Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing

Dissertation zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Abdou Gomaa Abdou ElSharawy

B.Sc., M.Sc.

Kiel, November 2008

Referent:…….. **Prof. Dr. Frank Kempken**

Koreferent:…….. **Prof. Dr. Stefan Schreiber**

Tag der mündlichen Prüfung: Kiel, den 11.11.08……...

Zum Druck genehmigt: Kiel, den ..…….

.....

Der Dekan

To

my **parents**

my **wife** and

my sons **Ahmed** & **Amr**

TABLE OF CONTENTS

[*Footnote*: for simplicity, I use in text the nucleotide symbol T to denote thymidine in DNA as well as uridine in RNA sequence].

1 INTRODUCTION

1.1. Single nucleotide polymorphisms: Biology and functional relevance

The Human Genome Project and the many population-based scientific projects that followed have provided valuable resources for a better understanding of the evolutionary and biomedical importance of human genetic variation. For instance, chimpanzee and humans share 99% of their genomes (Chen *et al.*, 2001; Ast, 2005) and there is only an 0.1% difference between two individual humans. Single nucleotide polymorphisms (SNPs), as the most abundant form of genetic variation, are mostly biallelic and therefore easy to assay once they are described. Given their abundance in the human genome (approximately one SNP every 100-300 bp (Sachidanandam *et al.*, 2001; Ke *et al.*, 2008)) and their ease of high-throughput typing, SNPs progressively replace microsatellites as first-choice genetic markers in association and linkage studies (Hiller *et al.*, 2006b; Reumers *et al.*, 2008). Although the majority of these variations probably result in neutral phenotypic outcomes, i.e. functionally silent (Teufel *et al.*, 2006), certain SNPs contribute significantly to phenotypic individuality, disease susceptibility, as well as to drug treatments (Wangkumhang *et al.*, 2007). Nevertheless, the 'neutral' SNPs can serve either as genetic markers or tagging SNPs.

In the current release of dbSNP database, more than 12 million germline genetic variants have been recorded and the advent of next generation sequencing technologies will likely lead to a complete assessment of the inventory of human genetic polymorphisms in the foreseeable future. For this information to become fully useful, however, a functional annotation of the known DNA sequence variations will also be required. Although much interest focuses on coding SNPs (cSNP), since those SNPs impair the normal sequence and function of proteins and can be readily interpreted, SNPs can also influence pre-mRNA splicing (ElSharawy *et al.*, 2006), which usually have a more dramatic effect on the resulting protein than the alteration of a single codon. Splicing mutations have been suspected to be the most frequent cause of hereditary diseases (Lopez-Bigas *et al.*, 2005). Although relatively hard to interpret, non-coding promoter SNPs may also disrupt functional sites on the transcriptional level (Reumers *et al.*, 2008). Thus, the identification of functional SNPs and their roles promises to provide important information not only for biochemical studies, but also for the study of other phenotypes of high relevance (Cavalli-Sforza, 2005; Teufel *et al.*, 2006).

1.2. Pre-mRNA splicing: Mechanism and challenges

Pre-mRNA splicing is an essential and a critical step in eukaryotic gene expression. Despite their relative large sizes, introns are co-transcriptionally removed by splicing with great accuracy and fidelity, although contrary to our expectations, currently known signals required for pre-mRNA processing are very degenerate and redundant (Soller, 2006). Therefore, the initial fundamental step in metazoan pre-mRNA splicing is the identification of exon-intron boundaries by direct interactions between the basal splicing machinery, the spliceosome, and pre-mRNA signature elements. In general, an acceptor splice site (ss) has a highly conserved AG dinucleotide, a preceding polypyrimidine tract and a branch point 'A', whereas, the donor ss has a highly conserved GT and an extended intronic consensus sequence (Zhuang and Weiner, 1986; Wu *et al.*, 1999). The basic biochemical mechanism of pre-mRNA splicing (Figure 1.1) mainly depends upon these canonical ss signals.

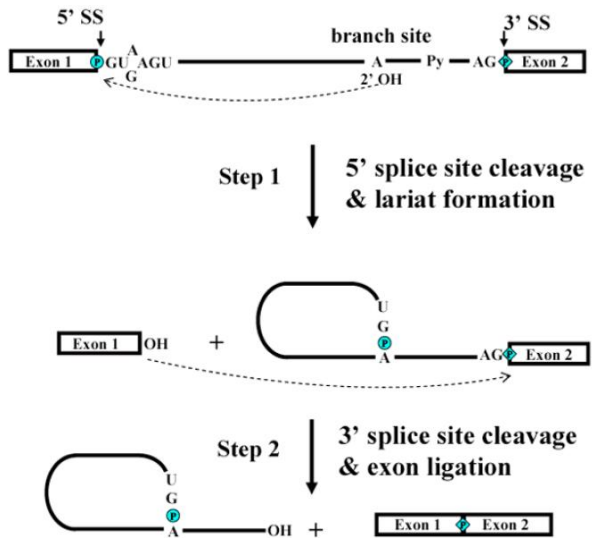

Figure 1.1 The biochemical mechanism of pre-mRNA splicing.
Pre-mRNA splicing occurs via a two-step transesterification mechanism, which ends with the ligation of the flanking exons and releases the intron in the form of a lariat (Pagani and Baralle, 2004). The phosphodiester linkages are indicated by the letter p inside a circle or a diamond. In the first step, the 2′-hydroxyl group of the A residue at the branch site attacks the phosphate at the GU 5′-ss. This leads to cleavage of the 5′ exon from the intron and the formation of a lariat intermediate. In the following transesterification reaction, which involves the phosphate (p) at the 3′ end of the intron and the 3′-hydroxyl group of the detached exon, ligates the two exons. This reaction releases the intron, still in the form of a lariat (Pagani and Baralle, 2004). *Illustration from (Mordes et al., 2006)*.

In fact, the pre-mRNA splicing process is far more complicated as given in Figure 1.1. It is now clear that exon recognition is accomplished by the accumulated recognition of multiple weak signals, resulting in a network of interactions across exons as well as across introns (Faustino and Cooper, 2003). After initial ss recognition and pairing (Reed, 1996; Lim and Hertel, 2004), the catalytic components of the spliceosome are activated through extensive structural rearrangements, ultimately resulting in intron removal (Staley and Guthrie, 1998; Hertel, 2008). The building blocks of the spliceosome are uridine-rich small nuclear RNAs (U snRNAs) packaged as ribonucleoprotein particles (snRNPs) that function in conjunction with

over 300 distinct non-snRNP auxiliary proteins (Jurica and Moore, 2003; Chen *et al.*, 2007). The major U2-type spliceosome, which consists of U1, U2, U4, U5, and U6 snRNPs, catalyzes the removal of introns with canonical (GT-AG) ss. The minor U12-type spliceosome that contains U11, U12, U4atac, U5, and U6atac snRNPs recognizes a small percentage of introns (<1% in *Arabidopsis* and humans) with noncanonical ss (Reddy, 2007).

In fact, the core splicing signals lack sufficient information content for the splicing machinery to distinguish correct pairs of ss from cryptic ss, which are vastly more abundant than correct ss (Senapathy *et al.*, 1990; Sun and Chasin, 2000). In this regard, additional *cis*-acting sequences are vitally required (Cartegni *et al.*, 2002; Matlin *et al.*, 2005). The final signal of the interactions between these various *cis*-acting layers results in guiding the spliceosome to the correct nucleotides (nt) for exon joining and intron removal. These elements make up what is now recognized as a 'cellular splicing code', which appears to be particularly dense within and around exons (Wang and Cooper, 2007) (Figure 1.2). The first layer of the 'splicing code' consists of consensus ss sequences positioned at exon-intron boundaries that are essential for the splicing of all exons. It is this RNA-RNA base-pairing that specifies which nucleotides are involved in the precise cut-and-paste reactions that join exons. Consequently, mutations in the pre-mRNA that disrupt this base pairing decrease the efficiency of exon recognition. A second layer of information is an extensive and complex array of diverse intronic and exonic splicing enhancer (ISE and ESE) and suppressor (ESS and ISS) elements, which direct the spliceosome to the appropriate sites and inhibit use of potential cryptic ss (Wang and Cooper, 2007). ESEs promote splicing by binding to the SR protein family, whereas ESSs and ISSs repress splicing by binding to heterogeneous nuclear ribonucleoproteins (hnRNPs) (Cartegni *et al.*, 2002). Enhancers and silencers tend to be short (~5-10 nt), degenerate consensus sequences (Matlin *et al.*, 2005) and working in a context-dependent manner (Pagani *et al.*, 2003). Interestingly, the position of a splicing-factor binding site relative to the exon can determine whether they act positively or negatively (Kanopka *et al.*, 1996; Ule *et al.*, 2006). The role of ESE-bound SR proteins in ensuring the correct linear order of exons in mature mRNA has also been reported (Ibrahim el *et al.*, 2005). To meet the physiological requirements of cells and tissues, most human genes are differentially spliced (Johnson *et al.*, 2003) enabling proteomic diversity, which indeed adds another challenge to the spliceosome to appropriately regulate this more complex process in a comprehensive manner. Again, these challenges are met through several intercommunications between layers of *cis*-acting elements (Wang and Cooper, 2007) (details in next section 1.3).

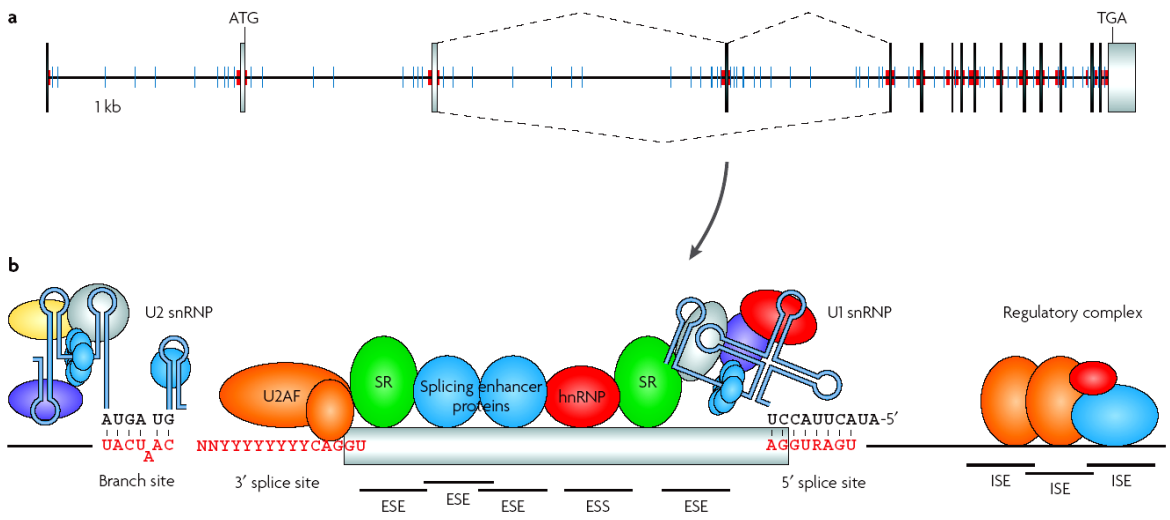

Figure 1.2 The cellular splicing code.
a) Pre-mRNA as it might appear to the spliceosome. Thick (or red) indicates consensus ss sequences at the intron-exon boundaries. Intronic thin (or blue) indicates additional intronic *cis*-acting elements that make up the splicing code. **b)** *cis*-elements within and around an alternative exon are required for its recognition and regulation. The 5′ ss and branch site serve as binding sites for the RNA components of U1 and U2 snRNPs, respectively. Exons and introns contain diverse sets of enhancer and suppressor elements that refine *bone fide* exon recognition. HnRNPs can inhibit exon definition by sterically blocking SR or U2AF interaction with the substrate (House and Lynch, 2008). HnRNPs also exert their actions on pre-mRNA differential splicing through either multimerization or looping-out mechanisms (Blencowe, 2006; Martinez-Contreras *et al.*, 2006). *Illustration from (Wang and Cooper, 2007).*

Even with the recent progress in identification of the precise consensus sequence of ss (Gao *et al.*, 2008), the mechanism of ss recognition is not yet fully understood. The current two models depend on intron length and the initial steps of the spliceosome assembly. In the "intron definition" or traditional model, ss of introns <250 nt in length are recognized across the intron. The formulation of this model depends on the direct identification of the 5′ and 3′ ss of introns as the splicing unit, and spliceosomal components assembled around the intron that will be excised (Hertel, 2008). In the "exon definition" or new model, ss of long introns are usually recognized across the exon. Here, an interior exon is first recognized by the paired binding of Ul and U2 snRNPs and associated splicing factors to the 5′ and 3′ ss, followed by the juxtaposition of neighboring exons in the correct order (de Almeida and Carmo-Fonseca, 2008; Hertel, 2008). It further assumes that processing of the last exon involves interaction between splicing components at the 3′ ss and the polyadenylation complex, whereas recognition of the first exon is thought to be mediated by interactions of the nuclear cap-binding complex with the spliceosome (de Almeida and Carmo-Fonseca, 2008). One mechanistic difference between the two models of ss selection may be the requirement of an additional exon juxtaposition step during exon definition (Hertel, 2008). Recent evidence also

indicates that intron excision from pre-mRNAs of higher eukaryotes requires a ‘transition’ from ss recognition across short exons to organization of the spliceosome across long introns (Schellenberg *et al.*, 2008).

1.3. Alternative splicing and biological complexity: One gene, many proteins

One of the most remarkable observations stemming from the sequencing of genomes of diverse species is that the number of protein-coding genes in an organism does not correlate with its overall cellular complexity. From where does complexity spring if not from the number of genes in an organism? Alternative pre-mRNA splicing is believed to be a major mechanism to bridge the gap between the gene and protein number (Graveley, 2001; Maniatis and Tasic, 2002), thereby allowing the expansion of the proteome and regulation of gene expression in higher eukaryotes. Alternative splicing is also known to play numerous critical roles in both normal and disease processes (Blencowe, 2006; Gabut *et al.*, 2008). By definition, AS is the process by which pairs of ss are differentially selected to generate multiple mRNA variants from a single precursor (pre-) mRNA (Gabut *et al.*, 2008). The greater frequency of AS events in mammals than in vertebrates again reflects the contribution of AS to this biological complexity (Kim *et al.*, 2004). Furthermore, it has been estimated that 40-60% of all human genes (Brett *et al.*, 2002; Boue *et al.*, 2003) and 74% of multi-exon human genes (Kapranov *et al.*, 2002; Johnson *et al.*, 2003) are alternatively spliced. In fact, large fraction of AS undergoes cell-specific regulation in which splicing pathways are modulated according to cell type, developmental stage, gender, or in response to external stimuli (Faustino and Cooper, 2003). Despite the growing list of mammalian protein factors known to regulate AS (Gabut *et al.*, 2008), we still lack the information that allows us to predict cell- and tissue-specific AS or even which protein factors are most likely to target which exons (Blencowe, 2006).

1.3.1. Patterns of alternative pre-mRNA splicing

In a typical multi-exon mRNA, the splicing pattern can be altered in many ways (Figure 1.3). Most exons are constitutive; they are always spliced or included in the final mRNA. When a constitutive ss is put in a competitive context with other ss, often little is needed to switch a particular ss from a constitutive to an alternative one. Numerous examples of this scenario have been described leading to either alternative 5′ or 3′ ss usage, skipping of an exon (Cartegni *et al.*, 2002; Black, 2003) or acquisition of new exons from repetitive *Alu* elements

(Lev-Maor *et al.*, 2003; Sorek *et al.*, 2004). A regulated exon that is sometimes included and sometimes excluded from the mRNA is called a cassette exon, which represents the most common type of AS, accounting for 33 to 53% (Thanaraj and Stamm, 2003; Blencowe, 2006). In certain cases, multiple cassette exons are mutually exclusive-producing mRNAs that always include one of several possible exon choices but no more; these type of exons are interchangeably used in the alternative transcripts (Malousi *et al.*, 2007). The 5′-terminal exons of an mRNA can be switched through the use of AS and alternative promoters, which are primarily an issue of transcriptional control. Similarly, the 3′-terminal exons can be switched by combining AS with alternative polyadenylation sites. Control of polyadenylation appears mechanistically similar to control of splicing (Colgan and Manley, 1997). Finally, some important regulatory events are controlled by the failure to remove an intron from the transcript, a splicing pattern called intron retention (Black, 2003; Malousi *et al.*, 2007). AS events are also classified into simple and complex depending on whether the exons flanking an alternatively spliced exon undergo a specific type of the aforementioned AS events (Thanaraj and Stamm, 2003; Malousi *et al.*, 2007). Diverse silencer sequences, as well as some ESEs, play an important role in controlling the selection of alternative 5′ and 3′ ss and a specific class of silencers may also function to regulate intron retention events (Wang *et al.*, 2006). Moreover, the transcription factors acting at the level of initiation and elongation can impact ss selection. In particular, factors resulting in reduced rates of RNA polymerase II (Pol II) elongation can increase the inclusion of alternative exons (Kornblihtt, 2006).

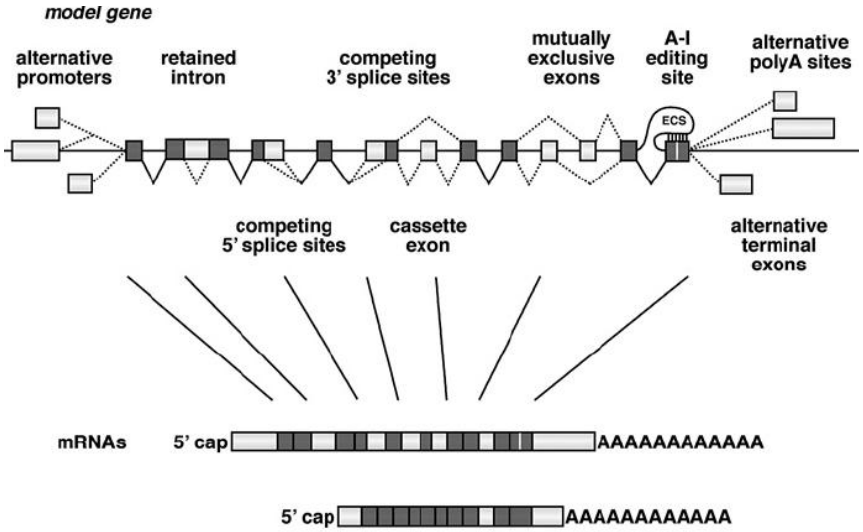

Figure 1.3 Modes of alternative splicing.
Exons are shown as boxes and introns as lines. Gene regions with AS processing choices are illustrated in white and connected with dashed lines, while constitutive parts are depicted in black and connected with solid lines. For adenosine to inosine editing (A to I) an editing site complementary sequence (ECS) located in an intron pairs with the edited site in the exon. *Design from (Soller, 2006)*.

1.3.2. Splicing regulatory mechanisms at genomic dimensions

Splicing at short-distance tandem sites. Alternative splicing at donor or acceptor sites located just a few nucleotides apart is widespread in many species (Hiller and Platzer, 2008). For instance, NAGNAG tandem acceptors occur in ~30% and are functional in at least 5% of human genes, and 1.3% of the splice donors allow AS at both GY (underlined) of the unusual motifs GYNGYN (Hiller *et al.*, 2004; 2006a; Hiller *et al.*, 2006b). Both types of tandems enable subtle protein variations (Hiller *et al.*, 2006b). Several of these tandem splice events contribute to the repertoire of functionally different proteins (advantageous), whereas many are neutral (being tolerated) or deleterious (may be causing disease). Remarkably, some of the functional events are differentially spliced in tissues or developmental stages, whereas others exhibit constant splicing ratios, indicating that function is not always associated with differential splicing (Hiller and Platzer, 2008). A large fraction may arise as a consequence of stochastic binding of the spliceosome at neighbouring ss (Chern *et al.*, 2006).

Splicing at long-distance. A common feature of genes in higher eukaryotes is the presence of very large introns, often extending over tens of kilobases. In the human *neurexin 3* gene, which spans 1600 kb, the largest constitutively and alternatively spliced introns are 292 kb (between exon 16 and 17) and 347 kb (between exon 1 and 5), respectively (Tabuchi and Sudhof, 2002). Frequently hidden in such large introns are very short alternatively spliced exons (cassette exons), such as the 12 nt long exon 4 in the *neurexin 3* gene. There are three major mechanisms to facilitate splicing of large introns: 1) looping out of intronic sequences to bring ss into proximity; 2) recursive splicing, which occurs when the 5′ ss is regenerated after splicing of an intron and is used again; 3) intra-splicing, which occurs in nested genes that are transcribed in the same direction (details in (Soller, 2006)). Beyond *cis*-splicing at a single locus, there is evidence for specialized *cis*-splicing that results from read-through transcription of adjacent loci followed by splicing to generate transcription-induced chimeras from two genes, as in the *TNSF12*/*TNSF13* chimera expressed in human T cells (Pradet-Balade *et al.*, 2002). In contrast to these *cis*-splicing events, *trans*-splicing joins exons from separate pre-mRNA transcripts. These transcripts can be encoded by different DNA strands at the same locus, as in *trans*-splicing of the *mod(mdg4)* gene in *Drosophila*, or by different alleles at the same locus, as for the *lola* gene, also in *Drosophila* (Horiuchi and Aigaki, 2006).

Alternative donor ss selection. A widely accepted mechanism for alternative donor splicing is the differential binding of the U1 snRNA to one of the potential donor sites. According to

this ss competition model, AS happens when one donor is sufficiently able to compete with the other donor for U1 binding. Constitutive splicing at a tandem motif (exclusive selection of only one donor) occurs when either donor is much stronger and consequently outcompetes the other. Apart from the intrinsic strength of donor sites, SR proteins and hnRNPs affect ss selection. SF2/ASF and other SR proteins promote splicing at the intron- proximal donor site, whereas hnRNP A1 promotes the distal site. The relative concentration of SR proteins and hnRNPs affects donor selection, and tissue-specific variations in this ratio might lead to tissue-specific splicing patterns (Caceres *et al.*, 1994) (reviewed in (Hiller and Platzer, 2008)).

Alternative acceptor ss selection. *In vitro* experiments found evidence for different modes of acceptor AG selection that depend on the distance of the AG to the branch point. If the branch point is more than ~ 20-35 nt away from the AG, the AG selection occurs by a scanning mechanism that starts from the branch point and usually selects the intron-proximal AG (Smith *et al.*, 1993; Chen *et al.*, 2000). The proximal AG can be bypassed if it is too close to the branch point or if it is in competition with a more distal AG. This competition can lead to AS and depends on (1) the distance between the AGs (shorter distances lead to a higher competition), (2) the nucleotide upstream of the AGs (C and T are preferred over A and especially over G) and (3) the sequence between both AGs (Smith *et al.*, 1993; Chen *et al.*, 2000; Chua and Reed, 2001; Dou *et al.*, 2006). Scanning does not occur if the distance to the branch point is short (<20 nt) (Chen *et al.*, 2000). In these cases, a distal AG can efficiently compete with a proximal AG given the distance between both AGs <6 nt (Chua and Reed, 2001). Similar to donor selection, SR proteins were shown to promote proximal acceptor ss, whereas hnRNP A1 promotes distal sites (Bai *et al.*, 1999). Thus, although AS is often regulated at the early splicing step, alternative acceptor selection can be regulated at the early and the late step (Lallena *et al.*, 2002; Dou *et al.*, 2006) (reviewed in (Hiller and Platzer, 2008)).

1.3.3. Global functions and communication of alternative splicing

Protein isoforms, produced by AS, can differ in various aspects, including ligand binding affinity, signaling activity, protein domain composition, subcellular localization, and protein half-life (Stamm *et al.*, 2005). In coordination with non-sense mediated decay (NMD), alternatively spliced transcripts can be degraded rapidly, providing a regulation and fine-tuning mechanism of the adjustment of the protein level (Lewis *et al.*, 2003). About 25% of transcripts will be switched off by pre-mature stop codons (PTCs) caused by AS and NMD.

This process that termed RUST, for regulated unproductive splicing and translation, currently represents the function of AS with the most obvious biological consequences (Stamm *et al.*, 2005). The coupling between transcription, mRNA processing and mRNA surveillance avoids the wasteful production of nonfunctional mRNAs with potentially deleterious effects for the cell. Current models suggest that RNA processing factors, such as U1 snRNP and SR proteins, are loaded onto the C-terminal domain of RNA Pol II and deposited on native transcripts as they are synthesized, thereby, promoting rapid cotranscriptional spliceosome assembly (de Almeida and Carmo-Fonseca, 2008). These models also imply that specific promoters might differentially affect AS processing by interaction with different factors (Soller, 2006).

Although several observations suggest that splice variants may have a biological role, the mere presence of a splice variant in tissues does not mean that it has a biological function. In many cases AS occurs in genes that encode multidomain proteins where splice variants encode proteins that differ in their domain organization and hence are likely to differ in function. Thus, all splice variants deserve close scrutiny to determine if they have a regulatory role before they are ignored as artefacts. There is also a caveat in assuming that only conserved AS types are meaningful. AS events that are not evolutionarily conserved are not necessarily unimportant as they may be specific to one organism and reflect the biology of that organism and/or might have evolved more recently and contributed to the diversification of species (Reddy, 2007). Once again, variable splicing ratios do not always imply functional importance (Hiller and Platzer, 2008), and the tissue-specific expression of a gene does not always imply a tissue-specific function (Cajiao *et al.*, 2004).

1.3.4. Components influencing exon recognition and alternative splicing

Given the complexity of higher eukaryotic genes and the relatively low level of ss conservation, the precision and flexibility of the spliceosome to identify and process exons within a given pre-mRNA is impressive. Indeed, multiple factors interact in these processes and include parameters such as, ss strength, the presence/absence of splicing regulators, RNA secondary structures, the exon/intron architecture, and the synthesis of pre-mRNA by RNA pol II itself (Figure 1.4). The relative contributions of each of these parameters control how efficiently ss are recognized and flanking introns are removed. Examples include: 1) greater complementarity with U1 snRNA and longer polypyrimidine tracts translate into higher affinity binding sites for these spliceosomal components and, thus, more efficient exon

recognition. 2) The interplay between activating and repressing *cis*-acting elements modulate the probability of exon inclusion. 3) Splice-site recognition is more efficient when introns or exons are small. 4) Exon skipping is also more likely to occur when exons are flanked by long introns in the human genome, most likely reflecting the influence RNA transcription exerts on pre-mRNA splicing. 5) Local RNA structures can either interfere with spliceosomal assembly, by either masking splicing repressor binding sites, looping out the exon, or preventing its recognition. 6) Like 5′ capping and 3′ polyadenylation, intron removal is physically and temporally linked to RNA transcription. That is, the ss of an exon can be identified by the spliceosome while downstream exons still await their synthesis by RNA Pol II (Hertel, 2008).

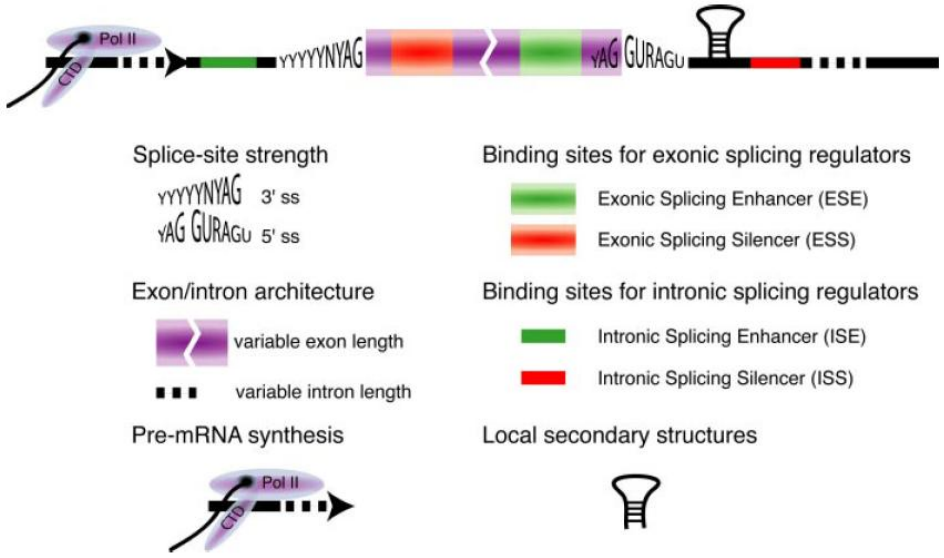

Figure 1.4 Several components influencing exon definition. *Illustration from (Hertel, 2008).*

Perturbations of post-translational modifications that are essential for optimal activity of many regulatory splicing factors, such as alterations in the phosphorylation state of specific SR proteins, also modulate AS. By influencing protein/protein and protein/RNA interactions, reversible protein phosphorylation modulates the assembly of regulatory proteins on pre-mRNA and therefore contributes to the splicing code. Different kinases and protein phosphatase 1 are identified as the molecules that control reversible phosphorylation, which controls not only ss selection, but also the localization of SR proteins and mRNA export. Protein phosphatase 1 moves between cellular compartments, depending on the activity of the cell. This dynamic behavior links splicing to other activities of the cell and provides evidence as to how cellular signals modulate gene expression by influencing AS (Stamm, 2008).

1.4. Pre-mRNA (mis)splicing as a primary cause of disease

The physiological importance of keeping mRNA biogenesis under tight quality control is well-illustrated by the growing number of human diseases known to be caused by errors in mRNA processing (de Almeida and Carmo-Fonseca, 2008). The AS processes are well regulated, but when mutations disrupt the ss or regulatory elements, disease can occur (Hiller *et al.*, 2006a; Solis *et al.*, 2008). Mutations can also cause disease through aberrant transcript production (Hiller *et al.*, 2006a; Solis *et al.*, 2008). Missplicing of cellular genes can either be a symptom of an underlying molecular defect, or the actual cause of the disease. It has recently been proposed that 60% of mutations that cause disease do so by disrupting splicing (Lopez-Bigas *et al.*, 2005) and wrong ss usage has been observed in numerous diseases. Changes in AS are frequently observed in cancer, where they are probably the result of the cellular transformation. In several genetic diseases, such as FDTP-17 or spinal muscular atrophy (SMA), a change in splicing is caused by mutations and is the actual cause of the disease. Furthermore, regulated AS events control apoptosis and are necessary for the replication of many viruses, such as HIV. One way to treat such diseases would be to influence AS pathways and to send undesired cells into apoptosis or stop viral replication (Stamm, 2008).

The distinction between *cis*- and *trans*-acting effects has important mechanistic implications. Effects in *cis* have a direct impact on the expression of only one gene, whereas effects in *trans* have the potential to affect the expression of multiple genes (Faustino and Cooper, 2003; Wang and Cooper, 2007). *Cis*-acting mutations can affect the use of constitutive or alternative ss. Disrupted constitutive splicing most often results in the loss of gene expression due to aberrant splicing. Alternatively, a *cis*-acting mutation that (in)activates one of two alternatively used ss will force expression of one of the AS patterns. Although a natural mRNA is expressed, its expression in an inappropriate tissue or developmental stage might result in disease. Then again, *trans*-acting splicing mutations can affect the function of the basal splicing machinery or factors that regulate AS. Mutations that affect the basal splicing machinery have the potential to affect splicing of all pre-mRNAs, whereas mutations that affect a regulator of AS will affect only the subset of pre-mRNAs that are targets of the regulator (Faustino and Cooper, 2003).

1.4.1. *Cis*-acting mutations: Possible dramatic effects upon the splicing code

Transcript analyses from specific disease genes such as *NF1*, *ATM*, and others lead to the striking conclusion that as many as 50% of disease mutations in exons may impact on splicing (Cartegni *et al.*, 2002; Blencowe, 2006; Pagenstecher *et al.*, 2006). Of the mutations in the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2003) that are not within ss, 78% are SNPs within exons (56.9%) or in microdeletions or microinsertions of up to six nt that occur primarily within exons (21.7%) (Wang and Cooper, 2007). Recently, the neural network efforts (Krawczak *et al.*, 2007) have shown that within a splice site, SNPs and disease-associated (HGMD) mutations outside the obligate dinucleotides also differ from each other, both for acceptor and donor ss (Figure 1.5).

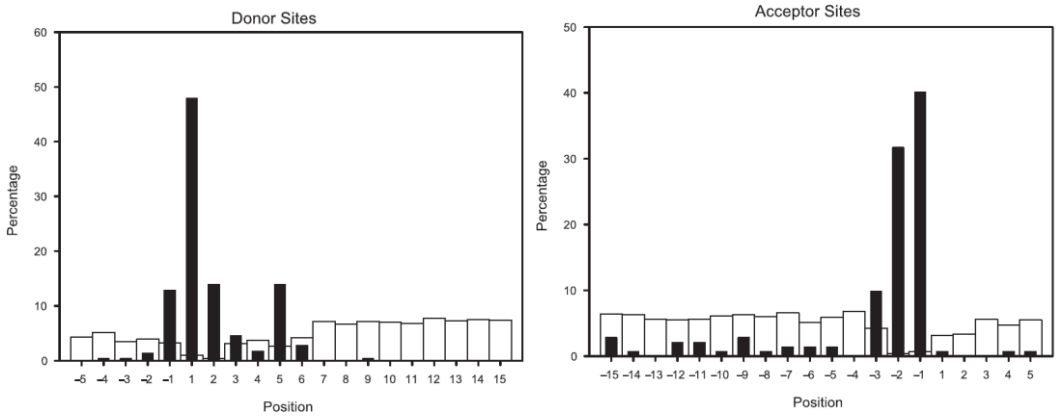

Figure 1.5 Distribution of SNPs and splicing-relevant disease-associated mutations outside the obligate dinucleotide of splice-sites.
Open bars represent SNPs, whereas solid bars represent splicing mutations. *Charts from (Krawczak et al., 2007).*

Coding SNPs can disrupt (or eventually create) ESE and ESS; create new ss or strengthen cryptic ones; alter pre-mRNA secondary structures important for exon-definition; and, conceivably, modify the pausing architecture of a gene, provoking changes in RNA Pol II processivity, which might in turn affect ss choice. These defects are not exclusive of cSNPs: missense, nonsense and translationally silent mutations as well as exonic deletions or insertions can affect AS in similar ways. More than 50% of such mutations have been shown to disrupt at least one of the target motifs for the SR proteins (SF2/ASF, SRp40, SRp55 and SC35) found in ESEs (Liu *et al.*, 2001). Even one-quarter of synonymous substitutions of exons 9 and 12 of the *CFTR* gene, which is mutated in cystic fibrosis, affected splicing (Pagani *et al.*, 2005). Other examples of human disease genes, where (non)synonymous

mutations often affect exonic splicing control elements, include *BRCA1* (breast cancer 1, early onset), *HPRT1* (hypoxanthine phosphoribosyltransferase 1), and *MAPT* (microtubule-associated protein tau) (Cartegni *et al.*, 2002).

Mutations located in noncoding regions, such as those affecting 5′ and 3′ ss, branch sites or polyadenylation signals, are frequently the cause of hereditary disease. Approximately 15% of mutations that cause genetic disease affect pre-mRNA splicing (Krawczak *et al.*, 1992). In particular, 28% of the NAGNAG SNPs occur in known disease genes (Hiller *et al.*, 2006a). Splice acceptors with the genomic NAGNAG motif may cause NAG insertion-deletions and can result in the gain/loss of a PTC in transcripts (Hiller *et al.*, 2004; 2006a). Once more, many human diseases, including Fanconi anemia, hemophilia B, neurofibromatosis, and phenylketonuria, can be caused by 5′ ss mutations that are not predicted to disrupt splicing. It is likely that such mutations disrupt the conserved pairwise dependencies between 5′ ss nucleotides, as some human SNPs appear to alter splicing. The longer span and more plastic organization of 3′ ss suggest that the pairwise associations at 3′ ss will not reveal as many biases as the associations at 5′ss (Roca *et al.*, 2008).

It is also clear that most SNPs and/or mutations exert their effect by changing splicing regulatory elements, and the rest can be explained on the basis of secondary structure rearrangements. Likewise, secondary structure can explain why mutations that change splicing motifs sometimes show no splicing effect. Most likely, the affected motifs are highly double-stranded in these cases (Hiller *et al.*, 2007b). SNPs are capable of inducing *in vivo* different structural folds in mRNA structures and can ultimately affect biological function (Shen *et al.*, 1999). For example, a silent mutation in human *CFTR* exon 12 that reduces exon inclusion from 80%–25% (Pagani *et al.*, 2005) does not create or destroy splicing motifs, but leads to a higher single-strandedness of existing ESSs and a lower single-strandedness of an existing ESE (Hiller *et al.*, 2007b). The mechanistic explanations may involve the occlusion/exposure of key *cis*-acting elements or the spatial modification of the distance between these elements (Buratti and Baralle, 2004).

1.4.2. *Trans*-acting mutations: Disruption of the splicing machinery

There are several genetic diseases in which a mutation disrupts the machinery of splicing, through either the constitutive components of the spliceosome or auxiliary factors that regulate AS (Faustino and Cooper, 2003). Two diseases, spinal muscular atrophy (SMA)

(Briese *et al.*, 2005) and retinitis pigmentosa (Mordes *et al.*, 2006), are caused by mutations in genes involved in snRNP assembly and function, respectively. SMA is an autosomal recessive disorder affecting motor neurons and is caused by loss of the survivor of motor neuron-1 (*SMN1*) gene product, which is required for the assembly of core snRNPs in the cytoplasm before final maturation and nuclear import. Retinitis pigmentosa is one of the most common forms of blindness. Surprisingly, three dominant retinitis pigmentosa disease genes (pre-mRNA-processing factor gene homologues *PRPF31*, *PRPF8* and *HPRP3*) encode proteins required for proper assembly and function of the u4•u5•u6 tri-snRNP, a core and essential component of the spliceosome (McKie *et al.*, 2001; Vithana *et al.*, 2001; Chakarova *et al.*, 2002). The disease is probably due to disruption of spliceosome function, because it seems unlikely that all three genes or the u4•u5•u6 tri-snRNP have alternative functions. A third example is the Prader-Willi Syndrome (PWS), which is the first known example of a genetic disease in which pathogenesis might be due to mutation of a gene encoding a splicing regulatory factor. The PWS is a congenital disease that is caused by the loss of paternal gene expression from a maternally imprinted region on chromosome 15. Kishore et al. (Kishore and Stamm, 2006) provide evidence that HbII-52 snoRNA (small nucleolar RNA) exhibits sequence complementarity and regulates splicing of exon Vb of the serotonin receptor 5-HC_{2C}R. Loss of HbII-52 snoRNA expression in PWS results in aberrantly regulated splicing of 5-HC_{2C}R. These results show that a snoRNA regulates the processing of a mRNA expressed from a gene located on a different chromosome and further indicate that a defect in pre-mRNA processing contributes to the PWS phenotype (Wang and Cooper, 2007).

1.5. Study of allele-dependent splicing: Motivations and Perspectives

Many facts are now becoming clearer to us. Alternative pre-mRNA splicing is a widespread phenomenon that affects approximately 75% of human genes (Moore and Silver, 2008). At the same time, many SNPs are known to be located in genomic regions of splicing relevance, including canonical ss, ESE, ISE, and other DNA sequence motifs (Fairbrother *et al.*, 2004a; Pagani and Baralle, 2004; Kralovicova *et al.*, 2005). These polymorphisms may lead to a disruption of the ‘splicing code’, thereby causing the splicing apparatus to utilize cryptic ss nearby or to skip one or more exons. This means that common genetic variation can result in substantial, phenotypically relevant variation at the protein level (Wang and Cooper, 2007). Indeed, several examples of splicing-relevant SNPs underlying human diseases have been reported (Cartegni and Krainer, 2002; Colapietro *et al.*, 2003). Variation in splicing patterns is known to be tissue specific, and for a small number of genes has been shown to vary among

individuals (Hull *et al.*, 2007). Nevertheless, what is not clear is whether allele-dependent splicing phenomenon is an important mechanism by which common genetic variation affects gene expression and to what extent. Therefore, the screen of allele-dependent splicing occurrence in the context of the present study was motivated from many perspectives.

1.5.1. Genomic and mechanistic perspectives

After the sequencing of the human genome, one of the key questions in the field is the correlation of genetic and phenotypic variation. The population of mRNA splice products generated from a specific (DNA) haplotype can be regarded as an 'intermediate phenotype', or 'mRNA-phenotype', and the detailed characterization of this phenotype is clearly one of the prerequisites for being able to draw a link between variation at the DNA sequence and protein level (Graveley, 2008). As discussed from a disease perspective below, it is unlikely that the marked phenotypic diversity of complex organisms can be explained on the basis of single amino acid substitutions or frame shift mutations alone. It is hypothesized here that variation in splicing-relevant sequence motives may be an important factor for transcriptome variability. On the other hand, the sequence elements that control splicing are all relatively short and show little local (or evolutionary) conservation. Only the AG and GT dinucleotides present at virtually all splice acceptors and donors, respectively, and the branch site appears to be mandatory for correct and efficient splicing. The role of related control sequences like enhancers (ESE, ISE) or silencers (ESS, ISS) is less well understood. Owing to the scarceness of data, the splicing effects of individual genetic variants can still only be predicted with limited accuracy (Pagani and Baralle, 2004; Wang *et al.*, 2004b), and most of the information on splicing motifs has been obtained from the alignment of such motifs to either ESTs or known structural gene models (Krawczak *et al.*, 1992; Clark and Thanaraj, 2002). Overall, the available empirical data on allele-dependent splicing is limited and has mostly been generated in a non-systematic fashion.

1.5.2. Disease relevance of allele-dependent splicing

There is a growing realization that splicing efficiency is a significant contributor to phenotypic variability (Marden, 2008), and the contribution of splicing to phenotype has become particularly apparent through its impact on modifying the severity of human disease (Nissim-Rafinia and Kerem, 2002) and its contribution to disease susceptibility (Wang and Cooper, 2007). Indeed, a key factor in the motivation for the present study was the recent

positional cloning of the first sarcoidosis (a polygenic autoimmune disorder of the lungs (Valentonyte *et al.*, 2005)) disease gene, namely *BTNL2*. Here, the main genetically associated SNP was predicted to cause an amino acid exchange. However, the functional impact exerted through that mutation could not explain the profound genetic effect observed. Indeed, the SNP-effect was strictly genotype-specific. This SNP, rs2076530, caused a substantial "weakening" of ss and thus lead to use of a cryptic donor ss. This leads to a loss of 4 bases in the transcript and a subsequent frame-shift and protein truncation (loss of transmembrane domain). Indeed, an estimated 20%–30% of disease-causing mutations is believed to affect pre-mRNA splicing (Faustino and Cooper, 2003), which is consistent with the recent suspicion that splicing mutations to be the most frequent cause of hereditary diseases (Lopez-Bigas *et al.*, 2005). Consequently, an increasing number of SNPs have been described that cause diseases, both monogenic and polygenic, by a change or disruption of the normal splicing pattern (Cartegni *et al.*, 2002; Garcia-Blanco *et al.*, 2004). These splice-relevant SNPs can alter important mRNA secondary structures affect donor and acceptor ss, branch points, exonic as well as intronic splicing enhancers and silencers (Hiller *et al.*, 2006a; Hiller *et al.*, 2007b). For example, the G allele of the silent coding SNP rs17612648 in the *PTPRC* gene that is associated with multiple sclerosis disrupts an ESS and abolishes the skipping of exon 4 (Lynch and Weiss, 2001). In fact, several splicing mutations of known disease relevance (Stenson *et al.*, 2003) have also been studied in controlled *in vitro* experiments (Pagani *et al.*, 2000; Wang *et al.*, 2004a; Zuccato *et al.*, 2004).

Traditionally, researchers who want to track down the molecular basis of monogenic ("classic Mendelian") disorders have focused on frame-shifts or mutations that directly change the amino acid composition of proteins. Increasingly, mutations that influence the protein sequence or expression through effects on splicing are being recognized as causative factors in Mendelian disorders (Teraoka *et al.*, 1999; Ars *et al.*, 2000). On the other hand, the clarification of polygenic ("complex") disorders through systematic positional cloning is a rapidly evolving field. Within the last few years, positional cloning successes have been published for around ten genes in various disorders including rheumatoid arthritis, inflammatory bowel disease, Parkinson's disease, but to name a few. For instance, the common allelic variation, which has been correlated with lower transcript levels of the soluble alternative splice form of *CTLA4* gene, contributes to autoimmune tissue destruction (Ueda *et al.*, 2003). The pace of discovery has significantly increased over the last few years, due to the availability of the annotated human genome sequence, the rapid development of

novel, more cost-efficient genotyping technologies, and the availability of large, well-characterized patient cohorts.

An important – and still very "low-throughput" step in this process is the establishment of tangible links between genetic variation and transcript function. Obvious changes in gene function (e.g. amino acid exchanges at an active site, frame shift mutations) are present only in a fraction of cases. In-house expertise and reports in the literature show the difficulty of translating a clear genetic finding (i.e. a SNP with consistently positive association to disease in multiple populations) into a functional meaning. It is not clear what proportion of phenotypically relevant gene alterations is the result of mutation-driven splicing effects. Studies, which systematically addressed this issue in the monogenic disorders of ataxia telangiectasia and neurofibromatosis type I, have shown that splice mutations were involved in up to 50% of cases (Teraoka *et al.*, 1999; Ars *et al.*, 2000). On the other hand, the wide range for the predicted frequency of splicing mutations (15–60%) reflects our incomplete knowledge of the splicing code and the fact that mRNAs from mutant alleles are rarely assayed for splicing abnormalities. A long-term goal of deciphering the splicing code is to acquire the power to predict which disease-associated nucleotide alterations are likely to affect splicing (Wang and Cooper, 2007). Therefore, methodology and database resources that allow a rapid assessment of allele-dependent splicing for transcripts of interest (e.g. in an associated region) would greatly enhance the efficiency of gene finding experiments (Pagani and Baralle, 2004). In the meantime, systematic surveys of various disease-causing mutations for aberrant splicing would be enlightening, both for the individual mutations that are analyzed and for a broad-based analysis of the impact of splicing as a primary mechanism of disease. The main limitation will be obtaining RNA from the disease-relevant tissues (Wang and Cooper, 2007).

1.5.3. Recent surveys approaching allele-dependent splicing

While splicing defects have been well-studied in the context of rare diseases, the extent to which common SNPs influence mRNA processing is still relatively unknown. Recent large-scale studies have suggested that a relatively high proportion of human genes show variation in expression due to allele-dependent splicing. In the few experiments approaching the last phenomenon from a transcript-based perspective, known SNPs identified as candidates for a splicing effect were tested with isoform-specific PCR (Hull *et al.*, 2007), or chip-based methods were used to seek evidence for alternative splicing (Kwan *et al.*, 2007; Kwan *et al.*,

2008). Hull et al. (2007) analyzed the splicing patterns of 250 exons in 22 individuals who had been previously genotyped as part of the HapMap project. Consistent allele-dependent splicing was identified for six of these exons and the strongest effects were observed for SNPs close to an intron-exon boundary. In a genome-wide screening experiment using an exon tiling microarray, Kwan et al. were able to show co-segregation of isoforms and haplotypes (Kwan *et al.*, 2007) and to correlate splicing patterns to genotypes at adjacent SNPs (Kwan *et al.*, 2008). Another concurrent effort combined a genome-wide scan of publicly available EST and exon array data and showed evidence of allele-specific splicing events closely linked to SNPs (Nembaware *et al.*, 2008).

1.5.4. Biomedical perspective and drug design strategies

Aside from its intrinsic biological concern, there is also a major biomedical interest in understanding the functional role of gene isoforms, as targeting the wrong isoform may result in unexpected damaging effects (Talavera *et al.*, 2007). On the other hand, altered splicing patterns can serve as markers of the altered cellular state associated with disease even when they are not in the primary pathway of the disease mechanism and still have the potential to provide diagnostic and prognostic information (Faustino and Cooper, 2003). Since several drugs have been demonstrated to be significantly modulated by single nucleotide changes, the analysis of allele-dependent RNA splicing repertoires can help in this respect. The various therapeutic approaches that utilize splicing can either alter the splicing patterns of target genes or target specific splice variants at the RNA or protein level to achieve a therapeutic effect. For example, antisense RNA and DNA, small interference RNA (siRNA), and ribozymes are ectopic oligonucleotides that can be designed to recognise target aberrant mRNA molecules and elicit their cleavage (recently reviewed in (Pajares *et al.*, 2007; Wang and Cooper, 2007)).

1.5.5. Predicting effects of splice-relevant SNPs

Soon after the discovery of exons and introns in adenovirus 2 genes in 1978 (Berget *et al.*, 1977; Chow *et al.*, 1977) Walter Gilbert (Gilbert, 1978) postulated that: (1) different combinations of exons could be joined together to produce multiple mRNAs from a single gene; and (2) mutations at exon-intron junctions and at silent codon positions could influence pre-mRNA splicing modulation and lead to functionally different proteins. Although this hypothesis is in part proved in many recent genome-wide projects, it is still a challenging to deeply understand how single nucleotide change cause molecular alterations and expression

changes in gene products through pre-mRNA splicing process. The missing link between the generation of genomics data and their analysis by conventional biological approaches (Teufel *et al.*, 2006) as well as the difficulty of the experimental approach (in particular, in its high-throughput version), leaves ample room for the development of many bioinformatic tools that can provide a first picture of the problem (Talavera *et al.*, 2007). Given that the impact of SNPs on splicing is also hard to predict *in silico*, silent or intronic SNPs that may cause a disease phenotype by changing splicing patterns are often not investigated (Pagani and Baralle, 2004). Furthermore, although modifications of the highly conserved AG/GT dinucleotides at the ends of introns are usually classified as deleterious, the impact of nucleotide variations at loosely defined positions and the creation of cryptic ss are more challenging. Unfortunately, testing all nucleotide modifications at the RNA level is a tremendous task that cannot be performed in a routine diagnostic setting. In this respect, assessment of the potential splicing effect of a specific mutation is important for decision-making in molecular diagnostics (Houdayer *et al.*, 2008).

The use of algorithms allowing correct and reliable predictions of the impact of SNPs upon splicing process would therefore be of utmost importance. Several tools have been devised to undertake such assessment, including splice site prediction by a neural network (Reese *et al.*, 1997) (NNSplice 0.9: *http://www.fruitfly.org/seq_tools/splice.html*), Human Splicing Finder (Version 2.3: *http://www.umd.be/HSF/*), MaxEntScan (Yeo and Burge, 2004) (MES: *http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html*), automated splice site analysis (Nalla and Rogan, 2005) (ASSA: *http://splice.uwo.ca/* - service discontinued), the AST web application (*http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm*), ESE Finder (Cartegni *et al.*, 2003) (*http://rulai.cshl.edu/tools/ESE/*) and Relative Enhancer and Silencer Classification by Unanimous Enrichment (Fairbrother *et al.*, 2002) (RESCUE-ESE: *http://genes.mit.edu/burgelab/rescue-ese/*).

The tools for the analysis of the canonical splice sites such as Alex's splice site score calculator, NNSplice, SSF and MES are based on the Shapiro and Senapathy matrices (Shapiro and Senapathy, 1987) but use different computational methods for splice site prediction. In principle, Shapiro and Senapathy carried out a systematic analysis of the RNA splice junction sequences of eukaryotic protein coding genes using the GENBANK databank. Thereby, they were able to identify splice junction consensus sequences from different classes of organisms. In turn, this led to the identification of potential ss in raw DNA sequences and

the finding of new ss and exons in known gene sequences, which may yield AS products in different *in vivo* situations. The recently reported neural network (NN) used in the present study (Krawczak *et al.*, 2007) used all annotated RefSeq exon boundaries for training. It runs distinct algorithms that generate a score matrix for each ss (donor and acceptor). The RESCUE-ESE analysis tool searches for hexanucleotide sequences as potential ESE motifs that have been identified through a computational method that assessed the relative abundance of hexanucleotide sequence stretches in exons as compared to introns (Fairbrother *et al.*, 2002; Fairbrother *et al.*, 2004b). ESEFinder uses systematic evolution of ligands by exponential enrichment (SELEX) and weight/position matrices to score ESEs responsive to the four most common human SR splicing factors proteins, namely SF2/ASF, SC35, SRp40 and SRp55 ESEs (Cartegni *et al.*, 2003). To ensure correct interpretation of the effects of disease-associated point mutations or polymorphisms, ESEfinder was freshly subjected to further refinement (Smith *et al.*, 2006), which resulted in a score matrix with increased specificity for the prediction of SF2/ASF-specific ESEs. The recent interest in allele-dependent splicing is also emphasized by a computational study (Lee and Shatkay, 2008) that yielded a list of potential splice SNPs through the combined use of 16 different bioinformatics tools and databases (*http://compbio.cs.queensu.ca/F-SNP/*).

1.6. Aims of the study

The main aim of the present study was to systematically determine the extent to which common SNPs at splice sites influence pre-mRNA (alternative) splicing.

To address this aim, the following specific sub-aims were addressed:

- **Establishment of a high-throughput methodology to assess the impact of naturally occurring SNPs on differential splicing.**

- **Evaluation of the current prediction tools of allele-dependent splicing by investigating four specific classes of putative splice-SNPs:**
  - **the mutational imbalance at the authentic acceptor and donor consensus.**
  - **variation in tandem acceptors (NAGNAG).**
  - **the exonic-splicing enhancers (ESEs).**

- **Assessment of the overall importance and efficiency of allele-dependent splicing.**

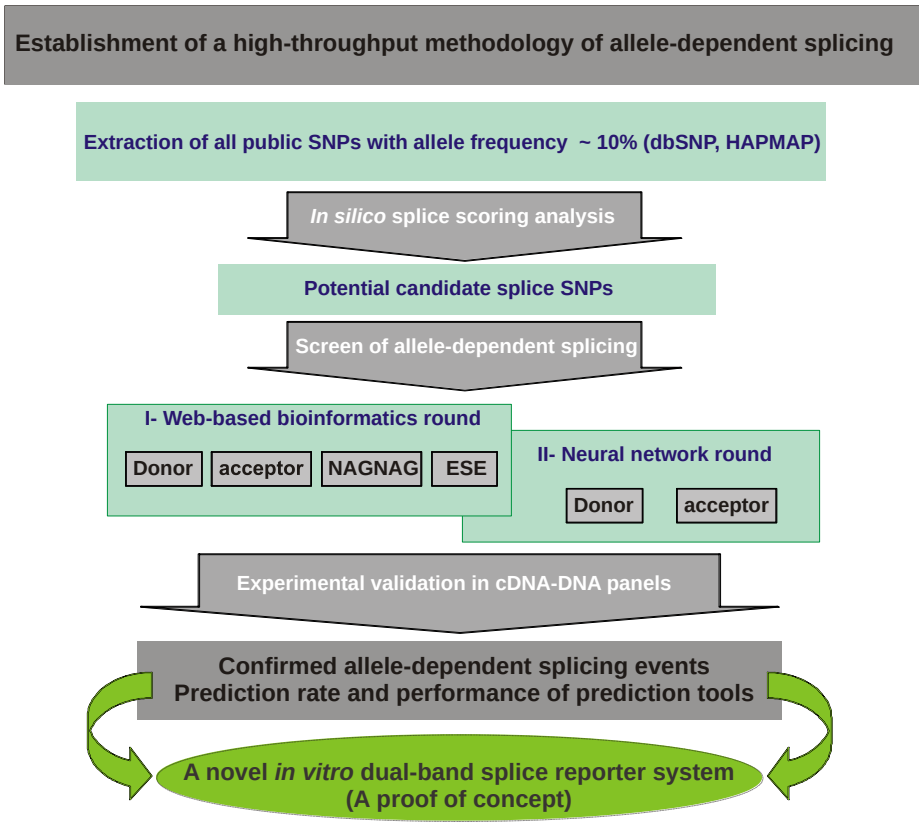

Figure 1.6 Flow diagram summarizes the experimental approach used in the present study.

2 METHODS

2.1. Selection of putative splice SNPs

In order to avoid unnecessary overlap, the methods used for the selection of putative splice SNPs are described in detail in the 'Results' section.

2.2. General methods

2.2.1. DNA extraction and quality control

DNA was isolated from different resources according to manufacturer's protocols. While 'MagAttract DNA Blood M48' (Qiagen, Hilden, Germany) was applied for lymphoblastoid cell lines, 'QIAamp® DNA Micro Kit' (Qiagen, Hilden, Germany) was used for extraction of DNA from brain tissues. From EDTA whole peripheral blood samples, DNA was extracted using the 'Invisorb® Blood Universal Kit' (Invitek, Berlin, Germany) with few modifications to the manufacturer's protocol in order to get the best DNA quality, as briefly described below.

Fresh blood samples were stored at -80°C and thawed in a cold water bath before use. Erythrocyte lysis was achieved mainly by incubating 9 ml of blood for 10 min with 30 ml of Buffer EL at room temperature. Then the obtained suspension was centrifuged for 3 min at 3,000 rpm and the supernatant was carefully discarded. This step was repeated with 20 ml buffer EL until the leucocyte-containing pellet was free of haem, which could cause problems in downstream experiments since haem can inhibit PCR reactions (Heath *et al.*, 1999). The obtained pellet was resuspended in 3 ml of Lysis Buffer HL and 50 µl of Proteinase K and incubated for 2 hours at 65°C in a water bath under continuous shaking (95 rpm). The latter step leads to the lysis of the leukocytes and their nuclei to facilitate the release of DNA into the suspension, and agitation at 65°C improves lysis efficiency. To separate the DNA from cell and protein fragments, 1.8 ml of Precipitation Solution was added with vigorous mixing until white flakes of DNA precipitate became visible and then incubated on ice for 5 minutes. In case of unsuccessful DNA precipitation, the tube was incubated at -20°C for at least 2 hours. Using 1 ml-pipette, DNA flakes were drawn out of the tube and transferred into a fresh Eppendorf containing 1 ml of 70% ethanol, and the tube was carefully inverted several times. The DNA was pelleted by centrifugation for 2 min at 13,000 rpm. If the DNA pellets were very loose, centrifugation was either prolonged or the speed was increased. The supernatant

was very carefully removed. The purified genomic DNA was then resuspended in 1200 µl 1x TE buffer for normal size pellet or in 600 µl in case of small pellet. DNA was completely dissolved by overnight incubation at 60°C. Continuous shaking or inverting of the tube from time to time during the incubation was recommended to increase the dissolving efficiency. Finally, the obtained DNA was stored either at +4°C for a short time (few days) or at -20°C for longer periods (over months). The quality of the extracted DNA samples was checked by agarose gel electrophoresis and quantified using PicoGreen® assay (section 2.2.6.1).

2.2.2. Total RNA extraction and quality control

In order to obtain the suitable conditions for isolation of total RNA, all reagents, glassware and laboratory utensils were specially treated in order to avoid RNA degradation by RNAses. All plastic-ware was purchased as UV-sterilized consumables (50 ml conical tubes, pipet tips with aerosol filters) or RNAse-free consumables (microfuge tubes). All glassware, ceramic mortar and pestles, Teflon pestles and metal spatulas were cleaned with common laboratory washing detergent, rinsed thoroughly in distilled water and air-dried before wrapping in aluminum foil and baking at 180°C for 12-16 h before use, in order to inactivate any contaminating RNAses. In addition, solutions were prepared with 0.1% DEPC-treated distilled water and sterilized by autoclaving.

Total RNA from peripheral blood and cell lines was isolated using the RNeasy (Qiagen) system. For RNA extraction from lymphoblastoid cell lines as well as from snap-frozen surgical specimens of brain tissue, the TRIZOL reagent (Invitrogen, Karlsruhe, Germany) was used according to the supplier's protocols. RNeasy technology simplifies total RNA isolation by combining the stringency of guanidine-isothiocyanate lysis with the speed and purity of silica-gel-membrane purification. On the other hand, the TRIZOL reagent, a mono-phasic solution of phenol and guanidinium isothiocyanate, is an improvement to the single-step RNA isolation method developed by Chomczynski and Sacchi (Chomczynski and Sacchi, 1987; 2006). During sample homogenization or lysis, TRIZOL reagent maintains the integrity of the RNA, while disrupting cells and dissolving cell components. Addition of chloroform followed by centrifugation separates the solution into an aqueous phase and an organic phase. Total RNA remains exclusively in the upper aqueous phase, while most of DNA and proteins remain either in the interphase or in the lower organic phase. After transfer of the aqueous phase, the RNA is recovered by precipitation with isopropyl alcohol. After removal of the aqueous phase, the DNA and proteins in the sample can be recovered by sequential

precipitation. This technique performs well with small quantities of human tissue (50-100 mg) and cells (5 x 10^6), and yields 5-15 µg of RNA from 1 x 10^6 cultured cells.

