either the long or the short splice variant.
- Reference nucleotide incorporations have to be defined, which are homogeneously incorporated in both splice variants functioning as reference values.
- Control nucleotides have to be defined, which cannot be incorporated since they are not part of the actual sequence.
- Homopolymeric runs are undesirable but often not avoidable.

In our test cases, we explored the regions of inserted or deleted sequence with respect to these design rules. Figure 4 shows dispensation orders with resulting programs.

The splice ratios were determined by analyzing the variant-specific peaks in relation to the reference peaks. For calculation, we considered the areas of all peaks to minimize inconsistencies between signals in individual pyrograms. Ergo, the values from variant-specific incorporations and from reference incorporations were averaged and used for the following final calculation: splice variant proportion = avg [isoform-specific values] / avg [reference values].

We furthermore established quality-control criteria: (i) low relative error between multiple isoform-specific peaks; (ii) the sum of isoform-specific signals should be close to the total signal; (iii) control nucleotides in the dispensation scheme produce low noise signals. Any individual experiment that fails the thresholds for these criteria was repeated.

Validation of PAGE-ED, PSQ and CE-LIF for splice variant quantification

To test the accuracy and reproducibility of data produced with each technique, we performed validation experiments with defined mixtures of cloned splice variants. We amplified with the optimized PCR setup template mixes containing 100%, 98%, 95%, 90%, 80%, 50%, 20%, 10%, 5%, 2%, 0% of the Δ13+ and Δ12+ isoforms, respectively. Across the three

methods, the same products of the first round of nested-PCR were used in order to minimize sample-to-sample variation.

With PAGE-ED we obtained satisfying measurements if the proportion of the long isoform is between 10% and 90% (fig. 2A). Extremely low and high portions of the long isoforms could not be accurately determined. Particularly, the amount of a rare isoform is underestimated. Leaving out the background-correction resulted in the opposite effect, i.e. the amount of a rare isoform was overerestimated, and overall, the level of accuracy did not improve (data not shown). The experimental accuracy is fluctuating, especially in the case of Δ12 measurements for the Δ12+ proportions of 20%, 80%, and 90%. Overall, the reproducibility (±4.6% and ±11.0% maximum SEM for Δ3 and Δ12, respectively) and the experimental accuracy are relatively low (Δ3: R^2=0.9917, Δ12: R^2=0.9837).

In case of PSQ, the provided mixtures of Δ3 splice variants could be accurately quantified (fig. 5A, B). Dispensation orders for both sequencing directions were suitable to determine accurate splice variant proportions (forward: R^2=0.9907, reverse R^2=0.9900; maximum SEM ±6.8%). In contrast, for the Δ12 case only the reverse dispensation order provided accurate data (R^2=0.9822, fig. 5D), whereas the forward direction did not allow for accurate measurements (fig. 5C), which is indicated by the coefficient of determination R^2=0.8832. Consequently, data obtained by the forward sequencing were left out from the overall evaluation and further analysis (section “Quantification of RT-PCR products from various human tissues”). Moreover, we noticed detection limits for test samples with a long isoform portion <5% and >95%. For example, in the case of Δ3, the 0% and 2% samples were measured 5.8%±1.7% and 5.3%±0.6%, respectively (fig. 5A). Analogously, the Δ12+ proportions of 98% and 100% were measured as 104.1%±0.4% and 108.6%±1.3% (fig. 5D). Overall, the reproducibility in both tested cases (leaving Δ12 forward out), is comparable to the PAGE-ED measurements (Δ3 maxiumun SEM ±6.8%, Δ12 maximum SEM ±3.8%).

Splice variant ratios measured with CE-LIF showed a remarkable high accuracy and reproducibility (fig. 3C; ∆3: R^2=0.9948, maximum SEM ±3.3%; ∆12: R^2=0.9963, maximum SEM ±2.8%). In both test cases, the defined mixtures of splice variants were exactly quantified. Remarkably, good results were also achieved for the extremely low and high fractions of the long isoforms.

Quantification of subtle splice ratios from various human tissues

Finally, as a proof of principle we quantified splice ratios of *CNBP* (Δ3) and *ARS2* (Δ12) in various human tissues (brain, liver, placenta, heart, leukocytes) with each of the three techniques (table 1). Overall, the *ARS2* measurements fluctuate more than those of *CNBP*, as discussed later. We obtained systematic deviations in the determined fractions between PAGE-ED and the other methods, i.e. PAGE-ED yielded extremely high (*ARS2*) or extremely low measurements (*CNBP*). Moreover, the standard deviations of the PAGE-ED results were the highest, followed by those of PSQ. CE-LIF showed a remarkably low experimental fluctuation, consistent with the data from the validation experiments.

Discussion

Since research on AS as a major source of transcriptome and proteome complexity gains in importance from basic research to clinical topics (37), reliable methods and standards for quantification are becoming a high priority. In this study we evaluate and compare the three techniques PAGE-ED, PSQ, and CE-LIF for quantification of splice variants with small length differences. Our aim was to give a methodological overview considering assay design and workflow, accuracy, and reproducibility of either technique.

Important aspects to consider primarily for quantification analyses are RT-PCR conditions and/or a low number of template molecules potentially leading to noisy measurements of splice variant proportions. For the latter, we aimed to simulate the situation of genes with low

transcription levels leading to rare transcript isoforms. This bears the danger of stochastic effects leading to higher error rates in isoform ratio measurements. Nevertheless, we could demonstrate the applicability of nested RT-PCR for splice variant quantification. This remarkably extends the range of quantitative analyses, as the nested approach provides significantly higher sensitivity in comparison to a conventional PCR and reduces biases due to a decrease in polymerase capacity over incubation time during large cycle numbers (38). Our measurements were very reproducible with 1,000 template molecules and the standard deviation in case of 500 template molecules was still acceptable (SEM=2.1, Supplementary fig. 1), suggesting that even rare transcript ratios can be reliably determined. Consistent with this, we observed an overall higher experimental variance of *ARS2* compared to *CNBP* in the analyses of RT-PCR products obtained from human cDNA, inversely correlated with the genes' expression levels (Supplementary fig. 2). The tests of different PCR cycle numbers revealed accurate measurements for the Δ3 and Δ12 cases, whereas a drastic bias towards short PCR products was seen for Δ195+ fractions >20%. Obviously, the size differences of 3 and 12 nt are too small for an amplicon size-dependent PCR bias. The causes of such biases can be nucleotide or primer depletion, too short elongation times, stability and capacity of polymerase. This shows that it is essential to carefully evaluate the PCR setup for quantitative analyses of splice variants with larger size differences (larger than the 12-bp difference which was found to be unproblematic). Since the analysis of PCR-biases is not in the scope of our study, we did not consider the Δ195 case in the validation experiments.

PAGE-ED is one of the mainstream techniques used for quantitative assessments of splice variants. PCR amplicons are separated, visualized and subsequently densitometrically quantified. This approach does not require any gene-specific setup regarding the assay, and the preparation including optimization of gel concentration, running time and voltage is negligible. The main advantage of this approach is that one can roughly characterize the proportion or the existence of splice variants in a short time. However, this method is not

applicable for automated medium-scale throughput and does not yield very accurate measurements.

The PSQ assay design is much more difficult and bears labor and cost intensive challenges due to various parameters like amplicon size, sequencing primer design, sample preparation and nucleotide dispensation order. However once the assay design is optimized, this technique is routinely applicable (table 2). Nevertheless, data sorting, quality curation and the final calculation of splice ratios are time-consuming with the standard PSQ software. Gharizadeh *et al.* recently addressed challenges including varying PCR parameters, sequencing primer design, sample preparation and nucleotide dispensation (39). We strongly recommend before routine application, to approve a newly designed assay by validation experiments. In some cases, sequencing directions have different efficiencies and qualities which cannot be circumvented by redesign of primers and/or the dispensation order (results not shown). This was also found for other targets, which are not presented herein. In such cases, values obtained from one sequencing direction may turn out to be sufficiently reliable for ratio calculation. Ideally, the variant-specific peak heights directly represent their relative proportions, but several factors like homopolymeric runs, background signal and signals produced by dATPαS affect peak heights (25,40). However, we did not observe changes in experimental accuracy and reproducibility by excluding the peak heights obtained from homopolymeric runs and exclusion of A-values (results not shown).

Beyond its high accuracy and reproducibility, the CE-LIF approach bears several advantages. It does not require a laborious setup in contrast to PSQ were the assay design is demanding and time-consuming (table 2). If the PCR is once optimized, one of the primers may be easily substituted by a FAM-labelled oligonucleotide and the respective products can be directly analyzed using a capillary sequencer. Moreover, software support for data acquisition and the calculation of splice ratios is convenient. As mentioned in the results paragraph, we observed in some cases stutter peaks. This is supposedly due to polymerase

slippage at homopolymers or more likely due to contaminations of primer oligonucleotides with shorter byproducts. We observed more stuttering by usage of *Pfu* DNA Polymerase, arguing for the first option (fig. 3). Moreover, usage of *Taq* DNA Polymerase causes n+1 peaks due to the 3' adenosine overhang characteristic for this enzyme. Because of those additional peaks it can be difficult to accurately quantify splice variants differing in 1-2 nt. To minimize the interference by the terminal overhang and to prevent doubled peaks in the electropherogram, we chose in the 2nd PCR round a final elongation time of 30 minutes to allow quantitative addition of the unspecific dATP.

In our validation experiments, the commonly used PAGE-ED method was the most imprecise, as accurate results were obtained only for minor isoform frequencies ≥10%. However the reproducibility is relatively low compared to the other techniques (table 2). PSQ provided accurate quantification only ≥5%. In contrast, we demonstrated that CE-LIF is the most accurate and precise method, which quantified all preset splice ratios with high accuracy and reproducibility. An overall low standard deviation was observed even at <5% fractions of the minor splice variant. In all, the PAGE-ED technique gives a rough estimation of isoform ratios but if a more precise analysis is required, techniques like PSQ and CE-LIF are more appropriate. Consistent to that are the results of the quantification of splice ratios in several human tissues which reproduced the characteristics of the methods obtained with the test cases.

Concerning the spectrum of applicability, CE-LIF and PAGE-ED are able to identify yet unknown isoforms. This is different with PSQ where nucleotide inclusions beside the expected variant-specific incorporations lead to uninterpretable data. Furthermore, CE-LIF and PAGE-ED are capable of detecting and quantifying AS events leading to large size differences, such as exon skipping or intron retention. In theory, PSQ is suited for differential quantification of such variants, but has inherent problems with amplicons that exceed a size of ~300 bp as they bear multiple potential sites for mispriming. Moreover, CE-LIF and PAGE-

ED can quantify components in very complex isoform mixtures. To a lesser extent, complex mixtures can be resolved by PSQ also (19). On the other hand, solely PSQ is able to detect sequence variations like single-nucleotide polymorphisms (SNPs), and thus allele-specific AS events. However, the detectability of those coupled events is again distance-limited. Analysis of such coupled events using CE-LIF and PAGE-ED would require allele-specific restriction of the RT-PCR products. Moreover, one major advantage of CE-LIF and PSQ is their applicability in automated medium-scale throughput, which is not possible for PAGE-ED. Importantly, with all three techniques, splice variants with small length differences can be analyzed with a high-resolution range of 1-2 nt (table 2).

Overall, nested PCR allows for reliable quantification of splice ratios even for low-expressed subtle transcript isoforms. PAGE-ED gives a rough quantification but fails if the fraction of the minor isoform is small. If a more precise analysis, a resolution of extremely skewed splice ratios or a higher throughput is required, techniques like PSQ and CE-LIF are highly recommended. CE-LIF is the most precise method for quantification of subtle splice variants, and it is also the most labor- and time-efficient method.

Acknowledgement

This work was supported by grants from the German Ministry of Education and Research (0313652D) and the Deutsche Forschungsgemeinschaft (SFB604-02) to M.P. The authors declare no competing interests.

References

1. **Blencowe, B.J.** 2006. Alternative splicing: new insights from global analyses. Cell *126*:37-47.
2. **Lareau, L.F., R.E. Green, R.S. Bhatnagar and S.E. Brenner.** 2004. The evolving roles of alternative splicing. Curr Opin Struct Biol *14*:273-282.
3. **Stamm, S., S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T.A. Thanaraj and H. Soreq.** 2005. Function of alternative splicing. Gene *344*:1-20.
4. **Black, D.L.** 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*:291-336.
5. **Faustino, N.A. and T.A. Cooper.** 2003. Pre-mRNA splicing and human disease. Genes Dev *17*:419-437.
6. **Charlet, B.N., R.S. Savkur, G. Singh, A.V. Philips, E.A. Grice and T.A. Cooper.** 2002. Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing. Mol Cell *10*:45-53.
7. **Schmucker, D., J.C. Clemens, H. Shu, C.A. Worby, J. Xiao, M. Muda, J.E. Dixon and S.L. Zipursky.** 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*:671-684.
8. **Hiller, M., K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen and M. Platzer.** 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet *36*:1255-1257.
9. **Hiller, M. and M. Platzer.** 2008. Widespread and subtle: alternative splicing at short-distance tandem sites. Trends Genet *24*:246-255.
10. **Dou, Y., K.L. Fox-Walsh, P.F. Baldi and K.J. Hertel.** 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA *12*:2047-2056.
11. **Szafranski, K., S. Schindler, S. Taudien, M. Hiller, K. Huse, N. Jahn, S. Schreiber, R. Backofen, et al.** 2007. Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns. Genome Biol *8*:R154.
12. **Schindler, S., K. Szafranski, M. Hiller, G.S. Ali, S.G. Palusa, R. Backofen, M. Platzer and A.S. Reddy.** 2008. Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes. BMC Genomics *9*:159.
13. **Maderazo, A.B., J.P. Belk, F. He and A. Jacobson.** 2003. Nonsense-containing mRNAs that accumulate in the absence of a functional nonsense-mediated mRNA decay pathway are destabilized rapidly upon its restitution. Mol Cell Biol *23*:842-851.
14. **Sambrook, J. and D. Russel.** 2000. Molecular cloning: A laboratory manual. New York: Cold Spring Harboor Laboratory Press.
15. **Hod, Y.** 1992. A simplified ribonuclease protection assay. Biotechniques *13*:852-854.
16. **Saccomanno, C.F., M. Bordonaro, J.S. Chen and J.L. Nordstrom.** 1992. A faster ribonuclease protection assay. Biotechniques *13*:846-850.
17. **Atkinson, T.P. and Y. Dai.** 2007. Activation-induced changes in alternate splice acceptor site usage. Biochem Biophys Res Commun *358*:590-595.
18. **Vandenbroucke, II, J. Vandesompele, A.D. Paepe and L. Messiaen.** 2001. Quantification of splice variants using real-time PCR. Nucleic Acids Res *29*:E68-68.
19. **Frey, U.H., H. Nuckel, D. Dobrev, I. Manthey, I.E. Sandalcioglu, A. Eisenhardt, K. Worm, H. Hauner, et al.** 2005. Quantification of G protein Gaalphas subunit splice variants in different human tissues and cells using pyrosequencing. Gene Expr *12*:69-81.
20. **Johnson, J.M., J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, et al.** 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*:2141-2144.
21. **Castle, J.C., C. Zhang, J.K. Shah, A.V. Kulkarni, A. Kalsotra, T.A. Cooper and J.M. Johnson.** 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet *40*:1416-1425.
22. **McCullough, R.M., C.R. Cantor and C. Ding.** 2005. High-throughput alternative splicing quantification by primer extension and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Nucleic Acids Res *33*:e99.
23. **Cooper, T.A.** 2005. Use of minigene systems to dissect alternative splicing elements. Methods *37*:331-340.

