2012

Identification and Characterization of Site-Directed A-to-I RNA Editing Targets

Christina Priska Godfried Sie
Lehigh University

Identification and Characterization of Site-Directed A-to-I RNA Editing Targets

by

Christina Priska Godfried Sie

A Dissertation

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

the Department of Biological Sciences

Lehigh University

January 2012

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Christina Priska Godfried Sie

Identification and Characterization of Site-Directed A-to-I RNA Editing Targets

September 30th, 2011

Defense Date

October 28th, 2011

Approved Date

Dr. Michael Kuchka
Dissertation Director

Committee Members:

Dr. Mark R. Macbeth

Dr. Linda J. Lowe-Krentz

Dr. Robert Skibbens

Acknowledgements

First I want to thank Dr. Michael Kuchka, my advisor and mentor. By an unusual twist of events I was so fortunate to come to fully profit from your professionalism, the whole breadth and depth of your knowledge, and your hopeless and contagious optimism. You stepped forward when I was most in need of support and advice and selflessly committed much of your time and many resources to help me through the intense final year of my thesis work, immersing yourself whole-heartedly into the strange world of RNA editing. Thank you for everything.

Willemijn, the most valuable outcome of this endeavor is your friendship, which makes everything worthwhile. I could have done a PhD in other places, but knowing what I do now I would still choose this path again, even if meeting you were the only reason. More than anyone else you sharpened my scientific mind by relentlessly asking critical questions and making pointed comments. Above all, I value your honesty on both the professional and personal level. As one of the hardest working people I know, you have set the bar high indeed at which I aim to perform. True friends are hard to find, I am very lucky.

If I have done one thing right in the last five years, it was choosing my committee. All of you have provided me with precious advice and guided me towards the successful completion of my doctorate degree. Dr. Lowe-Krentz, thank you for answering all of my questions anytime and anywhere, astounding me over again with your incredible wealth of knowledge that you have acquired over the years. I will always remember your classes as some of the most challenging and interesting I have ever had. Dr. Macbeth, thank you

for all the dearly needed expert feedback and encouragement throughout the process. Dr. Bob Skibbens, what can I say, if there is one thing I regret, it is ever setting a foot out of your lab. Thank you for always believing in me, challenging me, and thank you for taking me back on!

Many special thanks go to Mike Kearse for being such a courteous and reliable colleague and friend. You are always there to answer any of my questions, ponder with me over results, devise devious strategies and let me vent without restraint. I have great respect for you as scientist and admire your work ethics and tireless efforts toward excellence. I feel lucky to have been working next-door to you, which has allowed me to learn a great deal from you. Do not think that just because I am almost done here and you are almost out of the door, my constant pestering you with questions and asking for your opinion will simply go away…

I must thank Stefan Maas for accepting me into his lab and providing me with the resources needed to conduct my research. Thanks also to present and former members of the Maas lab, who have accompanied me along the way: Emaan and Jessica, who helped me analyze REDS candidates; Nicholas, Nikki, Verena, and Marina, who started and helped me with the FLNA project; Steve Hesler, who helped me tremendously getting the proteolysis experiments up and running; Sanaz Farajollahi for bringing some light into the lab with her good nature; and finally Dylan Dupuis with whom I have shared the ups and downs of the lab over the years.

They say it takes a village to raise a child, and in my case it is certainly true that it took a department to help me achieve my goals. Many thanks to the faculty members who have so nobly committed their time, advice, and resources to a fledgling scientist outside

of their own labs. The words of encouragement as well as scientific and career advices from Dr. Lynne Cassimeris, Dr. Vassie Ware, Dr. Mike Burger and Dr. Barry Bean may not have seemed much to you at the time, but it meant everything for me those days. Thanks to Dr. Bill Coleman for providing me with mice, Dr. Kathy Iovine for zebrafish specimen, Dr. Samollow from the Texas A&M University for the opossum tissues, Dr. Lynne Cassimeris and Dr. Vassie Ware for antibodies, and Dr. Robert Skibbens for the restriction enzymes. I am very grateful to Dr. Murray Itzkowitz for pledging the support of the department when I was most desperate – it really made life easier for me. I will also be forever grateful for the department's trust and confidence in my work by awarding me the Nemes Fellowship. Finally, Maria Brace, Lee Graham, Vicki Waldron, Heather Sohara, Carol Esposito, and Joann Deppert, who have kept and continue to keep the departmental wheel running ever so smoothly, deserve more than a thank you. We all know that without your diligent work we scientists would be hopelessly lost.

I regret not having recognized earlier how important the support of graduate students by graduate students is. Better late than never, I would like to let the graduate student body at this department know how much I appreciate their advice on life in general and in these hallways in particular and how much I value our animated discussions on how things could or should be. We have wandered the same path as many before us, and there will be countless to come after us, but to walk it together with you guys has been a real pleasure. The future is ours, and one day I hope some of us will remember how things could or should have been and strive to make a difference for the generations after us.

Over the years I have been privileged to get to know a number of fine people. My childhood friend Manuela Duxenneuner has been tremendously supportive and my connection to the “Heimat” in times when I thought no one over there can possibly know or understand me anymore. Thank you for being such a strong and sympathetic anchor. Former graduate students have been instrumental in guiding me with invaluable advice during my first timid years here at Lehigh. I am glad to have met Meron Mengistu, who has inspired me not only with her extraordinary thesis work, but also with her level of professionalism, lack of fear to break unknown grounds, contagious enthusiasm and tranquil confidence. Thanks also to Anna Gumpert and Marie Maradeo, who have given me many good suggestions especially on how to deal with grad student life. Victoria Caruso, Jeremy Brozek, Joe Leese, Josh Slee, Andrew Black – thanks for not only talking the talk but also walking the walk, so often brightening up my days and completely understanding when I was stressed out and unreasonably grumpy. I know you will all be doing great, wherever the future may take you.

This small thank-you note would be incomplete without acknowledging the many great teachers and advisors I have been blessed with during my life before Lehigh. Some of my early teachers are to be credited with infusing me with a curiosity for the life sciences. My professors at the Swiss Federal Institute of Technology, especially Professor Amrhein and Professor Andreas Schaller, have been instrumental in building the basis of my scientific understanding. Dr. Rupert Hagg and Dr. Roberto Tommasini sought to instill professionalism and work ethics in a freshly baked M.Sc. with a lot of patience.

Tiefste Dankbarkeit empfinde ich für meine Eltern, deren Liebe und Güte keine Grenzen kennt. Ihr seid die besten Vorbilder, die sich eine Tochter vorstellen kann. Eure kontinuierliche Unterstützung in allem, was ich mir zum Ziel setze, ermöglicht es mir erst, den Weg in Angriff zu nehmen. Obwohl mich meine Träume so weit von Euch weggeführt haben, fühle ich mich Euch näher als je zuvor. Eure unerschöpfliche Liebe zeigt sich mir in allem, was Ihr sagt und tut. Ich hoffe, dass die Art und Weise, wie ich mein Leben führe, Euch mit Genugtuung erfüllt; wissend, dass ich meine Werte und Lebensansichten ganz Euch verdanke.

And last but certainly not least, I do not know the words how to properly thank my husband and companion Osvaldo who has been walking by my side, supporting me and putting my feet firmly back onto the ground whenever I needed it. You never allowed me to forget what is really important in life, and for that I love you. Realizing how our bond has grown stronger with each challenge it faced, understanding that our lives are now fully intertwined and combined, I look to our future calmly and joyfully.

So many hands have reached out to help me advance one step or keep me from falling off a precipice. So many friends, old and new, never stopped believing in me when doubts were creeping into my mind. So many obstacles have served me well by showing me my limits. It is a privilege to be able to do the work that we do and I am thankful to have had the opportunity to walk this path, honored by all the support, and mystified by Nature's countless remaining secrets.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

$5HT_{2c}R$	Serotonin receptor
A	Adenosine (in nucleic acid context), Alanine (in amino acid context)
ADAR	Adenosine deaminase acting on RNA
A-to-I	Adenosine-to-Inosine
ATP6VOE2	ATPase, H+ transporting V0 subunit e2
bp	Base-pair
BC10	Bladder cancer-associated protein
C	Cytidine
C1QL1	Complement component 1, q subcomponent-like 1
CDK13	Cyclin dependent kinase 13
cDNA	Complementary DNA
CNS	Central nervous system
COPA	Coatomer protein complex subunit alpha
CRD	Cysteine-rich domain
DGCR8	DiGeorge syndrome critical region gene 8
DNA	Deoxyribonucleic acid
DRBM	Double-stranded RNA binding motif
ds	Double-stranded
dsRNA	Double-stranded RNA
ECS	Editing complementary sequence
EST	Expressed sequence tag

FLNA	Filamin A
G	Guanosine
gDNA	Genomic DNA
GluR	Glutamate receptor
I	Inosine
IGFBP7	Insulin-like growth factor binding protein 7
K	Lysine
kDa	Kilo Dalton
LINE	Long interspersed nuclear element
miRISC	MicroRNA-induced silencing complex
miRNA	Micro-RNA
mRNA	Messenger RNA
ncRNA	non-coding RNA
NEIL1	Endonuclease VIII-like 1
NES	Nuclear export signal
NLS	Nuclear localization signal
NoLS	Nucleolar localization signal
nt	Nucleotide(s)
PCR	Polymerase chain reaction
Q	Glutamine
R	Arginine
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction

SerpinA3	Serpin peptidase inhibitor clade A, member 3
SINE	Short interspersed nuclear element
SN	Staphylococcal nuclease
SNP	Single nucleotide polymorphism
ss	Single-stranded
ssRNA	Single-stranded RNA
T	Thymidine (in nucleic acid context), Threonine (in amino acid context)
TENR	Testis nuclear RNA binding protein
TRBP	Trans-activation-responsive RNA-binding protein
Tudor-SN	Tudor staphylococcal nuclease
U	Uracil
UTR	Untranslated region

Abstract

Posttranscriptional processes increase protein diversity by expanding the information content of the transcriptome and thus contribute to the tremendous complexity of higher organisms. One widespread posttranscriptional modification is the selective deamination of adenosines (A) to inosines (I) in RNA. In protein-coding transcripts, the edited codon can alter the protein's amino acid sequence, which quite often has critical consequences on protein stability, localization, and/or functional properties. The production of several protein isoforms in the same cell increases proteome diversity from a limited number of genes. This recoding process is necessary for the proper development and physiological functions of higher eukaryotes, underscoring the importance of editing in higher organisms. The goal of this thesis was to identify previously unknown A-to-I RNA recoding events and characterize their implications for the alternative protein variants.

Only few protein-coding RNA editing sites were known in 2006, most of which had been identified by chance. However, their often dramatic impact on protein function warranted a genome-wide search strategy to systematically identify previously unknown A-to-I RNA editing recoding targets and evaluate their prevalence in the human transcriptome. Here, we experimentally validate several novel A-to-I RNA recoding targets identified by bioinformatics screening methods as potential candidate sites. From these, we selected several targets to assess consequences of the elicited amino acid changes on protein function using molecular and protein biology methods tailored to the respective targets. We show that A-to-I RNA editing has significant and biologically

relevant consequences for at least one protein. Finally, we analyzed the regulatory aspects of A-to-I RNA editing in a specific target that is promiscuously edited in the pre-messenger RNA, but shows no editing in the spliced, mature RNA. This thesis work underscores the importance of post-transcriptional A-to-I RNA editing on proteome diversity in higher organisms and uncovers a previously undescribed mechanism of regulation that opens up a new area of research.

1 Introduction

1.1 RNA versatility

RNA may well be the most underrated biological molecule of the 20th century. Apart from being extensively studied for its central roles in spliceosomes and ribosomes during transcriptional processing and translation, respectively, it has by and large led a wallflower existence in the wider research community. For a long time, the main function of RNA was thought to be serving as intermediary between DNA and protein, facilitating the portability and translation of genetic information. However, this view has been revised in recent years and the prominence of RNA, particularly within regulatory networks, has started to surface (Mercer et al., 2009; Hung & Chang, 2010). In fact, RNA is extraordinarily well suited to serve as regulatory molecule, since it can provide both exquisite one-dimensional sequence-specificity and versatile tertiary structures that can either be recognized by proteins or have catalytic activity themselves (Hung & Chang, 2010). RNA therefore resembles a converter of digital to analog signals, directly bridging two worlds, that of informational content stored in DNA and the functional output embodied by proteins (Kohler et al., 1993; Mattick, 2004; St Laurent & Wahlestedt, 2007). Classic examples of RNAs with these properties are tRNAs, whose tertiary structures interact selectively with both aminoacyl tRNA synthetases and ribosomes, while specifically base-pairing to the respective codons of an mRNA. The versatility of functional RNAs, such as ribozymes, enables them to catalyze enzymatic reactions and allows for certain long non-coding RNAs (ncRNAs) to nucleate and propagate epigenetic silencing across chromatin (Forster & Symons, 1987; de la Pena & Garcia-Robles, 2010; Hung & Chang, 2010).

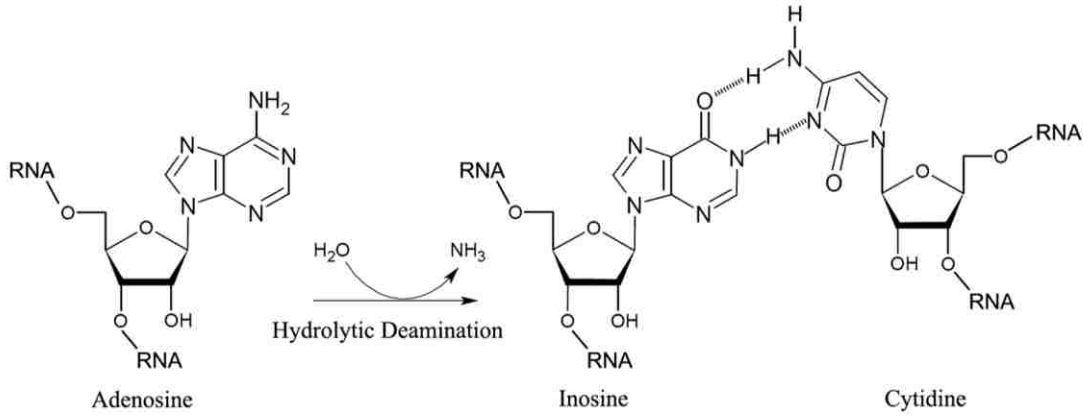


Figure 1: Adenosine-to-inosine RNA editing

Hydrolytic deamination of adenosine yields inosine. Inosine forms Watson–Crick base-pairing with cytidine, analogous to guanosine and inosine appears as a guanosine in cloned cDNA. Figure from Godfried Sie & Kuchka, 2011.

Another layer of versatility is added by the hydrolytic deamination of adenosine to inosine (Figure 1), or A-to-I RNA editing, of nuclear-encoded RNAs of metazoa, which enables a cell to change genetic information at the post-transcriptional level (Bass, 2002). Since inosine base-pairs with cytosine, this deamination effectively results in a change in sequence with potential consequences for downstream events, as most enzymes recognize inosine as guanosine (Bass, 2002). For instance, if this sequence change takes place within the coding region of a pre-mRNA, it can result in a codon change which subsequently leads to production of a protein with a different amino acid sequence (Figure 2).

The first of a number of accounts of such RNA editing recoding events was reported twenty years ago (Sommer et al., 1991). Fast excitatory glutamate receptor channels in the brain were found to contain either an arginine or a glutamine residue at a critical position in the channel-forming segment, with profound effects on ion flow. Intriguingly, the genomic sequence only encoded a glutamine, and thus the research

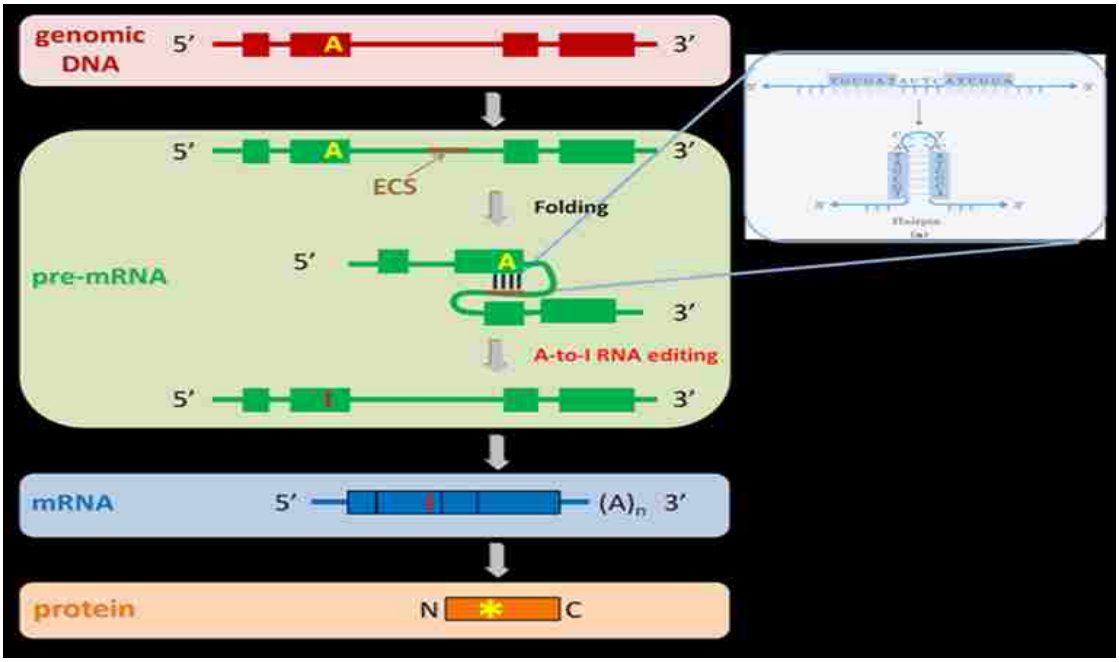


Figure 2: Schematic representation of the flow of genetic information and RNA editing

Double-stranded structures of a certain length formed by RNA (shown here is pre-mRNA) are potential substrates for the enzymes mediating A-to-I RNA editing, the ADAR family of proteins.

group proposed the existence of an RNA editing machinery that modifies the glutamine to an arginine codon. This was confirmed when a previously identified protein with 'double-stranded (ds) RNA unwinding activity' (Bass & Weintraub, 1987) was found to have deaminase activity and was therefore named adenosine deaminase acting on RNA (ADAR1) (Kim & Nishikura, 1993). A homologue was later cloned (Melcher et al., 1996), and termed ADAR2. ADAR3 and TENR (testis nuclear RNA binding protein) complete the family in humans. However, ADAR3, which is solely expressed in the brain, and TENR, expressed exclusively in testis and essential for proper spermatogenesis (Connolly et al., 2005), do not seem to have deaminase activity (Chen et al., 2000).

1.2 ADARs

1.2.1 ADAR family members are essential for higher organisms

All types of RNA transcripts can be subject to modification by A-to-I editing, which converts adenosine (A) to inosine (I) through hydrolytic deamination in double-stranded RNA (dsRNA). Inosine preferentially base-pairs with cytidine and is thus interpreted as guanosine in most cases (Figure 1). Consequently, editing provides a means by which cells can manipulate primary sequence readouts as well as higher order RNA structures. Editing is catalyzed by members of the ADAR family of proteins, which are found in most metazoans, but have not been detected in plants, yeast, and fungi (Jin et al., 2009). There are four ADAR family members in mammals (ADAR1-3 and TENR = testis nuclear RNA binding protein), one in *Drosophila* (dADAR) and two in *Caenorhabditis elegans* (adr-1 and adr-2). ADAR deficiencies lead to a wide range of phenotypes, particularly with effects on functions of the central nervous system, which underlines their importance for proper development and behavior in different species. The dADAR$^{-/-}$ *Drosophila* show incapacitated coordination of locomotion and abnormal behavior (Palladino et al., 2000). Homozygous deletion of ADARs in *C. elegans* results in defective chemotaxis (Tonkin et al., 2002). ADAR2$^{-/-}$ mice die within three weeks *post natum*, after repeated episodes of epileptic seizures (Higuchi et al., 2000), and ADAR1$^{-/-}$ mice have an embryonically lethal phenotype associated with liver disintegration and defects in hematopoiesis (Wang et al., 2000; Hartner et al., 2004; Hartner et al., 2009; XuFeng et al., 2009).