Before using isolated RNA in subsequent cDNA synthesis, the concentration was measured using NanoDrop (section 2.2.6.2) and the integrity was checked on gel electrophoresis (section 2.2.4). DNA contamination in isolated RNA was tested by amplification of a housekeeper gene, glyceraldehyde-3-phosphate dehydrogenase (G3PDH), as described in section (2.2.7). PCR product was then checked on 1.5% agarose-gel electrophoresis. The presence of a band at 983 bp in positive control, and its absence in both RNA samples and water control, indicated that isolated RNA was free from genomic contamination. Otherwise, if detected, RNA contaminated with DNA was treated with DNase enzyme (Qiagen, Hilden, Germany) according to supplier's recommendations, and checked again with G3PDH-PCR test until genomic contamination was no longer detected.

2.2.3. First-strand cDNA synthesis and quality control

Five hundred nanograms of isolated RNA was reverse-transcribed into complementary DNA (cDNA) using either random hexamers, oligo(dT) primers, or gene-specific primers using the RevertAid H Minus First Strand cDNA Synthesis Kit from Fermentas Life Sciences (St. Leon-Rot, Germany), or the Clontech's Advantage® RT-for PCR kit system, according to the supplier's instructions. The success of the first strand cDNA synthesis was checked using 2 µl (1:10) diluted cDNA reaction as a template for PCR amplification with G3PDH, as described in section (2.2.7). The amplification of an intensive intact band at 983 bp indicated successful first-stand cDNA synthesis. Because the quantification of the prepared cDNA was essential in our downstream nested RT-PCR assays, cDNA concentrations were also measured using NanoDrop technique as described in section (2.2.6.2).

2.2.4. Gel electrophoresis

The percentage of agarose in the gel varied depending on the expected size(s) of the fragment(s) to be applied. The smaller the size, the higher the concentration of agarose was used. In each case, ethidium bromide solution (10 mg/ml; 1µl/100 ml agarose gel) was included in the gel to enable fluorescent visualization of the DNA fragments under UV light. According to standard protocol, agarose gels were then submerged in TBE electrophoresis buffer in a horizontal gel apparatus. The DNA samples were mixed with 5µl (2x) loading

buffer and loaded into the sample wells. Depending on the desired separation and size of the gel chamber (BioRad, Munich, Germany), different electric parameters were as standard applied. Size markers were also co-electrophoresed with DNA samples, when appropriate for fragment size determination. In general, two size markers were used, namely the 100 bp DNA ladder (Invitrogen, Karlsruhe, Germany) and Smartladder (Eurogentec, Cologne, Germany). After electrophoresis, the DNA was viewed under UV-illumination and documented with the Gel Doc XR Gel Documentation System (BioRad, Munich, Germany).

2.2.5. Elution of DNA fragments from agarose and PCR clean-up

DNA fragments were excised and eluted from agarose gels using Wizard® SV Gel and PCR clean-up system from Promega, according to the supplier's procedure. When desired, PCR products were also purified using the same kit system. The purified DNA was finally eluted in 25 µl nuclease-free water.

2.2.6. Measurement of DNA/RNA concentrations

2.2.6.1. PicoGreen® assay

For genotyping purposes, the concentrations of all DNA samples were quantified with PicoGreen® assay (Ahn *et al.*, 1996; Rengarajan *et al.*, 2002). PicoGreen is a very sensitive fluorescent dye with very low own fluorescence used for quantitative assays of double-stranded DNA (dsDNA) in solution. This dye enables the detection of as little as 25 pg/ml of dsDNA (Singer *et al.*, 1997). Using a single dye concentration, the PicoGreen assay has a detection range extending from 25 pg/ml to 1 µg/ml dsDNA. While free dye is essentially nonfluorescent, it exhibits >1000-fold fluorescence enhancement upon binding to dsDNA (with excitation and emission maxima of ~500 nm and ~520 nm, respectively) (Singer *et al.*, 1997). The fluorescence enhancement of the PicoGreen is exceptionally high; little background occurs since the unbound fluorophore has virtually no fluorescence. Moreover, PicoGreen is very stable to photobleaching, allowing longer exposure times and assay flexibility (Ahn *et al.*, 1996).

Using a TECAN pipetting robot (Genesis RSP 150), full automation of pipetting, dilution, and normalization steps were established at ICMB (Kiel, Germany). The measurement of 96 DNA samples in parallel was formulated with the help of an in-house implemented software, namely SampleTool. Thirty-two DNA samples were arranged in duplicate in a 96-well optical

Sarstedt plate. Therefore, the worktable of the robot had capacity for three of such plates. An adjusted protocol was applied to measure DNA concentrations with PicoGreen® dsDNA quantification reagent. The DNA samples and PicoGreen reagent were left to equilibrate to room temperature. The PicoGreen reagent was diluted with 1x TE buffer. The measurement plates were prepared to contain 4 different standard DNA solutions (1, 10, 100 and 500 ng/ml), and 32 DNA samples in 1:40 dilution, in addition to negative controls and blank wells (1x TE buffer). One-hundred microliters from the diluted PicoGreen solution was then added to each well in the measurement plates, which contained 1:400 diluted DNA samples. Contents of all wells were well-mixed and incubated in a dark place for 5 min. Using a TECAN Spectrafluor Plus Fluorometer, DNA concentrations of the contents of each measurement plate were obtained with excitation wavelength 485 nm. The raw data files were exported for each plate separately. Concentrations and dilution factors were calculated using the average of the two measurements for each DNA and a standard curve. Normalized DNA concentrations were verified by a second measurement. If single tubes were measured without SampleTool, a special script was used. Depending on these data, required dilutions and volumes for subsequent whole-genome amplification were optimized and utilized.

2.2.6.2. NanoDrop assay

The concentrations of small volumes preparations, such as RNA, cDNA, plasmid DNA, and proteins, were measured in the present study by applying a 'NanoDrop' technique using either NanoDrop® ND-1000 Spectrophotmeter or IMPLEN Nanophotemeter. Two µl of diluted sample was pipetted directly onto the active measurement window of the device and directly measured against 2 µl nuclease-free water as blank. Optical density (OD) ratio (A260/A280 and A260/A230) was also measured every time for purity estimation. Pure DNA and RNA were typically had A260/A280 ratios between 1.8 and 2.10. For pure RNA samples, i.e. free from genomic contamination, displayed ratio (A260/A230) values >2.0.

2.2.7. Polymerase chain reaction

Polymerase chain reaction (PCR), a method that rapidly produces numerous copies of a desired piece of DNA, was widely used for many purposes in the present study. For assays (SNPs) that did not fit to the design of high-throughput genotyping methods (SNPlex/TaqMan), direct genomic sequencing of PCR products were carried out using the mother protocol and thermal cycler settings outlined in Table 2.1 and Table 2.2, respectively.

Here, specificity was enhanced by using a touchdown thermoprofile as previously recommended (Don *et al.*, 1991). To determine the optimal primer annealing temperature (the so-called, primer optimization step), a gradient (12°C across 12 positions) PCR was carried out using similar reaction mix (Table 2.1); the PCR products were then loaded an onto 1.5% agarose gel for electrophoresis, and the best annealing temperature of each tested primer pair was chosen and applied.

Table 2.1 PCR protocol for direct genomic sequencing purposes

Component	Volume (µl)/reaction	Final Concentration
GeneAmp® 10x PCR buffer II	2.50	1X
$MgCl_2$ [25 mM]	2.00	2 mM
dNTPs [10 mM]	0.50	200 µM
Forward primer [10 µM]	1.00	0.04 µM
Reverse primer [10 µM]	1.00	0.04 µM
PCR-water	16.85	-
Good mixing		
DNA [5 ng/µl]*	1.00	0.75 U
AmpliTaq Gold® [5 U/µl]	0.15	0.03 U/µl
Gentle mixing		
Total volume	25.00	

- *: DNA used here was obtained from 1:5 diluted whole genome amplification (WGA) product (see section 2.4.1).

Table 2.2 Thermal cycling conditions of PCR for direct genomic sequencing

Event	Temperature (°C)	Time	No. of Cycles
Initial melting step/ Taq Polymerase activation	95	5 min	1
Denaturation	95	30 sec	td=-0.5°C/cycle
Annealing of primers	65*	30 sec	
Extension (1kb/min)	72	1 min	Repeat 16 cycles
Denaturation	95	30 sec	
Annealing of primers	57*	30 sec	Repeat 20 cycles
Extension	72	1 min**	
Final extension: filling up the ends (recommended for TA cloning as well)	72	10 min	1
Hold at	4	∞	1
Storage	-20	-	-

- *: the optimal annealing temperature of each primer pair was obtained after performing primer optimization PCR.
- **: elongation time depends on length of amplicon: 1 kb/min.

In order to test for genomic contamination in prepared RNA/cDNA, the so-called G3PDH-PCR was carried out. As a standard, PCR amplification of a housekeeper gene glyceraldehyde-3-phosphate dehydrogenase (G3PDH) from 2 µl of each prepared RNA/cDNA sample was performed with Human G3PDH Amplimers (10 mM) using the

GoTaq polymerase (Promega). The thermal cycler conditions were: initial melting at 96°C for 2 min, then 40 cycles at 96°C for 2 min, 55°C for 30 sec, and 72°C for 1 min, followed by 5 min final extension at 72°C. On the other hand, to select white clones with correct insert size after TA-cloning procedure (section 2.2.9.1), PCR with M13 universal primers (M13-F: GTAAAACGACGGCCAGTG; M13-R: AACAGCTATGACCATG) was carried out using standard protocol of Taq DNA Polymerase (Qiagen; 5 U/µl). The thermal conditions were: initial melting at 95°C for 5 min, then 30 cycles at 95°C for 1 min, 53°C for 1 min, and 72°C for 2 min, followed by 10 min final extension at 72°C. The DNA used here was extracted by heating 5 µl from each grown white clone in a PCR machine at 95°C for 10 min. Five µl from each M13-PCR product was then applied to a 1.8% agarose gel for electrophoresis. After cDNA preparation (section 2.2.3), reverse-transcription (RT)-PCR technique was carried out for transcript detection (Table 2.3 and Table 2.4). In each round of amplification, positive and negative controls were included to better monitor transcript expression and to verify the performance of reagents as well.

Table 2.3 RT-PCR general protocol

Component	Volume (µl)/reaction
Nuclease-free water	15.80
GoTaq green buffer [5X]	5.00
dNTPs [10 mM]	1.00
Forward primer [10 µM]	1.00
Reverse primer [10 µM]	1.00
Good mixing	
cDNA	1.00
GoTaq polymerase [5 U/µl]	0.20
Gentle mixing	
Total volume	25.00

Table 2.4 Thermal cycling conditions for general RT-PCR protocol

Event	Temperature (°C)	Time	No. of Cycles
Initial melting step/ Taq Polymerase activation	96	2 min	1
Denaturation	96	30 sec	Repeat 26 cycles
Annealing of primers	55*	30 sec	
Extension	72	1 min**	
Final extension	72	5 min	1
Hold at	4	∞	1
Storage	-20	-	-

- *: annealing temperature differs from one primer pair to another.
- **: elongation time depends on length of amplicon: 1 kb/min.

2.2.8. Digestion of DNA with restriction endonucleases

It is well-known that restriction endonucleases recognize short DNA sequences and cleave double-stranded DNA at specific sites within or adjacent to their recognition sequences. The basic single- and double-digestions were carried out at 37°C for 1-2 hours in an end volume of 10 (or 50) µl using appropriate buffer(s) conditions, as recommended from the supplier (New England Biolabs).

2.2.9. Cloning

2.2.9.1. TA Cloning for PCR products

PCR products were cloned using the pCR 2.1 TOPO TA Cloning Kit (pCR® II, pCR® 2.1) from Invitrogen according to standard procedure. This cloning system is based on the fact that *Taq* polymerase has a nontemplate-dependent terminal transferase activity that adds a single deoxyadenosine (A) to the 3´ ends of PCR products. The linearized vector supplied in this kit has single, overhanging 3´ deoxythymidine (T) residue. This allows PCR inserts to ligate efficiently with the vector (Zhou *et al.*, 1995). In order to check for white clones with correct insert size, PCR with M13 universal primers was carried out as outlined in section (2.2.7). Corresponding M13-PCR products, which showed correct insert size from at least 30 clones, were sequenced according to standard procedure described below in section (2.2.12). The resulting sequence traces were aligned and analyzed using Sequencher (version 4.5) software, followed by manual verification of the alignments.

2.2.9.2. Cloning using T4-DNA ligase

The standard ligation reaction used 100 ng of vector. In order to calculate the best insert to vector ration, which is 3:1, the following formula was applied: [ng of insert= ((size of insert x 100 ng of vector)/size of vector) x 3]. To control insert and vector concentrations after restriction digestion, they were applied to agarose gel electrophoresis parallel to smartladder. The DNA content was then eluted from gel with the Promega's kit system described above. Thereafter, the standard ligation reaction was done using 1 µl from each ligation buffer (10 x) and T4-DNA ligase, corresponding volumes of insert and vector to fit with the 3:1 ration, and the reaction was equilibrated to final volume of 10 µl with PCR-water. The ligation reaction was then incubated overnight at 14°C. Quick ligation, when required, was also achieved by incubating the ligation reaction at room temperature for 2-3 hours.

2.2.9.3. Transformation

Five µl of each ligation reaction was then gently pipetted into 25 µl of *Escherichia coli* TOP 10 competent cells and incubated on ice for 30 min. The reaction was then incubated, without mixing or shaking, for exactly 30 seconds in the 42°C water bath (heat shock step). The transformed mixture was immediately placed again on ice for at least 3 min. Next, 250 µl of pre-warmed S.O.C. medium was added to each reaction tube, and the mixture was incubated at 37°C for 1-2 hours at 200 rpm in a shaking incubator. The content of each transformation vial was spreaded on separate, labelled LB agar plates containing appropriate antibiotics. The plates were inverted and incubated at 37°C overnight. For each plate, 30 clones were picked and individually overnight cultured in 3 ml LB medium with the same antibiotics at 37°C with shaking. Plasmid DNAs were then isolated (section 2.2.11).

2.2.10. Site-directed mutagenesis

Site-directed mutagenesis was performed using QuikChange® Lightning Site-Directed Mutagenesis Kit (Stratagene), according to supplier's protocol. The mutagenic primer design was done using 'Quickchange Primer Design Program' at Stratagene web site and primers were ordered as HPLC-purification grade from Microsynth Laboratory. The basic procedure utilizes a supercoiled dsDNA vector (10-100 ng/reaction) with an insert of interest and two synthetic oligonucleotide primers (125 ng/µl), both containing the desired mutation. Extension of the oligonucleotide primers generates a mutated plasmid containing staggered nicks. Following temperature cycling, each amplified product was treated with 2 µl *Dpn* I at 37°C for 10 min. The *Dpn* I endonuclease (target sequence: 5´-Gm^{6}ATC-3´) is specific for methylated and hemimethylated DNA and is used to digest the parental DNA template and to select for the synthesized DNA containing the mutation. The nicked vector DNA containing the desired mutation is then transformed into XL10-Gold® ultracompetent cells. The transformation and cloning procedure were carried out as described with *E. Coli* Top 10 competent cells (see previous section: 2.2.9.3).

2.2.11. Plasmid DNA purification

Minipreps of DNA plasmids were isolated from 1.5 ml of overnight culture using Wizard® *Plus* SV Minipreps DNA Purification System from Promega, according to the manufacturer's protocol. For fast purification of large-scale transfection grade DNA, plasmid DNA was

purified from 200 ml of overnight culture with the help of the QIA Filter™ Plasmid Maxi Kit from Qiagen, according to the supplier's procedure. The success of either application (Mini/Maxi) was checked by measuring the concentration of 2 µl of the isolated plasmid DNA using the IMPLEN Nanophotometer (see section 2.2.6.2).

2.2.12. DNA Sequencing

For DNA sequencing, the BigDye® Terminator v1.1 Cycle Sequencing Kit from Applied Biosystems was used in the present study. Eight µl of PCR product were transferred to a 96-well Costar plate and subjected to an enzymatic digestion ('digest step'), in order to remove primer-dimers, superfluous primers and free dNTPs. Highly concentrated PCR products, determined by agarose gel electrophoresis (section 2.2.4), were diluted 1:5 with PCR-water before the digestion reaction. Two µl of the digest mixture, which consisted of 0.30 µl *SAP* (Shrimp Alkaline Phosphatase; 1U/µl), 0.075 µl *ExoI* (Exonuclease I; 20U/µl) and 1.625 µl DDW, were then added. The digest reactions were then incubated at 37°C for 15 min. Here, *ExoI* digests single-stranded DNA molecules (Berthold and Geider, 1976) and remaining dNTPs were also destroyed, since a dephosphorylating *SAP* was used (Sauer *et al.*, 2000). Thus, a potential dNTP/ddNTP imbalance, which could disturb the sequencing reaction, was avoided. The digestion reaction was then stopped by heating at 72°C for 15 min. The heating step inactivates SAP that could negatively affect the subsequent sequencing reaction. Next, in a separate 96-well plate, 2 µl of the digested product was mixed with 8 µl sequencing mixture (Table 2.5). Separate sequencing reactions were performed for forward and reverse primer, in order to confirm sequences as well as to monitor artefacts. The thermocycling settings, as outlined in Table 2.6, were applied for sequencing reaction. For isolated plasmid DNA as well as DNA purified from agarose gel, the 'digest step' was skipped and sequencing reaction was directly carried out following the protocol outlined in Table 2.7, with the same thermocycler settings provided in Table 2.6.

Table 2.5 Components for DNA sequencing reaction

Component	Volume (µl)
Water (HPLC grade)	4.80
Sample Buffer [5x]	1.50
Sequencing forward or reverse primer [3.2 pmol/µl]	1.00
Big Dye™ Terminator Ready Reaction Mix [v1.1]	0.70
Total volume	8.00

Table 2.6 Sequencing Thermoprofile

Event	Temperature (°C)	Time	No. of Cycles
Initial melting step	96	1 min	
Denaturation	96	10 sec	25 cycles
Optimized annealing temperature of nested primer	55*	5 sec	
Chain termination reaction	60	4 min	
Pausing at	10	∞	1
Storage	-20	-	-

- *: The BigDye reaction is optimized to a temperature range of 50 to 60 °C, in order to avoid any premature termination stops above 60°C (Wen L., 2001).

Table 2.7 Components for Plasmid DNA sequencing reaction

Component	Volume (µl)
Water (HPLC grade)	3.50
Plasmid (100 ng)	3.00
Sequencing forward or reverse primer [4.8 pmol/µl]	1.00
Sample Buffer [5x]	1.00
Big Dye™ Terminator Ready Reaction Mix [v1.1]	1.50
Total volume	10.00

To generate high quality DNA sequence data, unincorporated dye terminators must be removed from sequencing product prior to capillary electrophoresis. MultiScreen Separation Plates, combined with easy column-loading and packing protocols, provide a very cost-effective and high performance means for parallel processing of 96 sequencing reactions by gel filtration. In the present study, excess Fluorescent Dye Terminator, primers and unincorporated nucleotides were removed from the sequencing reactions based on gel filtration using modified cross-linked dextran, namely G-50 Sephadex Spin Columns. The Sephadex gel filtration matrix was prepared according to the standard protocol using a 96-well Multiscreen Column loader (MAHVN 4550). The sequencing products were diluted with 20 µl DDW and applied to the prepared Sephadex plate. The filtration plate was fitted on top of a MicroAmp® Optical 96-well reaction plate and the sequencing products were eluted through the filtration column into the MicroAmp® Optical 96-well by centrifuging at 2100 rpm for 5 min. Finally this plate was sealed with an aluminum adhesive cover, to prevent contamination, and analyzed on a 3730xl DNA Analyzer.

2.2.13. Transfection of cultured HeLa cells using FuGene 6 Reagent

HeLa S3 cervical cancer cell line (ACC 161) from the German Collection of Microorganisms and Cell Cultures (DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany) was cultured as monolayers and maintained in RPMI 1640

(PAA Laboratories GmbH, Pasching) containing 10% fetal bovine serum (PAA Laboratories GmbH, Pasching), 100 U/ml penicillin and 100 µg/ml streptomycin (PAA Laboratories GmbH, Pasching) in a humidified atmosphere with 5% CO_2 at 37°C. To prepare cells for transfection, adherent cells to be subcultured were first washed twice with 2 ml 1X PBS (Phosphate-Buffered Saline) and trypsinized (2 ml 1X sterile trypsin-EDTA solution; enough to cover the cell monolayer) to loosen adherent cells from the growth surface. Five minutes at 37°C incubator was enough in most cases to detach all cells (microscopic examination). After detaching, 2 ml RPMI 1640 medium with 10% FCS was added to the cells to inactivate the trypsin. After centrifugation for 5 minutes at 1000 rpm, the supernatant was decanted and cells were resuspended again in 2-10 ml of the same medium depending on the size of the cell pellet. The cells were then diluted 1:25 (1 µl cell suspension: 24 µl Trypan blue/PBS) and counted using a hemocytometer (Brand, Germany), according to standard procedure. The cell count per ml was obtained from the following formula: (total number counted/4) x 50 x 10,000. A total of 0.4- 0.8 x 10^6 cells were plated per well in a 6-well plate for ~80% confluency the day of transfection.

On the next day of plating, the transfection was performed using a cationic lipid-based transfection reagent (FuGENE 6; Roche) that complexes with DNA and transports it into the cell during transfection. For each well, transfection was done using 3 µl FuGENE 6 reagent in a total volume of 100 µl serum-free RPMI 1640 medium. One microgram vector DNA solution (0.5-50 µl) was then added to the pre-diluted FuGENE 6 reagent from the previous step. The tubes were very gently mixed by tapping. After 15 minutes incubation at room temperature, the content of each complex mixture (~100 µl) was dropwise pipetted to the corresponding well in the 6-well plate. The transfected cells were then returned to incubator at 37°C. All of these steps were performed under sterile conditions (in a laminar flow hood). Typically, 24-30 hours post-transfection, cells were rinsed and harvested in each time with 1 ml cold 1X PBS.

2.2.14. Protein lysate preparation and Western blotting

Protein lysate preparation. Lysed protein extracts were prepared from transfected HeLa cells by boiling the tissue homogenates for 5 min in denaturing extraction buffer containing 1% SDS, 10 mM Tris (pH 7.4), and 1% phosphatase inhibitor mixture II (Sigma-Aldrich Chemie, Germany). After sonicating twice (Bandelin Sonoplus GM 70) for 5 seconds, insoluble material was removed by centrifugation for 15 min at 16,000 X *g* at 4°C (Waetzig *et*

al., 2002). Protein extracts were snap-frozen in liquid nitrogen and stored at - 80°C. For all western blotting experiments, total protein was determined using 'DC-Protein Assay' from BioRad. Thereafter, protein concentrations were directly measured at 750 nm (IMPLEN Nanophotemeter) from a calibrated standard curve using the *Lowry* method parameters.

Western blot analysis. Western blotting experiments were mainly performed as described by Waetzig and co-workers (Waetzig *et al.*, 2002) with minor changes. Protein extracts (standardized to 10 µg of total protein in 1X SDS-loading buffer/lane) were separated by 12% denaturing SDS-PAGE (Multigel Whatman Biometra; 30 min at 20 mA for stacking gel and 1 hr at 50 mA for separating gel) and transferred to a polyvinylidene difluoride 'PVDF' membrane (Hybond-P; Amersham Pharmacia Biotech) by semidry blotting (80 min at 40 mA: 0.65 mA/cm^2) using an electroblotter (PeqLab Biotechnologie). Following transfer, membranes were blocked at 4°C overnight with 5% non-fat milk proteins suspended in 1X TBST on a Roller Mixer (Stuart). These inert milk proteins bind to the unoccupied membrane sites (without displacing the target protein from the membrane), thereby blocking non-specific binding of antibodies to PVDF membrane. After blocking step, the PVDF membrane was incubated for 2 hrs with primary antibody (DsRed monoclonal antibody (1:500) or *Aequorea victoria* (A.v.) GFP monoclonal antibody (JL-8) (1:1000), Clontech). After being washed in 1X TBST (three times for 15 min), membrane was incubated for 30 min with a HRP-conjugated secondary antibody (Rabbit polyclonal to Mouse IgG antibody - H&L (HRP) (ab6728) (1:2000), Abcam). Membranes were subsequently washed 3 times with TBST (three times for 15 min), incubated with ECL-Plus Detection Reagent (3,9 ml solution A plus 100 µl solution B; Amersham Biotech) for 3-5 min. Western blots were documented using Bio-Rad ChemiDoc XRS System, and pictures were taken using its CCD camera**.** As an internal control, PVDF membranes were stripped by a 4 min incubation at room temperature in 10 ml mixture of 0.1% SDS and 0.2% M NaOH, and then the blots were probed again with 1:1000 mouse anti-β-actin monoclonal antibody (Clone AC-15, Sigma).

2.2.15. Fluorescent Activated Cell Sorter (FACS)-Analysis

In the present study, fluorescence-activated cell sorter (FACS) analysis was carried out using a FACSCalibur™ cytometer (Becton-Dickinson, San Jose, CA) with filters suitable for both enhanced green and red fluorescent proteins (EGFP and RFP, respectively). Excitation of both proteins resulted from use of an air-cooled argon-ion laser (Excitation; 488 nm), detection parameters were set to the maximum emission wavelength for EGFP (Fluorescence channel

FL1, Emission: 530±15nm) or RFP (Fluorescence channel FL2, Emission: 585±15nm), respectively. For FACS analyses, HeLa cells were transfected as described (section 2.2.13) and after 24h cells were washed in ice-cold PBS and transferred into cytometer tubes in a final volume of 500 µl PBS. Untransfected HeLa (mock) cells were used at first to equilibrate the FACS system and to define cell-type specific instrument settings. Fluorescence was separately detected in channels FL1 (for EGFP) and FL2 (for RFP) in 20,000 cells/sample using a measurement speed of <1000 cells/sec. Fluorescence intensity for both individual parameters was quantified by applying the CellQuestPro^{TM} software package (Beckton Dickinson).

2.3. Generation of matching DNA- cDNA pairs

2.3.1. Recruitment

The analysis of allelic splicing required a resource of cDNA from individuals with pre-determined genotypes. Thus at the beginning of the present study, a total of 170 matching pairs of DNA and cDNA were extracted from either whole blood (N=25), lymphoblastoid cell lines (N=135) or brain tissue (N=10) obtained from surgical specimens. Proband recruitment protocols were approved by the institutional ethics committees at all participating institutions and written informed consent was obtained from all participants prior to the study. Samples were anonymized according to the local data protection regulations.

Lymphoblastoid cell lines: A resource of 135 lymphoblastoid cell lines from normal individuals (N=58) and patients with inflammatory bowel disease (N=54) and sarcoidosis (N=23) was generated. Lymphoblastoid cell lines were obtained from B lymphocytes, which can be grown indefinitely in the laboratory after special treatment of the cells with Epstein-Barr virus (EBV) (Cavalli-Sforza, 2005). Human B lymphocytes have a receptor for EBV and once infected, they can become immortalized with a high rate of success to produce a cell line. EBV remains *episomal* (nonintegrated) and therefore does not alter the endogenous genome (Strachan and Read, 2004).

Corresponding resource of human leucocytes cDNA and DNA: As part of the regional POPGEN biobank project (Krawczak *et al.*, 2006), 25 normal healthy controls were used to obtain corresponding samples of DNA and cDNA from peripheral blood. This resource served to minimize potential artefacts introduced through the viral transformation in the cell lines.

The area from which these normal control individuals came is thought to consist of a very homogenous Northern German population (Krawczak *et al.*, 2006).

Corresponding resource of human brain cDNA and DNA: In cooperation with the Department of Neuropathology, University of Bonn (Prof. Dr. Albert Becker), we obtained DNA and cDNA from 10 human brain samples derived from stereotactic surgery. Brain tissue biopsies were obtained from patients with chronic pharmacoresistant temporal lobe epilepsy in the Epilepsy Surgery Program at Bonn University. Surgical removal of epileptogenic tissue was necessary to achieve seizure control in all patients after standardized pre-surgical evaluation using a combination of non-invasive and invasive procedures (Kral *et al.*, 2002). This resource was used to address potential tissue-specificity of the observed splicing effects.

2.3.2. Plate layout

For each sample, in addition to a label on the lid, each 2 ml tube received a barcode for tracking it in the laboratory information management system (LIMS), which is an in-house available database used to systematically store and retrieve information at ICMB. To be suitable for subsequent genotyping and RT-PCR analyses, these samples were arrayed in labelled 96-well microtiterplates (MTP). An overview of the corresponding DNA/cDNA pairs of control and patients samples recruited in this study is provided in Figure 2.1. Four wells were used for internal controls and quality control. These included three empty wells (no template controls [NTCs]), one positive control and the so-called CEPH cell-line control (Dausset *et al.*, 1990). Negative controls were used to reveal potential contamination.

Control Plate

	1	2	3	4	5	6	7	8	9	10	11	12
A	ZLN3001	ZLN3009	ZLN3017	ZLN3023	ZLN3031X	ZLN3040	ZLN3048	ZLN3057	Muc0005	Muc0013	Muc0020	
B	ZLN3002	ZLN3010	ZLN3018	ZLN3024	ZLN3032X	ZLN3041	ZLN3049	ZLN3058	Muc0006	Muc0014	Muc0021	
C	ZLN3003	ZLN3011	positive	ZLN3025	ZLN3033X	ZLN3042	ZLN3051	ZLN3059	Muc0007	Muc0015	Muc0022	
D	ZLN3004	ZLN3012	Empty	ZLN3026	ZLN3035	ZLN3043	ZLN3052	ZLN3060	Muc0008	Empty	Muc0023	
E	ZLN3005	ZLN3013	ZLN3019	ZLN3027	ZLN3036	ZLN3044	ZLN3053	Muc0001	Muc0009	Muc0016	Muc0024	
F	ZLN3006	ZLN3014	ZLN3020	ZLN3028	ZLN3037	ZLN3045	ZLN3054	Muc0002	Muc0010	Muc0017	Muc0025	
G	ZLN3007	ZLN3015	ZLN3021	ZLN3029	ZLN3038	ZLN3046	ZLN3055	Muc0003	Muc0011	Muc0018		
H	ZLN3008	ZLN3016	ZLN3022	ZLN3030	ZLN3039	ZLN3047	ZLN3056	Muc0004	Muc0012	Muc0019		Empty

ZLN-	cell-lines normal controls (N=58)
Muc-	Blood normal controls (N=25)
positive	CEPH postive control
Empty	Empty well

Patient Plate

	1	2	3	4	5	6	7	8	9	10	11	12
A	EB1029/02	EB1921/03	ZLI1009	ZLI1015	ZLI1027	ZLI1038	ZLI1048	ZLI1057	ZLI1075	ZLS2009	ZLS2021	ZLS2041
B	EB2573/02	EB2330/03	ZLI1010	ZLI1016	ZLI1028	ZLI1039	ZLI1049	ZLI1058	ZLI1077	ZLS2010	ZLS2023	ZLS2045
C	EBN2629/02	ZLI1001	positive	ZLI1018	ZLI1031	ZLI1041	ZLI1050	ZLI1061	ZLS2003	ZLS2012	ZLS2025	
D	EB1425/02	ZLI1002	Empty	ZLI1019	ZLI1032	ZLI1042	ZLI1052	ZLI1062	ZLS2004	Empty	ZLS2026	
E	EB2348/03	ZLI1003	ZLI1011	ZLI1023	ZLI1033	ZLI1043	ZLI1053	ZLI1068	ZLS2005	ZLS2015	ZLS2027	
F	EB2556/03	ZLI1005	ZLI1012	ZLI1024	ZLI1034	ZLI1044	ZLI1054	ZLI1069	ZLS2006	ZLS2017	ZLS2035	
G	EB59/04	ZLI1006	ZLI1013	ZLI1025	ZLI1035	ZLI1045	ZLI1055	ZLI1070	ZLS2007	ZLS2019	ZLS2039	
H	EB1899/03	ZLI1007	ZLI1014	ZLI1026	ZLI1037	ZLI1047	ZLI1056	ZLI1074	ZLS2008	ZLS2020	ZLS2040	Empty

EB-	Brain samples (N= 10)
ZLI-	IBD cell line samples (N= 54)
ZLS-	Sarcoidosis cell line samples (N= 23)
positive	CEPH postive control
Empty	Empty well

Figure 2.1 Plate layout of matching DNA and cDNA controls and patients samples labelled with ICMB-specific codes.

2.3.3. Quality control checkups

After recruitment of samples, matching pairs of cDNA and genomic DNA (gDNA) were prepared from the respective tissue as described above (section 2.3.1). To ensure the quality of the prepared gDNA and cDNA, all the quality control checkups were carried out. The quality of all extracted DNA samples was checked on agarose gels (section 2.2.4), quantified using PicoGreen® assay (section 2.2.6.1), and used for whole genome amplification. One intact band with high molecular weight was observed for qualified DNA, while a smear or weak bands indicated degraded DNA or low DNA concentration (i.e., failure of DNA extraction). This checkpoint was very important, since degraded DNA or a DNA concentration below 10 ng/µl was not suitable for whole-genome amplification using multiple displacement reaction. The isolated RNA samples were also subjected to quality control checkups as well. These included quantification using NanoDrop technique and G3PDH-PCR test for genomic contamination (section 2.2.7). The isolated RNA was reverse-transcribed into complementary DNA (cDNA) using random hexamers (section 2.2.3) and the success of the first strand cDNA synthesis was also checked using PCR amplification with G3PDH.

2.4. Whole-genome amplification and genotyping

2.4.1. Whole-genome amplification

To overcome the restrictions of poor DNA yield and limited amounts of available samples, which was not suitable for wide-scale genotyping experiments, a new whole genome amplification (WGA) method, namely GenomiPhi DNA amplification kit (Amersham Biosciences), was used in the present study. This method produced microgram quantities of high molecular weight DNA from as little as 1 ng of genomic DNA. The GenomiPhi assay utilizes bacteriophage Phi29 DNA polymerase from *Bacillus subtilis* - a unique, highly processive enzyme with excellent strand displacement activity- in combination with random-sequence hexamer primers to amplify DNA in an isothermal (30°C) process. Therefore, in standard displacement reaction, thermal cycling was not required (Lizardi *et al.*, 1998; Dean *et al.*, 2001). Because of the proofreading activity of Phi29 DNA polymerase (Esteban *et al.*, 1993; Nelson *et al.*, 2002), DNA replication in this method is extremely accurate.

The GenomiPhi method relies on the multiple displacement amplification (MDA) WGA reaction that was first described by (Dean *et al.*, 2002), and has been recently considered

developed technique for high performance WGA (Lovmar and Syvanen, 2006). The basic idea of MDA is that, the random hexamers anneal to the single stranded target molecule and as the DNA polymerase elongates the primer, the upstream DNA strands are displaced. The displaced DNA strands can then serve as templates for new priming events, which results in primer elongation in the opposite direction. The MDA reaction continues, and new DNA strands are displaced to produce new templates and a hyperbranched structure, generating an abundance of copies of the original DNA molecule (Lovmar and Syvanen, 2006). Varying DNA concentrations in the initial sample will plateau during MDA, which is a potential benefit for MDA in large-scale genotyping applications because it unifies and increases the DNA concentrations of the samples (Lovmar and Syvanen, 2006). The performance of MDA is dependent upon the quality of the input DNA (Lage *et al.*, 2003). In this regard, the quality of the DNA yield from degraded DNA templates is poor and often not suitable for genotype analyses. A degraded DNA template has fewer primer binding sites per DNA molecule for initiation of replication, and will thus undergo fewer hyperbranching events. In this case, the high processivity of the MDA reaction will not be fully utilized, which lowers the yield (Lovmar and Syvanen, 2006). For that reason, the extracted DNAs (section 2.2.1) were checked on agarose gel to insure their integrity.

It was highly recommended to not use more than 1-2 µl DNA volume in the Phi29 reaction. Therefore, DNA samples with higher concentration were diluted with 0.1x TE buffer to final concentrations of 10-30 ng/µl. Two different versions of GenomiPhi kits, v1 and v2, were used for amplification of DNA. While version 1 is an overnight reaction, version 2 produces the same yield after two hours of reaction time. Moreover, there is no random amplification in empty wells thus no artefacts are generated. The procedure generated fragments between 10 and 100 kb long. Briefly, 1 µl of a template DNA to be amplified was added to 9 µl of sample buffer on ice, and then heated to 95°C for 3 min to denature the template DNA. Afterwards, the sample was immediately cooled, mixed with 9 µl of reaction buffer and 1 µl of enzyme mix, and incubated at 30°C (16-18 hours for Kit-v1 or 2 hours for v2). After amplification, Phi29 DNA polymerase was heat-inactivated in a 10 min incubation at 65°C. The quality of the amplified DNA was checked by agarose gel electrophoresis, to be sure that the WGA reaction had worked and resulted in one intact high molecular weight band (not a smear) for each reaction. The resultant wgaDNA was quantified using PicoGreen® assay, in order to calculate the suitable volume from each sample to be used in downstream genotyping assays. Finally, the final 20 µl (~5 µg) reaction volume was diluted 1:5 with 1x TE-buffer (a final

volume of 100 µl; ~50 ng/µl). The 100 µl was split (2x 50 µl) into two fresh 96 well MT plates: one plate was used for SNPlex genotyping and the other for TaqMan plate production.

2.4.2. Genotyping of amplified DNA samples

As the experimental program was critically dependent on the availability of heterozygote transcript carriers, the DNA was used for cost-efficient genotyping and selection of individuals to reduce the number of cDNA-based experiments. The amplified DNA samples were genotyped in the present study using the high-throughput SNP genotyping platform at the ICMB, which was supported by an in-house LIMS and automation system (Hampe *et al.*, 2001; Teuber *et al.*, 2005) and in-house expertise to develop project-specific software for data analysis (Hampe *et al.*, 2001). Ten nanograms of amplified genomic DNA were dried in 96-well plates and genotyped using either SNPlex technology or TaqMan (Applied Biosystems, Foster City, CA), depending upon technical feasibility (Appendix Table 8.1). For assays that failed, direct genomic sequencing with automated calling (Manaster *et al.*, 2005a) was carried out. For the direct sequencing method, PCR was performed using 5 µl from each diluted wgaDNA (1ng/µl) (section 2.2.7).

2.4.2.1. SNPlex™ Genotyping: An advanced high-throughput technology

In the present study, the selected candidate splice SNPs were mostly genotyped using the recently developed SNPlex™ genotyping system. This system represents an attractive alternative to existing genotyping methodologies, as it requires only three unlabeled probes per SNP, consumes very little genomic DNA, can be highly multiplexed, and uses widely available capillary electrophoresis (CE) instruments. The SNPlex™ assay (Figure 2.2) involves eight main steps (De la Vega *et al.*, 2005; Tobler *et al.*, 2005), which were performed in three consecutive days, according to the established protocol at ICMB. The only SNP-specific components of the assay were the ligation probes that participate in the oligonucleotide ligation (OLA), which was the key allele-discriminating step. Currently, up to 48 SNPs can be addressed simultaneously in one OLA reaction over a 384 well MTPs. The SNPlex™ system is based on an OLA-PCR assay with a universal ZipChute™ probe detection for high-throughput SNP genotyping. In this method, fluorescently labeled ZipChute™ probes are hybridized to complementary ZipChute™ sequences that are part of genotype-specific amplicons. These ZipChute™ probes are then eluted and detected by electrophoretic separation on Applied Biosystems 3730 or 3730xl DNA Analyzers.

In experiment, the wgaDNA was first fragmented for 5 min, diluted 1:2 with 1x TE-buffer to a final volume of 100 µl (~25 ng/µl), and then aliquots of 5 µl were dispensed using a 384-channel Robbins Scientific Hydra microdispenser to fresh 384 MT PCR plates. After the plates were left to dry overnight in a closed cupboard, the plates were well-sealed and labelled with a unique barcode for database tracking. At this point, the plates were ready-to-use for SNPlex™ genotyping (preferably used within 6-12 months for optimal results). On the first day of genotyping, after an initial kinase step to phosphorylate linkers and ligation probes, the activated oligonucleotides were combined with fragmented wgaDNA (100–150 ng per well, i.e. 2–3 ng per assay) to perform genotyping in separate reactions. Since the design of the present study depended on the 96-well plate format to shape with the downstream cDNA-arraying and RT-PCR experiments, a modified pipetting script was applied. By this means, DNA samples in 96-well plate format were four times pipetted into a 384-MTPs format, thereby allowing genotyping of four different 48-SNPlex pools in one run, which was cheaper and improved the robustness and throughput of the assay four-fold. The selected candidate SNPs were also frequent enough (≥ 10% heterozygosity) to be enriched in these 96 DNA samples. Next, the OLA reaction was prepared for each 48-SNPlex pool (Table 2.8), transferred to the pre-designed quarter of the 384-well SNPlex plate, and PCR (Table 2.9) was performed overnight.

Table 2.8 Phosphorylating and Ligating Probes to gDNA (OLA reaction)

Component	Single reaction (µl)	210 reactions (µl)
Nuclease-free water	2.30	483.00
Oligonucleotide-Ligation-MasterMix	2.50	525.00
Universal Linkers 48-plex	0.05	10.50
dATP (100x)	0.05	10.50
SNPlex System Ligation Probes	0.10	21.00
Total	5.00	1050.00

- Each SNPlex pool was prepared on a 210-reaction scale and dispensed, in the appropriate quarter of the 384 well plate.

Table 2.9 Running the OLA reactions on the thermal cycler

Step	Step type	Temperature (°C)	Time
1	Hold	48	30 min
2	Hold	90	20 min
3	25 cycles	94	15 sec
		60	30 sec
		51 *3% ramp*	30 sec
4	Hold	99	10 min
5	Hold	4	∞

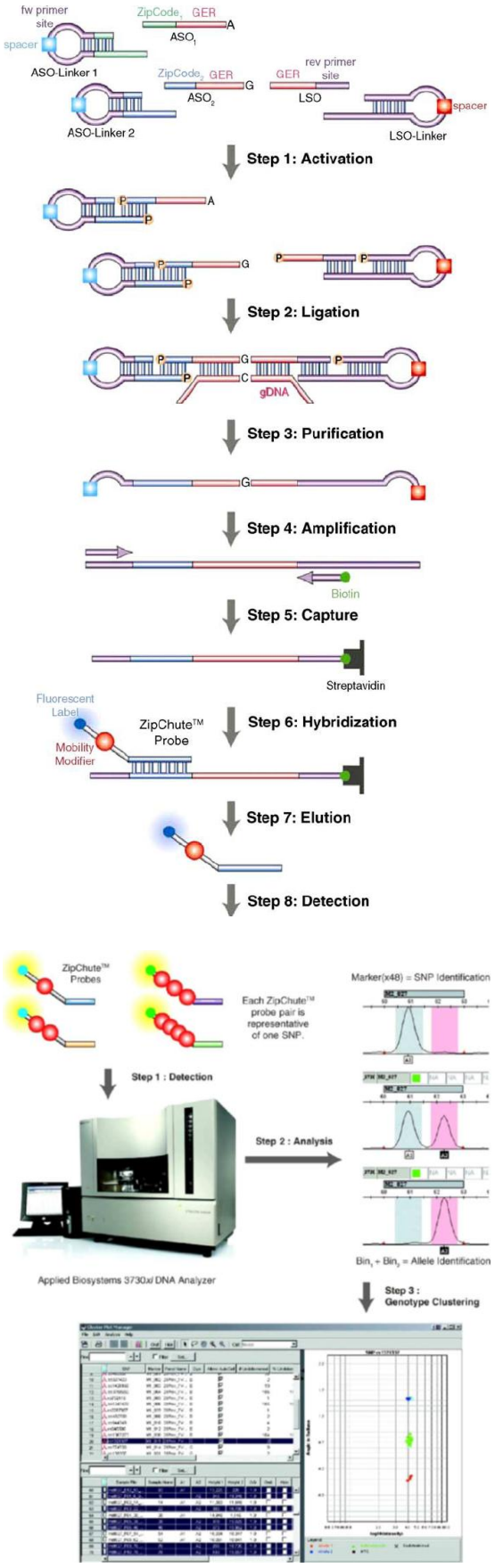

Figure 2.2 SNPlex™ Genotyping system workflow. This involves eight main steps. (1) Activation of ligation probes and linkers by phosphorylation. (2) Ligation of linkers and probes on the genomic DNA target. During allele-specific OLA reaction, allele-specific oligonucleotide (ASO) probes and locus-specific oligonucleotide (LSO) probes hybridize to the genomic target sequence. These allele-specific and locus-specific probes ligate when they are hybridized to a perfectly matching sequence at the SNP site. Simultaneously, universal linkers are ligated to the distal termini of the ASO and LSO ligation probes. These linkers contain universal PCR primer–binding sequences as well as sequences complementary to ASO and LSO probes. A unique ZipCode sequence is attached at the 5′ end of the genomic equivalent sequence within each ASO. Consequently, by virtue of the ZipCode sequence, the OLA step encodes the genotype information of every SNP into unique ligation products. (3) Removal of unligated or incompletely ligated oligonucleotides and genomic DNA by exonucleases. (4) Simultaneous amplification of ligation products by PCR, using a set of universal primers. (5) Capture of biotinylated amplicons on streptavidin-coated plates, and removal of the unbound strand. (6) Hybridization of the universal set of ZipChute probes to the complementary ZipCode product sequence on the captured PCR strand. (7) Release of specifically hybridized ZipChute probes. (8) Detection of fluorescent ZipChute probes by CE; after elution, ZipChute probes are electrophoretically separated on an Applied Biosystems 3730*xl* DNA Analyzer. The intensities of specific signals in each bin of a marker are automatically converted into cluster plots using GeneMapper® Analysis Software. *Illustration adapted from (De la Vega et al., 2005).*

On the second day, unligated and incompletely ligated oligonucleotides, as well as the genomic DNA templates, were removed by digestion with exonuclease I and λ-exonuclease. This so-called "purification step" governs the reduction of signal background noise. Five µl of purification master mix, Table 2.10 , was pipetted into each well of the OLA reaction plate. The plate was sealed, shortly vortexed, centrifuged, and transferred to the thermal cycler (Table 2.11). Then, 15 µl of nuclease-free water was added to each well, mixed, spinned down, and subjected to PCR (Table 2.12 and Table 2.13) with two universal primers, one of which was biotinylated. The last PCR step was carried out in the thermal cyclers of the PCR laboratory, and not of the OLA laboratory, to avoid any chance of genomic contamination, as universal primers were used for amplification of ligated OLA reaction products.

Table 2.10 Purifying Ligated OLA Reaction Products

Component	Single reaction (µl)	500 reactions (µl) (For one 384 well plate)
Nuclease-free water	4.20	2100.00
SNPlex Exonuklease Buffer	0.50	250.00
SNPlex Lambda Exonuklease	0.20	100.00
SNPlex Exonuklease I	0.10	50.00
Total	5.00	2500.00

Table 2.11 Thermal cycle program for purification step

Step	Step type	Temperature (°C)	Time
1	Hold	37	90 min
2	Hold	80	10 min
5	Hold	4	∞

Table 2.12 Amplification protocol of ligated OLA reaction products by PCR

Component	Single reaction (µl)	500 reactions (µl) (For one 384 well plate)
Nuclease-free water	2.42	1331.0
SNPlex Amplification MasterMix (2X)	5	2750.0
SNPlex Amplification Primers (20x)	0.5	275.0
Total	7.92	4356.0

Table 2.13 Thermal cycle profile for PCR amplification of ligated OLA products

Step	Step type	Temperature (°C)	Time
1	HOLD	95	10 min
2	30 cycles	95	15 sec
		70	60 sec
5	Ho	4	∞

After PCR, biotinylated amplicons were incubated with streptavidin-coated microtiter plates in the same PCR laboratory, using a special robot-program that mixed 17.5 µl of the hybridization mixture (Table 2.14) with 3 µl PCR-product. These mixtures were incubated for 15 min at room temperature in a shaking incubator (TiMix Control) at 600 rpm. Using another program, the non-biotinylated amplicons were detached by mixing 50 µl of 0.1 M NaOH in each well and direct incubation at room temperature for 5min in the shaking incubator at 800 rpm. Upon removal of the non-biotinylated amplicon strands, 25 µl of a mixture of 102 pre-optimized, universal ZipChute™ probes (Table 2.15) was added to each well for hybridization and decoding of the genotypic information. Of these, 96 ZipChute™ probes corresponded to all 96 possible alleles of the 48 addressable SNPs in the multiplex. The six remaining ZipChute™ probes were needed for internal controls, such as the positive and the negative hybridization control (PHC/NHC). ZipChute™ probes are fluorescently labeled oligonucleotides, with each probe having a unique size (so-called mobility modifiers). The plates were then directly incubated at 37°C in the shaking incubator at 600 rpm for 60 min. After stringent washing, the ZipChute™ probes were eluted using 17.5 µl sample loading master mix (Table 2.16), and incubation time of 10 min at 37°C in the shaking incubator (at 800 rpm). The universal ZipChute™ probes were finally detected by electrophoretic separation following to supplier's recommendations on Applied Biosystems 3730xl DNA Analyzers. An allelic ladder containing all available ZipChute™ probes was analyzed in parallel to correct run-to-run sizing variations. GeneMapper® software was used for analyzing the raw CE data and calling SNP genotypes. Because one SNP is typically characterized by two possible alleles, two fluorescent peaks in a CE electropherogram represent the two alleles of a specific SNP (Figure 2.2). GeneMapper® analysis software assigns individual genotypes, based on the intensity and location of peaks. Auto-calls of GeneMapper® were manually checked for faulty genotype assignments before the data was exported from GeneMapper® and imported into the in-house database 'ibdbase'.

Table 2.14 Hybridization Buffer

Component	Single reaction (µl)	1568 reactions (µl) (For two 384 well plate)
SNPlex Hybridization Buffer	17.491	27425.90
Positive Hybridization Control	0.009	14.100
Total	17.500	27440.00

Table 2.15 ZipChute™ hybridization master mix

Component	Single reaction (µl)	1528 reactions (µl) (For two 384 well plate)
ZipChute Mix	0.05	76.40
Denaturant SNPlex System	11.25	17190.00
SNPlex ZipChute Dilution Buffer	13.7	20933.60
Total	25.00	38200.00

Table 2.16 Sample loading master mix

Component	Single reaction (µl)	1682,285714 reactions (µl) (For two 384 well plate)
SNPlex size standard	0.54	908.40
SNPlex sample loading Reagent	16.96	28531.60
Total	17.50	29440.00

2.4.2.2. TaqMan® genotyping assay: A fluorogenic 5′ nuclease assay

Although SNPlex™ technology is preferred as a high-throughput system, TaqMan® is still very valuable and robust for genotyping a small number of SNPs that failed with SNPlex™ due to assay design failures. The TaqMan® SNP Genotyping Assay is a single-tube PCR assay that exploits the 5′ exonuclease activity of AmpliTaq Gold® DNA polymerase. The assay includes two locus-specific PCR primers that flank the SNP of interest, and two allele-specific oligonucleotide TaqMan® probes (Figure 2.3). These probes have a fluorescent reporter dye at the 5′ end, and a non-fluorescent quencher (NFQ) with a minor groove binder (MGB) at the 3′ end (De la Vega *et al.*, 2005). An intact probe emits minimal fluorescent signal when excited, because the close physical proximity of the 5′ fluorophore to the 3′ quencher causes the fluorescent resonance energy transfer (FRET) effect to quench the fluorescence emitted by the fluorophore. A fluorescent signal is generated when the intact probe, which is hybridized to the target allele, is cleaved by the 5′ exonuclease activity of AmpliTaq Gold® DNA polymerase during each cycle of the PCR reaction. The PCR primers amplify a specific locus on the genomic DNA template, and each fluorescent dye-labelled hybridization probe reports the presence of its associated allele in the DNA sample (Figure 2.3). In each PCR cycle, cleavage of one or both allele-specific probes produces an exponentially increasing fluorescent signal by freeing the 5′ fluorophore from the 3′ quencher. The use of two probes, one specific to each allele of the SNP and labelled with two fluorophores, allows detection of both alleles in a single tube (De la Vega *et al.*, 2005).

TaqMan® probes were labelled with the fluorescent dyes FAM™ (6-carboxyfluorescein) or VIC® (proprietary dye from Applied Biosystems) and with the quencher TAMRA™ (6-carboxytetramethylrhodamine, succinimidyl ester). The passive reference dye ROX (6-carboxy-X-rhodamine, succinimidyl ester) was included in every well for normalization. Fluorogenic probes with an MGB produce enhanced allelic discrimination, because the MGB stabilizes the double-stranded probe template structure, thereby increasing the probe melting temperature (Tm) without increasing probe length (Kutyavin *et al.*, 2000). This provides enhanced mismatch discrimination between these shorter probes, resulting in improved allele specificity. In the present study, most of the performed TaqMan® genotyping assays were Assays-on-Demand, a pre-designed and validated assay format offered directly by the manufacturer. However, if pre-designed assays were not available, Assays-by-Design was ordered, i.e. assays were custom-ordered from ABI according to a user-defined sequence. Both types of assays required no further optimization.

In experiment, 5 µl from wgaDNA was further diluted 1:80 to a final volume of 4 ml (~0.63 ng/µl). Then, 5 µl was pipetted into the corresponding wells of the 96-well TaqMan plate and dried down at 60°C for 2 hours, subsequently sealed, labelled with a unique barcode for database tracking. These dried TaqMan® PCR plates could be kept for two years before use. The reaction components were mixed in a final volume of 5 µl as demonstrated in Table 2.17. Typically, 5 µl of this reaction mix was added to the 96-well plates with the dried genomic DNA either manually or by a TECAN Genesis RSP 150 multipipetting robot. This process was carried out and tracked with the in-house software Pipettor, which was part of the integrated LIMS (Hampe *et al.*, 2001; Teuber *et al.*, 2005). The applied two-step PCR thermal cycling protocol is provided in Table 2.18. The endpoint read of fluorescence was performed with ABI Prism® 7700 Sequence Detection System and allele calling for each assay/plate was done manually to ensure data quality. A call rate of 95% was considered successful. Each successful assay produced three separated clusters/clouds representing the three genotypes, homoyzygotes for allele 1 were shown in red, heterozygotes (12) in green (both dyes are measured), and homozygotes for allele 2 in blue. The genotypes scattered between these specific clouds were considered as undefined genotypes and subsequently excluded from downstream analyses. Assays that did not show a cluster plot with these three recognizable clouds were deemed unsuccessful and genotyped by direct genomic sequencing.

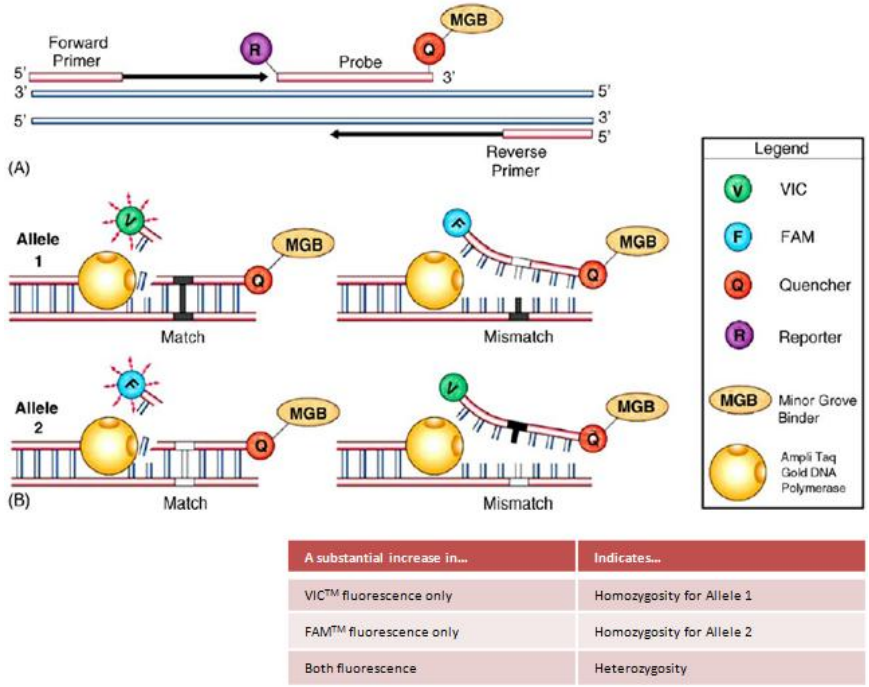

Figure 2.3 Principle of TaqMan® assay.
(A) Probe binding and primer extension in a TaqMan® SNP Genotyping Assay.
(B) Allelic discrimination is achieved by the selective annealing of matching probe and template sequences, which generates an allele-specific signal. *Illustration modified from(De la Vega et al., 2005)*.

Table 2.17 TaqMan® reaction mixture

Component	Assays-on-Demand	Assays-by-design
	Volume (µl)/reaction	Volume (µl) /reaction
TaqMan® master mix	2.500	2.500
Read-to-use-assay mix	0.250	0.063
Water	2.250	2.437
Total volume	5.000	5.000

Table 2.18 Thermal cycling conditions for TaqMan® genotyping

Event	Temperature (°C)	Time	No. of Cycles
Activation of Ampli TaqGold®	95	10 min	1
Denaturation	95	15 sec	45
Annealing, elongation, nucleolytic cleavage of hybridized probes	60	1 min	
Storage	4	∞	1

2.5. Arraying of corresponding cDNA

In principle, the cDNAs were transferred using a TECAN-robot with cooled 96-well plate holders, in order to avoid degradation. In the present study, a customized software, namely

'SpliceTool', was specifically developed to translate the genotypes into the plate position for cDNA-transfer (details in Results section 3.1.2). A schema of genotyping and cDNA-transfer is illustrated in (Figure 2.4).

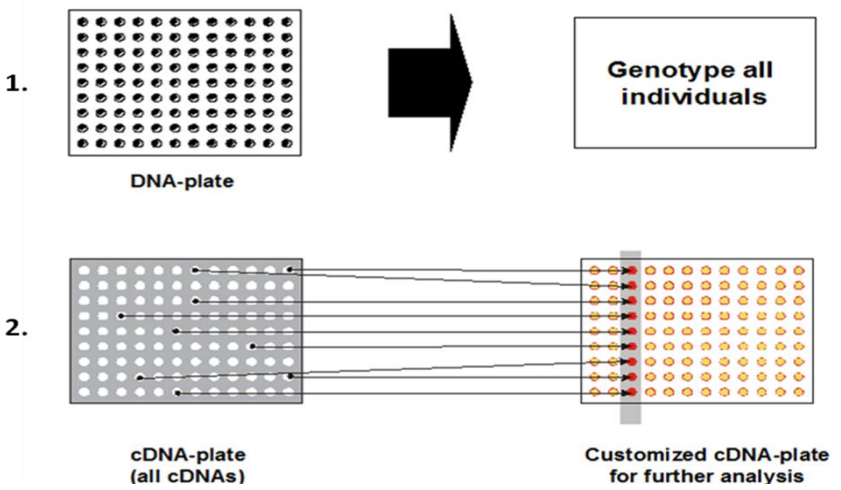

Figure 2.4 Schema of genotyping and cDNA selection.
In **step 1**, two corresponding plates with DNA and cDNA were prepared with an identical position of each individual. In **step 2**, the cDNAs were selected on the basis of the genotypes and transferred to one column of the delivery plate (each delivery plate can thus hold the cDNA for 12 different subsequent sequencing assays).