24. **Tadokoro, K., M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, M. Ozaki, M. Ono, et al.** 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.
25. **Gruber, J.D., P.B. Colligan and J.K. Wolford.** 2002. Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing. Hum Genet *110*:395-401.
26. **Diggle, M.A. and S.C. Clarke.** 2004. Pyrosequencing: sequence typing at the speed of light. Mol Biotechnol *28*:129-137.
27. **Langaee, T. and M. Ronaghi.** 2005. Genetic variation analyses by Pyrosequencing. Mutat Res *573*:96-102.
28. **Ronaghi, M. and E. Elahi.** 2002. Pyrosequencing for microbial typing. J Chromatogr B Analyt Technol Biomed Life Sci *782*:67-72.
29. **Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen and P. Nyren.** 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem *242*:84-89.
30. **Utting, M., J. Hampe, M. Platzer and K. Huse.** 2004. Locking of 3' ends of single-stranded DNA templates for improved Pyrosequencing performance. Biotechniques *37*:66-67, 70-63.
31. **Butler, J.M., E. Buel, F. Crivellente and B.R. McCord.** 2004. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis *25*:1397-1412.
32. **Wenz, H., J.M. Robertson, S. Menchen, F. Oaks, D.M. Demorest, D. Scheibler, B.B. Rosenblum, C. Wike, et al.** 1998. High-precision genotyping by denaturing capillary electrophoresis. Genome Res *8*:69-80.
33. **Richards, M.P., C.M. Ashwell and J.P. McMurtry.** 1999. Analysis of leptin gene expression in chickens using reverse transcription polymerase chain reaction and capillary electrophoresis with laser-induced fluorescence detection. J Chromatogr A *853*:321-335.
34. **van Eekelen, J.A., F.V. Shammas, L. Wee, R. Heikkila and A. Osland.** 2000. Quantitative analysis of cytokeratin 20 gene expression using RT-PCR and capillary electrophoresis with fluorescent DNA detection. Clin Biochem *33*:457-464.
35. **Tsai, K.W. and W.C. Lin.** 2006. Quantitative analysis of wobble splicing indicates that it is not tissue specific. Genomics *88*:855-864.
36. **Cartegni, L., S.L. Chew and A.R. Krainer.** 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet *3*:285-298.
37. **Venables, J.P., R. Klinck, A. Bramard, L. Inkel, G. Dufresne-Martin, C. Koh, J. Gervais-Bird, E. Lapointe, et al.** 2008. Identification of alternative splicing markers for breast cancer. Cancer Res *68*:9525-9531.
38. **Lawyer, F.C., S. Stoffel, R.K. Saiki, S.Y. Chang, P.A. Landre, R.D. Abramson and D.H. Gelfand.** 1993. High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity. PCR Methods Appl *2*:275-287.
39. **Gharizadeh, B., M. Akhras, N. Nourizad, M. Ghaderi, K. Yasuda, P. Nyren and N. Pourmand.** 2006. Methodological improvements of pyrosequencing technology. J Biotechnol *124*:504-511.
40. **Wasson, J., G. Skolnick, L. Love-Gregory and M.A. Permutt.** 2002. Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology. Biotechniques *32*:1144-1146, 1148, 1150 passim.

Tables

Table 1: **Splice variant proportions of RT-PCR products from several human tissues**. The fractions of the long splice variants for the CNBP and ARS2 genes (Δ3+ and Δ12+) are depicted.

	[%] CNBP (Δ3+)			[%] ARS2 (Δ12+)		
tissue	**PAGE-ED**	**PSQ**	**CE-LIF**	**PAGE-ED**	**PSQ**	**CE-LIF**
brain	**12.9** ± 5.8	**25.4** ± 5.0	**18.3** ± 0.6	**72.6** ± 7.0	**63.4** ± 2.7	**58.9** ± 1.9
liver	**18.9** ± 6.3	**32.5** ± 4.6	**25.3** ± 1.8	**86.5** ± 2.6	**68.4** ± 2.2	**72.6** ± 2.3
placenta	**19.8** ± 3.1	**29.3** ± 2.8	**18.3** ± 0.5	**87.9** ± 2.0	**78.6** ± 7.3	**74.1** ± 3.4
heart	**27.8** ± 0.1	**31.6** ± 1.8	**32.8** ± 0.4	**81.6** ± 3.7	**76.5** ± 5.2	**77.5** ± 0.7
leukocytes	**44.1** ± 3.5	**41.5** ± 6.3	**43.1** ± 0.2	**82.3** ± 2.3	**70.0** ± 8.6	**63.8** ± 2.3

Table 2: **Comparison of technical characteristics**

(#) after established PCR (agarose gel)

	PAGE-ED	PSQ	CE-LIF
assay design #	0.5 d	> 5 d	0.5 d
assay	6-8 h	4 h	2 h
sample size	20	96	96
amplicon size	60 - 600 nt	40 - 100 nt	60 - 600 nt
resolution	2 nt	1 nt	2 nt
average accuracy (R^2)	0.9877	0.9876	0.9956
reproducibility (max SEM)	± 11.0 %	± 6.8 %	± 3.3 %

Figures

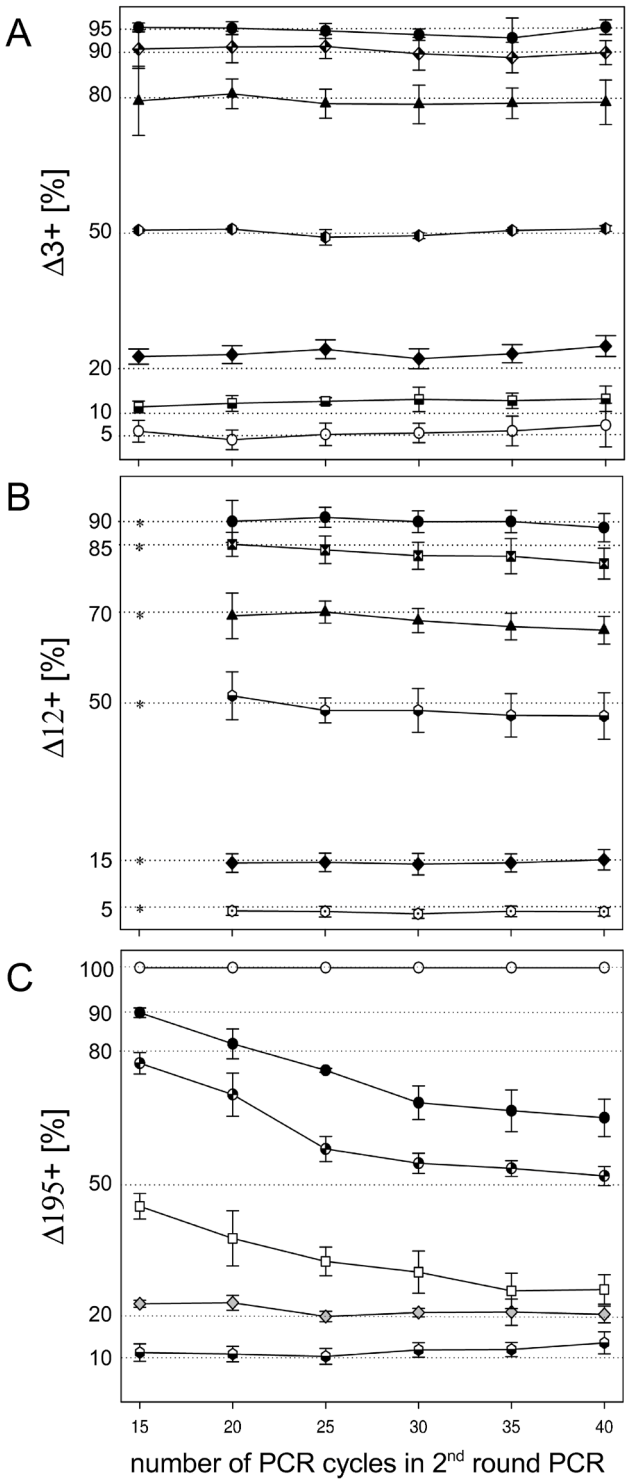


Figure 1:

Effect of different numbers of PCR cycles in the 2nd round of nested PCR. (A) The cloned Δ3 splice variants were mixed with a 5%, 10%, 20%, 50%, 80%, 90%, 95% fraction of the longer isoform (indicated by the dotted lines). (B) For Δ12, the mixed fractions of the longer isoform were 5%, 15%, 50%, 70%, 85%, 90% (indicated by the dotted lines). In case of Δ12, 15 cycles were not sufficient to yield quantifiable products marked by an asterisk. (C) shows the effect of PCR conditions on targets with large differences in amplicon size. The cloned Δ195 splice variants were mixed with a 10%, 20%, 50%, 90%, 100% fraction of the longer isoform (indicated by the dotted lines). PCR fragments were separated and quantified by CE-LIF.

Figure 2:

Validation of PAGE-ED with preset mixtures of Δ3 and Δ12 splice variants. (A) The experimentally measured fractions of the long splice variant are plotted against the expected ones. The diagonal indicates a perfect match between expectations and measurements. (B) The polyacrylamide gels were stained by ethidium-bromide, and the fragments were visualized and quantified by the GeneTools gel image analysis software (Syngene). The illustrated fractions represent the proportion of the Δ3+ and Δ12+ variants.

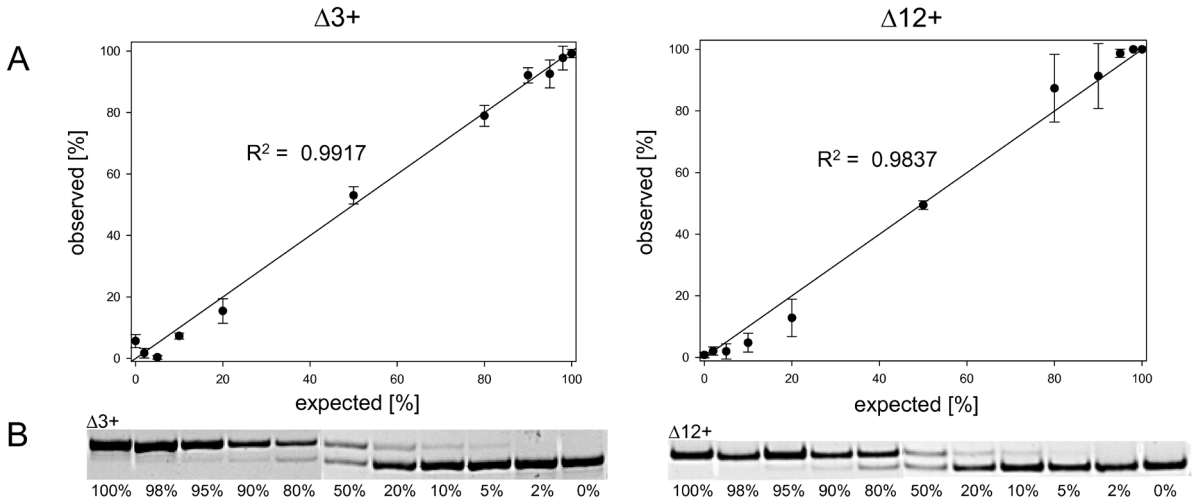


Figure 3:

Separation of Δ3 and Δ12 splice variants by CE-LIF and validation with preset mixtures. (A, B) The separated fragments are visualized by the GeneMapper 4.0 software. Fragment sizes are determined throughout size calibration using an internal standard (GeneScan 500 LIZ, Applied Biosystems). The peaks in the electropherograms are labeled with the corresponding fragment sizes. Δ3 (left) and Δ12 (right) PCR products were generated by *Taq* DNA Polymerase (A) or by *Pfu* DNA Polymerase (B). (C) The experimentally measured fractions of the long splice variant are plotted against the expected ones. The diagonal indicates perfect match between expectations and measurements. The illustrated fractions represent the proportion of the Δ3+ (left) and Δ12+ (right) variants.

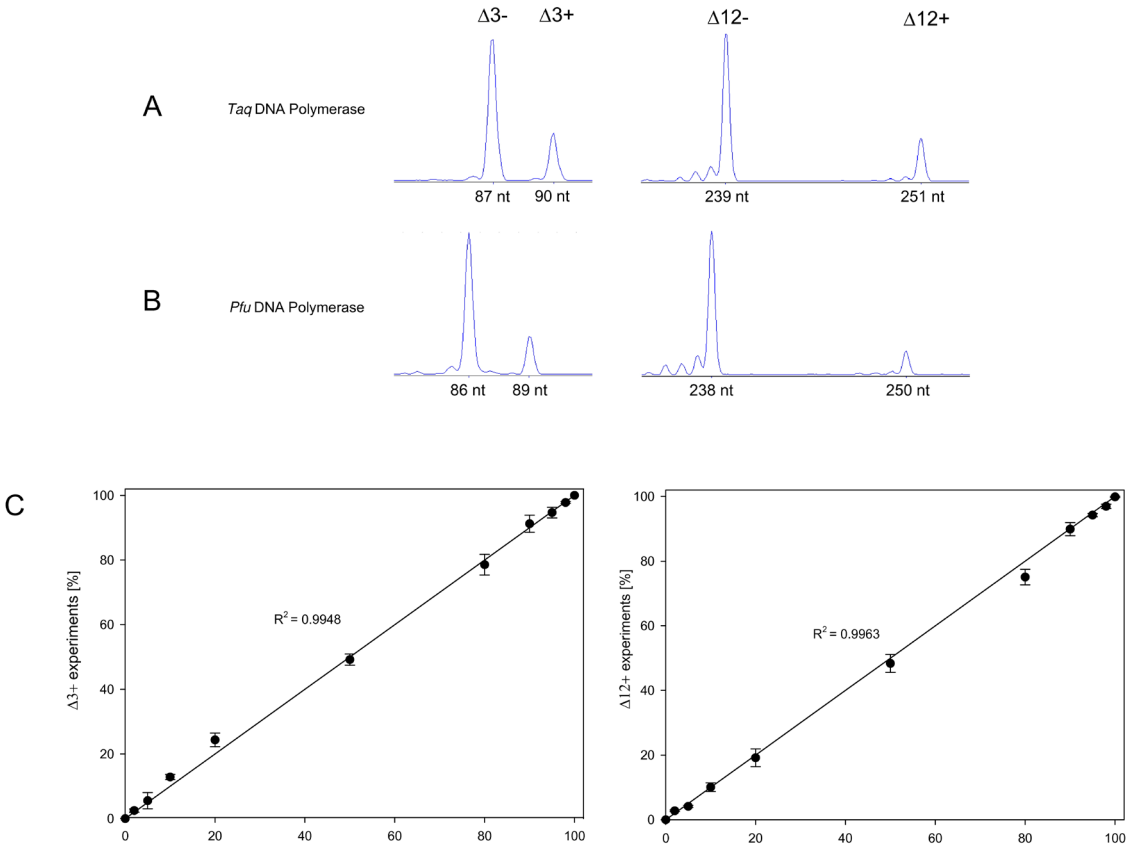


Figure 4:

Pyrograms of the Δ3 and Δ12 test cases. (A, E) The corresponding dispensation orders of the reverse sequencing directions are illustrated. The numbers indicate the quantity of nucleotides which are incorporated in either variant. Red highlighted numbers specify the incorporations into the long, blue highlighted numbers into the short isoform. (B, F) Pyrograms originating from incorporations into the long isoform, (C, G) pyrograms derived from the short isoform, or (D, H) from a 1:1 mixture of both. The nucleotide sequences of the respective splice variants are shown on the right. The 3' part of the pyrosequencing primer is indicated by horizontal arrows.

