Expression changes confirm predicted single nucleotide variants affecting mRNA splicing

Eliseos J. Mucaki and Peter K. Rogan

Departments of Biochemistry and Computer Science

University of Western Ontario,

London ON

Correspondence:

Peter K. Rogan, Ph.D.

SDRI 201A, Schulich School of Medicine and Dentistry

The University of Western Ontario

London ON Canada N6A 5C1

E: progan@uwo.ca

T: 519-661-4255

Abstract

Mutations that cause genetic diseases can be difficult to identify if the mutation does not affect the sequence of the protein, but the splice form of the transcript. However, the prediction of deleterious changes caused by genomic variants that affect splicing has been shown to be accurate using information theory-based methods. We made several such predictions of potential splicing changes that could be caused by SNPs which were found to cause natural and/or cryptic splice site strength changes. We evaluated a selected set of 22 SNPs that we predicted by information analysis to affect splicing, validated these with targeted expression analysis, and compared the results with genome-scale interpretation of RNAseq data from tumors. Abundance of natural and predicted splice isoforms were quantified by q-RT-PCR and with probeset intensities from exon microarrays using RNA isolated from HapMap lymphoblastoid cell lines containing the predicted deleterious variants. These SNPs reside within the following genes: *XRCC4*, *IL19*, *C21orf2*, *UBASH3A*, *TTC3*, *PRAME*, *EMID1*, *ARFGAP3*, *GUSBP11* (*Fλ8)*, *WBP2NL*, *LPP*, *IFI44L*, *CFLAR*, *FAM3B*, *CYB5R3*, *COL6A2*, *BCR*, *BACE2*, *CLDN14*, *TMPRSS3* and *DERL3*. 15 of these SNPs showed a significant change in the use of the affected splice site. Individuals homozygous for the stronger allele had higher transcription of the associated gene than individuals with the weaker allele in 3 of these SNPs. 13 SNPs had a direct effect on exon inclusion, while 10 altered cryptic site use. In 4 genes, individuals of the same genotype had high expression variability caused by alternate factors which masked potential

effects of the SNP. Targeted expression analyses for 8 SNPs in this study were confirmed by results of genome-wide information theory and expression analyses.

Keywords

Allele specific gene expression, mRNA splicing, single nucleotide polymorphism, mutation, alternative splicing, information theory

Introduction

Accurate and comprehensive methods are needed for predicting impact of non-coding mutations, in particular splicing defects, which are prevalent in causing genetic disease (Krawczak *et al.* 1992; Teraoka *et al.* 1999; Ars *et al.* 2003). This class of mutations may account for as much as 62% of point mutations (López-Bigas *et al.* 2005). While bioinformatic approaches can identify potential splicing mutations, predicting the relative abundance of various transcripts has proven to be challenging because of the complexity of splice site recognition.

The selection of splicing signals involves exon and intron sequences, complementarity with snRNAs, RNA secondary structure and competition between spliceosomal recognition sites (Berget 1995, Moore and Sharp 1993). U1 snRNP interacts with the donor (or 5') splice site (Séraphin *et al.* 1988, Zhuang *et al.*1986) and U2 (and U6) snRNP with the acceptor and branch sites of pre-mRNA (Parker *et al.* 1987, Wu and Manley, 1989). Although both of the U1 and U2 spliceosomal interactions involve base pairing to RNA, the majority of human splice donors (5') and acceptors (3') that are not precisely complementary to these sequences (Rogan *et al.* 2003).

Additional exonic and intronic cis-regulatory elements can promote or suppress splice site recognition through recruitment of trans-acting splicing factors. These factors contain RNA-recognition motifs (RRM) and a carboxy-terminal domain enriched in Arg/Ser dipeptides (SR domain; Birney *et al.* 1993). SR proteins function in splice site communication by forming an intron bridge needed for exon recognition (Zuo and Maniatis 1996). Binding of RRMs in pre-mRNA enhances exon recognition by promoting interactions with spliceosomal and other proteins (Fu and Maniatis 1992).

Splicing mutations affect normal exon recognition by altering the strengths of natural donor or acceptor sites and proximate cryptic sites either independently or simultaneously. Weaker splice sites can reduce of kinetics of mRNA processing, leading to overall decrease in full length transcripts, increased exon skipping, cryptic splice site activation within exons or within adjacent introns, intron retention, and inclusion of cryptic, pseudo-exons (Buratti 2006; Talerico and Berget, 1990; Carothers *et al.* 1993). Aberrant mRNAs may alter the reading frame, resulting in unstable and degraded or truncated proteins. The kinetics of splicing at weaker cryptic sites is slower than at authentic splice sites (Domenjoud *et al.* 1993). Mutations strengthen cryptic sites either by increasing resemblance to consensus sequences (Nelson and Green 1990) or by modulating the levels of SR proteins contributing to splice site recognition (Cáceres *et al.* 1994; Mayeda and Krainer 1992). Splicing mutations at splicing regulatory elements (Dietz *et al.* 1993; Richard and Beckmann 1995) may also occur through disruption of trans-acting SR protein interactions (Staknis and Reed 1994) with distinct exonic and intronic cis-regulatory elements (Black 2003).

Information theory-based (IT-based) models of donor and acceptor mRNA splice sites reveal the effects of changes in strengths of individual sites (Rogan *et al.* 1998; Rogan *et al.* 2003). This facilitates prediction of phenotypic severity (Rogan and Schneider 1995; von Kodolitsch *et al.* 1999; von Kodolitsch *et al.* 2006). The effects of splicing mutations can be predicted *in silico* by information theory (Rogan and Schneider 1995; Rogan *et al.* 1998; O'Neill *et al.* 1998; Allikmets *et al.* 1998; Kannabiran *et al.* 1998; Khan *et al.* 1998; von Kodolitsch 1999; Vockley *et al.* 2000; Svojanovsky *et al.* 2000; Khan *et al.* 2002; Rogan *et al.* 2003; Lamba *et al.* 2003; Khan *et al.* 2004; von Kodolitsch *et al.* 2006, Viner *et al.* 2014; Dorman *et al.* 2014; Shirley *et al.* 2018) and predictions confirmed *in vitro* by experimental studies (Vockley *et al.* 2000; Rogan *et al.* 2003; Lamba *et al.* 2003; Susani *et al.* 2004; Hobson *et al.* 2006). Strengths of one or more splice sites may be altered and, in some instances, concomitant with amino acid changes in coding sequences (Rogan *et al.* 1998). Information analysis has been a successful approach for recognizing non-deleterious variants (Rogan and Schneider 1995), and for distinguishing of milder from severe mutations (Rogan *et al.* 1998; von Kodolitsch *et al.* 1999).

Recent genome wide association studies have associated common SNP variants with expression levels of linked genes (Stranger *et al.* 2007). Functional evaluation of SNPs can potentially accelerate recognition of allelic variants associated with inherited predisposition to disease. We are applying quantitative methods to prioritize such studies by inferring which SNPs impact gene expression levels and transcript structure. Models of nucleic acid binding sites based on information theory have successfully been used to detect cis-acting promoter modules in the genome (Vyhlidal *et al.* 2004; Bi and Rogan, 2004) and splicing signals in transcripts. In the

present study, we explicitly predict and validate SNPs that influence mRNA structure and levels of expression of the genes containing them.

The robustness of information analysis in predicting splicing mutations for Mendelian disorders justifies the use of this approach to identify SNPs that are likely to have a measurable impact on mRNA splicing. Others have used exon microarrays to compare different cellular states and then confirm suggested abnormalities from the expression data using q-RT-PCR (Thorsen *et al.* 2008). We hypothesized that the predicted effect of SNPs on expression of the proximate exon would correspond to the expression of exon microarray probes of genotyped individuals in the HapMap cohort. We used the dose-dependent expression of the minor allele to qualify SNPs for subsequent information analysis consistent with alterations of mRNA splicing. These predicted mutations were then analyzed by q-RT-PCR to validate the accuracy of the bioinformatic predictions.