1.2.2 ADAR domain structure

All ADARs share a similar domain structure, with one to three double-stranded RNA binding motifs (DRBMs) and a C-terminal deaminase domain (Figure 4). However, of the four mammalian ADAR family members, only ADAR1 and ADAR2 have a catalytically active deaminase, while ADAR3 and TENR (also called ADAD1 = adenosine deaminase domain containing 1) are presumed to be deamination deficient as catalytically critical residues are not conserved and the enzymes are unable to modify any of the known substrates (Lai et al., 1995; Mian et al., 1998; Chen et al., 2000). Sometimes ADAR1 and ADAR2 have overlapping target-specificity, but more often an adenosine within a given sequence is primarily edited by one or the other (Lehmann & Bass, 2000; Wong et al., 2001). Preliminary evidence suggested that this target-specificity can mostly be credited to the respective deaminase domains (Wong et al., 2001). However, editing specificity is mediated to a considerable extent also by the DRBMs, as shown by hydroxyl radical footprinting (Yi-Brunozzi et al., 2001; Stephens et al., 2004) and recently confirmed by the crystal structure of DRBMs 1 and 2 of ADAR2 bound to a pre-mRNA target (Stefl et al., 2010). There are at least two types of ADAR substrates – dsRNAs of 50 or more base pairs that are promiscuously edited, and partially double-stranded RNAs with loops and bulges that are edited at specific adenosines only (Lehmann & Bass, 1999; Bass, 2002). In the former, nonspecific editing occurs because ADARs bind long dsRNA via their DRBMs without sequence preference. The presence of loops and bulges in the dsRNA, conversely, mediates the defined positioning of the active site on the substrate through sequence-specific interactions of the DRBMs and provides for precise targeting of one or few adenosines within the

tertiary RNA structure (Enstero et al., 2009; Stefl et al., 2010). Thus, DRBMs and active sites unite to promote exquisitely controlled editing in some contexts, while allowing promiscuous editing in others.

Both deaminase domains and DRBMs contribute to preferences for certain nucleotides that adjoin the edited adenosine. Such nearest-neighbor preferences were first observed in editing of Alu repeat elements (Athanasiadis et al., 2004) and were subsequently analyzed in a more quantitative manner (Lehmann & Bass, 2000; Dawson et al., 2004; Eggington et al., 2011). Nearest-neighbor preferences of human ADAR1 and ADAR2 are shown in Figure 3.

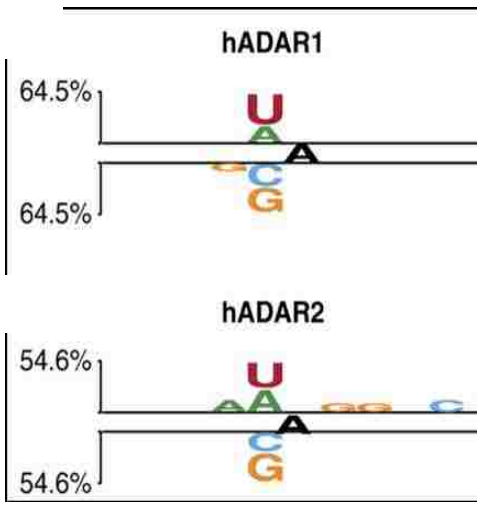


Figure 3: Nearest-neighbor preferences of human ADAR1 and ADAR2
Schematic displaying enriched bases above top line and depleted bases below bottom line on both sides of the central edited adenosine determined among 406 adenosines. Level of enrichment/depletion shown by letter heights with reference to scale on the left. Figure taken and adapted from Eggington et al., 2011.

1.2.3 Regulation of ADARs

Our understanding of how ADAR activity is regulated is still incomplete, but new findings continue to paint an increasingly complex picture. ADARs are expressed in a large number of tissues but editing is most abundant in the central nervous system (CNS) (Paul & Bass, 1998). ADAR expression and editing activity are regulated in a tight spatio-temporal manner and generally increase during development (Rula et al., 2008;

Jacobs et al., 2009; Wahlstedt et al., 2009). On the other hand, a decrease of editing is seen in specific subsets of ADAR targets during differentiation of human embryonic stem cells into the neural lineage (Osenberg et al., 2010). Several lines of evidence thus point toward complex and precise regulation of editing.

Drosophila expresses different ADAR isoforms through alternative splicing of one ADAR transcript. Substrate dsRNA induces cooperative interaction between two dADAR monomers to promote dimerization and heterodimerization of different dADAR isoforms, which affects editing efficiency (Gallo et al., 2003). Homodimerization of ADAR1 or ADAR2 in human is also required for enzymatic activity (Cho et al., 2003), but heterodimerization between ADAR1 and ADAR2 was not detected. However, this analysis did not take into account heterodimerization between alternative splice isoforms of ADAR1 and ADAR2, respectively (Figure 4), creating a possible regulatory mechanism that remains to be explored.

Transcription from alternative promoters leads to the expression of variant ADAR isoforms, as indicated in Figure 4. Aside from allowing regulation of protein levels in response to external stimuli, transcription from alternative promoters can result in different translation start sites and produce proteins with additional domains and/or untranslated regions (UTRs). ADAR1, for example, exists in two isoforms. The constitutive ADAR1p110 harbors a nuclear localization signal (NLS) within the DRBM3 and is thus mostly nuclear (Eckmann et al., 2001). In contrast, ADAR1p150 is transcribed from an interferon-inducible promoter (George & Samuel, 1999b, a; George et al., 2008) and not only contains an additional Z-DNA binding domain, but also a nuclear export signal (NES), which allows its nucleocytoplasmic shuttling (Eckmann et

al., 2001; Poulsen et al., 2001; Strehblow et al., 2002; Fritz et al., 2009). Of course, both alterations in functional domains and cellular localization allow ADARs to access different sets of substrates, and so it is believed that the main targets of ADAR1p150 are dsRNA from viral infections present in the cytoplasm (Samuel, 2011). Furthermore,

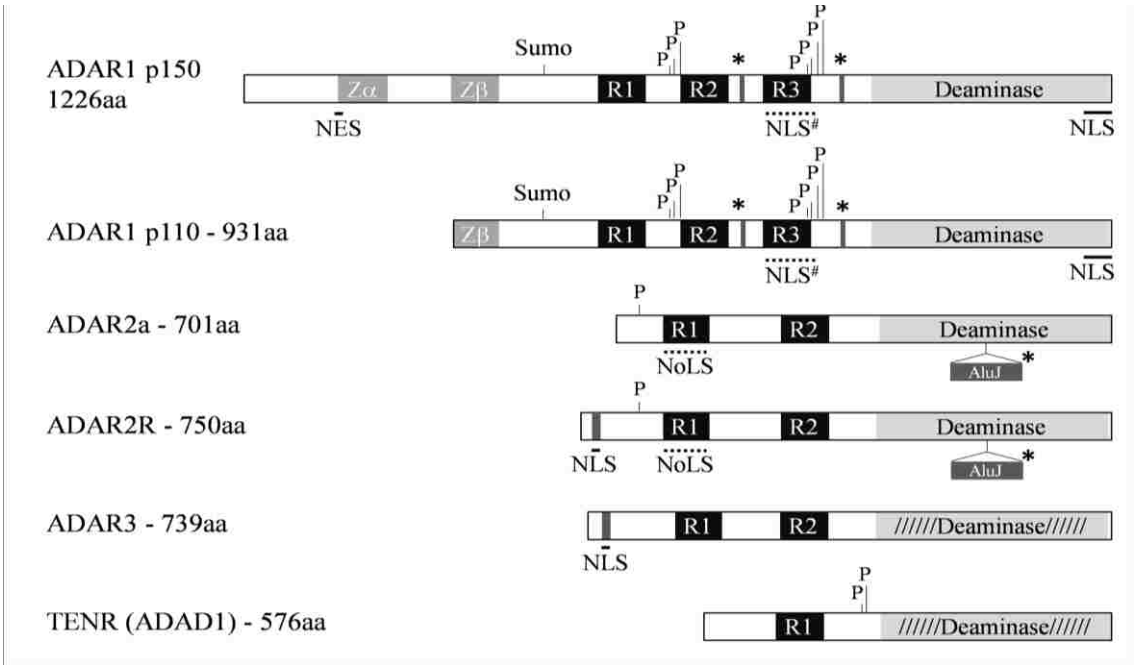


Figure 4: Adenosine Deaminases acting on RNA family of proteins

Hydrolytic deamination of adenosines is catalyzed by the ADAR family of proteins. ADARs are probably regulated on many different levels such as use of alternative promoters, alternative splicing, regulation of sub-cellular localization, and post-transcriptional modifications. ADAR2 is known to edit its own transcripts in a negative feedback loop. The deaminase domains of ADAR3 and TENR are believed to be non-functional as positions important for catalytic activity are not conserved (Chen et al., 2000; Saunders & Barber, 2003). For more detailed information, see text. Abbreviations: Zα and Zβ, Z-DNA binding domains; R1-R3, double-stranded RNA binding motifs; NES, nuclear export signal; NLS, nuclear localization signal; NoLS, nucleolar localization signal; Sumo, sumoylation site; P, phosphorylation site; AluJ, AluJ alternative cassette; NLS# overlaps with NoLS. Figure from Godfried Sie & Kuchka, 2011.

ADAR1 dynamically associates with the nucleolus due to the presence of a nucleolar localization signal (NoLS), which overlaps with the NLS (Desterro et al., 2003).

Besides subcellular localization, ADAR1 is likely regulated on a number of additional levels as well. Alternative splicing of ADAR1 leads to differential spacing between DRBMs and the deaminase domains (marked with an asterisk * in Figure 4), potentially altering target specificity and/or activity of homodimers. Post-translational modifications also play a role as sumoylation of K418 reduces editing activity, possibly due to stereochemical inhibition of dimerization (Desterro et al., 2005). And finally, ADAR1 has several phosphorylation sites, but their functional relevance is currently unknown.

ADAR2 is apparently also subjected to regulation on multiple levels, which is not yet fully understood. Two autoinhibitory mechanisms allow partial regulation of enzyme activity. First, the N-terminal DRBM ensures binding of ADAR2 to dsRNA of a certain length by inhibiting binding to shorter dsRNAs that cannot accommodate interactions with both DRBMs (Macbeth et al., 2004). Second, ADAR2 editing of its own pre-mRNA creates a 3′-acceptor site, adding a 47 nucleotide insertion to the 5’ end of the second coding exon. This leads to a frameshift and the creation of a premature termination codon, providing for a negative autoregulatory mechanism (Rueter et al., 1999). Additional regulatory mechanisms include alternative splicing, which results in the insertion of an AluJ cassette into the deaminase domain, reducing enzyme activity approximately two-fold (Gerber et al., 1997). Use of an alternative promoter extends the N-terminus with a NLS, which may also function as single-stranded (ss)RNA binding domain (Maas & Gommans, 2009). ADAR2 is primarily nuclear and like ADAR1

dynamically associates with the nucleolus due to the presence of a NoLS (Desterro et al., 2003). One phosphorylation site with unknown function is present. Considering the controlled spatio-temporal regulation, tissue-specificity, and responsiveness to environmental cues, ADAR activity must undergo tight regulation.

1.3 Consequences of editing

1.3.1 Pre-mRNA editing in coding sequences

As mentioned earlier, site-specific editing in coding regions was first discovered in brain-specific mRNAs, where the resulting codon changes lead to the expression of protein isoforms with altered amino acid sequences, as inosine is interpreted as guanosine by the translational machinery (Sommer et al., 1991; Burns et al., 1997; Berg et al., 2001). One editing site of the glutamate receptor 2 (GluR2) changes a glutamine to an arginine codon (Q/R site), which renders the channel Ca^{2+}-impermeable (Sommer et al., 1991). This is also the only site known to date to undergo 100% editing in mammals. In fact, the lethal ADAR2$^{-/-}$ phenotype has been entirely ascribed to the lack of editing at the GluR2 Q/R site (Higuchi et al., 2000). Another thoroughly studied editing target is the G-protein coupled serotonin receptor ($5HT_{2c}R$), where editing occurs at five adenosines that change the coding of three amino acids. These are located in the second intracellular loop, important for receptor activity, and the combinatorial editing of these five adenosines alters G-protein interaction, agonist affinity, and receptor trafficking (Burns et al., 1997; Herrick-Davis et al., 1999; Berg et al., 2001; Marion et al., 2004).

Glutamate and serotonin receptor pre-mRNA editing have been extensively studied, but much still needs to be learned about the regulation and consequences of editing in these complex targets. The profound effects of editing on the functions of these two exemplary frontrunners have inspired the intensive search for additional recoding sites (Clutterbuck et al., 2005; Eisenberg et al., 2005a; Levanon et al., 2005; Gommans et al., 2008; Li et al., 2009).

The most recent recoding target for which a functional consequence has been established is NEIL1. NEIL1 is a DNA repair enzyme that catalyzes the cleavage of oxidized base lesions (David et al., 2007). It can remove a wide array of modified DNA bases that arise from oxidative stress during inflammation, radiation, or after exposure to toxic agents and endogenous metabolic activity (Neeley & Essigmann, 2006). NEIL1 was found to be edited, changing codon 242 from lysine (K) to arginine (R) (Li et al., 2009). The recoding site is located in its lesion recognition loop. Indeed, the two variants of the NEIL1 protein have distinct enzymatic properties with changes observed for both glycosylase activity and lesion specificity (Yeo et al., 2010). Most notably, NEIL1 mRNA recoding is regulated by interferon through upregulation of ADAR1, suggesting a unique regulatory mechanism for DNA repair.

Only few examples of edited targets that produce an alteration in protein function are known (Table 1). A comprehensive list of all currently known recoding targets has recently been discussed by Pullirsch and Jantsch (Pullirsch & Jantsch, 2010). Recoding editing sites that are edited to high levels are relatively rare. In many cases, however, editing seems to occur at low levels. This speculation was based on limited dataset analyses, and has recently been more strongly substantiated by deep-sequencing results (Clutterbuck et al., 2005; Levanon et al., 2005; Gommans, 2008; Li et al., 2009; Pullirsch & Jantsch, 2010). It has been proposed that editing of many targets at low levels may provide a mechanism for the continuous probing of potentially advantageous editing events, without compromising the genomic information content (Gommans et al., 2009). Such a mechanism would manifest itself in the observed low levels of editing in many transcripts. The fluctuating nature of pre-mRNA secondary structure may allow such

Table 1: Editing events leading to codon changes

List of editing events that lead to codon changes and where a functional impact has been determined. Table from Godfried Sie & Kuchka, 2011.

Gene name	Function	AA change	Functional impact of editing	References
GluR-2	Glutamate-gated ion channel subunit 2	Q606R, R763G	Decreased Ca^{2+} permeability (Q606R); faster recovery from desensitization (R763G)	(Sommer et al., 1991; Lomeli et al., 1994)
GluR-3	Glutamate-gated ion channel subunit 3	R775G	Faster recovery from desensitization	(Lomeli et al., 1994)
GluR-4	Glutamate-gated ion channel subunit 4	R765G	Faster recovery from desensitization	(Lomeli et al., 1994)
GluR-5	Glutamate-gated ion channel subunit 5	Q621R	Variation in ion permeability	(Sommer et al., 1991)
GluR-6	Glutamate-gated ion channel subunit 6	I567V, Y571C, Q621R	Q621R: increased Ca^{2+} permeability if I/V and Y/C are edited.	(Sommer et al., 1991)
$5HT_{2c}R$	G-protein coupled serotonin receptor	I156V/M, N158S/D/G, I160V	Altered G-protein coupling, agonist affinity, receptor trafficking	(Burns et al., 1997)
KCNA1	Voltage-gated potassium channel	I400V	Increased recovery rate from inactivation	(Bhalla et al., 2004)
Gabra-3	γ-aminobutyric acid gated chloride channel subunit	I342M	Altered receptor sensitivity and deactivation rate	(Rula et al., 2008; Nimmich et al., 2009)
ADAR2	Adenosine deaminase acting on RNA	creates 3' splice acceptor site in intron	Frameshift, premature termination codon, negative feedback	(Rueter et al., 1999)
NEIL1	DNA repair enzyme	K242R	Changes lesion specificity	(Li et al., 2009; Yeo et al., 2010)

continuous probing of potentially beneficial editing sites. Accordingly, an editing event could become engraved if it conferred an adaptive advantage under a given selection pressure, and only then would be edited at relatively high levels (Gommans et al., 2009).

The observed tissue-specificity (He et al., 2011), developmental regulation (Wahlstedt et al., 2009) and responsiveness to certain environmental cues (Gan et al., 2006; Yeo et al., 2010) may be a further impediment to arrive at a better appreciation of the prevalence of recoding editing events that have functional consequences for the resulting protein variant(s). However, the impact of editing on the regulation and functionality of known targets calls for efforts to uncover novel recoding editing sites on a genome-wide scale. The vastness of the transcribed genome makes such an undertaking the literal equivalent of searching for a needle in the haystack and thus necessitates a systematic and well-directed approach. Our laboratory employed such a systematic search and was able to find new editing recoding sites, as discussed in Chapters 2 and 3.

1.3.2 Pre-mRNA editing in non-coding sequences

While the number of high-level, site-selective recoding editing sites appears to be relatively low, bioinformatic screenings of human sequence databases have revealed that most (>85%) pre-mRNAs are edited at least at one position. However, in primates the bulk of editing takes place in Alu repeat elements within introns and untranslated regions (UTRs) (Athanasiadis et al., 2004; Blow et al., 2004; Kim et al., 2004; Levanon et al., 2004). Alu elements belong to the short interspersed nuclear elements (SINE) family of transposable elements and comprise about 10% of the human genome (International Human Genome Sequencing Consortium, 2004). They are approximately 300 nucleotides long and are found in gene rich regions. Due to their high copy number, they are often present multiple times in one transcript, sometimes in opposite orientation to one another. Alu elements have a conserved sequence owing to their relatively recent expansion in the primate genome (Batzer & Deininger, 2002). Therefore, transcribed oppositely oriented

Alu elements form long dsRNA structures and are potent substrates for ADARs and are edited promiscuously, i.e. at multiple adenosines within the dsRNA structure. Alu elements appear to be evolutionarily important for primates, especially in conjunction with A-to-I RNA editing (Eisenberg et al., 2005b; Hasler & Strub, 2006; Hasler et al., 2007; Chen & Carmichael, 2008; Mattick & Mehler, 2008). This notion has been supported by the finding that editing can lead to the exonization of an Alu element by changing an AA dinucleotide into an AG 3′ acceptor splice site (Lev-Maor et al., 2007; Moller-Krull et al., 2008).

RNA editing in introns and UTRs also appears to affect transcripts in other ways, such as retention in the nucleus, altered stability, and inhibition of transcription and translation, although conflicting results foretell multi-layered regulatory mechanisms involved in determining the fate of inosine-containing pre-mRNAs (Zhang & Carmichael, 2001; Prasanth et al., 2005; Scadden & Smith, 2001; Chen et al., 2008; Scadden, 2007; Chen et al., 2008; Hundley & Bass, 2010; Morse & Bass, 1999; Hundley et al., 2008).