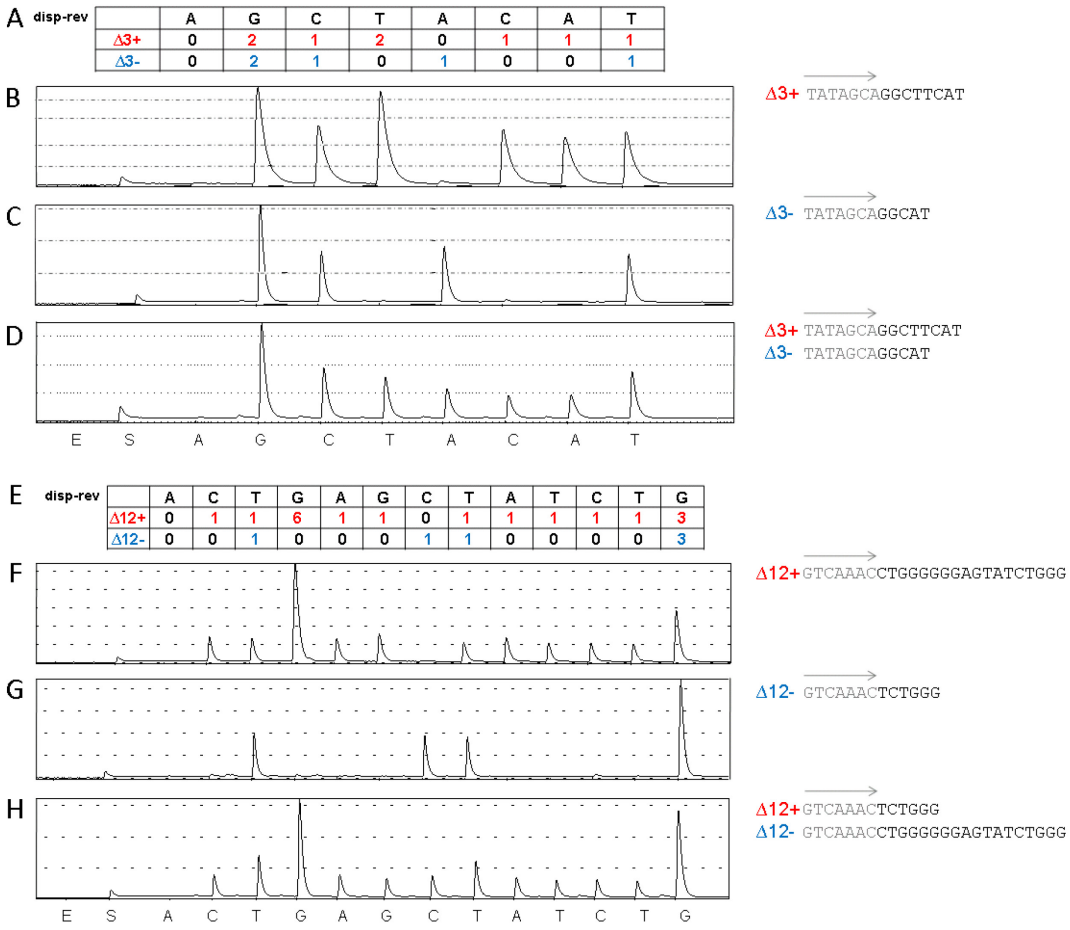
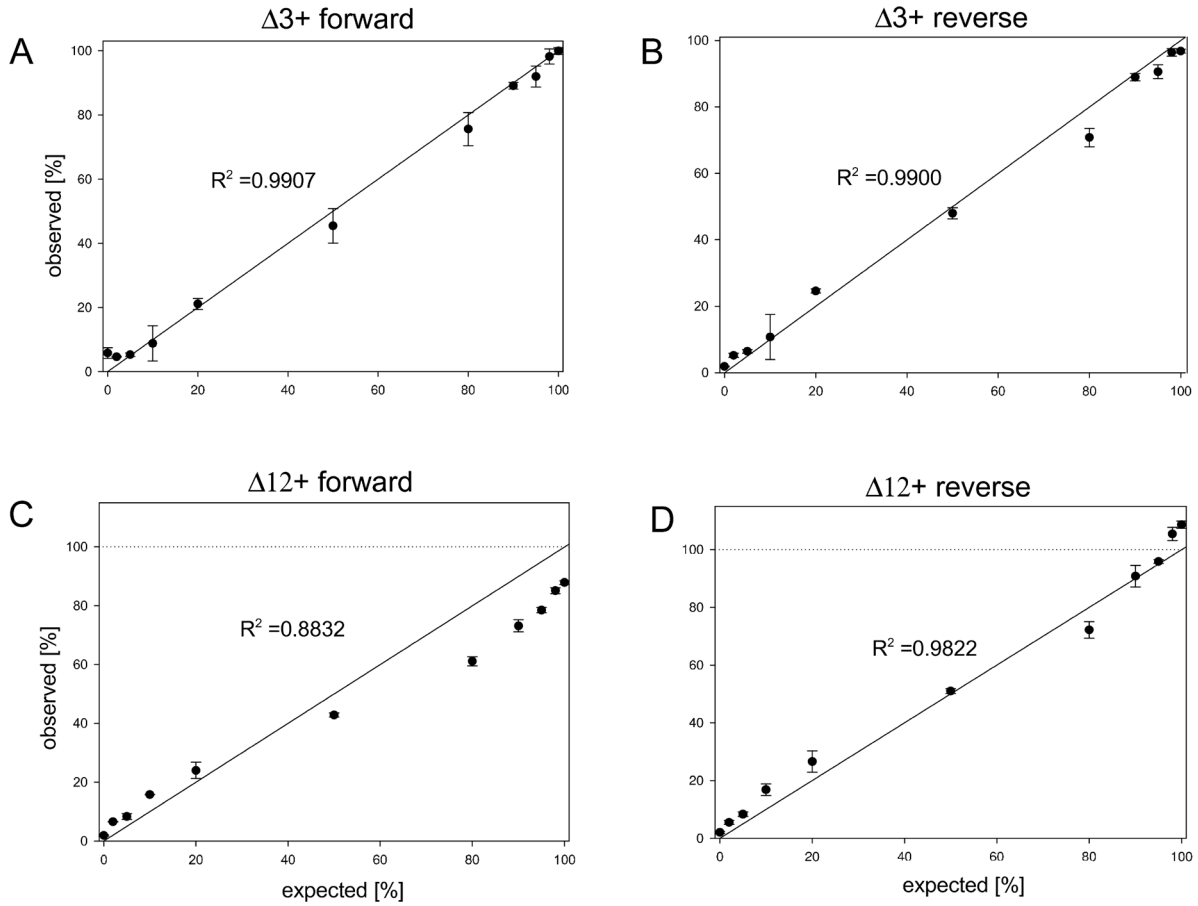


Figure 5:

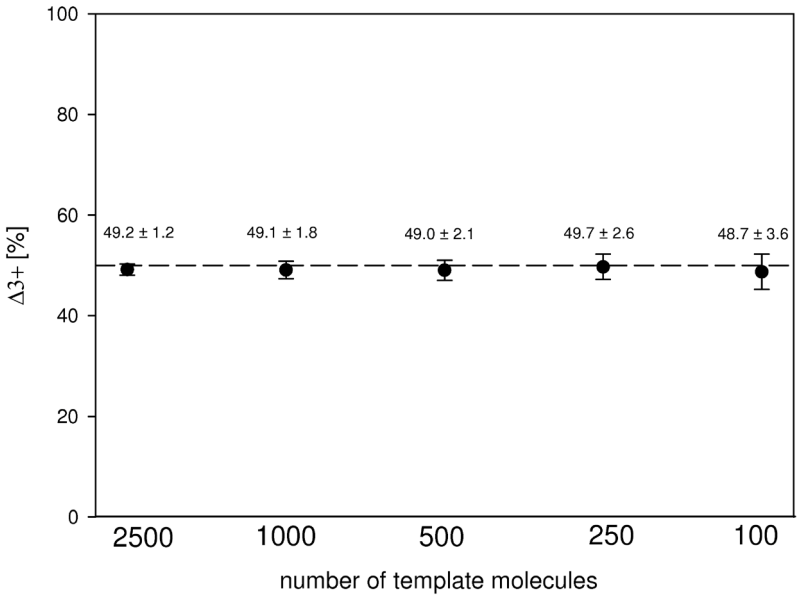
Validation of PSQ with preset mixtures of Δ3 and Δ12 splice variants. The measured fractions of the long splice variants are plotted against the experimentally expected fractions (A, B: Δ3; C, D: Δ12). The diagonal indicates perfect match between expectations and measurements. (A) and (C) show the plots for the forward, (B) and (D) for the reverse sequencing direction. The illustrated fractions represent the proportion of the Δ3+ and Δ12+ variants.



Supplementary Material

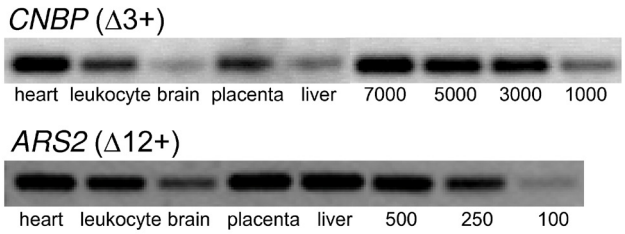
Supplementary Figure 1:

Nested PCR with varying template amounts. The Δ3 isoforms were mixed in a proportion of 50% (indicated by the dashed line) and PCR amplified. Isoform ratios were quantified by CE-LIF and values were obtained by at least three independent measurements.



Supplementary Figure 2:

RT-PCR products from various human tissues. cDNA samples from human heart, leukocytes, brain, placenta, and liver tissues were RT-PCR amplified and compared to PCR-products obtained from cloned splice variants with preset molecule numbers. The agarose gels show *CNBP* (Δ3) and *ARS2* (Δ12) RT-PCR products originating from ~7,000 to ~1,000, and ~200 to ~500 template molecules, respectively. Due to the low resolution of agarose gels the splice variants cannot be identified as separate bands.



DISKUSSION

AS ist ein wichtiges Element der Genregulation in höheren Eukaryoten. Die dadurch erzeugte Variabilität des Spleißprozesses ist eine Hauptquelle genetischer Diversität und trägt entscheidend zur Komplexität des Transkriptoms und Proteoms einer Zelle bei [8]. Die zentrale Thematik dieser Dissertationsschrift ist das subtile AS. Anhand von Analysen von NAGNAG- und TG/AG-Tandem-Spleißstellen wurden Untersuchungen bezüglich der Verbreitung des subtilen AS durchgeführt und Beiträge zu dessen Analytik geleistet [15,16]. Verschiedene Herangehensweisen in Datenakquisition, -verarbeitung, und -interpretation wurden etabliert und die im Zuge meiner Studien gesammelten Erfahrungen im Bereich der Erhebung quantitativer Daten zusammengefasst und diskutiert [14]. Auf meinen experimentellen Ergebnissen aufbauend, habe ich schließlich versucht, regulatorische und mechanistische Aspekte dieser Klasse von Spleißereignissen besser zu verstehen.

Verschiedene Studien der letzten Jahre haben gezeigt, dass zahlreiche alternative 5‘- und 3‘-Spleißstellen sowohl in Säugetieren [9,10,12,56,57] als auch in Pflanzen [15,28,73] nur wenige Nukleotide voneinander entfernt sind. Dabei beläuft sich der am häufigsten auftretende Abstand zwischen alternativen 3‘-Spleißstellen auf 3 nt [10,28]. In meiner Studie „Alternative Splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes“ [15] konnte ich nicht nur die häufige Präsenz von NAGNAG-Tandem-Spleißstellen in der Modellpflanze *A. thaliana* mithilfe von Transkript-Daten und einer sequenzbasierten Vorhersagemethode nachweisen, sondern auch deren alternative Verwendung bei der Expression von SR-Protein-kodierenden Genen experimentell validieren. Bemerkenswerterweise konnten wesentliche humane Befunde [12] im Reich der Pflanzen reproduziert werden: Sowohl die Häufigkeit des genomweiten Auftretens von NAGNAG-3‘-Spleißstellen, als auch die Überrepräsentation in der Gruppe der SR-Protein-kodierenden Genen ist mit dem Menschen vergleichbar. Bei der experimentellen Analyse fiel außerdem auf, dass die Spleißvarianten-Verhältnisse an NAGNAG-Tandem-Spleißstellen der untersuchten SR-Proteine eher organ- und bedingungsspezifisch als genspezifisch sind.

In meiner zweiten Arbeit „Violating the splicing rules: TG dinucleotides function as alternative 3’ splice sites in U2-dependent introns“ [16] habe ich eine bis dato nur von Einzelfällen [107-109] bekannte Form des subtilen AS im Mensch systematisch untersucht und experimentell an humanen RNA-Proben validiert: Wir konnten eine kleine Population von 36 Introns identifizieren, die TG-Dinukleotide als alternative 3’-Spleißstellen verwenden. Diesen TG-Spleißstellen kommt eine Sonderstellung zu, da sie der GY-AG-Regel für Intron-Termini widersprechen. Unseren Ergebnissen zufolge werden TG-3‘-Spleißstellen niemals als alleinige 3’-Spleißstellen, sondern ausschließlich zusammen mit einer alternativen

AG-Spleißstelle gefunden, von der sie maximal 28 nt entfernt sind. Interessanterweise ist die flankierende intronische Sequenz bei etwa der Hälfte der orthologen TG-3'-Spleißstellen zwischen Säugetieren auffällig stark konserviert - in einigen Fällen bis zu Huhn, Frosch oder Pufferfisch.

Schließlich habe ich in meiner dritten Arbeit „Comparison of methods for splice variant quantification" [14] die von mir angewandten, für die Quantifizierung von Spleißvarianten mit kleinen Längendifferenzen adaptierten Methoden der PSQ [110] und CE-LIF [115,116] vorgestellt und im Vergleich zu der weit verbreiteten PAGE-ED [70,117] in einer systematischen Analyse hinsichtlich Reproduzierbarkeit und Genauigkeit der gemessenen quantitativen Daten, sowie Experimentaufbau, Datenanalyse und Anwendungsspektrum evaluiert. Außerdem habe ich den Einfluss von niedrigen Template-Konzentrationen und Amplikon-Längendifferenzen in der RT-PCR untersucht, die zu zufälligen bzw. systematischen Fehlern führen können.

Die zur Identifizierung und Lokalisierung von Spleißereignissen verwendeten Datensätze (RefSeq-, mRNA-, EST-Daten) werden unter der Annahme verwendet, dass sie ein Spektrum natürlicher Transkriptvarianten repräsentieren. Die Identifizierung von TG/AG-Tandem-Spleißstellen im Menschen bzw. die Analyse von NAGNAG-Tandem-Spleißstellen in *A. thaliana* erfolgte mittels verschiedener Herangehensweisen, da für jede Aufgabenstellung unterschiedliche Herausforderungen der vorhandenen Transkript-Datensätze bewältigt werden mussten:

(A) Für einen ersten Ansatz bei der Suche nach TG/AG-Tandem-Spleißstellen in den humanen EST-Daten bestand die Gefahr, dass in den sehr umfangreichen humanen Datensätzen durch Sequenzierfehler bzw. durch irreführende Alignments Spleißvariationen vorgetäuscht werden. Es musste also eine Strategie gewählt werden, die zum einen diese potentiellen Fehlerquellen umgeht, zum anderen aber sensitiv genug ist, um Ereignisse mit kleinen Variationen und Häufigkeiten zu detektieren. Die Abundanz der EST-Daten ermöglichte es, statistische Methoden in Kombination mit entsprechenden Filterkriterien anzuwenden, um natürlich vorkommende Spleißvarianten von den o.g. Artefakten zu trennen. Damit ist es gelungen, neue subtile alternative Spleißereignisse von irreführenden Daten zu unterscheiden, die mit einer Erfolgsrate von 92% experimentell validiert werden konnten [16]. In einer zweiten unabhängigen Analyse wurden, bezugnehmend auf frühere Studien [97,100], ausschließlich RefSeq-Daten herangezogen. Neben sechs TG/AG-Tandem-Spleißstellen, die in beiden Datensätzen detektiert wurden, konnten die RefSeq-Daten nur für drei Fälle Evidenz liefern. Es erscheint zunächst paradox, dass die Analyse der

EST-Daten bessere Resultate erzielen konnte (27 exklusive Fälle) als die Analyse qualitätskontrollierter Datensätze. In Übereinstimmung mit anderen Studien [118] liegt dieser Befund darin begründet, dass RefSeq-Daten trotz Qualitätskontrolle nicht frei von Annotations-Artefakten sind [100], die durch Filterkriterien schwer erfasst werden können bzw. aufgrund der mangelnden Redundanz keine Filterung zulassen. Nichtsdestotrotz besteht die Wahrscheinlichkeit, dass weitere subtile alternative 3'-Spleißstellen vorerst unentdeckt bleiben, weil sie in Regionen mit nicht ausreichender EST-Abdeckung lokalisiert sind. Desweiteren soll erwähnt werden, dass auch ein Teil der Fälle, die die angewandten Filter- und Validierungskriterien nicht passiert haben, trotzdem Spleißvarianten repräsentieren können.

(B) Die Analyse der NAGNAG-Tandem-Spleißstellen in *A. thaliana* stellte insofern eine Herausforderung dar, weil die Abdeckung des *A. thaliana*-Genoms mit Transkriptdaten im Vergleich zu Mensch oder Maus niedrig ist. EST-basierte Belege für alternative Spleißereignisse sind jedoch von der Häufigkeit der Transkriptisoformen abhängig, denn die Wahrscheinlichkeit, seltene Transkriptisoformen zu identifizieren, sinkt mit der Anzahl der ESTs eines Gens. Daher ist die Identifizierung von alternativen Spleißereignissen in *A. thaliana* wesentlich schwieriger als in Organismen, die über umfangreichere Datensätze verfügen. Fälle von AS können daher unentdeckt bleiben und der tatsächliche Anteil an Spleißereignissen wesentlich höher sein. EST-Evidenz für AS existierte bis dato für lediglich 5% aller NAGNAG-3'-Spleißstellen in *A. thaliana* [15]. Um den Mangel an ausreichender Transkript-Abdeckung in *A. thaliana* zu kompensieren, wurde die Analyse des subtilen AS an NAGNAG-Tandem-Spleißstellen mit einer prognostischen Methode vervollständigt. Diese basiert auf dem unmittelbaren Sequenzkontext der NAGNAG-Spleißstellen, der dafür bekannt ist, relevante Informationen für eine Vorhersage von AS zu enthalten [9,119]. Die Wahrscheinlichkeit, dass ein Sequenzkontext AS begünstigt, wurde mithilfe der vorhandenen Transkriptevidenz ermittelt. Mit dieser Methode konnte für die Fälle, die über ungenügende Transkript-Abdeckung verfügen, die Wahrscheinlichkeit alternativ gesplissen zu werden, vorhergesagt werden. Unsere prognostische Methode hat sich als spezifisch herausgestellt, da 53% (8/15) der als AS positiv prognostizierten und für die Validierung ausgewählten Fälle experimentell bestätigt werden konnten. Basierend auf unserer Vorhersage und der ermittelten Validierungsrate kann angenommen werden, dass ca. 17% der NAGNAG-Tandem-Motive in *A. thaliana* dem AS unterliegen. Limitierte Transkriptdaten können also mithilfe einer geeigneten Vorhersagemethode kompensiert werden. Diese bioinformatische Methode umfasste jedoch keine Kriterien wie beispielsweise Gewebespezifität. Eine

Prognose, die derartige Aspekte des subtilen AS vorhersagt, stellt eine neue bioinformatische Herausforderung dar.