2.6. Transcript analysis using nested RT-PCR in genotyped cDNA samples

2.6.1. Primer design criteria and semi-automation

To verify the putative splicing effects and to avoid excessive PCR optimization, nested primers were designed (Farrell, 1999) according to the following criteria:

i) Primers were preferably positioned in conserved exons flanking to the candidate SNP.

ii) An additional exon was included on each side of the potentially affected exon if the amplicon length was short enough (600-800 nt long).

iii) At least 50 nt of exonic sequence in the amplification direction was included in the amplification product to create sufficient overlap for subsequent sequence alignment (by SNPSplicer software; section 2.8).

iv) All nested primers were chosen to have similar annealing temperature (around 55-57 °C) in order to be able to perform subsequent RT-PCR experiments in parallel for different candidate SNPs that were occupied the same 96-well plate.

v) Because Repeatmasker, which is a tool for screening submitted DNA sequence against a broad library of repetitive elements, tends to be quite aggressive in its annotation of sequences (Phillips, 2007), when designing nested primers for RT-PCR experiments, it was safer to include masked sequence and then check the resulting designs for specificity in BLAST and right spans of primers in BLAT search (searching if a given primer is unique in the genome).

vi) As SNPs located within probes may affect their hybridization to target DNA sequence (Kwan *et al.*, 2007), all probes containing SNPs were conservatively masked out to circumvent this problem (using SNP-BLAST search at NCBI).

In order to fulfil the above-mentioned designing criteria and make this process fast and precise, another helping software tool, namely "SkippedExonPrimer", was developed (details in Results section 3.1.3). Obtained primers were then manually double-checked using Primer Express software (Applied Biosystems). All primers (Appendix Table 8.1) were produced by Metabion (Martinsried, Germany). Primer stocks were diluted to a concentration of 100 µM with double-distilled water (DDW). These stocks were further diluted to 10 µM, aliquoted, and stored at –20˚C.

2.6.2. Nested RT-PCR

Nested RT-PCR reactions were performed using a chemical hot start enzyme, AmpliTaq Gold® DNA Polymerase. While one microliter of a 1:10 dilution of the reverse transcriptase reaction was used in the first round of amplification, 1 µl of the first round amplification product was amplified in the second-round PCR using the same thermocycling protocol (Table 2.19 and Table 2.20). Here, the specificity of obtained products was enhanced by a touchdown thermoprofile (Don *et al.*, 1991) in a nested PCR protocol (McPherson and Moller, 2000). All the primers used for nested RT-PCRs are provided in Appendix Table 8.1. PCR products from the second round were separated on 1.5% agarose gels and visualized under UV-illumination with the Bio-Rad Gel Doc XR gel documentation system according to the co-migrating DNA-size standards (100 bp DNA ladder). In case of more than one band, i.e. more than one splice variant, separate bands were excised from the gel, and extracted using the Minielute Gel Extraction kit (QIAGEN). Because genotypes were known, 16

different cDNA samples were initially considered for transcript analysis. However, in case of a signal for positive allele-splice effect, further independent cDNA samples were analyzed to confirm the effect.

Table 2.19 Protocol of external and nested round RT-PCR

Component	Volume (µl)/reaction
PCR-water	18.05
GeneAmp® 10x PCR Buffer II	2.50
$MgCl_2$ solution [25 mM]	2.50
dNTPs [10 mM]	0.50
External forward primer [10 µM]	0.150
External reverse Primer [10 µM]	0.150
cDNA 1:10 diluted (or first-round product*)	1.00
AmpliTaq Gold® DNA Polymerase [5 U/µl]	0.15
Total volume	25.00

Table 2.20 Thermal cycling conditions for external and nested round RT-PCR

Event	Temperature (°C)	Time	No. of Cycles
Initial melting step/AmpliTaq Gold® activation	94	2 min	1
Denaturation	94	15 sec	(td =- 0.5°C/ cycle) For 12 cycles
Annealing of primers	63	15 sec	
Extension	72	1 min*	
Denaturation	94	15 sec	Repeat 25 cycles
Annealing of primers	57	15 sec	
Extension	72	1 min*	
Final extension: filling up the ends**	72	10 min	1
Hold at	4	∞	1
Storage	-20	-	-

- td: 'touchdown' PCR; *: elongation time depends on length of amplicon: 1 kb/min; **:especially needed for TA-cloning

2.7. Direct sequencing

A major concern in the investigation of AS is the detection of potentially down-regulated splice variants. Nonsense-mediated mRNA decay (NMD) is a well-established mechanism that can lead to the down-regulation of transcripts carrying PTCs: If an intron is located > 50 nt downstream of the stop codon, then termination codon is recognized as nonsense or premature (PTC) and the transcript will be down-regulated by NMD (Green *et al.*, 2003; Lewis *et al.*, 2003; Maquat, 2005). However, previous in-house experience in detection of AS at *CARD15* locus indicated that direct sequence approach would work even in the presence of NMD. The presence of a second, alternatively spliced transcript was detected down to the

90:10% (Appendix Figure 8.1). Eight µl of PCR products from the second round were directly sequenced using the internal (nested) primers with the Big Dye™ Terminator chemistry (Applied Biosystems) according to the customized protocol described in section (2.2.12), and analyzed on an automated, high-throughput 96-capillary fluorescence detection system, the 3730xl DNA Analyzer from Applied Biosystems. Sequencing was performed for both orientations (forward and reverse), to circumvent sequencing artefacts. The resulting sequence traces were assessed for evidence for allele-dependent splicing with a newly developed specialized tool- the SNPSplicer software (ElSharawy *et al.*, 2006).

2.8. Analysis Software: SNPSplicer

As previously mentioned, direct sequencing of PCR products from cDNAs with sources of known genotype was used as an analysis tool within the experimental framework. Consequently, in order to make this an efficient process, the analysis needed to be supported by appropriate software. Therefore, a public-domain software solution, namely SNPSplicer (ElSharawy *et al.*, 2006), was developed. SNPSPlicer aids in the rapid interpretation of allele-dependent splicing of such screening experiments and, in turn, helps in the functional annotation of SNPs in a more high-throughput fashion than existing on-line tools.

In order to use SNPsplicer software, a folder containing the following files for each SNP must be prepared:

- a GenBank file of the sequence for a reference cDNA.
- a text file containing the sequences of the two primers (forward and reverse).
- several ABI or SCF trace files.

The trace file names must end in ".ab1" or ".scf". Each name must also include a classification string to indicate that file's genotype and read direction (i.e., _11_F; _12_F; _22_F; _11_R; _12_R; _22_R). Allele-dependent splicing was concluded to be present if a consistent pattern of alternative splicing was observed in all samples (including hetrozygotes and homozygotes). Interpreting the display of SNPSplicer is described in Results section (3.1.4) to avoid redundancy. For more details of the software package see (ElSharawy *et al.*, 2006) or visit the homepage: www.ikmb.uni-kiel.de/snpsplicer/. The software comes with a complete user manual and is open-source software licensed under the GNU Lesser General Public License.

2.9. Validation of allele-dependent splicing by cloning

All allele-dependent splicing effects were verified by cloning. To this end, PCR products from the second round of amplification were separated on agarose gels, excised and extracted using the Minielute Gel Extraction kit from QIAGEN (Hilden, Germany). Extracts were cloned using Invitrogen (Carlsbad, CA, USA) TOPO TA Cloning Kit (pCR®II, pCR®2.1). For each genotype, 30 clones were picked and sequenced. The resulting sequence traces were aligned using Sequencher (www.genecodes.com), followed by manual verification of the alignments.

2.10. Development of an *in vitro* splice reporter system

To develop a novel *in vitro* splice reporter assay for alternative splicing, a test genomic region, comprising exons 7, 8 and 9 and the intronic regions, was chosen from *PGM2L1* gene (chr 11; NM_173582). This genomic region was PCR-amplified using platinum Taq polymerase with primers (PGM_SacII_F at exon 7; PGM_R_BamH at exon 9) with 5′ restriction site of *SacII* and 3′ restriction site of *BamHI*, respectively. The purified PCR product (1884 bp) was cloned with TA- cloning kit (Invitrogen) (section 2.2.9). As a standard, the retrieved genomic region of *PGM2L1* was cloned at the *SacII*/*BamHI* restriction sites, at the MCS of the pEGFP-N1 vector (Clontech). The insertion of the genomic region was confirmed by double digestion with *SacII*/*BamHI* and verified by sequencing. To avoid an internal translation initiation site, the start codon of GFP of 'PGM2L1-pEGFP-N1' construct was eliminated using QuikChange® Lightning Site-Directed Mutagenesis Kit (Stratagene) (section 2.2.10) with the mutagenic primers 'GFP_rem_ATG_F' and 'GFP_rem_ATG_R'. The ATG-removal was verified by sequencing the purified plasmid DNAs (Maxi) using the primer 'NM173582_PCR_f'. The produced vector, namely PGM2L1-pEGFP-N1 (Figure 2.5), was then subjected to FACS analysis (section 2.2.15).

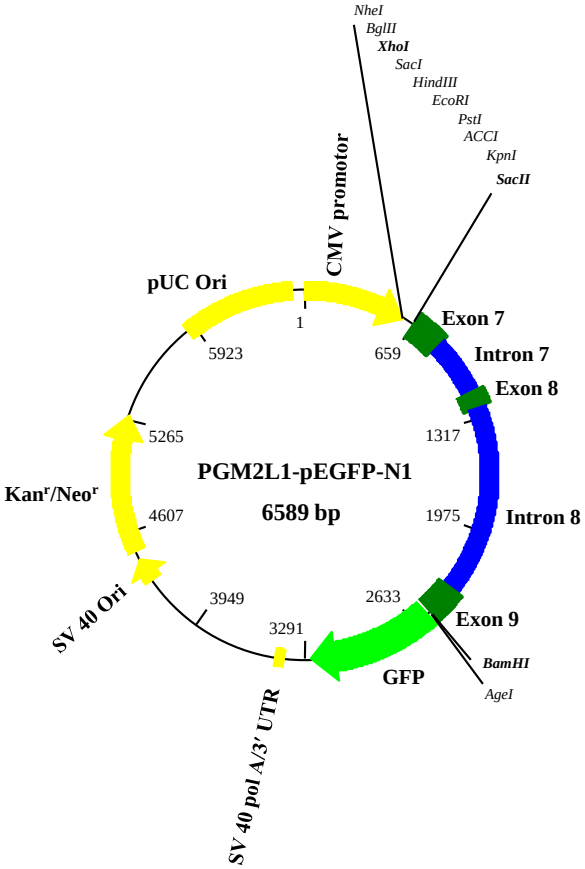

Figure 2.5 Insertion of the test genomic region into the MCS of pEGFP-N1 vector.
The genomic region that comprises exon7-exon9 (1864 bp) of *PGM2L1* gene was inserted at the MCS of pEGFP-N1 vector using the *SacII/BamHI* restriction sites. To avoid internal translation initiation, the ATG-start codon of the resulting hybrid vector (at positions 2538-2540 bp) was eliminated using site directed mutagenesis.

Next, the coding region of RFP (672 bp) was PCR-amplified using a high fidelity PWO Taq polymerase (Roche) from pDsRed2-N1 vector (Clontech) with primers (dsRed_Xho_f; dsRed_XhoI_r) that ended with recognition site of *XhoI* endonuclease. Here, the forward primer (dsRed_Xho_f) was designed to begin at the start codon of RFP and the reverse primer (dsRed_XhoI_r) to delete the TAG-stop codon of RFP to have a read-through when inserted into at *XhoI* in the PGM2L1-pEGFP-N1 vector in next steps. The purified PCR product and PGM2L1-pEGFP-N1 vector (Clontech) were separately digested with *XhoI* enzyme using standard protocol (section 2.2.8). PGM2L1-PEGFP-N1 was dephosphorylated using 1 µl alkaline phosphatase (New England Biolabs) in the digestion reaction. The purified DNA products were ligated using T4-ligase enzyme, transformed into *E.coli* Top 10 competent cells as described in section (2.2.9.3). The insertion of the cds of RFP was confirmed using digestion with *XhoI* enzyme. The insertion of both regions, cds of RFP and the test genomic region of *PGM2L1*, were validated by sequencing using three primers at different locations (dsRed2-578-f: located at RFP and read through *PGM2L1*; NM173582_PCR_f: started

reading at 430 bp of in *PGM2L1*; PGM_GFP_F_Seq: located at intron 8 of *PGM2L1* and read through GFP of the parent vector). The resulting construct, namely 'RFP-*PGM2L1*-pEGFP-N1' vector, was then subjected to FACS analysis as described in section (2.2.15).

To test the utility of the splice construct (RFP-PGM2L1-pEGFP-N1) in detecting ss variation, both acceptor and donor ss of test exon 8 (of the inserted genomic region of PGM2L1) were separately mutated using QuikChange® Lightning Site-Directed Mutagenesis Kit (Stratagene) (section 2.2.10). While the mutagenic primers ('PGM_int8_CA_F' and 'PGM_int8_CA_R') were used to mutate the obligatory GT-donor consensus to CA dinucleotides, the primers ('PGM_int7_TC_F' and 'PGM_int7_TC_R') were used to mutate the obligatory AG-dinucleotides at acceptor ss (at the end of intron 7) to TC-dinucleotides. After standard transformation into XL10-Gold® Ultracompetent cells and cloning procedure (described in section 2.2.9.3), plasmid DNA of the mutant clones were purified to transfection grade with the help of the QIA Filter™ Plasmid Maxi Kit from Qiagen (section 2.2.11). To verify the success of the mutagenesis, 2 µl of each isolated plasmid DNA was separately sequenced using two different primers (NM173582_PCR_f and dsRed2-578-f). To monitor any alteration in the functional mode of the produced construct, FACS analysis and protein evaluation by SDS-PAGE/western blotting were carried out in duplicate for both wild type and mutant constructs. Briefly, for each type of analysis, 1 µg of each plasmid DNA (from maxi preparations) was transfected in duplicate to HeLa cells (section 2.2.13) in the presence of separate positive (pEGFP-N1 and pDsRed2-N1 vectors) and negative controls (untransfected mock cells). After one day of incubation at 37°C, the transfected HeLa cells were washed and harvested in 1 ml PBS buffer. FACS analysis was carried out as described in (section 2.2.15). On the other hand, protein lysates were prepared from harvested HeLa cells using denaturing lysis buffer as described previously (section 2.2.14), assayed using BioRad DC-Protein Assay, and concentrations were measured at 75 nm using IMPLEN Nanophotemeter. Ten µg of total protein extract were then separated by SDS polyacrylamide gel electrophoresis, with subsequent electrotransfer to a suitable PVDF membrane, and subjected to Western blot analysis as described by (Waetzig *et al.*, 2002) and in section (2.2.14). GFP was detected using a murine *A.V.* GFP monoclonal antibody (JL-8) (1:1000; Clontech), and an anti-mouse antibody conjugated to HRP (horseradish peroxidase; ab6728; 1:2000) from Abcam was used as a secondary antibody. The blot was then stripped and reported for β-actin, as an internal control, using mouse anti-β-actin monoclonal antibody (1:1000; Clone AC-15, Sigma) (see section 2.2.14).

3 RESULTS

3.1. A high-throughput assay for the investigation of allele- dependent splicing

In the present study, a high-throughput method was optimized and established to facilitate the screening of allele-dependent splicing (ElSharawy *et al.*, 2006); Figure 3.1. In addition to the available high-throughput genotyping facilities at ICMB (Kiel, Germany), the method integrated a package of four new software tools that has been created and implemented in the in-house database infrastructure (described below).

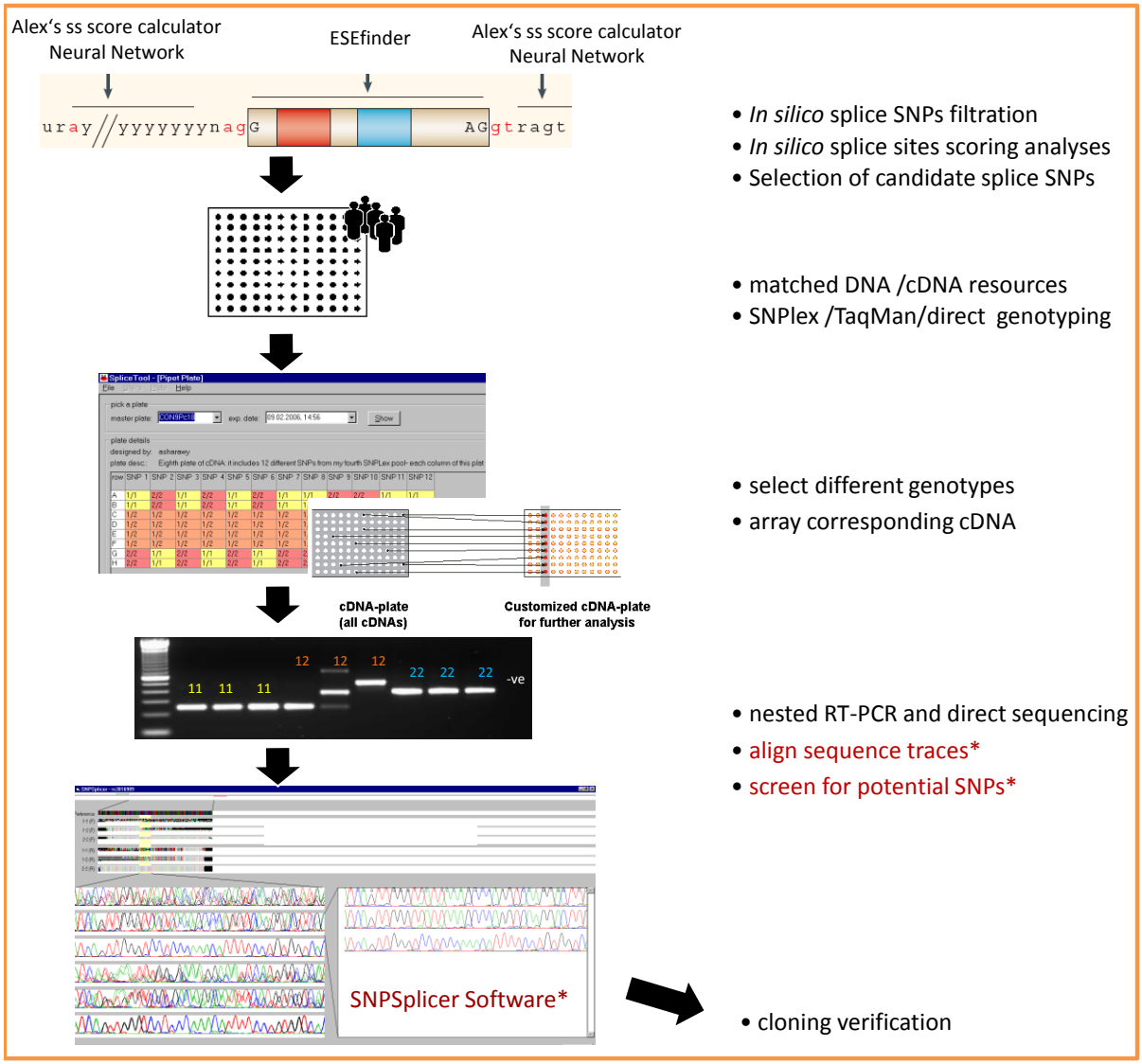

Figure 3.1 Established workflow of the streamlined methodology used in the present study to screen for allele-dependent splicing instances.

3.1.1. MotifSNPs Tool: Extraction of splice SNPs from public database

In order to carry out a genome-wide screen for candidate genes with potential splice SNPs, an online SNP evaluation tool, namely 'Motif SNPs Input Page' (available at www.ikmb.uni-kiel.de/motifsnps), was created in the framework of the present study. MotifSNPs tool is a standalone application that rapidly displays how a splice motif score changes in a biallelic SNP of interest. The results are displayed in a user-friendly html format. Figure 3.2 provides an illustrative sketch of the basic flow that has been used to extract biallelic SNPs that mainly reside in specific splice motifs, such as ESEs, donor and acceptor sites. This application rebuilt annotated genes *in silico* and analyzed the sequences (Figure 3.2 I): the tool obtained the annotation data of a specific cDNA using the cDNA annotation table "*RefSeqAli*" from the UCSC Genome Browser homepage, extracted the sequences from the chromosome file, and rebuilt the gene *in silico*. The database build used in the present study was hg17 with 24292 annotated sequences. Annotation mistakes, such as two exons without intronic sequences in-between or intronic sequences shorter than 16 bp, were also corrected at this stage. In such cases, all sequence parts were merged to yield one correct exon annotation sequence.

Next, every SNP in each sequence part (intron/exon) was filtered (Figure 3.2: II) and for each part, the application built two sequences (Figure 3.2: III). The first sequence contained the allele from the chromosome files and the second sequence with the second allele of every SNP inside this part. Finally the program collected the "*ChromFA"* (human genomic sequence in FASTA format) files from the UCSC Genome Browser homepage and scored ESE motifs matrices using the ESEfinder homepage (Cartegni *et al.*, 2003). By combining these resources the program could easily find SNP(s) inside ESE motifs and ss (Figure 3.2: IV). Then, the application searched for motifs in both sequences and stored the results in two different lists. Also, ss sequences from the intron-exon and exon-intron changeover were extracted. The results containing the highest motif scores of each type in both sequences were then stored (Figure 3.2: V). The tool described here can be easily updated and adapted to populate additional splice motifs, because it uses configuration files that define the motifs. One can add a splice motif by opening the configuration file with a text editor and inserting the new scoring matrix. The file format is described in the file, *readme.txt*, which is included in the downloadable zip files. Splicing variants that are not annotated in the public database can be edited as new entries in this application. This is valid for *de novo* detected SNPs as well. Instead of using the cDNA annotation data, the application can also be operated with the table "*all_est*" (Expressed Sequence Tags annotation).

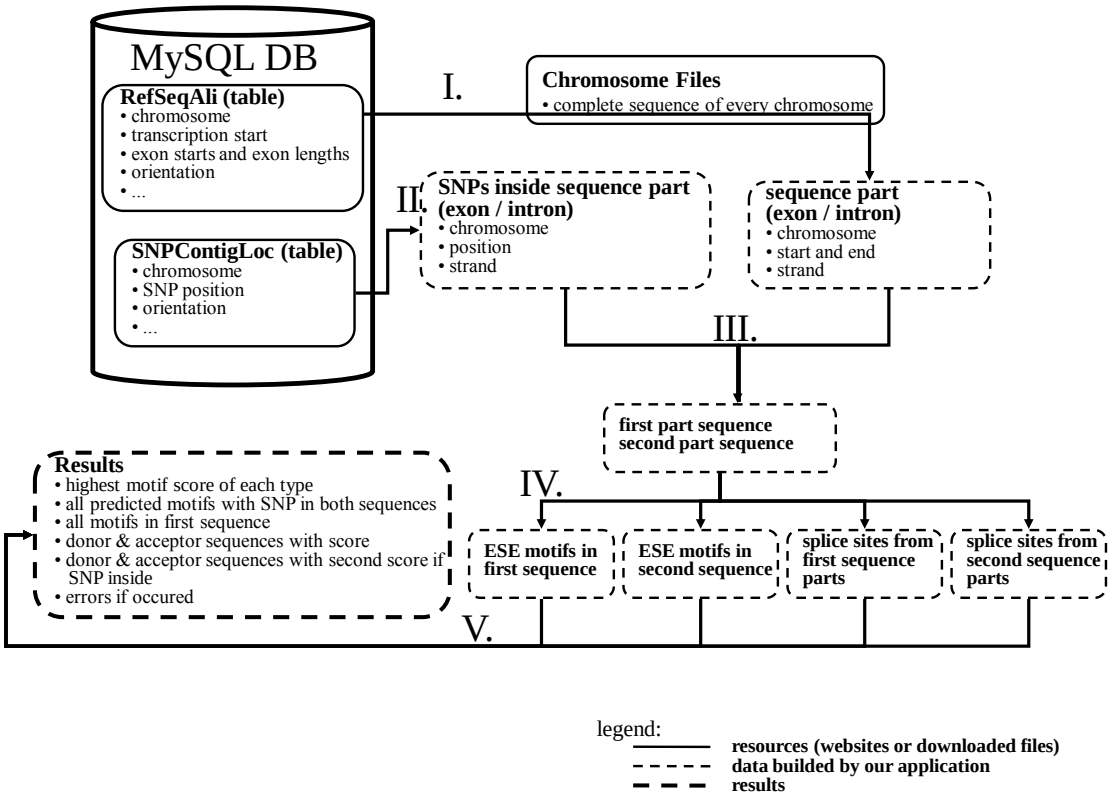

Figure 3.2 A Workflow of extraction of splice SNPs from public database using MotifSNPs tool.

3.1.2. SpliceTool software: Arraying of cDNA

As transcripts with defined genotypes at potential splice SNPs were required and the selected candidate SNPs were frequent enough with about 10% heterozygosity, genotyping of an initial sample set of 92 gDNAs was performed in the present study. Because the principal existence of the SNPs after genotyping has been verified (section 2.4), only 16 different cDNAs needed to be analyzed in the subsequent steps to investigate allele-dependent splicing effects. Individuals representing the three possible SNP genotypes, 4 homozygotes for each allele and 8 heterozygotes, were chosen and the corresponding cDNA samples were picked by a robot and arrayed into 96-well plates for subsequent RT-PCR experiments. If possible, each genotype was represented by different tissue resources. For rare SNPs, where no samples homozygous for the minor allele were present, at least four heterozygotes had to be available in order a SNP to be included in the subsequent analyses. If evidence for allele-dependent splicing was obtained, additional independent cDNA samples were analyzed to confirm the effect (ElSharawy *et al.*, 2008).

To facilitate automated arraying of cDNA samples of corresponding genotypes of each SNP, a SpliceTool software was created in the present study. Screenshots of the SpliceTool interface and utilization are shown in Figure 3.3. Technically, SpliceTool was developed as a user-friendly client program to work with data, which was stored in a Microsoft SQL Server 7 database. The tool was written in Visual Basic 6 and implemented on a Microsoft Windows 2000 system. It connects to the database via TCP/IP using the OleDb provider. For pipetting, SpliceTool remote controls the Gemini software (Tecan) via its named pipe interface. Using SpliceTool, individuals representing the three possible genotypes (denoted 11, 12, and 22) were identified and corresponding cDNA samples were robotically picked and arrayed into 96-well plates for subsequent transcript RT-PCR analysis (ElSharawy *et al.*, 2006).

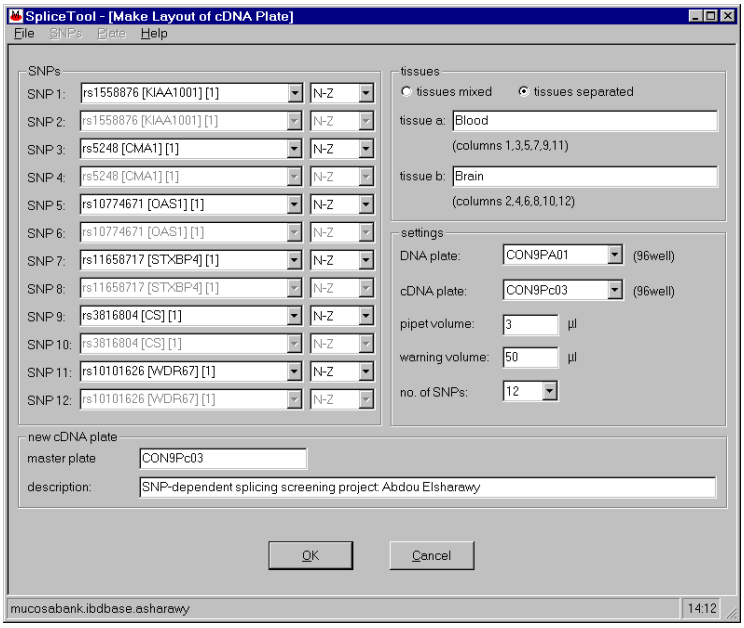

Figure 3.3 SpliceTool interfaces and arraying of cDNA samples.
The upper screenshot shows the common interface window of the SpliceTool software that used for designing cDNA-plate layouts. This screenshot illustrates the ease of using SpliceTool to facilitate choosing and arraying cDNA according to the corresponding genotypes of desired candidate splice SNPs in specified 96-well plate format. One can annotate tissue resources, required pipetting volume, warning volume to alert about consumed volume of cDNA samples, number and ID of SNPs to be arrayed, and other different database details. The lower screenshot of the program shows another window of the SpliceTool interface that helped for recalling genotyping data. Another useful option here, as shown at the bottom of the second screenshot, is to provide with a short report about allele frequency and heterozygosity of each genotyped splice SNP, which in turn, helped in designing downstream RT-PCR experiments.

3.1.3. SkippedExonPrimer Tool: Semi-automation of designing nested primers

In order to facilitate design of nested primers, an assisting software tool, namely "SkippedExonPrimer", was created in the present study. As indicated from its name, designing nested primers on skipped exons flanking the candidate SNP of interest was not preferred, but rather on common or conserved ones. In principle, this tool was connected to a MySQL Server to get the annotation data of the UCSC table RefSeqAli. The annotation was then used to collect the relevant exon sequences from the UCSC "*ChromFA*" files. Technically, this software was written in Visual Basic 6.0 to run on Microsoft platforms, e.g. Windows NT. The only prerequisite to start this software was a tab-delimited text file containing three columns: SNP ID, refseq annotation accession number and genomic position of each SNP. Using this (tab-deliminted text) input file, SkippedExonPrimer enabled accession to different required web interfaces: (1) 'Primer3' (Rozen and Skaletsky, 2000) for picking desired primers with annotated criteria; (2) 'SPIDEY' at NCBI, which is an mRNA to genomic alignment tool; (3) 'UCSC Genome Browser' to check whether the obtained primers spanned conserved exons flanking the potential splice SNP's region and ensure specificity. A list of the primers used in the present study is provided in Appendix Table 8.1.

3.1.4. SNPSplicer: A screening tool for allele-dependent splicing signals

In the present study, a key software solution, SNPSplicer, was developed in order to support experiments that used corresponding pairs of gDNA and cDNA. This specialized new piece of software allowed a rapid interpretation to determine if a potential site-specific splice effect was present (ElSharawy *et al.*, 2006). The basic display of SNPSPlicer (Figure 3.4) is described next, followed by three different practical examples (Figure 3.5- Figure 3.7) of the investigation of allele-dependent splicing.

In its alignments, SNPSplicer relies on the fact that at least the initial 20-30 bases are located in a nondifferentially spliced exon and thus is able to anchor the sequences. Technically, SNPSplicer builds on a software library previously used for mutation detection and SNP genotyping software (Manaster *et al.*, 2005a; Manaster *et al.*, 2005b). SNPSplicer reads several files from one folder. These files consist of a GenBank file of the sequence for a reference cDNA, a text file containing the sequences of the two primers, and ABI or SCF trace files named to indicate their genotype and read direction. To show splicing differences between cDNA of individuals representing different genotypes, SNPSplicer groups the cDNA

sequences by genotype and read direction (Figure 3.4). The sequences are represented as horizontal strips in six sequence charts aligned underneath a strip representing the cDNA reference sequence; bases in the sequences are pale where they match the reference and dark where they differ. This makes systematic differences between groups easy to identify. Each base in the sequence occupies just a single pixel of width. Above the sequence charts is a horizontal line that represents the entire length of the cDNA reference sequence, with arrows identifying the primer locations. Below the sequence charts are trace groups for each genotype and read direction. In the trace group, traces from all sequences of the group are overlaid together. Clicking a trace group shows its individual traces in a panel on the right. Clicking a sequence chart refocuses the traces around the clicked base.

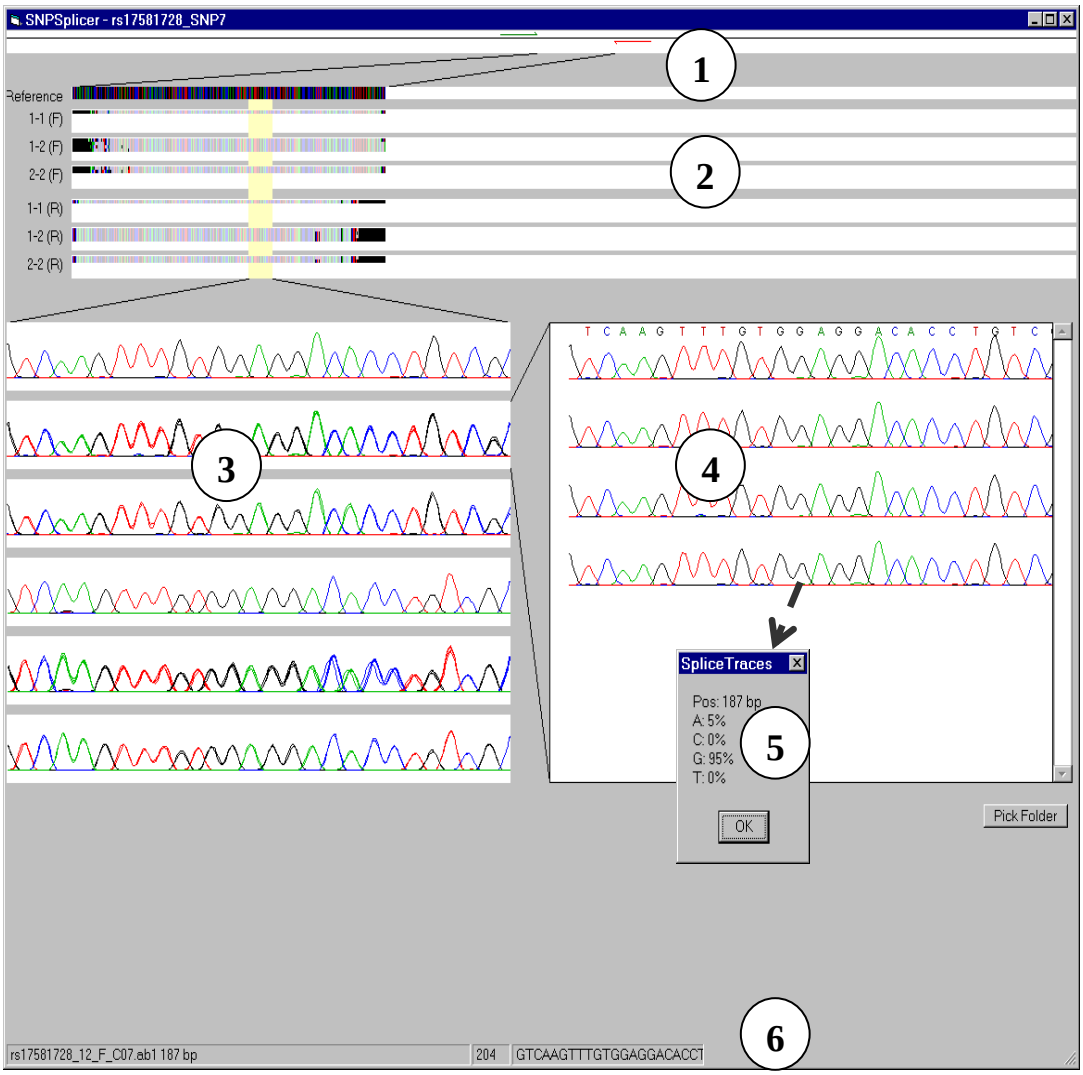

Figure 3.4 A screenshot of the SNPSplicer interface and utilization.
The interface of the program shows a schematic drawing of the reference sequence and the primer positions (1). The aligned traces of the forward and reverse sequences of the second round RT-PCR products for all three genotypes denoted 1-1, 1-2 and 2-2 at rs17581728, are displayed in the stacked sequence chart (2). This is a simple representation of sequences in a vertical stack; each sequence is one pixel high and each base is one pixel

wide. The colour of each pixel indicates the base it represents and whether it agrees with the reference sequence. Bright colours indicate agreement and dark pixels show deviation from the reference sequence. A trace view of sequences from one genotype plotted in the same chart is provided (3) together with the individual traces for a particular genotype group (4). Clicking a base on window (4) gives a short report (5) about the base position and quantifies the areas under the peak (curve); this helped to quantify the relative amounts of more than one splice variants, if present, in the heterozygotes. 'Mouse-over' across traces in window (4) also reports the name of the individual sequence at the bottom toolbar (6) together with the bp position. The 'Pick Folder' button on the lower right-hand corner was added to the SNPSplicer results' window in order to move smoothly and pick quickly another SNP-folder for analysis without restarting the program.

SNPSplicer does not do its own base calling; it relies on the calls in the trace files. Alignment to the reference sequence is done through a simple word match using a window of 20 bases, starting from each primer site. From that point on, it advances through the entire sequence, inserting and deleting bases as needed to preserve a fivebase sequence matching the reference. The utility of the approach is exemplified in next subsections.

3.1.4.1. Example of the use of SNPSplicer showing a splicing-nonrelevant SNP

The primary feature of SNPSplicer is the collection of six sequence charts in the upper half of the display screen. These show the groups of sequences with bases that differ from the reference sequence emphasized by darker shading. If the investigated SNP has no effect upon splicing, all the groups will look approximately the same; a few dark pixels indicate the presence of simple variation at these positions. Indeed, this represented the most frequent outcome in the present larger scale experiments and, in turn, served to focus resources on more promising splice SNPs for future functional and/or mechanistic analyses.

A practical example of a SNP that showed no impact on pre-mRNA splicing process is illustrated in the output shown in Figure 3.5. Panel A of this figure shows the underlying genomic sequence surrounding exon 8 of caspase 5 (*CASP5*) and the position of the putative splice SNP rs540819:T>A in the intronic donor sequence of exon 8. Panel B shows the analysis results in SNPSplicer after the sequencing of PCR products of cell-lines with all three genotypes at rs540819:T>A (ElSharawy *et al.*, 2006). From Panel B, it is evident that all genotypes display the same cDNA sequence, i.e., there is no evidence for an allele-dependent splice effect. As shown in panel C, the junction between exon 8 and 9 spliced without any impact from the tested SNP in all tested genotyped cDNA samples.

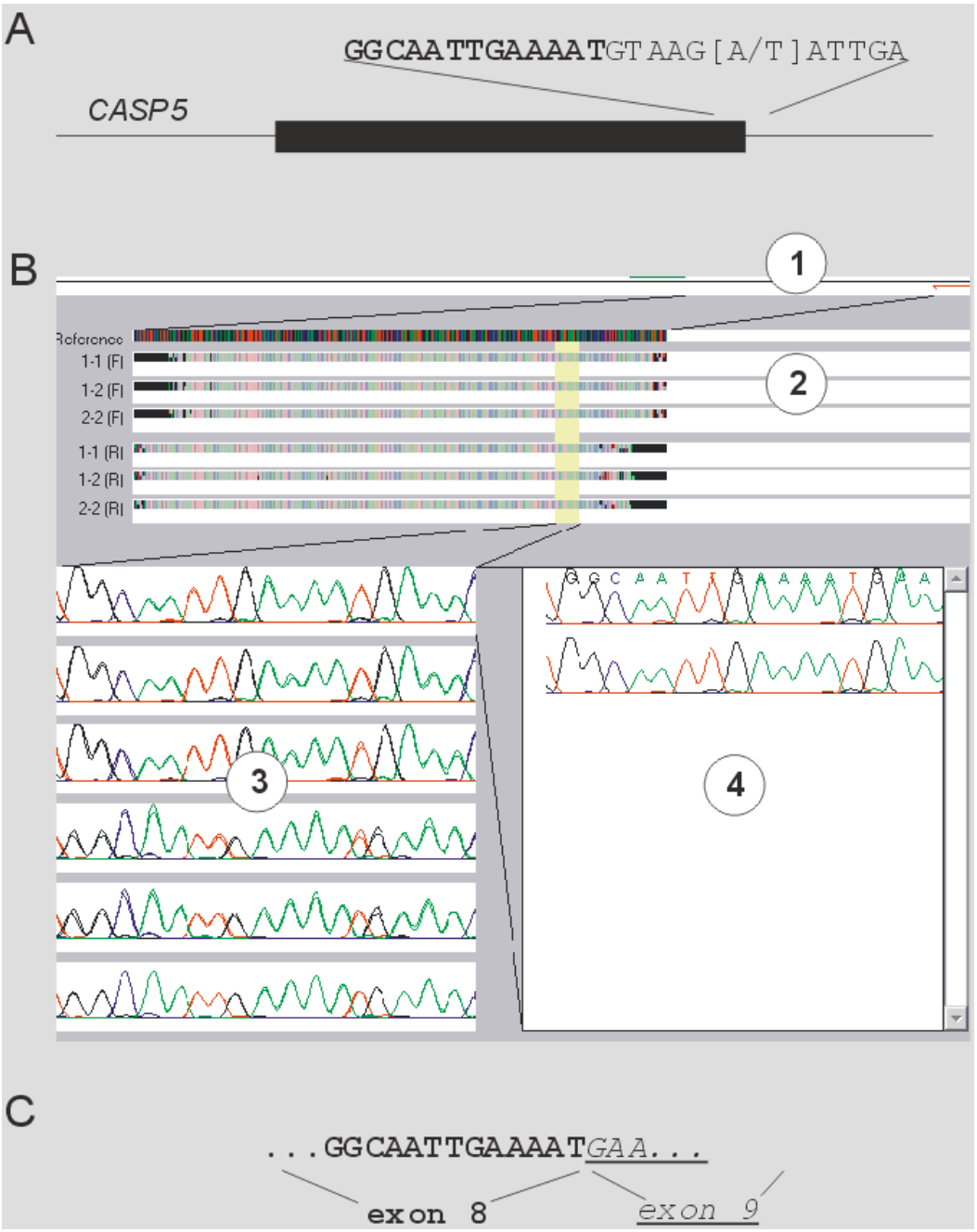

Figure 3.5 The use of SNPSplicer: a negative example involving a splicing-nonrelevant SNP.
Panel A shows the annotation of the SNP (rs540819:T>A) in the intronic donor sequence of exon 8 of caspase 5 (*CASP5*) gene. In panel B, the primers used in RT-PCR to experimentally validate the SNP effect on splicing are annotated on cDNA reference sequence (1). The stacked sequence chart shows the aligned traces of the RT-PCR products for all three genotypes at rs540819:T>A, denoted 1-1 (AA), 1-2 (AT), and 2-2 (TT) in both forward 'F' and reverse 'R' sequence read. It is clear in panel B that there is no evidence of allele-dependent splicing effect (no genotype-specific splicing effect is visible). This is because all genotyped groups at rs540819:T>A display the same cDNA sequence and look the same (ElSharawy *et al.*, 2006). Dark pixels at both margins show deviation of the experimentally generated sequences from the reference sequence, which is a typical output at the start of sequencing reads from both direction (F and R). A trace view of sequences from one genotype plotted in the same chart is provided (3) together with the individual traces for a particular genotype group (4). On the top of the individual traces, the reference sequence is shown (4). Panel (C) shows the exon annotation of the observed cDNA sequence in area (4) of Panel (B).

3.1.4.2. Simple positive example of the use of SNPSplicer

If the SNP affects splicing in a simple manner, then sequence beyond the splice site (to the right in forward traces and to the left in reverse traces of SNPSplicer display window) will show dark bands in one of the homozygous groups and possibly to some extent in the heterozygous groups.

Figure 3.6 shows a practical example where a ‘‘simple’’ splice effect is detected: a genotype-specific deviation from the consensus sequence is detected at one site in both the forward and reverse sequences. Panel A shows the underlying genomic sequence of surrounding exon 5 of the butyrophillin-like protein 2 (*BTNL2*) gene and the position of the putative splice SNP rs2076530:A>G in the exonic donor sequence of exon 5. Panel B shows the analysis results in SNPSplicer after the sequencing of PCR products of cell-lines with all three genotypes at rs2076530:A>G. The aligned traces of the forward and reverse sequences of the RT-PCR products for all three denoted genotypes, 1-1 (AA), 1-2 (AG), and 2-2 (GG), are displayed in the stacked sequence chart (2).

It is evident from Figure 3.6 that at a specific site (indicated by the blue arrow), the homozygous trace stacks and some of the heterozygous trace stacks start to differ from the reference sequence in the respective read directions. This region is selected in the group trace view (3). The individual traces show the splice pattern of allele A (4). The blue shading in area (4) was added in the figure to indicate the junction of exon 5 to exon 6. Outside the blue box, the traces no longer correspond to the reference sequence, due to the deletion of four bases. The blue shading in panel C corresponds to area (4) in Panel (B). The putative splice effect was confirmed by cloning and sequencing of PCR products from the respective genotypes (ElSharawy *et al.*, 2006). Panel C shows the genotype-specific splice pattern: the A-allele leads to a loss of 4 bases on exon 5 and subsequent frame shift of the underlying protein later in exon 6, as previously described in by Valentonyte and co-workers (Valentonyte *et al.*, 2005).

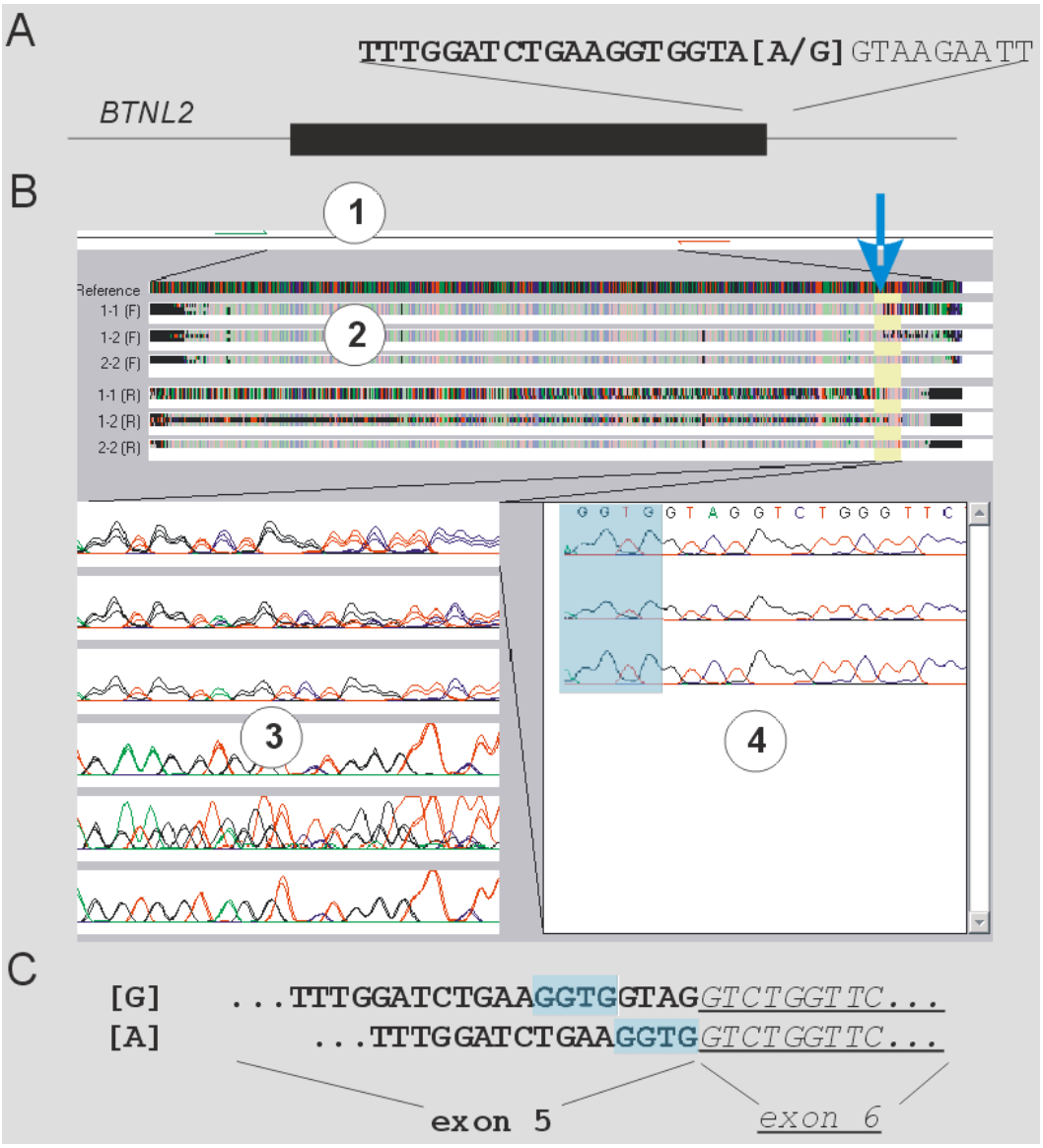

Figure 3.6 "Simple positive example" of the use of SNPSplicer: allele-dependent splicing at *BTNL2* locus. Panel A shows the sequence surrounding the splice SNP in *BTNL2*. In panel B, it is evident that the homozygous trace stacks and some of the heterozygous trace stacks start to differ from the cDNA reference sequence in the respective read directions (blue arrow). After cloning and sequencing of PCR products from the respective genotypes (ElSharawy *et al.*, 2006), the genotype-specific splice pattern is confirmed: allele A at rs2076530:A>G, is associated with 4 bases (GTAG) deletion (panel C) and using of an alternative ss upstream of donor ss of exon5 (Valentonyte *et al.*, 2005).

3.1.4.3. Complicated example of the use of SNPSplicer

If SNPSplicer yields a more complex result, this may indicate the interplay of more than one SNP on splicing process at the region under investigation. Figure 3.7 exemplifies the case. Panel A in this figure shows the genomic structure of *FLJ40873* (or *TCTEX1D1*; Tctex1 domain containing 1 gene). The primer positions of the RT-PCR are indicated with small red arrows in the genomic structure. The SNP rs3816989:G>A in the donor sequence of exon D was selected as a putative splice SNP. The SNP sequence is provided in Panel A. Panel B

shows the analysis results in SNPSplicer after the sequencing of PCR products of cell-lines with all three genotypes at rs3816989:G>A. The aligned traces of the forward and reverse sequences of the RT-PCR products for all three genotypes denoted 1-1 (AA), 1-2 (AG) and 2-2 (GG) are displayed in the stacked sequence chart (2). Here, apparent genotype- specific differences are observed at two sites, which are highlighted with blue bold arrows. One of the regions is selected in the group trace view (3). The individual traces show the splice pattern of the allele A/G heterozygotes (4). The blue shading in area (4) was added in the figure to indicate the junction of exon C to exon D, outside of the blue box. The upper two traces no longer correspond to the reference sequence, while the bottom trace matches the reference. The underlying pattern only became clear after the underlying RT-PCR products were cloned and sequenced. Here, two different splicing effects were observed. The first splicing effect was the association of the A-allele at rs3816989:G>A with skipping of exon D, because it disrupts the GT consensus sequence at its donor ss (ElSharawy *et al.*, 2006). From the cloned sequences, an additional exon (X), which is not present in the RefSeq annotation, was observed. This was associated with the presence of an additional SNP (rs2274987:T>C) that generates an ESE sequence in the intron between exons B and C. The C-allele created a novel ESE motif for SF2/ASF with an ESEfinder score of 4.01. However, the T-allele at rs2274987:T>C was no longer recognized by ESEfinder using the default cutoffs in ESEfinder (Table 3.1). The haplotypes of rs2274987:T>C and rs3816989:G>A together with the corresponding cDNA composition as observed in the cloned RT-PCR products are shown in Panel C. The [C-A] haplotype, i.e., the C-allele at rs2274987:T>C and A-allele at rs3816989, was associated with the insertion of new exon (X) and skipping of exon D, respectively. In the same order, the [T-A] haplotype was correlated with the absence of both exons X and D from the transcript. The [C-G] haplotype was not observed as a result of linkage disequilibrium (ElSharawy *et al.*, 2006). The respective sequences have been submitted to GenBank (DQ411321).

Table 3.1 ESEfinder analysis of rs2274987:T>C at FLJ40873/ TCTEX1D1

(SF2/ASF: Threshold as given by ESEfinder = 1.956)			
Position*	**Motif**	**Score**	**Splicing events** (see Figure 3.7**)**
23 (-72)	CA**C**ACAA	4.01135	Exon X insertion
23 (-72)	CA**T**ACAA	↓↓	No insertion

- *: The position refers to the motif location within the inserted exon (94 bases), as given by ESEFinder.
- The two alleles of the SNP are underlined and presented in bold print.

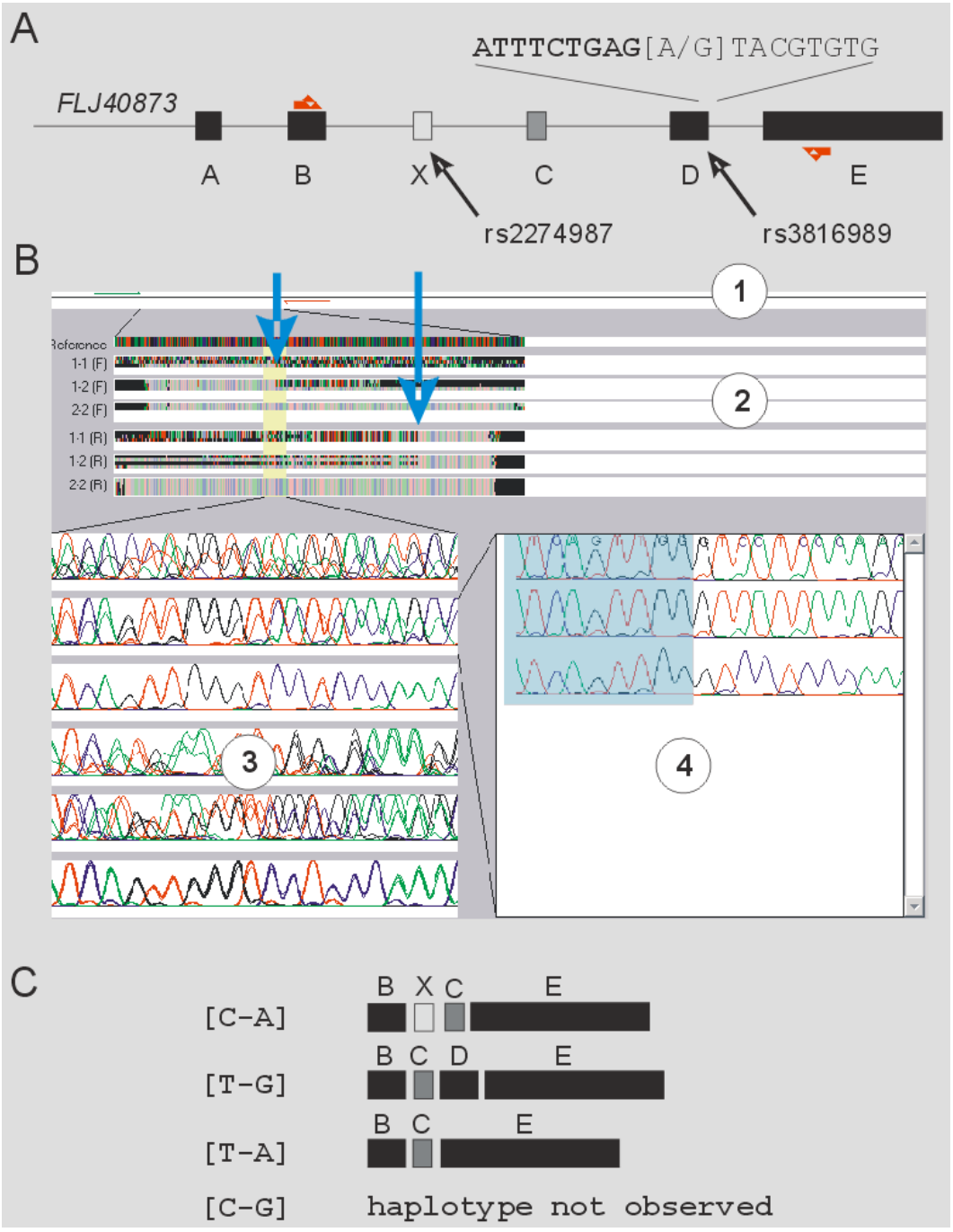

Figure 3.7 "Complicated example" of the use of SNPSplicer: two different concurrent allele-dependent splicing events.
As shown in panel B, the deviation from cDNA reference sequence of *FLJ40873* (*TCTEX1*) gene was observed at two different positions indicated by the blue bold arrows (compare to Figure 3.5 and Figure 3.6: panel B). These results cannot be solely attributed to the originally targeted putative splice SNP rs3816989:G>A. The compound splice effect was revealed after cloning and sequencing of PCR products from the respective genotypes (Panel C). While the A-allele at rs3816989:G>A disrupts the obligatory GT consensus sequence at the donor of exon D and leads to its exclusion from transcript, the C-allele at a second SNP (rs2274987:T>C) creates a novel ESE motif for SF2/ASF splicing factor as revealed by ESEfinder (Cartegni *et al.*, 2003) and leads to insertion of a new exon X (ElSharawy *et al.*, 2006).

3.1.5. Direct sequencing approach

The results from the present study showed that direct sequencing is a robust and sensitive screening tool even in the presence of nonsense- mediated mRNA decay. This is concluded from the results presented in the previous section (3.1.4), and supported from the previous in-house experience of AS analysis of *CARD15* locus (Appendix Figure 8.1); see also (Hiller *et al.*, 2006b)). Thus, the sensitivity for a qualitative detection of alternatively spliced transcript was detected down to the 80:20% range. Such transcript ratios were readily detectible in heterozygous state (ElSharawy *et al.*, 2006). The analysis of homozygote cell lines, one for each allele, also allowed unambiguous detection of allele-dependent splicing (ElSharawy *et al.*, 2008).

3.2. First screening-round of allele-dependent splicing: Web-based tools

In the first screening-round of allele-dependent splicing, freely available web-based *in silico* tools, which are designed to predict the impact of nucleotide variations on splicing, were utilized. While Alex's splice site score calculator (Shapiro and Senapathy, 1987; Senapathy *et al.*, 1990), available online at http://violin.genet.sickkids.on.ca/~ali/splicesitescore.html, was used to select candidate splice SNPs at canonical donor and acceptor ss, ESEfinder tool (Cartegni *et al.*, 2003), available at (http://rulai.cshl.edu/cgi-bin/tools/ ESE3/ esefinder. cgi? process=home), was used for SNPs at ESE sites. In parallel, candidate splice SNPs at NAGNAG tandem acceptor ss were selected from the study of Hiller and co-workers (Hiller *et al.*, 2004), owing to the difficulty to predicting their functional consequences by available *in silico* tools without experimental settings (Hiller *et al.*, 2006a).

3.2.1. Candidate SNPs for canonical and NAGNAG splice sites

Release 125 of dbSNP was screened for variants located within 3 nt of exonic or 6 nt of intronic DNA sequence surrounding a canonical ss in an annotated Refseq in UCSC hg17. This choice of sequence length is attributed to the sequence window used by Alex's splice site scoring tool. This includes 2 nt of the exon and 6 nt of the intron for donor ss. For acceptor ss, it requires 14 intronic nt and one exonic nt. To be included in subsequent analyses, a SNP had to have a minor allele frequency of ≥ 10% in the HapMap CEU samples. All SNPs in the highly conserved AG and GT dinucleotides were discarded except for nine SNPs used as 'positive controls' (Table 3.2). The variants were scored in three categories: acceptor SNPs, donor SNPs and NAGNAG SNPs (Hiller *et al.*, 2004). For acceptor ss containing a

NAGNAG motif (Hiller *et al.*, 2004), the identity of the obligatory AG dinucleotide was not clear *a priori*. In total, 1096 putative donor splice SNPs, 1451 putative acceptor splice SNPs and 28 SNPs within NAGNAG motifs were identified (step 1 in Figure 3.10) (ElSharawy *et al.*, 2008).