We recently described several deleterious single nucleotide polymorphisms in dbSNP that affect splicing and at least one of these is common (Nalla and Rogan 2005). This analysis used the NCBI Entrez query engine which conservatively defines splicing-related SNPs as only those variants involving the dinucleotides immediately adjacent to exon boundaries. Given that constitutive splicing mutations can arise at other locations within pre-mRNA sequences and can involve cryptic splicing, we addressed whether other genomic variants might be a source of common mutation. To test the feasibility of this hypothesis, we used information analysis to examine the potential impact of SNPs mapped predominantly onto the genome sequences of chromosomes 21 and 22 on splicing.

Materials and Methods

IT-based analysis. The protein-nucleic acid interactions intrinsic to splicing can be analyzed using information theory, which comprehensively and quantitatively models functional sequence variation based on a thermodynamic framework (Schneider 1997). Donor and acceptor splice site strength can be predicted by the use of IT-based weight matrices, which are derived from known functional sites (Rogan *et al.* 2003). The Automated Splice Site server (ASSA; https://splice.uwo.ca) is an online resource based on the April 2003 build of the genome to calculate splice site changes (Nalla and Rogan, 2005). The ASSA was scaled to carry out batch processing of multiple dbSNP entries.

Creation of exon array database. The download and PLIER (Probe Logarithmic Intensity Error) normalization of the Affymetrix Human Exon 1.0 ST microarray data testing 176 genotyped HapMap cell lines (Huang *et al.* 2007, Gene Expression Omnibus accession no. GSE 7792) has been previously described (Nembaware *et al.* 2008). Probes which overlap SNPs were determined by intersecting dbSNP129 with probe coordinates (obtained from X:MAP [Yates *et al.* 2007]) using the Galaxy Browser [Giardine *et al.* 2005]), and were removed from analysis.

A MySQL database was created and contains the PLIER normalized intensities, as well as CEPH and Yoruba genotypes for Phase I+II HapMap SNPs (downloaded Nov. 17Th 2008). Perl scripts were written to create tables which linked a SNP to its nearest like-stranded probeset (to within 500nt), and to associate probesets to the exons they may overlap (transcript and exon tables from Ensembl version 51). A MySQL query was used to create a table of the splicing index (SI; intensity of a probeset / gene intensity) of each probeset for each HapMap individual.

A Perl script which queried the database was written to find significant SI changes of an

exonic probeset based on the genotype of a SNP it is associated to (SNP within natural donor/acceptor region of exon). Probesets displaying a stepwise change in mean SI (where mean SI of homozygous rare is < 90% of homozygous common with a simultaneous decrease in heterozygotes) were found with another script. Splicing Index boxplots were created using R, where the x and y-axis are genotype and SI, respectively (Supplemental Figure 1). This gives a visual representation of all 176 individuals (if genotyped), and allows one to quickly analyze the effect a SNP has on a particular probeset.

SNPs of varying strength changes ($\geq$ 0.5 bits) were chosen to be further analyzed by q-RT-PCR. ΔR_i < 1 bit were included to determine if these small changes lead to detectable changes in splicing. Additional SNPs tested for splicing effects were predicted in previous publications (Nalla & Rogan, 2005).

Cell culture & RNA extraction. EBV-transformed lymphoblastoid cell lines of HapMap individuals with our SNPs of interest (homozygous common, heterozygous and homozygous rare when available) were ordered from the Coriell Cell Repositories (CEU: GM07000, GM07019, GM07022, GM07056, GM11992, GM11994, GM11995, GM12872. YRI: GM18855, GM18858, GM18859, GM18860, GM19092, GM19093, GM19094, GM19140, GM19159). Cells were grown in HyClone RPMI-1640 medium (15% FBS [HyClone], 1% L-Glutamine and 1% Penicillin:streptomycin [Invitrogen]; 37ºC, 5% CO_2). RNA was extracted with Trizol LS (Invitrogen) from 10^6 cells and treated with DNAase (20mM $MgCl_2$ [Invitrogen], 2mM DTT [Sigma-Aldrich], 0.4U/uL RNasin [Promega], 10µg/mL DNase [Worthington Biochemical] in 1x TE buffer) at 37ºC for 15 minutes. The reaction was stopped with EDTA (0.05M; 2.5% v/v), and heated to 65ºC for 20 minutes, followed by ethanol precipitation (resuspended in 0.1% v/v

DEPC-treated 1x TE buffer). DNA was extracted using a Puregene Tissue Core Kit B (Qiagen).

Design of real-time expression assay. Sequences were obtained from UCSC and Ensembl. DNA primers used to amplify a known splice form or one predicted by information analysis were designed using Primer Express (ABI). DNA primers were obtained from IDT (Coralville, IA, USA), and dissolved to 200 uM; sequences in Supplemental Table 1. Primers were designed to amplify the wildtype splice form, exon skipping (if a natural site is weakened), and cryptic site splice forms which were previously reported (UCSC mRNA and EST tracks) or those predicted by information analysis (where R_i cryptic site $\geq R_i$ weakened natural site).

Two types of reference amplicons were used to quantify allele specific splice forms. These consisted of intrinsic products derived from constitutively spliced exons with the same gene and external genes with high uniformity of expression among HapMap cell lines. Reference primers internal to the genes of interest were designed 1-4 exons adjacent from the affected exon (without evidence of variation from the UCSC Genome Browser), placed upstream of the SNP of interest whenever possible. Two advantages to including an internal reference in the q-RT-PCR experiment include: potential detection of changes in total mRNA levels; and account for inter-individual variation of expression.

External reference genes were chosen based on consistent PLIER intensities with low coefficients of variation in expression among all 176 HapMap individuals. The following external controls were selected: exon 39 of *SI* (PLIER intensity 11.4 ± 1.7), exon 9 of *FRMPD1* (22 ± 2.81), exon 46 of *DNAH1* (78.5 ± 9.54), exon 3 of *CCDC137* (224 ± 25) and exon 25 of *VPS39* (497 ± 76). The external reference chosen for an experiment was matched to the intensity of the probeset within the exon of interest. This decreased potential errors in $\Delta\Delta C_T$ values and

proved to be accurate and reproducible for most genes.

Primers were placed over junctions of interest (whenever possible) to amplify a single splice form. T_m ranged from 58-65ºC, and amplicon lengths varied from 69-136nt. BLASTn (refseq_rna database) was used to reduce possible cross-hybridization.

PCR and quantitative RT-PCR. M-MLV reverse transcriptase (Invitrogen) converted 1µg of DNase-treated RNA to cDNA with 20nt Oligo-dT (25µg/mL; IDT) and rRNAsin (Promega). Precipitated cDNA was resuspended in water at 20ng/µL of original RNA concentration.

All designed primer sets were tested with conventional PCR to ensure a single product at the expected size. PCR reactions were prepared with 1.0M Betaine (Sigma-Aldrich), and were heated to 80ºC before adding Taq Polymerase (Invitrogen). Optimal T_m for each primer set was determined to obtain maximum yield.

Quantitative PCR was performed on an Eppendorf Mastercycler ep Realplex 4, a Bio-Rad CFX96, as well as a Stratagene Mx3005P. SYBR Green assays were performed using the KAPA SYBR FAST qPCR kit (Kapa Biosystems) in 10µL reactions using 200µM of each primer and 24ng total of cDNA per reaction. For some tests, SsoFast Eva Green supermix (Bio-Rad) was used with 500µM of each primer instead.

When testing the effect of a SNP, all primers designed for that SNP as well as the gene internal reference and external reference, were run simultaneously. C_t values obtained from these experiments are normalized to its external reference using the Relative Expression Software Tool (REST; http://www.gene-quantification.de/rest.html; Phaffl *et al.* 2002).

Taqman Assay. Two dual-labeled Taqman probes were designed to detect the two splice forms of *XRCC4* (6nt deletion). Probes were placed over the sequence junction of interest where

variation would be near the probe middle (Supplemental Figure 1). The assay was performed on an ABI StepOne Real-Time PCR system using ABI Genotyping Master Mix. Experiment was run in 25µL reactions (300nM each primer, 400nM probe [5'-FAM or TET fluorophore with a 3' Black Hole quencher; IDT], and 80ng cDNA total). Probes were tested in separate reactions.