Zur experimentellen Analyse der Spleißereignisse habe ich unterschiedliche Quantifizierungsmethoden angewendet: CE-LIF [115,116], PSQ [110] und Klonierung/Sequenzierung von RT-PCR-Produkten [120]. Letztgenannte Methode ist für die Aufklärung neuer Spleißvarianten sehr gut geeignet, weil die Sequenzinformation von Exon-Übergängen generiert wird [121]. Für routinemäßige quantitative Aufgaben ist sie jedoch zu arbeits- und zeitaufwendig, da erst mit einer großen Zahl von Klonen (≥100) eine Sensitivität >95% für AS-Ereignisse mit einer Häufigkeit von >3% der seltenen Spleißvariante erzielt werden kann [16]. Für einen höheren Probendurchsatz sind die Methoden der PSQ und CE-LIF besser geeignet (Tab. 1), die ich für die Spleißvarianten-Quantifizierung adaptiert habe. Neben anderen etablierten Quantifizierungsmethoden, wie z.B. „Northern-Blotting“ [122,123] oder „Real-Time-quantitative-PCR“ [124-126], bieten beide Methoden vor allem den Vorteil, dass sie mit einer maximalen Auflösung von 1-2 nt ausgezeichnet für die Analyse von Spleißvarianten mit kleinen Längenunterschieden infolge von subtilem AS geeignet sind (Tab. 1).

Tabelle 1: Vergleich technischer Charakteristika von PAGE-ED, PSQ und CE-LIF. (#) nach etablierter RT-PCR. Tabelle aus dem Manuskript Schindler *et al.* 2009 dieser Dissertation [14].

	PAGE-ED	PSQ	CE-LIF
Assay Design #	0.5 d	> 5 d	0.5 d
Assay	6-8 h	4 h	2 h
Probenanzahl	20	96	96
Amplikongröße	60 - 600 nt	40 - 100 nt	60 - 600 nt
Auflösung	2 nt	1 nt	2 nt
durchschnittliche Genauigkeit (R^2)	0.9877	0.9876	0.9956
Reproduzierbarkeit (max. Abweichung vom Mittelwert)	± 11.0 %	± 6.8 %	± 3.3 %

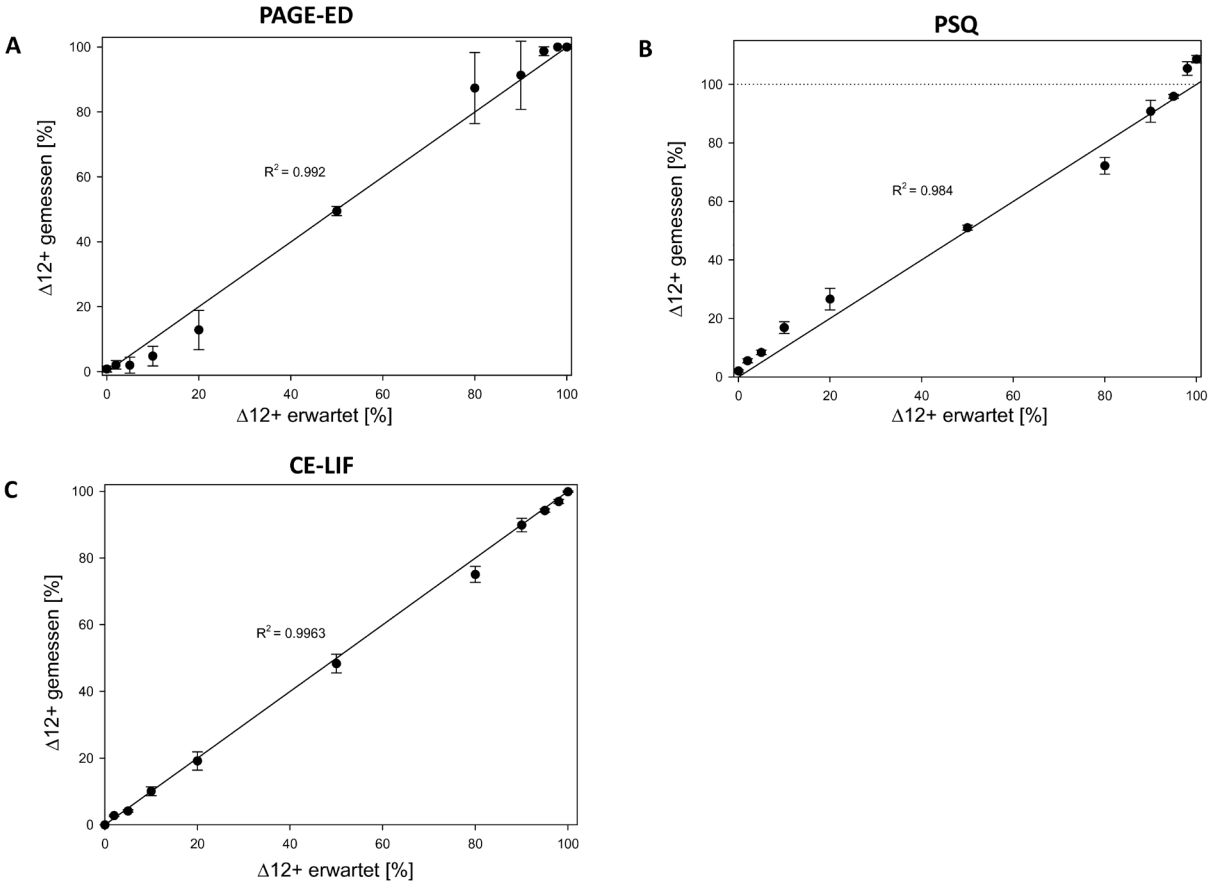


Abbildung 7: Validierung von PAGE-ED, PSQ und CE-LIF mittels voreingestellter Gemische von Spleißvarianten mit 12 nt Größendifferenz. (A) PAGE-ED, (B) PSQ, (C) CE-LIF. Gemische mit definierten Verhältnissen an langer und kurzer Isoform wurden voreingestellt und PCR-amplifiziert. Die Diagonale markiert eine perfekte Übereinstimmung von erwarteten und gemessenen Anteilen. Das Bestimmtheitsmaß (R^2) gibt die Streuung der Messwerte von den erwarteten Werten und somit die Genauigkeit des Experiments an. Die Standardabweichung der gemittelten Messergebnisse (Fehlerbalken) zeigt die Reproduzierbarkeit der Messwerte. Die experimentell gemessenen Anteile an langer Spleißvariante (Δ12+) sind gegen die erwarteten Anteile aufgetragen. In Anlehnung an das Manuskript Schindler *et al.* 2009 dieser Dissertation [14].

PSQ und CE-LIF wurden in einer systematischen Analyse in Bezug auf Reproduzierbarkeit und Genauigkeit der gemessenen quantitativen Daten, sowie auf Experimentaufbau, Datenanalyse und Anwendungsspektrum evaluiert. Als Vergleichsmethode habe ich die in vielen Laboren gebräuchliche Methode der PAGE-ED herangezogen [70,117]. Ich stellte fest, dass CE-LIF die höchste Genauigkeit und Reproduzierbarkeit erzielte und gleichzeitig die arbeits- und zeiteffizienteste Methode darstellte [14] (Abb. 7, Tab. 1). Experimentplanung und -aufbau sind sehr übersichtlich: Es muss lediglich eine RT-PCR mit Fluoreszenz-markierten Primern etabliert werden, damit die Produkte direkt auf einem ABI3730xl-Kapillarsequenzierer analysiert werden können. Im Vergleich dazu sind Etablierung und Datenanalyse von PSQ-Assays mit einem hohen Genauigkeitsgrad sehr

zeit- und arbeitsaufwendig [14,127] (Tab. 1). Ist der Assay jedoch konzipiert, kann diese Methode routinemäßig für die Quantifizierung von Spleißvarianten-Verhältnissen verwendet werden. Die Methode der PAGE-ED gibt in kürzester Zeit eine ungefähre Einschätzung von Spleißvarianten-Verhältnissen (Abb. 7A). Im Gegensatz zu CE-LIF oder PSQ, ist diese für hoch genaue Messungen subtiler quantitativer Unterschiede sowie für Hochdurchsatz-analysen jedoch weniger gut geeignet (Tab. 1). Auch das Anwendungsspektrum unterscheidet sich zwischen den beschriebenen Quantifizierungs-Methoden [14]: Beispielsweise können mittels CE-LIF und PAGE-ED unbekannte Transkriptisoformen identifiziert und quantifiziert werden (Tab. 1) – sogar in sehr komplexen Nukleinsäuregemischen. Im Gegensatz dazu können PSQ-Assays nur für bekannte alternative Spleißereignisse etabliert und eingesetzt werden.

Wichtige Aspekte, die vor jeder Spleißvarianten-Quantifizierung beachtet werden sollten, sind die Reaktionsbedingungen der RT-PCR [128-131], eine niedrige Zahl von Template-Molekülen und Amplikons mit großen Längendifferenzen. Diese Faktoren können zu falschen Messergebnissen der Spleißvarianten-Verhältnissen führen [132]. Die von mir verwendete Methode der verschachtelten PCR („nested-PCR“), bei der das Produkt einer ersten PCR als Matrize für eine zweite Amplifikation mit „inneren“ Primern verwendet wird, ist dabei folgendermaßen von Vorteil: (i) Sie verfügt über eine höhere Sensitivität als eine Standard-PCR. (ii) Werden Zyklenzahlen erhöht, können Verzerrungen reduziert werden, die durch nachlassende Polymerase-Leistungsfähigkeit über lange Inkubationszeiten auftreten können [133]. Bei der Analyse von PCR-Amplikons mit verschiedenen Längendifferenzen (3 nt, 12 nt, 195 nt) zeigten die Testfälle, die sich infolge von subtilem AS in nur 3 nt bzw. 12 nt unterscheiden, auch bei hohen Zyklenzahlen genaue Messergebnisse [14]. Eine deutliche Verfälschung der Spleißvarianten-Verhältnisse war hingegen bei PCR-Amplikons erkennbar, die wesentlich größere Längendifferenzen (195 nt) aufwiesen [14]. Ursachen solcher Verzerrungen ist das bekannte reziproke Verhältnis von PCR-Effektivität und Template-Länge. Optimierungs-Ansatzpunkte können hierbei beispielsweise die Zyklenzahl und/oder die zugegebene Menge an Nukleotiden sein, die im Falle längerer Amplikons schneller erschöpft ist bzw. eine Optimierung der Elongationszeit, die zu vorzeitigen Kettenabbrüchen führen kann, wenn diese zu kurz ist. Die Resultate meiner Analyse von niedrigen Template-Konzentrationen zeigten erwartungsgemäß, dass die experimentelle Schwankung mit abnehmender Molekülzahl zunimmt [14]. Niedrige Template-Molekülzahlen treten bei niedriger exprimierten Genen auf, da in einer RNA-Probe dann nur wenige Transkript-Moleküle eines niedrig exprimierten Gens enthalten sind. Wird davon eine Probenmenge entnommen und für die RT-PCR eingesetzt, ist es möglich, dass in dieser Probenmenge

Moleküle einer Spleißvariante nicht erfasst worden sind bzw. der relative Anteil einer Spleißvariante zu niedrig oder zu hoch ist. Das Verhältnis der Spleißvarianten nach Amplifikation ist dann nicht repräsentativ für das tatsächliche Verhältnis. Wenn diese Messungen nicht oft genug repliziert werden und somit stochastische Schwankungen in einer Fehlerstatistik nicht erfasst werden können, kann im Vergleich mit einer Messung desselben Gens in einem anderen Gewebe Gewebespezifität vorgetäuscht werden. Gerade in solchen Fällen ist eine ausreichende Anzahl an Mess-Replikaten ($n \geq 3$) mit anschließender Fehlerstatistik notwendig, um die Genauigkeit des Ergebnisses abzuschätzen. Mit diesem Schema habe ich eine solide analytische Grundlage geschaffen, um gewebespezifische Unterschiede von Spleißvarianten-Verhältnissen in systematischen Studien zu klären. Es ist weiterhin bekannt, dass Spleißvarianten-Verhältnisse inter-individuelle Schwankungen aufweisen können. Aktuelle Schätzungen gehen davon aus, dass etwa 21% aller alternativ gesplissenen Gene von Polymorphismen betroffen sind, die die relativen Häufigkeiten einiger alternativer Spleißvarianten ändern [134]. Um interindividuelle Einflüsse zu minimieren, habe ich in meinen Quantifizierungsexperimenten vorzugsweise gepoolte RNA verschiedener Individuen verwendet.

Die Quantifizierungsdaten meiner Studie [16] zeigen für die untersuchten TG/AG-Fälle *ARS2* („arsenate resistance protein 2“), *CNBP* („CCHC-type zinc finger, nucleic acid binding protein“) und *BRUNOL4* („bruno-like 4, RNA binding protein“) statistisch signifikante gewebe-spezifische Unterschiede der Spleißvarianten-Verhältnisse. Bei *CNBP* liegt die Schwankungsbreite des Anteils an TG-Spleißvariante zwischen 20% (Gehirn) und 41% (Leukozyten), bei *ARS2* zwischen 48% (Testis) und 63% (Thymus) [16]. In der Analyse von *BRUNOL4* konnte ein Anteil von 20% an TG-Spleißvariante im Gehirn und 0% in der Lunge gemessen werden [16]. Die Verwendung der TG/AG-Tandem-Spleißstellen der analysierten Fälle wird also differentiell reguliert, d.h. dass die Häufigkeit der Benutzung dieser Tandem-Spleißstellen je nach Gewebe variiert wird. Da die Quantifizierungsdaten von *BRUNOL4* jedoch mittels Auszählung sequenzierter RT-PCR-Klone in einer einzigen Messung ermittelt wurden und *BRUNOL4* in der Lunge niedrig exprimiert wird, sollte dieser Fall nach aktuellem Wissensstand nochmals mit dem entwickelten Schema von Mehrfachmessungen und anschließender Fehlerstatistik überprüft werden.

Differentielle Regulation von Spleißvarianten-Verhältnissen wird oft als Indiz für Funktion angesehen. Bemerkenswerterweise haben jedoch verschiedene Tandem-Spleißstellen mit klaren funktionellen Auswirkungen konstante Spleißvarianten-Verhältnisse [25,59,62,63,65,67,70,74,77,78]. Ein Beispiel ist das *ING4*-Gen („inhibitor of growth family,

member 4“): Die vier Spleißvarianten, die durch AS an 5’- und 3’-Tandem-Spleißstellen im selben Intron in konstantem Verhältnis gebildet werden, unterscheiden sich in ihren Transkriptions-regulatorischen Eigenschaften und unterdrücken damit Zellwachstum und -migration [67,131]. Ein weiteres Beispiel ist die 5’-Tandem-Spleißstelle des *WT1*-Gens (siehe Einleitung), dessen abgeleiteten Proteinisoformen in einem konstanten Verhältnis gebildet werden [68,69,135,136], jedoch funktionell unterschiedlich sind [63]. Abweichungen von diesem Verhältnis haben ausgeprägte phänotypische Konsequenzen [68]. Konstante Isoformen-Verhältnisse an Tandem-Spleißstellen sprechen also keineswegs gegen eine funktionelle Bedeutung. Im Gegenzug gibt es Studien, die keine Funktionalität für subtile alternative Spleißereignisse ermitteln konnten, bei denen Spleißvarianten in variablen Verhältnissen produziert werden. Ein Beispiel sind die Spleißvarianten der 3‘-Tandem-Spleißstellen des *PIMT2*-Gens von *A. thaliana*, deren Verhältnis entwicklungsspezifische und hormonbedingte Veränderungen wiederspiegelt [137].