In fact, no interpretation guidelines were available at Alex's splice site calculator, which meant that it was left to the user to decide when a prediction is positive (i.e., expected splice-relevant SNP) or negative (SNP has no influence on splicing process). This absence of interpretation guidelines can be explained by the fact that splicing outcome does not only depend on variation in ss consensus sequences but rather on combinatorial control of many factors involved in the splicing process (Hertel, 2008). In order to facilitate this *in silico* interpretation, score variations were considered rather than the scores themselves. Next, the absolute score difference (Δ_S) between the two alleles was calculated using Alex's splice site score calculator (Shapiro and Senapathy, 1987; Senapathy *et al.*, 1990). The absolute difference between scores was used because the presumed 'wild-type' allele yielded a lower score than the other allele in 48% of cases so that any assignment of wild-type status to one allele or the other would have highly been ambiguous (ElSharawy *et al.*, 2008). The following step was to set a limit of significance for score variations. Based on the distribution of the absolute allelic difference (Δ_S) of the splice site scores, shown in Figure 3.8, and scores of published donor ss SNPs (Roca *et al.*, 2005), all 58 SNPs with $\Delta_S>8$ were selected as candidate splice SNPs. Three donor SNPs (rs482082:C>T, rs820329:A>T and rs540819:A>T) with $8\geq\Delta_S>5$ were included because of their proximity to protein domains (WD40, IG and CARD, respectively) that are duplicated in many gene families. Proximity to such duplicated domains was thought to increase the probability of allele-dependent splicing (ElSharawy *et al.*, 2008). Domain contexts containing such repetitive protein domains are involved in ligand and pathogen recognition. These and other domains are primarily directed towards inflammation and innate immunity, and are of great in-house interest in the study of inflammatory barrier diseases (Schreiber *et al.*, 2005). Therefore, SNPs in these domain contexts were also favored for in-depth investigation. All 28 NAGNAG SNPs were also retained as candidates, thereby yielding a total of 58+3+28=89 SNPs.

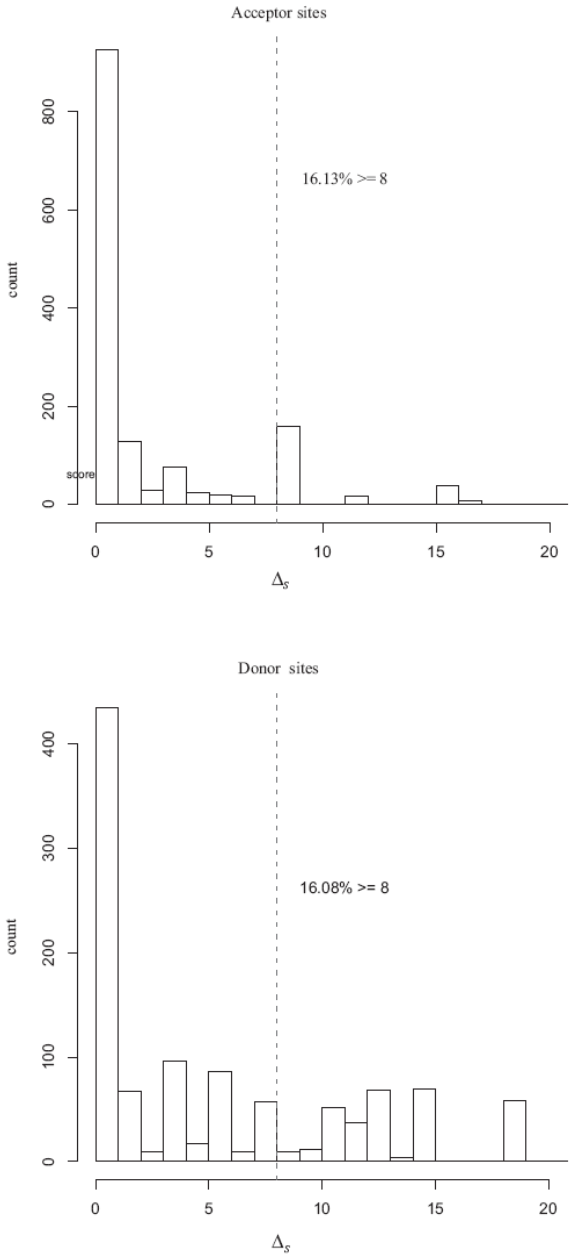

Figure 3.8 Distribution of the absolute allelic difference Δ_S of the splice site scores.
This was obtained for 1115 acceptor and 798 donor SNPs using Alex's splice score calculator (Shapiro and Senapathy, 1987; Senapathy *et al.*, 1990). The tails of the distributions were used to select candidate SNPs for allele-dependent splicing.

All 89 candidate splice SNPs retrieved from dbSNP were genotyped in a panel of 92 gDNAs using either SNPlex, TaqMan or direct sequencing, depending upon technical feasibility (step 2 in Figure 3.10; see also Appendix Table 8.1). Twenty-four donor, nine acceptor and 24 NAGNAG SNPs were found to be sufficiently polymorphic in the DNA panel so that at least four heterozygotes were present, leaving 24+9+24=57 SNPs for RT-PCR evaluation. All nine acceptor SNPs were located at position –3 whilst the 24 donor SNPs occupied positions –2 to +6 (ElSharawy *et al.*, 2008). Sixteen cDNAs were selected on the basis of SNP genotypes as

described in the Methods section. If available, one or two homozygotes for the rare allele were included in the tested 16 matching cDNAs of known genotype. These cDNAs were subjected to nested RT-PCR using the primers reported in Appendix Table 8.1. The resulting PCR-products were sequenced directly and screened for differential splicing as described in the Methods section (ElSharawy *et al.*, 2006). Putative splicing effects were confirmed by cloning and sequencing of the respective PCR-products. Allele-dependent splicing was considered to be established if the effect of a given polymorphisms could be demonstrated in all investigated samples. An overview of the selection procedure and output from the first screening round is provided in Table 3.2.

As expected, all the nine selected SNPs impacting the highly conserved AG or GT dinucleotides at ss, i.e. the positive controls, were correctly predicted as deleterious and showed allele-dependent splicing patterns (Table 3.4). None of the nine candidate SNPs at acceptor splice sites exhibited allele-dependent splicing of the adjacent transcript. Of the 24 donor SNPs, two (i.e. 8%) occupied positions -1 and +5 exerted an influence upon splicing (ElSharawy *et al.*, 2008). Four confirmed splice effects (4/24, corresponding to a positive predictive value of 17%) were observed for SNPs at acceptor sites containing a NAGNAG motif. *Post hoc* analysis of their splice site scores revealed that splicing was only affected if the AG dinucleotide with the higher impact on the splice score was changed by the SNP (ElSharawy *et al.*, 2008). An overview of the NAGNAG SNPs and the corresponding splice site score differences is given in Table 3.5. The six SNPs with a confirmed splicing effect are listed in Table 3.4.

3.2.2. Putative splice SNPs at ESEs

As a fourth category of SNPs, polymorphisms located in ESE motifs were also evaluated. All exonic SNPs located in RefSeqs in UCSC hg17 were screened with ESEfinder and a prediction was made regarding their likely effects upon splicing (Cartegni *et al.*, 2003). Owing to the large number of potential ESEs, candidate splice SNPs were further required to be located within 30 nt of the nearest exon-intron boundary, as suggested previously (Fairbrother *et al.*, 2004a); see Appendix Figure 8.2. The scores for the different ESE motifs were normalized to unity, using the respective score thresholds proposed by the ESEfinder tool (Cartegni *et al.*, 2003), and the respective score differences Δ_{ESE} were calculated in analogy to signal differences at canonical ss. Based on the obtained distribution of the absolute allelic difference (Δ_{ESE}) of SNPs at ESE sites (Figure 3.9) and the established ESE

splice SNPs from literature (Liu *et al.*, 2001; Cartegni and Krainer, 2002; Colapietro *et al.*, 2003; Zatkova *et al.*, 2004), an arbitrary Δ_{ESE} cut-off of 0.8 was chosen for the selection of SNPs for experimental follow-up. This resulted in the inclusion of 106 SNPs in the genotyping stage 2, which was carried out as described above. Five of the 106 SNPs (rs2228173:T>C, rs3763840:G>A, rs974144:C>T, rs2188383:C>G and rs736795:G>A) with $0.8 \geq \Delta_{ESE} > 0.5$ were also included on the basis of their proximity to WD40, LRR and DEATH domains (ElSharawy *et al.*, 2008). These potentially repetitive protein domains were thought to be more likely to be subject to allele-dependent splicing and involved in ligand and pathogen recognition (Appendix Table 8.1). Forty-two of the genotyped SNPs were sufficiently frequent in order to be evaluated by RT-PCR.

For none of the 42 investigated SNPs at ESE sites could an effect upon splicing be observed (ElSharawy *et al.*, 2008). One instance of putative allele-dependent splicing at an ESE was detected by chance (Table 3.4): SNP rs2274987:T>C was identified as a splice SNP when analyzing nearby SNP rs3816989:G>A, located at a canonical donor splice site (ElSharawy *et al.*, 2006). SNP rs2274987:T>C itself is located 25 nt downstream of the acceptor splice site, of the newly inserted exon (Figure 3.7), but creates a novel ESE, as suggested by ESEfinder (Cartegni *et al.*, 2003; Smith *et al.*, 2006).

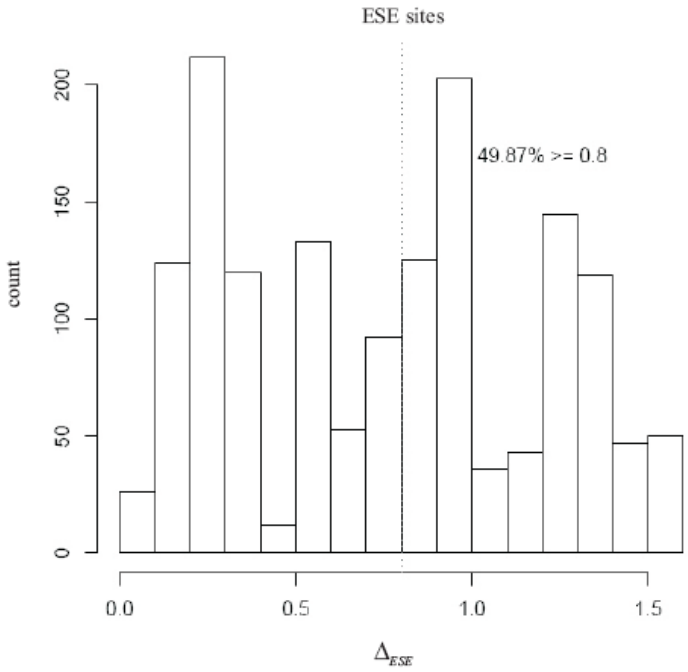

Figure 3.9 Distribution of the absolute allelic difference (Δ_{ESE}) of SNPs at ESE sites.
The tail of the distribution was used to select candidate SNPs for allele-dependent splicing. A random Δ_{ESE} cut-off of 0.8 was chosen for the selection of SNPs for experimental follow-up.

Table 3.2 Overview of the selection stages and output from the first screening round (web-based tools)

Genome-wide filtration	137*	~ 63,000 SNPs				Total
Stage I	28	1115	798	1495	-	3436
Stage II	28	23	38	106	-	195
Stage III	24	9	24	42	9	108
Location	**NAGNAG***	**Acceptor**	**Donor**	**ESE**	**AG/GT****	**-**
Allele-dependent splicing effect	**4 (17%)**	**0 (0%)**	**2 (8%)**	**0 (0%)**	**9 (100%)**	**-**

- SNPs at selection stage (I) were fulfilling the following criteria: i) HapMap validated, ii) Caucasian, iii) ≥ 10% Heterozygosity, iv) SNPs at acceptor and donor ss located within 9-nt window (6 nt intronic and 3 nt exonic, 3 nt exonic and 6 nt intronic, respectively), and V) SNPs at ESE sites were chosen from normalized scores from ESEfinder that located in a 30-nt window of exon borders.
- At selection stage II top-scoring SNPs were selected after applying the suggested cut-offs of each group.
- To be included in the subsequent analyses, at least four heterozygotes had to be available for each genotyped SNP (stage III).
- *: SNPs located at NAGNAG-tandem acceptor ss were obtained from Hiller *et al.* (Hiller *et al.*, 2004).
- **: Expected positive control SNPs, as they disrupt the obligatory AG or GT dinucleotides at canonical acceptor and donor ss, respectively.

3.3. Second screening-round: Neural network assessment of canonical splice sites

First-round screening results of allele-dependent splicing revealed that the performance of *in silico* web-based tools, i.e., Alex's splice score calculator and ESEfinder tool, was weak (ElSharawy *et al.*, 2008). Therefore, we decided to carry out a second screening second round using a recently reported neural network (Krawczak *et al.*, 2007) in order to predict the splicing effects of SNPs at canonical donor and acceptor ss. The neural network applied here has shown the ability to facilitate the recognition of higher-order sequence features that would not be detectable by the sequential consideration of consensus sequences. Moreover, it was highly efficient at recognizing the effect of SNPs at ss, achieving 91.3% and 96.1% sensitivity on SNP-containing acceptor and donor ss, respectively (Krawczak *et al.*, 2007). In fact, the applied neural network achieved sensitivity and specificity values that are comparable to those provided in other reports of neural network–based ss recognition (e.g., (Ogura *et al.*, 1997; Ho and Rajapakse, 2003)). Technically, a multilayer, back-propagation neural network (Wasserman, 1989) was trained for the purpose of ss recognition (Krawczak *et al.*, 2007) in the present study. The term 'back propagation' refers to the principle that calculation of synaptic weight changes proceeds in the reverse direction (from the output layer towards the input layer). More technical details that defined the basic mathematical algorithms of the artificial neural network are reviewed in (Papik *et al.*, 1998).

The neural network emits a signal between 0 and 1, where 1 corresponds to the classification of a sequence motif as a functional splice site. Similar to the score-based approach, SNPs located in a 20-nt window around a canonical splice site were extracted from dbSNP build

125, using essentially the same criteria as above (see Table 3.4; Figure 3.10). This choice of sequence length represented a compromise between the constantly improving efficiency of neural networks, which was obtained by taking an increasing number of nucleotide positions into account, and the fact that the specificity of the DNA sequence context of ss outside the chosen range was marginal (Zhang, 1998; Eden and Brunak, 2004). The number of SNPs selected for genotyping (N=202, 101 each for donor and acceptor) was chosen so as to match the available laboratory resources. For both splice site types, the selected 101 SNPs came from the top (N=52), middle (N=23) and bottom (N=26) range of the absolute allelic signal difference Δ_N for the neural network. The 52 top SNPs corresponded to a Δ_N range between 0.998 and 0.1224 for donor sites, and between 0.993 and 0.140 for acceptor sites. The SNPs from the middle range were chosen at random and comprised Δ_N values between 0.026 and 1×10^{-4} for donor sites, and between 0.060 and 2×10^{-4} for acceptor sites. For the bottom category, the 26 SNPs from the lower end of each Δ_N distribution were chosen, all of which had a Δ_N value of zero. All of the selected SNPs were subjected to genotyping. As described above, all variants with at least four heterozygotes in the DNA panel were further investigated with RT-PCR and validation. These were included 81 SNPs at donor splice sites (43 at top, 13 at middle, and 25 SNPs at bottom range) and 70 SNPs at acceptor splice sites (35 at top, 17 at middle, and 18 SNPs at bottom range) (ElSharawy *et al.*, 2008). The selected nine positive controls were also occurred to produce, as expected, high Δ_N ranged which from 0.80319 to 0.99659 (Table 3.4). A complete list of SNPs and the corresponding Δ_N values are provided in Appendix Table 8.1.

Here, allele-dependent splicing could be demonstrated for two donor site SNPs (5%) and three acceptor site SNPs (9%) from the top range of the neural network signal difference. Two acceptor site SNPs from the middle range (rs1558876:C>G and rs5248:A>G) exhibited allele-dependent splicing too, although the actual Δ_N values were small (0.00305 and 0.00244) (Table 3.4). Both SNPs were located at NAGNAG motifs. No allele-dependent splicing was observed for donor site SNPs with a Δ_N value below the threshold for the top range (0.1224) (ElSharawy *et al.*, 2008). Five instances of experimentally verified allele dependent splicing were predicted by both the neural network screen and Alex's splice site score calculator. As expected, no impact upon splicing could be detected for any of the tested SNPs at the bottom range for both donor and acceptor ss. An overview of the selection procedure and results from the neural network analysis (second round) is provided in Table 3.3.

Table 3.3 Overview of the selection procedure and output from the neural network (second round)

Genome-wide filtration	~ 8,000 SNPs		Total
Stage I	4039	3940	7979
Stage II	2209	2255	4464
Stage III	1240	1311	2551
SNPs at top-middle-bottom list	52- 23- 26	52- 23- 26	104- 46- 52
At least 4 heterozygotes	43- 13- 25	35- 22- 18	78- 35- 43
Location	**Donor**	**Acceptor**	**-**
Allele-dependent splicing effect	**2 (5%)– 0 (0%)– 0 (0%)**	**3 (9%)– 2 (12%)– 0 (0%)**	**-**

- At stage I, the neural network was operated for SNPs located at 20- nt window of splice site (15 nt intronic and 5 nt exonic at acceptor ss; 5 nt exonic and 15 nt intronic at donor ss).
- At selection stage II only unique donor and acceptor ss were considered; i.e., repetitive Genbank accession numbers were thus removed.
- SNPs were filtered to stage III after fulfilling the following criteria: i) HapMap validated, ii) Caucasian, iii) ≥ 10% Heterozygosity.
- Candidate SNPs were chosen from top, middle, and bottom of scoring list of respective donor and acceptor splice site, to better evaluate the performance of neural network and to match to the available laboratory resources.
- Genotyped SNPs with at least 4 heterozygotes in the tested DNA panel were experimentally validated by nested RT-PCR and direct sequencing.

3.4. Combined outputs and observations from both screening rounds

A schematic overview of the selection procedure of candidate splice SNPs and results from both screening rounds of allele-dependent splicing are simultaneously presented in Figure 3.10. A total of 344 SNPs were genotyped in the panel of 92 DNAs. The numbers in the figure sum up to 397 (89 from Alex' splice site score calculator, 101 for each acceptor and donor from the neural network and 106 SNPs from the ESEfinder). The difference is due to 53 SNPs that were retrieved both with Alex' splice site score calculator and the neural network. As a result of genotyping, 223 non-redundant variants (including the 9 'positive controls' at AG or GT canonical dinucleotides, 99 and 115 SNPs from the first and second round, respectively), which were frequent (i.e., at least 4 heterozygotes were available for each SNP) in the tested DNA panel, were tested by nested RT-PCR and direct sequencing at step 3 (ElSharawy *et al.*, 2008). All the investigated SNPs and primers used for nested RT-PCRs in both screening rounds are provided in Appendix Table 8.1.

3.4.1. Observed splice effects

Five of the 18 instances of allele-dependent splicing (28%) resulted in exon skipping. Insertions and deletions consequent to alternative or cryptic ss utilization were observed in the remaining 13 cases (72%). In particular, the majority of differential splicing events at donor sites comprised of exon skipping (4/7 = 57%) whereas cryptic splice site usage was

predominant at acceptor splice sites (10/10 = 100%). As for the occurrence, however, the computational prediction of the consequences of allele-dependent splicing was found to be poor (ElSharawy *et al.*, 2008). An overview of all confirmed allele-dependent splicing events is given in Table 3.4. In spite of the mutational imbalance at the lawful AG or GT ss (positive controls), the results outlined in Table 3.4 may indicate that mismatch at positions (+3, +4, +6, +14, +15, and –2) to donor and (–3, –6) to acceptor ss were not critical to splicing process compared to other positions (+1, +5, –1 and +2) and (–1, –2, –4, –7, –9, +2 and +3), respectively. The experimentally observed splicing effect of theses 18 SNPs has been submitted to dbSNP, and the technical prerequisites to accommodate this information are currently being established at dbSNP.

3.4.2. Allele-dependent splicing at NAGNAG tandem acceptors

In agreement with previous reports (Hiller *et al.*, 2004; 2006a; Hiller and Platzer, 2008), confirmed splicing effects of SNPs at tandem acceptor sites with a NAGNAG motif resulted in 3-nt insertions or deletions due to the use of the alternative AG dinucleotide. Four of the acceptor site SNPs chosen at NAGNAG tandem motifs exhibited allele-dependent splicing. In fact, *post hoc* analysis of the ss scores revealed that splicing was only affected if the stronger of the two AG dinucleotides was changed by the SNP (ElSharawy *et al.*, 2008). An overview of the NAGNAG SNPs and the corresponding ss score differences is given in Table 3.5.

3.4.3. Evaluation of the performance of F-SNP tool

The present study provided an opportunity to assess the postulation as to whether or not the combination of the 16 integrated bioinformatics tools and databases in F-SNP tool (Lee and Shatkay, 2008) (with each tool running its own distinct algorithms) was mandatory to achieve maximum sensitivity and was overall sufficient as a decision-making tool to screen for allele-dependent splicing, since their relative strengths might be additive while compensating for their weaknesses (Houdayer *et al.*, 2008). Thus, all the confirmed allele-dependent splicing effects in the present study (Table 3.4) were fed into the web-based server of F-SNP and the output was concurrently outlined in the same Table 3.4. As for the occurrence, the experimentally verified effects of allele-dependent splicing coincided with the computational predictions by F-SNP for only five of the 18 SNPs (28%) (ElSharawy *et al.*, 2008).

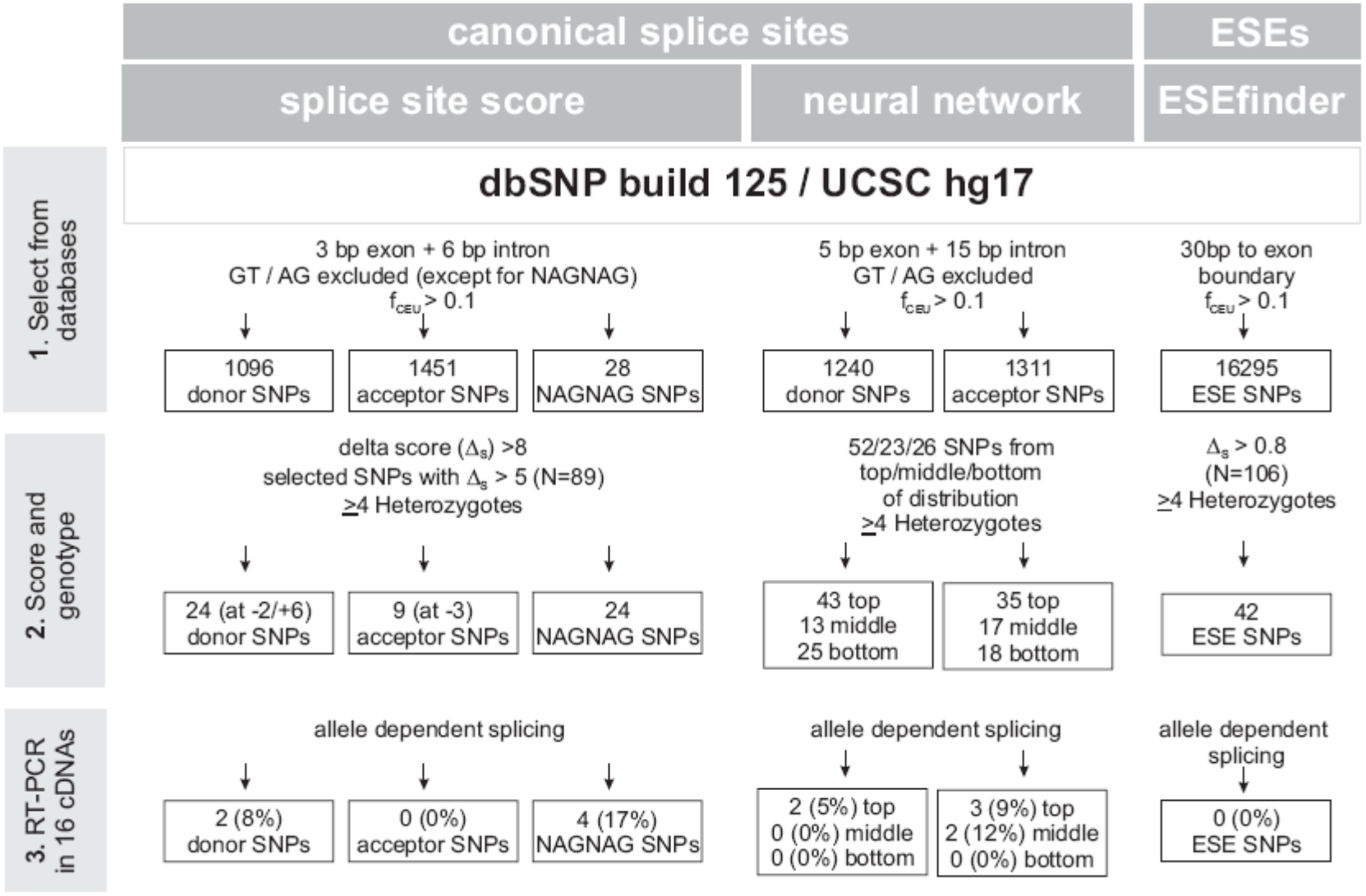

Figure 3.10 Graphical overview of splice SNP prediction in both screening rounds.
Different classes of putative splice SNPs are described in different columns; the time line of the study flows from top to bottom. The experimental steps and the main selection criteria employed in each step are given in the boxes on the left. In ***step 1***, SNPs were selected from dbSNP if they (i) had a minimum allele frequency of 0.1 in Caucasians and

(ii) were located near splice sites according to the criteria given in the top row. In ***step 2***, SNPs were scored using the bioinformatic tools given in the open block arrays in the middle panel. A total of 344 SNPs were genotyped in the panel of 92 DNAs. The numbers in the figure sum up to 397 (89 from Alex' splice site score calculator, 101 for each acceptor and donor from the neural network and 106 SNPs from the ESEfinder). The difference is due to 53 SNPs that were retrieved both with Alex' splice site score calculator and the neural network. As a result of genotyping, 223 non-redundant variants (including the 9 'positive controls') were tested by RT-PCR and sequencing in ***step 3***. Five of the positive splice SNPs overlap between the splice site scoring and neural network selection, which corresponds to the 13 events listed in this figure (8 unique splice events + 5 duplicates = 13). The ESE SNP (rs2274987:T>C) identified by chance is not marked in this figure but listed in Table 3.4. A detailed overview is provided in Appendix Table 8.1. The SNPs at the experimental validation stage in this figure add up to 255, which is due to the fact, that certain SNPs (N=41) overlapped between the neural network and splice scoring evaluations. Of the 24 donor SNPs, two (i.e. 8%) exerted an influence upon splicing. A notably higher rate of differential splicing (4/24, corresponding to a positive predictive value of 17%) was observed for SNPs at acceptor sites containing a NAGNAG motif. For none of the 42 investigated SNPs at ESE sites and 9 SNPs at position -3 of acceptor ss, an effect upon splicing could be observed. The positive predictive value of the neural network ranged from 5% (2 of 43) for donor to 9% (3 of 35) for acceptor ss SNPs. Two out of the 17 acceptor site SNPs from the middle range of the neural network (rs1558876:C>G and rs5248:A>G) were identified to exert an effect upon splicing despite a small allelic signal difference (0.00305 and 0.00244, respectively) (Table 3.4). In contrast to acceptor site SNPs from the middle range, no allele-dependent splicing was observed for donor site SNPs with a Δ_N value below the threshold for the top range (i.e. 0.1224). All the tested donor and acceptor SNPs from the bottom range of the neural network showed no effect upon splicing.

Table 3.4 Confirmed allele-dependent splicing events

The number given in the first column corresponds to the numbering used in Appendix Table 8.1. "NN scoring" refers to the neural network scoring approach. "Donor scoring" refers to the donor splice site score approach.

#	SNP ID	Δ_S	Δ_N	Splice Effect **- simple annotation** *- annotation according to HGVS*	F-SNP prediction **(Lee and Shatkay, 2008)**	Site	Position relative to splice site	Screening Category	Exon number/ size of adjacent exon	Gene	Refseq
15	rs1152522:A>G	16	0,97073	**3-nt deletion** *r.394_396del*	stop_gained	NAGNAG	**-2**	AG-variation at NAGNAG and NN scoring at top-acceptor	4/102	*C14orf105*	NM_018168
3	rs1558876:C>G	n/d	0,00305	**3-nt deletion** *r.705_707del*	frameshift_ coding; ESE-changed	NAGNAG	**+3**	AG-variation at NAGNAG and NN scoring at middle-acceptor	5/197	*ARSG*	NM_014960
11	rs2290647:G>A	n/d	0.0023	**3-nt deletion** *r.1070_1072del*	synonymous	NAGNAG	**+2**	AG-variation at NAGNAG	10/144	*GRAMD1A*	NM_020895
17	rs5248:A>G	0	0,00244	**3-nt insertion** *r.[210-3_210-1ins; 210-4a>g]*	frameshift_coding	NAGNAG	**-4**	AG-variation at NAGNAG and NN scoring at middle-acceptor	3/136	*CMA1*	NM_001836
35	rs12857479:G>A	16.1	0,98135	**278-nt insertion**[*] *r.[313-278_313-1ins; 313-1g>a]*	stop_gained	Acceptor	**-1**	Positive control at acceptor	4/157	*C13orf26*	NM_152325
36	rs10774671:G>A	16.1	0,97416	**98-nt deletion** *r.1039_1136del*	no functional information	Acceptor	**-1**	Positive control at acceptor	6/514	*OAS1*	NM_016816
37	rs3818780:C>G	16	0,89773	**2-nt deletion** *r.-10_-9del*	stop_gained	Acceptor	**-1**	Positive control at acceptor	2/297	*AVPI1*	NM_021732
38	rs1805377:G>A	16	0,80319	**6-nt deletion** *r.894_899del*	stop_gained	Acceptor	**-1**	Positive control at acceptor	8/589	*XRCC4*	NM_022406
34	rs11658717:G>A	0.1	0,99341	**6-nt ins** *r.[288-7_288-1ins; 288-7a>g]*	stop_gained	Acceptor	**-7**	NN scoring at top-acceptor	6/211	*STXBP4*	NM_178509
71	rs330924:G>C	3.8	0,14027	**8-nt insertion** *r.[-17-8_-17-1ins; -17-9c>g]*	conserved	Acceptor	**-9**	NN scoring at top-acceptor	2/5,414	*PPP1R3B*	NM_024607
91	rs3816989:G>A	18.2	0,99659	**Exon 4 skipping** *r.212_336del*	stop_gained	Donor	**+1**	Positive control at donor	4/125	*TCTEX1D1*	NM_152665
117	rs764497:T>A	18.2	0,9901	**Exon 1 skipping**	stop_gained	Donor	**+2**	Positive control at donor	1/136	*CCDC149*	NM_173463

				r.-238_-103del							
90	rs10101626:G>T	18.2	0,98826	**Exon 19 skipping** *r.2641_2835del*	stop_gained; conserved	Donor	**+1**	Positive control at donor	19/195	*WDR67*	NM_145647
119	rs2276611:G>A	18.2	0,97768	**7-nt insertion** *r.[-70+1_-70+7ins; -70+1g>a]*	stop_gained	Donor	**+1**	Positive control at donor	1/151	*PPIG*	NM_004792
120	rs482308:G>A	18.1	0,94543	**111-nt deletion** *r.6512_6622del*	stop_gained	Donor	**+1**	Positive control at donor	35/305	*ZAN*	NM_003386
96	rs2298839:A>G	14.4	0,78973	**Exon 7 skipping**[**] *r.714_843del*	frameshift_coding	Donor	**+5**	Both NN and donor scoring	7/130	*AFP*	NM_001134
99	rs2076530:A>G	12.5	0,12612	**4-nt deletion** *r.1075_1078del*	nonsynonymous; ESE-changed	Donor	**-1**	Both NN and donor scoring	5/348	*BTNL2*	NM_019602
-	rs2274987:T>C	-	-	**New exon insertion from intron2** *r.119_120ins119+903_119+996*	no functional information	ESE	**+25**	Donor SNP (rs3816989:G>A)	3^{new}/94	*TCTEX1D1*	NM_152665

- del: deletion; ins: insertion; ss: splice site; skip: skipping; NN: neural network; Refseq: reference sequence; HGVS: The Human Genome Variation Society; nt: nucleotides.
- Δ_S : absolute allelic difference ss scores as calculated for each SNP using Alex's online splice score tool; Δ_N: absolute signal difference as calculated from the signals emitted from the neural network.
- n/d: score cannot be determined because the SNP is located outside of the scope of Alex's splice site score calculator, which includes only one exonic nucleotide for the acceptor;
- **[**]:** A 116- nt insertion (*r.[843+1_843+116ins; 843+5g>a])* was also observed but in only one heterozygote and exon 7 skipping *(r.714_843del)* was observed in all other cDNA samples with allele A. In total, five homozygotes for both alleles and six heterozygotes were tested.
- [*]: this 278 nt insertion was only seen in one heterozygote with rare allele A at rs12857479 that disrupts the canonical acceptor ss of exon 4 at *C13orf26* gene.
- SNP rs2274987:T>C is the only instance of allele-dependent splicing at an ESE site; it was identified by chance while analyzing a nearby donor site at rs3816989:G>A.
- HUGO HGNC-approved gene symbols (http://www.genenames.org/) were used in this table.
- The mutation and splice effect nomenclature appears in this table to follow the format indicated in the HGVS (see the website http://www.hgvs.org/mutnomen/).

Table 3.5 Functional effects and splice site scores of NAGNAG SNPs

A splice site score was calculated for both alleles. In the column with the *post hoc* score, the experimentally verified splice site was used for the prediction. The difference in scoring can be attributed to the sequence window used by the splice site scoring tool, which includes 2 nt of the exon and 6 nt of the intron for donor sites and 14 and one nt for intron and exon at the acceptor, respectively. Thus, the score depends on the *a priori* position of the splice site. One column comments on the *post-hoc* scores in many instances, an alternative splice site as compared to the one annotated in the RefSeq was used. This is described in the respective column. However, these splice events were mostly invariable (except #19) and not genotype-related. The number given in the first column corresponds to the number in the Appendix Table 8.1.

#	SNP ID	Observed splice effect	Exon No.	SNP in scored sequence	Initial scores* Allele1/allele2/ Δ_S	*Post-hoc* sequence	*Post-hoc* Scores* Allele1/allele2/ Δ_S	Comment on *post-hoc* scores	Gene symbol
15	rs1152522:A>G	**CAG del** *(r.394_396del)*	4	GTTGTCTTTCAT**R**GC	81.7/65.7/16	GTCTTTCAT**R**GC**AG**G	91.7/65.7/26.0	Scores after CAG del	*C14orf105*
17	rs5248:A>G	**CAG ins** *(r.[210-3_210-1ins; 210-4a>g])*	3	CTTCTTCCTCAC**AR**C	75.4/91.4/16	CTTCCTCACA**R**C**AG**G	75.4/91.6/16.2	Scores after CAG ins	*CMA1*
11	rs2290647:G>A	**CGG del** *(r.1070_1072del)*	10	TCTGTCTCCAGC**R**GA	73.7/89.8/16.1	TCCTCTGTCTCC**AG**C	73.7/87.7/14.0	Score after CGG ins	*GRAMD1A*
3	rs1558876:C>G	**CAC del** *(r.705_707del)*	5	TCCTGTTTCAGC**A**SC	70.4/86.4/16	TTCTCCTGTTTC**AG**C	70.4/89.6/19.0	Score after CAC ins	*ARSG*
1	rs17105087:A>G	None	7	CTCTTCTGCAGC**AR**C	69.5/85.5/16	TCTCTCTTCTGC**AG**C	89.6	Score after CAR ins	*SLC25A21*
2	rs11597439:C>G	None	2	TGTCCCTTCAGA**A**SA	61.0/77.0/16	CTGTGTCCCTTC**AG**A	90.9	Score after AAR ins	*CUEDC2*
4	rs9606756:A>G	None	2	TCTTTTCTAAGA**AR**T	63.1/79.1/16	TTTTCTTTTCTA**AG**A	83.9	Score after AAR ins	*TCN2*
5	rs1152888:A>G	None	5	TCCTTCCTAAGG**AR**T	58.2/74.2/16	CTTTCCTTCCTA**AG**G	83.6	Score after GAR ins	*IRAK3*
6	rs17036879:G>A	None	8	AATAACTTTAGG**AR**C	62.1/46.0/16.1	GTAAATAACTTT**AG**G	66.4	Score after GAR ins	*TSEN2*
7	rs2156634:G>A	None	3	TTTTGCTGCAGG**AR**A	75.4/59.4/16	CAGTTTTGCTGC**AG**G	89.7	Score after GAR ins	*GRIK4*
8	rs3014960:G>A	None	14	TCTTTATACAGC**AR**A	87.4/71.4/16	ATTTCTTTATAC**AG**C	91.0	Score after CAR ins	*COG3*
9	rs4822258:G>A	None	8	CCCGTCACCAGG**AR**G	70.7/54.6/16.1	TTTCCCGTCACC**AG**G	92.2	Scoe after GAR ins	*TTLL1*
10	rs2243603:C>G	None	5	CTGATTTCCAGA**A**SC	54.8/70.8/16	TCCCTGATTTCC**AG**A	88.8	Score after AAS ins	*SIRPB1*
12	rs2273431:G>A	None	10	TACTCATGCAGA**RG**A	51.3/67.3/16	CTCTACTCATGC**AG**A	88.1	Score after ARG ins	*NID2*
13	rs7862221:A>G	None	14	TTTCTCTTCAGA**RG**A	80.4/64.4/16	TTGTTTCTCTTC**AG**A	94.5	Score after ARG ins	*TSC1*
14	rs2275992:A>G	None	5	CTTATTTTTAGT**RG**T	81.2/65.2/16	TTACTTATTTTT**AG**T	81.5	Score after TRG ins	*ZFP91*

16	rs2307130:A>G	None	2	TTCAAATCCTCT**R**G**A**G	76.1/60.1/16	TTTTTGTTTCAT**AG**G	90.7	Score after 60 nt ins	*AGL*
18	rs2292402:T>A	None	2	TGTGTTTGG**W**GC**AG**T	81.1/75.0/6.1	-	Exon 2 skipping	Scoring is not possible	*ACPL2*
19	rs2071558:C>T	None	6	AGTGTCCC**Y**AGC**AG**G	84.5/86.0/1.5	CACAGTGTCCC**Y**AG**C**	73.3/65.0/8.3	Scores after CAG ins; Exon skipping and intron retention independent of the SNP genotype were observed	*AMHR2*
20	rs12905385:C>T	None	20	CTTCACTGATA**Y**AG**G**	87.2/78.9/8.3	CACTGATA**Y**AGG**AG**A	60.7/62.1/1.4	Scores before GAG ins	*CDAN1*
21	rs2250205:C>T	None	5	TCTTTGATTGA**Y**AG**G**	92.4/84.0/8.4	TTGATTGA**Y**AGG**AG**A	64.6/66.0/1.4	Scores before GAG ins	*EIF6*
22	rs2174769:T>C	None	3	TGTTTGAATTT**Y**AG**G**	81.5/89.9/8.4	TTGAATTT**Y**AGG**AG**C	68.5/67.1/1.4	Scores before GAG ins	*SNIP1*
23	rs12944821:G>C	None	3	CTTTATATTTTC**AG**S	97.0/91.3/5.7	TATATTTTCAG**S**AG**G**	79.2/91.0/11.8	Scores before SAG ins	*AP1GBP1*
24	rs879022:G>A	None	3	TCCCAGGACAG**R**AG**G**	62.7/63.0/0.3	TTTTCCCAGGAC**AGR**	90.2/86.8/3.4	Scores after RAG ins	*REG1P*

- Ins: insertion; del: deletion; Δ_S : absolute allelic difference ss scores.
- * : Splice site scores are provided here according to allele-order occurrence from left to right as annotated in SNP-ID column.
- The splice site score is calculated with Alex's splice site score calculator (http://violin.genet.sickkids.on.ca/~ali/splicesitescoreForm.html)
- The consensus AG dinucleotides in RefSeq, as listed in supplementary Table S1, are given in bold type.
- The SNP ambiguity codes are underlined and given in bold type.
- HUGO HGNC-approved gene symbols (http://www.genenames.org/) were used in this table.
- The mutation and splice effect nomenclature appears in this table to follow the format indicated in the the Human Genome Variation Society (see the website for HGVS: http://www.hgvs.org/mutnomen/).

3.5. Establishment of a novel *in vitro* splice reporter system

In order to overcome the shortage of the currently available prediction tools of allele-dependent splicing, a new reporter system was designed in the present study. The reporter system should meet three main required features. First, it should be suitable for high-throughput screening of alterantive splicing. Second, it should have a broad dynamic range, allowing measurement of impact of *cis*-acting DNA variations at different splice-related locations, such as donor and acceptor splice sites, ESE, etc. Third, the system should distinguish changes in AS patterns from changes in transcription and translation. Toward this end, a number of pilot experiments were done in the present study.

3.5.1. Insertion of test genomic region and coding sequence of RFP: Optimization

The test genomic region, which comprised exon 7 to exon 9 (1864 bp) of *PGM2L1* gene was inserted at the MCS of pEGFP-N1' vector (Figure 2.5). The ATG-start codon of GFP of the resulting hybrid vector was then eliminated using site-directed mutagenesis, in order to avoid any internal translation initiation. The FACS analysis of the produced construct ('PGM2L1-pEGFP-N1' vector) revealed no green fluorescence. This might be due instability of the transcript or a mis-splicing of the inserted genomic region. To overcome this problem, the coding sequence of the RFP was amplified by PCR from pDsRed2-N1 vector and inserted at the *XhoI* site at the 5´ end of the genomic region in the MCS of the produced construct. The plasmid map of this construct, namely 'RFP-PGM2L1-pEGFP-N1', is provided in Figure 3.11. Using FACS analysis, the last construct showed the typical green-fluorescence pattern of GFP, which confirmed the correct splicing of the inserted test genomic region and the 'stabilizing' function of the cds of RFP.

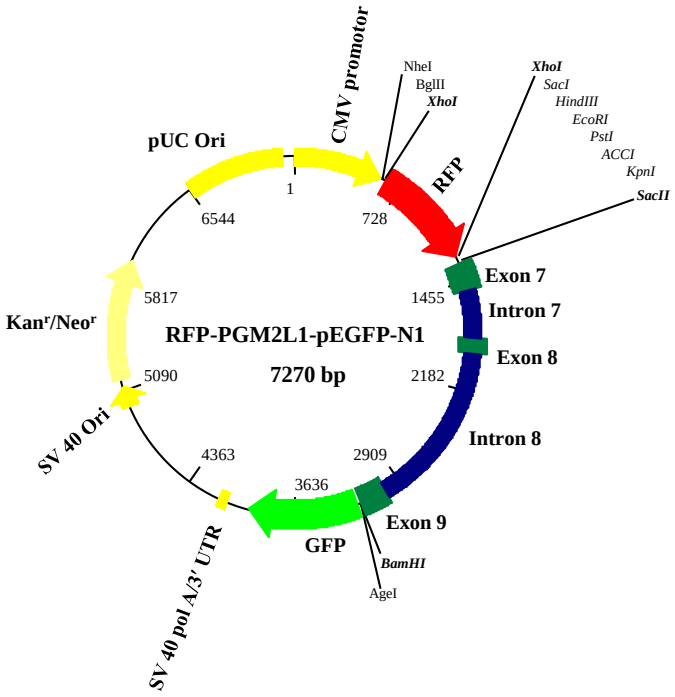

Figure 3.11 Insertion of the cds of RFP into the MCS of pEGFP-N1 vector.
The cds of RFP (672 bp) amplified by PCR from pDsRed2-N1 vector was inserted at the unique *XhoI* restriction site at MCS of the produced PGM2L1-pEGFP-N1 vector. Here, the MCS is located between 591 and 1343 bp. The resulting hybrid vector (7.270 kb) contains the cds of RFP at position 619-1290 bp and of GFP at 1354-2073 bp.

3.5.2. Functional validation: A fluorescence-based detection method for comprehensive analysis of splice site mutations

To test the validity of the developed splice reporter construct (RFP-*PGM2L1*-pEGFP-N1), respective donor and acceptor ss of the test exon 8 of the inserted *PGM2L1* genomic region, were separately mutated. In one construct, the original GT dinucleotides at donor ss (at the start of intron 8) were knocked out and mutated to CA dinucleotides. This mutation was predicted (by Alex's Splice ss calculator) to diminish the conservation level at this donor ss with 36.5 points (donor ss score with obligatory GT was 76.3 and with CA-dinucleotide was 39.8). In the second construct, the obligatory AG dinucleotides at acceptor ss (at the end of intron 7) were mutated and converted to TC dinucleotides. Due to the last modification, a similar decrease of the saturation at this acceptor ss was also predicted by Alex's ss score calculator (acceptor ss score with AG was 85.6 and with TC was 53.6 points; resulting delta score was 32 points). After transfection of the constructs into human Hela cells, the FACS-analysis of the respective fluorescence signals (Figure 3.12) indicated that, transcripts that retained wild-type *PGM2L1* expressed functional GFP. In contrast, both types of ss mutations (Δ-acceptor and Δ-donor of exon 8 of PGM2L1) exhibited almost completely abolished

expression levels of GFP as a result of the frame shift introduced into ORF of GFP. To confirm these results, expression of the protein, which spans RFP, PGM2L1, and GFP, was confirmed by immunoblot analysis (Figure 3.13). Here, the expected protein was only expressed with the wild-type *PGM2L1*, and not with other mutated constructs (Δ-acceptor and Δ-donor of exon 8 of PGM2L1; Figure 3.13. Thus, the modulation of GFP expression level can be readily interpreted and used as a sensitive screening tool to access the impact of variations at different splice-relevant positions (exonic and intronic) on measuring splicing efficiency.

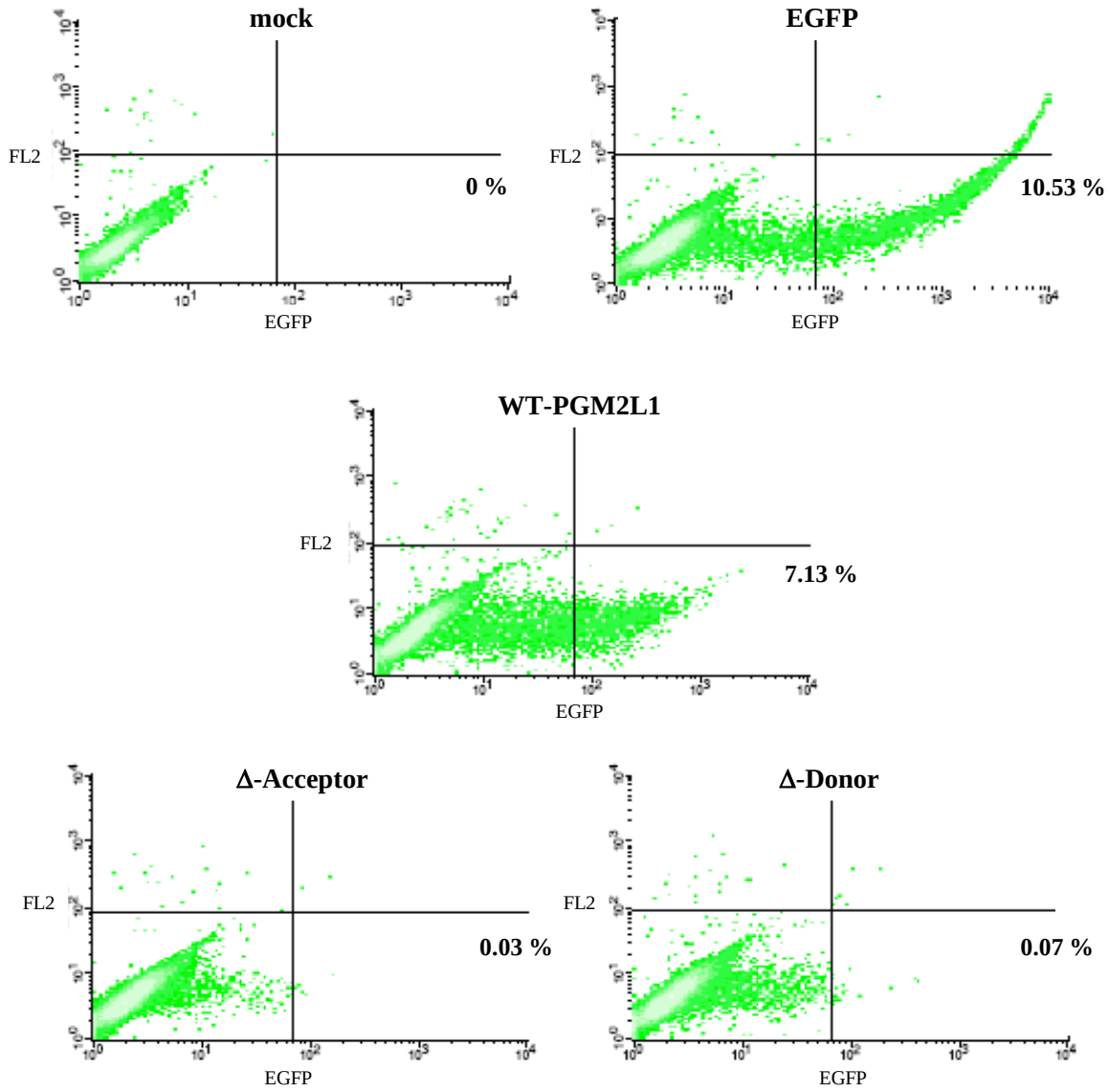

Figure 3.12 A fluorescence-based detection method for comprehensive analysis of splice site mutations: Results of FACS analysis.
HeLa cells were separately transfected with wild-type (WT-PGM2L1 of RFP-PGM2L1-pEGFP-N1 construct) and mutant (Δ-acceptor and Δ-donor of exon 8 of PGM2L1) constructs, EGFP, and dsRed2 vectors. Untransfected (mock) cells were used as a negative control. Fluorescence channel FL1 for detection of GFP

(emission 530±15 nm) is plotted on the X-axis against the FL2 fluorescence channel (emission: 585±15nm) on the Y-axis. The percentages of cells exhibiting a defined fluorescence signature were calculated by applying quadrant statistics. This data indicate that, 1) GFP-fluorescence level can be used to differentiate between wild type and mutant constructs (mutant of either acceptor or donor ss markedly reduced, or abolished, the number of cells with GFP green fluorescence as shown in the lower right quadrant of each window); 2) the red fluorescence of RFP was not detected in channel FL2 in the presence of GFP.

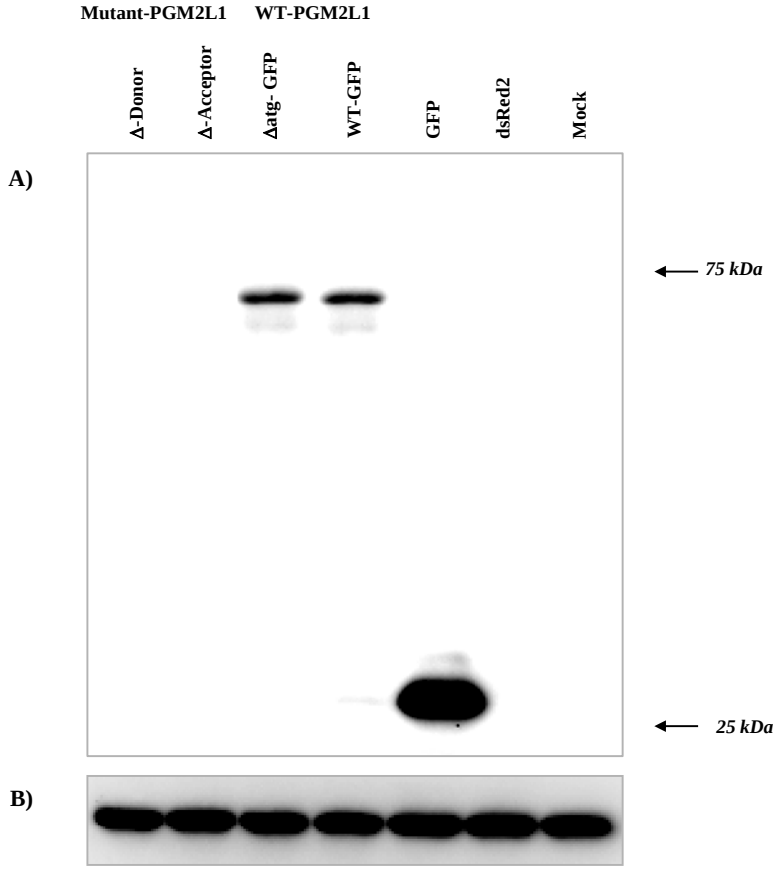

Figure 3.13 Immunoblot analysis of wild-type and mutant constructs.
The HeLa cells were separately transfected with wild-type splice reporter (WT-PGM2L1 of RFP-PGM2L1-pEGFP-N1 construct) with and without the ATG-start codon of the GFP coding sequence, and mutant splice reporter (with Δatg of GFP) at acceptor (Δ-Acceptor) and donor (Δ-Donor) ss of exon 8 of PGM2L1. In addition, EGFP vector was transfected as a positive control and dsRed2 vector and untransfected HeLa cells (mock) were used as negative controls. After protein lysate preparation, western blot analysis was performed using *A.V.* GFP monoclonal antibody (JL-8; 1:1000) **(A)** and the blot was then stripped and reported for β-actin **(B)**. The blot **(A)** shows the expression of the GFP-fusion protein (69 kDa; spans dsRed2 cds-PGM2L1 ex7/ex8/ ex9- GFP cds) from the WT reporter and not from the mutant ones.

4 DISCUSSION

4.1. Characteristics of the applied approach

After the sequencing of the human genome and the ongoing large-scale SNP-discovery programs, the annotation of SNPs with putative biological effects is one of the largest remaining genomic challenges. Here, the impact of SNPs on the RNA-phenotype is one of the major mechanisms under intense investigation. In the present study, a systematic, SNP-centered approach has been followed in order to identify germline genetic variations that have a potential effect upon pre-mRNA splicing (ElSharawy *et al.*, 2008). The applied approach is similar to that of an epidemiologist. Whereas a genetic epidemiologist first identifies candidate mutations in a disease gene and then looks for functional interpretation, the applied approach identifies the SNP and evaluates the SNP's effect on splicing. These marker-driven research paradigms pose a daunting challenge. Genome-wide association studies, e.g. for colon and prostate cancer (Tomlinson *et al.*, 2007; Zanke *et al.*, 2007; Zheng *et al.*, 2008), have discovered a multitude of genetic loci, only a few of which have lent themselves to an immediate functional interpretation.

Despite many focused studies on the functions and mechanisms of alternative splicing (AS) that are associated with specific transcripts, high-throughput experimental approaches for systematically elucidating the extent of functionally relevant AS events are only very recently beginning to be used (Blencowe, 2006). The possible impact of germline polymorphisms on mRNA splicing has previously been analysed from a transcript or exon perspective. In these early studies, transcripts were screened for evidence for AS by bioinformatics or experimental means, and candidate variants were tested for AS using isoform-specific PCR (Hull *et al.*, 2007) or chip-based methods (Kwan *et al.*, 2007; Kwan *et al.*, 2008). Integrated analysis of genomic polymorphisms at EST and exon array data also revealed evidence of allele-specific splicing (Nembaware *et al.*, 2008). These studies have unequivocally established the relevance of common SNPs for mRNA splicing. Other systems-wide experimental profiling methods of AS, such as splice junction or tiled genomic arrays, have very recently started to define how global splicing regulation shapes complex biological properties and pathways (Ben-Dov *et al.*, 2008; Moore and Silver, 2008). The main goals of such efforts are producing comprehensive catalogs of splice variants in different organisms and cell types, defining *cis/trans*-acting splicing factors, and characterizing the response of splicing to signaling pathways, differentiation, and disease states (Moore and Silver, 2008). However, variation in

splicing pattern across tissues is probably controlled by the availability of *trans*-acting splicing factors, and the use of public transcript data, such as ESTs, to estimate either the tissue-specificity or the allele-dependent splicing of transcript isoforms is complicated by the fact that multiple overlapping ESTs from the same individual are present in dbEST (Nembaware *et al.*, 2004). EST data often have poor coverage (i.e. only a small number of ESTs from a given tissue for a region of interest in a gene) and many sampling artefacts. For example, there can be dramatically different numbers of ESTs from different libraries or tissues, creating sample bias (Xu *et al.*, 2002). Despite this drawback, EST sequences provide information on the structure of alternative isoforms and include data from different gene expression contexts. However, this information is highly biased towards ends of genes and is sparse for all but the most highly expressed genes (Nembaware *et al.*, 2008).

To this end, a high-throughput methodology (Figure 3.1) based on using a panel of 92 matching pairs of individual-specific DNA and cDNA samples (ElSharawy *et al.*, 2006), was established in the present study in order to: i) assess the impact and overall importance of naturally occurring SNPs on differential splicing; and ii) to evaluate (ElSharawy *et al.*, 2008) and improve, current prediction tools of allele-dependent splicing if necessary. The established approach combined the use of optimized wet-lab protocols and computational facilities. Combining technologies has been very fruitful to identify regulatory sequences, including motifs involved in tissue-specific AS (Ben-Dov *et al.*, 2008). The applied method is supported by a package of four helpful softwares, which were created during the course of the present study (see Methods and Results sections). Indeed, correlating splicing patterns and SNPs is a labor-intensive undertaking. There are several competing methods available for the evaluation of potential splicing effects of naturally occurring genetic variations. The direct use of cDNAs from sources with known genotypes provides a very attractive option, especially for the investigation of SNPs in larger scale systematic experiments (ElSharawy *et al.*, 2008). Due to the current lack of precise prediction methods for the splicing effects of SNPs, and in order to make the selection of candidate SNPs an efficient process, it was necessary to develop new software. Therefore, SNPSplicer screening tool was created (ElSharawy *et al.*, 2006) to evaluate whether a potential site-specific splice effect is present in a given sequence. If allele-dependent splice variation occurs, homozygotes show clean traces that substantially differed depending on the underlying splicing effect. Heterozygote traces show an apparent breakdown of sequence quality starting from the site of splice effect, due to an overlay of two very different sequences. Depending on the relative amounts of the alternative transcripts, the

trace pattern may resemble one of the underlying homozygote sequences (ElSharawy *et al.*, 2006). Because of the potentially large deviations of the observed sequence from the cDNA reference, alignment and interpretation of such trace collections is extremely labor intensive with standard sequence alignment software (e.g. Sequencher). SNPSplicer facilitates easy visualization of potential splicing patterns, which can then be experimentally verified by subcloning and sequencing of PCR products, especially in the case of complex splice effects.

The presented experimental approach therefore incorporates several required advantages:

- Modern high-throughput genotyping methods (SNPlex/TaqMan) are used to determine genotypes. cDNAs corresponding the respective genotypes were robotically or automatically selected (using SpliceTool). This minimizes the required number of corresponding matched cDNA samples for RT-PCR (ElSharawy *et al.*, 2008), thereby making the correlation of an RNA-phenotype with a SNP-allele easier, cheaper and faster.
- SNPSplicer software rapidly interprets whether or not an allele-dependent splicing signal is present. This helps to focus resources on the most promising splice SNPs for future functional and/or mechanistic analyses (ElSharawy *et al.*, 2006).
- The splicing effects are directly evaluated in the tissues of interest. This eliminates the confusion resulting from the induction of *trans*-acting splicing regulatory factors (Xu *et al.*, 2002; Grosso *et al.*, 2008). Here, the splicing patterns can be evaluated for robustness across tissues (ElSharawy *et al.*, 2006).
- The insert size restrictions for construct-based splicing assays do not apply; therefore potentially more complex splicing events can be detected and investigated (ElSharawy *et al.*, 2006; ElSharawy *et al.*, 2008).