Evidently, A-to-I RNA editing has been shown to effectuate a wide range of consequences for pre-mRNAs, and it is expected that additional functions remain to be discovered. Figure 5 summarizes known and potential ways A-to-I RNA editing may influence pre-mRNA fate.

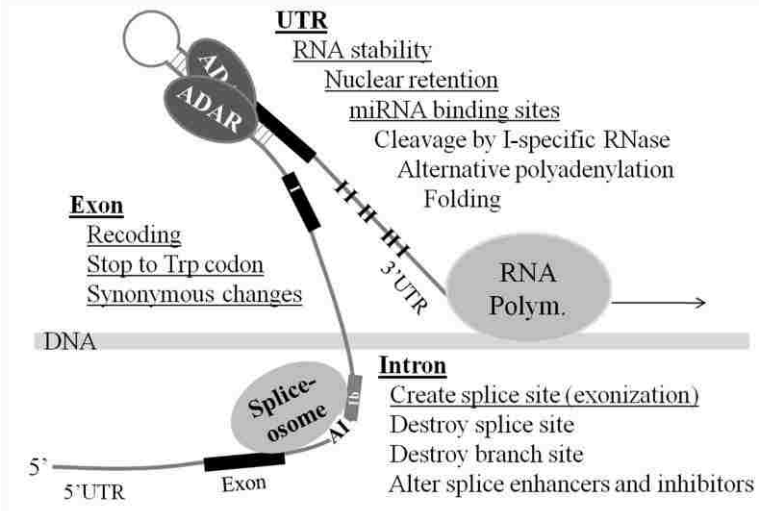


Figure 5: Possible effects of editing on pre-mRNA

Effects which have been demonstrated in vivo *are underlined (for references see text), while the other indicated consequences are hypothetical (Godfried Sie & Kuchka, 2011).*

1.4 Editing and the microRNA pathway, a lesson about impact of editing on RNA fate

MicroRNAs (miRNAs) are a class of small non-coding RNAs of approximately 22 nucleotide length. A single miRNA can regulate the expression of dozens or even hundreds of genes by repressing translation or promoting transcript destabilization (Djuranovic et al., 2011). MiRNAs have important roles in development, differentiation, proliferation, and apoptosis. The miRNA database [www.mirbase.org (Griffiths-Jones et al., 2008)] currently reports 1424 miRNAs in human, which together regulate thousands of protein-coding genes (Friedman et al., 2009). MiRNAs are usually produced from long primary miRNAs (pri-miRNAs) by two processive steps (for review see Bartel, 2004). In the nucleus, a protein complex containing Drosha and DGCR8 recognizes characteristic hairpin structures within the pri-miRNAs and cleaves them at the base, resulting in stem-loop precursor miRNAs (pre-miRNAs) of about 70 nucleotide length. Pre-miRNAs are transported by exportin-5 and RanGTP into the cytoplasm, where they are further processed by the Dicer/TRBP-containing protein complex into 22-bp miRNA/miRNA duplexes consisting of 5′ (miR-5p) and 3′ strands (miR-3p). The strand of the duplex with lower base-pairing stability at the 5′-end selectively associates with one of the Argonaute (AGO) proteins, which are part of the multi-protein microRNA-induced silencing complex (miRISC). The charged miRISC complexes are guided by the seven nucleotide "seed" region of the miRNA located at its 5′-end to partially complementary sequences most commonly located in the 3′-UTRs of target mRNAs, where they direct translational repression and/or mRNA degradation (Djuranovic et al., 2011). Transcriptional regulation can only partly determine miRNA levels, as pri-miRNAs are often transcribed

in clusters or as part of introns and UTRs of protein-coding genes. Their expression is therefore governed by the promoter activity of other genes (Baskerville & Bartel, 2005; Miyoshi et al., 2010). The multi-step maturation path provides several focal points where regulatory mechanisms find leverage to stop or divert production, and/or adjust the amount of final product (Slezak-Prochazka et al., 2010).

The stem–loop structures of both pri- and pre-miRNAs are targets for editing. Pri-miRNAs have been shown to undergo editing, often at more than one adenosine (Luciano et al., 2004; Yang et al., 2006; Kawahara et al., 2007a; Kawahara et al., 2007b; Kawahara et al., 2008). Yang and coworkers analyzed editing in eight randomly chosen pri-miRNAs, and found that four were edited. Notably, editing at two adenosines close to the Drosha cleavage site (positions +4 and +5 from the 5′-end of pri-miR-142) inhibits processing of pri- to pre-miR-142. Inhibition also occurs when these positions are mutated to guanosines, showing that Drosha/DGCR8 recognizes A→G and A→I changes alike. Surprisingly, the authors could not detect an increase of edited pri-miR-142 in this cell culture system, but pri-miR-142 with adenosines mutated to guanosines accumulated. Therefore, inosine-containing pri-miR-142 RNAs are short-lived, whereas pre-edited pri-miR-142 (with guanosines at positions +4 and +5) is stable and accumulates. Tudor-staphylococcal nuclease (Tudor-SN) has been identified as a ribonuclease, or a critical component thereof, that is specific to inosine-containing dsRNA *in vitro* (Scadden & Smith, 1997; Scadden, 2005). Higher levels of editing in pri-miR-142 render them progressively sensitive to Tudor-SN, and as such edited pri-miR-142 accumulates in cells treated with a Tudor-SN specific inhibitor. Moreover, endogenous miR-142 levels are

increased in ADAR null mice, underscoring that ADAR activity significantly modulates proper miR-142 output.

Drosha cleavage is not the only miRNA biosynthesis step that can be modulated by editing. Editing in pri-miR-151 at a position close to the Dicer cleavage site is accomplished in a tissue-specific manner (Kawahara et al., 2007a). Detailed analysis of the editing levels in human pri-, pre-, and mature miR-151 revealed that in this case, processing by Dicer, but not Drosha, is inhibited by editing of miR-151 precursors. Unlike inosine-containing pri-miR-142, high levels of editing are detected in pre-miR-151, indicating that the presence of inosine in pre-miR-151 does not render it unstable, in contrast to what happens to pri-miR-142.

A systematic interrogation of editing of known pri-miRNA across different tissues was first performed by Blow et al. (Blow et al., 2006). Medium to high (10-70%) levels of editing that varied across tissues were observed in six out of 99 analyzed pri-miRNAs. The authors caution that low editing levels escape detection by their analysis, and thus their findings may underestimate the number of edited pri-miRNAs. Another study found that 16% (47 of 209) of pri-miRNA in human brain are edited to 10-100% at 86 sites (Kawahara et al., 2008). Editing in the seed sequences of mature miRNA was only detected in four cases. It was observed that the fraction of edited miRNA is lower than that of edited pri-miRNA. This reveals that editing almost always inhibits miRNA maturation, which was confirmed by *in vitro* processing analysis of a subset of edited pri-miRNA. Interestingly, in two of the six analyzed pri-miRNAs, editing increases Drosha cleavage rate as much as twofold, thus promoting processing.

Deep sequencing studies corroborate the findings that editing can rarely be detected in mature miRNAs (Schulte et al., 2010). In just one case so far it has been shown that editing can change the target specificity of a miRNA: ADAR2 re-directs the edited miR-376a2-5p to a different set of targets. At the same time ADAR2 binding to pri-miR-376a2 also specifically competes with Drosha processing, independent of its editing activity (Heale et al., 2009). Predominantly, however, ADARs seem to modulate miRNA biogenesis and steady-state levels. The fact that 16% of pri-miRNAs are edited demonstrates the relevance of posttranscriptional A-to-I modification on miRNA-based silencing in human. This impact may be extended even further when considering that binding alone of pri- or pre-miRNAs by ADAR, without the actual editing event taking place, can interfere with Drosha or Dicer processing due to substrate competition. Deaminase-independent effects and degradation of inosine-containing RNAs thus probably further increase the percentage of miRNAs affected by ADARs compared to what has been reported so far.

Taken together, since editing of miRNA precursors can modulate miRNA fate, it thus functions as a regulatory step in the miRNA pathway. Editing can have one of several effects on the quality and/or quantity of the functional mature miRNA. Working on a case-by-case basis, editing can either lead to destabilization and degradation of pri-miRNA by Tudor-SN, inhibition or promotion of cleavage by Dicer or Drosha, change of strand-selection or re-direction to a distinct set of target mRNAs. Furthermore, binding of ADAR alone can interfere with the components of the miRNA pathway and thus affect miRNA levels.

1.5 Biocomputational analysis methods for the identification of novel A-to-I RNA editing targets

As discussed, ADAR binding and editing can have a strong impact on the fate and functionality of diverse RNA species and/or the resulting proteins. This thesis work focuses on examples of editing that lead to non-synonymous codon changes in protein-coding transcripts, so-called recoding events. A number of reasons contribute to the elusiveness of A-to-I RNA recoding events. First, there is no primary sequence signature that is targeted by ADARs. In fact, recognition primarily relies on intramolecular double-stranded secondary structure of RNA, which is formed between partially complementary sequences that may be several kilobases apart. Second, comparison of annotated mRNA or EST sequence data with the genomic counterpart identifies A/G discrepancies (inosine appears as a guanosine in cloned cDNA). However, these may also be due to genomically encoded SNPs. Third, editing has in many cases been shown to be tissue-specific and temporally regulated and thus may evade identification due to limitations in the sampling spectrum. The first recoding events were all chance discoveries. Yet, the significant impact on the affected proteins of non-synonymous amino acid changes attributable to editing justify a more systematic and directional approach to uncover further editing events that lead to recoding. Advances on different fronts have made it possible to attempt such studies: the complete sequencing of the human genome and the collection of a vast amount of expression data provide the basis for genome-wide analysis (Lander et al., 2001; International Human Genome Sequencing Consortium, 2004); the speed of data processing achieved in recent years by advancements in computer technology grants the necessary horse-power to plow through immense databases at a manageable rate; and the

accumulation of data for the statistical analysis of the inclinations of ADARs toward certain sequence signatures such as nearest-neighbor preferences may streamline the biocomputational analyses sufficiently to achieve a viable signal-to noise ratio in the predictions, enabling the discovery of novel A-to-I RNA editing recoding targets. Furthermore, with the ever-increasing amount of data available from a number of species, it is possible to analyze conservation of potential editing target sites and, most importantly, the partially complementary sequence often present in adjoining introns, the so-called editing complementary sequence (ECS). High conservation of sequences within introns suggests the presence of functional information, which could be due to conservation of an ECS in order for the transcript to retain the capacity of being recognized by ADARs.

Double-strandedness is an essential feature for ADAR-targets; however, other determinants are less well defined, which makes the search for site-selective editing targets particularly elusive (Seeburg, 2002). Notwithstanding, several groups attempted to identify novel editing targets in coding regions either experimentally (Ohlson et al., 2005; Pokharel & Beal, 2006) or by first scanning the human genome database in search of candidate genes and then verifying *in vivo* targets experimentally (Blow et al., 2004; Bundschuh, 2004; Levanon et al., 2004; Clutterbuck et al., 2005; Eisenberg et al., 2005a; Levanon et al., 2005). Alas, these screens missed several of the already known targets, reflecting the difficulty of the task. Furthermore, the candidate lists overlap little between studies and there was invariably a high background with only few novel bona fide editing targets actually being discovered. The bioinformatics approaches apply algorithms that allow filtering of the candidate genes to exclude genomically encoded SNPs and zero in

on true editing targets. ADARs show some selectivity with respect to nearest neighbor nucleotides and, of course, the ability of the RNA to form double-stranded structures of a certain length. This permits, on the one hand, the reduction of the number of candidate genes in the output list and, on the other hand, the ranking of the remaining putative targets as true *in vivo* substrates for ADARs. However, even though novel targets were found using strategic approaches, the significant amount of work invested does not seem to justify the low success rate. Alternative and more versatile algorithms with learning capabilities may be required for a systematic and directional approach for the identification of novel recoding events. One avenue exploring such an approach was implemented by our lab and is discussed in Chapters 2 and 3.

1.6 Identification and characterization of site-directed A-to-I RNA editing targets

The specific aims outlined at the beginning of the thesis research were aimed at answering two questions. First, can we identify previously unknown recoding editing events using algorithms based on the cumulative predictive force of known features of such recoding events? Second, do specific recoding events have functional consequences for the resulting protein isoforms(s)?

Our lab employed a two-stage approach in an effort to identify novel ADAR targets to answer the first question. In a pilot-study outlined in Chapter 2, we implemented and experimentally validated a filter algorithm that allowed for the identification of recoding editing within a subset of human genes, those present in the SNP database. This search algorithm was then used as basic principle in a fully automated computer program that allowed for a genome-wide analysis, as discussed in Chapter 3. Again, we were able to identify previously undetected A-to-I RNA editing events, validating and corroborating the basic principle of the approach. Such a global analysis, coupled with high-throughput sequencing technology to analyze different tissues and developmental time points, could greatly advance our understanding of the extent and prevalence of the "inosinome".

Three A-to-I RNA editing targets were then selected to determine whether the elicited codon change(s) have consequences on protein function in these cases. Several tools were employed to answer this question. For example, evolutionary conservation is a strong indication for the relevance of a recoding event in terms of functionality. Tissue-specific regulation of editing levels may be yet be another sign of additional capacity that

editing confers upon a certain target. And finally, well-directed functional protein assays based on information reported in the literature about the protein can directly answer whether an editing event is likely to be physiologically relevant or not. In Chapter 4, the evolutionary conservation of A-to-I RNA editing of the Complement component 1, q subcomponent-like 1 (C1QL1) mRNA was assessed by analyzing the predicted RNA folds and editing in diverse species. In Chapter 5, we discuss functional effects editing could have on FilaminA (FLNA). Editing in this target has previously been shown to be evolutionary conserved. Here we analyzed the effect of editing on the protein-binding capacity of FLNA. And in Chapter 6 we characterized both evolutionary conservation and tissue-specificity of editing of insulin-like growth factor binding protein 7 (IGFBP7). This study was complemented by a functional protein assay that showed differences in proteolytic processing of IGFBP7 editing isoforms with possibly significant physiological effects. The characterization of these A-to-I RNA editing targets showed that editing has an effect on protein function, possibly with dramatic consequences *in vivo*.

Finally, the predicted candidate ranked highest in the genome-wide analysis screen (Chapter 3) was shown not to be edited. Based on our current knowledge of editing targets, this observation was intriguing. In fact, the lack of editing in this Serpin peptidase inhibitor clade A, member 3 (SerpinA3) presented an opportunity to unveil possible regulatory mechanisms that prevent editing, a question that has heretofore not been addressed in the literature. We therefore sought to better understand the reasons for the lack of editing in this highly predicted target using minigene constructs and appropriate editing assays and concluded that editing does occur, yet it affects pre-mRNA

stability. This finding is unanticipated and opens new doors through which we can pursue to enhance our understanding on regulatory mechanisms that intersect with and modulate the A-to-I RNA editing pathway.

2 Screening of human SNP database identifies recoding sites of A-to-I RNA editing

2.1 Abstract

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that can affect the expression or function of genes. As a result, they may lead to phenotypic differences between individuals, such as susceptibility to disease, response to medications, and disease progression. Millions of SNPs have been mapped within the human genome providing a rich resource for genetic variation studies. Adenosine-to-inosine RNA editing also leads to the production of RNA and protein sequence variants, but acts on the level of primary gene transcripts. Sequence variations due to RNA editing may be mis-annotated as SNPs if this annotation is based solely on expressed sequence data instead of genomic material. In this study, the human SNP database was screened for potential cases of A-to-I RNA editing that cause amino acid changes in the encoded protein. The search strategy scores candidate sites with regard to five molecular features displayed by known editing sites. It identifies all previously known cases of editing present in the SNP database and successfully uncovers novel, bona fide targets of adenosine deamination. Our approach sets the stage for effective and comprehensive genome-wide screens for A-to-I editing targets.

2.2 Introduction

This chapter is part of the publications by Gommans, W.M. et al, *RNA* (2008), 14:2074–2085, describing a study undertaken by our lab, and Sie, C.P. and Maas, S. *FEBS Letters* (2009) (Gommans et al., 2008; Sie & Maas, 2009). At the time of the study, the total number of single nucleotide polymorphisms (SNPs) reported in public databases exceeded 9 million (Sherry et al., 2001), making SNPs the most frequently occurring genetic variation in the human genome, appearing every 100 to 300 bases in coding and non-coding regions of the genome. SNPs are DNA sequence variations that occur when a single nucleotide in the genome sequence differs between members of a biological species. For a variation to be considered a SNP, it must occur in at least 1% of the population. Many SNPs have no effect on cell function, but some are important molecular markers that link sequence variations to phenotypic differences. The characterization of these SNPs advances the understanding of human physiology and the molecular bases of diseases (Taylor et al. 2001). Furthermore, SNPs that involve an amino acid change (recoding SNPs) are of interest for clinicians and researchers, since they often strongly influence the function of the resulting gene product.

It is imperative to distinguish DNA-based single nucleotide variations (true SNPs) from sequence alterations in gene products (RNA or protein) caused by editing events. SNPs are generally annotated based on the sequence analysis of chromosomal DNA from many individuals and subsequent determination of the ratio of the alleles within the population for each site. However, among the millions of validated genomic SNPs, some polymorphisms have been annotated based solely on the analysis of expressed sequences derived from mRNA (Buetow et al., 1999; Irizarry et al., 2000). In the absence of

additional genomic confirmation, it is possible that such sequence variations may not represent true SNPs, but instead are a result of RNA editing. Indeed, a few previously annotated SNPs located within non-coding sequences were recently shown to be single nucleotide sequence variations caused by RNA editing (Eisenberg et al., 2005a). Eisenberg and co-authors identified these editing sites because they coincide with Alu repeat elements that had previously been shown to undergo RNA editing at other positions (Athanasiadis et al., 2004; Kim et al., 2004; Levanon et al., 2004).

In contrast to thousands of editing sites identified in Alu repeat elements (see "1.3.2: Pre-mRNA editing in non-coding sequences"), only a small number of site-selective recoding events are known (Pullirsch & Jantsch, 2010), most of which were identified serendipitously. Only few cases of recoding in mammals were found through bioinformatics-driven approaches (Clutterbuck et al., 2005; Levanon et al., 2005; Ohlson et al., 2007). However, a major limitation of systematic searches for edited genes in mammals has been a low signal-to-noise ratio (Morse et al., 2002; Morse, 2004; Gommans, 2008). Recently, high-throughput sequencing allowed analysis of a large number of predicted target sites. These efforts determined that high levels of recoding editing are relatively rare, but can occur at low levels at many sites (Li et al., 2009). Despite these advancements in technology and analysis methods, scientists are still restricted to analyzing a pre-determined pool of candidates, requiring sophisticated screening to enrich this sample set with true editing sites.

We hypothesized that a bioinformatics approach, which employs a strategic filtering of available databases and ranking of candidates, would increase the proportion of true editing targets within high-ranking putative targets. Briefly, A/G discrepancies

between genomic and expressed sequences that are not genomically validated SNPs are retrieved from available databases. These candidates are then ranked with respect to their compliance with key attributes of a reference set of known targets. Even though there is no single attribute such as a specific sequence or secondary structure that determines an editing site, several features combined may produce a sufficiently high signal-to-noise ratio in the output dataset to allow for a sophisticated sampling of the gene pool and the retrieval of novel bona fide editing targets. The goal of this study was to experimentally sample candidate lists produced by this bioinformatics approach and determine whether or not the employed concept provides a basis for effective and comprehensive screening for A-to-I RNA recoding targets on the genomic scale.

The screening protocol was shown to identify all of the previously known editing targets with SNP annotations as high-scoring candidates. Furthermore, here I experimentally show the *in vivo* occurrence of recoding RNA editing in human brain tissue for two additional genes that are among the highest scoring candidates from our screen, IGFBP7 (Gommans et al., 2008) and C1QL1 (Sie & Maas, 2009). The highest-ranking candidate, SRp25 nuclear protein isoform 3, was also shown to be edited by Dr. W.M. Gommans in our lab.