Auch für NAGNAG-Tandem-Spleißstellen in den SR-Protein-kodierenden Genen in *A. thaliana* wurden konstante (*U11-35K*, At2g24350, At5g59950, At3g54230) bis numerisch geringe, jedoch statistisch signifikante (*SR33, RS41,* At4g35785*, U2AF6*) organspezifische Schwankungen der relativen Spleißvarianten-Anteile gemessen [15]. Es fällt jedoch auf, dass sich die organ-, stress- sowie entwicklungsspezifischen Spleißvarianten-Verhältnisse der verschiedenen Gene ähneln (Abb. 8). So wurde bei den meisten Genen nach Hitzeschock die Expression der langen Transkriptvariante verringert und nach Kältestress gesteigert, bzw. wies der Blütenstand in nahezu allen Fällen den geringsten Anteil an langer Transkriptisoform auf [15]. Die differentiellen Effekte auf das AS von NAGNAG- Tandem-Spleißstellen scheinen also eher organ- und bedingungsspezifisch als genspezifisch zu sein. Ob dies auch für andere NAGNAG-Tandem-Spleißstellen in *A. thaliana* zutrifft oder ein Charakteristikum SR-Protein-kodierender Gene ist bzw. welche funktionelle Bedeutung damit verbunden ist, erfordert weiterführende experimentelle Ansätze. Interessanterweise sind von AS an NAGNAG-Tandem-Spleißstellen in Säugetieren und Pflanzen vorzugsweise polare Aminosäure-Reste betroffen [12,138]. Polare Serin-Reste in SR-Proteinen sind wichtige Phosphorylierungspunkte und zahlreiche Studien belegen, dass der Phosphorylierungs-Status von SR-Proteinen für deren Aktivität kritisch ist [84,86,139,140]. SR-Proteine könnten an diesen Stellen gezielt subtil verändert werden, um deren spleißregulatorische Aktivität zu beeinflussen. Dies muss jedoch bewiesen werden.

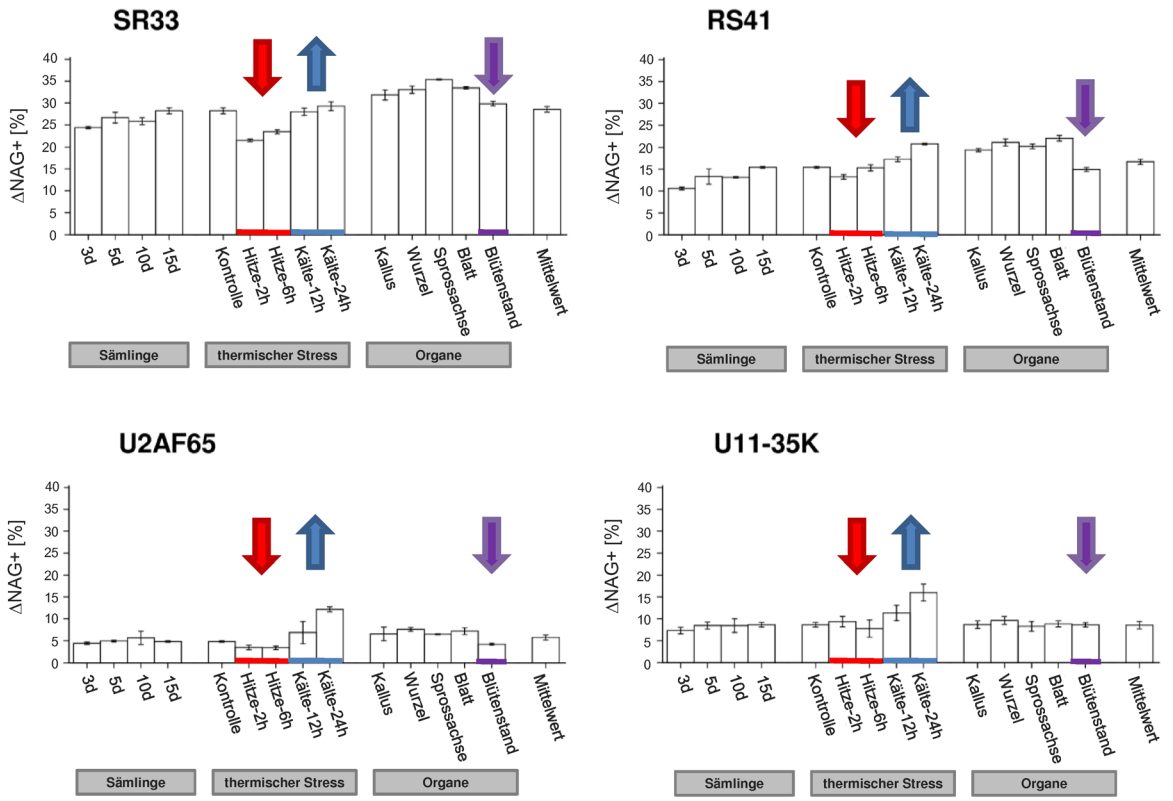


Abbildung 8: Effekte verschiedener Bedingungen auf die Spleißvarianten-Verhältnisse von NAGNAG-Tandem-Spleißstellen in SR-Protein-kodierenden Genen der Pflanze *A. thaliana*. Der relative Anteil der langen Spleißvariante (NAG+) ist dargestellt. Nach Hitzeschock wird die Expression der langen Transkriptvariante herunter reguliert (roter Pfeil), nach Kälteschock hoch reguliert (blauer Pfeil). Der Blütenstand zeigt im Vergleich zu den anderen Pflanzenorganen den geringsten Anteil Anteil der Isoform mit NAG-Insertion (lila Pfeil). In Anlehnung an Schindler *et al*. 2008 dieser Dissertation [15].

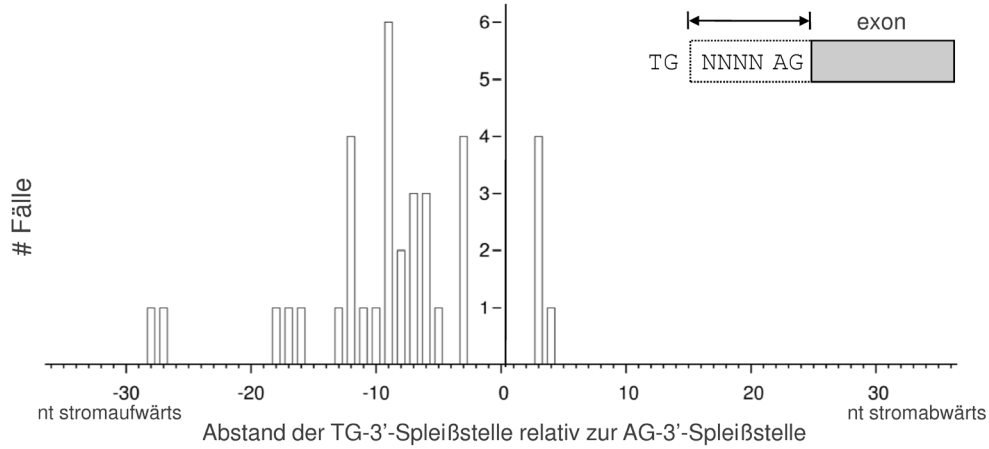


Abbildung 9: Distanz-abhängiges Auftreten von TG/AG-Tandem-Spleißstellen. Bei den Fällen des negativen Bereiches der x-Achsen-Skalierung ist die TG-3'-Spleißstelle stromaufwärts von der AG-3'-Spleißstelle lokalisiert, in den Fällen des positiven Bereiches stromabwärts.

In Bezug auf TG/AG-Tandem-Spleißstellen habe ich mich besonders mit der Frage nach den mechanistischen Grundlage dieser Form des AS beschäftigt [16]. Dabei fielen folgende Besonderheiten auf: (i) Ein TG-Dinukleotid benötigt zwingend den Kontext einer alternativen AG-Spleißstelle, um als 3'-Spleißstelle fungieren zu können; (ii) die maximale Distanz zwischen TG- und AG-Spleißstelle ist 28 nt (Abb. 9); (iii) die Konservierung von TG-Dinukleotiden als alternative 3'-Spleißstelle zwischen entfernt verwandten Vertebraten und (iv) die für AS generell typische höhere Konservierung von flankierender Sequenz (z.B. *RYK*, Abb. 10) im Vergleich zu konstitutiven Spleißorten. Letzteres weist auf die Beteiligung *cis*-, möglicherweise auch *trans*-Elemente hin, die die Wahl der TG/AG-Spleißstellen vermitteln. Unter bestimmten Bedingungen wird dem Spleißosom also die Flexibilität verliehen, TG als 3'-Spleißstelle zu wählen. Gleichzeitig ist es jedoch auf eine benachbarte AG-Spleißstelle angewiesen, da ein Distanzlimit zwischen TG- und AG-Dinukleotid besteht und konstitutives Spleißen an TG-Spleißstellen nicht detektiert werden konnte. Es kann deshalb angenommen werden, dass die AG-Spleißstelle, die sich im typischen Kontext von „Branchpoint"-Motiv [141] und Polypyrimidintrakt befindet, für die qualitative Definition des intronischen 3'-Terminus im ersten Schritt der Spleißreaktion unverzichtbar ist. Die positionsgenaue Definition der 3'-Spleißstelle erfolgt dann in einem späteren Schritt. Diese Vermutung wird durch Studien untermauert, die gezeigt haben, dass der dem basalen Spleißosom zugehörige essentielle Spleißfaktor U2AF35 in Kombination mit anderen Faktoren die Spezifität des Spleißosoms für ein AG als intronischen 3'-Terminus vermittelt [142]. Nachdem der Spleißfaktor U2AF vom spleißosomalen Komplex nach dem ersten Schritt dissoziiert ist [143,144], könnten andere Faktoren die endgültige Wahl der Spleißstelle im zweiten Schritt beeinflussen. Studien der AG/AG-Tandem-Spleißstellen des Intron 2 im Drosophila *sxl*-Gen [145] sowie des Intron 1 der β-Globin-Mutante β^{110} [146,147] zeigen, dass die stromabwärts gelegene AG-Spleißstelle für das Spleißen zwingend notwendig ist, während das entbehrliche stromaufwärts gelegene AG im zweiten Schritt des Spleißvorgangs optional gewählt werden kann. Die Wahl dieser Spleißstelle erfolgt über den Spleißfaktor SPF45, der an das stromaufwärts gelegene AG bindet [145]. Ob SPF45 oder andere Faktoren auch an der Wahl der TG-Spleißstelle beteiligt sind, muss in Zukunft untersucht werden.

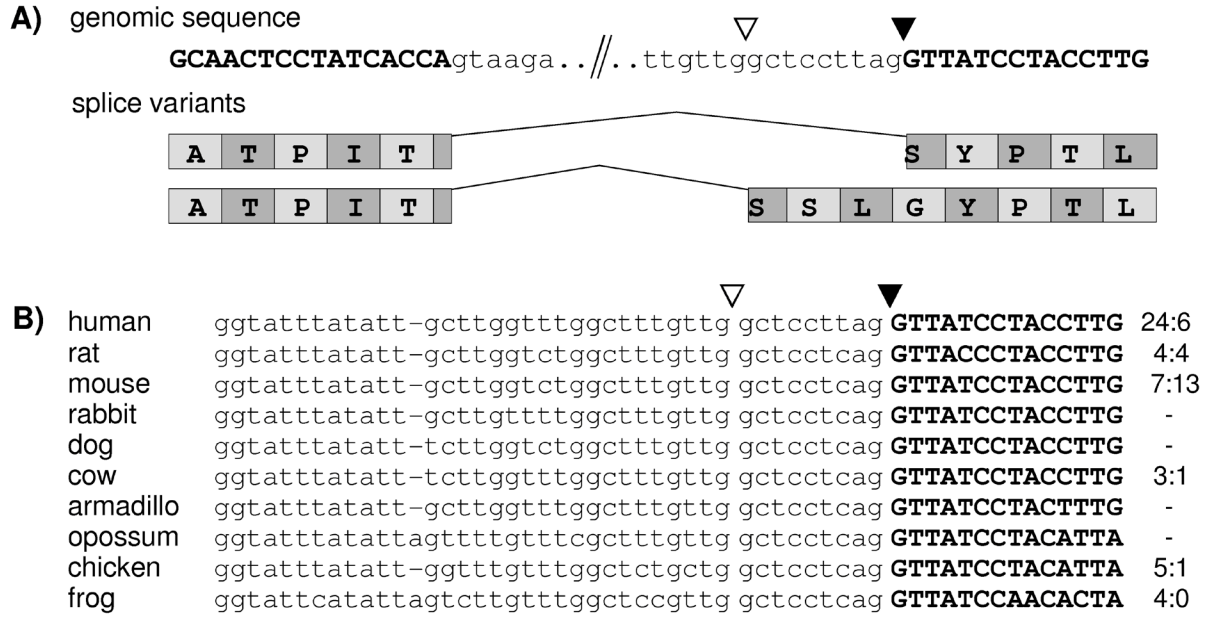


Abbildung 10: Konservierung der alternativen TG-3'-Spleißstelle und der flankierenden Sequenz im *RYK*-Gen Intron 7. Abbildung aus der Publikation „Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns" dieser Dissertation [16]. (A) Humane genomische Sequenz mit abgeleiteten Spleißvarianten. (B) Alignierung der orthologen Intron-Exon-Grenzen verschiedener Vertebraten-Genome. Die rechts angezeigten Verhältnisse zeigen die Anzahl der Transkriptevidenzen aus den jeweiligen EST-Daten für die TG- und AG-abgeleitete Spleißvariante. (A)+(B) Die TG-3'-Spleißstelle ist mit einem weißen Dreieck, die AG-3'-Spleißstelle mit einem schwarzen Dreieck markiert. Abbildung aus Schindler *et al*. 2007 dieser Dissertation [16].

Der Anteil an alternativ gesplissenen TG/AG-Tandems ist mit 0.01% im Vergleich zu den konstitutiv am AG gesplissenen TG/AG-Intron-Exon-Grenzen sehr gering. In Übereinstimmung mit der überdurchschnittlich hohen Sequenz-Konservierung der Hälfte aller TG-Spleißstellen liegt die Vermutung nahe, dass benachbarte Sequenz-Signale zur Definition der TG-Spleißstelle beitragen. Aus anderen Studien ist bekannt, dass die TG-3'-Spleißstelle im Intron 3 des *GNAS*-Gens [148] von drei putativen Bindemotiven für Spleißregulatoren flankiert ist und die Wahl der TG-3'-Spleißstelle durch die Spleißregulatoren SF2/ASF und hnRNPA1 moduliert wird [107]. Für die anderen Introns mit alternativ verwendeten TG/AG-Tandem-Spleißstellen konnten bisher keine Sequenzmotive identifiziert werden, die mit der Wahl der TG-3'-Spleißstelle assoziiert sind. Aufgrund der relativ geringen Anzahl von Introns mit TG/AG-Tandem-Spleißstellen haben aktuell verfügbare Methoden [149,150] für die Suche von gemeinsamen Sequenzmotiven nur limitierte Detektionsstärke, vor allem wenn sich *cis*-Elemente über einen größeren Bereich verteilen [151] oder mehrere Elemente kooperieren. Es ist ebenso möglich, dass jede

TG/AG-Spleißstelle ein individuelles Ensemble von Spleißregulatoren rekrutiert, um die Wahl der TG-Spleißstelle zu forcieren.

Ein Anteil von 39% aller TG/AG-Tandem-Spleißstellen ist im jeweils orthologen Maus-Intron konserviert, und zeigt dort gleiche Spleißvarianten. Für die Gene *FBXL10* („F-box and leucine-rich repeat protein 10“) und *BRUNOL4* existieren Transkiptevidenzen für AS an orthologen TG/AG-Tandem-Spleißstellen im Frosch (*Xenopus tropicalis*), für *RYK* („receptor-like tyrosine kinase“) sogar im Pufferfisch (*Takifugu rubripes*). Ein gemeinsamer Tetrapoden-Vorfahre hat in *FBXL10* und *BRUNOL4* also sehr wahrscheinlich TG/AG-Tandem-Spleißstellen besessen, der von Pufferfisch und Tetrapoden sogar im *RYK*-Gen. Interessanterweise steigt in einem Vergleich Mensch-Maus die Häufigkeit der Verwendung der TG-Spleißstelle mit der Konservierung von Spleißstelle und flankierender Intron-Sequenz [16]. Introns, deren TG-Spleißstelle mit einer Häufigkeit von >10% benutzt wird, stimmen in der Sequenz ihrer proximalen Flanken (50 nt stromaufwärts der Tandem-Spleißstelle) zu mehr als 80% überein, im Vergleich zu 65% Übereinstimmung bei konstitutiven AG-Introns. Mit geringer Häufigkeit (<10%) benutzte TG-Spleißstellen unterscheiden sich in ihrem Konservierungs-Grad (64%) nicht von AG-Introns - ein Phänomen, was auch in Studien der Konservierung alternativer Exons beobachtet wurde [93]. Diese Ergebnisse legen die Vermutung nahe, dass TG-Spleißstellen im Zuge von neutraler Evolution entstanden sind und zunächst mit einer geringen Effizienz benutzt und toleriert werden. Auf dieser Basis könnten Isoformen funktionelle Bedeutung erlangen und sich unter positiver Selektion Tandem-Spleißorte mit einer höheren Effizienz etablieren, die später durch negative Selektion erhalten werden.