Clearly, minigene constructs provide a more defined and experimentally controlled system, which may need to be used for ultimate mechanistic clarification to determine the effects of SNPs on splicing (Niksic *et al.*, 1999; Pagani *et al.*, 2002; Baralle *et al.*, 2003; Lewandowska *et al.*, 2005). This classic mechanistic approach has certain limitations, especially low throughput, limited insert size, and thus incomplete detection of long-range effects. The problems associated with the PCR amplifications of cDNA samples containing differently spliced isoforms, are also known. In such cases, a preferential amplification of one or few, mostly the shorter isoforms, may be observed (Zhu *et al.*, 2003). However, especially for

SNPs with weak *a priori* evidence of a splicing effect, the approach described above may represent a very effective screening method (ElSharawy *et al.*, 2008). On the other hand, microarray-based approaches can analyze the splicing patterns of many thousands of exons and have been used to distinguish splicing patterns seen in different tissues. Interpretation is complex, and for some arrays sensitivity is low and false positive rates are high. Although it is likely that the technology will improve, these approaches have not yet been shown to have the sensitivity to detect the level of variation particularly for low-abundance isoforms (Hull *et al.*, 2007). The advantage of the system described in the present study is targeted amplification of the splicing event of interest, which may provide greater sensitivity. On the other hand, depending on the exact location of probesets in a given gene, many of the transcript isoforms that occur, particularly those that affect donor or acceptor sites but do not cause exon skipping or inclusion, are undetectable using exon arrays. When alternative isoforms are distinguishable using the exon arrays, they still provide little information on the nature of the isoforms, and this may need to be inferred either by integrating information from other sources or experimentally (Nembaware *et al.*, 2008). The combination of the array-based approach, supplemented with splice junction probes and replication of the produced positive hits using the approach developed in the present study, comprises a strategy to improve interpretation of results, increase sensitivity and provide a means of accessing the causes of differential gene expression in a genome-wide scale.

Direct sequencing of PCR products from cDNAs with sources of known genotype was used as a screening tool within the established methodological pipeline (ElSharawy *et al.*, 2006; ElSharawy *et al.*, 2008). The application of this approach was mainly based on: (1) the observation that the degree of NMD is moderate and only one-third of reliably inferred alternative mRNA isoforms are suggested to be candidates of NMD (Stamm *et al.*, 2000; Lewis *et al.*, 2003); and (2) the in-house control experiments, which showed that potentially down-regulated (minor) splice variants with a frequency as low as 20%-10% of the total transcripts can be clearly detected by direct sequencing (Appendix Figure 8.1). In the present study, the presence of a second, alternatively spliced transcript was readily detectible in heterozygote state down to the 80%:20% or 90%:10% range (Results section 3.1.4; (ElSharawy *et al.*, 2006)). Even if this was not true in some instances, the analysis of homozygotic cell lines, one for each allele, allowed clear-cut detection of allele-dependent splicing (ElSharawy *et al.*, 2008). Thus, direct sequencing is a robust and sensitive screening tool, even in the presence of NMD. In fact, the use of this approach helped to reduce the

number of cloning and clone-based sequencing experiments, thereby saving time and labor. Combining RT-PCR and direct sequencing-based approach has also been successfully applied in many other studies. For instance, it has been used to distinguish functional from non-functional GYNGNY tandem donors, without affecting the detection of alternative transcripts that were expressed at a low level (down to 10%) (Hiller *et al.*, 2006b).

The method presented here is not restricted to variations in particular components of the splicing recognition sequences. As long as a robust RT-PCR spanning the site of a potential splicing effect can be designed, the approach presented in the current effort is applicable in principle. This is demonstrated in the Results (section 3.1.4) for donor ss and ESE variation (ElSharawy *et al.*, 2006); but, in addition, intronic SNPs and long-range *cis*-acting effects were also accessible as long as a putative site of effect in the transcript could be predicted (ElSharawy *et al.*, 2008). The requirement of a robust RT-PCR, however, precludes the analysis of the transcription start site (i.e., promoter inactivation through SNPs) and variation affecting the polyadenylation site (i.e., RNA stability in general). For these SNPs, methods such as the analysis of allelic imbalance through pyrosequencing, for instance, are more appropriate (Cowles *et al.*, 2002; Wojnowski and Brockmoller, 2004).

The methodology of the present study was careful to avoid bias against certain candidate genes. Thus candidate splice SNPs were evaluated regardless the reported splice variants at specific loci, suggesting that the scoring-based approach together with the laboratory-observed RT-PCR products from different genotypes is the key direction for correlating any splice variation to a SNP under investigation (ElSharawy *et al.*, 2008). Depending only on available gene/mRNA annotations, or the exclusion of genes with only one documented or observed transcript isoforms as has been considered in a similar recent study (Hull *et al.*, 2007), might negatively influence the detection of splice-relevant SNPs. For instance, splicing variation around the location (exon 35) of one of the confirmed allele-dependent splicing effect in the present study (rs482308 at zonadhesion (*ZAN*) gene) is neither reported (Gasper and Swanson, 2006), nor yet documented at UCSC genome browser. The resulting splice effect is quite predictable, as the minor A-allele at rs482308 disrupted the obligatory G nt at position +1 of exon 35 and associated with the utilization of a cryptic donor ss 111 nt upstream of this exon (ElSharawy *et al.*, 2008) (Table 3.4). *ZAN* protein is also of biological importance, since it is involved in the acrosome reaction, which is a key recognition event in animal fertilization process (Vacquier, 1998; Tanphaichitr *et al.*, 2007). Yet again, the wide

range for the predicted frequency of splicing mutations (15-60%), which expose the fact that mRNAs from mutant alleles are rarely assayed for splicing abnormalities (Wang and Cooper, 2007), promotes the application of a rational strategy similar to that of the present study. Moreover, a significant fraction of mutated alleles in both recessive and dominant conditions has not been identified, and the availability of RNA samples from affected individuals and their families is often problematic (Buratti *et al.*, 2007).

Taking the advantages of the applied approach in mind, broad application of this methodology is anticipated to help in the functional annotation of SNPs, thereby providing an important contribution to the understanding of the impact of genetic variation on natural phenotypic variability and disease susceptibility in a high-throughput fashion.

4.2. Prediction rate of allele-dependent splicing

Predicting potential splicing effects of SNPs is currently complicated by several factors. Most information on the sequences involved in splicing has been obtained from the alignment of genomic sequences to expressed sequence tags (ESTs) and known gene models (Krawczak *et al.*, 1992; Clark and Thanaraj, 2002). Splice mutations of known disease relevance have also been investigated (Stenson *et al.*, 2003). Specific ss have been studied mechanistically in depth (Pagani *et al.*, 2000; Wang *et al.*, 2004a; Zuccato *et al.*, 2004), mostly because the affected genes were of particular disease importance. Overall, the available empirical data on allele-dependent splice variation is still limited.

Using current *in silico* splice prediction tools to decide if a biallelic SNP has an impact on splicing process requires an analytical attitude, given that no interpretation guidelines are available. This means that the user must decide himself when a prediction is reliable (i.e., likely allele-dependent splicing effect) or not (no expected influence on splicing). This absence of interpretation guidelines is in part explained by the fact that efficient recognition of ss in higher eukaryotes by the spliceosome is mediated through multiple parameters other than the strength of the ss, such as the exon/intron architecture, the presence or absence of splicing enhancers or silencers, the presence or absence of local RNA secondary structures, and the process of pre-mRNA synthesis by RNA polymerase II. Each component contributes to the overall affinity of spliceosomal components to the exon, and thus, the level of exon inclusion (Hertel, 2008). As a result, splicing outcome does not only depend on the nucleotide variation of these consensus sequences (Houdayer *et al.*, 2008). In this respect, delta splice

site scores (Δ_S and Δ_{ESE} for web-based tools, and Δ_N for neural network) were considered to determine whether a SNP had an effect upon splicing, rather than looking at the ss scores *per se* (ElSharawy *et al.*, 2008). The score may not reflect the true strength of the ss because of the other contributing components involved in splicing process. In particular, this is especially important when interpreting effects of SNPs at loosely defined positions (i.e., with expected low splice scores). A similar score strategy has been considered and applied in a concurrent recent study (Houdayer *et al.*, 2008) for evaluating *in silico* splice tools for decision-making in molecular diagnosis.

In the present study, the non-redundant 223 candidate splice SNPs retrieved from dbSNP were experimentally tested using the established approach in a panel of 92 matching gDNAs and cDNAs (ElSharawy *et al.*, 2006). The rate at which allele-dependent splicing was correctly predicted in the present study was low; the positive predictive value of the respective bioinformatics tools ranged from 0% to 9% (ElSharawy *et al.*, 2008). At least in part, these surprisingly small success prospects may be explicable in terms of the design of the present experiment. Since naturally occurring genetic variation formed the basis of the analysis, only candidate splice SNPs with a sufficiently high degree of heterozygosity could be verified experimentally in the utilized gDNA-cDNA panel. Furthermore, the type of source tissue (mainly lymphoblastoid cell lines and peripheral blood) may also have played an important role (ElSharawy *et al.*, 2008). However, the existence of allele-dependent splicing was readily confirmed for nine SNPs affecting the conserved AG or GT dinucleotides of canonical ss, thereby corroborating the scientific rational of the applied approach (ElSharawy *et al.*, 2006). The concordance between the *in vitro* splicing findings and *in silico* prediction tools at the obligatory AG/GT dinucleotides, is not surprising, as anomalous splicing is highly expected when the disruption occurs at these canonical dinucleotides. This deleterious impact has widely been recognized for a long time and this knowledge was taken into account when designing the algorithms running in these tools (Houdayer *et al.*, 2008). As outlined in Table 3.4, Alex's splice score tool, as well as the neural network approach, provided reliable scores; all mutant canonical donor and acceptor ss showed strong score variations ($18.2 \geq \Delta_S \geq 16.0$ and $0.99659 \geq \Delta_N \geq 0.80319$, respectively) compared to the applied arbitrary threshold in both screening rounds ($\Delta_S > 8.0$, and $\Delta_N > 0.1224/0.14$, respectively) and splicing defects were supported by *in vitro* evidence.

The performance of the bioinformatics tools used in the present study was primarily dependent upon the degree of conservation of the corresponding target sequence. Splicing-relevant motifs are often short and poorly conserved. This inherent drawback is highlighted by the fact that a notably higher positive predictive value was obtained when a better-defined sequence like the tandem acceptor (NAGNAG) is included in a screen for splice SNPs. Thus, 17% of the SNPs initially selected at NAGNAG acceptor ss showed allele-dependent splicing. This result was likely due to the alteration of the highly conserved AG dinucleotide motif in least some cases. In fact, a *post-hoc* analysis of the ss scores at acceptor sites showed that splicing was only affected if the stronger of the two AG dinucleotides was altered by the SNP (Table 3.5). For the less well-defined canonical donor and acceptor ss lacking a tandem structure (Hiller *et al.,* 2006b; Hiller *et al.,* 2007a), the positive predictive value of the bioinformatics tools was generally poor (ElSharawy *et al.,* 2008) (Table 3.4, Figure 3.10).

The lowest positive predictive value was obtained for putative splice SNPs in exonic splicing enhancers (ESEs). Indeed, the single instance of allele-dependent splicing due to such a SNP was serendipitously found while analyzing a nearby donor site variant (rs3816989:G>A) (ElSharawy *et al.,* 2008). As argued above, the failure to detect splicing-relevant SNPs at ESE is most likely due to the poor definition of enhancer motifs, both with regard to their sequence and position. The available information is integrated in ESEfinder (Liu *et al.*, 2001; Cartegni and Krainer, 2003; Cartegni *et al.*, 2003; Smith *et al.*, 2006). Retrospectively, this software tool would have been capable of identifying the one splicing effect that may have resulted from SNPs interfering with ESE functionality because one of the two alleles generated a *de novo* ESE motif. Indeed, other recent studies (Pfarr *et al.,* 2005; McVety *et al.,* 2006) argue against the efficiency of ESEfinder to predict functional outcomes of splice SNPs. For example, McVety et al. (McVety *et al.*, 2006) were able to authenticate their ESE-dependent splicing mutation at the 5′ end of exon 3 of *MLH1* gene by *in vitro* splicing assay, which was not recognized by all available motif-scoring matrices including ESEfinder. Likewise, the allele-specific skipping of exon 5, at phosphomannomutase 2 gene (*PMM2*), is due to a SNP that disrupts an ESE that was not detected by ESEfinder (Nembaware *et al.,* 2008). On the other hand, a more recent study revealed that, a cryptic ss usage in exon 7 of the human fibrinogen beta-chain (*FGB*) gene is regulated also by a naturally ‘silent’ SF2/ASF binding site within this exon (Spena *et al.,* 2006). This may also imply that, not all ESE motifs are actual functional splicing enhancers (Cartegni *et al.,* 2002) and not all nucleotide variations in

functional ESEs disrupt their function (Cartegni and Krainer, 2002; Fackenthal *et al.*, 2002; Pollard *et al.*, 2002).

4.3. Efficiency of *in silico* splice SNP prediction tools

For practical reasons, the discussion has so far focused upon the limitations of the available *in silico* tools in terms of their positive predictive value, i.e. of the proportion of predicted splice SNPs that indeed showed allele-dependent splicing. It is worthwhile remembering, however, that the poor performance observed in the current study does not *per se* devaluate the tools in question. Both algorithms could still have a high sensitivity and specificity even if the prior probability of differential splicing was simply too small in the present study for them to be able to make reliable positive predictions (ElSharawy *et al.*, 2008). In view of the recent global assessment of alternative splicing using exon tiling arrays, allele-dependent splicing at SNPs indeed seems to be a relatively infrequent event (Kwan *et al.*, 2007; Kwan *et al.*, 2008). Out of 17,897 genes screened (Kwan *et al.*, 2008), only 324 exhibited a significant association between transcript levels and flanking SNPs. Of these instances, 55% involved isoforms that we would have considered the result of 'allele-dependent splicing' in the context of the present study. Therefore, assuming an average number of 10 exons per gene, the prior probability of alternative splicing at a given exon would be approximately 10^{-3}. This implies that, even with a specificity as high as 99%, a positive predictive value of 8% would still correspond to a sensitivity of ~80% or higher. With a lower specificity, even higher sensitivity values would be compatible with the small positive predictive values observed in the present study, given that the prior probability of allele-dependent splicing was indeed of the order of 10^{-3}. Finally, it must be remembered that the present study was confined to relatively frequent SNPs which, due to a likely absence of strong evolutionary pressure, may have had a lower *a priori* probability of allele-dependent splicing anyway.

As mentioned in the introduction, a variety of bioinformatics tools for the prediction of splice-related SNPs other than those used in the present study are now available, including RESCUE-ESE, ExonScan and MaxEntScan (Fairbrother *et al.*, 2004b; Yeo and Burge, 2004; Nalla and Rogan, 2005). However, the choices of suitable tools had to be made at the beginning of the experimental validation, which was in early 2005. Three years later, the need for further experimental data on SNP allele-dependent splicing remains- a recently published splicing tool (F-SNP; (Lee and Shatkay, 2008)) combines use of 16 different bioinformatics tools and databases and still lacks predictive efficacy (ElSharawy *et al.*, 2008). Other recent

studies (Buratti *et al.*, 2007; Houdayer *et al.*, 2008) similarly concluded that existing *in silico* predictions are neither adequate to identify allele-dependent splicing effects particularly at loosely defined consensus positions, nor to classify unknown variants as deleterious or neutral especially at exonic sites. The shortage of available bioinformatics tools were also viewed in the context of the results from Hull et al. (Hull *et al.*, 2007). In the Hull study, allele-specific alternative splicing was observed in 6 out of the finally selected 70 exon-skipping events. However, sequence analysis of the relevant ss and of the regions surrounding SNPs correlated with the splicing events, observed in the Hull study, failed to identify any predictive bioinformatic signals.

4.4. Current understanding of allele-dependent splicing

Earlier studies suggested that gene expression constituted an important piece of human variation, and although it remains a significant aspect, the added complexity of transcript-processing variations and the potential outcome of these differences greatly alter our earlier perceptions (Kwan *et al.*, 2008). Genetic variation, through its effects on gene expression, influences many aspects of the human phenotype. Understanding the impact of genetic variation on human disease risk has become a major goal for biomedical research and has the potential of revealing both novel disease mechanisms and novel functional elements controlling gene expression. Recent large-scale studies have suggested that a relatively high proportion of human genes show allele-specific variation in expression. Effects of common DNA polymorphisms on mRNA splicing are less well-studied. Variation in splicing patterns is known to be tissue-specific, and for a small number of genes has been shown to vary among individuals. What is not known is whether allele-dependent splicing is an important mechanism by which common genetic variation affects gene expression (Hull *et al.*, 2007).

A careful reading through the accessible findings and observations, from the present study (ElSharawy *et al.*, 2008) and three other concurrent related studies (Hull *et al.*, 2007; Kwan *et al.*, 2007; Kwan *et al.*, 2008) that have focused on finding a clear relationship between genotype and splice phenotype constitute an important change in way we view the effects of common genetic variation in humans:

- It is likely that allele-dependent splicing is a vast underestimate of the true extent of this phenomenon (Graveley, 2008). This can be explained on the basis of the nature of these studies on the one hand, and the (mis)interpretation of microarrays on the other.

The present study considered only relatively frequent splice SNPs for evaluation (ElSharawy *et al.*, 2008). Excluding the perfect matching at canonical AG/GT dinucleotides of canonical ss would reduce the prediction rate to an half (4% of 8/214, and 8% of 17/223, after and before exclusion, respectively). Hull et al. (2007) and Kwan et al. (2007) only identified nine exons that are differentially spliced in an allele-specific manner that correlates with SNPs that are common in the human population (Hull *et al.*, 2007; Kwan *et al.*, 2007). While Hull et al. focused on only one form (exon skipping) of splicing variation in a relatively small number of genes (250), Kwan et al. validated only a small subset (20 of ~ 1000 candidate events) from their exon-based arrays. Similarly, Kwan et al. (2008) were able to show significant association between transcript levels and flanking SNPs of only 324 out of 17,897 genes screened using a global exon tilling arrays (Kwan *et al.*, 2008). Of these events, only 55% represented splicing-associated isoform changes. The problems that complicate interpretations of the results from microarrays are also well known. First, these arrays have difficulty in identifying cases where the splicing changes are subtle, even though they might be significant, both statistically and functionally. Second, the arrays can be 'noisy' or have a high degree of false positives and false negatives - for instance, the study by (Kwan *et al.*, 2007) had a 55% false-discovery rate (Graveley, 2008). Another point to bear in mind is that, allele-dependent splicing analysis was restricted in all of the previously mentioned studies to SNPs, as they constitute the most common type of genetic variation in humans. However, other types of allele-specific splicing events could be due to other types of polymorphisms such as indels (Romano *et al.*, 2002), polymorphism at VNTR (variable number tandem repeat) such as G/A substitution at position +8 in the coding sequence of exon 2 of *MUC1* gene (Ligtenberg *et al.*, 1991; Pratt *et al.*, 1996), or allele-specific polyadenylation due to differential CpG island methylation (Wood *et al.*, 2008). The estimation of allele-specific splicing is further complicated by the observation that, specific haplotype differences is correlated with differential expression and alterantive splicing such as that of microtubule-associated protein tau *(MAPT)* locus (Caffrey *et al.*, 2007; Caffrey and Wade-Martins, 2007).

- The possibility that allele-dependent splicing effects may be at least as prevalent in the genome as those on overall gene expression is raised from the work of Kwan et al. (2008). Kwan and co-authors classified their studied 324 genes from the exon-based arrays on the basis of expression changes at the exon and/or transcription level. They

found that 26% of genes showed changes at alternative splicing of a cassette exon, versus 39% reflected changes at the whole transcript level. The rest was either transcription initiation or termination changes (11% or 17%, respectively), or complex changes of multiple event types (7%). This means that about 55% of gene expression variation was isoforms-based (Kwan *et al.*, 2008).

- Although the contribution of heritable variation to the observed diversity of mRNA splice isoforms is well established from these studies, the resulting gene expression variation patterns from the Kwan study (Kwan *et al.*, 2008) further indicate that the regulatory effects of genetic variation in a normal human population are far more complex than previously observed. Thus, it is postulated here that allele-dependent splicing phenomenon is not uncommon in the human population, but it seems that splice SNPs exert their impact rather through complex effects. A recent survey showed association of 21% alternatively spliced genes with closely linked SNPs (Nembaware *et al.*, 2004) and among these events, there is evidence of two different types of allele-specific splicing: 1) pure (complete) allele-dependent splicing, in which one allele gives rise to one isoform and another results in the alternative form (as detected in the present study). This type was later suggested to be less common (Nembaware *et al.*, 2008). 2) Complex (partial) allele-specific splicing in which different alleles result in distinct relative isoform abundance. In fact, all of the readily available studies only examined RNA isolated from small number of different cell lines, which means that many of the common human haplotypes were not examined.
- Identifying SNPs that correlate with heritable changes in alternative splicing but do not cause disease added a new twist to the link between genetic variation and pre-mRNA splicing. This suggests that allele-dependent splicing is a mechanism that accounts for individual variation in the human population (Graveley, 2008). Furthermore, SNP-driven transcript variation may serve to increase proteome variability and maintain a heterozygote advantage on the population level. On the other hand, SNPs that predict splicing phenotypes are likely to be important markers to include in genetic association studies of complex diseases (ElSharawy *et al.*, 2008), since estimates from monogenic disorders, as mentioned in the introduction, indicate that up to 30% of phenotypically relevant mutations actually act through allele-dependent splicing.
- Allele-dependent splicing events are likely more frequent around exon-intron junction. For most uncovered instances in the various studies, SNPs with the strongest

correlation were those closest to the intron-exon boundaries of the splicing events. The present study was rationally looking for such events at ss junctions, since assembly of the splicing machinery around the ss comprise the foundation for efficient exon definition (Hertel, 2008) and mutations in the pre-mRNA that disrupt RNA-RNA base pairing at ss will, in turn, decrease the efficiency of exon recognition. Kwan et al. (Kwan *et al.*, 2007) also found a SNP located at the 5′ ss of the affected exon in *CAST* gene, suggesting that this SNP most likely impacts the efficiency of U1 snRNP binding. In addition, Hull et al. (Hull *et al.*, 2007) found that for five out of six of these events, the strongest correlation was found with the SNP closest to the intron-exon boundary. Knowing that the ratio of SNPs affecting splicing located intronic in a very tight window, i.e. at exon-intron borders, increase the possibility of SNPs near this junction higher influence splicing. The SNPs that reside outside this frame may increase the extensive flexibility of spliceosome to identify and process within a given pre-mRNA and AS (Hertel, 2008). A certain level of variability is still tolerated, which leads the splicing process to occur normally even if the extent of base pairing is not fully satisfied, and this variability can be compensated by recognizing different ss with different spliceosomal factors (Rekha and Mitra, 2006). The contributions of the other parameters will vary significantly from case to case, augmenting or reducing the overall affinity of the splicing machinery (Hertel, 2008).

In actual fact, regulation of splicing is incompletely characterized and complicated by the fact that additional *cis*-elements that control splicing are still being discovered (Yeo *et al.*, 2007), and allele-dependent splicing needs also to be considered with regard to inter-population variation (Jakobsson *et al.*, 2008)—common splice SNPs in Caucasian populations, for instance, are not necessarily frequent in other populations. This highlights again the need for larger-scale whole genome studies investigating all possible splicing patterns/motifs, to determine the actual extent of SNP-associated splicing phenotypes in different populations.

4.5. Remarks on the impact (nature) of the observed splice-relevant SNPs

Another important aspect of allele-dependent splicing, in addition to its mere occurrence, is the need to predict its outcome in terms of either exon skipping or cryptic ss utilization. In accordance with previous reports (Nakai and Sakamoto, 1994; Baralle and Baralle, 2005; Krawczak *et al.*, 2007), the majority of differential splicing events at donor sites in the present study were comprised of exon skipping (4/7 = 57%) whereas cryptic splice site usage was

predominant at acceptor splice sites (10/10 = 100%) (ElSharawy *et al.*, 2008). Reported results from the recent neural network (Krawczak *et al.*, 2007) also indicated that donor ss mutations, screened in a region of 50 nt upstream of all the affected donors, were basically leading to exon skipping by a total of 85%. In accordance with previous knowledge (Krawczak *et al.*, 2007; Houdayer *et al.*, 2008), the dramatic effect of splice SNPs at a donor (exon skipping) rather than at an acceptor ss (indel), would support the view that the correct recognition of the donor ss represents the key step in splicing (exon recognition). Furthermore, the disruption of conservation balance at donor ss is quite noisy in a distance-dependent manner. Once again, this high probability of alternative 3'-ss activation in close proximity of the dominant 3'-ss suggests that the second step of the splicing may be prone to violating splicing fidelity (Dou *et al.*, 2006). On the other hand, it seems that the probability of cryptic ss utilization increases as a function of the saturation of the local DNA sequence environment with such motifs. For example, only one (rs330924:G>C) out of the 4 SNPs (including rs3763131:A>G, rs181390:T>C, and rs3745503:A>C) that have been screened at position (-9), was able to create a cryptic ss with a novel 'AG' consensus with surrounding nucleotides, while the rest resulted in a 'non-AG' consensus (C[A/G]T, C[C/T]G, C[A/C]C motifs, respectively). Likewise, the choice between exon skipping and cryptic ss utilization upon 5′ splice site abolition could depend on the presence of a strong putative cryptic 5′ splice site and/or the degree of local saturation of cryptic motifs for 5′ ss (Krawczak *et al.*, 2007; Wimmer *et al.*, 2007).

In the present study the vast majority (78%; 14/18) of the detected allele-dependent splicing events was occurred in intronic sequences, namely 6 of 7 (85.71%) at donor, 6 of 6 (100%) at acceptor, and 2 of 4 (50%) at NAGNAG-tandem acceptors (Table 3.4) (ElSharawy *et al.*, 2008). At first, this indicates that non-coding SNPs are potentially contributing to ss alterations (ElSharawy *et al.*, 2006; Skotheim and Nees, 2007). A similar lower tendency (14.9%; 71/478) of exonic disease-causing single base-pair substitution within a verified effect upon splicing is recently reported from the neural network (Krawczak *et al.*, 2007). Once more, four of the identified allele-dependent splicing events in the present study were occurred at loosely defined consensus positions, namely positions (-7 and -9) upstream of acceptor and (-1 and +5) up- and downstream of donor ss, respectively (see Table 3.4). Taking both of these observations together, this may imply that the information required for splicing is contained in the consensus outsized 6-8 nt at both regions, contrary to what has been suggested in previous reports (Rekha and Mitra, 2006; Koren *et al.*, 2007). The results

from a comparative study suggest that the conserved intronic elements— 100 bases in length flanked of the alternatively spliced exons— possibly function in alternative splicing regulation (Sorek and Ast, 2003). However, another study revealed that the most common alternative acceptor or donor ss used in the human genome are located within 6 nt of the dominant ss (Dou *et al.*, 2006). Nevertheless, the number of detected cryptic 5′ ss decreased with increasing distance from the authentic 5′ ss (Roca *et al.*, 2003). It seems that, a plausible tendency for the splicing apparatus to use a cryptic ss depending very much upon the distance from the site of mutation for donor, but not for the acceptor (Krawczak *et al.*, 2007). This was explained on the basis that the successful functional recognition of an acceptor ss depended upon the presence of DNA sequence elements that have a less stringent consensus than donor ss (e.g., the polypyrimidine tract or the branch point).

Buratti et al. (Buratti *et al.*, 2007) provided statistical evidence that the frequency of intronic position +5 of donor ss is significantly higher than that observed in the Human Gene Mutation Database, suggesting that alterations of this position are particularly prone to aberrant splicing, possibly due to a requirement for sequential interactions with U1 and U6 snRNAs. Buratti and co-workers also showed that all point mutations at position +5 of authentic 5′ ss that activated cryptic 5′ ss were substitutions of G, and not any other nucleotide, raising the possibility that 5′ ss with +5G are more susceptible to aberrant ss activation than 5′ ss with +5H (non-G). Furthermore, the same study provided evidence that for cryptic donor ss, point mutations appeared in the following order: +1 (39.4%); +5 (21.6%); +2 (14.7%); -1 (14.3%); and (+3, +4, +6, -2) <3%. In fact, these results are consistent with the findings from the present effort. The present study showed that the mismatches at positions -2, +3, +4, and +6 to donor ss were not critical to splicing compared to mismatches at other positions (-1, +1, +2, and +5) (ElSharawy *et al.*, 2008). This finding is well documented in other reports (Zhuang and Weiner, 1986; Stephens and Schneider, 1992). In addition, Krawczak et al. (Krawczak *et al.*, 2007) found that the disease-associated mutations clustered more closely around the exon-intron junction in donor ss, with 70% of the 110 lesions being located at either exonic position -1 or intronic position +5. Furthermore, position +5 has recently gained great concern as a hot spot of disease-causing mutation. For example, a famous mutation at +5 of donor ss (IVS3+5GC, 5-GUAACU-3) and resultant exon 3 skipping was reported as a disease-causing mutation in the *NF1* gene (Baralle *et al.*, 2003). Also, exclusion of exon 2 at *HMSD* gene due to AS was completely controlled by an intronic SNP (rs9945924) at IVS+5 (Kawase *et al.*, 2007).

To further investigate position -3 upstream to acceptor ss, a total of 24 SNPs were screened at this location in both screening rounds, namely 9 SNPs (with delta scores >8 points) at the web-based round and 15 SNPs at the neural network round (of them 4 SNPs located at the top and 11 located at the middle range of the generated scoring list). All of the experimentally verified instances (24/24; 100%) had no effect on splicing outcome (Appendix Table 8.1). This finding may indicate that this acceptor position (-3), albeit its close neighbourhood to the consensus AG dinucleotides, has less effects on splicing events than expected from the recent neural network prediction efforts (Krawczak *et al.*, 2007). The Krawczak study observed 14 of the 40-acceptor ss mutations (35%) to occur within the 38 analyzed genes at intronic position -3. Further analysis is thus required to preclude this controversy and provide a more comprehensive view of the effect of genomic variations at this and other splice-relevant locations.

Another different concept to bear in mind, especially in explaining allele-dependent splicing occurrence at ESE sites, is the natural selection. As natural selection removes deleterious mutations from the population, variations that persist as SNPs were largely suggested to be neutral and appeared to avoid "functional" elements, such as ESEs (Pfarr *et al.*, 2005). It is proposed that a coding exon is subjected to at least three different selection pressures: (1) preserving the coding sequence, (2) preserving the sequence of splicing motifs, and (3) preserving an appropriate structural context for these splicing motifs. Selection on the coding sequence is likely to be the strongest pressure (Hiller *et al.*, 2007b). Analyzing the set of SNPs that overlap RESCUE-ESE hexamers showed that, nearly one-fifth of the mutations that disrupt predicted ESEs have been eliminated by natural selection. This selection was strongest for the predicted ESEs that were located near ss (Fairbrother *et al.*, 2004a). Evidence of purifying selection against synonymous mutations in mammalian ESEs has been recently reported (Parmley *et al.*, 2006). A unique discovery that a synonymous SNP in exon 5 of *MCAD* gene protects from deleterious mutations in a flanking ESE, suggests yet another complication of evaluation of potential deleterious effects of mutations on splicing in the context of the relevant haplotype (Nielsen *et al.*, 2007). Additionally, recent results demonstrated that a decision to include or exclude sequences adjacent to splicing mutations in mature transcripts is influenced by their ESS/ESE frequencies (Kralovicova and Vorechovsky, 2007).

4.6. Hypotheses on the functional consequences of putative splice SNPs

Natural genetic variations in the splicing machinery might contribute to the predisposition of different individuals to human diseases and to the severity of their phenotype. An estimated 20%–30% of disease-causing mutations is believed to affect pre-mRNA splicing (Faustino and Cooper, 2003), through the disruption of ss, exonic and intronic splicing enhancers and silencers, or RNA secondary structure. Deviations from a normal AS pattern—either through isoform expression imbalance or presence of aberrant isoforms—initiate many diseases (Caceres and Kornblihtt, 2002; Garcia-Blanco *et al.*, 2004; Lopez-Bigas *et al.*, 2005; Garcia-Blanco, 2006). The current progress in understanding the role of splicing modulation as a genetic modifier opens new avenues towards developing treatments for many human diseases and availability of functional annotations for these events will in turn lead to targeting the correct splice isoforms (Talavera *et al.*, 2007). For example, splicing modulation therapy has been used in the treatment of Duchenne's muscular dystrophy (DMS). An antisense-mediated exon skipping approach was used in a clinical trial to restore dystrophin synthesis in the muscles of patients with DMS. In this approach, local intramuscular injection of a 20-nt antisense oligoribonucleotide induced exon skipping in exon 51, which subsequently restored the disrupted reading frame, and thus introduced dystrophin protein in the muscle in all 4 patients who received therapy (van Deutekom *et al.*, 2007). The introduction of dystrophin protein would convert a severe DMS into a milder Becker muscular dystrophy phenotype (Aartsma-Rus and van Ommen, 2007).

The present study identified 18 splice SNPs, of which 15 were novel and 3 had known functional relevance (ElSharawy *et al.*, 2008). Splice SNPs with a known phenotypic impact include rs2076530, which is located in the *BTNL2* gene (butyrophilin-like 2) and which has been shown to be associated with sarcoidosis (Valentonyte *et al.*, 2005). SNP rs1805377 is located in the *XRCC4* gene (X-ray repair complementing defective repair in Chinese hamster cells 4) and has been reported to be associated with bladder cancer (Figueroa *et al.*, 2007). Host susceptibility to viral infection in type I diabetes has been shown to be associated with variation in the *OAS1* gene (oligoadenylate synthetase 1), for which a splice SNP, rs10774671, was identified (Field *et al.*, 2005). The remaining 15 splice SNPs have not been reported to be associated with a specific phenotype, but the information provided here may contribute to a better understanding of the functional relevance of the respective loci, particularly since an increasing number of disease-associated loci are being identified in hypothesis-free genome-wide association studies (ElSharawy *et al.*, 2008). Three of the novel

splicing polymorphisms, representing acceptor, NAGNAG tandem and donor ss, are described in more detail below.

The first candidate SNP is rs11658717 at *STXBP4* (syntaxin binding protein 4 or synip) gene. First, this SNP located at a loosely defined position (-7) at acceptor ss of exon 6 (refseq NM_178509) and its minor allele 'G' is associated with 6 nt insertion upstream of that acceptor. Second, the SNP minor allele-G generated a competitive alternate or cryptic acceptor ss with score of 83.7, whereas the other SNP allele-A produced only a score of 67.7. Third, the 6 nt insertion is predicted to create a novel SC35 motif (GGTTAGAA; ESEfinder score: 2.79115) with the best score at the upstream half of exon 6, which might support its insertion. Fourth, analogous to NAGNAG tandem, a '*NAG*NAC*NAG*' motif is identified, which might offer plasticity at the acceptor ss. Finally, using domain-prediction SMART Tool, this 6-nt insertion is predicted to result in a shorter PZD domain (by 6 amino acids) and consequently leucine 99 (with hydrophobic side chain) replaces serine 99 (which in turn converted to serine 101). Generally, synip protein contains an N-terminal PDZ domain, a central region with EF and coiled-coiled domains, and a C-terminal WW motif (Min *et al.*, 1999). It is known that synip protein represents a potential target of insulin signaling, which may regulate the fusion of glucose transporter 4 (GLUT4) storage vesicle (GSV) with the plasma membrane, where the transporter facilitates the diffusion of glucose into striated muscle and adipocytes, and thereby, enhances glucose uptake (Watson and Pessin, 2007). According to the existing model, synip undergoes phosphorylation at 'serine 99' in response to insulin stimulation, and this leads to the dissociation of the synip-syntaxin 4 complex, thus freeing syntaxin 4 and allowing productive VAMP2-syntaxin4 complex formation and subsequent fusion (Yamada *et al.*, 2005; Okada *et al.*, 2007). This data, therefore, highlights the need for further investigation of the functional impact of rs11658717 at glucose uptake pathway and opens many questions of biological interest to be addressed: 1) does the splice effect regulate synip/syntaxin4 interaction, thereby modulating GLUT4 translocation and glucose uptake, or further modulate human insulin resistance? 2) Is serine 101 phosphorylated by insulin? And if it does, does synip still dissociate from syntaxin 4? Last but not least, 3) does the modified PDZ domain predicted here pose normal function?

The second candidate SNP is rs5248, which is located at NAGNAG acceptor ss of exon 3 at the chymase (*CMA1*) gene of chr 14. This SNP affected the upstream G nucleotide at the NAGNAG tandem and encouraged 3-nt (CAG-motif) insertion, which generated a new

competitive acceptor ss with score (91.4) similar to that of the downstream one (score: 91.6). The insertion itself seems to be advantageous as it led to a simultaneous creation of a novel SC35 (GGTCTATA) motif with the highest predicted exonic score (3.3968) by ESEfinder, which might encourage the insertion incidence. Indeed, a differential expression of tandem allele at rs5248 was also noticeable in a recent study (Hiller *et al.*, 2006a)**.** In addition, *CMA1* locus is reported in a famous linkage, at 14q11-12, to inflammatory bowel disease (IBD) (Duerr *et al.*, 2000). Particularly, it might have an impact in susceptibility to Crohn's disease (CD) in active mucosa (Andoh *et al.*, 2006). Moreover, *CMA1* has been suggested to play a role in modification of the functional outcome of pulmonary sarcoidosis (Kruit *et al.*, 2006), susceptibility to atopic asthma (Sharma *et al.*, 2005)**,** and has been pointed out as a candidate gene for atopic eczema (Weidinger *et al.*, 2005).

The third novel splice SNP is rs482308. Here, the minor allele 'A' disrupted the obligatory G nt at position +1 to exon 35 of zonadhesion (*ZAN*) gene (NM_003386), and resulted in the utilization of a cryptic donor ss 111 nt upstream of exon 35. Using Expasy Translation Tool and domain-prediction SMART Tool, this 111-nt deletion was predicted to result in deletion of 37 aa and the possibility of skipping of two Pfam TIL (Trypsin inhibitor like cysteine rich) domains and remodeling of VWD (homologous to the D domains of the von Willebrand factor) domains. The VWD domains are recently reported to be involved in direct contact with zona pellucida (ZP; egg's extracellular matrix) in a species-specific manner (Gasper and Swanson, 2006). Yet, there is no reported splicing variation around this location at UCSC, which supports the novelty and significance of the detected splice effect. Indeed, the current status indicates that the human ZAN protein exists as six splice variants among exons 41-43 (not around exon 35), ranging in length from 2,600 to 2,724 codons and many of these variants are derived from testis EST data (Gasper and Swanson, 2006). Despite the multiple (20–30) candidate sperm proteins that have been proposed over the years (Brewis *et al.*, 2005; van Gestel *et al.*, 2007), zonadhesion is still considered the major sperm membrane protein that has ZP binding ability (Tanphaichitr *et al.*, 2007). Based on the facts that 1) ZAN-ZP binding is essential for acrosome reaction, which occurs in the acrosome of the sperm as it approaches the ZP, and 2) the glycoprotein nature of the ZP, it would be interesting to further investigate the effect of the predicted domain remodulation of *ZAN* protein in the fertilization process. It is hypothesized here that the novel splice finding might alter the efficiency of acrosome reaction by changing the ability of the sperm to fuse to the oocyte, thereby lowering the chance of fertilization. Another observation is that the minor allele (A) at rs482308 is

distributed differently among populations (see dbSNP), and shows more abundance in Caucasian populations. Again, what is the impact of this abundance in Caucasian? Translating the identified splice effect at *ZAN* into comprehensive biological meaning using functional analyses might therefore provide a new perspective in tackling the issue of primary sperm-zona interaction.

4.7. Need for an alternative system: A proof of concept and outlook

Correlating naturally occurring human genetic variations to functional impacts on pre-mRNA splicing presents a current challenge, which is accompanied by the development of numerous systems and tools to achieve this goal. The putative impact of unknown variants on splicing is also one of the routine challenges faced by molecular geneticists in their everyday practice. Unfortunately, RNA studies cannot be performed in each case and a compromise must be found between the time and cost required by RNA analysis and the risk of missing a deleterious mutation (Houdayer *et al.*, 2008). Indeed, the current task is further complicated by the impact of many other factors. First, the available *in silico* bioinformatic tools for prediction of allele-dependent splicing effects are limited by their derivation from mostly EST data from different and/or disease tissues. Second, there is a lack of an effective high-throughput screening assay to identify potential positions of splice-relevant SNPs and differentiate between SNPs that cause primary pathogenic effects and SNPs that simply modulate plasticity of the 'splicing-code'. Third, tissue resources are limited and the availability of RNA samples from disease-relevant tissues and affected individuals and their families is often problematic (Buratti *et al.*, 2007; Wang and Cooper, 2007). Fourth, in spite of efforts to discover *trans*-acting tissue-specific splicing signatures (Xu *et al.*, 2002; Grosso *et al.*, 2008), correlation of *cis*-regulatory motifs occurrences with gene expression and AS levels across tissues (Yeo *et al.*, 2004; Das *et al.*, 2007) is still in its infancy. To overcome these difficulties, a novel *in vitro* expression reporter system for alternative splicing was designed and tested in the present study. This system is mainly based on a splice-dependent expression model of GFP and provides a tool to screen a randomly mutagenized plasmid bank by FACS sorting and subsequent anaylsis using the second generation sequencing technology.

The inserted test genomic region was chosen to meet several requirements: relatively small exon (190, 98 and 181 bp for exon 7, 8 and 9, respectively) and intron (318 and 1082 bp for intron 7 and 8, respectively) sizes that fitted to the cloning strategy. Second, the internal test exon (exon 8) had to be accommodated with its flanking introns. Third, the genomic region

had to be inserted while preserving the ORF of GFP. The site-directed mutagenesis experiments and data obtained from FACS-based (Figure 3.12) and immunoblot analyses (Figure 3.13) confirmed the efficiency of the current construct (RFP-PGM2L1-pEGFP-N1; Figure 3.11) to report splicing modulation as a result of splice-specific variations. Thus, GFP (the experimental reporter gene) represents an indicator of splicing efficiency of the inserted genomic region from *PGM2L1* gene—correct splicing of primary transcript leads to expression of functional GFP (green fluorescence). Although the RFP, which is located in the front of the genomic region (Figure 3.11), stabilized the splicing of the test genomic region, there is still a need for an independent transfection control. It is imperative that the second reporter gene to be expressed by the same vector, in order to allow the normalization of transfection efficiency and cell number. Small perturbations in the growth conditions for the transfected cells can dramatically affect gene expression and transfection efficiency. Thus, the second reporter would help to determine if the effects are due to the treatment of the cells or a response from the experimental reporter. Indeed, our initial experiments showed that dsRed1, but not dsRed2, could serve as transfection control, but the use of RFP for this purpose requires cloning of all of its regulatory elements (promoter, Kozak consensus, poly A tail), which will result in a very large vector. Another ongoing strategy is to fuse either c-myc (using the mutagenic primers EGFP_myc_f and EGFP_myc_r) or a FLAG tag (using the mutagenic primers (EGFP_flag_f and EGFP_flag_r) to the N-terminal of neomycin resistance gene of the produced construct (Figure 3.11). This will allow an independent expression of the fused protein using the SV40 promoter. This epitope tagging technique provides an efficient means for recognition of the obtained fused protein by readily commercially available tag-specific antibodies (mouse monoclonal anti-c-myc (clone 9E10) IgG or mouse monoclonal anti-FLAG® (clone M2) IgG; Columbia Biosciences). Both of these antibodies are conjugated with red-Phycoerythrin dye (RPE) (excitation max. λ: 565>498 nm; emission max. λ: 578 nm) that is also suitable for flow cytometry/FACS analysis, for which standard protocols are adaptable and available.

Indeed, several other strategies to identify splicing-regulatory factors are currently in wide usage. These include RT-PCR, reporters producing luciferase or GFP, and a topoisomerase I phosphorylation assay. Each of these assays has limitations in the high-throughput screening of large chemical libraries. RT-PCR is costly and scales up poorly. Most *in vivo* splicing reporters have poor dynamic range or do not distinguish compounds affecting splicing from those altering transcription or translation (Stoilov *et al.*, 2008). Indeed, three recent studies

demonstrate the utility of dual-color reporter systems in improving the dynamic range and discriminating changes in alternative splicing from changes in transcription or translation. Two of these systems (Newman *et al.*, 2006; Orengo *et al.*, 2006) may require modification of a test exon to adapt it to the reporter, which may change its regulatory properties (Stoilov *et al.*, 2008). The last evolved system argued for its flexibility to accommodate a variety of test exons from different genes in a high-throughput trend (Stoilov *et al.*, 2008). However, it is not always possible to house whole flanking introns to each test exon, especially in case of flanking introns exhibiting large sizes. In this regard, the reporter system engineered in the present study is advantageous, since the inserted test genomic region, comprises an integral 'permanent' part of the reporter system.

Incorporating the advantage of using second generation sequencing technology, together with the use of the FACS-based reporter system with its dichromatic readouts, would meet many several requirements and features.

- The presence of all possible (un)known splicing regulatory motifs around the test middle exon 8 of *PGM2L1* gene are favorable. In addition to its ability to distinguish changes in AS patterns from changes in transcription and translation, the natural and proper assembly of the splicing machinery around the nascent pre-mRNA transcripts only allows the impact of *cis*-acting splicing motifs to arise and to be measured.
- The system has a broad dynamic range, allowing ease of access impact of *cis*-acting variations in a variety of splicing sites (such as at donor ss, acceptor ss, ESE, ESS, and others) at the mRNA level—it is not restricted to variation in particular components of the splicing recognition sequences.
- It combines the advantages of minigenes constructs (a defined and experimentally controlled system) and second generation sequencing technology (ultra-high-throughput), and therefore it provides a means for quantitative analysis of sequence-dependent splice variations. As a result, the abundance of splicing motifs could be correlated to corresponding variation and splicing efficiency.
- Ease of manipulation; the same construct can be tested in different tissue panels. Therefore, tissue-specific splice motifs/variations, context-dependent weight matrix of splice motifs (donor, acceptor, ESEs, etc.) together with the potential position of splice-relevant DNA variation could be more precisely identified.

- The present system has the ability, although not recommended, to accommodate other test exons or other specific motifs-containing SNPs obtained from association studies in order to explore their impact on mRNA phenotype and/or corresponding protein.

In this way, the described experimental system in the present study provides a suitable high-throughput screening tool of variations that modulate AS and presents an improvement of prediction tools of allele-dependent splicing. This, in turn, would improve our understanding of mammalian ss anatomy and invent a means for future mechanistic and functional analyses.

5 SUMMARY

Background: The evolutionary and biomedical importance of differential mRNA splicing is well established, especially with regard to pathophysiological conditions. Up to 60% of mutations that contribute to disease development have been proposed to do so by disrupting splicing events. Erroneous splice site usage is also observed in numerous diseases.

Problem: Identification and functional annotation of single-nucleotides polymorphisms that interfere with splicing mechanisms (‘splice SNPs') is a major challenge and needs to be supported by an efficient method.

Solution:

1) A high-throughput methodology was established to facilitate the screening of allele-dependent splicing in a high-throughput fashion (ElSharawy *et al.*, 2006). The method integrated a package of four new software tools and was mainly based on using a panel of 92 matched pairs of individual-specific gDNA and cDNA samples. For each SNP, 16 cDNAs providing a balanced representation of the genotypes at the respective SNP were investigated by nested RT-PCR and subsequent sequencing. Putative allele-dependent splicing events were verified by cloning and sequencing.

2) A systematic, SNP-centered approach was followed and the database dbSNP was screened to filter a group of common SNPs at either canonical splice sites or ESEs that were classified as putatively splicing-relevant by bioinformatics tools. This was completed in two screening rounds using web-based tools (Alex’s splice site score calculator and ESEfinder) and neural network, respectively. A group of SNPs at NAGNAG tandem repeat sites was also tested (ElSharawy *et al.*, 2008).

Results and conclusion: As a result of genotyping, the 223 non-redundant candidate SNPs were experimentally tested, and 18 allele-dependent splicing events were identified, of which 15 were novel and 3 exhibited an already known functional relevance. However, the positive predictive value of the bioinformatics tools turned out to be low, ranging from 0% for ESEFinder to 9% (in the case of acceptor site SNPs) for the neural network. Overall, the currently available bioinformatics tools contribute little to the understanding as to how common genetic variation impacts mRNA splicing. Therefore, there is a need for an alternative system.

A proof of concept and outlook: The present study made some preliminary steps to develop a novel *in vitro* fluorescence-based splice reporter system. The ongoing systematic and hypothesis-driven experiments, which combine the advantages of FACS-based reporter constructs with a dichromatic readout method (a defined and experimentally controlled system) and ultra-high-throughput second generation sequencing technology, will serve to establish an efficient means to address many splice-related topics, and thus, would improve our understanding of mammalian splice site anatomy.

6 ZUSAMMENFASSUNG

Hintergrund: Die weitreichende evolutionäre und physiologische Bedeutung des differentiellen mRNA-Spleißens ist allgemein bekannt, besonders im Hinblick auf pathophysiologische und biomedizinische Fragestellungen. Man geht davon aus, dass bis zu 60 Prozent aller krankheitsverursachenden Mutationen auf eine Zerstörung von Spleißstellen zurückzuführen sind. Eine fehlerhafte Nutzung von vorhandenen Spleißstellen ist bereits für eine Vielzahl von Krankheiten bekannt.

Problemstellung: Die Identifizierung und funktionelle Annotation spleißrelevanter SNPs stellt eine große Herausforderung dar und bedarf der Unterstützung durch eine effiziente Methodik.

Lösungsansatz:

1) Zur Erleichterung des Screenings nach allelabhängigen Spleißereignissen wurde eine neue Hochdurchsatzmethodik entwickelt (ElSharawy *et al.*, 2006). Diese umfasst vier neue Software-Anwendungen und basiert hauptsächlich auf der Nutzung eines Panels von 92 übereinstimmenden Paaren individuenspezifischer gDNA- und cDNA-Proben. Für jeden der zu untersuchenden SNPs wurden 16 cDNAs mittels RT-PCR und anschließender Sequenzierung untersucht. Allelabhängige Spleißereignisse wurden durch Klonierung und Sequenzierung verifiziert.

2) In einem systematischen, SNP-zentrierten Ansatz wurden häufige SNPs an kanonischen Spleißstellen sowie an ESEs aus dbSNP gefiltert und mittels webbasierter Anwendungen als potentiell spleißrelevant klassifiziert. In einem zweiten Ansatz erfolgte die Klassifizierung der SNPs mittels eines neuronalen Netzwerkes. Die als spleißrelevant klassifizierten SNPs wurden im Anschluß mit der oben beschriebenen Methode untersucht. Zusätzlich wurde eine Gruppe von SNPs an NAGNAG Tandems Repeats getestet (ElSharawy *et al.*, 2008).

Ergebnisse und Schlussfolgerungen: Insgesamt wurden 223 nicht redundante Kandidaten-SNPs experimentell getestet. Dabei wurden 18 allelabhängige Spleißvorgänge identifiziert, von denen 15 neuartig waren und für 3 die funktionelle Relevanz bekannt ist. Dabei stellte sich die korrekte positive Vorhersagefähigkeit der bioinformatischen Tools als äußerst gering heraus - von 9% (für Spleißakzeptor-SNPs) für das neuronale Netzwerk bis zu 0% für den „ESEFinder“. Zusammenfassend konnten die verwendeten bioinformatischen Anwendungen nur wenig zum Verständnis beitragen, wie häufige genetische Variationen das mRNA-Spleißen beeinflussen.

Ausblick: In der vorliegenden Arbeit wurden entscheidende vorläufige Schritte zur Entwicklung eines neuartigen fluoreszenzbasierten *in vitro*-Spleißreportersystems geleistet, welches zur Zeit getestet und im Hinblick auf gezielte Fragestellungen validiert wird. Die momentan durchgeführten systematischen und hypothesenorientierten Experimente kombinieren die Vorteile FACS-basierter dichromatischer Reportersyteme mit denen der Hochdurchsatz-Sequenziertechnologie (*second generation sequencing technology*) und könnten ein effizientes Mittel zur Aufklärung vieler spleißrelevanter Fragestellungen darstellen und unser Verständnis des allelabhängigen Spleißens entscheidend verbessern.