Overall, the experimental analysis of 68 predicted sites from four scoring groups revealed a high accuracy of predicting bona fide editing sites. In the highest scoring group, four out of seven sites (57%) are real editing substrates. It is important to keep in mind that our experimental analysis is limited to one human brain cDNA and as such to one tissue specimen and one specific developmental time-point. It cannot be ruled out that candidates undergo editing in different tissues (He et al., 2011), at distinct

developmental time points (Rula et al., 2008), or in response to external stimuli, as recently shown for Neil1 (Yeo et al., 2010). Furthermore, the detection limit of the assay may prevent identification of bona fide targets that are edited to low levels. Despite these experimental limitations, the results underscore the validity of the applied search algorithm and ranking system. The validated algorithm was used as the underlying basis for the development of an automated algorithm (Chapter 3) for the genome-wide analysis and prediction of editing target genes.

2.3 Materials and Methods

2.3.1 Databases and data analysis

The database search and data analysis was conceived and performed by other members of the lab (W.M. Gommans, N. Tatalias, S. Maas). For reference and in order to give a complete picture of the study, the search strategy is outlined briefly here and in detail in Appendix A, even though my contribution was restricted to the experimental analysis of candidate genes. Annotations for human single-nucleotide polymorphisms (SNPs) from the dbSNP database (Sherry et al., 2001) build 125 were downloaded using the UCSC genome table browser (Kuhn et al., 2007). RNA secondary structures were predicted using the M-fold algorithm (Zuker, 2003) and multiple sequence alignments were obtained with the online program clustal W 1.8 (Jeanmougin et al., 1998).

The dbSNP database contained a total of > $5x10^6$ mapped SNPs. Figure 6 depicts the subsequent filtering steps that were performed to narrow down the list of SNPs to only those that might represent recoding RNA editing sites within known genes. First, we extracted all annotations that were based solely on expressed sequence data using the UCSC genome table browser (Kuhn et al. 2007) to yield ~30,000 sites. Second, all variations other than A/G or G/A were removed. Subsequently, only the entries where adenosine is present in the genomic sequence at the putative SNP position were retained, whereas those with G in the genomic sequence were removed. Fourth, we filtered the sites located within the known coding sequence of genes from sites in non-coding exons, since we wanted to focus on recoding events. This step eliminates potential A-to-I RNA editing sites in small regulatory RNAs such as miRNAs and editing events affecting 5'- and 3'-untranslated regions of mRNAs (for review, see Nishikura 2006; Gommans et al.

2008). In the next step we removed the sites that produce synonymous changes, i.e. the codon changes caused by the discrepancy leaves the protein sequence unaltered. This narrowed the number of potential editing sites down to 984. Finally, we selected among these 984 positions the ones located within known genes, thus removing entries with ‘‘hypothetical’’ and ‘‘unknown protein’’ annotations. The resulting list of 863 sites constituted the starting point for our bioinformatics analysis to rank the entries in order to identify the ones that have a high probability of representing bona fide RNA editing sites.

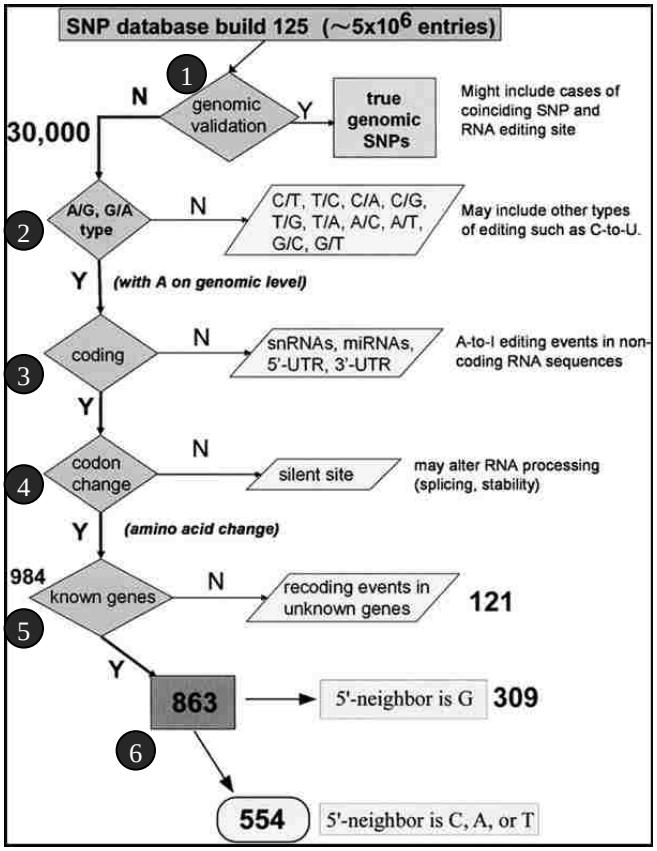


Figure 6: Filtering of data derived from SNP db build 125

Sequential removal of candidates from list due to:

1. *SNPs annotated based on genomic sequence*
2. *discrepancies other than A/G and G/A*
3. *genes where discrepancy occurs in non-coding regions*
4. *discrepancies that are silent codon changes*
5. *unknown genes*
6. *sequences displaying a G as 5’-neighbor*

Figure adapted from Gommans et al., 2008.

The molecular features used to rank and filter each of the 863 potential editing/SNP sites are derived from known properties of previously characterized mammalian RNA editing sites. We downloaded and evaluated the bases at the -1 and +1 positions relative to the predicted editing site in order to score the entries according to the

main 5'- and 3'-base preferences of ADARs (Bass, 2002; Athanasiadis et al., 2004). First, since guanosine rarely if ever precedes an editing site, we removed all entries with a G at -1 from the list. Second, the assigned values for the -1 position are 1 for A or T, and 2 for C and for the +1 position 1 for G and 2 for either A, T or C. Third, we manually assigned a value for cross-species conservation based on the PhastCons program (Siepel et al., 2005). The value captures how strongly the potential target site itself as well as the surrounding sequence is conserved (including mouse, rat, chicken and zebrafish). Only candidates exhibiting medium to high conservation were used for further analysis. Fourth, evidence for *in silico* editing was analyzed for each of the remaining sites, whereby 30nt upstream and 30nt downstream of the predicted nucleotide were blasted against the nucleotide (nr/nt) and human EST databases (NCBI) and the percentage of sequences carrying a G instead of an A was recorded. Only candidates showing $\geq 1\%$ *in silico* editing were used for the last analysis step. Finally, up to 2.5kb of genomic sequence in both directions from the putative editing site were inspected for RNA foldback structures using M-fold (Zuker, 2003) and structural scores were determined for each fold.

For each of the described features individual scores were computed using a LODs scoring method and combined to yield an overall score (S) (Gommans et al., 2008). For each feature the values of a reference set (positive control regions from known cases of RNA editing) were compared to the values of the sample set (all A/G discrepancies) to rank the sample set. For a more detailed description of the search and ranking strategy see "2.5: Discussion" and Appendix A.

2.3.2 RNA editing analysis

For experimental validation, human brain total RNA and gDNA isolated from the same specimen (Biochain) were purchased and processed using standard protocols for reverse transcription and PCR (Appendix A). Even though ADARs are expressed in almost all tissues (O'Connell et al., 1995), editing levels have been shown to be highest in the brain. Therefore, brain cDNA was used for initial analysis even if the transcript for the gene in the database that carries the G was derived from another tissue. In those cases where candidate gene transcripts are specifically expressed in tissue other than the brain, tissue-specific cDNA and gDNA pairs from other human tissues were analyzed. Gene-specific fragments of cDNA as well as genomic regions were amplified by PCR and subjected to dideoxy sequencing (Geneway) as described previously (Athanasiadis et al. 2004). Sequence traces of PCR products were inspected for mixed reads, with the ratio of the peak heights giving a first indication of editing levels.

2.3.3 Statistical analysis

To determine if the chance of finding a novel recoding editing site within the various scoring groups was significantly different from random chance we used Fisher's exact test.

2.4 Results

2.4.1 Identification of novel sites of A-to-I RNA editing among high scoring candidates

Four groups of genes that span the entire spectrum of the ranked candidates were selected in order to estimate the signal-to-noise ratios across the whole range of the sample set. At least 10 individual genes from each of the four groups were experimentally analyzed yielding a total of 68 analyzed genes (Table 2).

No editing was detected in any of the gene candidates from the lower three groups (Table 2: group 2: score ranks III–X [22 of 47 analyzed]; group 3: score ranks XI–XV [31 out of 83 analyzed]; group 4: score ranks XVI–XXIX [12 out of 52 analyzed]) by our RT-PCR and sequencing screening method. It is important to note that this does not prove that editing cannot or does not occur at those positions. Rather, it shows that

Group	Score	Total *n*	Tested	Edited (% of tested)
1	5-2.5	4	4	75
2	2.5-0.0	47	18 (22)	0
3	0.0-(-)2.5	83	31	0
4	<(-)2.5	52	12	0

Table 2: Statistical analysis of experimental validation

Group numbers and score ranges correspond to those in Table 4. *n = number of sites in group; tested = number of sites experimentally tested in group; edited = percentage of sites edited in vivo of sites tested per group. Fisher's exact test: group 1 versus group 2: p = 0.0026; group 1 versus group 3: p = 0.00061; group 1 versus group 4: p = 0.00714. Not included in this analysis are the four additional genes analyzed after publication (all in group 2) (Gommans et al., 2008).*

editing at these positions is not detectable using the RT-PCR screening method in a specific tissue sample isolated at a single time point from a single individual. Neither can it be ruled out that editing occurs at a rate below the detection threshold of this method. When we analyzed the top four highest scoring sites that constitute group 1 (score ranks I+II), we detected RNA editing in human brain at three of the four sites. These were located within two genes; the splicing factor SRp25 isoform 3 (analyzed by W.M. Gommans) and Insulin-like growth factor binding protein 7 (IGFBP7). Therefore, within this highest scoring group (summary score of > 2.5) four out of seven (57%) sites are real positives (Table 4). Table 2 summarizes the validation data and the statistical evaluation of expected versus observed outcomes based on the screening results in Table 4.

Since the apparent editing level for the SRp25 transcripts based on the RT-PCR sequencing assay was low, the PCR amplicon was subcloned by Dr. W. Gommans and a total of 100 individual clones were sequenced. This revealed that 7(±1)% of cDNAs carried a G instead of an A at the predicted position. In addition to the main site there may be additional minor editing sites within the same exon. The main editing site results in a lysine-to-arginine change within a basic region of the protein. Interestingly, the entire computer-predicted RNA fold-back structure is composed of exonic RNA sequences.

We showed that two of three predicted positions in IGFBP7 (Figure , Table 4), all residing within the same exon, are true editing sites. It had previously been suggested that these two sites might be subject to A-to-I modification based on database evidence and cDNA sequencing (Eisenberg et al., 2005a), without experimental evidence of an RNA-based mechanism. Our results from analysis of matched cDNA and genomic DNA from the same tissue specimen confirm that both adenosines are subject to RNA editing

Table 3: Screen using reference set of 15 known editing sites

	Gene	Sum	Rank
1	SRp25 nuclear protein isoform 3	3.963	I
2	Insulin-like growth factor binding protein 7	3.963	
3	*Bladder cancer-associated protein (Bc10)*	*3.963*	
4	YY1 transcription factor	2.844	II
5	Insulin-like growth factor binding protein 7	2.844	
6	*Bladder cancer-associated protein (Bc10)*	*2.844*	
7	*Filamin A*	*2.085*	III
8	3'-phosphoadenosine 5'-phosphosulfate synthase	2.085	
9	Insulin-like growth factor binding protein 7	2.085	
10*	C1q-related factor precursor*	2.085	
11	Glioma-associated oncogene homolog (zinc finger protein)	1.915	IV
12	Thymidylate kinase	1.915	
13	Component of oligomeric golgi complex 1	1.915	
14	Vacuolar protein sorting 4B (yeast)	1.915	
15	RARS – arginyl-tRNA synthetase	1.915	
16	S-adenosylhomocysteine hydrolase	1.915	
17	Myxovirus resistance 1, interferon-inducible protein p78	0.967	V
18	Neogenin homolog 1 (chicken)	0.967	
19	Actin related protein 2/3 complex, subunit 3	0.967	
20	Heme oxygenase (decycling) 2	0.967	
21	Phospholipid transfer protein	0.796	VI
22	Pyruvate dehydrogenase complex, component X	0.796	
23	Eukaryotic translation initiation factor 2, subunit 3 gamma	0.796	
24	Similar to Proliferating-cell nucleolar antigen AK021577	0.796	
25	Amyloid beta precursor protein binding protein 2	0.796	
26	Mitochondrial ribosomal protein L28	0.796	
27	ABCD3 protein ATP-binding cassette (ABC) transporters	0.796	
28	Ubiquitin protein ligase E3 component n-recognin 1	0.796	
29	Thioredoxin-like protein p19 precursor	0.796	
30	Methionyl aminopeptidase 2	0.796	
31	Vacuolar protein sorting 29 (yeast)	0.796	
32	Excision repair cross-complementing rodent	0.796	
33	Tumor suppressor candidate 3 isoform a	0.796	
34	Signal-induced proliferation-associated 1 like 1	0.796	
35	Sterol carrier protein 2	0.796	
36	Protein phosphatase 1, catalytic subunit, gamma isoform	0.796	
37	Sorcin isoform a	0.796	
38	Tripeptidyl peptidase II	0.796	
39	Brix domain containing 1	0.796	
40	EBNA-2 co-activator variant	0.796	
41	CYFIP	0.530	VII

Table 4: Screen using reference set of 19 known editing sites

	Gene	Sum	Rank
1	SRp25 nuclear protein isoform 3	4.239	I
2	**Insulin-like growth factor binding protein 7**	4.239	
3	YY1 transcription factor	2.776	II
4	**Insulin-like growth factor binding protein 7**	2.776	
5*	**C1q-related factor precursor***	2.455	III
6	3'-phosphoadenosine 5'-phosphosulfate synthase	2.455	
7	**Insulin-like growth factor binding protein 7**	2.455	
8	Glioma-associated oncogene homolog (zinc finger protein)	1.607	IV
9	Thymidylate kinase	1.607	
10	**Component of oligomeric golgi complex 1**	1.607	
11	Vacuolar protein sorting 4B (yeast)	1.607	
12*	**RARS – arginyl-tRNA synthetase***	1.607	
13*	**S-adenosylhomocysteine hydrolase***	1.607	
14	Myxovirus resistance 1, interferon-inducible protein p78	0.992	V
15	Neogenin homolog 1 (chicken)	0.992	
16	Actin related protein 2/3 complex, subunit 3	0.992	
17	Heme oxygenase (decycling) 2	0.992	
18	Phospholipid transfer protein	0.516	VI
19	Mitochondrial ribosomal protein L52, isoform a	0.221	VII
20	ATP synthase H+ transporting mitochondrial F1, subunit	0.221	
21	Lysozyme-2	0.221	
22	Hyaluronan synthase 1	0.221	
23	Bromodomain containing 1	0.221	
24	Male-specific leghal 3-like 1 (Drosophila)	0.221	
25	Inositol polyphosphate-5-phosphatase A	0.221	
26	Zinc finger protein 358	0.221	
27	KIAA0741 protein	0.221	
28	Glycoprotein hormones, alpha polypeptide	0.221	
29*	**Zinc finger protein 289, ID1 regulated***	0.221	
30	Tropomyosin-binding subunit of the troponin complex	0.221	
31	Sterol carrier protein 2	0.221	
32	Ras-related C3 botulinum toxin substrate 1	0.221	
33	Pyruvate dehydrogenase complex, component X	0.144	VIII
34	Eukaryotic translation initiation factor 2, subunit 3 gamma	0.144	
35	Similar to Proliferating-cell nucleolar antigen AK021577	0.144	
36	Amyloid beta precursor protein binding protein 2	0.144	
37	Mitochondrial ribosomal protein L28	0.144	
38	ABCD3 protein ATP-binding cassette (ABC) transporters	0.144	
39	Ubiquitin protein ligase E3 component n-recognin 1	0.144	
40	Thioredoxin-like protein p19 precursor	0.144	
41	Methionyl aminopeptidase 2	0.144	
57*	**C1Q-like tumor necrosis factor 5***	-0.177	

in human brain and are not genomic polymorphisms. At site B, which is edited to 30%, the resulting lysine-to-arginine (K95R) amino acid substitution affects the IGFBP7 protein sequence within a region representing a heparin binding site that also overlaps a protease cleavage site (Sato et al., 1999; Ahmed et al., 2003). Site A is also subject to RNA editing with a level of modification around 10% (Figure 7) and encodes glycine instead of arginine in its edited state (R78G).

EST database analysis can provide additional insight on editing levels at a specific site, as it often contains large numbers of annotated expressed sequences of a given gene. Within the IGFBP7 gene, 36 out of 302 (12%) of human ESTs carry a G at site R78G and 132 of 302 (=43.7%) at site K95R. Such *in silico* editing also suggests that position C may be edited, with eight of 302 ESTs (=2.6%) carrying a G at this site.

Legends to Tables 3 and 4 (previous pages):

Table 3: List of gene names, score ranks, and summary score values obtained after screening using the basic reference set of 15 known editing sites. Entries with the same summary score S are grouped together. Three sites of previously annotated SNPs that are known editing sites are italicized. Candidates shaded in gray are true editing sites as shown in this or previous studies.

Table 4: List of genes, score ranks and summary score values obtained with the expanded reference set of 19 sites. Candidates experimentally tested by our lab are shaded: validated RNA editing sites in dark and candidates with no evidence of editing in light gray. Candidates that were experimentally analyzed by me are highlighted in bold. Asterisks indicate analysis after publication of the study by Gommans et al., 2008.