Auf der überdurchschnittlichen Sequenzkonservierung basierend, könnten 40% (14 von 36) der bekannten humanen TG-Spleißstellen eine funktionelle Relevanz zugesprochen werden. In einem Einzelfall ist die biologische Funktion der durch AS an TG/AG-Tandem-Spleißstellen subtil veränderten Proteinisoformen belegt [152]. Die Protein-Isoformen der α_{1A}-Untereinheit eines Calcium-Kanals (*CACNA1A*, „calcium channel, voltage-dependent, P/Q type, alpha 1A subunit“) haben unterschiedliche Eigenschaften in der neuronalen Erregbarkeit [152]. Neben dem möglichen Einfluss auf die Proteinsequenz, erlauben NAGATG-Tandem-Spleißstellen außerdem die Insertion eines Start-Codons. Dies scheint im Falle des *PCGF2*-Gens („polycomb group ring finger 2“) realisiert zu sein, dessen Spleißvarianten sich in der Präsenz eines zusätzlichen stromaufwärts gelegenen Leserahmens voneinander unterscheiden [153] (Abb. 11). Solche sogenannten „upstream ORFs (open reading frames)“ haben meist regulatorischen Charakter und beeinflussen die Translations-

Effizienz des Haupt-Leserahmens [154]. Um den Einfluss des „upstream ORF“ zu untersuchen, habe ich Experimente durchgeführt, die belegen, dass die Insertion des ATG tatsächlich einen Effekt auf die translationale Effizienz des Haupt-Leserahmens aufweist (Daten nicht publiziert). Außerdem habe ich nach *in vitro*-Transkription/Translation von „full-length“ cDNA-Klonen beider Transkriptvarianten eine zusätzliche um etwa 10 kDa größere Proteinisoform mittels Immunopräzipitierung und Western-Blot detektiert. Diese Isoform wurde ausschließlich vom Klon mit ATG-Insertion gebildet. Ob dieses Protein eine N-terminal verlängerte Proteinisoform darstellt, bzw. an welcher Stelle die Translation dieser Isoform initiiert und welcher Zusammenhang mit dem Vorhandensein des zusätzlichen Leserahmens oder Sekundärstrukturen besteht, ist zum aktuellen Zeitpunkt nicht geklärt. Es wäre lohnenswert, die Auswirkungen dieser *PCGF2*-Spleißvariation weiter zu untersuchen, da diese anscheinend weit über die Insertion/Deletion von drei nicht-protein-kodierenden Nukleotiden hinausgehen. Eine Vielzahl der Gene, die von AS an TG/AG-Tandem-Spleiß-stellen betroffen sind, repräsentieren Regulatoren der Chromatinstruktur (*PCGF2, GPBP1, SAP30, SUV420H2, SSRP1, SMARCA4*) sowie Spleißfaktoren und translationale Regulatoren (*CNBP, BRUNOL4, HNRPR, PCBP2*). Zusammen mit einer anderen Untergruppe, die mit Rezeptor-vermittelter Signaltransduktion assoziiert sind (*GNAS, DRD2, FREQ, IL21, RYK, DLG4, RRAD, PTPN11, SYTL2, MARK3, SH3D19*), können diese mit informationsweiterleitenden Prozessen in der Zelle in Verbindung gebracht werden [16]. Es wäre ebenfalls denkbar, dass AS an TG/AG-Tandems nicht die Funktion hat, alternative Transkript- oder Proteinisoformen zu generieren, sondern einen regulatorischen Flaschenhalses für die Reifung des Transkripts darstellt, wie es auch für die U12-Introns bekannt ist [155].

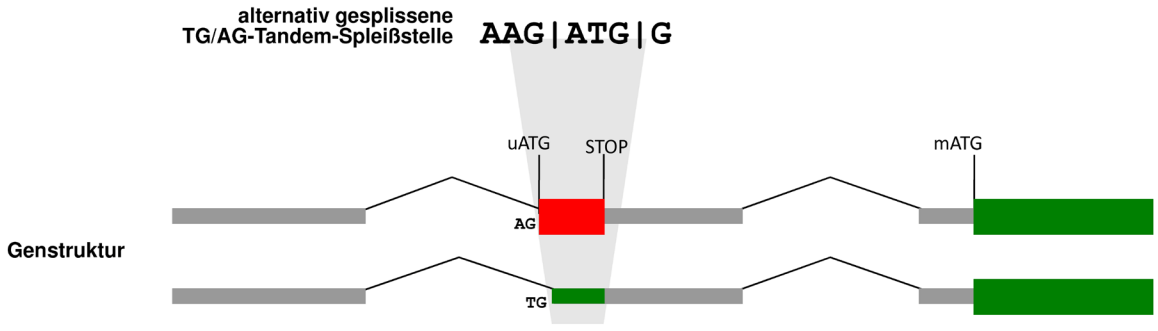


Abbildung 11: Struktur der 5‘ UTR des *PCGF2*-Gens. Bei Benutzung der stromaufwärts lokalisierten AG-3’-Spleißstelle des TG/AG-Tandem-Motivs wird ein zusätzliches Start-Codon (uATG) inseriert. Wird an der 3 nt stromabwärts gelegenen TG-3’-Spleißstelle gesplissen, wird dieses Start-Codon deletiert (grün). Da sich 24 nt stromabwärts der AG-Spleißstelle ein Stop-Codon befindet, kann zusätzlich zum 106 nt entfernt gelegenen Hauptleserahmen (mATG) ein 27 bp großer „upstream open reading frame“ (rot) durch ATG-Insertion generiert bzw. durch ATG-Deletion inaktiviert werden.

Subtiles AS an NAGNAG- und TG/AG-Tandem-Spleißstellen gibt Anlass eine aktuelle Debatte aufzugreifen, ob die Mehrzahl von alternativen Transkripten mit biologischer Relevanz oder infolge eines fehlerhaften Prozesses entstehen [156]. Im Zuge der Studien an NAGNAG-Tandem-Spleißstellen wurde von Chern *et al.* die These vertreten, dass AS an diesen Motiven eine Form von „Rauschen" darstellt [9,157], d.h. Variationen oder Schwankungen, die keiner zellulären Kontrolle unterliegen. Eine Erscheinung, die auch in anderen genregulatorischen Prozessen zu finden ist [158-160]. Durch Konformationsänderungen der spleißosomalen Kernkomponenten oder Konzentrationsschwankungen von regulatorischen Spleißfaktoren verursacht, kann sich das Spleißosom von der bevorzugten zu einer benachbarten Spleißstelle bewegen und diese nicht deterministisch, sondern stochastisch auswählen [9]. Es wird angenommen, dass diese „Ungenauigkeit" des Spleißosoms in der Zelle größtenteils als Rauschen toleriert wird [9,157], weil es weder vorteilhaft noch schädlich und somit irrelevant für biologische Funktionen ist. Unabhängig von dieser Wertung, ist die Wahlmöglichkeit des Spleißosoms zwischen konkurrierenden Spleißstellen von Natur aus ein stochastischer Prozess [9,161]. Über eine rein stochastisch gewählte Spleißstelle hinaus können Faktoren die zufällige Wahl in Richtung einer deterministischen Wahl verschieben. Die Möglichkeit, zwischen zwei alternativen Spleißstellen zu wählen, wird damit immer geringer, bis zur rein deterministischen Nutzung von nur einer Spleißstelle im Fall des konstitutiven Spleißens. Von überwiegend stochastischen Spleißereignissen könnte man also erwarten, dass z.B. in verschiedenen Geweben ähnliche Spleißvarianten-Verhältnisse erzeugt werden, wie in einigen Studien beobachtet wurde [13,25,62,67,68,70,73,78,131]. Dagegen können gewebespezifische Unterschiede von *trans*-agierenden Spleißfaktoren die Spleißvarianten-Verhältnisse stochastisch gewählter Spleißstellen verschieben. Die AS regulierenden Spleißfaktoren der SR-Proteine oder der heterogenen nukleären Ribonukleoproteine (hnRNP) [162] sind Aspiranten, die Wahl von Tandem-Spleißstellen zu regulieren [25,163], was für die TG/AG-Tandem-Spleißstellen des humanen *GNAS*-Gens gezeigt wurde [107]. Ein Einfluss der Spleißfaktoren hSlu7 [164,165] und SPF45 [145,164] auf die Wahl von AG/AG-Tandem-Spleißstellen ist außerdem bekannt. Genaue Mechanismen sind bis dato jedoch unklar. Das *GNAS*-Gen zeigt außerdem, dass eine Regulation auch auf einer anderen Ebene stattfinden kann: Das Vorhandensein des stromaufwärts gelegenen Exons verschiebt die Wahl der TG/AG-Tandem-Spleißstellen - ein Befund der in den EST-Daten und in einer anderen Studie erkennbar ist [125] und von mir experimentell in Mausfibroblasten (NIH3T3) reproduziert wurde (Daten nicht publiziert). In diesem Beispiel ist die Wahl der 3'-Spleißstellen kontextabhängig und zeigt, dass nicht nur die unmittelbare Sequenzumgebung der Spleißstelle ausschlaggebend ist. Hinter den Mechanismen der Tandem-Spleißstellen-Wahl verbirgt sich also weitaus mehr, als nur ein

einfaches, auf dem unmittelbaren Sequenzkontext der Spleißstelle basierendes Modell für eine spleißosomale Ungenauigkeit [9,157]. Schließlich zeigt auch eine erst kürzlich publizierte Studie, dass das Spleißosom mit erstaunlich hoher Genauigkeit konstitutive Spleißstellen erkennt, aber auch alle anderen möglichen Kombinationen der konstitutiven Spleißorte mit geringer Frequenz prozessiert [166]. In Bezug auf die biologische Relevanz der Mehrzahl alternativer mRNA-Isoformen wird in dieser Studie geschlussfolgert, dass der hohe Grad an AS in höheren Eukaryoten nicht auf einer intrinsischen Ungenauigkeit des Spleißosoms basiert. Vielmehr sind im Zuge der Evolution gezielt Spleißstellen mit einem schwachen Bindungs-Potential für spleißosomale Kernkomponenten oder Spleißregulatoren entstanden, um einen hohen Grad an AS und somit eine Erhöhung des kodierenden Potentials des Genoms zu ermöglichen.

Letztendlich schließt stochastisches und keiner Regulation unterworfenes AS keineswegs funktionelle Bedeutung aus [157]. Eine nur auf den Kernkomponenten des Spleißosoms basierende stochastische Wahl der Spleißstellen würde vor allem in solchen Situationen vorteilhaft sein, in denen beide Proteinisoformen ubiquitär benötigt werden, da beide Varianten nahezu unabhängig von anderen Bedingungen, die AS regulieren, produziert werden können [25]. Dies könnte z.B. bei den Tandem-Spleißstellen der Transkriptionsfaktor-kodierenden Gene *PAX3* und *PAX7* im Menschen der Fall sein [78,167], deren Proteinisoformen in konstantem Verhältnis gebildet werden und wichtige funktionelle Rollen spielen. Ob konstante Spleißvariantenverhältnisse durch stochastische Ereignisse oder durch spezifische Mechanismen aufrechterhalten werden, ist unklar. Im diesem Zusammenhang ist erwähnenswert, dass in *WT1* eine Pyrimidin-reiche intronische „Enhancer“-Sequenz identifiziert wurde, die die Wahl der 5’-Tandem-Spleißstelle beeinflusst [168]. Es wird vermutet, dass der Spleißfaktor TIA-1 an dieses *cis*-Element bindet und für die Aufrechterhaltung des konstanten Spleißvarianten-Verhältnisses verantwortlich ist. Stochastisches Spleißen spielt darüber hinaus auch für die 48 „mutually exclusive“ Exons (sich ausschließende Exons) des *Dscam*-Gens in *D. melanogaster* eine entscheidende Rolle, und ist für die Leitung und Zielfindung der Axone („axon guidance“) während der Entwicklung des Nervensystems essentiell [20,168]. Es ist allgemein bekannt, dass im Rahmen vieler biologischer Prozesse stochastische Ereignisse funktionelle Relevanz erlangt haben [158], ein Phänomen, dass auch als „kultiviertes Rauschen“ bezeichnet wird [169]. Funktionell wichtige Spleißvarianten, die durch subtiles AS an Tandem-Spleißstellen in konstantem Verhältnis entstanden sind, könnten also auch eine Form dieses „kultivierten Rauschens“ darstellen [25].

Gerade im Fall der TG/AG-Tandem-Spleißstellen kann gemutmaßt werden, dass diese Form des AS aus einer Ungenauigkeit des Spleißvorgangs resultiert, wenn man bedenkt, dass diese im Vergleich zu AG/AG-Tandem-Spleißstellen extrem selten vorkommen. Unsere Befunde befürworten jedoch, dass diese Spleißstellen eine funktionelle Rolle spielen: Alternative TG-3'-Spleißstellen und deren Sequenzkontext werden evolutionär aufrechterhalten; einige TG-3'-Spleißstellen werden mit großer Häufigkeit benutzt und sind in einigen Fällen sogar die bevorzugten Spleißstellen [16]. Auf die Beteiligung regulatorischer Prozesse deutet der Einfluss der Spleißfaktoren SF2/ASF und hnRNPA1 auf die Wahl von TG- bzw. AG-3'-Spleißstelle im *GNAS*-Gen hin [107], außerdem sind funktionelle Unterschiede der durch TG-Spleißstellen vermittelten Proteinisoformen des *CACNA1A*-Gens bekannt [152]. Schließlich ist auch das *PCGF2*–Gen ein vielversprechender Kandidat für eine Analyse translationaler Effekte [16]. Für NAGNAG-Tandem-Spleißstellen gibt es gezielte experimentelle Studien, die für konkrete Fälle eine funktionelle Relevanz zeigen, [70,74,76,78,170]. Der Annahme nicht widersprechend, dass die Wahl einer Vielzahl von NAGNAG-Tandem-Spleißstellen stochastischen Prozessen unterliegt [9,157], und eine Vielzahl dieser Spleißstellen wahrscheinlich als phänotypisch neutral toleriert wird, ist mittlerweile belegt, dass ein kleine Population der Tandem-Spleißstellen durch negative Selektion aufrechterhalten wird und demnach einen vorteilhaften Phänotyp bedingt, dessen Identität es in den meisten Fällen noch zu erforschen gilt [171].

Das subtile AS an NAGNAG-Tandem-Spleißstellen in *A. thaliana* ist auch aus der Perspektive bemerkenswert, dass Befunde im Mensch in einer Pflanze reproduziert werden konnten. Trotz der großen evolutionären Distanz ist bekannt, dass das Spleißosom und die Ausstattung der Zelle mit Spleißregulatoren zwischen Pflanzen und Säugetieren wesentliche Übereinstimmungen aufweisen [139,172]. Andere Publikationen belegen, dass der letzte gemeinsame Vorfahre der Eukaryoten bereits ein komplexes Spleißosom besessen hat [173]. Da AS an NAGNAG-Tandem-Spleißstellen im Tierreich verbreitet ist und ebenfalls in Pflanzen stattfindet, kann vermutet werden, dass auch der gemeinsame Vorfahre der Eukaryoten bereits NAGNAG-Tandem-Spleißstellen besessen hat. Bislang noch unveröffentlichte Studien zeigen, dass Transkripte des Pilzes *Arthroderma benhamiae* häufig an 3'-Tandem-Spleißstellen alternativ gesplissen werden (G. Glöckner, persönliche Mitteilung), und befürworten diese Annahme. Eine Ausnahme im eukaryotischen Stammbaum stellt der Fadenwurm *C. elegans* dar, in dem zwar NAGNAG-3'-Spleißstellen zu finden sind, die den EST-Daten zufolge aber kaum alternativ gesplissen werden [80]. Überraschenderweise lassen aktuelle Vorhersagen auch kein AS an diesen Motiven erwarten (unveröffentlicht). In nahezu allen NAGNAG-Motiven, die erwartungsgemäß

alternativ gesplissen werden müssten, wird das intronisch-proximal gelegene AG als 3'-Spleißstelle benutzt [12,157]. Die Benutzung des intronisch-distal gelegenen AGs scheint aufgrund des speziellen Mechanismus der Wahl der 3'-Spleißstelle im Fadenwurm kaum möglich zu sein: Das U2AF-Dimer des *C. elegans*-Spleißosoms bindet an das strikte Motiv UUUUCAG, wobei Abweichung von diesem Konsensus dessen Bindung beeinträchtigen [174,175]. Dies könnte der Grund dafür sein, dass das alternative intronisch-distale AG nicht erkannt wird und zeigt gleichzeitig, dass eine Flexibilität der U2AF-Bindung für AS an 3'-Tandem-Spleißstellen entscheidend ist [25].