Results

Selection of candidate SNPs affecting splicing. Publically available exon microarray data was used to find exons affected by splice site strength-altering SNPs. A change in the mean SI of individuals of differing genotypes may suggest the possibility of altered splicing. A stepwise decrease (where the mean SI for the heterozygote is in between the two homozygotes) could reflect an increase in the allelic effect. There were 9328 HapMap-annotated SNPs within donor/acceptor regions of known exons which contained at least one probeset. Of 987 SNPs that are associated to exonic probesets which differ in mean SI between the homozygous common and rare HapMap individuals, 573 caused a decrease in natural site R_i. Leaky mutations (reduction in information content where final $R_i \geq R_{i,\text{minimum}}$) comprise 40-60% of the total and also exhibit reduced SI values. These results indicate that the proposed approach will detect severe, as well as moderate, splicing mutations with reduced penetrance and milder phenotypes, consistent with our previous reports (von Kodolitsch *et al.* 1999; von Kodolitsch *et al.* 2006).

These SNPs were analyzed by information theory to find those which caused a potential splice-affecting R_i change. Of the 9328 HapMap SNPs within the natural splice sites of exon probeset-containing exons, 112 (1.2%) and 235 (2.5%) were found on chromosome 21 and 22, respectively. Of those, 21 SNPs on chr21 (0.23% total, 18.8% of chr21) and 34 on chr22 (0.36%

of total, 14.5% of chr22) associated with a stepwise decrease in probeset intensity. 7 of the 21 chr21 SNPs (33.3%) and 9 of the 34 chr22 SNPs (26.5%) caused information changes which satisfied either of the following criteria: a natural site $\Delta R_i \geq 0.5$ bits, or a change in strength to a potential cryptic site(s) with an R_i comparable than the neighbouring natural site, or where mRNA/EST data supported cryptic site use. While a minimum ΔR_i of 0.5 bits (1.4 fold) was chosen, the actual minimum change resulting in a detectable splicing effect is not known (ΔR_i's range from 0.5 to 7.8 bit). The 16 SNPs are listed in order of decreasing ΔR_i: rs2075276 (*MGC16703*), rs2838010 (*FAM3B*), rs3747107 (*GUSBP11*), rs2070573 (*C21ORF2*), rs17002806 (*WBP2NL*), rs3950176 (*EMID1*), rs1018448 (*ARFGAP3*), rs6003906 (*DERL3*), rs2266988 (*PRAME*), rs2072049 (*PRAME*), rs2285141 (*CYB5R3*), rs2252576 (*BACE2*), rs16802 (*BCR*), rs17357592 (*COL6A2*), rs16994182 (*CLDN14*), *TMPRSS3* (*rs8130564*).

We report q-RT-PCR results for 13 out of the 16 SNPs (primers to test the affect of rs16994182, rs2075276 and rs3950176 were not suitable for q-RT-PCR, or gave ambiguous results), along with 8 other candidate SNPs found in previous publications (Nalla and Rogan, 2005) and are not found in chromosomes 21 and 22; (rs1805377 [*XRCC4*], rs2243187 [*IL19*], rs2835585 [*TTC3*], rs2865655 [*TTC3*], rs1893592 [*UBASH3A*], rs743920 [*EMID1*], rs13076750 [*LPP*] and rs1333973 [*IFI44L*] and rs10190751 [*CFLAR*]).

Limitations in detecting altered splicing with exon microarrays. Detection of splice forms is limited by the probeset placement of the array, as they often avoid small alternative splicing events (see UCSC Genome Browser image in Supplemental Figure 1.6, 1.7 and 1.9). Even where genotype-directed SI changes are very distinct, some individuals with the common allele have equivalent SI values individuals with the rare allele (*C21orf2* rs2070573; Suppl. Figure 1.1).

Limited availability of individuals with a particular genotype can be insufficient for statistical significance (rs2243187; Suppl. Figure 1.4). Although exon microarrays can be used to find potential alternate splicing and give support to our predictions, q-RT-PCR must be employed to confirm the splicing effect.

Accuracy of predictions. There were 22 total SNPs chosen for analysis by q-RT-PCR. Primers were developed to amplify known and information-theory predicted splice forms. 15 out of 22 SNPs tested showed a measurable change in splicing consistent with information-theory predictions. Of these 15 sites, 10 lead to an increase in alternate splice site use (2 of which increased strength of cryptic site, 8 increased use of unaffected pre-existing site), 6 lead to a change in exon retention (5 increased exon skipping), 3 which increased the use of an alternative exon, and 4 which appears to decrease total mRNA levels of that gene. We did not detect altered splicing in 6 SNPs, 2 of which caused information changes > 1 bit. Three of the four SNPs where $\Delta R_i < 1$ were hampered by high variability of gene expression between individuals.

Change in the information content of a splice site (a measure of binding affinity) was used to predict experimentally-derived change in splice isoform levels. In 12 out of the 15 SNPs which caused measurable effects, the change in splice site strength predicted by information theory were consistent with the changes measured by q-RT-PCR. The 3 exceptions are rs2070573 (*C21orf2*), rs17002806 (*WBP2NL*), and rs2835585 (*TTC3*). Fold changes predicted to reduce strength > 100 fold were experimentally found to reduce expression by 38 to 58 fold. Predicted changes in strength below 8 fold were not consistently detectable on wildtype splicing, though changes in less abundant splice forms were regularly observed (i.e. rs2835585 altered exon skipping levels by ~3-9 fold, but the wildtype splice form predominated).

Experimental results are shown in Table 1, and are briefly described in the following sections. For detailed descriptions for each individual SNP, refer to Supplemental Figure 1, which provides a summary of the predicted and experimentally-derived splicing effects for each SNP, a UCSC Genome Browser image of the relevant region, and boxplots showing exon microarray expression levels of each allele based on the relevant probesets.

SNPs affecting use of cryptic sites by weakening natural splice sites. When natural site strength is changed, the resulting mRNA splicing change depends on the strength and location of cryptic splice sites. We detected an increase in cryptic site use coinciding with a decrease in natural site strength caused by the following SNPs: rs1805377 (*XRCC4* exon 8 acceptor; 11.5 bits to 3.9 bits; 221-473 fold increase of 6nt downstream site detected by q-RT-PCR, complete discrimination in dual-labelled probe experiment); rs2243187 (*IL19* exon 5 acceptor; 7.3 to -0.3 bits; 1.8 fold increase of 3nt downstream site in heterozygote); rs3747107 (*GUSBP11* acceptor of 3' terminal exon of mRNA splice form BX538181; 8.9 to 1.4 bits; 31 and 42.8 fold increase of 114 and 118 nt upstream cryptic sites, respectively); rs17002806 (*WBP2NL* exon 6 donor; 10 to 6.5 bits; 34 fold increase of 25nt downstream site use); rs6003906 (*DERL3* exon 5 acceptor; 2.2 to 0.3 bits; double appearance of AK125830 mRNA splice form using a 123nt downstream acceptor); and rs13076750 (*LPP*, acceptor of rare exon within intron 1; 9.3 to -1.6 bits; 16 fold increase of 7nt downstream cryptic acceptor).

There are pre-existing cryptic sites near and with greater predicted information content of weakened natural splice sites which were not recognized or was not significantly altered in use; rs1893592 (*UBASH3A* exon 10 donor; 9.1 to 4.3 bits) did not increase the use of 7.0 and 6.1 bit sites 29 and 555nt downstream of affected donor; rs17002806 (*WBP2NL*, described earlier) did

not activate a 5.7 bit site 67nt downstream; rs3747107 (*GUSBP11*) strengthened a cryptic site 2nt downstream (1.6 to 7.5 bits) but no product was detected; rs2835585 (*TTC3* exon 3 acceptor; 6.4 to 4.4 bits) did not activate two stronger cryptic sites (6.9 and 7.2 bits, 60 and 87nt upstream respectively); It is clear that cryptic splice sites proximate to weakened natural splice site are not guaranteed to be activated and thus emphasizes the need for wet-lab experiments to confirm these bioinformatic predictions.