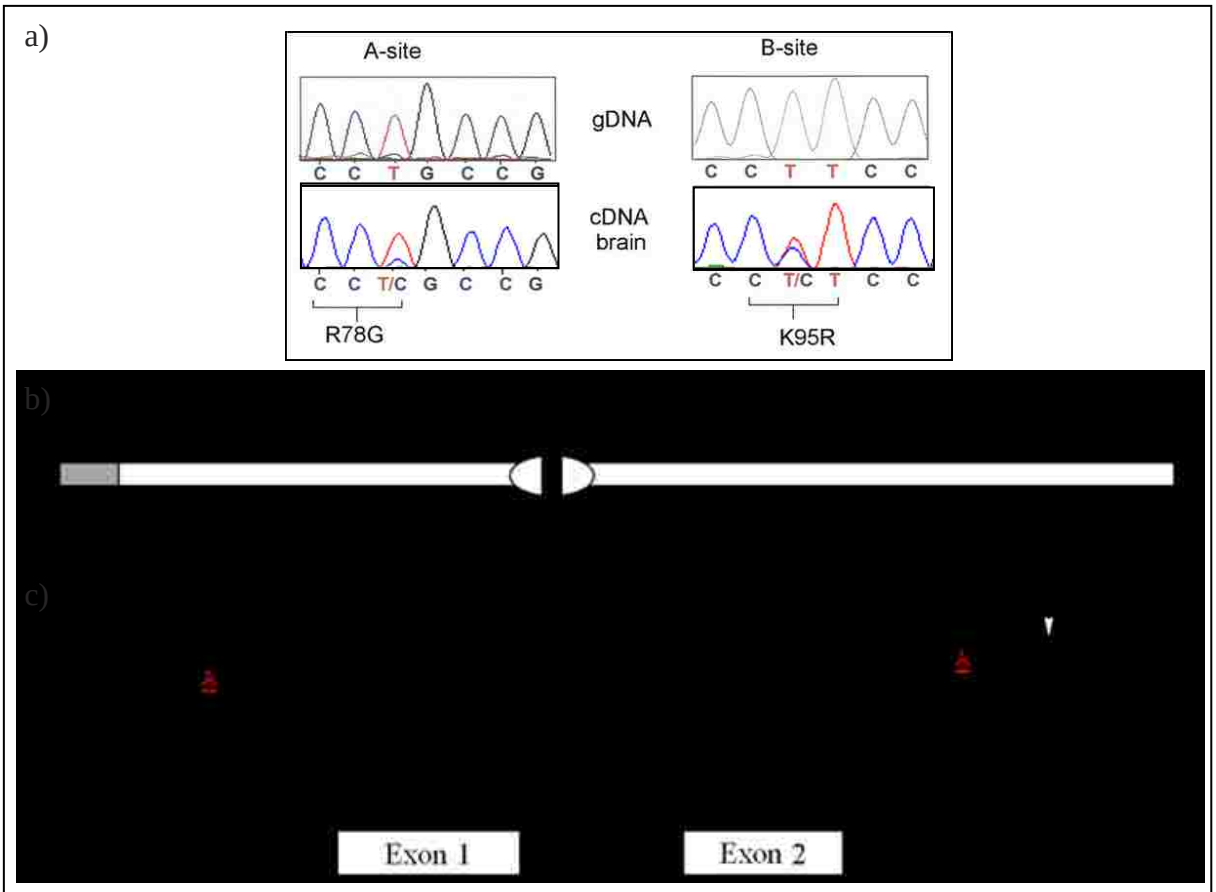


Figure 7: A-to-I RNA editing of IGFBP7 pre-mRNA

a) Sequence traces obtained from amplified genomic and matching cDNA samples encompassing the predicted RNA editing sites R78G and K95R within the IGFBP7 exon1 sequence. A mixed read of A and G at the same position indicates a mixed population of mRNAs in the cDNA samples due to post-transcriptional A-to-I RNA modification. Reverse complement shown. b) Schematic representation of functional domains of the IGFBP7 protein The approximate positions of the two amino acid changes resulting from RNA editing are indicated. SS = secretion signal (aa1 1-26); HBS = heparin binding site (aa 89-97); cleavage = protease processing site at amino acid position 97. c) Predicted RNA fold encompassing IGFBP7 editing sites (underlined and highlighted in red) within exon 1 (Zuker, 2003). Also labeled is a third potential minor editing site C that did not show evidence of editing in vivo *in our screen and might represent a true genomic SNP. Figure adapted from Gommans et al., 2008.*

Candidate position C in IGFBP7 did not show evidence of editing in our sequencing analysis and may therefore represent a genomic polymorphism, or RNA editing is restricted to specific cell types or occurs at a very low level. Inspection of mouse and rat mRNA and EST databases suggests that RNA editing at positions R78G and K95R also occurs in rodents. In mouse, 48 out of 85 ESTs (=56.5%) carry a G at position R78G, and a similar number carry a G at position K95R (=57.7%). In rat, 75.4% of ESTs carry a G at position R78G and 80% at position K95R.

The predicted editing site on rank 5, Complement component 1, q subcomponent-like 1 (C1QL1), changes codon 66 from glutamine to arginine (Q66R) when edited. It shows a high level of editing (Figure 8), but due to high background in the sequence track, the amplicon was subcloned for more definite determination of editing levels. This revealed that Q66R was edited to 56%. Furthermore, we discovered an additional editing site located ten nucleotides upstream of this major site, which is edited to a lower degree (18%), changing a threonine to an alanine codon (T63A). Interestingly, this minor site is always edited together with codon 66. Editing at sites located on the same face of an RNA duplex, as is the case here, is often coupled, whereby the site edited to a lower degree (termed minor site) is only edited when the major site is edited (Enstero et al., 2009). Unfortunately, the genomic counterpart for this human specimen was not available, and these results were thus not included in the publication by Gommans et al. After validating the editing in this target, it was published in separate manuscript (Sie & Maas, 2009).

Within score ranks I–VI, none of the sites lacking evidence for editing turned out to be genomic SNPs based on the analysis of the matched genomic DNA. The analysis of

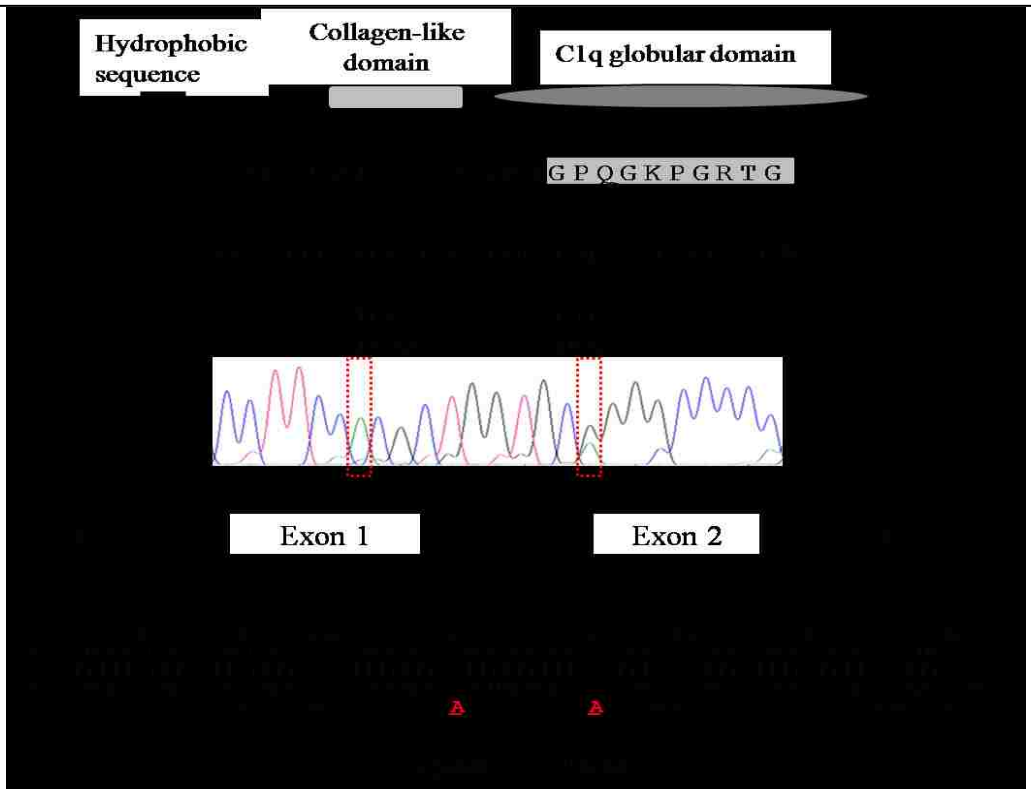


Figure 8: Editing of human Complement component 1, q subcompononent-like 1 (C1QL1)

a) Schematic representation of the C1QL1 protein with the three main functional domains indicated. The amino acid sequence surrounding the editing site is shown, and recoding events are indicated both at the amino acid and at the RNA sequence level. cDNA sequence track is shown, but the matching genomic sample was not analyzed (see Chapter 4 for further information). b) Predicted RNA secondary structure (Mfold) of the sequence encompassing the two editing sites of C1QL1 in exon1. Figure adapted from Sie & Maas, 2009.

genes from groups VII and VIII revealed three cases as genomic SNPs (UBE3, IP5PA, and AK021577). The presence of a genomic SNP does not rule out that the same position undergoes A-to-I editing in transcripts derived from an adenosine-bearing allele, but in the absence of evidence for editing after experimental validation we assume that editing does not occur there.

2.5 Discussion

Toward the long-term goal of a comprehensive analysis of the prevalence of A-to-I RNA editing in the human transcriptome, we developed a combined bioinformatics and experimental strategy. A critical component of such a strategy is to define selective criteria that capture as many of the true targets as possible while eliminating sequences that are not modified by ADARs *in vivo*. Although each individual feature does not strongly select for a bona fide editing target over background, the combination of scores using five distinct molecular determinants into a single weighted score has a much stronger predictive value.

2.5.1 Bioinformatics screen for A-to-I RNA editing candidates in the human SNP database

Interestingly, three out of four recently validated cases of A-to-I RNA editing (Clutterbuck et al., 2005; Levanon et al., 2005) affecting two genes (bladder cancer-associated protein BC10 and Filamin A) were previously annotated as SNPs and were ranked very high in our analysis (Table 3, position ranks 3, 6, 7). The fourth editing site located in the CYFIP coding sequence was ranked at position 41. These results clearly indicate that our search strategy is selecting for real editing targets. Furthermore, no known editing site is missed in our screen since there is no other previously reported recoding editing site among the >30,000 entries that constituted the starting point for our search.

Within the pre-filtered sample set of 554 human SNPs, all known editing sites previously annotated as SNPs that have been identified using various approaches were recaptured in our screen as high-scoring candidates. In fact, when including the novel

sites identified and validated in this study, 50% of candidates within the highest scoring ranks (I–IV) are known RNA editing targets (Table 3), whereas only a single known editing site (CYFIP; group VII) appears within all other tested medium and low scoring groups. Therefore, none of the real editing targets that have previously been characterized were missed or ranked lower than position 41 among the total of 554 entries of recoding, non-synomymous SNPs. For any of the candidates that did not show detectable editing activity in human brain, the occurrence of editing in brain or other tissues cannot be ruled out. It is in the nature of the experimental screening method applied here that editing events with levels below ~5% may be missed. Furthermore, to facilitate testing of large numbers of candidates, only one adult human specimen was analyzed. RNA editing events that are specific for certain cell types or developmental stages may thus also escape this initial screening.

Next we moved the four known cases of A-to-I editing (two in BC10 and one each in FLNA and CYFIP) that were contained in our candidate list into the reference set (now containing 19 sites). The ranking of the resulting list of highly scored candidates compared with the previous one showed no changes in order of the first 18 entries (Table 4). Apart from the sites within BC10, FilaminA, and CYFIP that we moved to the reference set, only minor changes with respect to the order of entries occur further down in the listing (Gommans et al., 2008; Supplementary Table S2).

2.5.2 Identification of novel recoding editing sites

Splicing factor SRp25 (also known as ADP-ribosylation-like factor 6 interacting protein 4), analyzed by W.M. Gommans, is a ubiquitously expressed protein of uncharacterized function (Sasahara et al., 2000). Because of its homology with SR-

splicing factors, it is believed to be a nuclear protein with a role in splicing regulation (Sasahara et al., 2000). The amino acid substitution due to RNA editing affects a basic region in the protein that has not been ascribed a specific function. Based on its sequence characteristic it may represent a nuclear localization sequence or a domain that interacts with the nucleic acid backbone. The lysine-to-arginine change does not alter the overall charge of the molecule, and represents a conservative change that may not affect the protein's function substantially. However, lysine residues can be sites of post-translational modification and thereby regulate protein function. For example, sumoylation of a specific lysine residue in tumor suppressor p53 activates its transcriptional response (Rodriguez et al., 1999). K-to-R mutation of this site blocks sumoylation of the protein while preserving the local charge in the protein (Sampson et al., 2001). Furthermore, another specific lysine residue in p53 has been found to be subject to methylation, which downregulates the protein's transcriptional activation activity (Shi et al., 2007). It will be interesting to see if the editing invoked K-to-R change in SRp25 also has a regulatory impact on SRp25 function.

The second gene that was detected in this study as a target for RNA editing is IGFBP7. Although editing in this gene had been postulated previously for two of the three sites (Eisenberg et al., 2005a), we provide experimental validation that the observed A/G discrepancy is in fact due to RNA editing by analyzing matched cDNA and genomic DNA sequences from the same tissue sample. IGFBP7 was initially identified as a gene differentially expressed in cancerous cells, and has been implicated in various forms of cancer, either as putative tumor suppressor (Sprenger et al. 2002; Wilson et al. 2002; Mutaguchi et al. 2003) with functions in apoptosis and senescence, or as a promoter of

angiogenesis in human tumor endothelium (St Croix et al. 2000; van Beijnum et al. 2006), and it is overexpressed in circulating endothelial cells (CECs) of metastatic cancer patients (Smirnov et al. 2006). The IGFBP7 protein has several functional domains in its N-terminal half, such as a leucine-rich sequence, a cysteine-rich domain (CRD), a heparin binding site, and a Kazal-type trypsin inhibitor domain (Collet and Candy 1998). The two editing sites A and B affect amino acid positions 78 (R-to-G) and 95 (K-to-R) of the full-length protein. Interestingly, the core sequence K_{89}SRKRR**K**GK_{97} (edited site in bold) has been proposed to function as a heparin binding site (Sato et al. 1999), and it was observed that cell-adhesion activities of IGFBP7 are inhibited by heparin (Akaogi et al. 1996). IGFBP7 is proteolytically cleaved after K97, which results in a two-chain form of the protein cross-linked by disulfide bridges. Proteolytic processing of IGFBP7 has been shown to modulate its growth-stimulatory activity (Ahmed et al. 2003). Furthermore, the heparin-binding activity of IGFBP7 is decreased upon proteolysis. The main editing site (K95R) not only lies within the proposed heparin-binding site of IGFBP7, but is also part of the recognition sequence for proteolytic cleavage. It will be interesting to explore the potential functional implications of RNA editing on heparin binding and/or proteolytic processing and its downstream effects regarding apoptosis, regulation of cell growth and angiogenesis.

The family of C1Q-domain proteins includes important signaling molecules with roles in inflammation, adaptive immunity and energy homeostasis (Ghai et al., 2007). The physiological function of C1QL1 has not been elucidated, but it is expressed highest within the brain and was suggested to be especially important for neurons involved in coordination and regulation of motor control (Berube et al., 1999). Furthermore, it may

be part of a neuroprotective immune response (Glanzer et al., 2007) and one study revealed upregulation of C1QL1 in response to kainic acid induced seizures (Hunsberger et al., 2005). The RNA editing sites in C1QL1 are located immediately upstream and at the beginning of a collagen-like domain. In other C1Q-domain proteins, such as the hormone adiponectin, this coincides with a region of protease-mediated processing (Waki et al., 2005). Future studies will show if the amino acid substitutions caused by RNA editing may alter post-translational processing of C1QL1, or if it affects other properties of the protein *in vivo*.

For the transcripts of all three genes, SRp25, IGFBP7, and C1QL1, the RNA fold-back structures that are predicted to mediate RNA editing involve solely exonic RNA sequences. This is in contrast to almost all other characterized recoding editing sites, which usually involve folds where the editing site complementary sequence is located within an intron. As more edited genes are identified, it will be interesting to see how often exon-only structures mediate editing compared to exon–intron fold-back structures, since it could have implications for the evolutionary mechanisms that lead to the emergence of novel editing sites and the changes in the extent of editing at individual sites over evolutionary time. Furthermore, RNAs that do not require the presence of intronic sequences for editing to occur could continue to undergo editing after the completion of nuclear pre-mRNA splicing.

2.6 Conclusions

The results of our limited screen indicate that the strategy is successful in identifying novel recoding targets. The algorithms for deriving each individual score, as well as the weighted combined score value reflect the current knowledge of the A-to-I editing mechanism and the properties of known targets. In previous database-driven studies, only A/G discrepancies that appear both in human sequences of a given gene as well as at the same position in another mammalian species were investigated (Clutterbuck et al. 2005; Levanon et al. 2005b). The latter is a valuable strategy for initial screens with little data on known targets. However, for a more comprehensive search the approach that is presented here is more suitable. In particular, current cDNA databases do not cover all genes and often do not have sufficient coverage across editing sites to reveal low-level editing events. Over time, improved and extended databases as well as additional insights into the RNA editing mechanism will lead to a refinement of the search algorithm. Biochemical approaches for performing target screens (Morse and Bass 1997; Ohlson et al. 2005) come with their separate set of biases that may favor the identification of certain types of editing targets but select against others.

At this point, the presented screen represents the most unbiased search for edited sequences in the human transcriptome with a reasonable signal-to-noise ratio. In the present study several of the selection steps were performed in a non-automated manner. A largely automated procedure will be needed to apply this approach to the complete transcriptome (Chapter 3). While such a genome-wide screening approach is expected to uncover more recoding RNA editing targets due to its comprehensiveness, recent advances suggest that many editing sites may be subjected to low levels of editing or are

regulated in a tissue- and time-specific manner (Li et al., 2009; He et al., 2011). Application of a transcriptome-wide search algorithm in conjunction with high-throughput sequencing of different tissues would enable an in-depth assessment of the prevalence of A-to-I RNA editing.

This pilot study showed that the applied filtering strategy not only identifies all previously known cases of editing present in the subset of genes analyzed, but also allowed us to identify three novel editing targets, SRp25, IGFBP7 and C1QL1. It is expected that application of a similar search algorithm on a much larger scale, namely in a genome-wide approach, would be similarly successful in identifying putative candidates with a high probability of being edited. Due to its scale, such an endeavor will need to be automated, a process which requires significant dedication of time and resources. However, the accomplishments of the strategy discussed in this chapter justified such an undertaking and therefore set the stage for effective and comprehensive genome-wide screens for A-to-I editing targets as outlined in Chapter 3.

3 Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites

3.1 Abstract

The search and filter strategy applied to the SNP database was successful in identifying novel A-to-I RNA editing sites but was restricted to a relatively small subset of genes. Here we administer an automated bioinformatics search strategy to the human and mouse genomes to explore the landscape of A-to-I RNA editing on a larger scale. In both organisms we find evidence for high excess of A/G-type discrepancies at non-polymorphic, non-synonymous codon sites over other types of discrepancies such as G/A, T/C or C/T, suggesting the existence of several thousand recoding editing sites in the human and mouse genomes. We experimentally validate recoding-type A-to-I RNA editing in a number of human genes with high scoring positions including those encoding the ATPase, H+ transporting V0 subunit e2 (ATP6V0E2) and the unknown protein BC027448. Others in the lab further identified mRNAs encoding the coatomer protein complex subunit alpha (COPA), as well as cyclin dependent kinase CDK13 as novel editing targets.

3.2 Introduction

Parts of this chapter have been published in Maas et al., 2011 (Maas et al., 2011). The pilot study analyzing the SNP database (Chapter 2) set the basis for a genome-wide application of the applied search strategy. In collaboration with Dr. D. Lopresti (Department of Computer Science and Engineering, Lehigh University), we developed a computer program (RNA Editing Dataflow System or REDS) that mines through the human genome, whereby it extracts genes that show A-to-G discrepancies between the genomic and mRNA sequences that lead to amino acid changes and that are not genomically validated SNPs. Furthermore, the targets are screened for possible self-complementary sequences around the predicted editing site.

Bioinformatic analysis methods were first capitalized on to systematically uncover novel A-to-I RNA editing recoding sites about eight years ago. Using the fruit fly as a model system, Hoopengardner and co-workers observed that sequences surrounding known recoding editing sites were conserved among *Drosophila* species (Hoopengardner et al., 2003). Hypothesizing that high conservation adjacent to editing sites arose from a selective constraint against mutations near sites of ADAR modification, this high degree of sequence identity among species was used as a potential signature of editing sites. Analysis of 914 genes encoding neurological proteins and transcription factors procured 41 genes with coding regions of unusually high sequence conservation. Editing indeed occurred in 16 of the 41 candidate mRNAs. Comparative genomics was therefore extraordinarily well suited for the identification of previously unknown editing targets in the fruit fly.