7 MATERIALS

Table 7.1 Kits, Enzymes, vectors, antibodies, and Chemicals

Product	Manufacturer
100 bp DNA ladder	Invitrogen; Karlsruhe, Germany
15 ml reaction tubes	Sarstedt, Nümbrecht, Germany
2.2 ml 96 deep well	MTP ABgene, Epsom, UK
2-Mercaptoethanol	Sigma, Munich, Germany
37% Formaldehyde	Sigma, Munich, Germany
384 deep well storage plate (max. 300 µl)	ABgene, Epsom, UK
384 well PCR MTP	Eppendorf, Cologne, Germany
384-well MT plates	Greiner Bio-One GmbH, Frickenhausen, Germany
384-well MT plates	Sarstedt, Nümbrecht, Germany
50 ml reaction tubes	BD Biosciences, Heidelberg, Germany
96-well MT plates	Costar Corning Incorporated, Cambridge, MA, USA
96-well MT plates	Sarstedt, Nümbrecht, Germany
Advantage RT-for-PCR (100 reactions)	BD Biosciences Clontech, Palo Alto, CA, USA
Aequorea victoria GFP (A.v. GFP) monoclonal antibody (JL-8)	Clontech, Heidelberg, Germany
Agarose	Eurogentec, Cologne, Germany
Alkaline phosphatise (CIP); 10,000 U/ml	New England Biolabs, Bad, Schwalbach
AmpliTaq Gold® with GeneAmp 10x PCR Buffer II & MgCl2 solution	Applied Biosystems, Foster City, CA, USA
BamHI enzyme	New England Biolabs, Bad, Schwalbach
BigDye® Terminator Ready reaction kit v1.1	Applied Biosystems, Foster City, CA, USA
Biosphere® Filter Tips (10/200/1000 µl)	Sarstedt, Nümbrecht, Germany
Bromphenol blue	Sigma, Munich, Germany
BSA (Bovine Serum Albumin)	New England Biolabs, Bad, Schwalbach
Cell culture flasks (250 ml; canted neck)	BD Biosciences, Heidelberg, Germany
DNase enzyme	Qiagen, Hilden, Germany
dNTP set (100 mM solutions, each 100 µM)	GE Healthcare UK Limited, Buckinghamshire, UK and Amersham, Piscataway, NJ
DsRed monoclonal antibody	Clonethech, Heidelberg, Germany
Dulbecco's PBS (1X)	PAA Laboratories GmbH, Pasching, Austria
Easy peel heat seal foil	ABgene, Epsom, UK
ECL-Plus Western Blotting Detection System	Amersham Pharmacia Biotech, Amersham Labs, UK
EDTA	Sigma, Munich, Germany
EDTA blood vial 10 ml	Sarstedt, Nümbrecht, Germany
Ethidium Bromide solution (10 mg/ml)	Invitrogen, Karlsruhe, Germany
FuGENE® 6 Transfection reagent	Roche, Mannheim, Germany
G3PDH (Human Amplimers; 200 µl: 10 µM each)	Clontech, Mountain View, USA
GenomiPhi v1, v2, and high yield WGA Kit	GE Healthcare UK Limited, Buckinghamshire, UK
Glycerol	Sigma, Munich, Germany
GoTaq DNA polymerase (2.500 u: 5u/µl)	Promega, Madison WI, USA
Invisorb Blood Universal Kit	Invitek, Berlin, Germany
Isopropanol	Merck, Darmstadt, Germany
LiChrosolv® double distilled water	Merck, Darmstadt, Germany
M13 universal primers	Carl Roth GmbH and Co.KG, Karlsruhe, Germany
MagAttract DNA Blood M48	Qiagen, Hilden, Germany
MicroAmp® optical 96-well reaction plate	Applied Biosystems, Foster City, CA, USA
MicroAmp® single strips	Applied Biosystems, Foster City, CA, USA
MicroAmp® single tubes	Applied Biosystems, Foster City, CA, USA
Microtiter plates, 96-well, round bottom w/ lid	Sarstedt, Nümbrecht, Germany

Minielute Gel Extraction kit from	Qiagen, Hilden, Germany
MOPS (10 X): 3-[N-morpholino] propanesulfonic acid	Sigma; Munich, Germany
Mouse anti-β-actin monoclonal antiobody, Clone AC-15 (A5441-2ML; 107K4800)	Sigma, Steinheim, Germany
pDsRed2-N1 vector	BD Biosciences Clontech, Palo Alto, CA, USA
pEGFP-N1 vector	BD Biosciences Clontech, Palo Alto, CA, USA
Phosphate inhibitor cocktail II	Sigma-Aldrich Chemie, Munich, Germany
PicoGreen®	Invitrogen, Karlsruhe, Germany
Pipette tips with filter (10 / 200 / 1000 µl)	Sarstedt, Nuremberg, Germany
Primers	Metabion, Martinsried, Germany Eurogentec, Seraing, Belgium Microsynth laboratory, Lindau, Germany
Proteinase K	Molecular Research Center, OH, USA
PWO SuperYield DNA polymerase PCR buffer	Roche Diagnostics GmbH, Mannheim, Gemrnay
PWO SuperYield DNA polymerase; 250 U (5U/µl)	Roche Diagnostics GmbH, Mannheim, Gemrnay
QIA Filter™ Plasmid Maxi Kit	Qiagen, Hilden, Germany
QIAamp® DNA Micro Kit	Qiagen, Hilden, Germany
QuikChange® Lightning Site-Directed Mutagenesis Kit	Stratagene, La Joll, CA
Rabbit polyclonal to Mouse IgG antibody - H&L (HRP) (ab6728)	Abcam, Cambridge, UK
RC-DC Protein Assay Kit	BioRad, Munich, Germany
Reaction tubes (0.5/1.5/2.0 ml)	Eppendorf, Cologne, Germany
RevertAid H Minus First Strand cDNA Synthesis Kit	Fermentas Life Sciences, St. Leon-Rot, Germany
RNeasy mini RNA extraction kit	Qiagen, Hilden, Germany
RPMI 1640 medium without FCS	PAA Laboratories GmbH, Pasching, Austria
SacII enzyme	New England Biolabs, Bad, Schwalbach
SAP shrimp alkaline phosphatase	Amersham Biosciences; Freiburg, Germany
Sephadex powder (G-50 superfine)	GE Healthcare UK Limited, Buckinghamshire, UK
Sephadex spin column plates MAHVN 4550	GE Healthcare UK Limited, Buckinghamshire, UK
Serological pipettes with filter (5/10/25 ml)	Sarstedt, Nümbrecht, Germany
SmartLadder DNA marker	Eurogentec, Cologne, Germany
SNPlex™ System Core Kit	Applied Biosystems, Foster City, CA, USA
T4 DNA ligase	New England Biolabs, Bad, Schwalbach
TAE Buffer 25x ready pack	Amresco, Solon, OH, USA
Taq DNA polymerase	Qiagen, Hilden, Germany
TaqMan® Universal PCR Master Mix	Applied Biosystems, Foster City, CA, USA
TBE Buffer 10x ready pack	Amresco, Solon, OH, USA
TRIZOL	Invitrogen, Karlsruhe, Germany
Wizard® *Plus* SV Minipreps DNA Purification System	Promega, Madison WI, USA
Wizard® SV Gel and PCR clean-up system	Promega, Madison WI, USA
XhoI enzyme	New England Biolabs, Bad, Schwalbach

Table 7.2 Primers used in establishment of the splice reporter system

Primer abbreviation	Primer sequence (5′-3′)
dsRed_Xho_f	cagcgactcgagATGGCCTCCTCCGAGAACGTC
dsRed_XhoI_r	ctggctcGAGGAACAGGTGGTGGCGG
dsRed2-578-f	TACTACGTGGACGCCAAG

GFP_rem_ATG_F	ccggtcgccaccgtgagcaagggc
GFP_rem_ATG_R	gcccttgctcacggtggcgaccgg
GFP_RT_F	TAAACGGCCACAAGTTCAGC
GFP_RT_R	CGGCCATGATATAGACGTTGT
NM173582_PCR_f	TTGTGCCTACATACAGGAAC
PGM_GFP_F_Seq	TACAGTGGTTGTTGGAAAGTT
PGM_int7_TC_F	ctttttgggtcactgtttctcgaactttccttgagactgg
PGM_int7_TC_R	ccagtctcaaggaaagttcgagaaacagtgacccaaaaag
PGM_int8_CA_F	gcagcagcagaacttcaggagaacaatgtagaatctgttatttga
PGM_int8_CA_R	tcaaataacagattctacattgttctcctgaagttctgctgctgc
PGM_R_BamH	CGGTGGATCCTCAAAATGAAATCCTTCTTTAAGTGC
PGM_SacII_F	GGACCCGCGGTTAAACTCGAAGACCACCTTGAA
EGFP_myc_f	caggatgaggatcgtttcgcatggagcagaaactcatctctgaaga ggatctgattgaacaagatggattg
EGFP_myc_r	gcgtgcaatccatcttgttcaatcagatcctcttcagagatgagtt tctgctccatgcgaaacgatcctcatcctg
EGFP_flag_f	caggatgaggatcgtttcgcatggattacaaggatgacgacgataa gattgaacaagatggattgcacgc
EGFP_flag_r	gcgtgcaatccatcttgttcaatcttatcgtcgtcatccttgtaat ccatgcgaaacgatcctcatcctg

Table 7.3 Solutions and Media

Name	Description
0.1% TBST	10x TBS = 200mM Tris pH7.6, 1.37M NaCl → 200ml 1M Tris pH7.6 , 80.1g NaCl, autoclaved dist water ad 1 l TTBS = 1x TBS+ 1ml/l Tween20 (=0.1%)
10 X DNA gel loading buffer	50% v/v glycerol, 0.1% bromophenol blue (w/v)
5X SDS-loading buffer (for SDS-PAGE electrophoresis)	5x SDS-Loading Buffer: 312.5mM Tris pH6.8, 10%SDS, 50% Glycerin, 10%β-ME , Brome-phenol-Blue → 1563 µl 1M TrisHCl pH6.8, 2.5 ml Glycerin, 0.5 g SDS, 500 µl β-ME (TOXIC), a few crystals Bromophenol-Blue , A.bidest ad 5 ml, aliquots of 500 µl → -20°C
DEPC treated water	1 mL DEPC in 1 L DDW, shake vigorously and autoclave
LB media	10g Tryptone, 5g Yeast Extract, 10 g NaCl (Carl Roth GmbH, Karlsruhe, Germany), complete dissolving in 1 litre distilled water and autoclaving for 15 min at 121°C.
Separating gel for Western blotting (12%)	3.5 ml autoclaved dist water, 2.5 ml 4x separation Buffer, 4 ml (Bis) acrylamid and 10 µl TEMED. Then, mix by inverting the tube 5x and quickly add 100 µl 10% APS. (4X separation buffer: 1.5M Tris pH8.8, 0.4%SDS→ 36.4g Tris in 140 ml A.bidest, pH with HCl → 8.8, + 8ml 10%SDS, autoclaved dist. Water ad 200 ml)
Stacking gel for Western blotting (12%)	3.9 ml autoclaved dist water, 1.5 ml 4x stacking buffer, 0.6 ml (Bis) acrylamid and 6 µl TEMED. Then, mix by inverting the tube 5x and quickly add 30 µl 10% APS. (4x Stacking Buffer: 0.5M Tris pH6.8, 0.4%SDS → 50 ml 1M Tris pH6.8, +4 ml 10%SDS, ad 100 ml)
TE (pH 7.5, 8.0)	10 mM Tris-HCl, 1 mM EDTA

Table 7.4 Machines

Name	Manufacturer
Centrifuges	
Heraeus Biofuge 'fresco' and 'pico'	Kendro, Hanau, Germany
Heraeus Labofuge 400	Kendro, Hanau, Germany
Heraeus Multifuge 3S-R	Kendro, Hanau, Germany
Heraeus Varifuge 3.2RS	Kendro, Hanau, Germany
Micro Centrifuge	Roth, Karlsruhe, Germany
Thermocyclers	
ABI Prism™ 7700 Sequence Detector	Applied Biosystems Inc., Foster City, CA, USA
ABI Prism™ 7900HT Sequence Detection System	Applied Biosystems Inc., Foster City, CA, USA
Biometra® T Gradient	Whatman Biometra GmbH, Göttingen, Germany
Biometra® T1 Thermocycler	Whatman Biometra GmbH, Göttingen, Germany
GeneAmp® PCR System 9700	Applied Biosystems Inc., Foster City, CA, USA
Electrophoresis and Western blotting	
Bandelin Sonopuls GM 70	Bandelin electronics, Berlin, Germany
BioDoc Analyzer	Biometra, Göttingen, Germany
Bio-Rad ChemiDoc XRS System	Bio-Rad, Munich, Germany
Gel Doc XR	Bio-Rad, Munich, Germany
Gibco BRL Electrophoresis Power Supply 250 EX	BioRad, Munich, Germany
Gibco BRL Horizontal Gel Electrophoresis Apparatus	BioRad, Munich, Germany
High Performance UV Transilluminator	VWR, Hamburg, Germany
Horizontal Electrophoresis Apparatus	Bio-Rad, Munich, Germany
KERN 440-47N scale	Kern & Sohn, Balingen, Germany
Microwave R-2V18	Sharp Electronics, Hamburg, Germany
Multigel SDS-PAGE Vertical Electrophoresis Apparatus	Whatman Biometra, Göttingen, Germany
Power Pac 300 Electrophoresis Power Supply	Bio-Rad, Munich, Germany
Roller mixer (Stuart; SRT6D)	Barlworld Scentific limited, Stone, UK
Semidry Electroblotter	PeqLab Biotechnologie GmbH, Erlangen, Germany
Shaking incubator (Stuart; S|500)	Barlworld Scentific limited, Stone, UK
Pipetting Robots	
Hydra 384 Robbins Scientific	Dunn Labortechnik, Asbach, Germany
Hydra 96 Robbins Scientific	Dunn Labortechnik, Asbach, Germany
Power Washer PW384	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Carousel for Evo 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Freedom Evo 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Freedom Evo 200	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis RSP 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis Workstation 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis Workstation 200	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Spectrafluor Plus	Tecan, Deutschland GmbH, Crailsheim, Germany
Te-MO	Tecan, Deutschland GmbH, Crailsheim, Germany
Te-MO with cooling rack	Tecan, Deutschland GmbH, Crailsheim, Germany
WRC96 washing station	Tecan, Deutschland GmbH, Crailsheim, Germany

Other Machines	
3700 DNA Analyzer	Applied Biosystems Inc., Foster City, CA, USA
3730xl DNA Analyzer	Applied Biosystems Inc., Foster City, CA, USA
Axiocam	Zeiss, Jena, Germany
Axiophot microscope	Zeiss, Jena, Germany
Bambi Compressor DT/23Q	Bambi, Birmingham, UK
FACSCalibur™ cytometer	Becton-Dickinson, San Jose, CA
GFL 1086 shaking waterbath	GFL, Burgwedel, Germany
Hemocytometer	Brand, Wertheim, Germany
Heraeus 3 incubator	Kendro, Hanau, Germany
IMPLEN Nanophotemeter	Implen GmbH, Munich, Germany
Mini Vortexer VM-3000	VWR, Darmstadt, Germany
NanoDrop® ND-1000 Spectrophotmeter	NanoDrop Technologies, Wilmington, DE, USA
PCR chambers	Bä-RO® Technology, Leichlingen, Germany
Platesealer ALPS-300	Abgene, Epsom, UK
Shaking incubator (GFL 3033)	Gesellschaft für Labortechnik mbH, Burgwedel, Germany
Thermomixer 5437	Eppendorf, Cologne, Germany
TiMix Control incl. TH15 hood	Edmund Bühler Labortechnik, Hechingen, Germany
Vortex-GENIE 2 G-560E	Scientific Industries, Bohemia, NY, USA

Table 7.5 Electronic Data Processing

Name	**Source**
Laboratory information management system (LIMS) at ICMB	Details and description of the used database system at ICMB (Kiel, Germany) are given in Hampe *et al. (Hampe et al., 2001)*, and Teuber *et al. (Teuber et al., 2005)*.
Software	
Alex's splice site score calculator	http://violin.genet.sickkids.on.ca/~ali/splicesitescoreForm.html
CellQuestPro™ software package	Becton-Dickinson, San Jose, CA
ESEfinder	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home
F-SNP	http://compbio.cs.queensu.ca/F-SNP/
Gemini 4.28	Tecan, Deutschland GmbH, Crailsheim, Germany
Genemapper 4.0	Applied Biosystems, Foster City, CA, USA
Motif SNPs Input Page (beta v1.0)	www.ikmb.uni-kiel.de/motifsnps
Primer Express 2.0	http://www.applied-biosystems.com Applied Biosystems, Foster City, CA, USA
Primer3 (v.0.4.0)	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
Sequence Detection System 2.1	Applied Biosystems, Foster City, CA, USA
Sequencher 4.2 and 4.5	http://www.genecodes.com (Gene Codes Corporation, Ann Arbor, MI, USA)
SNPSplicer	http://www.ikmb.uni-kiel.de/snpsplicer/ ICMB, Kiel, Germany
SpliceTool	ICMB, Kiel, Germany
Web Resources	
AB Applied Biosystems	http://www.appliedbiosystems.com/
Alternative splicing Gallery	http://statgen.ncsu.edu/asg/
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
CEPH	http://www.ceph.fr
Columbia Biosciences	http://www.columbiabiosciences.com/

EBI-Alternative splicing Database Project	http://www.ebi.ac.uk/asd/
Ensembl	http://www.ensembl.org
ExPASy Translation tool	http://www.expasy.org/tools/dna.html
Fast DB	http://www.fast-db.com/fastdb2/frame.html
FirstEF: first-exon and promoter prediction	http://rulai.cshl.org/tools/FirstEF/
Genecards	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
HapMap	http://www.hapmap.org
InterProScan sequence search	http://www.ebi.ac.uk/InterProScan/
Microsynth Laboratory	http://www.microsynth.ch
NCBI dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
RepeatMasker	http://woody.embl-heidelberg.de/repeatmask/
Smart	http://smart.embl-heidelberg.de
SNP-BLAST	http://www.ncbi.nlm.nih.gov/SNP/snp_blastByOrg.cgi
SPIDY, mRNA to genomic alignments	http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/
UniProt	http://www.uniprot.org

Table 7.6 TaqMan Assays and SNPlex Pools

TaqMan Assays		
rs ID	**Assay code**	
rs2076530	hcv2488471	
rs2228173	hcv11764349	
rs2295773	hcv2144407	
rs2298839	hcv3212434	
rs5248	hcv2796264	
rs540819	hcv962479	
SNPlex Pools		
Pool code (Applied Biosystems)	**Design Name**	**Designed SNP-count per pool**
For first (web based) round		
w0510100067-0001	Q1-A01	47 SNPs
w0510100067-0002	Q2-B01	48 SNPs
w0510100067-0003	Q3-A02	48 SNPs
w0510100067-0004	Q4-B02	47 SNPs
For second (neural network) round		
w0609104034-0001	Pool1-D2M	47 SNPs
w0609104034-0002	Pool2-D2M	40 SNPs
w0611104666-0001	Pool1-ACC2	48 SNPs
w0611104666-0002	Pool2-ACC2	29 SNPs

8 APPENDIX

8.1. All experimentally validated SNPs and primers used for the nested RT-PCR

Table 8.1 List of all tested candidate splice SNPs and primers used for the nested RT-PCRs

The 41 SNPs, that overlapped between the ss score and NN approach are marked with an rs-number in bold print.

#	SNP ID	**Site and experiment category**	Δ_S *or* Δ_{ESE}	Δ_N	*Position*	SNP-in-sequence*	**Exon #/ size of adjacent exon**	**Gene symbol**	**Refseq annotation**	Round 1 RT-PCR *Forward 5′-3′* **Reverse 5′-3′**	Round 2 RT-PCR *Forward 5′-3′* **Reverse 5′-3′**	**Ampli-con Size**	**Genotyp-ing Method**	**Genotypes availability (11-22-12)**
1	rs17105087:A>G	AG variation at NAGNAG	-	-	(+3)	TTCTCTCTTCTGC**AG**CA**R**CC	7/165	SLC25A21	**NM_030631**	GAAATTGCTGGGATATGTGTCA AAGTCTCATAATCTTGGGAAGCAG	TCTGGACTAACAGAAGCCATTGTA AAATCCCTTCTTCCTGATAGACTG	436bp	SNPLex	2-78-12
2	rs11597439:C>G	AG variation at NAGNAG	-	-	(+3)	TCTGTGTCCCTTC**AG**AA**S**AG	2/84	CUEDC2	NM_024040	TGGCAGAAGCTCCTCCTCA GGGATGAGAGCTGCACCG	AGCAGCCGAAGACCTAGTCTCT –	451bp	SNPLex	32-16-44
3	**rs1558876:C>G**	AG variation at NAGNAG/ NN-middle-acceptor	-	0.00305	(**+3**)	GTTCTCCTGTTTC**AG**CA**S**CA	5/197	ARSG	NM_014960	TACTGATACTCCAGGCTACAACCA CTGGTGACATTAACTGGAACTCTG	CTCCCTCTTTATGAAAACCTCAAC AGTTTGCCAAAATCCAGTGAAG	375bp	SNPLex	25-19-48
4	**rs9606756:A>G**	AG variation at NAGNAG/ NN-middle-acceptor	-	0.00494	(+3)	ATTTTCTTTTCTA**AG**AA**R**TA	2/193	TCN2	NM_000355	GAGGATTAATCAGTGACAGGAAGC GATCAATGTAGGTCTTGTGGTTCA	GATTCTTGCTCACTGCTCACC AAAGGTTCCACAGCATACAGAAGT	592bp	SNPLex	74-1-17
5	rs1152888:A>G	AG variation at NAGNAG	-	-	(+3)	ACTTTCCTTCCTA**AG**GA**R**TA	5/152	IRAK3	NM_007199	CAGTGTTGAGTCCTTCAGAGAAGA GTCTCTGTAAAATATGCAGCCAAC	GAAGAGTTATCAGGAAGGTGGATT CTCTGTAAAATATGCAGCCAACTC	361bp	SNPLex	0-77-15
6	rs17036879:G>A	AG variation at NAGNAG	-	-	(+3)	TGTAAATAACTTT**AG**GA**R**CC	8/139	TSEN2	NM_025265	GCAAACAAGATGCTCTCATCC CATACATTCTGGTGACTCCATTTC	GGTCTATGCTCTGGGATGTTTAAG AGAGACATTAACGGAAACTCTGCT	325bp	SNPLex	0-90-2
7	**rs2156634:G>A**	AG variation at NAGNAG/NN-middle-acceptor	-	0.01286	(+3)	CCAGTTTTGCTGC**AG**GA**R**AA	3/204	GRIK4	NM_014619	AGAGGTCCAACTACGCTTTGAA GAGTACTGGTTCACCAGGAGGTT	GGTCCAACTACGCTTTGAAAAT AGATAGGCTAGAAGCATGAAGAGC	556bp	SNPLex	2-63-25
8	**rs3014960:G>A**	AG variation at NAGNAG/NN-middle-acceptor	-	0.0044	(+3)	GATTTCTTTATAC**AG**CA**R**AT	14/106	COG3	NM_031431	ATGATATCCACCAGTGTTACCTTG TGATAGACTCTGACGCTCCAAGTA	TAGAGACTCTGTCGGAACTTTGTG –	540bp	SNPLex	71-2-19
9	**rs4822258:G>A**	AG variation at NAGNAG/NN-middle-acceptor	-	0.00496	(+3)	CTTTCCCGTCACC**AG**GA**R**GA	8/144	TTLL1	NM_012263	TTCCTTATCAACAAGCTCTCACAG GTAATTGCCGAGGACTTCCTTA	AGTACCAGTGAGCTGGACAACAT GTTGTACTTGAGGATTCGGTCATT	354bp	SNPLex	51-9-32
10	**rs2243603:C>G**	AG variation at NAGNAG/NN-middle-acceptor	-	0.00036	(+3)	TTCCCTGATTTCC**AG**AA**S**CA	5/115	SIRPB1	NM_006065	ATAGAGAACAAGGATGGCACCTAC AAGTCCTGGTGTGTTTAGATTTGG	CAAGCAGTCAGCAAAAGCTATG AGATTTGGAGTGTTTCTCACCTTC	449bp	SNPLex	4-57-31

11	rs2290647:G>A	AG variation at NAGNAG	-	-	(+2)	CTCCTCTGTCTCC**AG**C**R**GAC	10/144	GRAMD1A	NM_020895	GACAGACACAAGTAACTCCTCTTCA CCAGATGGTGGAAATAGTCTTCA	CACAAGTAACTCCTCTTCATCCAC GGCAGTGTAGAAGTAGTCCTGGTA G	394bp	SNPLex	41-6-40
12	rs2273431:G>A	AG variation at NAGNAG	-	-	(+2)	CCTCTACTCATGC**AG**A**R**GAT	10/144	NID2	NM_007361	ACTTGACCCAGAGAACTACCTGAG CTTCCAGGCAAGTTGATACATACA	CCAGAACATCACTTACCAGGTGT CCAGGCAAGTTGATACATACAGAG	366bp	SNPLex	0-79-13
13	rs7862221:A>G	AG variation at NAGNAG	-	-	(+2)	TTTGTTTCTCTTC**AG**A**R**GAG	14/105	TSC1	NM_000368	TACACAGACAACACCATCTTCTGA GCTTTCTTTAACAGCTCCTCAGTC	ACAACACCATCTTCTGAATGACAG AGGTCTATGGGAGTAAAGGCTTG	369bp	SNPLex	1-68-22
14	**rs2275992:A>G**	AG variation at NAGNAG/NN-middle-acceptor	-	0.0002	(+2)	TTTACTTATTTTT**AG**T**R**GTG	5/105	ZFP91	NM_053023	– TCTCACAACGGACATACTGGATT	TGGCGTAGTAGTAGGACATCTGTT GACATACTGGATTGGAGGCTTTT	480bp	SNPLex	47-5-40
15	**rs1152522:A>G**	AG variation at NAGNAG/ NN-top-acceptor	-	0.97073	(-2)	TGTTGTCTTTCAT**RG**CAGGT	4/102	C14orf105	NM_018168	AGAAAAGACTTCATATTCACTGGCAA TTCAAGATTTCATCAGGCAACATG	GCAGGACCAGAATAAAGCCTTGG TGGTTAGAAGGTCATGGTCATCAT TC	473bp	Direct sequencing	5-47-28
16	rs2307130:A>G	AG variation at NAGNAG	-	-	(-2)	TTTCAAATCCTCT**RG**AAGCC	2/90	AGL	NM_000644	GTCTACGGCAGCTATTCCAGAG CCTAGATAGTCCAAGAGTCTGCAA	ACTGCTTCCCTCTGTTCTCATCT GTAAAATGGGGTCCACAACTATGT	448bp	SNPLex	23-23-46
17	**rs5248:A>G**	AG variation at NAGNAG/ NN-middle-acceptor	-	0.00244	(-4)	TCTTCCTCACA**R**C**AG**GTCTA	3/136	CMA1	NM_001836	AGATGCTGCTTCTTCCTCTCC ATTGTGGTCAAAGTCTCTGAAGTG	CCTGCTGCTCTTTCTCTTGTG ATGAGTCTCAGCTTCACCTCTTG	489bp	SNPlex/ TaqMan	82-1-9
18	rs2292402:T>A	AG variation at NAGNAG	-	-	(-5)	GTGTGTTTGG**W**GC**AG**TGAGT	2/103	ACPL2	NM_152282	GGAGCTGGCGGCGAG TGGGAATGACATACAGTGGGTAC	– TGGCGAATGAACACATGC	585bp	SNPLex	4-74-14
19	**rs2071558:C>T**	N variation at NAGNAG/NN-middle-acceptor	-	0.00932	(-6)	CAGTGTCCC**Y**AGC**AG**GTAAT	6/231	AMHR2	NM_020547	CTGCTACAGCGAAAGAACTACAGA GTGGGCAATACCTGGTTTATATTG	TACAGCGAAAGAACTACAGAGTGC GGGCAATACCTGGTTTATATTGG	482bp	SNPLex	66-4-22
20	**rs12905385:C>T**	N variation at NAGNAG/ NN-middle-acceptor	-	0.04271	(-3)	ACTTCACTGATA**YAG**GAGAG	20/134	CDAN1	NM_138477	AGTTTCTCTCCTTTGCTGACCAT CTGGGAACACAAGATCTCCAAC	AATATTACCGGGACATCTTCACTC CTGCTCTTGGAGAAGTGACTCTG	569bp	SNPLex	2-66-24
21	**rs2250205:C>T**	N variation at NAGNAG/ NN-middle-acceptor	-	0.06068	(-3)	ATCTTTGATTGA**YAG**GAGAC	5/177	EIF6	NM_181469	CTTCGTTCGAGAACAACTGTGA TTCATTCAGCTTGAAGACACTCTC	GGAGGCTCAGAGAACTTCTACAGT ACTGACAGCTCTGTGCTGGTT	390bp	SNPLex	4-52-35
22	**rs2174769:T>C**	N variation at NAGNAG/ NN-middle-acceptor	-	0.00639	(-3)	CTGTTTGAATTT**YAG**GAGCG	3/599	SNIP1	NM_024700	AAGTCTCCTCGCAGTAAGAGAAAC CCTCGTCATCTTTCCTGTCTATTT	CACTCAACAGTCAAAGTGAAGCA GCAAGACGTATTCTCTGCTACTGA	805bp	SNPLex	3-68-21
23	rs12944821:G>C	N variation at NAGNAG	-	-	(+1)	TCTTTATATTTTC**AGS**AGGC	3/122	AP1GBP1	NM_007247	TTCATGTTTCCTGTTGCAGGT TCCAAAGCATCATCTCTACTCTTCT	CTGTTGCAGGTGGGATAAGAC GGTTTCACACTGCTCAAAAGTCT	385bp	SNPLex	1-65-26
24	rs879022:G>A	N variation at NAGNAG	-	-	(+1)	CTTTTCCCAGGAC**AGR**AGGC	3/114	REG1P	D56494	TGCTCCTTAAGCAAGAGATTCAC AGCTTATCTCGAAGAACCTATGGA	GCTCCTTAAGCAAGAGATTCACTG GATAATCAGGAGGTAGAAGATGCA G	383bp	SNPLex	81-1-10
25	**rs515071:T>C**	Acceptor scoring/ DEATH/ NN-middle-acceptor	8.3	0.00051	(-3)	CCCTTCCTTTCC**YAG**ATCAT	41/66	ANK1	NM_020475	GGAACAAGTACCATGACTGAAGG GCTTGTTTTCTATCCCTCTCTCTC	CAAGAACACCTTCACCCAAGT ATGCGTCTACAGTCAGTCATTCAT	354bp	SNPLex	45-10-36
26	**rs3793326:T>C**	Acceptor scoring/ NN-middle-	8.3	0.00019	(-3)	TCTCCTTTTCAA**YAG**GTTTT	12/157	CASD1	NM_022900	CCTGTATACATGCACATTCGAGTT ATAACGGTCTAACCTCCATCTGAA	GGGCATTTCTCATACTTTTGGA CCATCTGAACCACCATTCATATAC	372bp	SNPLex	16-29-47

		acceptor												
27	**rs9986447:T>C**	Acceptor scoring/ NN-middle-acceptor	8.3	0.01466	(-3)	TTATTTCTCTAA**YAG**AGGTA	2/164	PEX6	NM_000287	GCCACTTTGGCTTTTAATCTTG AAGGACACTGCTAGTTCCTGTCA	CCACTTTGGCTTTTAATCTTGG GAACAGGGCTCAGGGTAGAAC	435bp	SNPLex	34-16-42
28	**rs2291662:T>C**	Acceptor scoring/ NN-middle-acceptor	8.3	0.00774	(-3)	TTTGTTGCTTCA**YAG**CGGTG	50/107	SMG1	NM_015092	ATCACTCCCACCTTGAAAGAACT TCGCTTAACCAGACTCATCTACTG	CTGATGTCATGTCACAGAATGCTA CTCATCTACTGTCTTGCCATTCAC	352bp	SNPLex	12-37-42
29	**rs2070410:T>C**	Acceptor scoring/ NN-middle-acceptor	8.4	0.00037	(-3)	TTTTAATTTCTT**YAG**ATGCA	27/88	TIAM1	NM_003253	ATGAACAAGGTTGCCAGTCAC TCCTCATACTGAGCAAGATCAAAC	GCTGAACAGACTGGTGAGAAAAA GAGGAGCTGTCTTCTGTGCTTATC	540bp	SNPLex	14-24-53
30	**rs3811609:T>C**	Acceptor scoring/ NN-middle-acceptor	8.4	0.00741	(-3)	TGTTTGTTTTTA**YAG**GGGCT	4/99	FLJ20160	NM_017694	AGAATGCCTGGACTGTTCTCC AACTGAAAGTCCTCATTGAACTCC	ACTGTTCTCCCCATGGAAGTT GGGTGTGAGATGAGCAGGAT	798bp	SNPLex	6-42-43
31	rs790055:T>C	Acceptor scoring/IG	8.3	-	(-3)	TTTCTTTCTCCT**YAG**GCTCC	5/69	F11R	NM_144503	GAGGAAACTGTTGTGCCTCTTC CCTCACAGCTGTATTCTCCAGTATC	GGAAACTGTTGTGCCTCTTCATA CATCTTTGAACCAGGTGTATTCAG	483bp	SNPLex	65-6-21
32	**rs2255632:T>C**	Acceptor scoring/ NN-middle-acceptor	8.3	0.00365	(-3)	TCTTTACTCTCA**YAG**CGGCG	34/200	KIAA0467	NM_015284	CCTGGTGCATTACTGTGCAA TAGGTGGTTACCGGAGAGCTAGT	TTACTGTGCAACAGCCATGC TCCCATCAGTCTTGGTTTTAGG	365bp	SNPLex	16-39-37
33	**rs607755:T>C**	Acceptor scoring/ NN-middle-acceptor	8.3	0.0003	(-3)	TCCTTTTTCTTT**YAG**CTGAA	6/79	RELN	NM_005045	CAGTTTGGTAACCAGTTTATGTGC ATTTTCTCTAGCTGAATCCAGTCC	CACAACCAACCTCAGTTTCATCT ATGCTGGGGTCTGAATAACTAAAG	435bp	SNPLex	20-23-49
34	rs11658717:G>A	NN-top-acceptor	-	0.99341	**(-7)**	TCTTTTTA**R**TACT**AG**GTTAG	6/211	STXBP4	NM_178509	CAGCTGTTAAATCCAAGGCTACTT AAAGACACTGTCCCTTTTGAGTCT	GTTAAATCCAAGGCTACTTTGGTG GTCCCTTTTGAGTCTGCTTGTACT	778bp	SNPLex	50-4-36
35	rs12857479:G>A	NN-top-acceptor (Positive control at acceptor)	-	0.98135	**(-1)**	TGTTTGCTTTGATA**R**GACAT	4/157	C13orf26	NM_152325	CCAAAACGGTATCAGAAGATTAGG GTGTCACAGTGAGTACGAATAAAGC	– CAGTGAGTACGAATAAAGCGATCA	699bp[¶]	SNPLex	11-41-39
36	rs10774671:G>A	NN-top-acceptor (Positive control at acceptor)	-	0.97416	**(-1)**	GTCTCACCCTTTC**AR**GCTGA	6/514	OAS1	NM_016816	AAAAGTACCTGAGAAGGCAGCTC AAGTGAGGCTGTGGAGAATGTTAT	AGTACCTGAGAAGGCAGCTCAC GGCTGTGGAGAATGTTATCTATGA	542bp	SNPLex	41-15-36
37	rs3818780:C>G	NN-top-acceptor (Positive control at acceptor)	-	0.89773	**(-1)**	GCTTTGCATCCTC**AS**AGCAT	2/297	AVPI1	NM_021732	GATACCCTCTGCCATGCTCTT CTCTTTCCCTGGATCAGTGTCT	CTGGCCCTTGTAAGCACCT AGTGCAGATACTGCTCACTGGA	502bp	SNPLex	61-6-23
38	rs1805377:G>A	NN-top-acceptor (Positive control at acceptor)	-	0.80319	**(-1)**	TGATTTTCTTTTC**AR**TTCTA	8/589	XRCC4	NM_022406	TAAGGAAGCTTTGGAGACTGATCT TTTCAAAATCTTCCCAGACAGG	CTCAAGAACGAGAAAAGGACATC TCAAAATCTTCCCAGACAGGAT	743bp	SNPLex	2-71-19
39	rs2073193:C>G	NN-top-acceptor	-	0.64333	(-3)	CTTCCCCCAATT**SAG**GCCGA	3/99	IDH3B	NM_006899	AAACATGGCGGCATTGAG ATTGTTGTGCCGAGTCATATACC	– AAGTGACTTCACATGGACTACGTT	448bp	SNPLex	42-7-38
40	rs3793809:G>T	NN-top-acceptor	-	0.43985	(-12)	ATT**K**GTCCTAACT**AG**GGGAT	3/2198	EIF4EBP2	NM_004096	CGCAGCTACCTCATGACTATTG GAAAGTCAGAGTTGAACTGTTTTCC	GACGCTCTTCTCCACCACAC TGTTTTCCTCTAAGGGCTGCT	522bp	SNPLex	44-7-41
41	rs17155183:A>T	NN-top-acceptor	-	0.41291	(-3)	TTGTTGTATTTT**WAG**CGGTT	2/109	PTPN12	NM_002835	CAAGTGGAGATCCTGAGGAAA CCTCATTAAGCTTATCATGTCCAG	ATGAAGAGTCCTGACCACAATG –	603bp	SNPLex	47-6-39
42	rs2105702:T>G	NN-top-acceptor	-	0.34145	(-10)	TTGAT**K**CATATTT**AG**TGCCC	21/135	CTNNA3	NM_013266	AAAGAGTAAGCTGGATGCTGAGAT	ATGATCATGATGGAGATGACAGAC	368bp	SNPLex	39-7-44

										CCTCTGGCTTCTCTCTTTTAATCA	ATCTTGGTTGAGGCAATGTAAGAC			
43	rs1534904:A>C	NN-top-acceptor	-	0.3268	(-11)	CTTT**M**AAATCCTC**AG**TGGTT	6/186	SELE	NM_000450	ACGGTGAATGTGTAGAGACCATC TCACAAACTGGGATTTGCTGT	GGTGAATGTGTAGAGACCATCAAT AGGATGATTTGAAGGTGAACTCTC	538bp	SNPLex	43-13-34
44	rs591058:A>G	NN-top-acceptor	-	0.3262	(-14)	C**R**ATTTCTTAAAT**AG**GGACC	5/165	MMP3	NM_002422	CTGTTGATTCTGCTGTTGAGAAAG GCTCGTACCTCATTTCCTCTGATA	TGTTGAGAAAGCTCTGAAAGTCTG TCGTACCTCATTTCCTCTGATAGC	721bp	SNPLex	19-23-48
45	rs251683:C>A	NN-top-acceptor	-	0.32102	(-12)	CAC**M**TTCTCCCAA**AG**CTGCC	6/121	PLA2G4C	NM_003706	GCTCTGAAGAAGCTAAGGATTGAG CATAAACCTTTCAGGGTCAGATTC	CTCTGAAGAAGCTAAGGATTGAGG ATGACTTCAGTGTTACCAAGAGCA	658bp	SNPLex	34-14-42
46	rs786906:T>C	NN-top-acceptor	-	0.31819	(+1)	TTTATCTTCTATA**AGY**GATT	12/127	PKN2	NM_006256	GGGGAAGGCTAGTAAGAAGAGCTA TACTTCATCTCGAGCCACAATATC	ACAGTAAATCATTCTGGCACCTTC TTCTAGGCCTGACTGAGGAGTATC	324bp	SNPLex	13-27-50
47	rs2592828:C>G	NN-top-acceptor	-	0.29862	(-10)	AGTCT**S**TTTTAAA**AG**CCAAC	4/1080	TACC3	NM_006342	ACAGACGCACAGGATTCTAAGTC CTCTCTGCTGTTGGGGTCTC	ACAGGATTCTAAGTCCTAGCATGG –	1258bp	SNPLex	26-21-42
48	rs3763131:A>G	NN-top-acceptor	-	0.28899	(-9)	CCGTGC**R**TGGCCC**AG**AGCTG	16/128	GFPT2	NM_005110	GTCAGTTCATCTCTCTGGTGATGT ACTTGGAACTTTCAGTATCGTCCT	ACTACAAAACAGGAGGCAAGAGAT TATCGTCCTTGGAGCACAGTATAA	380bp	SNPLex	65-2-23
49	rs1077340:G>A	NN-top-acceptor	-	0.27314	(-6)	TGATTTGTT**R**AAA**AG**GTTTG	5/226	ELAVL1	NM_001419	TAGAAGACATGTTCTCTCGGTTTG GGCTTCTTCATAGTTTGTCATGGT	– GGTTGTAGATGAAAATGCACCAG	387bp	SNPLex	53-6-32
50	rs2296804:C>G	NN-top-acceptor	-	0.26896	(-12)	GGG**S**CTCTGTTAC**AG**TAAGT	6/346	GNMT	NM_018960	TCAGTGCTGATAGTGAACAACAAG TATATTGTTTACCCTTCCGTCTGTG	CACATGGTGACCCTGGACTATAC ACCCCAGCTGTAGTCTGCTCTA	354bp	SNPLex	27-19-45
51	rs535801:G>A	NN-top-acceptor	-	0.23791	(-6)	AATTATTTT**R**CAT**AG**GCAGA	7/142	MRE11A	NM_005590	TTATGGAGAAAGATGCAGTCAGAG CTTCCATGTTTACTCCTGTTCTGA	CCCTCAAGGAAAACATTACATACC CAAGAGTTCTCATCTTCCTTTGGT	431bp	SNPLex	37-7-47
52	rs435806:T>C	NN-top-acceptor	-	0.2335	(-3)	CTGCCCATTACC**YAG**ATAAT	16/151	TLE6	NM_024760	AGTCAAGAGTATCGTGGTCAAGG CCCCTCAGTAGGTGATCTGGT	ACCTCTGGAGTACCAATTCAAGTC GTTGTTGGAAGAGACGTCACAG	365bp	SNPLex	6-44-40
53	rs10510594:T>C	NN-top-acceptor	-	0.22505	(+2)	GTGATATTTATAT**AGA**Y**G**AT	22/122	NEK10	NM_152534	GATCATCTTGGAAGTGGAGCTTT ACTATTTTTGTAGCCAAGGACAGC	– AAGGACAGCATGTTAGTGCTGTAG	506bp	SNPLex	8-43-41
54	rs3819255:A>T	NN-top-acceptor	-	0.21015	(-11)	AAAA**W**TTTTTAAT**AG**GTATC	6/105	CHKA	NM_001277	AGAGCGTTATGTTTGCCATTCT CTGTTGTTTCTTGGTGGGATACTT	GCCAAAACTCTATGGCATCTTT –	547bp	SNPLex	13-42-35
55	rs2169456:C>A	NN-top-acceptor	-	0.19176	(-10)	CCGCA**M**TTCTGGC**AG**AATTA	47/161	KIAA1529	NM_020893	AGCAAAAAGATCCTGGAGTATCAG ATCAGGGGTTTCTCACTCTCTTC	AGGCAAATAAGTACCACAACTCCT GAGTTTCCTTCGTATGAGCATTG	494bp	SNPLex	16-24-52
56	rs2273540:T>G	NN-top-acceptor	-	0.19018	(-14)	C**K**TTTAAACTTTA**AG**GATAT	10/319	DEPDC7	NM_139160	ACTACTGTATTTCATGGCTGTTGC TGGCAGAAAGTTTTGAATCCTC	– GGCAGAAAGTTTTGAATCCTCA	382bp	SNPLex	18-27-47
57	rs181390:T>C	NN-top-acceptor	-	0.18899	(-9)	CCTGCC**Y**GAACAC**AG**ATTCT	7/140	BID	NM_001196	CATAAGGAGGAAGCGGGTAGT TCACTGTTGTGTGAAAGACATCAC	– ACTGTTGTGTGAAAGACATCACG	674bp	SNPLex	14-42-32
58	rs741932:C>T	NN-top-acceptor	-	0.18431	(-3)	CCTGCGGCCCCA**YAG**CGGGC	3/113	PQBP1	NM_005710.2	CTATGACGATGATCCTGTGGACTA GAAGCTTCAATCCTGCTGCT	TATGACGATGATCCTGTGGACTAC CTTCAATCCTGCTGCTTGGT	704bp	SNPLex	30-39-20
59	rs3745503:A>C	NN-top-acceptor	-	0.1786	(-9)	CCACAC**M**CTCCCC**AG**CTGGA	25/207	MYH14	NM_024729.3	GAAGAGGAGCGAGACCTGAAG CCTCATATTTGAGCCGTAGCTTAT	GACATCATCGTCTCCTTCCAG TAGCTTATTGAGGCTCTTGACCTT	722bp	SNPLex	71-3-17
60	rs2075863:G>A	NN-top-acceptor	-	0.16218	(-11)	GGGT**R**TCTTCTAC**AG**TGTTT	5/97	PFKFB2	NM_006212.2	CTTCCTCAGAACAGAACAACAACA TGTCATAGTTGTCTGGGTCAAGAG	CAAGAAACTAACACGCTACCTCAA CAAGAGGTCGGTAGGTAACTTTGT	449bp	SNPLex	15-27-48
61	rs3746003:C>T	NN-top-acceptor	-	0.15978	(-13)	AC**Y**GGGGTGTGGC**AG**GAGTG	5/165	CADM4	NM_145296	ACCGCAAGGAGCTGAAAG GCAAGTGTAGGTGCCGTTATC	– CAAGTGTAGGTGCCGTTATCC	391bp	SNPLex	3-66-21
62	rs1734432:A>G	NN-top-acceptor	-	0.15505	(-11)	GTCT**R**TTTAAATC**AG**GAGCT	10/73	PDIA6	NM_005742	CAGCTTCAGAAGTAAAAGAGCAGA CTAAGTCATCAAGCTCCACATCAC	GAAGTAAAAGAGCAGACGAAAGGA –	691bp	SNPLex	21-22-48
63	rs1382543:G>A	NN-top-acceptor	-	0.15276	(-7)	TATTTGCT**R**CCTA**AG**AGCCA	3/221	MSH3	NM_002439	ACAAAAGGAAGGAGGAAGTGATCT	AAAAGGAAGGAGGAAGTGATCTG	388bp	SNPLex	7-53-31

										TGTAATTCTAGCGGCGTATAGATG	-			
64	rs2990986:G>A	NN-top-acceptor	-	0.14914	(-14)	T**R**TAAAATTATAC**AG**AAAAC	13/86	CCDC7	NM_145023	AGCAGTGAAAAGTTGTGAAGCTCT TATCACTTGTCCTTTGTCCACCT	AATTCTTAGAAGCCCACTCAACTG GTCTCAGTTTTCCCCTCACTTTTA	538bp	SNPLex	8-44-40
65	rs7752421:A>T	NN-top-acceptor	-	0.14746	(-6)	CATCTTCAC**W**TTC**AG**TCTTC	16/110	SNAP91	NM_014841	TTGCTTGCTACAATGATGGTGT GGAGGTCCAGGAGATCACTAGAT	TGCTTGCTACAATGATGGTGTT TCCAGGAGATCACTAGATGGTTTA	429bp	SNPLex	27-16-47
66	rs783544:T>G	NN-top-acceptor	-	0.14702	(-11)	TCCT**K**TCCAAAAC**AG**ACTTC	3/189	CPEB1	NM_030594	AGGATAAAAGATTGCTGGGACA GATGATCTGATCCAGAGCTGAAG	TAACTGAGGGTGCTGGAAACTT TCTGGATTTCAGTAGAGTCTGCAC	368bp	SNPLex	4-46-42
67	rs920791:T>A	NN-top-acceptor	-	0.14529	(-12)	ATG**W**TTCTACCTT**AG**GAAAG	5/219	CCDC11	NM_145020	CTATCCATCAGAAGAAGGTGTGTG CTTCTTTGCCTTGTCTTCCTCTAA	AGACAGAAAGAGCTGATGGAGAAC CTTTCTCCTGAGCTTTTTCTTCCT	412bp	SNPLex	42-19-29
68	rs11024770:G>C	NN-top-acceptor	-	0.14326	(-15)	**S**GGACTTTGATCC**AG**CTGGA	5/100	IGSF22	NM_173588	GTGGAGTTCTTCAGCTTAGTGACC CAGTTCCATTATGCAGTCAAAGAC	TGGAGTTCTTCAGCTTAGTGACC CTTCTTCATCTCTTTGAGCTTCCT	539bp	SNPLex	5-63-22
69	rs2240340:T>C	NN-top-acceptor	-	0.14262	(-15)	**Y**TGATGGGATTTC**AG**AAATC	4/68	PADI4	NM_012387	AGATTTCATACTACGGACCCAAGA CTCACTTTGTCCATCTCAGACCT	GATTTCATACTACGGACCCAAGAC -	357bp	SNPLex	14-28-48
70	rs2607628:C>T	NN-top-acceptor	-	0.1415	(-4)	TAATCATTGCA**Y**T**AG**ATAAA	63/91	PKHD1L1	NM_177531	GACAGATGGATTGGACATAGATGA GGGCTACTTCCAACAATTAATGAG	CCAGGAACCTATCAGAACAGAAAA ATAAATGTGCACACTCTCTGTGGT	324bp	SNPLex	8-33-51
71	rs330924:G>C	NN-top-acceptor	-	0.14027	(**-9**)	TTTCCA**S**GGTTCT**AG**CCTGT	2/5.414	PPP1R3B	NM_024607	ACACCGACGCTCACGTAGT TTTCACGGTCTTCTCAAATGC	- GGTCTTCTCAAATGCGAGGT	588bp	SNPLex	46-11-33
72	rs10511687:T>C	NN-bottom-acceptor	-	0.00002	(+3)	TTTCATGTCTTAT**AG**GT**Y**AG	6/205	KIAA1797	NM_017794	CTCATCCTTTGATAACTGTGCTTG TTAAGCTGACTTCACTGACACACA	- GACTCATCTGGGTAAGCTGAATTT	408bp	SNPLex	37-11-43
73	rs1800774:C>T	NN-bottom-acceptor	-	0.00002	(-14)	G**Y**TTTTTTATTTC**AG**GATTA	13/34	CETP	NM_000078	CTTCAGTGATGGTGAAATTCCTCT CAATCTCCATCTCCGTACTCCTAA	AGACCAGCAACATTCTGTAGCTTA ATCTCCGTACTCCTAACCCAACTT	487bp	SNPLex	40-9-42
74	rs2243396:C>T	NN-bottom-acceptor	-	0.00002	(-6)	TGTCCCTCT**Y**TGC**AG**GGCCC	8/90	DTX1	NM_004416	ATCCGCATCGTCTATGACATC GGTAGGGAGGTTTCCAGTACTCTC	ATCGTCTATGACATCCCCACA GAGGTTTCCAGTACTCTCTGCTTG	576bp	SNPLex	4-53-28
75	rs2244182:G>C	NN-bottom-acceptor	-	0.00002	(-12)	TCT**S**TCTGTTAAC**AG**GACTT	8/607	FHL2	NM_201555	AGAGTTTCATCCCCAAAGACAA CAGAGCTGTAAATAACAACTGGTCA	CCTGCTATGAGAAACAACATGC GGGACTGAACTATCACAAAGCACT	563bp	SNPLex	3-63-25
76	rs2290124:T>C	NN-bottom-acceptor	-	0.00002	(-6)	ACTTCCTCC**Y**TCT**AG**ATCTC	10/41	ACTR1B	NM_005735	GACCGATTACTCAGTGAAGTGAAG AACTGGCTACCCAGGCATC	ACCGATTACTCAGTGAAGTGAAGA GTCCATGCAGCTCAATGTTACTT	399bp	SNPLex	4-39-46
77	rs2296160:A>G	NN-bottom-acceptor	-	0.00002	(+3)	TTGTCTTCCTTTT**AG**GT**R**CA	36/24	CR1	NM_000573	GAATGGAATCTCGAAGGAGTTAGA AGAACAGTGACCACCTTACAAACC	TGTGAAGATGGGTATACTCTGGAA ATGTCTGCGTCTCTGTGAAGTC	435bp	SNPLex	57-5-28
78	rs2445738:C>T	NN-bottom-acceptor	-	0.00002	(-8)	TTTTTTC**Y**GTTGT**AG**GACCT	3/18	GLDN	NM_181789	CGAGTGATGGTGGACCTGT ACTAGGGTATTAGAACCCCCTTTG	AGTGATGGTGGACCTGTGC CTAGGGTATTAGAACCCCCTTTGT	806bp	SNPLex	39-10-42
79	rs2833929:G>T	NN-bottom-acceptor	-	0.00002	(-7)	ATGTCTTT**K**TTCC**AG**GACGC	31/71	SYNJ1	NM_003895	TGACTATAGTGCTGAAGTGGAGGA TCAGTGGTTCAGGAAGGAAAGT	TGCCAGTCACCTACAATATCAGAG TTGGCTTTCAGGAGTCAGTCTT	522bp	SNPLex	3-47-37
80	rs4751995:A>G	NN-bottom-acceptor	-	0.00002	(+5)	TATTTCTTTGGAC**AG**GTTG**R**	11/112	PNLIPRP2	NM_005396	CTTTCCAAATGGAGGAAAGGA GCTCAGATAGATTTATCCCACGTT	GTATTACTCAAGCAGCGTCCTCA -	434bp	SNPLex	26-26-38
81	rs1205817:T>C	NN-bottom-acceptor	-	0.00001	(-6)	CCCTTGTCT**Y**TCC**AG**GCCTG	7/92	POU2F2	NM_002698	AGAAATGGACCAGACACTAATCATC AAAACTCTTCTCTAAGGCGAAGC	ATGGACCAGACACTAATCATCAGA -	794bp	SNPLex	5-56-26
82	rs12205497:A>G	NN-bottom-acceptor	-	0.00001	(+3)	GTTTTCTCATTTC**AG**AT**R**CC	5/149	CRISP1	NM_001131	GACCAATTTAATAAGCTCGTCACC GATGTATCAGGATGGGAGTTAAGG	ATTGTGATATGACAGAGAGCAACC CACAGACAAGTGGCTTTACAGAAT	490bp	SNPLex	3-65-23
83	rs1264894:T>C	NN-bottom-acceptor	-	0.00001	(-6)	CTTTGTTTC**Y**CAC**AG**GGCTG	2/30	OVGP1	NM_002557	TCACAGCTATCAGACCATTGAGAT CTTCCTAGAAAGCGCACATCATA	TATCAGACCATTGAGATGTGGAAG -	623bp	SNPLex	27-14-50

84	rs2932777:C>T	NN-bottom-acceptor	-	0.00001	(-8)	TTTTTTC**Y**TTTTC**AG**CGTGG	4/949	MRPS36	NM_033281	GAAGAGACAATCCTAAACCCAATG CCCTAAAACAGCAATTCTGACTG	– CCATGCCTAAACAGAAAGTAAAGT G	592bp	SNPLex	26-13-50
85	rs4148437:T>C	NN-bottom-acceptor	-	0.00001	(-5)	TTTTCCTCAT**Y**GT**AG**GTTCT	3/121	ABCC4	NM_005845	CCATAAACGGAGATTAGAGGAAGA TGGCCATGTTACTAAGACGAAGT	ATATGTATTCAGTGCTGCCAGAAG AATCATATGGCACATGGCTACTC	395bp	SNPLex	16-40-35
86	rs646348:G>A	NN-bottom-acceptor	-	0.00001	(+5)	TTCTTCTTTTCAC**AG**CTTC**R**	9/503	CCDC90B	NM_021825	AGGAAAGACATGGTCATCCTAGAG GGTCTGAGTGACAGCTTTTCAATA	CAAATCTGAGAGCAGAGAATGAGA ATAAAGACACACACCTGCACTCAC	565bp	SNPLex	14-28-48
87	rs2704766:C>T	NN-bottom-acceptor	-	0	(-6)	TCTTTCCTT**Y**TGT**AG**CATGT	3/135	KRR1	NM_007043	AGGACAATCCCAGAGGACTTTT TAGATCCTTTGGGACCAATAAGC	GACTTTTGGAGGAGAGCAGTTTC –	353bp	SNPLex	39-7-46
88	rs3737498:C>A	NN-bottom-acceptor	-	0	(-6)	TTATTTTTC**M**TAT**AG**GATTT	2/134	SCYE1	NM_004757	ACGGTTGTTACTGCTGTAGACTGT GACCTACATGCTTCCTGCTGT	ATAGAATTAGCGTGCAGTGGAGTA ACCTACATGCTTCCTGCTGTG	350bp	SNPLex	64-4-22
89	rs9822885:T>C	NN-bottom-acceptor	-	0	(-6)	TTGCTTTCC**Y**TAC**AG**TTACC	2/1426	ARPM1	NM_032487	CCCAGTTTATCTACCCGAACATTA ACAAAAGCTCTCCTTGATGTCTTC	– CAGACAGTAACCCTCAAAGATGG	431bp	SNPLex	52-6-33
90	**rs10101626:G>T**	Donor scoring/ NN-top-donor/ WD40 (Positive control at donor)	18.2	0.98826	**(+1)**	AGAAG**KT**AAAAAATAGTGTT	19/195	WDR67	NM_145647	ACCTAGAAATGAGACAGCTGGAAC CTTAATAAGATTCTGTCCCGTGGT	GATGCCTATAGACGAAAAGTGGAT GTTCTACAGCCGCATTTATCTCTT	400bp	SNPLex	13-8-67
91	**rs3816989:G>A**	Donor scoring/ NN-top-donor (Positive control at donor)	18.2	0.99659	**(+1)**	CTGAG**RT**ACGTGTGTGATTT	4/125	TCTEX1D1	NM_152665	CAGAGCAGCTCATTCATGGAAGA TCACTTTTAGGATCCCAGAGGCA	AATCATGAATTTTGGCGAAAGGAA TATGCTCTGCCTGTTCAGTTGTCC	352bp	Direct sequencing	59-2-20
92	**rs2275742:A>G**	Donor scoring/ NN-middle-donor	14.5	0.00687	(+5)	TTAAG**GT**AA**R**ACACATTGCT	14/126	RGS7	NM_002924	ATGTCAAAAGTCGCTGACAGTCTA AGGCACTGGATCTTATAAAACGTG	ATGTCAAAAGTCGCTGACAGTCT GAGCATCTTCAAATGTGTATCGTC	442bp	SNPLex	30-17-45
93	**rs2584627:A>G**	Donor scoring/ NN-middle-donor	14.5	0.00299	(+5)	AAGAG**GT**GA**R**AGGAGCTGGG	12/156	FTSJ3	NM_017647	GTCGCTACTAAACTGGAGAACAAA AGGTCACTATCCAGAGATGTGTCA	TCGCTACTAAACTGGAGAACAAAA C ATCCCCTTGTGTTACTTCCTCTAA	399bp	SNPLex	9-40-43
94	**rs3749234:A>G**	Donor scoring/ WD40/ NN-middle-donor	14.5	0.00171	(+5)	AGATG**GT**AA**R**TGAAGCATTT	8/64	TBL1XR1	NM_024665	CCCTCCTAATAAAGCTGTTGTGTT TGTACTACAAGAAGCAAAGGTGTTG	CTCCTAATAAAGCTGTTGTGTTGC TTGTCTACTCCAGCACTTAGGATG	391bp	SNPLex	59-4-29
95	**rs11046589:A>G**	Donor scoring/ NN-middle-donor	14.4	0.00004	(+5)	AACAG**GT**AA**R**TTTTACCGCA	5/33	MFAP5	NM_003480	GGAGTGGCTCTGTTCATCTTATTC TGTTCCTTACAGACAAGACGAGAG	AAAGTAGGAACAGCGTAAGAGGAG GAGTAGAGCCTTGTGCAGGTAAAT	368bp	SNPLex	6-41-45
96	**rs2298839:A>G**	Donor scoring/ NN-top-donor	14.4	0.78973	**(+5)**	ATGGG**GT**GA**R**GAGTCTTGCT	7/130	AFP	NM_001134	TCCTGTATGCACCTACAATTCTTC AACACTTCTCCAATAACTCCTGGT	CCTGTATGCACCTACAATTCTTCT T TGTTTGACAGAGTGTCTTGTTGAG	377bp	SNPlex / TaqMan	12-34-46
97	rs7314152:T>G	Donor scoring	14.4	-	(+5)	CCAAG**GT**AA**K**GAGCAGAGGA	8/78	SLC26A10	NM_133489	CTTCTGTGGACACAAGATACCAAG CAAAGTACAGTGGTGTTGGATAGC	TTCTGTGGACACAAGATACCAAGT AGTACAGTGGTGTTGGATAGCTCA	539bp	SNPLex	8-43-41
98	**rs3213591:T>G**	Donor scoring/ NN-middle-donor	14.4	0	(+5)	TGCAG**GT**GA**K**TGCTGGTGCC	17/266	MLLT4	NM_005936	CAGTCCACTTTAAGTTGTCCCCTA GTTGGGATAAAAGGCTCATCA	TTCTAGATGACCCTGAAGAGAACA GGAATGTCATCAGGTGCATACTTA	357bp	SNPLex	17-27-46
99	**rs2076530:A>G**	Donor scoring/NN-top-donor	12.5	0.12612	**(-1)**	GGTA**RGT**AAGAATTCTAGAT	5/348	BTNL2	NM_019602	CTGTTAACCTGCCAGCTACTCCC CTTAGCAATGTCTGCACGTGGA	GATGGAGTGGAGGTGACTGAGATG GCTGCATTTCTCCATCTTCTTGC	619bp	TaqMan	31-20-40
100	**rs1397548:A>G**	Donor scoring/	12.5	0.02413	(-1)	AACC**RGT**AAGCAACCTACAT	17/210	LPHN3	NM_015236	GCTCCTGACAACAAATAAGACACA	ACATGCTCTTGTAACCACCTAACA	399bp	SNPLex	10-46-36

		NN-middle-donor								AGTTGCTGGTCCTATAAAACTCCA	CCTACGTGAATGTTCACTCTCAAA			
101	rs7214723:A>G	Donor scoring	12.4	-	(-1)	GAGG**RGT**GAGTTGTCCACCC	12/74	CAMKK1	NM_032294	GAGGACAACCTCTATTTGGTGTTT CTGAGTTCTTAACCTCCTCCTCTG	ACATCAAGCCATCCAACCTG TGAGTTCTTAACCTCCTCCTCTGT	488bp	SNPLex	18-34-38
102	rs12690517:A>G	Donor scoring	12.4	-	(-1)	AAAC**RGT**AGGAATATTTTCC	17/150	ITGA4	NM_000885	GCAGAGTCTCCACCAAGATTCTAT ACTGTGATACTGAGGTCCTCTTCC	TCCAGCAGAGAAGCTAACTGTAGA CCAATACTGCAGTCAAGTTGTACC	491bp	SNPLex	15-25-52
103	**rs11021065:C>A**	Donor scoring/ NN-middle-donor	11.5	0.01185	(+4)	AAAAA**GTA**MGTGTTTCAGTA	9/145	SESN3	NM_144665	AGAAAGTCTTTTTGTGGTCTCTGG GACATTTTCCTTGGGTGATACTTC	TCGGATGGTCTACAATCTCACATA ACGAAGAGCATAAAGAAGTTCAGC	319bp	SNPLex	44-7-41
104	**rs2297889:A>T**	Donor scoring/ NN-middle-donor	11.3	0.00202	(+4)	GAAAG**GTA**WGTCCTTGCATC	5/154	TRIM9	NM_015163	GATGCCCTCAACAGAAGAAAAG GTATGTGCTGTTGAAGTGAAGACC	GTCAACAAGGAGCATGAGCAC GTTGTTGTGGGTACAACATTCCT	357bp	SNPLex	54-70-28
105	**rs3755906:A>T**	Donor scoring/IG/ NN-top-donor	11.3	0.41651	(+4)	ACAAG**GTC**WGTGGCACAGAC	2/110	IGFBP7	NM_001553	CTCTCCTCTTCCTCCTCTTCG CCATGACTACTTTTAACCATGCAG	CTCTTCCTCCTCTTCGGACAC CTCATATTCTCCAGCATCTTCCTT	670bp	SNPLex	29-13-50
106	rs2285666:G>A	Donor scoring	10.8	-	(+4)	ACCAG**GTA**RGCTACTAATTT	3/94	ACE2	NM_021804	TTCTCAGCCTTGTTGCTGTAACT GTCATAGCCATCTACCCCATTTAC	CCGAAGACCTGTTCTATCAAAGTT CCCATAGTCCTCATAATGATTTGC	494bp	SNPLex	66-11-14
107	**rs4681297:G>A**	Donor scoring/ NN-middle-donor	10.8	0.02655	(+4)	GAATG**GTA**RGAGAAACACCC	3/137	PLOD2	NM_182943	GGATTCCATCGATTTATGCAG CTACAGCTCCATTTAAGGTCTGGA	TTCCATCGATTTATGCAGTCAG ATACGGTTGACATATGGAGCATAG	383bp	SNPLex	4-56-32
108	**rs3736185:G>A**	Donor scoring/ NN-top-donor	10.7	0.97809	(+4)	AGGAG**GTC**RGCAGATACAAA	27/90	ITPR2	NM_002223	CAGGTGCAATTACTGGTGTCTAA GTCATTACCATATCCTGGCATTTC	AAAGTCAAGTGAAAGGTGGTGAAG TGTACAACTCTCTCGCTAATTTCG	454bp	SNPLex	9-48-32
109	rs27089:G>A	Donor scoring	10.7	-	(+4)	AAAAG**GTA**RAAAGGGGATTT	3/87	DIMT1L	NM_014473	CTCTTTGGTCCTCCTTGACG GGTCTGAAGTTATTCTTTCCCACT	GAGATGCCGAAGGTCAAGTC GCCAACAGCTGTGTATTAATTGAG	551bp	SNPLex	22-22-48
110	**rs10741752:C>A**	Donor scoring/IG/ NN-middle-donor	10	0.00035	(+3)	GCTGG**GTM**AGGGCAGGGCCA	14/297	IGSF22	NM_173588	CTGAGTTGTGTGTAGTGCTGAATG ACTCGGAATTCATATTCTGTGTCC	GGAAAGTGAAGCCTCTGTATTCAT GTCACGGCTTCTTTAGTCACATC	384bp	SNPLex	4-58-27
111	**rs2272500:C>A**	Donor scoring/ NN-top-donor	10	0.15594	(+3)	ACCAG**GTM**TGAAGTGGAGAA	3/123	SYT1	NM_005639	CATAGTCGCAGTCCTTTTAGTCCT GAATGACAACAGTCAGCTTACCAG	ATAGTCGCAGTCCTTTTAGTCCTG CACTTTCACGTAAGGATCAGATGT	357bp	SNPLex	44-13-35
112	**rs482082:G>A**	Donor scoring/ WD40/ NN-middle-donor	7.9	0.00003	(-2)	GCC**R**G**GT**AAGACTCAAGAGT	16/157	WDR78	NM_024763	GCACTGAAGAAGGTCATATTCACA ATGCTGATTGGTTTGACTTGG	GCAGATTGGGGTGTTATTATATGG TATGCTGATTGGTTTGACTTGG	373bp	SNPLex	19-31-42
113	**rs2057413:G>A**	Donor scoring/ NN-middle-donor	7.8	0.0002	(-2)	ACT**R**T**GT**AAGTAACAGCTGA	10/226	SETDB2	NM_031915	TGTACGCTGTCTAGATGACATTGA GTGTTGAAAAATGGCTGTCTTG	TAGATGACATTGACAGAGGGACAT GAAAAATGGCTGTCTTGGATTC	348bp	SNPLex	49-8-35
114	**rs820329:A>T**	Donor scoring/IG/ NN-middle-donor	5.7	0.00002	(+6)	GAATG**GT**GAG**W**TTCCCCCTG	12/135	MYLK	NM_053028	ATGTAATCTCAAAGGAGTCGAAGC AAAGTGGAAGTCCTCTGACTCTTG	GAGGTCAAGGAAAATCAAACTGTC TCACTCTTCCTGCTACTCTTCTTT T	353bp	SNPLex	5-44-32
115	rs540819:A>T	Donor scoring/ CARD	5.6	-	(+6)	AAAAT**GT**AAG**W**ATTGAGAGT	8/103	CASP5	NM_004347	ACTCTGGGTCAGAGACTCTCCA TGCAAGCTATACTGGTAAATGTGC	GCACTCATCTCTTCACAGTCATCT GCAAGCTATACTGGTAAATGTGCT C	378bp	TaqMan	38-11-39
116	rs1859143:G>A	NN-top-donor	-	0.99827	(+5)	CAGGG**GT**CC**R**TATCCGCTCG	1/475	COL25A1	NM_032518	AAAGAGGTGTCGGTCCTCTG GTTGGTTTTCACACCCAGGTA	GGAGTCGGAAGAGCTGTCTG –	382bp	SNPLex	20-27- 43
117	rs764497:T>A	NN-top-donor (Positive control at donor)	-	0.9901	(**+2**)	GTCAG**GW**AAAAATCCTTTCT	1/136	CCDC149	NM_173463	CGTACTAGAGAAGGGGGCCTTA CAAGCCTTTGCTGAAGTTCTTT	CCTTAGGGAAGTCTCAAAATGCT CGGTCCTGAGAATCTCTCAATAGT	334bp **	SNPLex	4-49-38
118	rs2255089:G>C	NN-top-donor	-	0.97818	(+3)	ACCAA**GTS**AGTAAGATGGGG	3/202	CHI3L2	NM_004000	ACCACCATGGACCAGAAGTC AAGTCAAAGGACAGGAGGTTGAT	CACCATGGACCAGAAGTCTCT CCAGTTTCTCAACTTGATAGCTGT T	590bp	SNPLex	29-18-45