Mutations which affect the strength and use of cryptic Sites. An increase in the strength of a cryptic site may increase the use of that site. The following two SNPs were found to strengthen lesser used splice sites, increasing their use: rs2070573 (*C21orf2* exon 6 extended form donor [mRNA BC031300]; 2.4 to 5.7 bits [natural 10.4 bits]; 4-22 fold increase in use, increased use supported by exon microarray [Supplemental Figure 1.1]); rs743920 (*EMID1* exon 4, 6nt downstream of acceptor [mRNA AJ416090]; 10.5 to 12.4 bits [natural 6.4 bits]; 2.3-5.8 fold increase in site use). Despite the difference in strength (17 to 64 fold), the upstream 6.4 bit site was detected earlier by q-RT-PCR (3 cycles earlier or 8 fold in C/C individual). Recall both *IL19* and *XRCC4* regions tested showed preference to the upstream acceptor as well, which is congruent with the processive mechanism of detecting acceptor splice sites (Robberson *et al.* 1990).

SNPs affecting exon retention. SNPs-directed increases in exon skipping were found to reduce natural site strength from 1.6 to 10.9 bits, and lead to increases in exon skipping ranging from 3 to 1911 fold between homozygotes of opposing genotypes. These SNPs include: rs2835585 (*TTC3* exon 3 acceptor; 6.4 to 4.4 bits; 2-9 fold); rs1018448 (*ARFGAP3* exon 12 acceptor; 10.6 to 12.8 bits; 1.5-2.6 fold); rs1333973 (*IFI44L* exon 2 donor; 9.5 to 5.0 bits; ~15

fold); rs2266988 (*PRAME* exon 3 donor; 8.9 to 7.3 bits; 4.6-8.8 fold); rs13076750 (*LPP* exon 2a acceptor; 9.3 to -1.6 bits; 10.3-26.6 fold). Exon microarray probesets for *ARFGAP3* exon 12 (ID 3962628) and *IFI44L* exon 2 (probeset ID 2343480) show a decreasing trend suggesting increased exon skipping (Supplemental Figure 1.3 and 1.11, respectively). SI changes were not distinct for *TTC3* (ID 3920395/6) *PRAME* (ID 3954440) and *LPP* (ID 2657307) boxplots and may be masked by increased cryptic site use (*LPP*) or large differences in the abundance of wildtype and exon skipping splice forms (*PRAME*, *TTC3*).

SNPs which changed natural site strength but were not found to affect exon retention changed information content ranging from 3 to 4.8 bits. Skipping was not detected in weak homozygotes for rs1893592 (*UBASH3A*) and rs17002806 (*WBP2NL*), and was detected not altered by rs2835655 (*TTC3* exon 39 donor; 12 to 9 bits). The SNP rs2243187 (*IL19*, described earlier) was found to decrease exon skipping (halved in heterozygote) while increasing the use of an alternate 3nt downstream acceptor. This is consistent with Rogan *et al.* (2003) where the creation of a strong splice site closely situated to a natural site was shown to facilitate an increase in exon skipping, an effect that the A-allele eliminates.

SNPs which facilitate alternate exon use. Decreases in natural site strength may facilitate the use of alternative exons up or downstream of the affected exon. The *CFLAR* SNP rs10190751 (acceptor; 9.9 to 17.4 bits; A>C) modulates the presence of the shorter c-FLIP(S) splice form (Ueffing *et al.* 2009). The use of this exon differed by 2^{17} fold between homozygotes tested. The exon microarray reflected this result (probeset ID 2522648; Supplemental Figure 1.2). The *CFLAR* (L) form uses an alternate downstream exon, found to be 2.4 fold higher in the weak homozygote (microarray support not available due to SNP overlap, see Methods). Two other

SNPs showing a probable increased preference for an alternate exon are rs3747107 (*GUSBP11*; Suppl. Figure 1.8) and rs2285141 (*CYB5R3*; Suppl. Figure 1.20), though neither exon microarray boxplot reflected this increase.

Potential effects of SNPs on mRNA levels. A change in the strength of a natural site of an exon can affect the total quantity of the gene transcript. Of the 22 SNPs tested, 2 showed a direct correlation between a decrease in natural splice site strength, reduced amplification of the internal reference by q-RT-PCR (testing multiple individuals) and a general decreasing trend of total gene intensity (Core transcript clusters of exon microarray): rs2072049 (*PRAME* exon 6 acceptor; 10.7 to 9.4 bits; 32-57% total mRNA compared to homozygous wild type genotype); rs1018448 (*ARFGAP3*; 12.8 to 10.6 bits; 31.9-68.5% compared to homozygous wild type genotype); In each case, genotypic differences in the exon microarray data follow the expected trend but are not large enough to be statistically significant. Due to the modest differences in the array data, additional individuals must be tested to confirm these effects.

Predicted deleterious SNP without detectable evidence of alternate splicing. There were 6 SNPs in this study which were predicted to disrupt natural splice sites, but where there was no detectable effect on splicing. Potential causes include inter-individual expression variability, small (< 1 bit) strength changes, limitations to RT-PCR primer design, and the failure to correctly predict the SNP's splicing effect due to limitations of the splicing models (for example, compensatory splicing regulatory enhancers). Splicing effects were not identified for 3 SNPs where the information change was <1 bit (2-fold). Genetic variability masked potential splicing effects of these SNPs: rs16802 (*BCR* exon 14 acceptor; 8.8 to 9.4 bits; individuals of like genotype varied by 8.3 fold), rs2252576 (of *BACE2* exon 5 acceptor; 9.0 to 9.6 bits; 12 fold) and

rs8130564 (*TMPRSS3* exon 5 acceptor; 6.3 to 6.8 bits; 112 fold). Various primer sets designed to detect an exon affected by the SNP rs17357592 where $\Delta R_i \leq 1$ bit (*COL6A2* exon 21 acceptor; 8.4 to 7.8 bits) failed to give a single product. Interpreting the results from the SNP rs16994182 (*CLDN14* exon 2 donor; 8.6 to 8.1 bits) was complicated by the lack of an adequate internal reference primer set. As this gene consists of only 3 exons, any internal reference must cover the potentially affected second exon. Any difference detected by these primers could be caused by altered splicing, and therefore cannot account for any variation in expression between individuals. Therefore, the splicing differences detected by q-RT-PCR (Table 1) are not conclusive.

In one case, a large R_i change failed predictions. The donor of a rarely used exon in IVS1 of *FAM3B* (mRNA AJ409094) is strengthened by rs2838010 (1.4 to 9.2 bits, A>T). This cryptic exon was not detected in any individual tested (including T/T). The microarray data supports this result (probeset ID 3922001; Supplemental Figure 1.16). Although it is likely that the T allele is required for the inclusion of this exon, additional factors (such as tissue type) are apparently also involved.

Discussion

Predicted deleterious SNP alleles that alter constitutive mRNA splicing are confirmed by expression and spliced EST data, and may be common in populations. The preponderance of leaky splicing mutations and cryptic splice sites, which often produce both normal and mutant transcripts, is consistent with balancing selection (Nuzhdin *et al.* 2004) or possibly with mutant loci that contribute to multifactorial disease. Minor SNP alleles are often present in > 1% of populations (Janosíková *et al.* 2005). This would be consistent with a bias against finding

mutations that abolish splice site recognition in dbSNP. Such mutations are more typical in rare Mendelian disorders (Rogan *et al.* 1998).

We note that work described in this manuscript was performed several years ago (ca. 2009), This study used the ASSA server to evaluate each mutation. It has been replaced by the ASSEDA (Automated Splice Site and Exon Definition Analyses; Mucaki *et al.* 2013) server which is now part of the MutationForecaster (http://www.mutationforecaster.com) interpretation system. The exon expression microarrays and quantitative RT-PCR results described confirm more recently developed approaches such as high-throughput transcriptome sequencing technology, RNAseq, in our laboratory and others (Viner *et al.* 2014; Dorman *et al.* 2014, Shirley *et al.* 2018).