Three notable systematic approaches investigating the prevalence of A-to-I RNA editing in human were published in 2004 (Athanasiadis et al., 2004; Blow et al., 2004; Levanon et al., 2004). Sequences from cDNA and EST databases or from a cDNA library were aligned to genomic sequences using bioinformatics tools and sequences with A-to-G discrepancies were retrieved. A major problem of such approaches is that mismatches between genomic DNA and expressed sequence can be due to low sequence quality, false alignment (for instance between pseudogene and gene), SNPs, or editing. Therefore, all candidate sequences underwent some sort of quality conformance test and genomically validated SNPs were removed. Alternatively, Levanon and co-workers restricted the search to potentially double-stranded regions in order to remove noise and facilitate identification of true editing sites. All three studies found that A-to-I RNA editing primarily occurred in non-coding regions, mostly in Alu repeat elements, now recognized as the main targets of ADARs.

By combining a bioinformatics approach and comparative genomics, Levanon and colleagues then selectively directed their search toward non-synonymous A-to-I RNA editing sites, thereby avoiding inundation of putative candidates of editing by Alu-repeat elements (Levanon et al., 2005). First, they obtained candidate editing sites from both the human and mouse genomes by aligning expressed sequences and genomic DNA and clearing the list of sequences with low quality. Next, mouse and human candidate sequences were aligned and non-synonymous nucleotide mismatches occurring at identical positions were identified. This approach successfully detected four previously unknown recoding editing targets. Interestingly, all four positions had been wrongly annotated in the SNP database.

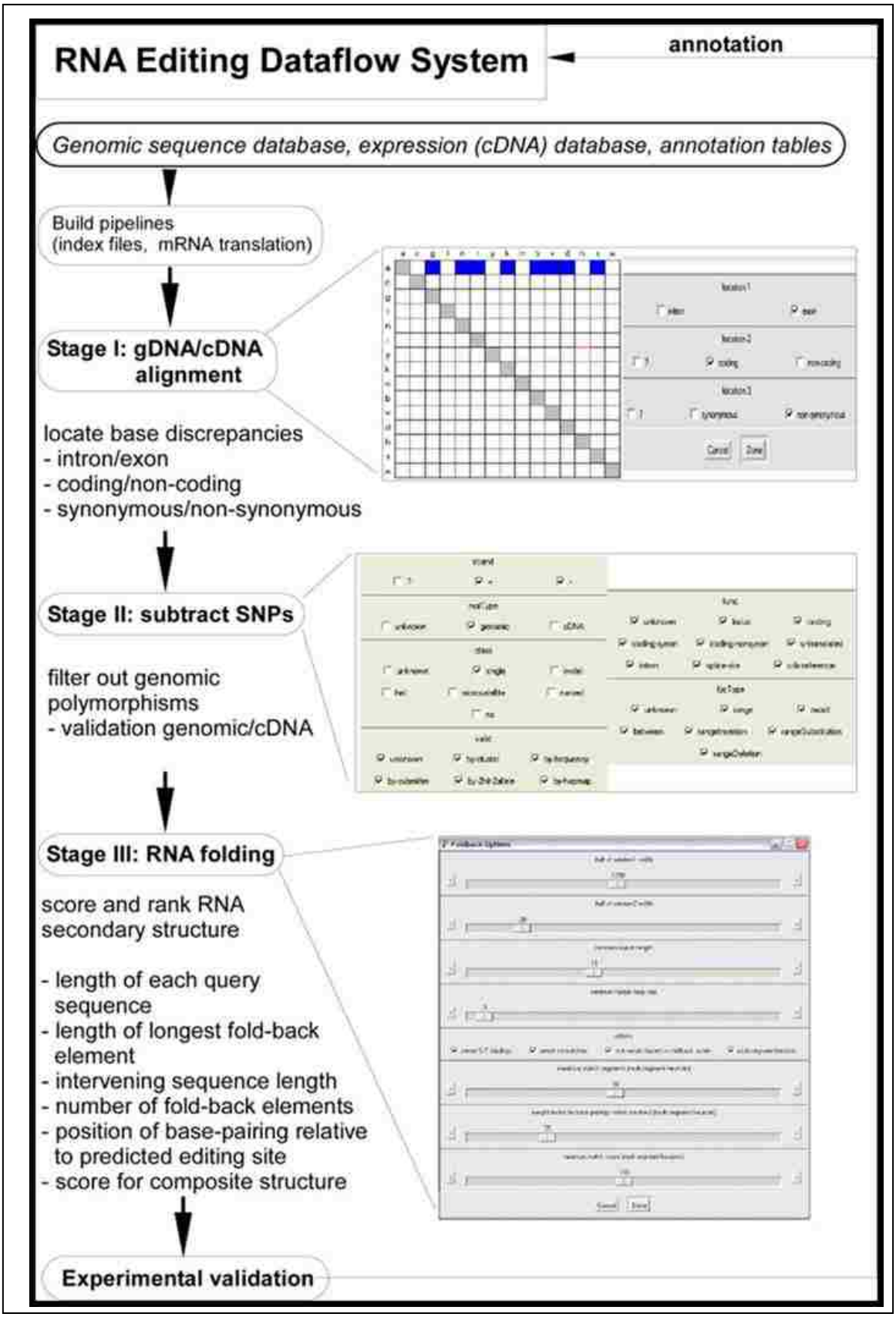
Due to limitations of sequence coverage in the databases or restriction of the search dataset in order to reduce noise, the bioinformatics approaches discussed above failed to identify known recoding editing sites as candidates. Furthermore, they had to rely on web-based tools such as blastn and Mfold for data analysis. A fast program, specifically designed to retrieve nucleotide discrepancies between expressed and genomic databases, malleable to user preferences, and optimizable in accordance with novel findings, would be a major improvement over bioinformatics approaches employed so far. REDS was designed precisely to allow such plasticity in the search strategy. It can be applied to any database, and could also be used for other purposes owing to its many adjustable parameter settings. The scripting language (Tcl/Tk – Tool Command Language/Toolkit) allows fast processing times, which permits running it on a normal desktop computer or as a web-based algorithm. Here we show that with REDS we are able to apply our proven search strategy to the genomic scale to analyze the overall landscape of base discrepancies in two species followed by experimental validation of novel recoding editing sites in the human transcriptome. In combination with standard or high-throughput experimental validation, REDS would facilitate mapping the A-to-I RNA editing landscape and define the overall impact of editing on gene expression.

3.3 Materials and Methods

3.3.1 Databases and REDS

The REDS program was conceived by S. Maas and D. Lopresti and written by D. Lopresti and his students in the scripting language Tcl/Tk (Tool Command Language, Toolkit). Human and mouse genomic DNA sequences, mRNA data files, SNP data, as well as table annotations were retrieved using the UCSC genome browser ftp site (assembly March 2006 for human and February 2006 for mouse) (Kuhn et al., 2007). REDS consists of three consecutive computational stages (Figure 9). Stage 1 aligns expressed sequences (UCSC mRNA database) from a given species to the corresponding genomic sequence. Coding sequences are translated, allowing for determination of non-synonymous codon positions within open reading frames (ORFs). According to user specifications (type of base difference, coding versus non-coding, synonymous versus non-synonymous), specific types of base discrepancies are mapped and recorded with chromosome location and position, mRNA accession and position, gene ID and description and affected amino acid. Previously known RNA editing sites are flagged. Stage 2 compares the list of base discrepancies to the species-specific SNP database (Sherry et al., 2001) and all positions that correspond to genomically validated SNPs are filtered out.

The third stage of the computational pipeline evaluates RNA folding characteristics for each of the remaining sites. The user defines several parameters: a first sequence window determines how much of the genomic sequence upstream and downstream of the putative editing site is analyzed. A second sequence window selects a small gene section surrounding the candidate site, for which the algorithm generates the


RNA Editing Dataflow System
annotation
Genomic sequence database, expression (cDNA) database, annotation tables
Build pipelines
(index files, mRNA translation)
Stage I: gDNA/cDNA alignment
locate base discrepancies
- intron/exon
- coding/non-coding
- synonymous/non-synonymous
Stage II: subtract SNPs
filter out genomic polymorphisms
- validation genomic/cDNA
Stage III: RNA folding
score and rank RNA secondary structure
- length of each query sequence
- length of longest fold-back element
- intervening sequence length
- number of fold-back elements
- position of base-pairing relative to predicted editing site
- score for composite structure
Experimental validation

reverse complement sequence with which to scan window 1 for matches (including G-T wobble base pairs and allowing for single, symmetric mismatches). Next, the cut-off value for the number of consecutive base pair matches within the RNA secondary structure are required to pass the filter and finally a minimum value for the length of the intervening sequence between the base-pairing sections are determined.

A multi-segment heuristic combines pairs of base-pairing segments within the sequence that may be part of one composite RNA secondary structure. A score is assigned to this composite structure, whereby base-pairs involving nucleotides within the inner window 2 are weighted more strongly according to a user-defined value (please see Appendix A for more detailed explanation). This biases the search for bona fide editing targets since base-paired segments of an RNA fold that supports editing include, or are in

Figure 9: Organization of the RNA Editing Dataflow System *(previous page) Flowchart describing the three-stage computational pipeline. For each stage a screen shot of the corresponding parameter options is shown. Stage I options include a matrix for selection of the type of discrepancies to search for and, because REDS translates all open reading frames, it is possible to search for exon/intron, non-coding/coding, and synonymous versus non-synonymous positions. In stage II all parameters annotated in dbSNP are selectable. Shown is an example that leads to the subtraction of any SNP sites that have been genomically validated. RNA secondary structure scores are computed in stage III according to the user-defined selection criteria. Candidate sites are ranked by structural score values in the output. Results from experimental analysis to validate bona fide editing events are annotated in REDS to flag known editing sites in future screens. Figure from Maas et al., 2011.*

close proximity to, the editing site. Finally, the output candidate sites are ranked by an overall score. For further information about the REDS program see Appendix A.

3.3.2 RNA and gDNA isolation from tissue

All experiments conformed to the guidelines laid down by the Lehigh University Animal Welfare Committee, in accord with international guidelines on the ethical use of animals. Two 4 month old male mice were sacrificed and their brains dissected. Cortex was cleared from white tissue with scissors, halved and weighed. One half was placed in 1ml of Trizol, the other minced and placed in 1ml of SNET (20mM Tris-HCl pH8, 5mM EDTA, 400mM NaCl, 1% SDS) for DNA extraction. Tissue in TRIZOL was homogenized and RNA isolated according to the manufacturer's protocol. 10μg of RNA were treated three times with DNase I (NEB) and RNA integrity verified by running an aliquot on a standard 1.5% formaldehyde gel. Yield before DNase treatments were 54-96μg RNA from 50-100mg of mouse brain tissue.

DNA was extracted by digesting tissue with 1mg/ml proteinase K at 55˚C overnight and subsequent RNaseA (1μg/ml) treatment at 37˚C for 2h. Protein was removed by phenol-chloroform extraction. Traces of phenol were cleared from the aqueous phase and interphase with 1ml chloroform/isoamylalcohol. After centrifugation, DNA in the aqueous phase was precipitated with 1ml isopropanol, spooled, washed successively in 70% and 100% ethanol and allowed to dry at room temperature. The DNA was eluted over night in 100μl Tris buffer. Yield from ~100mg tissue was 78-105μg of DNA.

3.3.3 RNA editing analysis

Human total RNA and gDNA isolated from the same specimen (Biochain) were processed using standard protocols for reverse transcription and PCR. Gene-specific fragments of cDNA as well as genomic regions were amplified by PCR and subjected to dideoxy-sequencing as described previously (Athanasiadis et al., 2004; Gommans et al., 2008). Primer information is given in Appendix B. Editing at the predicted positions was analyzed by inspecting the sequence traces for double peaks, with the ratio of the peak heights giving an indication of approximate editing levels. The occurrence of SNPs at candidate positions was excluded by analyzing the specimen-matched gDNA.

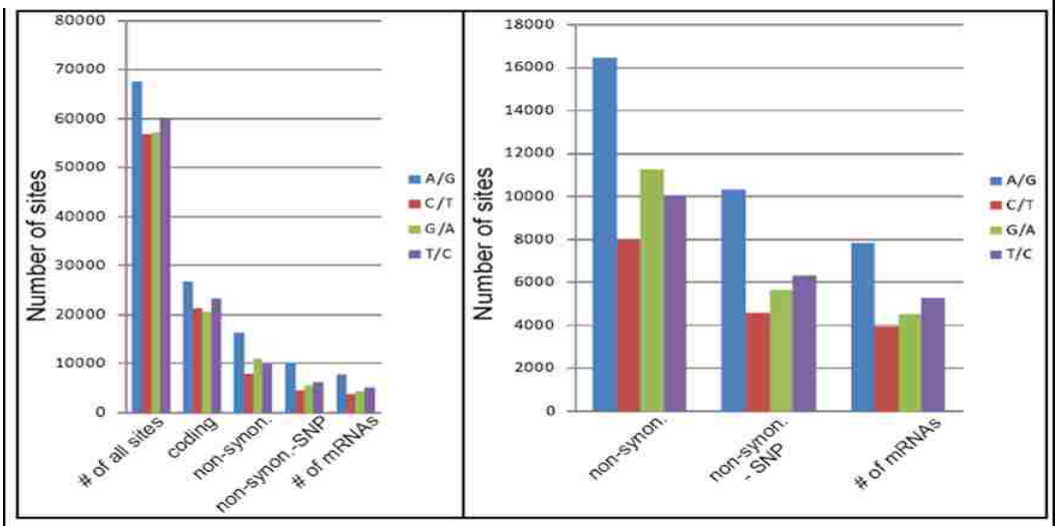
3.4 Results and Discussion

3.4.1 Evidence for abundant, site-selective recoding A-to-I editing in human and mouse

In eukaryotes, A-to-I RNA editing is the only known mechanism for generating inosine residues in RNA molecules. First we asked whether an excess of A/G discrepancies versus other types of base differences between cDNA and genomic DNA is detectable, even when excluding repetitive element mediated editing. Such a finding would support the hypothesis that many more editing sites within protein-coding sequences remain to be identified and may provide an estimate of the total number of existing sites. We aligned all sequences available in the human and mouse mRNA databases [from UCSC genome browser (Kuhn et al., 2007)] to their genomic counterparts and mapped the positions of A/G or other discrepancies between genomic and expressed sequence (Figure 10). Since there is no known mechanism for causing G/A or T/C transitions, we expect that such discrepancies are due to polymorphisms (SNPs) and/or sequencing errors and can therefore be considered background noise.

The combined coding and non-coding regions of mRNA sequences contain more A/G-type differences in human than either G/A, C/T or T/C discrepancies (68,000 A/G versus 57,000 G/A; 58,000 C/T and 60,000 T/C). In mouse, the total numbers of discrepancies are slightly fewer and the excess of A/G discrepancies is less striking (62,000 A/G versus 56,000 G/A; 58,000 C/T and 55,000 T/C). This difference can probably be accounted for by the widespread Alu-repeat mediated editing in primates, as we and others have previously mapped thousands of such editing sites in non-coding human mRNA sequences (Athanasiadis et al., 2004; Levanon et al., 2004).

a) Human



b) Mouse

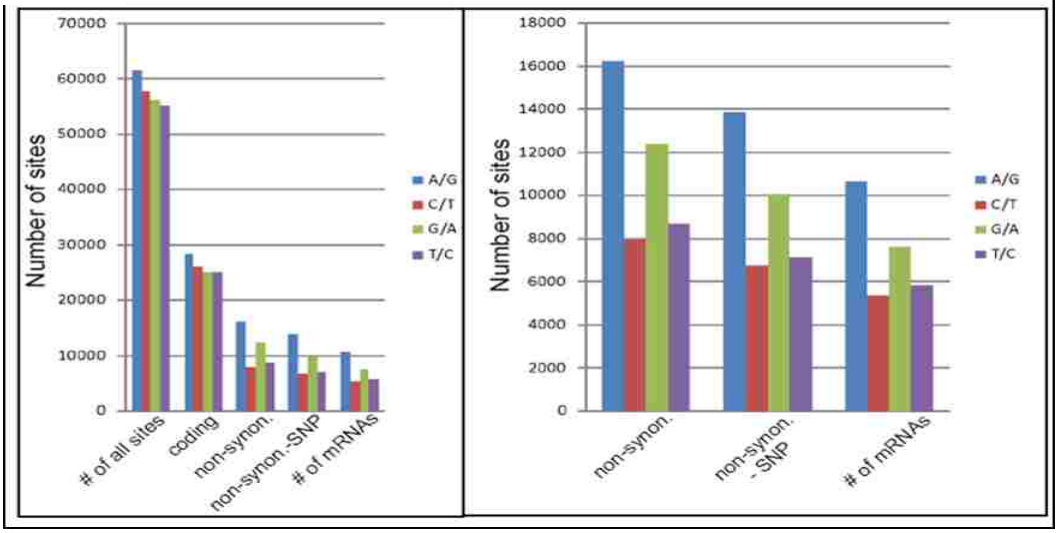


Figure 10: **Distribution of mRNA/gDNA base-discrepancies in protein-coding sequences**

a) Comparison of A/G (blue), C/T (red), G/A (green), and T/C (purple) base discrepancies between human cDNA and DNA within coding and non-coding transcripts (# of all sites), only protein-coding sequences (coding), only at codon positions that predict a non-synonymous codon change (non-synon.), only non-synonymous sites that are not annotated as genomic SNPs (non-synon. – SNP), and the number of distinct mRNAs with non-synonymous sites that are not gSNPs (# mRNAs). b) Same as in a) but for mouse. Part of the data is blown up to higher resolution on the right side of the figure. Figure adapted from Maas et al., 2011.

In contrast, rodents lack Alu-repeats and display only low levels of repeat element mediated editing. This conclusion is further supported when we exclude all discrepancies located within non-coding regions from our analysis (Figure 10). In both human and mouse, A/G discrepancies are still the most prominent type of discrepancy and the observed excess of A/G in human is comparable to that in the mouse dataset.

We observe a very strong A/G-discrepancy bias when further restricting our analysis to non-synonymous codon changes (Figure 10). Editing at such sites changes the meaning of the codon and leads to amino acid substitutions in the resulting protein. There is a substantial overrepresentation of A/G discrepancies in both human and mouse (1.4-2.1 fold in human and 1.3-2.0 fold in mouse) compared to all other types of transitions. When all known, genomically validated SNPs are subtracted from the list, the excess of A/G discrepancies further increases (1.6-2.2 fold in human and 1.4-2.1 fold in mouse). Moreover, when we count once sites identified multiple times in annotated RNA sequences (Figure 10; # of mRNAs), the excess of A/G discrepancies persists (1.5-2.0 fold in human and 1.4-2.0 fold in mouse). Therefore, A/G discrepancy sites are distributed across many genes and not dominated by a small number of highly expressed genes that are overrepresented in the expression databases.

Our analysis shows that even when just considering protein coding sequences, thereby eliminating the impact of repeat-mediated editing, there is still a substantial excess of A/G versus other transitions in the human and mouse transcriptomes. We see a surplus of ~4,700 (human) and 5,800 (mouse) unique sites compared to the average of all other types of changes. As many sites are edited to only a small extent or in a cell-type specific fashion *in vivo*, the number of discrepancies detected at this time may still be an

underestimate of the total number of editing sites in these species. Similarly, negative results from experimental evaluation of potential editing sites cannot rule out editing in another cell type or at another developmental time point. Alternatively, editing may occur to such a low level that the experimental assay is not sensitive enough for detection. In fact, two recent high-throughput sequencing studies suggest that the bulk of target sites may be edited to a low extent (Enstero et al., ; Li et al., 2009). Such a result is also predicted by the continuous probing (COP) hypothesis regarding the possible mechanism of how novel editing sites in the transcriptome emerge (Gommans et al., 2009). Due to the high accuracy sequence databases we utilized for the analysis (not including EST-type sequences), we expect that almost half of the predicted ~10,000 potential sites may reflect real RNA editing events. Since A-to-I RNA editing is the only eukaryotic mechanism known to generate A/G-discrepancies, the excess over the background of SNPs and sequencing errors points to the potential existence of thousands of additional editing sites to be characterized.