Die Mechanismen der Genregulation und das regulatorische Potential eukaryotischer Genome sind bislang bei weitem noch nicht vollständig erforscht und verstanden. Welche funktionellen Konsequenzen durch AS erzeugte Transkript- und Proteinisoformen haben, und welche Auswirkungen diese auf die Biologie der Zelle haben, ist im Fokus vieler experimenteller und bioinformatischer Studien [25]. In diesem Zusammenhang wurden Tandem-Spleißereignisse wegen ihrer subtilen Effekte zunächst nicht in Betracht gezogen. Mittlerweile gibt es viele Arbeiten, die beweisen, dass Tandem-Spleißstellen eine verbreitete Form des AS repräsentieren, die die Diversität der Proteome vieler Spezies erhöht. Funktionelle Effekte auf Proteinebene sind für einige Tandem-Spleißstellen bereits bekannt [59-67,70,74-79,152], andere Studien zeigen, dass die Proteinfunktion unbeeinflusst bleibt [73,137,176]. Das generelle Ausmaß ist jedoch noch weitgehend unbekannt. Zum Forschungsgebiet des subtilen AS konnte meine Arbeit einen Beitrag leisten, indem sie zum einen neue Erkenntnisse bezüglich der Verbreitung dieser Spleißereignisse liefert und auf experimentellen Ergebnissen basierend, mechanistische und biologische Einblicke gewährt, die zum Verständnis dieser Klasse von Spleißereignissen im Gesamtkontext der Genregulation in komplexen Organismen beitragen. Darüber hinaus wurden verschiedene grundlegende Herangehensweisen bezüglich Datenakquisition, -verarbeitung, und -interpretation zur Analytik von Spleißereignissen präsentiert, die es anderen Wissenschaftlern ermöglicht, diese für eigene Zwecke zu nutzen. Die Arbeit „Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns" [16] bietet interessante Angriffspunkte, die etablierten Spleißregeln, den spleißosomalen Komplex und dessen Regulation weiter zu hinterfragen, um in weiteren Studien Einblicke in den Mechanismus des Tandem-Spleißens zu gewähren. In einem nächsten Schritt könnte gezielt untersucht werden, inwieweit kontextabhängige Sequenzsignale zu diesem Spleißereignis beitragen bzw. welche Spleißfaktoren in diesen Vorgang involviert sind. Nicht zuletzt ist es für das Verständnis einer biologischen Funktion dieser Spleißereignisse wichtig, phänotypische Auswirkungen der subtilen Variationen auf

Transkript- und Proteinebene zu ermitteln. Welcher Anteil dieser Spleißereignisse ist am Repertoire funktionell unterschiedlicher Proteinisoformen beteiligt? Welche haben neutralen Charakter, sind vorteilhaft oder sogar schädlich? In diesem Zusammenhang ist die Analyse der konservierten TG/AG-Tandem-Spleißereignissen ein weiterer Angriffspunkt [16], bzw. bietet auch die Arbeit „Alternative Splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes“ [15] interessante Fragestellungen für weiterführende Experimente: Verleiht AS an NAGNAG-Tandem-Spleißstellen der SR-Proteine dem Spleißprozess tatsächlich eine zusätzliche Flexibilität? Werden NAGNAG-Tandem-Spleiß-stellen als Antwort auf Umwelteinflüsse konzertiert reguliert und wenn ja, welches sind molekularen Prozesse, die zum Spleißosom führen und dort die Wahl der 3’-Spleißstelle beeinflussen?

LITERATURVERZEICHNIS

1. Nurtdinov RN, Artamonova, II, Mironov AA, Gelfand MS: **Low conservation of alternative splicing patterns in the human and mouse genomes**. *Hum Mol Genet* 2003, **12**(11):1313-1320.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
4. IHGSC: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.
5. Anderson NG, Matheson A, Anderson NL: **Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology**. *Proteomics* 2001, **1**(1):3-12.
6. Consortium *Ce*S: **Genome sequence of the nematode C. elegans: a platform for investigating biology**. *Science* 1998, **282**(5396):2012-2018.
7. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**(5645):501.
8. Maniatis T, Tasic B: **Alternative pre-mRNA splicing and proteome expansion in metazoans**. *Nature* 2002, **418**(6894):236-243.
9. Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A simple physical model predicts small exon length variations**. *PLoS Genet* 2006, **2**(4):e45.
10. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site**. *RNA* 2006, **12**(12):2047-2056.
11. Ermakova EO, Nurtdinov RN, Gelfand MS: **Overlapping alternative donor splice sites in the human genome**. *J Bioinform Comput Biol* 2007, **5**(5):991-1004.
12. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity**. *Nat Genet* 2004, **36**(12):1255-1257.
13. Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Phylogenetically widespread alternative splicing at unusual GYNGYN donors**. *Genome Biol* 2006, **7**(7):R65.
14. Schindler S, Heiner M, Platzer M, Szafranski K: **Comparison of methods for splice variant quantification**. *Biotechniques* 2009(submitted).
15. Schindler S, Szafranski K, Hiller M, Ali GS, Palusa SG, Backofen R, Platzer M, Reddy AS: **Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes**. *BMC Genomics* 2008, **9**:159.
16. Schindler S, Szafranski K, Taudien S, Hiller M, Huse K, Jahn N, Schreiber S, Backofen R, Platzer M: **Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns**. *Genome Biol* 2007, **8**(8):R154.
17. Sharp PA: **The discovery of split genes and RNA splicing**. *Trends Biochem Sci* 2005, **30**(6):279-281.
18. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 2001, **17**(2):100-107.
19. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing**. *Gene* 2005, **344**:1-20.
20. Graveley BR: **Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures**. *Cell* 2005, **123**(1):65-73.
21. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.

22. Blencowe BJ: **Alternative splicing: new insights from global analyses**. *Cell* 2006, **126**(1):37-47.
23. Lareau LF, Green RE, Bhatnagar RS, Brenner SE: **The evolving roles of alternative splicing**. *Curr Opin Struct Biol* 2004, **14**(3):273-282.
24. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
25. Hiller M, Platzer M: **Widespread and subtle: alternative splicing at short-distance tandem sites**. *Trends Genet* 2008, **24**(5):246-255.
26. Burge CB, Tuschl T, Sharp PA: **The RNA World II**. *Cold Spring Harbor Laboratory Press* 1999:525–560.
27. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing**. *Nat Rev Genet* 2002, **3**(4):285-298.
28. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR: **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis**. *BMC Genomics* 2006, **7**:327.
29. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes**. *Nucleic Acids Res* 2007, **35**(1):125-131.
30. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annu Rev Biochem* 2003, **72**:291-336.
31. Garcia J, Gerber SH, Sugita S, Sudhof TC, Rizo J: **A conformational switch in the Piccolo C2A domain regulated by alternative splicing**. *Nat Struct Mol Biol* 2004, **11**(1):45-53.
32. Kamatkar S, Radha V, Nambirajan S, Reddy RS, Swarup G: **Two splice variants of a tyrosine phosphatase differ in substrate specificity, DNA binding, and subcellular location**. *J Biol Chem* 1996, **271**(43):26755-26761.
33. Rudenko G, Nguyen T, Chelliah Y, Sudhof TC, Deisenhofer J: **The structure of the ligand-binding domain of neurexin Ibeta: regulation of LNS domain function by alternative splicing**. *Cell* 1999, **99**(1):93-101.
34. Vilardell J, Chartrand P, Singer RH, Warner JR: **The odyssey of a regulated transcript**. *Rna* 2000, **6**(12):1773-1780.
35. Hilleren P, Parker R: **Mechanisms of mRNA surveillance in eukaryotes**. *Annu Rev Genet* 1999, **33**:229-260.
36. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans**. *Proc Natl Acad Sci U S A* 2003, **100**(1):189-192.
37. Hillman RT, Green RE, Brenner SE: **An unappreciated role for RNA surveillance**. *Genome Biol* 2004, **5**(2):R8.
38. Jumaa H, Nielsen PJ: **The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation**. *Embo J* 1997, **16**(16):5077-5085.
39. Lejeune F, Cavaloc Y, Stevenin J: **Alternative splicing of intron 3 of the serine/arginine-rich protein 9G8 gene. Identification of flanking exonic splicing enhancers and involvement of 9G8 as a trans-acting factor**. *J Biol Chem* 2001, **276**(11):7850-7858.
40. Sureau A, Gattoni R, Dooghe Y, Stevenin J, Soret J: **SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs**. *Embo J* 2001, **20**(7):1785-1796.
41. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements**. *Nature* 2007, **446**(7138):926-929.
42. Roberts AG, Redding SJ, Llewellyn DH: **An alternatively-spliced exon in the 5'-UTR of human ALAS1 mRNA inhibits translation and renders it resistant to haem-mediated decay**. *FEBS Lett* 2005, **579**(5):1061-1066.
43. Coutinho-Mansfield GC, Xue Y, Zhang Y, Fu XD: **PTB/nPTB switch: a post-transcriptional mechanism for programming neuronal differentiation**. *Genes Dev* 2007, **21**(13):1573-1577.
44. Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, Zeeberg BR, Kane D, Weinstein JN, Blume J, Darnell RB: **Nova regulates brain-specific splicing to shape the synapse**. *Nat Genet* 2005, **37**(8):844-852.
45. Zhang C, Zhang Z, Castle J, Sun S, Johnson J, Krainer AR, Zhang MQ: **Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2**. *Genes Dev* 2008, **22**(18):2550-2563.

46. Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, Allain FH: **Molecular basis of RNA recognition by the human alternative splicing factor Fox-1**. *Embo J* 2006, **25**(1):163-173.
47. Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL: **Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals**. *Mol Cell Biol* 2005, **25**(22):10005-10016.
48. Zhu H, Hasman RA, Barron VA, Luo G, Lou H: **A nuclear function of Hu proteins as neuron-specific alternative RNA processing regulators**. *Mol Biol Cell* 2006, **17**(12):5105-5114.
49. Sazani P, Kole R: **Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing**. *J Clin Invest* 2003, **112**(4):481-486.
50. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease**. *Biochim Biophys Acta* 2009, **1792**(1):14-26.
51. Kim E, Goren A, Ast G: **Insights into the connection between cancer and alternative splicing**. *Trends Genet* 2008, **24**(1):7-10.
52. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR: **The gene encoding the splicing factor SF2/ASF is a proto-oncogene**. *Nat Struct Mol Biol* 2007, **14**(3):185-193.
53. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease**. *Genes Dev* 2003, **17**(4):419-437.
54. Wirth B, Brichta L, Hahnen E: **Spinal muscular atrophy and therapeutic prospects**. *Prog Mol Subcell Biol* 2006, **44**:109-132.
55. Zhang Z, Lotti F, Dittmar K, Younis I, Wan L, Kasim M, Dreyfuss G: **SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing**. *Cell* 2008, **133**(4):585-600.
56. Sugnet CW, Kent WJ, Ares M, Jr., Haussler D: **Transcriptome and genome conservation of alternative splicing events in humans and mice**. *Pac Symp Biocomput* 2004:66-77.
57. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome**. *Genome Res* 2003, **13**(6B):1290-1300.
58. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C: **Assessing the impact of alternative splicing on domain interactions in the human proteome**. *J Proteome Res* 2004, **3**(1):76-83.
59. Rivers C, Levy A, Hancock J, Lightman S, Norman M: **Insertion of an amino acid in the DNA-binding domain of the glucocorticoid receptor as a result of alternative splicing**. *J Clin Endocrinol Metab* 1999, **84**(11):4283-4286.
60. Hu CA, Lin WW, Obie C, Valle D: **Molecular enzymology of mammalian Delta1-pyrroline-5-carboxylate synthase. Alternative splice donor utilization generates isoforms with different sensitivity to ornithine inhibition**. *J Biol Chem* 1999, **274**(10):6754-6762.
61. Yan M, Wang LC, Hymowitz SG, Schilbach S, Lee J, Goddard A, de Vos AM, Gao WQ, Dixit VM: **Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors**. *Science* 2000, **290**(5491):523-527.
62. Burgar HR, Burns HD, Elsden JL, Lalioti MD, Heath JK: **Association of the signaling adaptor FRS2 with fibroblast growth factor receptor 1 (Fgfr1) is mediated by alternative splicing of the juxtamembrane domain**. *J Biol Chem* 2002, **277**(6):4018-4023.
63. Wagner KD, Wagner N, Schedl A: **The complex life of WT1**. *J Cell Sci* 2003, **116**(Pt 9):1653-1658.
64. Takeda H, Matsuzaki T, Oki T, Miyagawa T, Amanuma H: **A novel POU domain gene, zebrafish pou2: expression and roles of two alternatively spliced twin products in early development**. *Genes Dev* 1994, **8**(1):45-59.
65. Treacy MN, Neilson LI, Turner EE, He X, Rosenfeld MG: **Twin of I-POU: a two amino acid difference in the I-POU homeodomain distinguishes an activator from an inhibitor of transcription**. *Cell* 1992, **68**(3):491-505.
66. Koenig Merediz SA, Schmidt M, Hoppe GJ, Alfken J, Meraro D, Levi BZ, Neubauer A, Wittig B: **Cloning of an interferon regulatory factor 2 isoform with different regulatory ability**. *Nucleic Acids Res* 2000, **28**(21):4219-4224.
67. Unoki M, Shen JC, Zheng ZM, Harris CC: **Novel splice variants of ING4 and their possible roles in the regulation of cell growth and motility**. *J Biol Chem* 2006, **281**(45):34677-34686.
68. Hammes A, Guo JK, Lutsch G, Leheste JR, Landrock D, Ziegler U, Gubler MC, Schedl A: **Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation**. *Cell* 2001, **106**(3):319-329.

69. Barbaux S, Niaudet P, Gubler MC, Grunfeld JP, Jaubert F, Kuttenn F, Fekete CN, Souleyreau-Therville N, Thibaud E, Fellous M, McElreavey K: **Donor splice-site mutations in WT1 are responsible for Frasier syndrome**. *Nat Genet* 1997, **17**(4):467-470.
70. Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, Toyoda M, Ozaki M, Ono M, Miki N, Miyashita T, Yamada M: **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* 2005, **50**(8):382-394.
71. Ferranti P, Lilla S, Chianese L, Addeo F: **Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein**. *J Protein Chem* 1999, **18**(5):595-602.
72. Rogina B, Upholt WB: **The chicken homeobox gene Hoxd-11 encodes two alternatively spliced RNA species**. *Biochem Mol Biol Int* 1995, **35**(4):825-831.
73. Li L, Howe GA: **Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway**. *Plant Mol Biol* 2001, **46**(4):409-419.
74. Condorelli G, Bueno R, Smith RJ: **Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics**. *J Biol Chem* 1994, **269**(11):8510-8516.
75. Joyce-Brady M, Jean JC, Hughey RP: **gamma -glutamyltransferase and its isoform mediate an endoplasmic reticulum stress response**. *J Biol Chem* 2001, **276**(12):9468-9477.
76. Lorkovic ZJ, Lehner R, Forstner C, Barta A: **Evolutionary conservation of minor U12-type spliceosome between plants and humans**. *RNA* 2005, **11**(7):1095-1107.
77. Makielski JC, Ye B, Valdivia CR, Pagel MD, Pu J, Tester DJ, Ackerman MJ: **A ubiquitous splice variant and a common polymorphism affect heterologous expression of recombinant human SCN5A heart sodium channels**. *Circ Res* 2003, **93**(9):821-828.
78. Vogan KJ, Underhill DA, Gros P: **An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity**. *Mol Cell Biol* 1996, **16**(12):6677-6686.
79. Caffrey JJ, Safrany ST, Yang X, Shears SB: **Discovery of molecular and catalytic diversity among human diphosphoinositol-polyphosphate phosphohydrolases. An expanding Nudt family**. *J Biol Chem* 2000, **275**(17):12730-12736.
80. Hiller M, Nikolajewa S, Huse K, Szafranski K, Rosenstiel P, Schuster S, Backofen R, Platzer M: **TassDB: a database of alternative tandem splice sites**. *Nucleic Acids Res* 2007, **35**(Database issue):D188-192.
81. AGI: **Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.
82. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.
83. Fu XD: **The superfamily of arginine/serine-rich splicing factors**. *RNA* 1995, **1**(7):663-680.
84. Graveley BR: **Sorting out the complexity of SR protein functions**. *RNA* 2000, **6**(9):1197-1211.
85. Schaal TD, Maniatis T: **Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences**. *Mol Cell Biol* 1999, **19**(3):1705-1719.
86. Manley JL, Tacke R: **SR proteins and splicing control**. *Genes Dev* 1996, **10**:1569-1579.
87. Blencowe BJ: **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases**. *Trends Biochem Sci* 2000, **25**(3):106-110.
88. Eldridge AG, Li Y, Sharp PA, Blencowe BJ: **The SRm160/300 splicing coactivator is required for exon-enhancer function**. *Proc Natl Acad Sci U S A* 1999, **96**(11):6125-6130.
89. Kan JL, Green MR: **Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor**. *Genes Dev* 1999, **13**(4):462-471.
90. Kalyna M, Barta A: **A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions?** *Biochem Soc Trans* 2004, **32**(Pt 4):561-564.
91. Palusa SG, Ali GS, Redddy ASN: **Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins and its regulation by hormones and stresses**. *Plant J,* 2007, **49**:1091-1107.

92. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes**. *Genome Res* 1999, **9**(12):1288-1293.
93. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs**. *Genome Res* 2002, **12**(12):1837-1845.
94. Thanaraj TA: **A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures**. *Nucleic Acids Res* 1999, **27**(13):2627-2637.
95. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes**. *Nucleic Acids Res* 2001, **29**(13):2850-2859.
96. Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human**. *Hum Mol Genet* 2002, **11**(4):451-464.
97. Jackson IJ: **A reappraisal of non-consensus mRNA splice sites**. *Nucleic Acids Res* 1991, **19**(14):3795-3798.
98. Levine A, Durbin R: **A computational scan for U12-dependent introns in the human genome sequence**. *Nucleic Acids Res* 2001, **29**(19):4006-4013.
99. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R: **Comprehensive splice-site analysis using comparative genomics**. *Nucleic Acids Res* 2006, **34**(14):3955-3967.
100. Burset M, Seledtsov IA, Solovyev VV: **Analysis of canonical and non-canonical splice sites in mammalian genomes**. *Nucleic Acids Res* 2000, **28**(21):4364-4375.
101. Brackenridge S, Wilkie AO, Screaton GR: **Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes**. *Embo J* 2003, **22**(7):1620-1631.
102. Dietrich RC, Incorvaia R, Padgett RA: **Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns**. *Mol Cell* 1997, **1**(1):151-160.
103. Wu Q, Krainer AR: **Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs**. *Rna* 1997, **3**(6):586-601.
104. Chong A, Zhang G, Bajic VB: **Information for the Coordinates of Exons (ICE): a human splice sites database**. *Genomics* 2004, **84**(4):762-766.
105. van Nimwegen E, Paul N, Sheridan R, Zavolan M: **SPA: a probabilistic algorithm for spliced alignment**. *PLoS Genet* 2006, **2**(4):e24.
106. Quan F, Forte MA: **Two forms of Drosophila melanogaster Gs alpha are produced by alternate splicing involving an unusual splice site**. *Mol Cell Biol* 1990, **10**(3):910-917.
107. Pollard AJ, Krainer AR, Robson SC, Europe-Finner GN: **Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice Site**. *J Biol Chem* 2002, **277**(18):15241-15251.
108. Stathakis DG, Udar N, Sandgren O, Andreasson S, Bryant PJ, Small K, Forsman-Semb K: **Genomic organization of human DLG4, the gene encoding postsynaptic density 95**. *J Neurochem* 1999, **73**(6):2250-2265.
109. Seeman P, Nam D, Ulpian C, Liu IS, Tallerico T: **New dopamine receptor, D2(Longer), with unique TG splice site, in human brain**. *Brain Res Mol Brain Res* 2000, **76**(1):132-141.
110. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P: **Real-time DNA sequencing using detection of pyrophosphate release**. *Anal Biochem* 1996, **242**(1):84-89.
111. Diggle MA, Clarke SC: **Pyrosequencing: sequence typing at the speed of light**. *Mol Biotechnol* 2004, **28**(2):129-137.
112. Gruber JD, Colligan PB, Wolford JK: **Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing**. *Hum Genet* 2002, **110**(5):395-401.
113. Langaee T, Ronaghi M: **Genetic variation analyses by Pyrosequencing**. *Mutat Res* 2005, **573**(1-2):96-102.
114. Ronaghi M, Elahi E: **Pyrosequencing for microbial typing**. *J Chromatogr B Analyt Technol Biomed Life Sci* 2002, **782**(1-2):67-72.
115. Butler JM, Buel E, Crivellente F, McCord BR: **Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis**. *Electrophoresis* 2004, **25**(10-11):1397-1412.
116. Wenz H, Robertson JM, Menchen S, Oaks F, Demorest DM, Scheibler D, Rosenblum BB, Wike C, Gilbert DA, Efcavitch JW: **High-precision genotyping by denaturing capillary electrophoresis**. *Genome Res* 1998, **8**(1):69-80.

117. Cooper TA: **Use of minigene systems to dissect alternative splicing elements**. *Methods* 2005, **37**(4):331-340.
118. Furey TS, Diekhans M, Lu Y, Graves TA, Oddy L, Randall-Maher J, Hillier LW, Wilson RK, Haussler D: **Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing**. *Genome Res* 2004, **14**(10B):2034-2040.
119. Tsai KW, Tarn WY, Lin WC: **Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3' tandem splice site selection**. *Mol Cell Biol* 2007, **27**(16):5835-5848.
120. Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, Roberts RG, Mathew CG, Platzer M, Schreiber S: **Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression**. *BMC Genomics* 2007, **8**:472.
121. Rosenstiel P, Huse K, Till A, Hampe J, Hellmig S, Sina C, Billmann S, von Kampen O, Waetzig GH, Platzer M, Seegert D, Schreiber S: **A short isoform of NOD2/CARD15, NOD2-S, is an endogenous inhibitor of NOD2/receptor-interacting protein kinase 2-induced signaling pathways**. *Proc Natl Acad Sci U S A* 2006, **103**(9):3280-3285.
122. Maderazo AB, Belk JP, He F, Jacobson A: **Nonsense-containing mRNAs that accumulate in the absence of a functional nonsense-mediated mRNA decay pathway are destabilized rapidly upon its restitution**. *Mol Cell Biol* 2003, **23**(3):842-851.
123. Sambrook J, Russel D: **Molecular cloning: A laboratory manual.** *New York: Cold Spring Harboor Laboratory Press* 2000.
124. Atkinson TP, Dai Y: **Activation-induced changes in alternate splice acceptor site usage**. *Biochem Biophys Res Commun* 2007, **358**(2):590-595.
125. Frey UH, Nuckel H, Dobrev D, Manthey I, Sandalcioglu IE, Eisenhardt A, Worm K, Hauner H, Siffert W: **Quantification of G protein Gaalphas subunit splice variants in different human tissues and cells using pyrosequencing**. *Gene Expr* 2005, **12**(2):69-81.
126. Vandenbroucke, II, Vandesompele J, Paepe AD, Messiaen L: **Quantification of splice variants using real-time PCR**. *Nucleic Acids Res* 2001, **29**(13):E68-68.
127. Gharizadeh B, Akhras M, Nourizad N, Ghaderi M, Yasuda K, Nyren P, Pourmand N: **Methodological improvements of pyrosequencing technology**. *J Biotechnol* 2006, **124**(3):504-511.
128. Schochetman G, Ou CY, Jones WK: **Polymerase chain reaction**. *J Infect Dis* 1988, **158**(6):1154-1157.
129. von Wintzingerode F, Gobel UB, Stackebrandt E: **Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis**. *FEMS Microbiol Rev* 1997, **21**(3):213-229.
130. Kanagawa T: **Bias and artifacts in multitemplate polymerase chain reactions (PCR)**. *J Biosci Bioeng* 2003, **96**(4):317-323.
131. Tsai KW, Lin WC: **Quantitative analysis of wobble splicing indicates that it is not tissue specific**. *Genomics* 2006, **88**(6):855-864.
132. Simpson CG, Fuller J, Maronova M, Kalyna M, Davidson D, McNicol J, Barta A, Brown JW: **Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts**. *Plant J* 2008, **53**(6):1035-1048.
133. Lawyer FC, Stoffel S, Saiki RK, Chang SY, Landre PA, Abramson RD, Gelfand DH: **High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity**. *PCR Methods Appl* 1993, **2**(4):275-287.
134. Nembaware V, Wolfe KH, Bettoni F, Kelso J, Seoighe C: **Allele-specific transcript isoforms in human**. *FEBS Lett* 2004, **577**(1-2):233-238.
135. Davies RC, Bratt E, Hastie ND: **Did nucleotides or amino acids drive evolutionary conservation of the WT1 +/-KTS alternative splice?** *Hum Mol Genet* 2000, **9**(8):1177-1183.
136. Perner B, Englert C, Bollig F: **The Wilms tumor genes wt1a and wt1b control different steps during formation of the zebrafish pronephros**. *Dev Biol* 2007, **309**(1):87-96.
137. Xu Q, Belcastro MP, Villa ST, Dinkins RD, Clarke SG, Downie AB: **A second protein L-isoaspartyl methyltransferase gene in Arabidopsis produces two transcripts whose products are sequestered in the nucleus**. *Plant Physiol* 2004, **136**(1):2652-2664.
138. Iida K, Shionyu M, Suso Y: **Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals**. *Mol Biol Evol* 2008, **25**(4):709-718.
139. Reddy AS, N.: **Alternative splicing of pre-messenger RNAs in plants in the genomic era**. *Annu Rev Plant Biol* 2007, **58**:267-294.

140. Ali GS, Reddy AS: **ATP, phosphorylation and transcription regulate the mobility of plant splicing factors**. *J Cell Sci* 2006, **119**(Pt 17):3527-3538.
141. Kol G, Lev-Maor G, Ast G: **Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation**. *Hum Mol Genet* 2005, **14**(11):1559-1568.
142. Soares LM, Zanier K, Mackereth C, Sattler M, Valcarcel J: **Intron removal requires proofreading of U2AF/3' splice site recognition by DEK**. *Science* 2006, **312**(5782):1961-1965.
143. Bennett M, Michaud S, Kingston J, Reed R: **Protein components specifically associated with prespliceosome and spliceosome complexes**. *Genes Dev* 1992, **6**(10):1986-2000.
144. Chiara MD, Palandjian L, Feld Kramer R, Reed R: **Evidence that U5 snRNP recognizes the 3' splice site for catalytic step II in mammals**. *Embo J* 1997, **16**(15):4746-4759.
145. Lallena MJ, Chalmers KJ, Llamazares S, Lamond AI, Valcarcel J: **Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45**. *Cell* 2002, **109**(3):285-296.
146. Zhuang Y, Weiner AM: **The conserved dinucleotide AG of the 3' splice site may be recognized twice during in vitro splicing of mammalian mRNA precursors**. *Gene* 1990, **90**(2):263-269.
147. Krainer A.R., Reed R., T. M: **Mechanisms of human-globinpre-mRNA splicing** *In Genetic Chemistry: the Molecular Basis of Heredity Edited by: Berg P Houston, TX: The Robert A Welch Foundation* 1985, **Conferences on Chemical Research**:353-382.
148. Kozasa T, Itoh H, Tsukamoto T, Kaziro Y: **Isolation and characterization of the human Gs alpha gene**. *Proc Natl Acad Sci U S A* 1988, **85**(7):2081-2085.
149. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR: **ESEfinder: A web resource to identify exonic splicing enhancers**. *Nucleic Acids Res* 2003, **31**(13):3568-3571.
150. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G: **Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers**. *Mol Cell* 2006, **22**(6):769-781.
151. Gomez-Skarmeta JL, Rodriguez I, Martinez C, Culi J, Ferres-Marco D, Beamonte D, Modolell J: **Cis-regulation of achaete and scute: shared enhancer-like elements drive their coexpression in proneural clusters of the imaginal discs**. *Genes Dev* 1995, **9**(15):1869-1882.
152. Bourinet E, Soong TW, Sutton K, Slaymaker S, Mathews E, Monteil A, Zamponi GW, Nargeot J, Snutch TP: **Splicing of alpha 1A subunit gene generates phenotypic variants of P- and Q-type calcium channels**. *Nat Neurosci* 1999, **2**(5):407-415.
153. Kozak M: **Regulation of translation via mRNA structure in prokaryotes and eukaryotes**. *Gene* 2005, **361**:13-37.
154. Meijer HA, Thomas AA: **Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA**. *Biochem J* 2002, **367**(Pt 1):1-11.
155. Patel AA, McCarthy M, Steitz JA: **The splicing of U12-type introns can be a rate-limiting step in gene expression**. *Embo J* 2002, **21**(14):3804-3815.
156. Sorek R, Shamir R, Ast G: **How prevalent is functional alternative splicing in the human genome?** *Trends Genet* 2004, **20**(2):68-71.
157. Hiller M, Szafranski K, Backofen R, Platzer M: **Alternative splicing at NAGNAG acceptors: simply noise or noise and more?** *PLoS Genet* 2006, **2**(11):e207; author reply e208.
158. Fedoroff N, Fontana W: **Genetic networks. Small numbers of big molecules**. *Science* 2002, **297**(5584):1129-1131.
159. McAdams HH, Arkin A: **It's a noisy business! Genetic regulation at the nanomolar scale.** *Trends Genet* 1999, **15**(2):65-69.
160. Thattai M, van Oudenaarden A: **Intrinsic noise in gene regulatory networks**. *Proc Natl Acad Sci U S A* 2001, **98**(15):8614-8619.
161. Smith CW, Chu TT, Nadal-Ginard B: **Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns**. *Mol Cell Biol* 1993, **13**(8):4939-4952.
162. Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code**. *Nat Rev Mol Cell Biol* 2005, **6**(5):386-398.
163. Bai Y, Lee D, Yu T, Chasin LA: **Control of 3' splice site choice in vivo by ASF/SF2 and hnRNP A1**. *Nucleic Acids Res* 1999, **27**(4):1126-1134.
164. Chua K, Reed R: **The RNA splicing factor hSlu7 is required for correct 3' splice-site choice**. *Nature* 1999, **402**(6758):207-210.

165. Lev-Maor G, Sorek R, Shomron N, Ast G: **The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons**. *Science* 2003, **300**(5623):1288-1291.
166. Fox-Walsh KL, Hertel KJ: **Splice-site pairing is an intrinsically high fidelity process**. *Proc Natl Acad Sci U S A* 2009.
167. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection**. *Rna* 2008, **14**(4):616-629.
168. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL: **Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity**. *Cell* 2000, **101**(6):671-684.
169. Rao CV, Wolf DM, Arkin AP: **Control, exploitation and tolerance of intracellular noise**. *Nature* 2002, **420**(6912):231-237.
170. Karinch AM, deMello DE, Floros J: **Effect of genotype on the levels of surfactant protein A mRNA and on the SP-A2 splice variants in adult humans**. *Biochem J* 1997, **321 (Pt 1)**:39-47.
171. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection**. *RNA* 2008(in press).
172. Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W: **Pre-mRNA splicing in higher plants**. *Trends Plant Sci* 2000, **5**(4):160-167.
173. Collins L, Penny D: **Complex spliceosomal organization ancestral to extant eukaryotes**. *Mol Biol Evol* 2005, **22**(4):1053-1066.
174. Hollins C, Zorio DA, MacMorris M, Blumenthal T: **U2AF binding selects for the high conservation of the C. elegans 3' splice site**. *Rna* 2005, **11**(3):248-253.
175. Zhang H, Blumenthal T: **Functional analysis of an intron 3' splice site in Caenorhabditis elegans**. *RNA* 1996, **2**(4):380-388.
176. Hosoda H, Kojima M, Matsuo H, Kangawa K: **Purification and characterization of rat des-Gln14-Ghrelin, a second endogenous ligand for the growth hormone secretagogue receptor**. *J Biol Chem* 2000, **275**(29):21995-22000.

EHRENWÖRTLICHE ERKLÄRUNG

Hiermit erkläre ich, dass mir die Promotionsordnung der Friedrich-Schiller-Universität Jena bekannt ist. Ich versichere, dass die eingereichte Dissertation selbständig und ohne die unzulässige Hilfe Dritter sowie nur unter der Verwendung der angegebenen Hilfsmittel und Literatur erstellt wurde. Weiterhin erkläre ich, dass die in der Arbeit enthaltenen Daten nur die originalen Daten enthalten und in keinem Fall inhaltsverändernder Bearbeitung unterzogen wurden. Ich habe nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen. Ich versichere, dass diese Dissertation nicht für eine staatliche oder andere wissenschaftliche Prüfung und nicht bei einer anderen Hochschule als Dissertation eingereicht wurde.

……

Stefanie Schindler