The ValidSpliceMut web-beacon (http://validsplicemut.cytognomix.com) is a splicing mutation variant database containing predicted and confirmed splice variants from the Cancer Genome Atlas (TCGA). These have been identified by the Shannon Pipeline (a high-throughput IT-based prediction tool based on ASSEDA), and validated by RNA-Seq data from matched tissues and tumors lacking these mutations (Shirley *et al.* 2018). The database reports whether a mutation was associated with an increase in exon skipping, cryptic site use, or intron inclusion RNAseq reads. Of the variants tested in this study, 8 were found in ValidSpliceMut (rs2070573 [*C21orf2*], rs10190751 [*CFLAR*], rs13076750 [*LPP*], rs2072049 [*PRAME*], rs2835585 [*TTC3*], rs2835655 [*TTC3*], rs1893592 [*UBASH3A*] and rs1805377 [*XRCC4*]. Variants rs1893592, rs2070573, rs13076750 and rs10190751 were flagged due to intron inclusion, which is supported by q-RT-PCR results. An independent study focused on cryptic site activating mutations with the

same TCGA data (Jayasinghe *et al.* 2018) failed to identify rs2070573 as a splicing mutation (which strengthens a cryptic splice site 360nt downstream of exon 6 in *C21orf2*). Furthermore, neither study flagged two other cryptic site-strengthening SNPs described in this manuscript (rs743920 [*EMID1*] and rs2838010 [*FAM3B*]). Interestingly, rs2835585 and rs1805377 were flagged due to intron inclusion, however q-RT-PCR experiments instead showed increased exon skipping and cryptic site use, respectively. Aberrant splicing was not was detected experimentally for rs2835655 or rs2072049, however these SNPs were flagged due to increased intron inclusion. The design of the q-RT-PCR experiments associated with these SNPs were not optimized to detect this form of abnormal splicing.

The splicing impact (and when known, the disease-association) of many of the discussed SNPs have been implicated subsequent to the development of this study. As previously mentioned, the *CFLAR* SNP rs10190751 is known to modulate the FLICE-inhibitory protein (c-FLIP) from its S-form to its R-form, and the latter form has been linked to increased lymphoma risk (Ueffing *et al.* 2009). Furthermore, increased exon skipping due to the *IFI44L* SNP rs1333973 has been reported in RNAseq experiments (Zhao *et al.* 2013a) and this alternate splice form has been implicated in a reduction in antibody response to the measles vaccine (Haralambieva *et al.* 2017). The splicing impact of *XRCC4* rs1805377 has been noted previously (Nalla and Rogan 2005), but to our knowledge has not previously been experimentally confirmed. This SNP has been implicated with an increased risk of gastric cancer (Chiu *et al.* 2008), pancreatic cancer (Ding and Li, 2015) and risk of gliomas (Zhao *et al.* 2013b). Similarly, the potential impact of *UBASH3A* SNP rs1893592 has been previously recognized (Kim *et al.* 2015) but not tested, and has been associated with arthritis (Liu *et al.* 2017) and type 1 diabetes

(Ge and Concannon, 2018). Hiller *et al.* (2006) described the 3nt deletion caused by *IL19* rs2243187, but did not describe the increase in exon skipping seen in this study. The SNP rs743920 (*EMID1*) was associated with change in allelic expression using EST data (Ge *et al.* 2005; impact on splicing not described). Conversely, studies which linked *TMPRSS3* variants to hearing loss did not find the SNP rs8130564 to be significant (Lee *et al.* 2013; Chung *et al.* 2014). Interestingly, the *BACE2* SNP rs2252576 (which was not found to alter splicing in this study) has been associated to Alzheimer's dementia in Down syndrome (Mok *et al.* 2014).

The compound effects of splicing mutations have been previously described. Krawczak *et al.* selected 38 genes known to have single-nucleotide mutations within donor and acceptor sites (from HGMD, as of January 2006) which have been associated with various diseases (Krawczak *et al.* 2007). Using neural networks, 87.4% of the mutations found in splice sites (n=430) were reported to cause exon skipping or cryptic site use. Of these splice-altering mutations (n=376), 56.9% were mutations in donor sites causing exon skipping, 13.6% resulted in the use of a cryptic donor site, 22.3% were acceptor site mutations leading exon skipping and 7.2% lead to cryptic acceptor use. Their data also suggested the possibility that exon skipping is less likely in the presence of nearby cryptic sites when a donor is weakened, but not acceptors. A region of only 50 bp surrounding the affected splice-sites was used to search for cryptic sites, and therefore there is a strong possibility that sites outside of this range may have been missed.

Why are so few natural splice sites strengthened by SNP-induced information changes? Most such changes would be thought to be neutral mutations, which are ultimately lost by chance (Fisher 1930). Those variants which are retained are more likely to confer a selective advantage (Li 1967). Indeed, the minor allele in rs2266988, which strengthens a donor splice site by 2.3 bits

(29.9 fold) at the 5' end of the open reading frame in *PRAME*, occurs in 25% of the population (~50% in Europeans). We have shown a number of instances where apparently simple changes in strength of splice sites that would be expected to have little or no impact on splicing of the associated exon in fact alters the degree of exon skipping of that exon.

SNPs producing significant changes in information at functional binding sites may be useful for selecting tag SNPs in disease association studies. Such an approach would be independent of measures of haplotype block (Zhang *et al.* 2002; Carlson *et al.* 2004; Pe'er and Beckmann 2004) or genotype diversity (Zhang *et al.* 2004), or proximity or correlations between neighboring SNPs. If SNP induced changes in individual information are associated with clinical phenotypes (because of their impact gene expression levels and transcript structure), the selection of these variants as tag SNPs should maximize statistical significance of the corresponding disease association. At a minimum, Bayesian strategies making use of this data may further refine of positions of candidate disease susceptibility loci.

Considering the number of constitutive splicing mutations found, it is unlikely that sequence variation alone can account for the extensive heterogeneity in mRNA transcript structures, given the relatively high proportion of genes known to exhibit tissue-specific alternative splicing (Modreck and Lee 2002). Nevertheless, this study raises questions regarding the degree to which alternative splicing is the result of inter-individual genomic sequence differences rather than purely regulatory mechanisms. Because much of the information required for splice site recognition resides within neighboring introns, it would be prudent to consider contributions from intronic and exonic polymorphism that produce structural exon variation.

While exon microarrays can be used to show alternate splicing differences based on

genotype, there are obvious limitations with this technology. Probesets were placed to detect wildtype splice forms with mRNA and EST support, but may not be adequately placed to detect rare splicing events that have not yet been reported or which have little evidence. Smaller nucleotide changes due to cryptic site use (*XRCC4*, *EMID1*) seemed to be explicitly avoided in these probesets, which could not detect the splicing change. Indeed, significant effort was required to design TaqMan assays that distinguished the isoforms generated by rs1805377 (*XRCC4*). While some genotype-specific differences in SI were quite significant (*CFLAR*, *IFI44L*), many showed only minor changes (*ARFGAP3, LPP*) and most had outlier individuals of one genotype with a comparable SI to the population of the second genotype (*IL19*).

rs2835585 increased exon skipping in *TTC3* nearly 10 fold but a resulting decrease in total wildtype splicing at the affected exon junction was undetectable, most likely due to the great difference in abundance between the wildtype and skipped splice isoforms. Whether or not this small increase would cross the threshold of allele-specific exon skipping that may contribute to disease predisposition and pathogenesis is in question. In the case of the E3 ubiquitin ligase, *TTC3*, the third exon does not include any definitively-assigned protein domain (Tsukahara *et al.* 1996, Suizu *et al.* 2009).

This study describes the prediction of validation of natural and cryptic splice site alterations caused by common SNPs. Individual information represents a continuous phenotypic measure that is well suited to the analysis of contributions of multiple, incompletely penetrant SNPs in different genes from the same individual, as typically seen in genetically complex diseases. This technique has complemented efforts to identify disease-associated protein coding (and non-coding) mutations in a comprehensive, high-throughput variant interpretation study of

breast cancer patient gDNA (Caminsky *et al.* 2016; Mucaki *et al.* 2016).