3.4.2 Synonymous versus non-synonymous changes

One unresolved question is whether or not ADAR target sites occur predominantly at synonymous, or silent, positions. Or in other words, we are interested in whether transcript modifications elicited by editing are tolerated more at synonymous sites, increased at non-synonymous sites, or occur equally at synonymous and non-synonymous sites. Until now, such an analysis was not possible due to limited information on the frequency of A-to-G discrepancies between genomic DNA and expressed sequences on a genome-wide scale. Such an analysis is also hampered by incomplete information on the prevalence of genomic SNPs. While still very difficult, the

information in Figure 10 allows an approximation of whether discrepancies between mRNA and gDNA lead to non-synonymous changes more or less frequently than would be expected by chance.

The expected amounts of synonymous versus non-synonymous codon changes due to discrepancies between mRNA and gDNA will depend on at least two things: first, the number of possibilities of a synonymous versus non-synonymous change (i.e. how many codons change the meaning when A is changed to G, and how many do not?). Second, it depends on the codon usage bias (i.e. how frequent is a codon and can it contribute to the synonymous versus non-synonymous pool of possible changes?). Let us only consider A-to-G changes in human as an example. For simplicity, only single nucleotide changes for each codon are included in the calculation. Taking into account the codon usage bias in human (downloaded from the codon usage bias database), A-to-G changes would elicit a codon change in 78.8% of all cases, if the changes occurred by chance (Table 5). Based on the number of discrepancies found in coding sequence (Figure 10), i.e. 26,500 for A-to-G changes, and discounting all SNPs (synonymous and non-synonymous, obtained from SNPdb build 132), 15,636 A-to-G discrepancies in the coding sequence are not based on genomically validated SNPs. Of these, we observed about 10,200 discrepancies that lead to non-synonymous changes. However, we would expect there to be 12,278 changes if all synonymous and non- synonymous positions were targeted by chance alone. Accordingly, non-synonymous changes occur 17.25% less likely than would be expected. Calculations were performed for all four types of discrepancies discussed in Figure 10 (Table 5). This underrepresentation of non-

Category	A→G	G→A	C→T	T→C
# codons with synonymous changes	13	15	16	16
# codons with non-synonymous changes	32	29	30	32
% non-synonymous changes	79%	68%	57%	63%
# discrepancies in coding sequence (observed)	26500	20000	21500	23000
# discrepancies, coding w/o SNPs (estimate)	15636	10171	13342	13882
# discrepancies, non-syn. w/o SNP (observed)	10200	5500	4600	6200
# discrepancies, non-syn. w/o SNP (expected)	12326	6912	7608	8799
% non-syn. changes underrepresented	**17.2%**	**20.4%**	**39.5%**	**29.5%**

Table 5: Observed and expected numbers of discrepancies

Possible discrepancies in each category were examined with regard to type of change and codon bias. This allowed calculation of a hypothetical percentage of discrepancies leading to non-synonymous changes. The resulting percentage was used to calculate the expected number of non-synonymous discrepancies. In all categories, these were underrepresented by the indicated percentage.

synonymous codon changes in the observed pool of discrepancies occurs in all categories analyzed and is statistically significant (T-test, $p<0.05$). Of course, in this pool of discrepancies we cannot distinguish between true genomic SNPs and editing, but assume that both SNPs and editing would affect synonymous vs. non-synonymous changes with a similar bias.

This finding makes sense in an evolutionary context, as non-synonymous codon changes might be deleterious compared to synonymous ones. One may speculate that the higher reduction seen for C-to-T and T-to-C changes might be due to higher evolutionary pressure to conserve these latter ones, possibly because they lead to a higher proportion of deleterious non-conservative changes. Likewise, A-to-G and G-to-A changes are possibly allowed more because they grant a higher proportion of conservative changes such as K/R.

This traditional view on neutral and conservative codon changes should not obscure more recent findings showing that synonymous changes can significantly impact

both mRNA maturation and protein folding and function. For instance, it has been shown that a neutral SNP in the multi-drug resistance 1 gene results in an altered conformation of the protein (Kimchi-Sarfaty et al., 2007). This is probably due to the presence of a rare codon, marked by the silent SNP, which affects the timing of co-translational folding, altering the structure of the sites of substrate and inhibitor interaction. Indeed, both transcript structure and the frequency of codons can affect translation rate and protein folding (see Angov, 2011 and references therein). Furthermore, another neutral polymorphism has been shown to inactivate an exonic splice silencer site (ESS), conferring immunity to deleterious silent mutations in an exonic splice enhancer (ESE) that antagonizes the ESS (Nielsen et al., 2007). Therefore, silent codon changes can have dramatic context-dependent effects through their involvement in mechanistic and/or regulatory aspects of gene expression. More research is required to better understand the impact of synonymous codon changes, marked by either a neutral SNP or introduced on the transcript level, on protein expression.

3.4.3 Experimental validation of novel editing sites in the human transcriptome

High-scoring sites predicted by REDS that were obtained when applying stringent parameter settings (window1=500nt, window2=20nt, minimal base-pairs=11) were experimentally analyzed (Table 6, Table 7). Out of 32 known and validated A-to-I editing sites (considered the ‘true positives’ for REDS analysis), 30 positions are detectable by REDS due to the presence of at least one mRNA sequence in the database that is of the edited variant (24 in case of mouse).

Gene	Chr	Position *	Codon change	% editing
COPA	1	158568868	I/V	31 (this study)
CDK13	7	39957073	Q/R	88 (this study)
ATP6V0E2	7	149206502	K/E	30 (this study)
	7	149206516	silent	29 (this study)
	7	149206525	I/M	69 (this study)
	7	149206589	R/G	59 (this study)
	7	149206599	H/R	39 (this study)
Unknown BC027448	20	4071900, 4071955, 4071957, 4071967, 4071987, 4071991, 4072030, 4072068	Unknown (reading frame unknown)	5-95% (this study)
HMCN1	1	184316976	K/E	Li et al.
CADPS	3	62398847	E/G	Li et al.
ATXN7	3	63942940	K/R	Li et al.
FBXL6	8	145550000	Stop/W	Li et al.
CRB2	9	125172441	T/A	Li et al.
RSU1	10	16898999	M/V	Li et al.
GANAB	11	62153917	Q/R	Li et al.
COG3	13	44988372	I/V	41 (Shah et al.)
NEIL1	15	73433139	K/R	Li et al.
MEX3B	15	80123700	Q/R	Li et al.
ZNF70	22	22417185	Y/C	Li et al.

Table 6: Validated human A-to-I editing sites predicted by REDS

Correctly predicted editing sites by REDS include 10 validated by Li et al., 2009, one by Shah et al., 2009 and 15 experimentally validated in our lab. This list does not include the 22 human validated RNA editing sites that were known before REDS. Table adapted from Maas et al., 2011.

**All chromosomal positions are from human NCBI Build 36.1/hg18*

Editing for several genes after parallel analysis of cDNA and gDNA from human specimen was confirmed *in vivo* (Table 6). In addition, 10 sites recently experimentally validated by Li and co-workers (Li et al., 2009) as well as one site validated by Shah and colleagues (Shah et al., 2009) are also predicted as high-scoring targets by REDS, as are the other human validated RNA editing sites known at the time. Table 7 lists the top ten candidates in human and mouse based on a single, continuously base-paired segment.

a)	Chr	Position of discrepancy	Gene	Accession mRNA	bp	Editing evidence
		hg18				
1	14	94155441	Serpin peptidase inhibitor A3	BC034554	90	In vitro
2	7	149206502	ATPase, H+ transporting V0 subunit	AK094602	72	In vivo
3	1	110057883	Glutathione S-transferase M5	AK127250	50	In silico
4	8	133969994	Thyroglobulin	X05615	46	In silico
5	2	132622172	Hypothetical protein LOC554226	BC045801	26	In silico
6	20	4071869	unknown	BC027448	23	In vivo
7	21	33916232	Crystallin, zeta-like 1	BX648547	22	In silico
8	16	57090613	SMAP-8	AK126729	22	ND
9	11	1599513	Keratin-associated protein 5.4	AB126073	21	In silico
10	11	70927021	Keratin-associated protein 5.8	AY360461	21	In silico

b)	Chr	Position of discrepancy	Gene	Accession mRNA	bp	Editing evidence
1	15	81961974	Coiled-coil domain protein 134	AK154557	36	In silico
2	2	157383876	Blcap; bladder cancer assoc. protein	AK018127	33	In vivo
3	6	51538382	Sorting nexin 10	AK152825	26	In silico
4	7	150745606	CARS, cysteinyl-tRNA synthetase	AK033328	26	In silico
5	2	83500344	Erythropoietin 4	AK217831	24	In silico
6	7	149697247	18-day embryo whole body cDNA	AK003147	24	In silico
7	17	24511429	AT P-binding cassette, sub-family A	AK168428	24	In silico
8	5	23958088	Centaurin gamma 3	AF459091	23	In silico
9	9	59516903	Pyruvate kinase, muscle	AK171106	22	In silico
10	17	45797892	Transmembrane protein 63b	AK217033	22	In silico

Table adapted from Maas et al., 2011.

Table 7: Candidate editing sites with extended continuous base-pairing in a) human and b) mouse

All sites show A/G discrepancy between cDNA and gDNA, exonic region involved in base-pairing, non-synonymous codon position, no mismatches in structure. Only exonic sequences are shown.

In vitro*: edited in transfection experiments of minigene constructs into HeLa cells;* in vivo*: validated in human/mouse specimen;* in silico*: database evidence but* in vivo *editing level not above detection limit of ~5%; ND= not determined due to lack of amplification. *All chromosomal positions are from NCBI Build 36.1/hg18 (human) and NCBI Build 37/mm9 (mouse).*

Figures 11 and 12 show two novel human editing targets that harbor a total of 13 newly validated editing sites. ATP6V0E2 is predicted to form an extended, highly base-paired dsRNA structure with two neighboring segments of 72 and 68bp, respectively, which are not part of any repetitive-type element. Five prominent A-to-I RNA editing sites were experimentally validated, four of which lead to non-synonymous codon changes (Figure 11). Editing levels are between 30% and 70% and the codons are all located within the alternatively spliced exon 3. Based on *in silico* analysis, only about 1.1% of transcripts (EST and mRNA sequences), all of which show evidence of multiple editing sites, contain the alternative exon, which appears to be brain-specific.

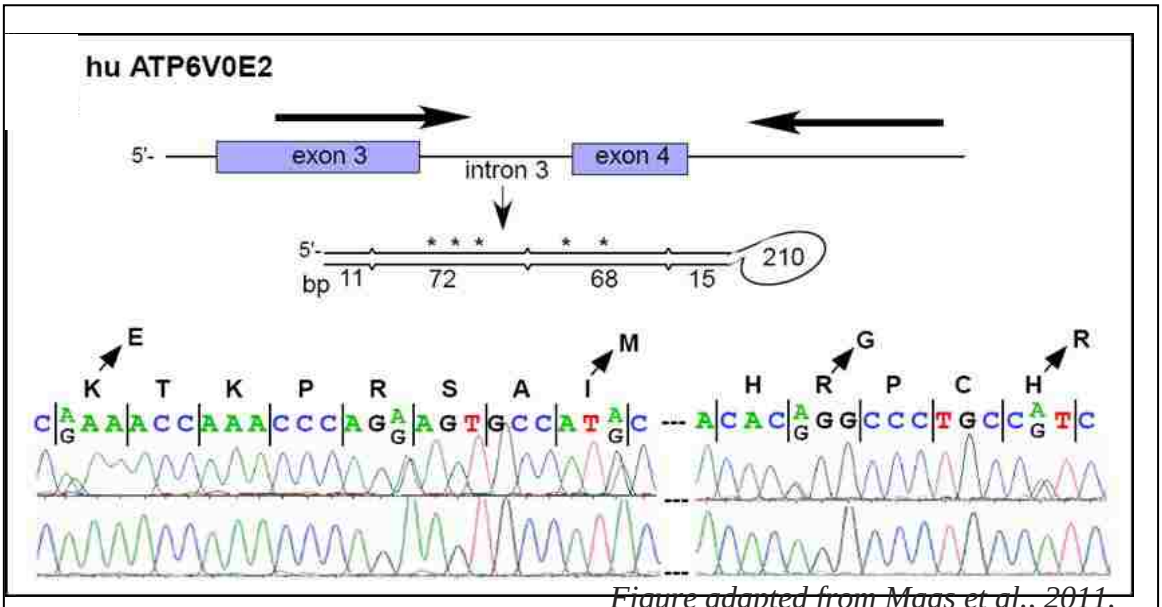


Figure adapted from Maas et al., 2011.

Figure 11: Recoding editing in human ATP6V0E2

Schematic of gene structure and predicted secondary structure of pre-mRNA with main base-paired segments indicated. Positions of editing sites are indicated with an asterisk and the base-pairing regions of exon 3, intron 3 and intron 4 are pointed out by arrows. The nucleotide and amino acid sequence with the five major editing sites (4 of them recoding) are given above the sequence traces for human brain cDNA (top) and gDNA (bottom).

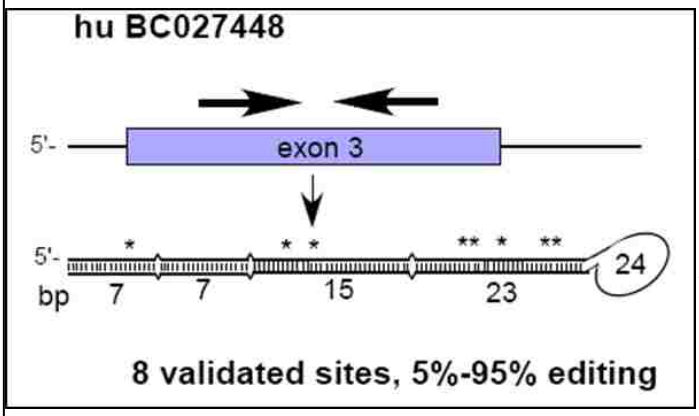


Figure adapted from Maas et al., 2011.

Figure 12: Editing in BC027448

Schematic of pre-mRNA corresponding to human BC027448 around exon 3 with location of major base-pairing features (arrows) and relative location of the eight editing sites that were experimentally detected.

Figure 12 documents editing in transcripts of a gene (BC027448) of unknown function at a total of eight sites with efficiencies between 5 and 95%. It is not known whether the mRNA is translated *in vivo*, or what part of the sequence may constitute a functional open reading frame. The predicted RNA secondary structure forms a hairpin with adjacent segments of 23 and 15 continuous base-pairs entirely positioned within the same exon.

Other members of the lab also identified the human coatomer protein complex subunit alpha (COPA) and cyclin dependent kinase CKD13 as novel editing targets. COPA was predicted as a candidate editing target from analysis of zebrafish databases and was experimentally validated (Maas et al., unpublished). With no edited version of COPA annotated in the human mRNA database, it would have evaded prediction by REDS, had we only screened the human database. Experimental analysis of human brain cDNA showed conserved editing at the equivalent position to that in zebrafish [31[±4]% individual 1; 16[±3]% individual 2; (Maas et al., 2011)]. The I-to-V substitution in COPA is positioned directly following one of its conserved WD-motifs. As editing of COPA transcripts is conserved in vertebrates from zebrafish to human, the COPA recoding event

is likely to have substantial impact on protein function or regulation. COPA is part of the heptameric coatomer complex that defines COPI-type transport vesicles (Nickel et al., 2002). These vesicles function mainly in the early secretory pathway, for instance in the retrograde transport of ER-specific luminal and membrane proteins from Golgi to ER (for review see Beck et al., 2009). COPA (or α-COP) is a WD-repeat (WD40) protein which interacts not only with its coatomer complex partners, but also with signal motifs of proteins sorted into COPI vesicles. The conserved I-to-V amino acid substitution which results from editing occurs exactly adjacent to the C-terminal end of the second WD40 domain and it will be interesting to investigate if coatomer complex formation or interaction of α-COP with other protein ligands is impacted by this recoding event.

Experimental evidence for site-specific recoding editing in the cyclin dependent kinase CDK13 was also obtained, where codon 103 within the ORF undergoes 88% editing in human brain changing a glutamine (Q) to an arginine (R) (Maas et al., 2011, data not shown). This high level of editing in CDK13 suggests a functional role for the predominant edited variant of the gene product. CDK13 has been implicated in splicing regulation based on its interaction with other splicing factors and its intracellular localization in speckles within the nucleoplasm (Even et al., 2006). Interestingly, the localization of CDK13 is dependent on the N-terminal sequence including an RS domain. The editing site maps to the N-terminus within a sequence immediately preceding the RS domain and may overlap with a monopartite nuclear localization sequence (Even et al., 2006).

Figure adapted from Maas et al., 2011.

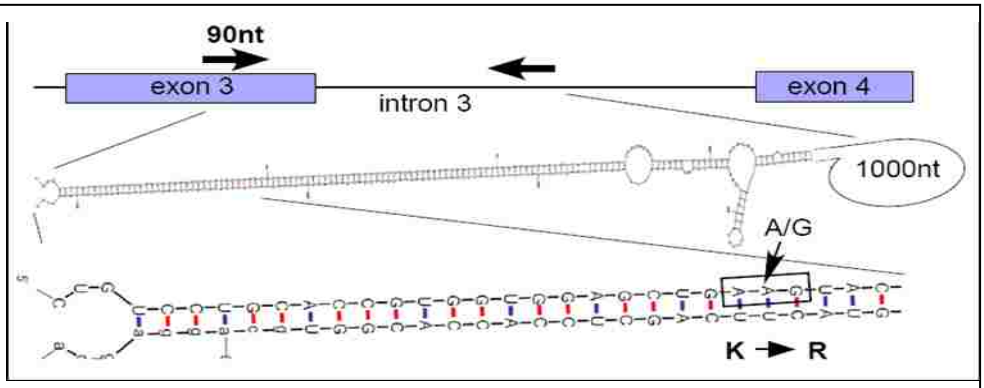


Figure 13: Serpin peptidase inhibitor, clade A, member 3 (SerpinA3)
Schematic of the SerpinA3 gene structure indicating the location of the two complementary sequence regions that yield a perfect 90bp RNA structure. A section that harbors an adenosine position predicted as editing site by REDS is enlarged.

Within the human genome, by far the longest perfectly base-paired RNA structure (90bp) that involves protein-coding sequences is part of the SerpinA3 transcript (Figure 13). Based on known RNA editing recoding targets in human and other organisms and the widespread editing of repetitive element induced RNA fold-back structures (Athanasiadis et al., 2004; Levanon et al., 2004), this extended double-stranded (ds)-structure should constitute a strong binding platform for ADARs, such that multi-site, high-level editing activity is expected, similar to that seen in ATP6V0E2. Surprisingly, no evidence of editing could be detected in SerpinA3 amplified from human tissues (brain and spleen) at the position predicted by the computational screen, nor elsewhere within the exonic sequence forming part of the 90bp structure. However, preliminary results show that when segments of the SerpinA3 gene are expressed in a cell culture system, editing takes place both by endogenous ADAR and to a higher degree when co-transfected with an ADAR2 expression plasmid. *In vivo*, editing may be prevented as a protective measure that ensures preservation of SerpinA3 function. Since the 90bp perfect

duplex would be heavily and promiscuously edited, the multiple amino acid changes and potential ORF disruptions would likely lead to a severely impaired or completely abolished function of SerpinA3. For analysis of SerpinA3 editing regulation see Chapter 7.