119	rs2276611:G>A	NN-top-donor (Positive control at donor)	-	0.97768	(**+1**)	TGCAG**RT**AAGTGGTATGAGG	1/151	PPIG	NM_004792	CTCCATGCCAGGACTGAGTT GTCACCACCTTGAACCATAAAATC	– CCTTGACAACTCTGTGAAAGAGAC	408bp	SNPLex	4-67-12
120	rs482308:G>A	NN-top-donor (Positive control at donor)	-	0.94543	(**+1**)	CTGCC**RT**GAGTGTGCCCTGC	35/305	ZAN	NM_003386	GACAGTGAATTTGTGAACAGTTGG GCAGTTGGTGTAGCTGCTGTAG	GAAAGATAAGGACATTGACCCAAG CAGTTGGTGTAGCTGCTGTAGG	366bp	SNPLex	30-16-45
121	rs17581728:G>A	NN-top-donor	-	0.9381	(+5)	TGGAG**GT**GC**R**GCATCTTCCA	21/144	UNC13D	NM_199242	GACGGTTGTGGGTGATGTAGT ACCAGTTTCTGGATGTGCTTG	GCTGCAGAAGACGTACAACG CAGCTGCTCCATGTCATTCA	336bp	SNPLex	30-4-39
122	rs366577:T>C	NN-top-donor	-	0.83599	(+3)	CCCAG**GTY**GGGTGAATCTTC	1/64	ENO3	NM_053013	ACTCGGAGCTCCATCCAAA CACTGGGAGTATGAGGTCAGG	– GGGAGTATGAGGTCAGGGTTC	479	SNPLex	16-27-43
123	rs13119659:T>C	NN-top-donor	-	0.78168	(-1)	CATT**YGT**ATCCTTTATGTTT	7/111	USO1	NM_003715	CTGTGCTGTTCTCTATTGTTTCCA GATGCTCTGGAATACAACTCATGT	GCCAGTTATTATGTGGAGGTTTGT CTCTGGAATACAACTCATGTTTGC	513	SNPLex	44-4-43
124	rs11944513:C>T	NN-top-donor	-	0.69791	(+6)	CTGTG**GT**AAA**Y**GAATGCAGC	8/105	ALPK1	NM_025144	CTCCATTGTAGGATATTTGGCACT ATTGCTTCCTTACAGAGCTGACTT	GTAGGATATTTGGCACTTCCTCAG GTGAAGCTCCTGTTTTCCAGTAAC	393bp	SNPLex	47-4-40
125	rs612862:C>T	NN-top-donor	-	0.65913	(-1)	AGAA**YGT**GAGGCTTCTGCGT	8/107	MCOLN1	NM_020533	TAACCTCCAGAGCCTCATCAATA ACAGAAGCAGTAGCCCAGGTAG	– AGCAGTAGCCCAGGTAGATGAC	602bp	SNPLex	44-9-34
126	rs6059183:T>C	NN-top-donor	-	0.52847	(+7)	ATTGA**GT**GAGT**Y**GTCCTAAA	3/160	PLUNC	NM_016583	GAGAGAGAGGAGACCAGGACAG AAGGCTTAGACCTTGATGACAAAC	AGGACAGCTGCTGAGACCTCTA AGGCTTAGACCTTGATGACAAACT	830bp	SNPLex	2-52-37
127	rs12148472:A>G	NN-top-donor	-	0.52378	(+4)	CTAAG**GT**C**R**GTGCCCAACTT	2/32	CTSH	NM_004390	TGAGCGCAAGAGCCAAG GTACCTCGAAGGTAGTTACTTTTGGT	– AGGTAGTTACTTTTGGTGGCTGAG	360bp	SNPLex	2-68-20
128	rs3214041:G>A	NN-top-donor	-	0.4852	(+10)	CACAG**GT**GTGGCTG**R**GAGAA	23/101	PLXNB1	NM_002673	ACCTGACTCTTTGCCTGAGTTC TGTCAGTGAAATACTCAAGCTTCC	CTGACTCTTTGCCTGAGTTCAC CTCTCCGCATACACCTTGTAGTC	358bp	SNPLex	23-32-30
129	rs3213451:A>G	NN-top-donor	-	0.48481	(-3)	CA**R**TG**GT**ATTCTTCATCTTC	23/149	MBTPS2	NM_015884	ACTGTCGTCTACCTGACCGACT AAAATATCCTTAGCTGCTGGACTG	CTGTCGTCTACCTGACCGACTT GAGTGGTGAACAGATCAACAAATG	579bp	SNPLex	37-33-20
130	rs3745779:A>G	NN-top-donor	-	0.47897	(+10)	TGGTG**GT**TAGTGCT**R**TTGTT	1/149	ZNF529	NM_020951	TATTGAGTTAAGCCTGGCGAGT CTGTAGTTCTCCATCATCACATCC	– CTCTGAGCAGAATCCAGATATTCC	376bp	SNPLex	60-3-28
131	rs3752703:C>T	NN-top-donor	-	0.45006	(-4)	G**Y**GGT**GT**ATGTTAGCTTGCT	28/123	PTPRB	NM_002837	AGTGAAGCTCTCCAATGTAGATGA TCTGTTGATGTAGTCCCTGACAGT	GACTACATCAATGCCAGCTACATC ACAGTTCTCACAAACTGGATCAGA	395bp	SNPLex	39-12-39
132	rs2290158:G>A	NN-top-donor	-	0.43618	(+3)	GAGAG**GTR**AAGCCACAAAAT	6/163	SCRN3	NM_024583	CCCAGACATGAGAAACTATGCTAA CCTGTTGATGTTTTTGGTAGAGTG	GCAGCATATTCCTATCTTGACACA GGTGTCTTCTGTCAGGCTTAAAAT	376bp	SNPLex	44-5-42
133	rs11161721:G>T	NN-top-donor	-	0.38675	(+9)	CTTCA**GT**AGGTTT**K**AACTTT	18/54	COL24A1	NM_152890	CCAGGTGACTTTGGAGACAGA CTTTAAGCCTTCTGGTCCTGGT	AGGGAATAAAGGACTACCTGGAAT –	406bp	SNPLex	8-47-32
134	rs3746657:C>T	NN-top-donor	-	0.3748	(+4)	CATGG**GT**G**Y**GTCCACGCAGT	12/124	OSBPL2	NM_014835	GATCCTGTTTCGTATGAATCCTTC GCGAACTCAGTCGTTAGAAAAGTT	GTATGAATCCTTCAAGAAGCAGGA GAAAAGTTGGTTTCTGCTGTGAC	603bp	SNPLex	30-15-46
135	rs217375:A>G	NN-top-donor	-	0.36602	(+14)	TCCAG**GT**CTGCCACAGCC**R**A	6/245	DDX56	NM_019082	CATGTCAGCTACTTTTAACGAGGA GCAGACACATGGTGGAAGTCTAT	GTACAAGCACTCAAGGAGCTGATA AGACACATGGTGGAAGTCTATGC	456bp	SNPLex	29-20-43
136	rs7298440:C>T	NN-top-donor	-	0.34257	(+7)	ATCGG**GT**AATC**Y**GGTTTGGT	12/81	TCTN2	NM_024809	GATCAACCCCTAGAATTGTGAATG TAAGATCAGCGTAATCGGAGTTG	CCCTAGAATTGTGAATGTGGAAG ACTTCTAACAGGCATCCAGAGAGT	351bp	SNPLex	39-9-43
137	rs913742:A>G	NN-top-donor	-	0.34138	(+12)	GCGAG**GT**GACTATTCT**R**CAT	14/54	C14orf101	NM_017799	GACCCAACCTGGAAAAAGAACTAT CACCTAAAATATCAGCTTGCTGTG	GGGTCTTGTGACAAATTAGTTCCT TTTCTGAGAGTACAGACGGAAATG	408bp	SNPLex	57-5-30
138	rs9352:C>T	NN-top-donor	-	0.31219	(-3)	TG**Y**GG**GT**GAGAAGGGCTGTA	14/97	CHAF1A	NM_005483	TCATTTCCGAGAACTCAGTGTATG GACAAAGTGCTCTTTACACAGGAA	ATCTCGCTGAAGAGGAAGTCAG GAGGACACCCTAAGCATTCTACAT	312bp	SNPLex	25-24-42

139	rs1584614:G>C	NN-top-donor	-	0.29227	(-5)	**S**TTAG**GT**AATCGCCGTCCTT	1/46	LSM5	NM_012322	ATGGCGGCTAACGCTACTAC CGGTTGTTTAAATGCACCTGT	GCGGCTAACGCTACTACCA CAAACTTTGTTCCACTGGCTACT	546bp	SNPLex	3-66-23
140	rs10134181:A>G	NN-top-donor	-	0.2896	(+10)	CAACT**GT**GAGTTTT**R**TTTTG	2/123	TC2N	NM_152332	TGGAAGTTTGTGTCTTTTGCTG TGGAAAGTTCTACCTTTCGATCTC	TGTGTCTTTTGCTGGATATTGG GAAAGTTCTACCTTTCGATCTCCA	378bp	SNPLex	22-20-50
141	rs759935:T>C	NN-top-donor	-	0.28446	(+5)	TGCAG**GT**GT**Y**GTATTGTCTT	15/220	GNPTAB	NM_024312	CTGGGAGGCAACTAAAAGATACAT CTGAGCATCTTTATGATTGTGGTC	ACATTTGCAGATTCCCTCAGAT CGATTTCTTCTTCTCCCATGAT	577bp	SNPLex	25-24-42
142	rs2049129:A>G	NN-top-donor	-	0.26322	(+9)	CAATG**GT**ATGTCA**R**TCCTAA	20/107	HPS5	NM_007216	TTTGGCCTATTTAGACAGTCTGGT CATGGCCTTAGCTAACAGAAGTG	ACCAGAGTCTTTAAGGTTGGATTG GTGCTCTTGCTCTGTATGAGATGA	378bp	SNPLex	66-2-23
143	rs2287761:G>C	NN-top-donor	-	0.22643	(+5)	ACAAG**GT**GC**S**GGGAGGACTC	9/135	PPFIA3	NM_003660	GAAGATATGGAGGAGCGGATTAC CTTCTTCATGTTGGCTATCTCCTC	GAACGAGTTAGCTAGCAAGGAGTC GCTCCTTGAGGTGAAGCTGTAA	365bp	SNPLex	4-56-29
144	rs3765115:A>G	NN-top-donor	-	0.21477	(+3)	TTCAG**GTR**CCAAGTATTGGT	28/238	SCAPER	NM_020843	CCTTTTAACAATCGAGTTCAGGAC GTTGTTGTAACAAGCAGCGATAAG	ATGTGTACACTGTGCTTTGCTGT GAAGGGAACAGTACTTTGATCAGC	524bp	SNPLex	18-38-36
145	rs4799570:A>C	NN-top-donor	-	0.21198	(-4)	G**M**TTT**GT**AAGTACTCAATTA	12/297	DSG4	NM_177986	ACATTCTTATGGGTCTCCGTTTAC CAATTCTCCAAGACTGCATCACT	ATGTGGGATGTCAGATCAACAA CCTTCTGGCTGTCTCTGTTTG	374bp	SNPLex	2-75-13
146	rs641018:G>C	NN-top-donor	-	0.18927	(+12)	GAGAG**GT**GCTCACCCT**S**TGG	2/199	FIBP	NM_004214	CAGTGAGCTGGACATCTTCGT TCCAAGGGTCCAGTTTTGG	CCTTATCGACGAGGACGTGTAT CTGATGTCATCCAGGTCTTTCTT	333bp	SNPLex	54-4-32
147	rs314359:A>G	NN-top-donor	-	0.16369	(-5)	**R**CATG**GT**GGGTTGCCCTAAT	10/65	EPHB4	NM_004444	GTCATTGTGGTCGCAGTTCTC CCACATCACAATCCCGTAACT	GTCGCAGTTCTCTGCCTCA CCGTAACTCCAGGCATCACT	749bp	SNPLex	16-31-45
148	rs263042:A>G	NN-top-donor	-	0.16243	(+10)	CTGAG**GT**AACCCAC**R**TCTGC	26/80	YEATS2	NM_018023	GGTTCAGTCATTTTCTACCAGCA TTGGTTTTGAGGGGAGTCAG	GTTCAGTCATTTTCTACCAGCAAG GTGGTGAAATCTCGGATTCTTCT	423bp	SNPLex	25-23-43
159	rs3764913:A>G	NN-top-donor	-	0.16176	(+11)	GCAAG**GT**GTGTAAAA**R**AGGA	5/67	ACADL	NM_001608	CCCAGGTTTTAGTATTCATTCAGG CTGTCTGTAGGTGAGCAACTGTTT	GTATTGGTGCAATAGCAATGACAG GAAGCCTTTATTCTCTTCTCCAAG	347bp	SNPLex	13-38-39
150	rs2075772:A>G	NN-top-donor	-	0.16032	(+13)	GGAAT**GT**ATCCTTCCTC**R**CC	4/75	VASH1	NM_014909	GTACAATCACACAGGGACACAGTT ACATCCTTCTTCCGGTCCTT	AATTAAGAAGAGCAGACCTCTGAC A CAGCTTCACCTTCTTGAGCAC	349bp	SNPLex	34-12-42
151	rs3739085:T>C	NN-top-donor	-	0.15407	(+4)	GAGTG**GT**A**Y**GTTTCCTAGAG	9/142	DPYSL5	NM_020134	ACACCAACACCTCAACCTACCT CTCTCTGGACCAGCTTCTTGTAG	CCAACACCTCAACCTACCTCAT TTCTCATACAGGTTGAAGTCTCCT C	376bp	SNPLex	29-11-51
152	rs1983764:C>T	NN-top-donor	-	0.15323	(+12)	AACAA**GT**GAGCCTACC**Y**GGT	31/114	NIN	NM_020921	CCAGAGATAGCTACTCATCCATCA GCACAAATATGAGTGTACCCTTTG	TCAGAAAGAACAATCTCCTGCTAA C ACTTCCAGAGCTTTCAACAACTG	528pb	SNPLex	9-42-39
153	rs4252120:T>C	NN-top-donor	-	0.14322	(+9)	AATGC**GT**ATGTCT**Y**TGATTT	10/160	PLG	NM_000301	TGAAGGGAACAGGTGAAAACTATC TCTACATCTGGAAGCAGGACAAC	ATAACAGGACACCAGAAAACTTCC GCTTCTGTTCCTGAGCATTTTT	457bp	SNPLex	7-45-39
154	rs1130638:C>T	NN-top-donor	-	0.13898	(-1)	GTAA**YGT**GAGCTCTTGCCCT	4/108	LASP1	NM_006148	AGAACTACAAGGGCTACGAGAAGA TCCTTGTAGCCACCATAGGACT	GAACTACAAGGGCTACGAGAAGAA TTGTAGCCACCATAGGACTGG	435bp	SNPLex	38-14-38
155	rs7495739:A>G	NN-top-donor	-	0.12707	(+9)	GCAAG**GT**GGACAC**R**GTTATA	5/183	MPI	NM_002435	ACAACCGCATCTCACAGAAGA GGTAAGCAGGTTCAGGAAGTAGAT	CTTCCTCTTCAAAGTGCTCTCAGT TAAGCAGGTTCAGGAAGTAGATGG	497bp	SNPLex	25-26-40
156	rs2074189:C>A	NN-top-donor	-	0.1224	(+9)	CCCAG**GT**GGGATT**M**TTAGAC	18/79	OSBPL7	NM_145798	GCCATGCAGAGTCTGAGAACTT AGCTCTGCTGTCAGCTCATTC	CATGCAGAGTCTGAGAACTTCG GTGCCACTTCCCAAAGAGTC	351bp	SNPLex	18-23-49
157	rs743128:T>G	NN-bottom-donor	-	0	(+14)	ATAAG**GT**GAGTGCTGGGG**K**C	6/88	ACTN1	NM_001102	ACAAGATCTCCAACGTCAACAAG GGAAGTCCTCCAGCTTCTGTT	TGGATTTCATAGCCAGCAAAG GTAGTCTTCCATAAGCTGCTCGTT	548bp	SNPLex	36-10-44
158	rs1866846:A>G	NN-bottom-donor	-	0	(+14)	GGAAG**GT**ATGTACCTGCT**R**G	3/87	KIAA1429	NM_015496	CTCATATAGATGTGGTTCGTTTTCC	TATCAATGAAGTCCGAGTCATACC	488bp	SNPLex	4-56-32

										TGGGCTCAAAGTAATCTTCTCTATG	CAGGATCATCTTCATCATCATCAG			
159	rs2114724:A>G	NN-bottom-donor	-	0	(+14)	CAGAG**GT**AAGGATGCGGC**R**G	22/188	DNMT1	NM_001379	TAAATGAATGGTGGATCACTGG CGAAGAAAGTATCGAAGATCTGGT	AGATCTACATCAGCAAGATTGTGG –	384bp	SNPLex	34-17-41
160	rs2276825:A>G	NN-bottom-donor	-	0	(+14)	TCACT**GT**AAGTGTTCCCT**R**T	3/96	TMEM110	NM_198563	CAGAGAACCAAAGCATGAAAGAC CCACTGAAGTATTAGGAGGACGAT	– AGTATTAGGAGGACGATGAAGACG	357bp	SNPLex	6-53-32
161	rs2277439:G>A	NN-bottom-donor	-	0	(+14)	AAAAG**GT**AAGTCCACATC**R**A	2/168	TNFSF11	NM_003701	GCCAGCAGAGACTACACCAAGTA AGGACAGACTCACTTTATGGGAAC	GCCCTGTTCTTCTATTTCAGAGC GTGGCATTAATAGTGAGATGAGCA	326bp	SNPLex	59-3-29
162	rs2279090:A>G	NN-bottom-donor	-	0	(+14)	ACCAG**GT**ACGTGGCCGCC**R**C	13/196	ADAM12	NM_003474	CCAGAAGTGTGGGAACAGATTT CGTGAACCCCAAAGACACTAATA	AACAGATTTGTGGAAGAAGGAGAG CATGTCATCGCCCAAGTACA	594bp	SNPLex	3-76-11
163	rs2303180:G>A	NN-bottom-donor	-	0	(+14)	CGCTG**GT**GAGTGGCTGTG**R**T	5/210	MARCH2	NM_001005415	CAGTATGTGGCACAGGTGACTT GAACTGTTGAAGTGGAAGTGTTGA	AGTATGTGGCACAGGTGACTTC GGATCTTCAGGCGAACTTTCT	552bp	SNPLex	41-6-42
164	rs4806711:A>G	NN-bottom-donor	-	0	(+14)	GAGAG**GT**GAGTGTGATGG**R**G	1/341	PRPF31	NM_015629	AAGGCCTTGCTTTCTTGTCTAAC TACTTATCCCGGATGAACTTATGG	– CTTTGGCTTGCTTGCTGATATAC	564bp	SNPLex	59-5--25
165	rs7300317:A>G	NN-bottom-donor	-	0	(+14)	AGACG**GT**GAGGACCATGG**R**G	4/96	KRT7	NM_005556	ACAGCTGCTGAGAATGAGTTTGT GGTTCATCTCTGAAATCTCATTCC	ATGAGTTTGTGGTGCTGAAGAAG TGAAATCTCATTCCGGGTATTC	353bp	SNPLex	12-24-54
166	rs11068780:G>A	NN-bottom-donor	-	0	(+14)	AGTCG**GT**GAGTCTCCAGC**R**T	5/101	WSB2	NM_018639	ATTGGAAGTCCAGCTGTGAAAC GCACACAGATCTCAGTGAGCTAAT	GACACTGCATCGTCAAACTGAT GTCCCACATAATCACATTGGTATC	665bp	SNPLex	61-3-26
167	rs12141283:A>G	NN-bottom-donor	-	0	(+14)	CCCAG**GT**GAGTGAAGGGG**R**A	4/106	NFASC	NM_001005387	AATTTGGGACGCTGGAGTTTA GTTACAACTGTAGTCGGTCTGCAT	AACAGAGCCTCCTCTGGTGTT CAGCATCACGTTGGAGAAGTATAG	735bp	SNPLex	35-19-34
168	rs31725:C>T	NN-bottom-donor	-	0	(+15)	GCCAG**GT**GAGGTCCGGGTA**Y**	17/922	PLEKHG2	NM_022835	GACCTCAATCACTGAAGAAATCCT ACAAAGGTATGGCAGCTTGAAC	– CAAAGGTATGGCAGCTTGAAC	1099bp	SNPLex	24-19-47
169	rs421587:A>G	NN-bottom-donor	-	0	(+15)	TCCAG**GT**AAGTCAACTCAA**R**	28/54	COL1A2	NM_000089	CAACATTGGATTCCCTGGAC GGTTACCAATTTCACCTCTGAGAC	– GCTTCCAATAGGACCAGTAGGAC	355bp	SNPLex	39-8-45
170	rs931479:A>G	NN-bottom-donor	-	0	(+15)	TACAG**GT**GAGGGGTTAGGC**R**	7/221	KRT4	NM_002272	CCTGAAGAACACCAAGAGTGAAAT AAGCCACTACTCAGGCCAAAC	AGATCGAGAACATCAAGAAGCAG ACTTCCTAATCCTCCGCTGAT	332bp	SNPLex	67-5-18
171	rs2042792:C>G	NN-bottom-donor	-	0	(+15)	CCCAG**GT**CAGTGCACAGGA**S**	11/144	SPAG16	NM_024532	CACCAGAAGGTCCTACTCAGAAAG CATGTGACCATAAAGTGACTGCTC	GATATGCAACCAAATCCAAACC TGTTCTTGCATCCCATATAGACAG	567bp	SNPLex	26-14-52
172	rs2070615:A>G	NN-bottom-donor	-	0	(+15)	GCCAG**GT**GAGAGTTGGGCC**R**	4/116	CACNB3	NM_000725	GGTTCAGCCGACTCCTACAC CGCTCAATGATGGTCCTCTT	GCCCATCTCTGGACTCAGAC TTGTTGAGCACAGATCGCTTT	622bp	SNPLex	17-27-48
173	rs2240999:C>G	NN-bottom-donor	-	0	(+15)	ACCAG**GT**GAGGAGGGAGTG**S**	7/110	SPHKAP	NM_016532	GACCTCATTATCTGGTTTGGAGAC TTGAAGAGTAGCTGACCATCATGT	– GTACGTCATGTGGCTGCTGTAG	366bp	SNPLex	3-67-20
174	rs2241920:G>A	NN-bottom-donor	-	0	(+15)	TGGGG**GT**AAGTCCTCTCCA**R**	1/210	MS4A14	NM_032597	ATGTTTGCTCACTCTTTCCCTTAC ACTGGAAAACTCTTCTTGAAGCAC	CCATAGAATCATGGAGTCAACATC GGCAGGTTGTTCACTATTTTGAGT	529bp	SNPLex	34-11-45
175	rs2248619:A>G	NN-bottom-donor	-	0	(+15)	AGGAG**GT**GAGGGGCTCTCA**R**	4/96	KRT83	NM_002282	TTTGCTGGCTACATCGAGACT GTCATCATACTGTGCCTTGATCTC	CTGGCTACATCGAGACTCTGC –	386bp	SNPLex	13-26-43
176	rs2287483:C>T	NN-bottom-donor	-	0	(+15)	CTCCG**GT**AAGGGGAAGTGT**Y**	1/129	GPR160	NM_014373	TGCAGTCCGGAGACGAA TCTATACAAGCTGTCAGGAAAACTG	– GGCAGATGTGGTATTTAGTGAACC	576bp	SNPLex	2-73-16
177	rs3740522:G>A	NN-bottom-donor	-	0	(+15)	AAAAT**GT**GAGTGTTGACGG**R**	5/227	ARHGAP19	NM_032900	ACCTGAGATTTTCACTGAGTTGGT ACATGAAGCTATGAGGTCAAGGTC	GCCCAGATATACCAACTGATTGA AAATAGTGCAATCTCGCACACTC	602bp	SNPLex	33-15-42
178	rs4957318:G>A	NN-bottom-donor	-	0	(+15)	AGCAG**GT**ATGTGCAAGTAC**R**	9/142	FYB	NM_001465	GAACAAGACAGTGAAGGAGAAACA ATCAGAGGTATCCACATCATCGTA	CTAACAGGCCCTATTCAAGTCATC CACATCATCGTAAACTTCATCTCC	585bp	SNPLex	48-9-34
179	rs6510801:G>A	NN-bottom-donor	-	0	(+15)	CCCAG**GT**ACGGGAGCCAGC**R**	2/360	TMIGD2	NM_144615	GGTGATGGGCCAGGAAT	–	607bp	SNPLex	32-12-45

										GAAGGAGGTTGAATAAATGCTCTG	GGTATAGGACGTTGCTGTAGAATG			
180	rs9644114:A>G	NN-bottom-donor	-	0	(+15)	CTAGG**GT**AAGTACCGGTCA**R**	5/150	BNIP3L	NM_004331	AGAAGATGGGCAGATCATGTTT ATGTAAAAACAGGGTGGTAGGTTG	GCAGATCATGTTTGATGTGGAA CATGCTTACAATGGTCTCAAGTTC	446bp	SNPLex	47-8-35
181	rs17688121:G>A	NN-bottom-donor	-	0	(+15)	CACTG**GT**AAGTTTCCACTG**R**	4/69	EVC2	NM_147127	CTGTCACTTTAAGACTGCAGTGGA CATGTTTCCCTTCAGACACTGATA	GTCTTCATCCCACTCTCAACTTCT GTTACGTTTTCTTCTGCTGTTATG G	545bp	SNPLex	73-2-16
182	**rs232518:T>C**	Normalized ESE scoring/IG/ NN-middle-acceptor	1.508	0.00508	(+5)	AATCGATGTAAAC**AG**AGCC**Y** GGACGGCCGTATCGAAGTCA AAGGG	9/151	NCAM2	NM_004540	GAGAATGGTCAAGTCACACTCGTA TCCAAAGTCATTGTCAGATGTAGG	GGAGCCTATTCCAGAAATCACTT GGTCGTGTTTTTAGCAGGTAAGAC	364bp	SNPLex	14-34-44
183	rs1348689:C>T	Normalized ESE scoring	1.508	-	(+15)	TTATTTTCTGTTT**AG**AATCC TGGCTATGC**Y**GGACGGCAGG AACTC	38/195	DNAH5	NM_001369	TGCACTGACAGGCTTGTAATAACT AATAGCTGGATGACCTTCAGTTTC	CGTATTGATCTACCAGTTCTCTCG TCCGAAGAACTGACAGAATGTTAC	385bp	SNPLex	38-11-43
184	**rs4964287:C>T**	Normalized ESE scoring/ NN-middle-acceptor	1.508	0.008	(+5)	TCAATGTTGTTGT**AG**GTCT**Y** GGAGGAACATGTGTGAATGT GGGTT	5/120	TXNRD1	NM_003330	AGGGCAGACTTCAAAAGCTACTAA CACTGCTGATGCAGTATTCTTTG	TTCCCAAGTCCTATGACTATGACC CCCATAAGCATTCTCATAGACGA	377bp	SNPLex	44-11-37
185	rs2274980:C>T	Normalized ESE scoring	1.508	-	(+29)	CCTTGTGGGTTTC**AG**GTTCT CTTAGTGCTCGATGTGACAA CTC**Y**G	3/136	LAMC2	NM_018891	GGAACTTCACAGACAAACTGGTAA –	CACAGACAAACTGGTAATGGATTC CATCTTGATGAAAGGTAGAGGTGA	514bp	SNPLex	5-66-21
186	rs3813795:A>G	Normalized ESE scoring	1.318	-	(+24)	CCACCTCCCGCAC**AG**CTTCG TGCTGCCTGATGACAGCC**R**G GCCAG	14/206	SYTL1	NM_032872	GAATCTGAATCCGGTTTTCAAC GTCCAGAGAGAGAGGCTTGGT	AACATCTTTCTGGGCGAAGTT AGGTTGGTTCTGAGGGGTAGA	592bp	SNPLex	10-44-33
187	rs2278211:G>A	Normalized ESE scoring	1.318	-	(-13)	ACGAAGCATGCACAGAC**R**GA GATCATTGAG**GT**GGGTGCTG GTGGT	5/119	INPP4A	NM_001566	CAAGAAGCACATCATCACCAAT ATTTCAGCAAAGTGTCCTTTCG	AAGAAGCACATCATCACCAATG TGATGTTACCTACACGGTCACTCT	526bp	SNPLex	65-4-23
188	rs7603997:G>A	Normalized ESE scoring	1.318	-	(+14)	TAACATTTTCTTT**AG**GACTC TGGCTGAC**R**TTGATGGTGAT GGACA	10/138	ITSN2	NM_006277	CAGCCTTTACCCATTCCTTATTCT CTTCCTTTTCTTTCTGGGCTTTAC	GCCTTTACCCATTCCTTATTCTTC ATTTCTTCTGAGGCTCCTCTTCTT	548bp	SNPLex	43-4-45
189	rs3177168:G>A	Normalized ESE scoring	1.318	-	(+15)	ATCTTTTACGTAC**AG**CGGAA AGAACACCC**R**GAAATGAAAG GCCAC	2/41	MRPS35	NM_021821	AAGGACTCTGCGTGCATTCT CTGATGAAACATAATCAGTGCTGTC	CATTCTCCACTGCCGTCTACT ATGCTTCTCACATTTCTCGTCAC	377bp	SNPLex	4-68-20
190	rs16890979:G>A	Normalized ESE scoring	1.318	-	(+30)	TGCCTCTGTTTGC**AG**CCTTC CAAACGTTCTTGGGTAAAGC AGAC**R**	4/188	SLC2A9	NM_020041	TCATCATGGGCATAGATGGAG CCCTGTACTCAAGGTGACGTATG	ATACCTGTTTGGAGTGATTGTGG CCTGTACTCAAGGTGACGTATGG	387bp	SNPLex	62-4-26
191	rs2289043:T>C	Normalized ESE scoring	1.29	-	(+26)	GGCATGTCATTAT**AG**GAAAT TTTACATCTTGAGAGACAGA **Y**GGGA	13/150	UNC5C	NM_003728	ATGGATGACTCTCAGACACTTTTG GTTTCTTTGAGATCCACTCCAAAC	ACTGGAAAATACTGCTCAAGAACC GAGATCCACTCCAAACATGGTAA	465bp	SNPLex	42-15-34
192	rs1566088:C>T	Normalized ESE scoring	1.29	-	(+21)	ACTATTCCATTTT**AG**GATGA TGACTTGGAAACAGA**Y**GTGA ACAAG	2/117	AGBL1	NM_152336	GATGGGCCAGTGCTATAATTTG CCATATAAAGATAGGGATCGTGGA	GGGCCAGTGCTATAATTTGG GGTACAGGTTCACTCCAGGATTT	371bp	SNPLex	34-17-41
193	rs761422:T>C	Normalized ESE scoring	1.29	-	(-17)	CAGTGTGTGCCCA**Y**GAGGAG CTCCTCCGAG**GT**AGGAAGGC CACCT	8/74	MFAP2	NM_002403	CGCCTCTACTCCATACACAGG GTAGGAAATCCAAGCAGACCAG	TGCAAACAGTGTCTCAACGAG CTGCAGTCCACTAACTTTTTCAGA	331bp	SNPLex	21-15-56

194	rs2227255:C>T	Normalized ESE scoring	1.29	-	(+24)	TGGCCCGCCCCCC**AG**GTCTA CGTGTTGAAGCGTCCTCA**Y**G TGGAT	5/93	CTDSP1	NM_021198	AGACAAGATCTGCGTGGTCAT AGCTCTGTGTCACTCATGTTGTC	CTTCATCATCCCTGTGGAGATT TGTGTCACTCATGTTGTCAAACC	357bp	SNPLex	13-35-44
195	rs767050:C>T	Normalized ESE scoring	1.29	-	(+24)	TTCCTGTGCTTCC**AG**ACCTC AAGCATCTGCAAGACAGC**Y**G TGCAC	14/134	CRISPLD2	NM_031476	CTGCAAAGACGAACCTTCCTAC CACTGAACATCAGTCAAAGGAAGT	CAAAGACGAACCTTCCTACTGG AAGGAAGTTTCCTGACTCTCCATA	350bp	SNPLex	33-15-44
196	rs2279103:C>T	Normalized ESE scoring	1.29	-	(-12)	GGTCCCGAGAATCTCAGA**Y**G AGAAAGAAAG**GT**GGGTAACC TCCTT	3/167	CTDP1	NM_048368	GGATAAGTCAAAGTCCAGGTCAGT AGGACTTCCTGGAGAAGATCG	CCATTTTCTAAAACGGGAAACC TAAGTCAAAGTCCAGGTCAGTGC	526bp	SNPLex	55-4-30
197	rs1465567:T>C	Normalized ESE scoring	1.29	-	(-28)	GC**Y**GGCCCAGTGACATCATC CGGACCCTCT**GT**GAGTACCA GGGTC	6/167	EGFLAM	NM_152403	TATCGTGTGAGCATAGCAGCTTA CCTAGAAATGGTCTTTGGGTTAGA	CGCCCCTATTCAGTACTATTCTGT TTCTTATTCCCTCCTTTGGTAGC	357bp	SNPLex	59-6-27
198	rs11065772:T>C	Normalized ESE scoring	1.29	-	(-28)	GA**Y**GTTAATCTGCCGGCCGC CCAGCTACAG**GT**GAGAAAAT GGGCT	10/171	ACACB	NM_001093	CAGTATGGGAATGCTGTGTCTCT TGAAGTAACCCCACACGTTCTT	GTATGGGAATGCTGTGTCTCTGT GGGTTTCAAAAGAAATGGGAGT	386bp	SNPLex	64-5-23
199	rs5749104:T>C	Normalized ESE scoring	1.29	-	(+19)	CCCTCCCCCTGAC**AG**GTGGC AGTTCTCATCTGA**Y**GGTGCG GACAT	11/170	SEC14L3	NM_174975	GTTTGCTGAAACTCATCAGTCCT TCACTAACGTCACAGAGTCAGGAG	AGGTGAAGACTCAGTACGAGCAC AGGGGTGAGCTCCTTATCATATTT	380bp	SNPLex	14-27-51
200	rs3741475:C>T	Normalized ESE scoring	1.29	-	(+23)	CTTGTACCTCCAC**AG**GTGTC ATCAGTAACTGGACAGA**Y**GA GCTCC	22/170	NOS1	NM_000620	ACCAGATGGTGAAAGTGGAACT GCACTTCATCAGGGTACATGTCT	AGATGGTGAAAGTGGAACTGCT ATCTGGATAGATGGGAACTCCTC	298bp	SNPLex	4-60-28
201	rs17612126:C>A	Normalized ESE scoring	1.281	-	(+25)	GCGGGTCTTCTCC**AG**GACAT CCGGCCACAGATCTGCCCA**M** GGAGG	14/173	IGHMBP2	NM_002180	AGGTTCATCACTGTGAGCAAGA GAGCTCACTCAGCTTCTTATCCA	GCAGCAGAAACTTCCAGAAAAG CAGCTTCTTATCCAGCCTCCT	361bp	SNPLex	9-46-37
202	rs3828323:G>A	Normalized ESE scoring	1.191	-	(+15)	CTGTTCTTGTTCT**AG**ATACT TCTGGACAC**R**GTGTAAATAC ATCTG	24/239	PLA2R1	NM_007366	TTCCAAGTCACAATACCACTGAAG GGAATAGATGTTTCTGAGCACAAC	GGTATTTTGAAGACTGTGGAAAGG GTGTTCAGATTGTCTTGTTTCAGG	485bp	SNPLex	23-25-42
203	rs2071624:C>T	Normalized ESE scoring	1.188	-	(+16)	ATCTCTCTCTGAC**AG**TTGCT GGGATACAAA**Y**GACCACAGT GTGCC	9/70	VIPR2	NM_003382	GTCAAGGACGACGTTCTCTACTC GCACAGCTCAAACAGTATCTGGTA	AAGCTGAGCCTGGTCTTCCT TATCTGGTATTTGGAGGAGATGCT	465bp	SNPLex	57-4-31
204	rs10950854:C>T	Normalized ESE scoring	1.188	-	(+13)	TCTTAATGTTTGT**AG**GCCAC CGTCAAA**Y**GAAAGGATAATA CTTCA	4/190	DNAH11	NM_003777	GTTGTCTCTTGGACATGTATCTGC ACTCCAGATCAGACAGATGGTATG	GACATGTATCTGCTTTCCTTGATG CCAGATCAGACAGATGGTATGAAA	647bp	SNPLex	26-20-46
205	rs12386051:G>A	Normalized ESE scoring/IG	1.115	-	(+8)	TCTCATCTCCTCT**AG**TTCCC AC**R**GTCCCTCCGGGCAATGT GCACG	12/116	SDK2	NM_019064	ATTCTCAAGGGTTACATCATCAGG GTGTTGGTTCGATTGTACTCCTC	GGGGTACCAGTTTAAGAACATCAC ATGTGTCCAGGATCTCACTGAAG	518bp	SNPLex	60-5-36
206	rs2295773:T>C	Normalized ESE scoring/ WD40	1.078	-	(-19)	GAAGTCCAGCA**Y**ATGAGACA GGCTGACAAG**GT**TTGAAGGG CTTGG	9/162	SEC31B	NM_015490	CAACACATTCTGTCTTCTGCTCA TTCAGATTCTGTGGTGACTTGACT	GGATCTCAGGAAGAATGAACCTAT TTCAGATTCTGTGGTGACTTGACT	582bp	TaqMan	60-4-28
207	rs1878061:G>A	Normalized ESE scoring/IG	0.981	-	(-26)	CACC**R**GGGAGGTGTTGACCC AAAATTCAGG**GT**AAGCACGC AGGGC	3/109	CD300E	NM_181449	AGAAGAGAAGGTGGAGAGGAATG CATCCAGTCTGAAAGGTTGACTC	AAATTCAGACAGTGTGGGTCCT ATCCAGTCTGAAAGGTTGACTCC	383bp	SNPLex	33-16-42

208	rs12593397:G>A	Normalized ESE scoring	0.977	-	(-11)	GGTGCCCAGAGCCTTCACC**R**CAAGCACAAG**GT**AATAGCCCTCTTC	24/187	SPTBN5	NM_016642	CCCTACAGAGCTCGGAAACACCAGTAAATGGCTAGTTCCTCCT	CATCCTGGAAGAGACCCAGACAGTAAATGGCTAGTTCCTCCTGT	567bp	SNPLex	36-12-29
209	rs7669741:C>T	Normalized ESE scoring	0.974	-	(-15)	CGAAGCCCTAGTGCA**Y**CTGCTCCACAGATG**GT**ATGAAGACTTTTT	18/110	KIAA0922	NM_015196	GAGCCTTTCTCTGGATCAATCTACTAACTTTTCCATAGTCAGCAGGTG	CAATCTACCTGGAATGTGGATTCTAACTTTTCCATAGTCAGCAGGTGT	516bp	SNPLex	49-11-32
210	rs3762672:G>T	Normalized ESE scoring	0.948	-	(-30)	**K**CTAGTAAACCAAGTGACATGTCAGTACAG**GT**GAGGCTAAAACCT	36/75	DNAJC13	NM_015268	ACTTCAGATGACCTCCTTTTCTCAAGCAAGTTTTCTAGCAACATAGGG	CAGAGCTAGCTTTCCATACTGTCATAGGTGTGTCTGTAGCCAGAAATC	347bp	SNPLex	25-21-46
211	rs842823:A>G	Normalized ESE scoring	0.93	-	(+11)	TTTCCTTCTTTAT**AG**GGCCTATTCC**R**CTAGAAGCAAGATGGCTGA	3/61	LOC26010	NM_015535	TTCAGGAACATTGCTGTGGATGATCTTCTTCACGGAACTATCCAC	TCAGGAACATTGCTGTGGATTGATAAGGGCTGGTTTTTCGTTAG	432bp	SNPLex	41-11-40
212	rs1001420:C>T	Normalized ESE scoring	0.893	-	(-22)	GATTTCTA**Y**AGGCAATTCTGTGTGAAACTG**GT**GAGTGTCCAATAG	40/132	FLJ40243	NM_173489	TATCATCAGAGGCCTGTATCACCCTAGTAACTCCACATATTGGCTGGT	CTAACAGGAAGAAGGTGGAAGATTAGTAACTCCACATATTGGCTGGTC	380bp	SNPLex	9-44-39
213	rs7303113:A>G	Normalized ESE scoring	0.892	-	(+19)	TCTGTTTTATTCC**AG**ACATGCTGGTGTCTACAC**R**GCAGAAGAAGT	5/168	C12orf41	NM_017822	GGTCTCAGGAACCTCTGTCTTGGTCTGGTCATTGGAAGAGACTGAT	GCTGAGCTCATATGCTAAGACAGAGCTTTAAGTTCTCTCGTTCTTTGG	416bp	SNPLex	8-37-46
214	rs10787428:A>G	Normalized ESE scoring	0.892	-	(-22)	TGTGACTG**R**AAATGTGCTGAACAGCAGTAG**GT**AAGGGCGGGGCAA	6/114	GPAM	NM_020918	ATCAGAATACAGTGTTGGTCGATGAGCTTTAACCATCTCAAGTTGACC	AATACAGTGTTGGTCGATGTAAGCACTTCTGCAATTGCCTCTTGTACT	375bp	SNPLex	40-13-39
215	rs9288938:G>T	Normalized ESE scoring	0.892	-	(-20)	CAGATTACAA**K**TTCTAAACAGATTAAACAG**GT**AAGAAAACCTTAA	18/150	SLC9A10	NM_183061	TGCCTGGAACATATTCGAGTTAGAGCAATTTCTGGGTGATCATACTC	CCTGGAACATATTCGAGTTAGCAGCAATTTCTGGGTGATCATACTCT	385bp	SNPLex	8-48-36
216	rs9438:G>C	Normalized ESE scoring	0.888	-	(+30)	ATGTTTTTATTTT**AG**GTATGTTCCAGAACAAAAAGAACACAGAT**S**	10/141	DHX36	NM_020865	CTGTACAACAGGAATCATCCTTCACCTCCATCTATCACATAAACGACA	GACCTTCTCAATTTTCGATCTGACTGTAGGCATCAGTGAATGTAAAGG	537bp	SNPLex	20-33-38
217	rs989902:T>G	Normalized ESE scoring	0.86	-	(-20)	TGCAGCAGGA**K**ACTCCTGTGGTCCAGGTAC**GT**GAACCAGATGAAT	39/138	PTPN13	NM_080684	TTCCTACAGTGTGGGGTCTTGGTGTATTTACCAGAGGGAAGCACT	GATAAAGGGATCACCAAACTTGACTGTTCTCTCTATTGGCAACTCATC	400bp	SNPLex	24-15-53
218	rs1898883:C>G	Normalized ESE scoring	0.854	-	(+20)	CTCCTCTCTTCCT**AG**CACCCAGACCAAGGCTGTG**S**CCCCTGAGGC	2/330	DISP2	NM_033510	GAAGGGGAGCAACGGCCACAGAAGAAGTTCTCTTGCCTTC	–CTCTCCTGGGCACTGCCT	713bp	SNPLex	47-7-38
219	rs2228173:A>G	Normalized ESE scoring/ WD40	0.669	-	(-19)	ACCATGGCAGA**R**TTGAATGCCATCATCGGG**GT**ACGTGGCCTACCA	6/75	TLE1	NM_005077	CCAGCCCTTCAAGTTCACTATCGAAGCTGGACTGCTAGAAGCAT	CTTCAAGTTCACTATCCCGGAGTTCCTCATTAGACACATCCACAACT	717bp	TaqMan	71-1-18
220	rs2188383:C>G	Normalized ESE scoring/ LRR	0.646	-	(+19)	TTCTCTAAACCAT**AG**TTTCATAAGTACAAGCCA**S**ACCTTTGATTT	2/121	MOSPD2	NM_001018113	GAGGTTTGGAGCAGATGGATACTTCCAGTTCTGAAATCTGACACAC	GGTTTGGAGCAGATGGATACCGGGTGTTTTTGTAGGCTCTTTATC	376bp	SNPLex	28-39-24
221	rs736795:C>T	Normalized ESE scoring /DEATH	0.58	-	(-15)	CCACCTGGGGCTTTG**Y**GGCATGAAGATCCG**GT**AGGAAGAGGGGTG	8/114	UNC5CL	NM_173561	GAGAATGAGGACTGTTCAGCACTACTCTTAGGCGTAGGAGAACAACC	GCTTGGAGACCAAGTATATGGAAAGTCCCACTCAGGTAGTTCTGGAT	361bp	SNPLex	4-61-27

222	rs3763840:C>T	Normalized ESE scoring/ WD40	0.559	-	(-8)	CCAAGGTGACAGATGAGACC TC**Y**GGCTGCT**GT**AAGTTGCC TCATA	12/70	PRPF19	NM_014502	CGACACCAACAAGATCCTCA GTAAAGTGAAGAATCTCCGTCCAT	AAGTTCTGAACAAATCCTGGCTAC ATCTTGATCTGAGAGTCCATGGTT	375bp	SNPLex	39-9-43
223	rs974144:C>T	Normalized ESE scoring/ WD40	0.479	-	(-16)	ATCCAAATCTTCTC**Y**TGTCA GTAAGTAAAG**GT**AAGTGAAG CAAAT	5/82	EED	NM_152991	CATGGACCTATGATAGCAATACGA AACCTCATGTACCAAATGTCACAC	CTGTAGCTGGATCTAGAGGCATAA ATCGCCCAAGAATAGTCACATTAG	576bp	SNPLex	12-35-45

- Δ_S : Absolute difference between the two allelic splice scores as calculated for each SNP using Alex's online splice score tool; Δ_{ESE}: absolute normalized ESE score differences as calculated using ESEfinder; Δ_N: absolute signal difference as calculated from the signals emitted from the neural network.

- *: The extracted sequence for ss SNPs represented here as 15 nt of intronic and 5 nt of exonic sequence for the canonical ss; extracted sequence for ESE-SNPs is represented as 30 nt exonic and 15 nt at nearby ss.

- SNP ambiguity codes, reference AG or GT dinucleotides and the overlapped SNPs ID in both rounds of investigation are represented in bold.

- **: To prove the skipping of exon 1 in the respective genotypes, another primer that spanned exon 2 (GGAAGCTGGAGAGTAAGAAGGAA) was used. The expected product size of using the later primer with the nested reverse primer (CGGTCCTGAGAATCTCTCAATAGT), in the round 2 RT-PCR, was 223 bp.

- ¶: To check for the presence of exon 4, another F primer spanning this exon (GAAAGAAGTTAACAAGGCACTATCAA) was used with the nested-reverse primer (CAGTGAGTACGAATAAAGCGATCA) resulting in an expected amplicon size of 450 bp.

- HUGO HGNC-approved gene symbols (http://www.genenames.org/) were used here in this Table;

- The mutation and splice effect nomenclature appears here to follow the format indicated in the HGVS (see the website http://www.hgvs.org/mutnomen/).

8.2. Detection of minor splice forms by direct sequencing of RT-PCR products

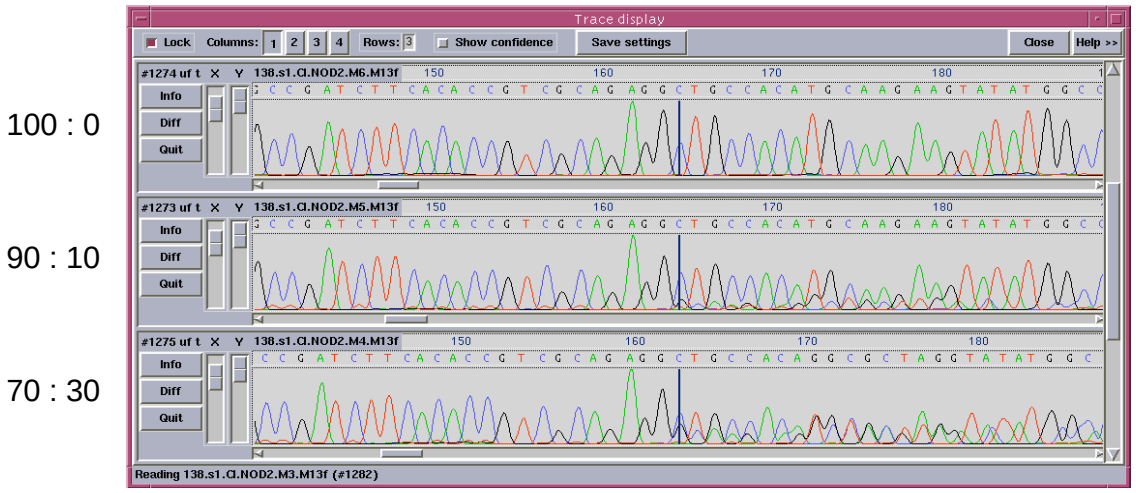

Figure 8.1 Evaluation of direct PCR-product sequencing as a qualitative detection method for alternative splicing: a control experiment.
Here, the sensitivity for a qualitative detection of AS was previously tested in splice variants from the *CARD15* locus. Diluted PCR products from two splice variants at the *CARD15* locus were mixed in different stoichiometric ratios, re-amplified by PCR and directly sequenced. It is evident here that the presence of a second, alternatively spliced transcript can be detected down to the 90:10% level. This means that minor splice forms with a frequency down to 10% of the total transcripts can be clearly detected. Thus, direct sequencing may be a robust screening tool even in the presence of NMD. *The data was obtained from previous in-house experience (see also comment in additional data file in (Hiller et al., 2006b)).*

8.3. Extraction of SNPs at ESE sites within a 30-nucleotides window of exon ends

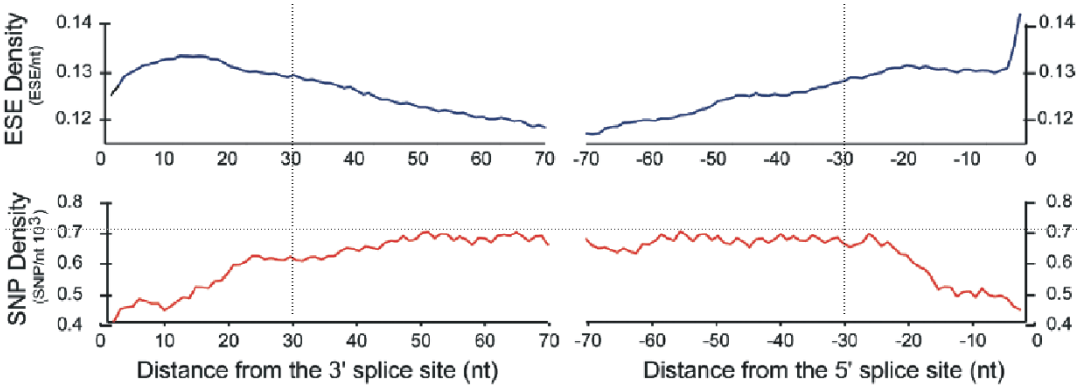

Figure 8.2 Density of Predicted ESEs and SNPs along Human Exons.
This figure illustrates that SNP density is approximately 20 –30% lower near both ss of human exons than in the interior of exons, and reaches a plateau at about 25-30 bases from the ss. The distribution of predicted ESE hexamers along exons had roughly an inverse relationship to the SNP density, with the highest density of ESEs observed near the ss junctions and a lower density in the interior of exons. Selective pressure is likely to be higher on ESEs located near splice junctions relative to ESEs in the interior of exons, which could explain the trend in ESE density shown in the above Figure. As a consequence of the increased density of ESEs near ss, SNPs/mutations that occur in exons near ss should have a higher likelihood of disrupting ESEs and therefore be more likely to be eliminated by purifying selection. Thus, SNPs located inside these 30-bp windows of the nearest exon-intron boundary from either ss might be more consistent to be considered for further analysis for allele-dependent splicing. *This figure modified from (Fairbrother et al., 2004a).*

8.4. Abbreviations and Lists of Figures and Tables

Table 8.2 List of abbreviations, units, symbols, and acronyms used in thesis text

Abbreviation	Description
∞	Forever
°C	degree Celsius
µ	micro; 10^{-6}
µl	microlitre(s)
µg	microgram
µM	micromolar (µmol/l)
aa	amino acid
AS	alternative splicing
BLAST	Basic Local Alignment Search Tool
bp	base pairs
cDNA	complementary DNA
cds	coding sequence
CE	capillary electrophoresis
CEPH	Centre d'Etude du Polymorphisme Humain (positive control cell lines)
chr	chromosome
ChromFA	Human genomic sequence files in FASTA format extracted from the UCSC home page.
conc	concentration
cSNP	coding SNP (located in coding regions)
ddNTP	dideoxynucleotide triphosphate
DDW	double distilled water
DNA(s)	deoxyribonucleic acid(s)
dNTP	2'-deoxynucleoside-5'-triphosphate
dsDNA	double stranded DNA
e.g.	exempli gratia
EDTA	ethylenediaminetetraacetic acid
ESE	exonic-splicing enhancers
ESS	Exonic Splicing Silencer
EST	expressed sequence tag
Exo I	exonuclease I
F	forward
FACS	Fluorescence-activated cell sorting
g	gram
gDNA	Genomic DNA
GFP	Green fluorescent protein
h	hour
HeLa	epithelial-like malignant cells derived from the cervix of Henrietta Lacks
hnRNP	heterogeneous nuclear ribonucleoproteins
i.e.	id est
ICMB	Institute of clinical molecular biology (Kiel, Germany)
Ig	immunoglobulin
ISE	Intronic Splicing Enhancers
ISS	Intronic Splicing Silencers
IVS	Intervening sequence
kb	kilobase
kDa	kiloDalton
l	liter
LIMS	laboratory information management system at ICMB, Kiel, Germany

LRR	leucine-rich repeat
m	milli; 10^{-3}
M	molar (mol/l)
Mb	mega base
MCS	Multiple cloning site
MDA	multiple displacement assay
mg	milligram
MGB	minor groove binder
$MgCl_2$	magnesium chloride
min	minute
ml	milliliter
mM	millimolar (mmol/l)
mRNA	messenger RNA
MTP	microtiter plate
NCBI	National Center for Biotechnology Information
ng	nanogram
NMD	nonsense mediated decay (of mRNA)
nmol	nanomole(s)
NN	neural network
nt	nucleotide(s)
OLA	oligonucleotide ligation reaction (in SNPlex™ genotyping procedure)
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	Phosphate-Buffered Salines
PCR	polymerase chain reaction
PGM2L1	Phosphoglucomutase-2-like 1
pH	potentia hydrogenii
pmol	picomol
Pol II	Polymerase II
Poly A	poly-adenosine tail
PPT	polypyrimidine tract
pre-mRNA	precursor- mRNA
PTB	polypyrimidinetract-binding protein
PTCs	Premature stop codons
pUC	Produced at the University of California (bacterial origin of replication)
R	reverse
RefSeq	Reference sequence
RFP	Red fluorescent protein
RNA	ribonucleic acid
RNABPs	RNA-binding proteins
rpm	rotations per minute
RS domain	arginine/serine dipeptides domain
RT	room temperature (roughly 21-23°C)
s	second
SAP	shrimp alkaline phosphatase
SNP(s)	single nucleotide polymorphism(s)
snRNAs/snRNPs	small nuclear ribonucleoprotein particles
SR proteins	serine/arginine-rich proteins
ss	splice site(s)
SV40 Ori	Simian virus 40 (mammalian origin of replication)
TAE	tris acetate EDTA
Taq	*Thermophilus aquaticus*
TaqMan®	commercial name for sequence variation detection assay (5'→3' exonuclease activity)
TBE	tris borate EDTA

TE	Tris-hydroxymethyl aminomethane buffer
Te-MO	Tecan multipipetting option
Tm	melting temperature (of primer)
Tris	tris-(hydroxymethyl)-aminomethane
UCSC	University of California at Santa Cruz (The Genome Browser homepage)
UV	ultraviolet
WGA	whole genome amplification
WT	wild type
DNA bases	
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
IUPAC- SNP ambiguity code	
A	A
B	C/G/T
C	C
D	A/G/T
G	G
H	A/C/T
K	G/T
M	A/C
N	G/A/T/C
R	A/G
S	C/G
T/U	T
V	A/C/G
W	A/T
Y	C/T

LIST OF FIGURES

LIST OF TABLES

9 REFERENCES

Aartsma-Rus, A. and G. J. van Ommen (2007). "Antisense-mediated exon skipping: a versatile tool with therapeutic and research applications." Rna 13(10): 1609-24.

Ahn, S. J., J. Costa, et al. (1996). "PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR." Nucleic Acids Res 24(13): 2623-5.

Andoh, A., Y. Deguchi, et al. (2006). "Immunohistochemical study of chymase-positive mast cells in inflammatory bowel disease." Oncol Rep 16(1): 103-7.

Ars, E., E. Serra, et al. (2000). "Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1." Hum Mol Genet 9(2): 237-47.

Ast, G. (2005). "The alternative genome." Sci Am 292(4): 40-7.

Bai, Y., D. Lee, et al. (1999). "Control of 3' splice site choice in vivo by ASF/SF2 and hnRNP A1." Nucleic Acids Res 27(4): 1126-34.

Baralle, D. and M. Baralle (2005). "Splicing in action: assessing disease causing sequence changes." J Med Genet 42(10): 737-48.

Baralle, M., D. Baralle, et al. (2003). "Identification of a mutation that perturbs NF1 agene splicing using genomic DNA samples and a minigene assay." J Med Genet 40(3): 220-2.

Ben-Dov, C., B. Hartmann, et al. (2008). "Genome-wide analysis of alternative pre-mRNA splicing." J Biol Chem 283(3): 1229-33.

Berget, S. M., C. Moore, et al. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A 74(8): 3171-5.

Berthold, V. and K. Geider (1976). "Interaction of DNA with DNA-binding proteins. The characterization of protein HD from Escherichia coli and its nucleic acid complexes." Eur J Biochem 71(2): 443-9.

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

Blencowe, B. J. (2006). "Alternative splicing: new insights from global analyses." Cell 126(1): 37-47.

Boue, S., I. Letunic, et al. (2003). "Alternative splicing and evolution." Bioessays 25(11): 1031-4.

Brett, D., H. Pospisil, et al. (2002). "Alternative splicing and genome complexity." Nat Genet 30(1): 29-30.

Brewis, I. A., R. A. Van Gestel, et al. (2005). "The spermatozoon at fertilisation: current understanding and future research directions." Hum Fertil (Camb) 8(4): 241-51.

Briese, M., B. Esmaeili, et al. (2005). "Is spinal muscular atrophy the result of defects in motor neuron processes?" Bioessays 27(9): 946-57.

Buratti, E. and F. E. Baralle (2004). "Influence of RNA secondary structure on the pre-mRNA splicing process." Mol Cell Biol 24(24): 10505-14.