The Veridical and Shannon pipeline software has been developed to perform and validated large-scale analysis of potential splicing variants in complete genomes using RNAseq data (Viner *et al.* 2014). These resources have been used to evaluate millions of variants in TCGA cancer patient genomes (Shirley *et al.* 2018). For the SNPs described in this paper, targeted functional splicing analyses, for the most part, reproduce the results of our multi-genome-wide surveys of sequence variations affecting mRNA splicing. This concordance increases confidence that these publicly (https://ValidSpliceMut.cytognomix.com) and commercially (https://MutationForecaster.com) available resources can help to identify mutations contributing to patient clinical phenotypes.

Acknowledgements

P.K. Rogan acknowledges support from The Natural Sciences and Engineering Research Council of Canada (NSERC) [Discover Grant 371758-09 and RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs, and CytoGnomix.

Table 1: Summary of q-RT-PCR Results

Summary of q-RT-PCR Results					SNP Effects (Increase/Decrease in fold change of homozygotes)					
Gene	**rsID**	**Splice Type**	**Information Change (ΔRi) (n –natural, c– cryptic site)**	**Fold Change**	**Natural Site**	**Cryptic Site**	**Exon Skipping**	**Alternate Exon**	**Total mRNA**	**In-Frame**
XRCC4	1805377	A	11.5 (G) -> 3.9 (A) (n)	194	38.4	47.3	-	-	n/a[b]	Y
		A	11.3 (G) -> 11.4 (A) (n)	1.1						
IL19	2243187	A	7.3 (G) -> -0.3 (A) (n)	194	1.8[a]	1.8[a]	2.1[a]	-	n/a[b]	Y
C21orf2	2070573	D	2.4 (A) -> 5.7 (C) (c)	9.9	NC	22.6	-	-	n/a[b]	Y
TTC3	2835655	D	12.0 (G) -> 9.0 (A) (n)	8.0	1.5	-	1.3[c]	-	1.5	-
TTC3[d]	2835585	A	6.4 (T) -> 4.4 (A) (n)	4.0	NC	-	8.8	-	n/a[b]	Y
WBP2NL	17002806	D	9.2 (G) -> 6.0 (A) (n)	9.2	23.8	34+[f]	-	-	n/a[b]	N
GUSBP11	3747107	A	8.9 (C) -> 1.4 (G) (n)	181	98.3	31/42.8[j]	-	1.6	n/a[b]	Y/N[e]
PRAME[g]	2266988	D	8.9 (G) -> 7.3 (A) (n)	3.0	NC	-	8.8	-	n/a[b]	N
PRAME[g]	2072049	A	10.7 (G) -> 9.4 (T) (n)	2.5	2.6[a]	-	-	-	3.1[a]	-
UBASH3A	1893592	D	9.1 (A) -> 4.3 (C) (n)	27.9	3.0	2.0/1.7[k]	N/A[i]	-	1.4	Y
DERL3	6003906	A	2.2 (A) -> 0.3 (T) (n)	3.7	NC	2.0	-	-	n/a[b]	N
ARFGAP3	1018448	A	12.8 (C) -> 10.6 (A) (n)	4.6	2.2	-	1.4	-	2.0	Y

CFLAR	10190751	A	17.4 (G) -> 9.9 (A) (n)	181	100000+	-	-	2.1	n/a[b]	Y
IFI44L	1333973	D	9.1 (A) -> 4.6 (T) (n)	22.6	15.4	-	15.6	-	3.9	N
LPP	13076750	A	9.3 (G) -> -1.6 (A) (n)	1910	5000+	15.8	26.6	-	-	Y
EMID1	743920	A	10.5 (C) -> 12.4 (G) (c)	3.7	n/a[b]	5.8	-	-	n/a[b]	-
CLDN14	16994182	D	8.6 (C) -> 8.1 (G) (n)	1.4	5.9	-	2.1	-	n/a[b]	Y
BCR	16802	A	8.8 (A) -> 9.4 (G) (n)	1.5	NC[h]	-	-	-	n/a[b]	-
TMPRSS3	8130564	A	6.3 (T) -> 6.8 (C) (n)	1.4	NC[h]	-	-	-	n/a[b]	-
BACE2	2252576	A	9.0 (C) -> 9.6 (T) (n)	1.5	NC[h]	-	-	-	n/a[b]	-
CYB5R3	2285141	A	6.1 (G) -> 5.0 (T) (n)	2.1	NC	-	-	1.8	n/a[b]	-
FAM3B	2838010	D	1.4 (A) -> 9.2 (T) (c)	223.0	-	-	-	N/A[i]	n/a[b]	Y

Red text indicates a decrease in the abundance of a particular splice form, while green text indicates an increase in abundance. A – Acceptor Splice Site Affected; D – Donor Splice Site Affected; NC - Not detectable (abolished). [a] Values from comparing heterozygote with homozygote common. [b] No allele specific difference in expression and splicing. [c] Change in splicing likely related to change in RNA level. [d] Intron 2-3 inclusion of *TTC3* amplified by PCR, but no allele specific change detected. [e] mRNA in-frame when alternate exon is used, and out of frame due to cryptic site use. [f] This splice form not at detectable levels in homozygote. [g] *PRAME* is a special case where two SNPs affect splicing of two separate exons. [h] High variation between individuals of the same genotype found by q-RT-PCR. [i] Splice form not detected by PCR. [j] Cryptic acceptor 114nt upstream of affected site / cryptic acceptor 118nt upstream of affected site. [k] Cryptic donor 555nt downstream of affected site / cryptic donor 29nt downstream of affected site.

References

Affymetrix Papers: **Exon Probeset Annotations and Transcript Cluster Groupings v1.0; Exon Array Background Correction v1.0; Guide to Probe Logarithmic Intensity Error (PLIER) Estimation; Alternative Transcript Analysis Methods for Exon Arrays v1.1.** [https://www.affymetrix.com/support/technical/whitepapers.affx]

Allikmets R, Wasserman WW, Hutchinson A, Smallwood P, Nathans J, Rogan PK, Schneider TD, Dean M. **Organization of the A*BCR* gene: analysis of promoter and splice junction sequences.** *Gene*. 1998 Jul 17;215(1):111-22.

Ars E, Kruyer H, Morell M, Pros E, Serra E, Ravella A, Estivill X, Lázaro C. **Recurrent mutations in the NF1 gene are common among neurofibromatosis type 1 patients.** *J Med Genet.* 2003 Jun;40(6):e82.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res.* 2009 Jan;37(Database issue):D885-90. Epub 2008 Oct 21.

Berget SM. **Exon recognition in vertebrate splicing.** J Biol Chem. 1995 Feb 10;270(6):2411-4.

Bi C, Rogan PK. **Bipartite pattern discovery by entropy minimization-based multiple local alignment.** *Nucleic Acids Res.* 2004 Sep 23;32(17):4979-91.

Birney E, Kumar S, Krainer AR. **Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors.** *Nucleic Acids Res.* 1993 Dec 25;21(25):5803-16.

Black DL. **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem.*

2003;72:291-336. Epub 2003 Feb 27.

Buratti E, Baralle M, Baralle FE. **Defective splicing, disease and therapy: searching for master checkpoints in exon definition.** *Nucleic Acids Res.* 2006 Jul 19;34(12):3494-510.

Cáceres JF, Stamm S, Helfman DM, Krainer AR. **Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors**. *Science.* 1994 Sep 16;265(5179):1706-9.

Caminsky NG, Mucaki EJ, Perri AM, Lu R, Knoll JH, Rogan PK. **Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known BRCA Mutations.** *Hum Mutat.* 2016 Jul;37(7):640-52.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *Am J Hum Genet.* 2004 Jan;74(1):106-20.

Carothers AM, Urlaub G, Grunberger D, Chasin LA. **Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells.** *Mol Cell Biol.* 1993 Aug;13(8):5085-98.

Chiu CF, Wang CH, Wang CL, Lin CC, Hsu NY, Weng JR, Bau DT. **A novel single nucleotide polymorphism in *XRCC4* gene is associated with gastric cancer susceptibility in Taiwan.** *Ann Surg Oncol.* 2008 Feb;15(2):514-8.

Chung J, Park SM, Chang SO, Chung T, Lee KY, Kim AR, Park JH, Kim V, Park WY, Oh SH, Kim D, Park WJ, Choi BY. **A novel mutation of TMPRSS3 related to milder auditory phenotype in Korean postlingual deafness: a possible future implication for a personalized auditory rehabilitation.** *J Mol Med (Berl).* 2014 Jun;92(6):651-63.

Dietz HC, Valle D, Francomano CA, Kendzior RJ Jr, Pyeritz RE, Cutting GR. **The skipping of**

constitutive exons in vivo induced by nonsense mutations. *Science.* 1993 Jan 29;259(5095):680-3.

Ding Y, Li LN. **Association between single nucleotide polymorphisms of X-ray repair cross-complementing protein 4 gene and development of pancreatic cancer.** *Genet Mol Res.* 2015 Aug 14;14(3):9626-32.

Domenjoud L, Kister L, Gallinaro H, Jacob M. **Selection between a natural and a cryptic 5' splice site: a kinetic study of the effect of upstream exon sequences.** *Gene Expr.* 1993;3(1):83-94.

Dorman SN, Viner C, Rogan PK: **Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer.** *Sci Rep.* 2014; **4**: 7063.

Fisher, R.A. 1930. **The genetical theory of natural selection.** Clarendon, New York.

Fu XD, Maniatis T. **The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site.** *Proc Natl Acad Sci U S A.* 1992 Mar 1;89(5):1725-9.

Ge B, Gurd S, Gaudin T, Dore C, Lepage P, Harmsen E, Hudson TJ, Pastinen T. **Survey of allelic expression using EST mining.** *Genome Res.* 2005 Nov;15(11):1584-91.

Ge Y, Concannon P. **Molecular-genetic characterization of common, noncoding UBASH3A variants associated with type 1 diabetes.** *Eur J Hum Genet.* 2018 Jul;26(7):1060-1064.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res.* 2005 Oct;15(10):1451-5. Epub 2005 Sep 16.

Haralambieva IH, Ovsyannikova IG, Kennedy RB, Larrabee BR, Zimmermann MT, Grill DE,

Schaid DJ, Poland GA. **Genome-wide associations of CD46 and IFI44L genetic variants with neutralizing antibody response to measles vaccine.** *Hum Genet.* 2017 Apr;136(4):421-435.

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. **Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing.** *Am J Hum Genet.* 2006 Feb;78(2):291-302.

Hobson GM, Huang Z, Sperle K, Sistermans E, Rogan PK, Garbern JY, Kolodny E, Naidu S, Cambi F. **Splice-site contribution in alternative splicing of PLP1 and DM20: molecular studies in oligodendrocytes.** *Hum Mutat.* 2006 Jan;27(1):69-77.

Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME. **A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity.** *Proc Natl Acad Sci USA.* 2007, Jun 5;104(23):9758-63.

Janosíková B, Zavadáková P, Kozich V. **Single-nucleotide polymorphisms in genes relating to homocysteine metabolism: how applicable are public SNP databases to a typical European population?** *Eur J Hum Genet.* 2005 Jan;13(1):86-95.

Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, Zhao Y, Climente-González H, Chai S, Wang F, et al. **Systematic Analysis of Splice-Site-Creating Mutations in Cancer.** *Cell Rep.* 2018 Apr 3;23(1):270-281.e3.

Kannabiran C, Rogan PK, Olmos L, Basti S, Rao GN, Kaiser-Kupfer M, Hejtmancik JF. **Autosomal dominant zonular cataract with sutural opacities is associated with a splice mutation in the betaA3/A1-crystallin gene.** *Mol Vis.* 1998 Oct 23;4:21.

Khan SG, Levy HL, Legerski R, Quackenbush E, Reardon JT, Emmert S, Sancar A, Li L,

Schneider TD, Cleaver JE, Kraemer KH. **Xeroderma pigmentosum group C splice mutation associated with autism and hypoglycinemia.** *J Invest Dermatol.* 1998 Nov;111(5):791-6.

Khan SG, Muniz-Medina V, Shahlavi T, Baker CC, Inui H, Ueda T, Emmert S, Schneider TD, Kraemer KH. **The human XPC DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function.** *Nucleic Acids Res.* 2002 Aug 15;30(16):3624-31.

Khan SG, Metin A, Gozukara E, Inui H, Shahlavi T, Muniz-Medina V, Baker CC, Ueda T, Aiken JR, Schneider TD, Kraemer KH. **Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk.** *Hum Mol Genet.* 2004 Feb 1;13(3):343-52.

Kim K, et al. **High-density genotyping of immune loci in Koreans and Europeans identifies eight new rheumatoid arthritis risk loci.** *Ann Rheum Dis.* 2015 Mar;74(3):e13.

Krawczak M, Reiss J, Cooper DN. **The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences.** *Hum Genet.* 1992 Sep-Oct;90(1-2):41-54.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. **Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing.** *Hum Mutat.* 2007 Feb;28(2):150-8.

Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J. **Heritability of alternative splicing in the human genome.** *Genome Res.* 2007 Aug;17(8):1210-8.

Lamba V, Lamba J, Yasuda K, Strom S, Davila J, Hancock ML, Fackenthal JD, Rogan PK, Ring

B, Wrighton SA, Schuetz EG. **Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression.** *J Pharmacol Exp Ther.* 2003 Dec;307(3):906-22.

Langmead B, Trapnell C, Pop M, Salzberg SL. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol.* 2009;10(3):R25.

Lee J, Baek JI, Choi JY, Kim UK, Lee SH, Lee KY. **Genetic analysis of TMPRSS3 gene in the Korean population with autosomal recessive nonsyndromic hearing loss.** *Gene*. 2013 Dec 15;532(2):276-80.

Li, C. C., 1967 Genetic **equilibrium under selection**. *Biometrics*. **23:** 397-484.

Liu D, Liu J, Cui G, Yang H, Cao T, Wang L. **Evaluation of the association of UBASH3A and SYNGR1 with rheumatoid arthritis and disease activity and severity in Han Chinese.** *Oncotarget*. 2017 Oct 17;8(61):103385-103392.

López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. **Are splicing mutations the most frequent cause of hereditary disease?** *FEBS Lett.* 2005 Mar 28;579(9):1900-3.

Margulis V, Lin J, Yang H, Wang W, Wood CG, Wu X. **Genetic susceptibility to renal cell carcinoma: the role of DNA double-strand break repair pathway.** *Cancer Epidemiol Biomarkers Prev.* 2008 Sep;17(9):2366-73.

Mayeda A, Krainer AR. **Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2.** *Cell*. 1992 Jan 24;68(2):365-75.

Modreck B, Lee C. **A genomic view of alternative splicing.** *Nat Genet.* 2002. 1: 13-19.

Molecular Diagnostic Laboratory, Department of Clinical Biochemistry, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark.

Mok KY, Jones EL, Hanney M, Harold D, Sims R, Williams J, Ballard C, Hardy J. **Polymorphisms in BACE2 may affect the age of onset Alzheimer's dementia in Down syndrome.** *Neurobiol Aging.* 2014 Jun;35(6):1513.e1-5.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature.* 2010 Apr 1;464(7289):773-7.

Moore MJ, Sharp PA. **Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing.** *Nature*. 1993 Sep 23;365(6444):364-8.

Mucaki EJ, Shirley BC, Rogan PK. **Prediction of mutant mRNA splice isoforms by information theory-based exon definition.** *Hum Mutat.* 2013; 34(4): 557–565.

Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JH, Rogan PK. **A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer.** *BMC Med Genomics.* 2016 Apr 11;9:19.

Nalla VK, Rogan PK. **Automated splicing mutation analysis by information theory.** *Hum Mutat.* 2005 Apr;25(4):334-42.

Nelson KK, Green MR. **Mechanism for cryptic splice site activation during pre-mRNA splicing.** *Proc Natl Acad Sci U S A.* 1990 Aug;87(16):6253-7.

Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C. **Genome-wide survey of allele-specific splicing in humans.** *BMC Genomics.* 2008 Jun 2;9:265.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. **Common pattern of evolution of gene expression level and protein sequence in Drosophila.** *Mol Biol Evol.* 2004 Jul;21(7):1308-17. Epub 2004 Mar 19.

O'Neill JP, Rogan PK, Cariello N, Nicklas JA. **Mutations that alter RNA splicing of the human HPRT gene: a review of the spectrum.** *Mutat Res.* 1998 Nov;411(3):179-214.

Parker R, Siliciano PG, Guthrie C. **Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA.** *Cell*. 1987 Apr 24;49(2):229-39.

Pe'er I, Beckmann JS. **Recovering frequencies of known haplotype blocks from single-nucleotide polymorphism allele frequencies.** *Genetics*. 2004 Apr;166(4):2001-6.

Pfaffl MW, Horgan GW, Dempfle L. **Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR.** *Nucleic Acids Res.* 2002 May 1;30(9):e36.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature*. 2010 Apr 1;464(7289):768-72.

Richard I, Beckmann JS. **How neutral are synonymous codon mutations?** *Nat Genet.* 1995 Jul;10(3):259.

Robberson BL, Cote GJ, Berget SM. **Exon definition may facilitate splice site selection in RNAs with multiple exons.** *Mol Cell Biol.* 1990;10(1):84–94. 10.1128/MCB.10.1.84.

Rogan PK, Schneider TD. **Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites.** *Hum Mutat.* 1995;6(1):74-6.

Rogan PK, Faux BM, Schneider TD. **Information analysis of human splice site mutations.** *Hum Mutat.* 1998. 12:153-171.

Rogan PK, Svojanovsky SR, Leeder JS. **Information theory-based analysis of CYP219,**

CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics*. 2003. 13(4): 207-218.

Schneider TD. **Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences.** *Nucleic Acids Res.* 1997. 25:4408-4415.

Séraphin B, Kretzner L, Rosbash M. **A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site.** *EMBO J.* 1988 Aug;7(8):2533-8.

Shirley BC, Mucaki EJ and Rogan PK. **Pan-cancer repository of validated natural and cryptic mRNA splicing mutations** [version 1; referees: 1 approved, 1 approved with reservations]. *F1000Research* 2018, 7:1908.

Staknis D, Reed R. **SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex.** *Mol Cell Biol.* 1994 Nov;14(11):7670-82.

Stephens RM, Schneider TD. **Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites.** *J Mol Biol.* 1992 Dec 20;228(4):1124-36.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science.* 2007 Feb 9;315(5813):848-53.

Suizu F, Hiramuki Y, Okumura F, Matsuda M, Okumura AJ, Hirata N, Narita M, Kohno T, Yokota J, Bohgaki M, Obuse C, Hatakeyama S, Obata T, Noguchi M. **The E3 ligase *TTC3* facilitates ubiquitination and degradation of phosphorylated Akt.** *Dev Cell.* 2009 Dec;17(6):800-10.

Susani L, Pangrazio A, Sobacchi C, Taranta A, Mortier G, Savarirayan R, Villa A, Orchard P, Vezzoni P, Albertini A, Frattini A, Pagani F. **TCIRG1-dependent recessive osteopetrosis: mutation analysis, functional identification of the splicing defects, and in vitro rescue by U1 snRNA**. *Hum Mutat.* 2004 Sep;24(3):225-35.

Svojanovsky SR, Schneider TD, Rogan PK. **Redundant designations of BRCA1 intron 11 splicing mutation; c. 4216-2A>G; IVS11-2A>G; L78833, 37698, A>G.** *Hum Mutat.* 2000 Sep;16(3):264.

Talerico M, Berget SM. **Effect of 5' splice site mutations on splicing of the preceding intron.** *Mol Cell Biol.* 1990 Dec;10(12):6299-305.

Teraoka SN, Telatar M, Becker-Catania S, Liang T, Onengüt S, Tolun A, Chessa L, Sanal O, Bernatowska E, Gatti RA, Concannon P. **Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences.** *Am J Hum Genet.* 1999 Jun;64(6):1617-31.

Thorsen K, Sørensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhøffer M, Laurberg S, Borre M, Wang K, Brunak S, Krainer AR, Tørring N, Dyrskjøt L, Andersen CL, Orntoft TF. **Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis.** *Mol Cell Proteomics.* 2008 Jul;7(7):1214-24.

Trapnell C, Pachter L, Salzberg SL. T**opHat: discovering splice junctions with RNA-Seq.** *Bioinformatics.* 2009 May 1;25(9):1105-11.

Tsukahara F, Hattori M, Muraki T, Sakaki Y. **Identification and cloning of a novel cDNA belonging to tetratricopeptide repeat gene family from Down syndrome-critical region 21q22.2.** *J Biochem.* 1996 Oct;120(4):820-7.

Ueffing N, Singh KK, Christians A, Thorns C, Feller AC, Nagl F, Fend F, Heikaus S, Marx A, Zotz RB, Brade J, Schulz WA, Schulze-Osthoff K, Schmitz I, Schwerk C. **A single nucleotide polymorphism determines protein isoform production of the human c-FLIP protein.** *Blood.* 2009 Jul 16;114(3):572-9.

Viner C, Dorman SN, Shirley BC, *et al.*: **Validation of predicted mRNA splicing mutations using high-throughput transcriptome data** [version 2; referees: 4 approved]. *F1000Res.* 2014; **3**: 8.

Vockley J, Rogan PK, Anderson BD, Willard J, Seelan RS, Smith DI, Liu W. **Exon skipping in IVD RNA processing in isovaleric acidemia caused by point mutations in the coding region of the IVD gene.** *Am J Hum Genet.* 2000 Feb;66(2):356-67.

von Kodolitsch Y, Pyeritz RE, Rogan PK. **Splice-site mutations in atherosclerosis candidate genes: relating individual information to phenotype.** *Circulation*. 1999 Aug 17;100(7):693-9.

von Kodolitsch Y, Berger J, Rogan PK. **Predicting severity of haemophilia A and B splicing mutations by information analysis.** *Haemophilia*. 2006 May;12(3):258-62.

Vyhlidal CA, Rogan PK, Leeder JS. **Development and refinement of pregnane X receptor (PXR) DNA binding site model using information theory: insights into PXR-mediated gene regulation.** *J Biol Chem.* 2004 Nov 5;279(45):46779-86.

Wu J, Manley JL. **Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing.** *Genes Dev.* 1989 Oct;3(10):1553-61.

Wu X, Gu J, Grossman HB, Amos CI, Etzel C, Huang M, Zhang Q, Millikan RE, Lerner S, Dinney CP, Spitz MR. **Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes.** *Am J Hum Genet.* 2006 Mar;78(3):464-79. Epub 2006 Jan 31.

Yates T, Okoniewski MJ, Miller CJ. **X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis.** *Nucleic Acids Res.* 2008 Jan;36(Database issue):D780-6. Epub 2007 Oct 11.

Zhang K, Calabrese P, Nordborg M, Sun F**. Haplotype block structure and its applications to association studies: power and study designs.** *Am J Hum Genet.* 2002 Dec;71(6):1386-94. Epub 2002 Nov 18.

Zhang W, Collins A, Morton NE. **Does haplotype diversity predict power for association mapping of disease susceptibility?** *Hum Genet.* 2004 Jul;115(2):157-64. Epub 2004 Jun 4.

Zhao L, Lu Z, Park JW, Zhou Q, Xing Y. **GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data.** *Genome Biol.* 2013; 14(7): R74. (a)

Zhao P, Zou P, Zhao L, Yan W, Kang C, Jiang T, You Y. **Genetic polymorphisms of DNA double-strand break repair pathway genes and glioma susceptibility.** *BMC Cancer*. 2013 May 10;13:234. (b)

Zhuang Y, Weiner AM. **A compensatory base change in U1 snRNA suppresses a 5' splice site mutation.** *Cell*. 1986 Sep 12;46(6):827-35.

Zuo P, Maniatis T. **The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing.** *Genes Dev.* 1996 Jun 1;10(11):1356-68.