Buratti, E., M. Chivers, et al. (2007). "Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization." Nucleic Acids Res 35(13): 4250-63.

Caceres, J. F. and A. R. Kornblihtt (2002). "Alternative splicing: multiple control mechanisms and involvement in human disease." Trends Genet 18(4): 186-93.

Caceres, J. F., S. Stamm, et al. (1994). "Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors." Science 265(5179): 1706-9.

Caffrey, T. M., C. Joachim, et al. (2007). "Haplotype-specific expression of the N-terminal exons 2 and 3 at the human MAPT locus." Neurobiol Aging.

Caffrey, T. M. and R. Wade-Martins (2007). "Functional MAPT haplotypes: bridging the gap between genotype and neuropathology." Neurobiol Dis 27(1): 1-10.

Cajiao, I., A. Zhang, et al. (2004). "Bystander gene activation by a locus control region." Embo J 23(19): 3854-63.

Cartegni, L., S. L. Chew, et al. (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." Nat Rev Genet 3(4): 285-98.

Cartegni, L. and A. R. Krainer (2002). "Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1." Nat Genet 30(4): 377-84.

Cartegni, L. and A. R. Krainer (2003). "Correction of disease-associated exon skipping by synthetic exon-specific activators." Nat Struct Biol 10(2): 120-5.

Cartegni, L., J. Wang, et al. (2003). "ESEfinder: A web resource to identify exonic splicing enhancers." Nucleic Acids Res 31(13): 3568-71.

Cavalli-Sforza, L. L. (2005). "The Human Genome Diversity Project: past, present and future." Nat Rev Genet 6(4): 333-40.

Chakarova, C. F., M. M. Hims, et al. (2002). "Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa." Hum Mol Genet 11(1): 87-92.

Chen, F. C., E. J. Vallender, et al. (2001). "Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences." J Hered 92(6): 481-9.
Chen, S., K. Anderson, et al. (2000). "Evidence for a linear search in bimolecular 3' splice site AG selection." Proc Natl Acad Sci U S A 97(2): 593-8.
Chen, Y. I., R. E. Moore, et al. (2007). "Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors." Nucleic Acids Res 35(12): 3928-44.
Chern, T. M., E. van Nimwegen, et al. (2006). "A simple physical model predicts small exon length variations." PLoS Genet 2(4): e45.
Chomczynski, P. and N. Sacchi (1987). "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction." Anal Biochem 162(1): 156-9.
Chomczynski, P. and N. Sacchi (2006). "The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on." Nat Protoc 1(2): 581-5.
Chow, L. T., R. E. Gelinas, et al. (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell 12(1): 1-8.
Chua, K. and R. Reed (2001). "An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing." Mol Cell Biol 21(5): 1509-14.
Clark, F. and T. A. Thanaraj (2002). "Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human." Hum Mol Genet 11(4): 451-64.
Colapietro, P., C. Gervasini, et al. (2003). "NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient." Hum Genet 113(6): 551-4.
Colgan, D. F. and J. L. Manley (1997). "Mechanism and regulation of mRNA polyadenylation." Genes Dev 11(21): 2755-66.
Cowles, C. R., J. N. Hirschhorn, et al. (2002). "Detection of regulatory variation in mouse genes." Nat Genet 32(3): 432-7.
Das, D., T. A. Clark, et al. (2007). "A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing." Nucleic Acids Res 35(14): 4845-57.
Dausset, J., H. Cann, et al. (1990). "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome." Genomics 6(3): 575-7.
de Almeida, S. F. and M. Carmo-Fonseca (2008). "The CTD role in cotranscriptional RNA processing and surveillance." FEBS Lett 582(14): 1971-6.
De la Vega, F. M., K. D. Lazaruk, et al. (2005). "Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System." Mutat Res 573(1-2): 111-35.
Dean, F. B., S. Hosono, et al. (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc Natl Acad Sci U S A 99(8): 5261-6.
Dean, F. B., J. R. Nelson, et al. (2001). "Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification." Genome Res 11(6): 1095-9.
Don, R. H., P. T. Cox, et al. (1991). "'Touchdown' PCR to circumvent spurious priming during gene amplification." Nucleic Acids Res 19(14): 4008.
Dou, Y., K. L. Fox-Walsh, et al. (2006). "Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site." Rna 12(12): 2047-56.
Duerr, R. H., M. M. Barmada, et al. (2000). "High-density genome scan in Crohn disease shows confirmed linkage to chromosome 14q11-12." Am J Hum Genet 66(6): 1857-62.
Eden, E. and S. Brunak (2004). "Analysis and recognition of 5' UTR intron splice sites in human pre-mRNA." Nucleic Acids Res 32(3): 1131-42.
ElSharawy, A., B. Hundrieser, et al. (2008). "Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing." Hum Mutat: In press.
ElSharawy, A., C. Manaster, et al. (2006). "SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs." Hum Mutat 27(11): 1129-34.
Esteban, J. A., M. Salas, et al. (1993). "Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization." J Biol Chem 268(4): 2719-26.
Fackenthal, J. D., L. Cartegni, et al. (2002). "BRCA2 T2722R is a deleterious allele that causes exon skipping." Am J Hum Genet 71(3): 625-31.
Fairbrother, W. G., D. Holste, et al. (2004a). "Single nucleotide polymorphism-based validation of exonic splicing enhancers." PLoS Biol 2(9): E268.
Fairbrother, W. G., R. F. Yeh, et al. (2002). "Predictive identification of exonic splicing enhancers in human genes." Science 297(5583): 1007-13.
Fairbrother, W. G., G. W. Yeo, et al. (2004b). "RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons." Nucleic Acids Res 32(Web Server issue): W187-90.
Farrell, R. E. (1999). RNA Methodologies. A laboratory guide for isolation and characterization. RNA Methodologies. A laboratory guide for isolation and characterization. San Diego, Academic Press.
Faustino, N. A. and T. A. Cooper (2003). "Pre-mRNA splicing and human disease." Genes Dev 17(4): 419-37.

Field, L. L., V. Bonnevie-Nielsen, et al. (2005). "OAS1 splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes." Diabetes 54(5): 1588-91.
Figueroa, J. D., N. Malats, et al. (2007). "Evaluation of genetic variation in the double-strand break repair pathway and bladder cancer risk." Carcinogenesis 28(8): 1788-93.
Gabut, M., S. Chaudhry, et al. (2008). "SnapShot: The splicing regulatory machinery." Cell 133(1): 192 e1.
Gao, K., A. Masuda, et al. (2008). "Human branch point consensus sequence is yUnAy." Nucleic Acids Res 36(7): 2257-67.
Garcia-Blanco, M. A. (2006). "Alternative splicing: therapeutic target and tool." Prog Mol Subcell Biol 44: 47-64.
Garcia-Blanco, M. A., A. P. Baraniak, et al. (2004). "Alternative splicing in disease and therapy." Nat Biotechnol 22(5): 535-46.
Gasper, J. and W. J. Swanson (2006). "Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution." Am J Hum Genet 79(5): 820-30.
Gilbert, W. (1978). "Why genes in pieces?" Nature 271(5645): 501.
Graveley, B. R. (2001). "Alternative splicing: increasing diversity in the proteomic world." Trends Genet 17(2): 100-7.
Graveley, B. R. (2008). "The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population." Trends Genet 24(1): 5-7.
Green, R. E., B. P. Lewis, et al. (2003). "Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes." Bioinformatics 19 Suppl 1: i118-21.
Grosso, A. R., A. Q. Gomes, et al. (2008). "Tissue-specific splicing factor gene expression signatures." Nucleic Acids Res 36(15): 4823-32.
Hampe, J., A. Wollstein, et al. (2001). "An integrated system for high throughput TaqMan based SNP genotyping." Bioinformatics 17(7): 654-5.
Heath, E. M., D. P. O'Brien, et al. (1999). "Optimization of an automated DNA purification protocol for neonatal screening." Arch Pathol Lab Med 123(12): 1154-60.
Hertel, K. J. (2008). "Combinatorial control of exon recognition." J Biol Chem 283(3): 1211-5.
Hiller, M., K. Huse, et al. (2004). "Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity." Nat Genet 36(12): 1255-7.
Hiller, M., K. Huse, et al. (2006a). "Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing." Am J Hum Genet 78(2): 291-302.
Hiller, M., K. Huse, et al. (2006b). "Phylogenetically widespread alternative splicing at unusual GYNGYN donors." Genome Biol 7(7): R65.
Hiller, M., S. Nikolajewa, et al. (2007a). "TassDB: a database of alternative tandem splice sites." Nucleic Acids Res 35(Database issue): D188-92.
Hiller, M. and M. Platzer (2008). "Widespread and subtle: alternative splicing at short-distance tandem sites." Trends Genet 24(5): 246-55.
Hiller, M., Z. Zhang, et al. (2007b). "Pre-mRNA Secondary Structures Influence Exon Recognition." PLoS Genet 3(11): e204.
Ho, L. S. and J. C. Rajapakse (2003). "Splice site detection with a higher-order markov model implemented on a neural network." Genome Inform 14: 64-72.
Horiuchi, T. and T. Aigaki (2006). "Alternative trans-splicing: a novel mode of pre-mRNA processing." Biol Cell 98(2): 135-40.
Houdayer, C., C. Dehainault, et al. (2008). "Evaluation of in silico splice tools for decision-making in molecular diagnosis." Hum Mutat 29(7): 975-982.
House, A. E. and K. W. Lynch (2008). "Regulation of alternative splicing: more than just the ABCs." J Biol Chem 283(3): 1217-21.
Hull, J., S. Campino, et al. (2007). "Identification of common genetic variation that modulates alternative splicing." PLoS Genet 3(6): e99.
Ibrahim el, C., T. D. Schaal, et al. (2005). "Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers." Proc Natl Acad Sci U S A 102(14): 5002-7.
Jakobsson, M., S. W. Scholz, et al. (2008). "Genotype, haplotype and copy-number variation in worldwide human populations." Nature 451(7181): 998-1003.
Johnson, J. M., J. Castle, et al. (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." Science 302(5653): 2141-4.
Jurica, M. S. and M. J. Moore (2003). "Pre-mRNA splicing: awash in a sea of proteins." Mol Cell 12(1): 5-14.
Kanopka, A., O. Muhlemann, et al. (1996). "Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA." Nature 381(6582): 535-8.
Kapranov, P., S. E. Cawley, et al. (2002). "Large-scale transcriptional activity in chromosomes 21 and 22." Science 296(5569): 916-9.

Kawase, T., Y. Akatsuka, et al. (2007). "Alternative splicing due to an intronic SNP in HMSD generates a novel minor histocompatibility antigen." Blood 110(3): 1055-63.
Ke, X., M. S. Taylor, et al. (2008). "Singleton SNPs in the human genome and implications for genome-wide association studies." Eur J Hum Genet 16(4): 506-515.
Kim, H., R. Klein, et al. (2004). "Estimating rates of alternative splicing in mammals and invertebrates." Nat Genet 36(9): 915-6; author reply 916-7.
Kishore, S. and S. Stamm (2006). "The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C." Science 311(5758): 230-2.
Koren, E., G. Lev-Maor, et al. (2007). "The emergence of alternative 3' and 5' splice site exons from constitutive exons." PLoS Comput Biol 3(5): e95.
Kornblihtt, A. R. (2006). "Chromatin, transcript elongation and alternative splicing." Nat Struct Mol Biol 13(1): 5-7.
Kral, T., H. Clusmann, et al. (2002). "Preoperative evaluation for epilepsy surgery (Bonn Algorithm)." Zentralbl Neurochir 63(3): 106-10.
Kralovicova, J., M. B. Christensen, et al. (2005). "Biased exon/intron distribution of cryptic and de novo 3' splice sites." Nucleic Acids Res 33(15): 4882-98.
Kralovicova, J. and I. Vorechovsky (2007). "Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition." Nucleic Acids Res 35(19): 6399-413.
Krawczak, M., S. Nikolaus, et al. (2006). "PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships." Community Genet 9(1): 55-61.
Krawczak, M., J. Reiss, et al. (1992). "The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences." Hum Genet 90(1-2): 41-54.
Krawczak, M., N. S. Thomas, et al. (2007). "Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing." Hum Mutat 28(2): 150-8.
Kruit, A., J. C. Grutters, et al. (2006). "Chymase gene (CMA1) polymorphisms in Dutch and Japanese sarcoidosis patients." Respiration 73(5): 623-33.
Kutyavin, I. V., I. A. Afonina, et al. (2000). "3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures." Nucleic Acids Res 28(2): 655-61.
Kwan, T., D. Benovoy, et al. (2008). "Genome-wide analysis of transcript isoform variation in humans." Nat Genet 40(2): 225-31.
Kwan, T., D. Benovoy, et al. (2007). "Heritability of alternative splicing in the human genome." Genome Res 17(8): 1210-8.
Lage, J. M., J. H. Leamon, et al. (2003). "Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH." Genome Res 13(2): 294-307.
Lallena, M. J., K. J. Chalmers, et al. (2002). "Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45." Cell 109(3): 285-96.
Lee, P. H. and H. Shatkay (2008). "F-SNP: computationally predicted functional SNPs for disease association studies." Nucleic Acids Res 36(Database issue): D820-4.
Lev-Maor, G., R. Sorek, et al. (2003). "The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons." Science 300(5623): 1288-91.
Lewandowska, M. A., C. Stuani, et al. (2005). "Functional studies on the ATM intronic splicing processing element." Nucleic Acids Res 33(13): 4007-15.
Lewis, B. P., R. E. Green, et al. (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans." Proc Natl Acad Sci U S A 100(1): 189-92.
Ligtenberg, M. J., A. M. Gennissen, et al. (1991). "A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA." Nucleic Acids Res 19(2): 297-301.
Lim, S. R. and K. J. Hertel (2004). "Commitment to splice site pairing coincides with A complex formation." Mol Cell 15(3): 477-83.
Liu, H. X., L. Cartegni, et al. (2001). "A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes." Nat Genet 27(1): 55-8.
Lizardi, P. M., X. Huang, et al. (1998). "Mutation detection and single-molecule counting using isothermal rolling-circle amplification." Nat Genet 19(3): 225-32.
Lopez-Bigas, N., B. Audit, et al. (2005). "Are splicing mutations the most frequent cause of hereditary disease?" FEBS Lett 579(9): 1900-3.
Lovmar, L. and A. C. Syvanen (2006). "Multiple displacement amplification to create a long-lasting source of DNA for genetic studies." Hum Mutat 27(7): 603-14.
Lynch, K. W. and A. Weiss (2001). "A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer." J Biol Chem 276(26): 24341-7.
Malousi, A., S. Kouidou, et al. (2007). "Detecting over-represented motifs in alternatively spliced exons using Gibbs sampling." Conf Proc IEEE Eng Med Biol Soc 2007: 139-42.

Manaster, C., R. Valentonyte, et al. (2005a). "SGCaller: a program to call and review genotypes measured by sequencing." Biotechniques 38(4): 544, 546.
Manaster, C., W. Zheng, et al. (2005b). "InSNP: a tool for automated detection and visualization of SNPs and InDels." Hum Mutat 26(1): 11-9.
Maniatis, T. and B. Tasic (2002). "Alternative pre-mRNA splicing and proteome expansion in metazoans." Nature 418(6894): 236-43.
Maquat, L. E. (2005). "Nonsense-mediated mRNA decay in mammals." J Cell Sci 118(Pt 9): 1773-6.
Marden, J. H. (2008). "Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms." Heredity 100(2): 111-20.
Martinez-Contreras, R., J. F. Fisette, et al. (2006). "Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing." PLoS Biol 4(2): e21.
Matlin, A. J., F. Clark, et al. (2005). "Understanding alternative splicing: towards a cellular code." Nat Rev Mol Cell Biol 6(5): 386-98.
McKie, A. B., J. C. McHale, et al. (2001). "Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13)." Hum Mol Genet 10(15): 1555-62.
McPherson, M. J. and S. G. Moller (2000). PCR. Oxford, UK BIOS Scientific Publishers Limited.
McVety, S., L. Li, et al. (2006). "Disruption of an exon splicing enhancer in exon 3 of MLH1 is the cause of HNPCC in a Quebec family." J Med Genet 43(2): 153-6.
Min, J., S. Okada, et al. (1999). "Synip: a novel insulin-regulated syntaxin 4-binding protein mediating GLUT4 translocation in adipocytes." Mol Cell 3(6): 751-60.
Moore, M. J. and P. A. Silver (2008). "Global analysis of mRNA splicing." Rna 14(2): 197-203.
Mordes, D., X. Luo, et al. (2006). "Pre-mRNA splicing and retinitis pigmentosa." Mol Vis 12: 1259-71.
Nakai, K. and H. Sakamoto (1994). "Construction of a novel database containing aberrant splicing mutations of mammalian genes." Gene 141(2): 171-7.
Nalla, V. K. and P. K. Rogan (2005). "Automated splicing mutation analysis by information theory." Hum Mutat 25(4): 334-42.
Nelson, J. R., Y. C. Cai, et al. (2002). "TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing." Biotechniques Suppl: 44-7.
Nembaware, V., B. Lupindo, et al. (2008). "Genome-wide survey of allele-specific splicing in humans." BMC Genomics 9: 265.
Nembaware, V., K. H. Wolfe, et al. (2004). "Allele-specific transcript isoforms in human." FEBS Lett 577(1-2): 233-8.
Newman, E. A., S. J. Muh, et al. (2006). "Identification of RNA-binding proteins that regulate FGFR2 splicing through the use of sensitive and specific dual color fluorescence minigene assays." Rna 12(6): 1129-41.
Nielsen, K. B., S. Sorensen, et al. (2007). "Seemingly neutral polymorphic variants may confer immunity to splicing-inactivating mutations: a synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer." Am J Hum Genet 80(3): 416-32.
Niksic, M., M. Romano, et al. (1999). "Functional analysis of cis-acting elements regulating the alternative splicing of human CFTR exon 9." Hum Mol Genet 8(13): 2339-49.
Nissim-Rafinia, M. and B. Kerem (2002). "Splicing regulation as a potential genetic modifier." Trends Genet 18(3): 123-7.
Ogura, H., H. Agata, et al. (1997). "A study of learning splice sites of DNA sequence by neural networks." Comput Biol Med 27(1): 67-75.
Okada, S., K. Ohshima, et al. (2007). "Synip phosphorylation is required for insulin-stimulated Glut4 translocation." Biochem Biophys Res Commun 356(1): 102-6.
Orengo, J. P., D. Bundman, et al. (2006). "A bichromatic fluorescent reporter for cell-based screens of alternative splicing." Nucleic Acids Res 34(22): e148.
Pagani, F. and F. E. Baralle (2004). "Genomic variants in exons and introns: identifying the splicing spoilers." Nat Rev Genet 5(5): 389-96.
Pagani, F., E. Buratti, et al. (2002). "A new type of mutation causes a splicing defect in ATM." Nat Genet 30(4): 426-9.
Pagani, F., E. Buratti, et al. (2000). "Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element." J Biol Chem 275(28): 21041-7.
Pagani, F., M. Raponi, et al. (2005). "Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution." Proc Natl Acad Sci U S A 102(18): 6368-72.
Pagani, F., C. Stuani, et al. (2003). "New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12." Hum Mol Genet 12(10): 1111-20.
Pagenstecher, C., M. Wehner, et al. (2006). "Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants." Hum Genet 119(1-2): 9-22.
Pajares, M. J., T. Ezponda, et al. (2007). "Alternative splicing: an emerging topic in molecular and clinical oncology." Lancet Oncol 8(4): 349-57.

Papik, K., B. Molnar, et al. (1998). "Application of neural networks in medicine - a review." Med Sci Monit 4(3): MT538-546.

Parmley, J. L., J. V. Chamary, et al. (2006). "Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers." Mol Biol Evol 23(2): 301-9.

Pfarr, N., D. Prawitt, et al. (2005). "Linking C5 deficiency to an exonic splicing enhancer mutation." J Immunol 174(7): 4172-7.

Phillips, C. (2007). "Online resources for SNP analysis: a review and route map." Mol Biotechnol 35(1): 65-97.

Pollard, A. J., A. R. Krainer, et al. (2002). "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 277(18): 15241-51.

Pradet-Balade, B., J. P. Medema, et al. (2002). "An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK-APRIL fusion protein." Embo J 21(21): 5711-20.

Pratt, W. S., I. Islam, et al. (1996). "Two additional polymorphisms within the hypervariable MUC1 gene: association of alleles either side of the VNTR region." Ann Hum Genet 60(Pt 1): 21-28.

Reddy, A. S. (2007). "Alternative splicing of pre-messenger RNAs in plants in the genomic era." Annu Rev Plant Biol 58: 267-94.

Reed, R. (1996). "Initial splice-site recognition and pairing during pre-mRNA splicing." Curr Opin Genet Dev 6(2): 215-20.

Reese, M. G., F. H. Eeckman, et al. (1997). "Improved splice site detection in Genie." J Comput Biol 4(3): 311-23.

Rekha, T. S. and C. K. Mitra (2006). "Comparative analysis of splice site regions by information content." Genomics Proteomics Bioinformatics 4(4): 230-7.

Rengarajan, K., S. M. Cristol, et al. (2002). "Quantifying DNA concentrations using fluorometry: a comparison of fluorophores." Mol Vis 8: 416-21.

Reumers, J., L. Conde, et al. (2008). "Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases." Nucleic Acids Res 36(Database issue): D825-9.

Roca, X., A. J. Olson, et al. (2008). "Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics." Genome Res 18(1): 77-87.

Roca, X., R. Sachidanandam, et al. (2003). "Intrinsic differences between authentic and cryptic 5' splice sites." Nucleic Acids Res 31(21): 6321-33.

Roca, X., R. Sachidanandam, et al. (2005). "Determinants of the inherent strength of human 5' splice sites." Rna 11(5): 683-98.

Romano, M., R. Marcucci, et al. (2002). "Regulation of 3' splice site selection in the 844ins68 polymorphism of the cystathionine Beta -synthase gene." J Biol Chem 277(46): 43821-9.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol 132: 365-86.

Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature 409(6822): 928-33.

Sauer, S., D. Lechner, et al. (2000). "A novel procedure for efficient genotyping of single nucleotide polymorphisms." Nucleic Acids Res 28(5): E13.

Schellenberg, M. J., D. B. Ritchie, et al. (2008). "Pre-mRNA splicing: a complex picture in higher definition." Trends Biochem Sci 33(6): 243-6.

Schreiber, S., P. Rosenstiel, et al. (2005). "Genetics of Crohn disease, an archetypal inflammatory barrier disease." Nat Rev Genet 6(5): 376-88.

Senapathy, P., M. B. Shapiro, et al. (1990). "Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project." Methods Enzymol 183: 252-78.

Shapiro, M. B. and P. Senapathy (1987). "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression." Nucleic Acids Res 15(17): 7155-74.

Sharma, S., U. M. Rajan, et al. (2005). "A novel (TG)n(GA)m repeat polymorphism 254 bp downstream of the mast cell chymase (CMA1) gene is associated with atopic asthma and total serum IgE levels." J Hum Genet 50(6): 276-82.

Shen, L. X., J. P. Basilion, et al. (1999). "Single-nucleotide polymorphisms can cause different structural folds of mRNA." Proc Natl Acad Sci U S A 96(14): 7871-6.

Singer, V. L., L. J. Jones, et al. (1997). "Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation." Anal Biochem 249(2): 228-38.

Skotheim, R. I. and M. Nees (2007). "Alternative splicing in cancer: noise, functional, or systematic?" Int J Biochem Cell Biol 39(7-8): 1432-49.

Smith, C. W., T. T. Chu, et al. (1993). "Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns." Mol Cell Biol 13(8): 4939-52.

Smith, P. J., C. Zhang, et al. (2006). "An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers." Hum Mol Genet 15(16): 2490-508.
Solis, A. S., N. Shariat, et al. (2008). "Splicing fidelity, enhancers, and disease." Front Biosci 13: 1926-42.
Soller, M. (2006). "Pre-messenger RNA processing and its regulation: a genomic perspective." Cell Mol Life Sci 63(7-8): 796-819.
Sorek, R. and G. Ast (2003). "Intronic sequences flanking alternatively spliced exons are conserved between human and mouse." Genome Res 13(7): 1631-7.
Sorek, R., G. Lev-Maor, et al. (2004). "Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons." Mol Cell 14(2): 221-31.
Spena, S., M. L. Tenchini, et al. (2006). "Cryptic splice site usage in exon 7 of the human fibrinogen Bbeta-chain gene is regulated by a naturally silent SF2/ASF binding site within this exon." Rna 12(6): 948-58.
Staley, J. P. and C. Guthrie (1998). "Mechanical devices of the spliceosome: motors, clocks, springs, and things." Cell 92(3): 315-26.
Stamm, S. (2008). "Regulation of alternative splicing by reversible protein phosphorylation." J Biol Chem 283(3): 1223-7.
Stamm, S., S. Ben-Ari, et al. (2005). "Function of alternative splicing." Gene 344: 1-20.
Stamm, S., J. Zhu, et al. (2000). "An alternative-exon database and its statistical analysis." DNA Cell Biol 19(12): 739-56.
Stenson, P. D., E. V. Ball, et al. (2003). "Human Gene Mutation Database (HGMD): 2003 update." Hum Mutat 21(6): 577-81.
Stephens, R. M. and T. D. Schneider (1992). "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites." J Mol Biol 228(4): 1124-36.
Stoilov, P., C. H. Lin, et al. (2008). "A high-throughput screening strategy identifies cardiotonic steroids as alternative splicing modulators." Proc Natl Acad Sci U S A 105(32): 11218-23.
Strachan, T. and A. P. Read (2004). Human Molecular Genetics3. London and New York, GS Garland science, Taylor and Francis Group
Sun, H. and L. A. Chasin (2000). "Multiple splicing defects in an intronic false exon." Mol Cell Biol 20(17): 6414-25.
Tabuchi, K. and T. C. Sudhof (2002). "Structure and evolution of neurexin genes: insight into the mechanism of alternative splicing." Genomics 79(6): 849-59.
Talavera, D., A. Hospital, et al. (2007). "A procedure for identifying homologous alternative splicing events." BMC Bioinformatics 8: 260.
Tanphaichitr, N., E. Carmona, et al. (2007). "New insights into sperm-zona pellucida interaction: involvement of sperm lipid rafts." Front Biosci 12: 1748-66.
Teraoka, S. N., M. Telatar, et al. (1999). "Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences." Am J Hum Genet 64(6): 1617-31.
Teuber, M., W. A. Koch, et al. (2005). "Improving quality control and workflow management in high-throughput single-nucleotide polymorphism genotyping environments." Journal of the Association for Laboratory Automation 10(1): 43-47.
Teufel, A., M. Krupp, et al. (2006). "Current bioinformatics tools in genomic biomedical research (Review)." Int J Mol Med 17(6): 967-73.
Thanaraj, T. A. and S. Stamm (2003). "Prediction and statistical analysis of alternatively spliced exons." Prog Mol Subcell Biol 31: 1-31.
Tobler, A. R., S. Short, et al. (2005). "The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping." J Biomol Tech 16(4): 398-406.
Tomlinson, I., E. Webb, et al. (2007). "A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21." Nat Genet 39(8): 984-8.
Ueda, H., J. M. Howson, et al. (2003). "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease." Nature 423(6939): 506-11.
Ule, J., G. Stefani, et al. (2006). "An RNA map predicting Nova-dependent splicing regulation." Nature 444(7119): 580-6.
Vacquier, V. D. (1998). "Evolution of gamete recognition proteins." Science 281(5385): 1995-8.
Valentonyte, R., J. Hampe, et al. (2005). "Sarcoidosis is associated with a truncating splice site mutation in BTNL2." Nat Genet 37(4): 357-64.
van Deutekom, J. C., A. A. Janson, et al. (2007). "Local dystrophin restoration with antisense oligonucleotide PRO051." N Engl J Med 357(26): 2677-86.
van Gestel, R. A., I. A. Brewis, et al. (2007). "Multiple proteins present in purified porcine sperm apical plasma membranes interact with the zona pellucida of the oocyte." Mol Hum Reprod 13(7): 445-54.
Vithana, E. N., L. Abu-Safieh, et al. (2001). "A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11)." Mol Cell 8(2): 375-81.

Waetzig, G. H., D. Seegert, et al. (2002). "p38 mitogen-activated protein kinase is activated and linked to TNF-alpha signaling in inflammatory bowel disease." J Immunol 168(10): 5342-51.
Wang, G. S. and T. A. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." Nat Rev Genet 8(10): 749-61.
Wang, J., Q. S. Gao, et al. (2004a). "Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors." J Neurochem 88(5): 1078-90.
Wang, Z., M. E. Rolish, et al. (2004b). "Systematic identification and analysis of exonic splicing silencers." Cell 119(6): 831-45.
Wang, Z., X. Xiao, et al. (2006). "General and specific functions of exonic splicing silencers in splicing control." Mol Cell 23(1): 61-70.
Wangkumhang, P., K. Chaichoompu, et al. (2007). "WASP: a Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations." BMC Genomics 8: 275.
Wasserman, P. D. (1989). Neural computing: theory and practice. . New York, Van Nostrand Reinhold Co.
Watson, R. T. and J. E. Pessin (2007). "GLUT4 translocation: the last 200 nanometers." Cell Signal 19(11): 2209-17.
Weidinger, S., L. Rummler, et al. (2005). "Association study of mast cell chymase polymorphisms with atopy." Allergy 60(10): 1256-61.
Wen L. (2001). "Two-step cycle sequencing improves base ambiguities and signal dropouts in DNA sequencing reactions using energy-transfer-based fluorescent dye terminators." Mol Biotechnol. 17(2): 135-42.
Wimmer, K., X. Roca, et al. (2007). "Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5' splice-site disruption." Hum Mutat 28(6): 599-612.
Wojnowski, L. and J. Brockmoller (2004). "Single nucleotide polymorphism characterization by mRNA expression imbalance assessment." Pharmacogenetics 14(4): 267-9.
Wood, A. J., R. Schulz, et al. (2008). "Regulation of alternative polyadenylation by genomic imprinting." Genes Dev 22(9): 1141-6.
Wu, S., C. M. Romfo, et al. (1999). "Functional recognition of the 3' splice site AG by the splicing factor U2AF35." Nature 402(6763): 832-5.
Xu, Q., B. Modrek, et al. (2002). "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome." Nucleic Acids Res 30(17): 3754-66.
Yamada, E., S. Okada, et al. (2005). "Akt2 phosphorylates Synip to regulate docking and fusion of GLUT4-containing vesicles." J Cell Biol 168(6): 921-8.
Yeo, G. and C. B. Burge (2004). "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals." J Comput Biol 11(2-3): 377-94.
Yeo, G., D. Holste, et al. (2004). "Variation in alternative splicing across human tissues." Genome Biol 5(10): R74.
Yeo, G. W., E. L. Van Nostrand, et al. (2007). "Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements." PLoS Genet 3(5): e85.
Zanke, B. W., C. M. Greenwood, et al. (2007). "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24." Nat Genet 39(8): 989-94.
Zatkova, A., L. Messiaen, et al. (2004). "Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1." Hum Mutat 24(6): 491-501.
Zhang, M. Q. (1998). "Statistical features of human exons and their flanking regions." Hum Mol Genet 7(5): 919-32.
Zheng, S. L., J. Sun, et al. (2008). "Cumulative association of five genetic variants with prostate cancer." N Engl J Med 358(9): 910-9.
Zhou, M. Y., S. E. Clark, et al. (1995). "Universal cloning method by TA strategy." Biotechniques 19(1): 34-5.
Zhu, J., J. Shendure, et al. (2003). "Single molecule profiling of alternative pre-mRNA splicing." Science 301(5634): 836-8.
Zhuang, Y. and A. M. Weiner (1986). "A compensatory base change in U1 snRNA suppresses a 5' splice site mutation." Cell 46(6): 827-35.
Zuccato, E., E. Buratti, et al. (2004). "An intronic polypyrimidine-rich element downstream of the donor site modulates cystic fibrosis transmembrane conductance regulator exon 9 alternative splicing." J Biol Chem 279(17): 16980-8.

10 CURRICULUM VITAE

PERSONAL INFORMATION

Name:	**Abdou Gomaa Abdou ElSharawy**
Date of birth:	August 25th, 1973
Place of birth:	El-Shoraa, Dameitta, Egypt
Citizenship:	Egyptian
Marital status:	Married with 2 children
Home address:	Germany: Hofholzallee 216. D-24109 Kiel Egypt*:* Magless EL-Mahaly Street, EL-Shoarra, Damietta
Work address:	Germany: Institute of Clinical Molecular Biology and University Clinic Schleswig-Holstein, Campus Kiel, Schittenhelmstraße 12. D-24105 Egypt: Chemistry Department, Faculty of Sciences (Mansoura University) at New Damietta City, Damietta
Email address:	a.sharawy@mucosa.de / el_sharawayabdou@yahoo.com

EDUCATION

1979 – 1984	6-years basic primary school, Elshoraa, Damietta, Egypt
1985 – 1987	3-years preparatory school, Elshoraa, Damietta, Egypt
1988–1990	3-years secondary school (Scientific section) at Damietta, Egypt (Egyptian university entrance qualification), which is equal to “Abitur" in Germany).

STUDY

1991 – 1992	Study of Biology/Geology, Faculty of Sciences, New Damietta City, Mansoura University, Egypt
1993-1995	Study of Biochemistry at faculty of sciences, New Damietta City, Mansoura Uni., Egypt
1995-1996	Preparatory courses for Master’s degree in Biochemistry at faculty of sciences, New Damietta City, Mansoura University, Egypt
1997-2000	Master’s degree in Biochemistry at Faculty of Sciences, New Damietta City, Mansoura University, Egypt. The title of research: Screening of Chromosomal Anomalies of Some Children in Dameitta Governorate.
2002-2003	PhD student (Biochemistry) at faculty of sciences, New Damietta City, Mansoura University, Egypt
2003/2004	Nominated for and awarded a personal scholarship from the Egyptian Government to study PhD abroad
2005- present	PhD student (Cell Biology) at Institute of Clinical Molecular Biology (Prof. Dr. Stefan Schreiber/PD. Dr Jochen Hampe) and Faculty of Mathematics and Natural Sciences at Christian-Albrechts-University of Kiel (Prof. Dr. Frank Kempken), Kiel, Germany. The title of my PhD project: Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing.

DEGREES

05.1991	Egyptian university entrance qualification/scientific (**Abitur**: 80.4%)
05.1995	**BSc**. (excellent with honor degree) in Biochemsitry/Chemistry
05.2000	**MSc**. in Biochemistry/Chemsitry (no grade system is given)

WORK HISTORY

12.1995-05.2000	Biochemistry Demonstrator, Chemistry Department, Faculty of Science at New Damietta City, Mansoura University, Egypt *.
06.2000- present	Biochemist Associate Lecturer (at the same place)*.

PROFESSIONAL SKILLS

Teaching

12.1995-09.2004	Teaching undergraduate chemistry/biochemistry students
01.2001-09.2004	Teaching graduate biochemistry (and diploma) students
1.07-05.08.2000	Attended and awarded "Teacher preparation/qualifying courses" at Faculty of Education, Mansoura University, Egypt.
Laboratory	Sequencing/mutation detection, genotyping, allele-dependent splicing, dual-band *in vitro* splicing assay, FACS analysis, cloning, western blotting, karyotyping, and others.
Computer	International Computer Driving License "**ICDL**" [under UNESCO supervision; Microsoft certified; authorized PROMETRIC testing center] at the Scientific Computer Center, Mansoura university, Egypt.
Languages	**Arabic:** Egyptian native tongue **German:** ZMB (Zentrale Mittelstufeprüfung Zeugnis). Total score: 104/120 (87%). Goethe Institute, Doqqi, Cairo, Egypt. **English:** International Computer-based TOEFL Exam; total score: 567(=227) points (Listening=22; Structure/Writing=23; Reading=23; Essay Rating=4.5). AMIDEST (America-Mideast Educational and Training Services INC, Dokki-Giza, Cairo, Egypt.

ACHIEVEMENTS

Awards

1991-1995	Outstanding student award Faculty of Science at Damietta, Mansoura University, Egypt-First place on the honor list
2003/2004	Awarded a four-year personal PhD-scholarship from the Egyptian Government (Ministry for Higher Education and Cultural Affairs)
DQ411321	Genbank submission: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=89212869

PERSONAL INTERESTS

Football, basketball, table-tennis, chess, cooking, and reading.

PUBLICATION LIST

1- **ElSharawy** A, Hundrieser B, Brosch M, Wittig M, Huse K, Platzer M, Becker A, Simon M, Rosenstiel P, Schreiber S, Krawczak M and Hampe J (2008). Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing. ***Hum Mutat.*** *In Press.*

2- Clemens Schafmayer, Stephan Buch, Henry Völzke, Witigo von Schönfels, Jan Hendrik Egberts, Bodo Schniewind, Mario Brosch, Andreas Ruether, Andre Franke, Micaela Mathiak, Bence Sipos, Tobias Henopp, Jasmin Catalcali, Stephan Hellmig, **Abdou ElSharawy**, Alexander Katalinic, Markus M Lerch, Ulrich John, Ulrich R. Fölsch, Fred Fändrich, Holger Kalthoff, Stefan Schreiber, Michael Krawczak, Jürgen Tepel, Jochen Hampe (2008). Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German case-control sample. ***Int J Cancer****;* 124(1):75-80.

3- Clemens Schafmayer, Henry Völzke, Stephan Buch, Jan Egberts, Annika Spille, Huberta von Eberstein, Andre Franke, Markus Seeger, Sebastian Hinz, **Abdou ElSharawy**, Dieter Rosskopf, Mario Brosch, Michael Krawczak, Ulrich R. Foelsch, Anton Schafmayer, Frank Lammert, Stefan Schreiber, Fred

Faendrich, Jochen Hampe, JuergenTepel (2007). Investigation of the *Lith6* candidate genes *APOBEC1* and *PPARG* in human gallstone disease. ***Liver International***; 27 (7): 910-919.

4- Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Hasler R, Little RD, Huse K, Ruether A, Balschun T, Wittig M, **Elsharawy A**, Mayr G, Albrecht M, Prescott NJ, Onnie CM, Fournier H, Keith T, Radelof U, Platzer M, Mathew CG, Stoll M, Krawczak M, Nurnberg P, Schreiber S (2007). Systematic Association Mapping Identifies NELL1 as a Novel IBD Disease Gene. ***PLoS ONE;*** 2:e691.

5- Buch S, Schafmayer C, Volzke H, Becker C, Franke A, von Eller-Eberstein H, Kluck C, Bassmann I, Brosch M, Lammert F, Miquel JF, Nervi F, Wittig M, Rosskopf D, Timm B, Holl C, Seeger M, **Elsharawy A**, Lu T, Egberts J, Fandrich F, Folsch UR, Krawczak M, Schreiber S, Nurnberg P, Tepel J, Hampe J (2007). A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. ***Nat Genet.***; 39(8):995-999.

6- Schafmayer C, Buch S, Egberts JH, Franke A, Brosch M, **El Sharawy A**, Conring M, Koschnick M, Schwiedernoch S, Katalinic A, Kremer B, Folsch UR, Krawczak M, Fandrich F, Schreiber S, Tepel J, Hampe J (2007). Genetic investigation of DNA-repair pathway genes PMS2, MLH1, MSH2, MSH6, MUTYH, OGG1 and MTH1 in sporadic colon cancer. ***Int J Cancer***; 121(3): 555-8.

7- **ElSharawy A**, Manaster C, Teuber M, Rosenstiel P, Kwiatkowski R, Huse K, Platzer M, Becker A, Nurnberg P, Schreiber S and Hampe J (2006). SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs. ***Hum Mutat.***; 27(11): 1129-34.

[The previous methodology paper is cited by: Skotheim et al., Int J Biochem Cell Biol (2007); Lin e tal., Oncogene (2007); Roca et al., Genome Res (2008);and Kim et al., BMC Bioinformatics (2008)]

8- Karawya EM; Abdel-Malak CA; Settin AA and **ElSharawy A** (2000). Chromosomal anomalies in mentally handicapped Children from Damietta Governorate. ***TheEgyptian Journal of Medical Sciences***; 21(1) June: 237-246.

ORAL PRESENTATIONS AT CONFERENCES AND MEETINGS

16.11.2006 Invited speaker at the NGFN Environmental Network Meeting, symposium VI 'From genomic variation to functional analysis', Heidelberg, Germany. Entitled: Establishment of a high-throughput methodology for the investigation of SNP-dependent splicing.

06.09.2006 Gave a lecture entitled "Systematic Investigation of SNP-dependent Splicing" in the project (section 3) of the laboratory course of: Stem Cell Biology: Future Perspectives and Possible Treatment Options. Department of General and Thoracic Surgery, section of Biotechnology and Transplantation Medicine. University Hospital, campus Kiel (UK-SH). Christian-Albrechts-University of Kiel. Kiel, Germany.

04.11.08 Invited speaker at the 'Alexander-von-Humboldt Foundation', Kiel, Germany. Entitled: Kiel aus Sicht einese ausländischen Gastwissenschaftler-Ehepaares (View of Kiel of a foreigner scientists married couple).

CONGRESS POSTERS AND ABSTRACTS

20-23.09.2006 Plant Genetics conference. Joint conference of the German Genetics Society and the German Society for Plant Breeding. Plant Institute, Kiel University, Kiel, Germany. [**Abdou ElSharawy**, Carl Manaster, Markus Teuber, Philip Rosenstiel, Ruta Kwiatkowski, Klaus Huse, Matthias Platzer, Albert Becker, Peter Nürnberg, Stefan Schreiber, Jochen Hampe: SNPSplicer- systematic analysis of SNP-dependent splicing in genotyped cDNA]

25-26.11.2006 Fifth National Genome Research Network (NGFN) conference; an official meeting of the German Federal Ministry for Education. German Cancer Research Center (DKFZ), Heidelberg, Germany. [**Abdou ElSharawy**, Klaus Huse, Markus Teuber, Andre Franke,

	Michael Wittig, Carl Manaster, Albert Becker, Michael Krawczak, Matthias Platzer, Stefan Schreiber, Jochen Hampe: Establishment of a high-throughput methodology for the investigation of SNP-dependent splicing]
12-14.07.2007	Inflammatory Diseases of Barrier Organs: Genetic Exploration leads to novel Therapies. A symposium of the national genome research network (NGFN), University Clinic of Schleswig-Holstein, AUDIMAX Lectures Hall, Christian-Albrechts-University (CAU), Kiel, Germany. Session: From genomics to function. [**Abdou ElSharawy**, Carl Manaster, Markus Teuber, Philip Rosenstiel, Ruta Kwiatkowski, Klaus Huse, Matthias Platzer, Albert Becker, Peter Nürnberg, Stefan Schreiber, Jochen Hampe: SNPSplicer- systematic analysis of SNP-dependent splicing in genotyped cDNA]
28.07 - 03.08.2008	Thirteenth Annual Meeting of the RNA Society 2008, Berlin, Germany: [**Abdou ElSharawy**, Bernd Hundrieser, Mario Brosch, Michael Wittig, Klaus Huse, Matthias Platzer, Albert Becker, Matthias Simon, Philip Rosenstiel, Stefan Schreiber, Michael Krawczak, Jochen Hampe: Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing]

ATTENDANCE AT CONGRESSES

21-24.09.2005	Joint Meeting. 36. Annual Meeting of the German Society of Immunology (DGFI) & 36. Annual meeting of the Scandinavian Society for Immunology (SSI). Kiel, Germany.
03-04.06.2005	Inflammatory Diseases of Barrier Organs. University Clinic Kiel, Germany.
24-25.08.2006	3rd International Symposium „Molecular and Clinical Aspects of Cellular Signaling“ of the SFB 415 “Specificity and Pathophysiology of Signal Transduction Pathways”. A joint symposium with NGFN. University of Kiel, Germany.
20.08.-01.09.2006	Stem Cell Biology: Future Perspectives and Possible Treatment Options. Department of General and Thoracic Surgery, section of Biotechnology and Transplantation Medicine. University Clinic Schleswig Holstein, Campus Kiel (UK-SH). Christian-Albrechts-University of Kiel. Kiel, Germany.
05-06.10.2006	International Symposium "RNAi in vivo Technologies". On behalf of NGFN RiNA organizes the in cooperation with the GSF-National Research Center for Environment and Health in Munich. GSF- National Research Center for Environment and Health, Munich, Germany.

ATTENDANCE AT TRAINING COURSES AND WORKSHOPS

22-26.03.1998	Techniques in Molecular Biology. Institute of Graduate Studies and Research, University of Alexandria , Egypt.
4-8.04.1999	Modern Techniques in Genetic Engineering. Institute of Graduate Studies and Research, Alexandria University, Egypt.
2-7.07.2000	PCR: Principles and techniques. Institute of Graduate Studies and Research, Alexandria University, Egypt.
8-10.01.2001	Detection of Genetic Modifications in Food and Feed. A collaborative activity of the Biotechnology Research Center at Suez Canal University (SCU) and the German Federal Institute for Health Protection and Veterinary Medicine (bgvv). Held at SCU, Ismailia, Egypt.
14-18.10.2001	Training Course on PCR: Basics and Applications. Institute of Graduate Studies and Research, Alexandria University, Egypt.
14-18.07.2002	Polymerase Chain Reaction (PCR): Methods and Applications. Institute of Graduate Studies and Research, Alexandria University, Egypt.
28.03.2007	Alternative Splicing - Regulation and Evolution. SFB604, workshop5. Thüringer Universitäts- und Landesbibliothek, Bibliotheksplatz 2, Jena, Germany.

11 DECLARATION

Declaration

Apart from the advice of my supervisors, this thesis is completely the result of my own work. No part of it has been submitted to any other board for another qualification. Most of the results have been or are about to be published (see below).

Erklärung

Hiermit erkläre ich, daß diese Dissertation, abgesehen von der Beratung durch meine akademischen Lehrer, nach Inhalt und Form meine eigene Arbeit ist. Sie hat weder im Ganzen noch zum Teil an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen. Die meisten Ergebnisse dieser Arbeit wurden zur Veröffentlichung eingereicht (siehe unten).

Kiel,.....(**Abdou ElSharawy**)

12 ACKNOWLEDGMENT

I would like to thank **Prof. Dr. Stefan Schreiber** for giving me the opportunity to work on my PhD in a very well-equipped scientific environment at ICMB, for his supervision, support and encouragement during the course of my study. I am especially grateful for his helpful discussion, sharing his wealth of experience, careful proof-reading of this thesis and for his intellectual input. Once again, I would like to thank him for his kind concern and support for me and my family and for his financial support.

I would like to thank **Prof. Dr. Frank Kempken** from the Botanical Institute of the Christian-Albrechts University of Kiel for taking on the role of my official supervisor in the Faculty of Mathematics and Natural Sciences. I would like to thank him for his motivation, giving me the possibilities to present the progress of my work during the laboratory meetings of his workgroup, careful proof-reading of this thesis and for his intellectual input. Indeed, I was very motivated after each meeting with him.

I would like to express my thanks to **PD. Dr. Jochen Hampe** for his supervision, great support and encouragement, and introducing me into alternative splicing and human genetics. His precious scientific work and project steering provided indeed the basis of this thesis. Furthermore, his time-intensive investments into the laboratory infrastructure and the LIMS are gratefully acknowledged. I would like to thank him for his careful proof-reading of this thesis and for his intellectual input. The DFG grant application (DFG Ha 3091/2-1) that he wrote was in fact the basis of my research program and allowed me to perform the large-scaled experiments. Once again, I would like to thank him for his kind concern and financial support especially during the last months of my PhD study in Germany.

I would like to thank **Prof. Dr. Michael Krawczak** for his great contribution in operating the neural work to select candidate splice SNPs and his valuable scientific ideas and comments.

I would like to thanks **Dr. Mario Brosch** for his laboratory guidance during establishment of the *in vitro* splice reporter system and for his careful proof-reading of this thesis and for his intellectual input.

I would like to thank all members of ICMB especially **Prof. Dr. Philip Rosenstiel, Dr. Andreas Till,** for their valuable contributions and kind support in performing the FACS analysis and support. I am grateful to **Michael Wittig** for writing the scripts of 'MotifSNps' and 'SkippedExonPrimer' tools, and to **Carl Manaster** for programming the SNPSplicer software, **Markus Teuber** for programming of the SpliceTool, and together with **Marcus Will** for their IT expertise and maintenance of the server and the network. Also I would like to thanks **Rainer Vogler and Birgitt Timm** for their help with the database.

I would like to thank our collaborative partners, **Dr. Klaus Huse, Dr. Matthias Platzer, Dr. Peter Nürnberg, and Dr. Albert Becker** for their efforts and contributions.

I would like to thank the members of the main office at ICMB, **Prof. Dr. Andre Franke, Dr. Ruta Kwiatkowski, Dr. Friederike Flachsbart, Sandra Their, Annegret Fischer, Rabea Kleindorp, Dr. Weiyue Zheng, and Dr. Almut Nebel** for the pleasant and warm atmosphere, though living together in a physically restricted office under occasional time pressure.

Special thanks go to **Dr. Nancy Mah** for her kind help in proofreading this thesis and for her professional comments. Thanks also to **Stephan Buch** for his kind friendship, help and translation of the summary into German.

Thanks also to **Ingelore Bäßmann, Meike Barche, Lena Bossen, Birthe Fedders, Tanja Wesse, Tanja Kaacksteen, Yasmin Brodtmann, Ilona Urbach, Catharina Fürstenau, Tanja Henke, Anita Dietsch, Melanie Friskovec, Susan Ehlers, Catharina von der Lancken, Sabine Sepke and Caro Lorenzen** for expert technical assistance.

I thank the **entire staff** and **assistants** of Prof. Dr. Schreiber at the ICMB and the **research group** of PD. Dr. Hampe for the very nice atmosphere and kind encouragement during the course of this study. Special thanks to **Dr. Andi Rüther, Dr. Christiane Wolf-Schwerin, Dr. Oliver Von Kampen, Dr. Clemens Schafmayer, and Alexander Hermann** for their kind concern and support.

This work was basically funded by the **Deutsche Forschungsgesellschaft (DFG)** of PD. Dr. Jochen Hampe (DFG Ha 3091/2-1) and Dr. Klaus Huse (DFG Hu 498/3-1).

It is also opportune here to thank the **Egyptian government** in the form of the Minister of Higher Education and Cultural Affairs and **Faculty of Sciences at Damietta/Mansura University**, where I am working in Egypt, for awarding me a full personal scholarship to study my PhD and learn more about the great culture of Germany, where I and my family have felt at home, won many kind friends and colleagues, and spent the most pleasant and beautiful time in our lives.

Finally, very special thanks to **my parents** for their constant motivation and support since the beginning of my life. They were always beside me with their prayers and they never feel unhappy or angry at me although I visited them only twice during all the 4 years I have been abroad. I would like to thank **my wife Dr. Ola Asker and sons Ahmed and Amr** for their love, kind companionship and support in every step during the course of my study, and their understanding for the long working hours especially during the weekends. I would also like to thank my wife (as my direct government) once again for her strong support and spending days in collecting the special topics of cell biology to prepare for my exam and for her delicious baking and Egyptian food.

List of publications and activities

*** Related to the thesis**	
Articles	
1	**ElSharawy A**, Hundrieser B, Brosch M, Wittig M, Huse K, Platzer M, Becker A, Simon M, Rosenstiel P, Schreiber S, Krawczak M and Hampe J (2008). Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing. **Hum Mutat.** *In Press* (has been accepted on 14.08.2008).
2	**ElSharawy A**, Manaster C, Teuber M, Rosenstiel P, Kwiatkowski R, Huse K, Platzer M, Becker A, Nurnberg P, Schreiber S and Hampe J (2006). SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs. **Hum Mutat**.; 27(11): 1129-34. *[This methodology paper is cited by: Skotheim et al., Int J Biochem Cell Biol (2007); Lin e tal., Oncogene (2007); Roca et al., Genome Res (2008);and Kim et al., BMC Bioinformatics (2008);...]*
Congress abstracts and posters	
1	Thirteenth Annual Meeting of the RNA Society 2008: [**Abdou ElSharawy**, Bernd Hundrieser, Mario Brosch, Michael Wittig, Klaus Huse, Matthias Platzer, Albert Becker, Matthias Simon, Philip Rosenstiel, Stefan Schreiber, Michael Krawczak, Jochen Hampe: Systematic Evaluation of the Effect of Common SNPs on Pre-mRNA Splicing]. (Berlin, Germany; 28.07 - 03.08.2008)
2	Inflammatory Diseases of Barrier Organs: Genetic Exploration leads to novel Therapies. A symposium of the national genome research network (NGFN), University Clinic of Schleswig-Holstein. Session: From genomics to function. [**Abdou ElSharawy**, Carl Manaster, Markus Teuber, Philip Rosenstiel, Ruta Kwiatkowski, Klaus Huse, Matthias Platzer, Albert Becker, Peter Nürnberg, Stefan Schreiber, Jochen Hampe: SNPSplicer- systematic analysis of SNP-dependent splicing in genotyped cDNA]. (CAU, Kiel, Germany; 12-14.07.2007)
3	Fifth National Genome Research Network (NGFN) conference; an official meeting of the German Federal Ministry for Education. [**Abdou ElSharawy**, Klaus Huse, Markus Teuber, Andre Franke, Michael Wittig, Carl Manaster, Albert Becker, Michael Krawczak, Matthias Platzer, Stefan Schreiber, Jochen Hampe: Establishment of a high-throughput methodology for the investigation of SNP-dependent splicing]. (German Cancer Research Center (DKFZ), Heidelberg, Germany; 25-26.11.2006)
4	Plant Genetics conference. Joint conference of the German Genetics Society and the German Society for Plant Breeding. [**Abdou ElSharawy**, Carl Manaster, Markus Teuber, Philip Rosenstiel, Ruta Kwiatkowski, Klaus Huse, Matthias Platzer, Albert Becker, Peter Nürnberg, Stefan Schreiber, Jochen Hampe: SNPSplicer- systematic analysis of SNP-dependent splicing in genotyped cDNA]. (Plant Institute, Kiel, Germany; 20-23.09.2006)
Oral representations	
1	Invited speaker at the NGFN Environmental Network Meeting, symposium VI 'From genomic variation to functional analysis'. Entitled: Establishment of a high-throughput methodology for the investigation of SNP-dependent splicing. (Heidelberg, Germany; 16.11.2006)
2	Gave a lecture entitled "Systematic Investigation of SNP-dependent Splicing" in the project (section 3) of the laboratory course of: Stem Cell Biology: Future Perspectives and Possible Treatment Options. Department of General and Thoracic Surgery, section of Biotechnology and Transplantation Medicine. (University clinic at Kiel (UK-SH), Kiel, Germany; 06.09.2006)
*** Other publications (applications, collaborations,..)**	
1	Clemens Schafmayer, Stephan Buch, Henry Völzke, Witigo von Schönfels, Jan Hendrik Egberts, Bodo Schniewind, Mario Brosch, Andreas Ruether, Andre Franke, Micaela Mathiak, Bence Sipos, Tobias Henopp, Jasmin Catalcali, Stephan Hellmig, **Abdou ElSharawy**, Alexander Katalinic, Markus M Lerch, Ulrich John, Ulrich R. Fölsch, Fred Fändrich, Holger Kalthoff, Stefan Schreiber, Michael Krawczak, Jürgen Tepel, Jochen Hampe (2008). Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German case-control sample. ***Int J Cancer***; 124(1):75-80.
2	Clemens Schafmayer, Henry Völzke, Stephan Buch, Jan Egberts, Annika Spille, Huberta von Eberstein, Andre Franke, Markus Seeger, Sebastian Hinz, **Abdou ElSharawy**, Dieter Rosskopf, Mario Brosch, Michael Krawczak, Ulrich R. Foelsch, Anton Schafmayer, Frank Lammert, Stefan Schreiber, Fred Faendrich, Jochen Hampe, JuergenTepel (2007). Investigation of the *Lith6* candidate genes *APOBEC1* and

	PPARG in human gallstone disease. ***Liver International***; 27 (7): 910-919.
3	Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Hasler R, Little RD, Huse K, Ruether A, Balschun T, Wittig M, **Elsharawy A**, Mayr G, Albrecht M, Prescott NJ, Onnie CM, Fournier H, Keith T, Radelof U, Platzer M, Mathew CG, Stoll M, Krawczak M, Nurnberg P, Schreiber S (2007). Systematic Association Mapping Identifies NELL1 as a Novel IBD Disease Gene. ***PLoS ONE;*** 2:e691.
4	Buch S, Schafmayer C, Volzke H, Becker C, Franke A, von Eller-Eberstein H, Kluck C, Bassmann I, Brosch M, Lammert F, Miquel JF, Nervi F, Wittig M, Rosskopf D, Timm B, Holl C, Seeger M, **Elsharawy A**, Lu T, Egberts J, Fandrich F, Folsch UR, Krawczak M, Schreiber S, Nurnberg P, Tepel J, Hampe J (2007). A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. ***Nat Genet.***; 39(8):995-999.
5	Schafmayer C, Buch S, Egberts JH, Franke A, Brosch M, **El Sharawy A**, Conring M, Koschnick M, Schwiedernoch S, Katalinic A, Kremer B, Folsch UR, Krawczak M, Fandrich F, Schreiber S, Tepel J, Hampe J (2007). Genetic investigation of DNA-repair pathway genes PMS2, MLH1, MSH2, MSH6, MUTYH, OGG1 and MTH1 in sporadic colon cancer. ***Int J Cancer***; 121(3): 555-8.
6	Karawya EM; Abdel-Malak CA; Settin AA and **ElSharawy A** (2000). Chromosomal anomalies in mentally handicapped Children from Damietta Governorate. ***TheEgyptian Journal of Medical Sciences***; 21(1) June: 237-246.
*** Attendance at other congresses**	
1	International Symposium "RNAi in vivo Technologies". On behalf of NGFN RiNA organizes the in cooperation with the GSF-National Research Center for Environment and Health in Munich. GSF- National Research Center for Environment and Health. (Munich, Germany; 05-06.10.2006)
2	Stem Cell Biology: Future Perspectives and Possible Treatment Options. Department of General and Thoracic Surgery, section of Biotechnology and Transplantation Medicine. (University Clinic Schleswig Holstein (UK-SH), CAU. Kiel, Germany; 20.08.-01.09.2006).
3	3rd International Symposium „Molecular and Clinical Aspects of Cellular Signaling“ of the SFB 415 “Specificity and Pathophysiology of Signal Transduction Pathways”. A joint symposium with NGFN. University of Kiel, Germany. (Date: 24-25.08.2006).
4	Inflammatory Diseases of Barrier Organs. (University Clinic UKSH, Kiel, Germany; 03-04.06.2005).
5	Joint Meeting. 36. Annual Meeting of the German Society of Immunology (DGFI) & 36. Annual meeting of the Scandinavian Society for Immunology (SSI). (Kiel, Germany; 21-24.09.2005).
*** Attendance at training courses and workshops**	
1	Alternative Splicing - Regulation and Evolution. SFB604, workshop5. (Thüringer Universitäts- und Landesbibliothek, Jena, Germany; 28.03.2007).
2	Polymerase Chain Reaction (PCR): Methods and Applications. (Institute of Graduate Studies and Research, Alexandria University, Egypt; 14-18.07.2002)
3	Training Course on PCR: Basics and Applications. (Institute of Graduate Studies and Research, Alexandria University, Egypt; 14-18.10.2001).
4	Detection of Genetic Modifications in Food and Feed. A collaborative activity of the Biotechnology Research Center at Suez Canal University (SCU) and the German Federal Institute for Health Protection and Veterinary Medicine (bgvv). (Held at SCU, Ismailia, Egypt; 8-10.01.2001).
5	PCR: Principles and techniques. (Institute of Graduate Studies and Research, Alexandria University, Egypt; 2-7.07.2000).
6	Modern Techniques in Genetic Engineering. (Institute of Graduate Studies and Research, Alexandria University, Egypt; 4-8.04.1999).
7	Techniques in Molecular Biology. (Institute of Graduate Studies and Research, University of Alexandria , Egypt; 22-26.03.1998).