3.5 Conclusion

In light of our findings and recent results from a few high-throughput sequencing studies, the majority of the recoding A-to-I modification sites seem to be subject to low to very low editing (Li et al., 2009; Enstero et al., 2010). In fact, the complexity of the current human and mouse mRNA databases is sufficiently high to expect that any recoding editing event with high penetrance and wide tissue distribution should be readily detectable through a screen, such as REDS, that is based initially on mapping A-to-G discrepancies. Indeed, our computational screen predicts many of the known validated editing sites in human. Also, compared to a recent high-throughput sequencing analysis in mouse (Enstero et al., 2010), REDS detected all of those sites located within eight genes that were validated experimentally (Gabra3, Matr3, Ube1x, Xpo7) except those for which no edited sequence variant is represented in the mRNA database. In summary, we conclude that probably few high-level editing sites exist, most of which have now been identified, but a large number of low-level modification events as well as some tissue- or time-specific editing sites remain to be validated.

Compared to previous algorithms used to identify novel A-to-I RNA recoding editing sites, REDS shows several advantages. First, unlike previous bioinformatics approaches, most known editing targets were present in the candidate list due to the greater sequence coverage and unbiased approach of the analysis. Second, user-defined parameter settings allow for optimization of the search algorithm, such that new insights of ADAR preferences can be immediately integrated. Third, REDS has expanded utility as it can also be used to analyze other nucleotide discrepancies.

While we were able to identify several previously unknown A-to-I RNA editing recoding sites, many of the tested candidates did not show evidence of editing in the human brain sample. Large-scale computational analyses such as REDS have recently been complemented by high-throughput sequencing methods in the laboratory. Analyses by REDS would benefit if combined with deep-sequencing analyses of several tissue specimens at different developmental time-points. As discussed previously, editing is regulated in a time- and tissue-specific manner and is amenable to external stimuli. Candidate editing sites with no evidence of editing in our analysis may be edited to a low degree or in a regulated manner, which would be detectable by deep-sequencing.

Analysis of a single feature of ADAR preference, namely the presence of a double-stranded RNA structure, may be insufficient for a ranking method with high predictive force (i.e. where highly ranked candidates are likely to be true editing targets). Improvements of the program should include the analysis of nearest-neighbors and weigh certain sequence signatures that appear to surround edited adenosines as shown recently through the structural analysis of ADAR bound to RNA (Stefl et al., 2010). Other optional features that could strengthen the program are the analysis of sequence conservation around candidate sites and the simultaneous analysis of databases from different species. Additional data information will increase the likelihood of finding true editing sites that are conserved in the candidate list. One good example is COPA, which appeared as a high-ranking candidate in our analysis of the zebrafish but not in the human database. As our understanding of the A-to-I RNA editing mechanism increases together with ever more information in the databases, so will our ability to predict true editing sites.

4 Conserved recoding RNA editing of vertebrate C1q-related factor C1QL1

4.1 Abstract

Complement component 1, q subcomponent-like 1 (C1QL1) was a high-scoring candidate (rank 5) of our SNP-based search and filter strategy (Chapter 2). Here we show that C1QL1 undergoes RNA editing *in vivo*, causing non-synonymous amino acid substitutions in human, mouse, as well as zebrafish. Remarkably, although editing of C1QL1 occurs in different vertebrate species, the predicted RNA secondary structure mediating editing involves different regions in zebrafish versus mammals. However, the predicted RNA folds of *X. tropicalis* and *M. domesticus* (Opossum) resemble neither those of mammals nor zebrafish and are not edited. The editing site could thus have evolved separately in zebrafish and mammals, or editing may have been lost in some species.

4.2 Introduction

Parts of this chapter have been published in Sie & Maas, 2009 (Sie & Maas, 2009). Complement component 1, q subcomponent-like 1 (C1QL1) emerged as a high scoring candidate editing site in our SNP database screen (Chapter 2, Table 4). We experimentally analyzed editing in C1QL1 in a cDNA sample and analysis of single clones derived from the amplicon revealed that the predicted site was edited to 56%, changing a glutamine to an arginine codon (Q66R). In addition to the predicted site, a second site ten nucleotides upstream was also edited (18%, T63A). As we lacked the matching genomic DNA to this cDNA, we refrained from publishing these findings together with the other validated sites from our screen (Gommans et al., 2008). Unfortunately, standard PCR did not allow productive amplification of the human C1QL1 cDNA fragment from other templates, probably due to its high G/C content. The limited amounts of human RNA sample available prompted us to analyze C1QL1 in other species such as the mouse, which afforded us ample material to allow optimization of the PCR reaction conditions.

Known editing sites are often conserved and edited in other species. Such conservation constrains the surrounding sequence, as it has to be maintained in order to facilitate editing by providing a double-stranded binding platform for ADARs. Indeed, bioinformatics screening methods often employ a conservation filter, with which highly conserved sequences surrounding a potential editing site are favored (Levanon et al., 2005; Gommans et al., 2008). We therefore sought to analyze RNA from other species in order to confirm editing of C1QL1.

The sequence surrounding the C1QL1 editing site as well as the corresponding predicted RNA folds are highly conserved in mammals. Analysis of the predicted RNA fold of zebrafish C1QL1 also revealed a highly base-paired region that could potentially serve as editing target. However, the sequence and secondary structure are vastly different from those in mammals. We therefore experimentally analyzed zebrafish cDNA and gDNA and found that the site corresponding to the mammalian Q66R is edited to a high level. To further evaluate conservation of editing at this particular site, we also investigated editing in *X. tropicalis* and *M. domesticus*, two distant species. The predicted RNA folds of these differ from both that of mammals and zebrafish and appear not to be amenable for editing, and indeed experimental validation revealed no editing of C1QL1 in these species. In summary, we show that C1QL1 undergoes A-to-I RNA editing within its open reading frame, leading to non-synonymous codon changes in human, mouse and zebrafish transcripts, but not in *X. tropicalis* and *M. domesticus*.

Even though the degree of conservation has been used as a filter-strategy to search for new editing sites, the way in which editing sites arise and evolve has been a vastly unexplored question in the field. For instance, differential editing occurs in potassium channels of octopus. RNA editing is regulated depending on the environment, specifically the temperature, where the octopus is living (Dr. J. Rosenthal, personal communication). The plastic nature of pre-mRNA secondary structure may allow continuous probing of potentially beneficial editing sites. Accordingly, an editing event would become engraved if it conferred an adaptive advantage under a given selection pressure, and only then would be edited at relatively high levels (Gommans et al., 2009).

4.3 Materials and Methods

4.3.1 Databases and data analysis

RNA secondary structures were predicted using the M-fold algorithm (Zuker, 2003) and multiple sequence alignments were done with clustal W 1.8. Expressed sequence tag (EST) analysis for human, mouse and rat were performed using the NCBI BLAST server.

4.3.2 cDNA and genomic DNA

Human brain total RNA and gDNA isolated from the same specimen were obtained from Biochain, CA. For analysis of mouse C1QL1, total RNA and genomic DNA from cortex and cerebellum of two adult mice were prepared using standard procedures (Appendix A). *Danio rerio* gDNA and total RNA were isolated from adults and 4 day hatchlings using the same procedures (Appendix A). RNA was treated with DNase and processed using a standard protocol for reverse transcription (Appendix A). One specimen of *X. tropicalis* was sacrificed by hypothermia and dissected and gDNA and total RNA were isolated as before. *M. domesticus* tissue samples in RNAlater were kindly provided by Dr. Samollow, Texas A&M University, and processed to obtain gDNA and RNA with the same standard procedures as described.

4.3.3 PCR

For experimental validation, gene specific fragments of cDNA as well as genomic regions from the same specimen were amplified by PCR, gel-purified (QIAEXII Gel Extraction Kit, Qiagen) and subjected to dideoxy sequencing (Geneway Research) (Athanasiadis et al., 2004).

For amplification of part of the human C1QL1 comprising the putative editing site, the reaction mixes contained 400nM of each primers (*Homo sapiens*: C1Q14D and C1Q13U; *Mus musculus*: mC1Q7D and mC1Q8U ; *Danio rerio*: drC1Q15D and drC1Q16U, *M. domesticus*: mdC1Q17D and mdC1Q18U; *X. tropicalis*: xtC1Q19D and xtC1Q20U; for primer sequences see Appendix B), 2µl Phire™ Hot Start DNA Polymerase (NEB), 400nM dNTP mix, 3% DMSO (5% for *Homo sapiens*), 1µl cDNA, and Phire polymerase buffer provided by the manufacturer in a total volume of 100µl. The reactions were carried out in an Eppendorf Mastercycler. Cycling conditions included a 2 minute initial denaturation step followed by 35 cycles of:

	H. sapiens	*M. musculus*	*D. rerio*	*M. domesticus*	*X. tropicalis*
Denaturing	98°C, 10”	98°C, 10”	98°C, 10”	98°C, 10”	98°C, 10”
Annealing	71°C, 5”	72°C, 5”	69°C, 5”	70°C, 5”	72°C, 5”
Extension	72°C, 10”	72°C, 15”	72°C, 12”	72°C, 12”	72°C, 10”

followed by a final extension step of 72°C for 1 minute. If yield was too low, a secondary PCR with the same conditions was performed, using 1µl of the first PCR as template. The PCR products were purified by phenol-chloroform extraction, ethanol-precipitated, and subjected to DNA gel electrophoresis on a 2% agarose gel. The band of the expected size was excised and purified using the QIAEX II Gel Extraction Kit (QIAGEN). The purified products were sequenced.

4.3.4 Subcloning

For further analysis, PCR products were subcloned into pBluescript II (Stratagene) vector and individual DNA templates were purified and sequenced. Briefly, a PCR reaction using 400nM of each of the primers (human: C1Q14D-Eco and C1Q13U-KpnI; mouse: mC1Q7D-Eco and mC1Q8U-KpnI; for primer sequences see Appendix B)

was performed on 2.5% of the purified amplicons, with PCR conditions as described above for human and mouse, respectively. The amplicons were restricted with 40U KpnI (NEB cat# R0142) for 3 hours at 37ºC, subsequently purified by phenol-chloroform extraction, ethanol-precipitated and restricted with 40U EcoRI (NEB) for 3 hours at 37ºC. Again, the cut fragments were purified by phenol-chloroform extraction and precipitated with ethanol, and then subjected to DNA gel electrophoresis on a 2.5% agarose gel. The bands of the expected size were excised, purified using the QIAEX II Gel Extraction Kit (QIAGEN) and then ligated into a pBluescript SK II vector also cut with EcoRI and KpnI and purified as described for the amplicon. The ligation reactions contained vector and insert in a 1:8 molar ratio and 0.5μl T4 ligase (Invitrogen) in a total volume of 10μl and was incubated at room temperature 4 hours to overnight. Z-competent DH5α cells were transformed with 5-10μl of the ligation. Heat-shock was performed for 1.5min at 37ºC. The transformed cells were plated on LB containing ampicillin. Individual recombinant clones were used to inoculate liquid LB-amp and the purified plasmids (QIAprep Spin Miniprep Kit, QIAGEN) were sequenced.

4.4 Results and Discussion

4.4.1 A-to-I RNA editing in mammalian C1QL1 is conserved

Within the highest scoring group of predicted target sites derived from our study outlined in Chapter 2, we showed that three out of four positions are bona fide RNA editing recoding sites which affect two genes, splicing factor SRp25 and insulin-like growth factor binding protein IGFBP7 (Gommans et al., 2008). None of the lower scoring candidates that we evaluated experimentally (total of 68 sites) showed detectable RNA editing in human brain tissue. As standard PCR did not allow productive amplification of the human C1QL1 cDNA fragment, probably due to its high G/C content, we used abundant total mouse RNA to optimize reaction conditions for a RT-PCR protocol that would allow RNA editing analysis. Not only is the mouse C1QL1 cDNA highly conserved to the human sequence, but the predicted RNA secondary structures of mouse, rat and human exon 1 sequences are also the same (Figure 14). Therefore, we hypothesized that RNA editing at the projected position in human would be conserved in its mouse orthologue.

Use of a DNA polymerase with high processivity, a buffer formulated to support amplification of GC-rich templates, and an optimized amplification protocol allowed us to obtain a specific amplicon for mouse C1QL1 cDNA. The analysis of purified cortex and cerebellum samples revealed three positions of RNA editing within the same exon of C1QL1, all of them effecting an amino acid substitution. The Q66R site had been predicted by our computational screen (Chapter 2, Table 4; Gommans et al., 2008), while the others alter a threonine (ACG) to an alanine (GCG) and a glutamine (CAG) to arginine (CGG) codon, respectively. Table 8 summarizes the editing levels measured at

the three sites within the two mice. Intriguingly, editing levels in cerebellum are substantially different from those in cortex, arguing for tissue-specific regulation of editing. The Q66R site is edited to 10% or 17% in cerebellum, whereas it is edited to only 1-3% in cortex. Q69R only showed evidence of editing in cerebellum (3-7%) and the T63A site is modified at 1-2% in both cortex and cerebellum. Taken together, we confirmed that all three sites in mouse C1QL1 undergo RNA editing *in vivo.*

Sample origin	Number of clones analyzed	T63A % editing	Q66R % editing	Q69R % editing
Cortex mouse 1	66	1.5	1.5	-
Cerebellum mouse 1	54	1.8*	17.0	3.6
Cortex mouse 2	38	-	2.6	-
Cerebellum mouse 2	40	2.5*	10.0	7.5

Table 8: Analysis of C1QL1 RNA editing in mouse specimen by subcloning

Amplicons from mouse cortex and cerebellum were subcloned into pBluescript vector. Sequencing of single clones shows differential editing at the Q66R site with 10-17% editing in cerebellum and only low levels in cortex. T63A and Q69R are edited to a low extent and may not be edited at all in certain tissue samples (for example cortex mouse 2).

**T63A is only edited together with Q66R site, which correlates with results from first analyzed human cDNA (Chapter 2).*

We then moved to the analysis of several samples of human cDNA, applying the optimized protocol. As discussed in Chapter 2, one human brain cDNA showed high levels of modification at T63A (18%) and Q66R (56%), respectively (Figure 8). As in mouse, codon 63 was always edited concomitantly with codon 66. However, the genomic counterpart for this human specimen was not available. Subsequent subcloning analysis of additional specimens from human brain together with the corresponding genomic DNA confirmed the occurrence of RNA editing at the Q66R position, as the genomic

samples displayed an adenosine (Figure 14). However, editing at the T63A site was not detectable in any of the samples (Table 9).

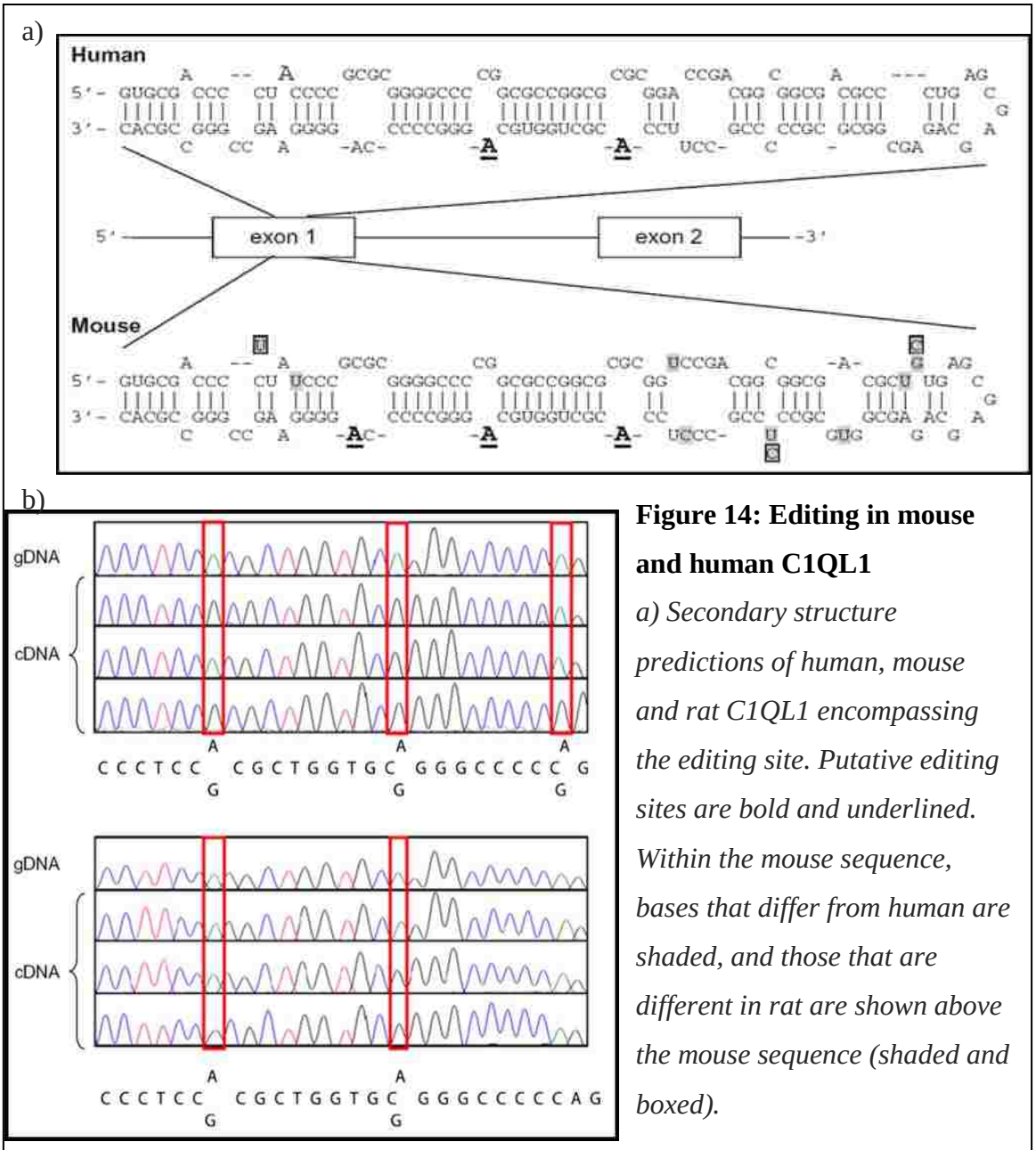


Figure 14: Editing in mouse and human C1QL1

a) Secondary structure predictions of human, mouse and rat C1QL1 encompassing the editing site. Putative editing sites are bold and underlined. Within the mouse sequence, bases that differ from human are shaded, and those that are different in rat are shown above the mouse sequence (shaded and boxed).

b) Representative sequence tracks from single clones of C1QL1 from mouse cerebellum (top) and human brain specimen (bottom). Genomic sequences are shown above cDNA sequence tracks. Editing sites are boxed.

Sample number	No. of clones	T63A site edited	Q66R site edited
unknown	45	18%	60%
B105092	42	-	14%
A802100	33	-	-
B105090	32	-	-
B105092	32	-	3%

Table 9: Analysis of C1QL1 RNA editing in human brain specimen by subcloning

Amplicons from human brain samples were cloned into pBluescript vector. Sequencing of single clones confirms editing at the Q66R site with 3-14% editing in amplicons derived from one specific specimen (B105092). The other two analyzed samples displayed no editing in any of the sequenced clones (A802100 and B105090).

Although we cannot rule out that the T63A site represents a previously unknown recoding gSNP in human, our results from analysis in mouse, the complete conservation of the predicted RNA secondary structure surrounding the editing sites, as well as the observed coupling between major and minor editing sites argue that, like in mouse, the minor position is also an A-to-I target in human.

The observed variation in editing levels at both the major and minor sites in human specimens may be due to regional and/or temporal regulation of C1QL1 editing, similar to what is observed for other recoding editing targets, such as glutamate receptor transcripts (Paschen & Djuricic, 1995). This assumption is supported by the tissue-specific pattern of editing in mouse brain tissue described above. We also analyzed human spleen, lung, kidney, heart, and pancreas RNA samples for editing in C1QL1, but did not detect editing above the detection limit of 5% for sequence track analysis Figure 15).

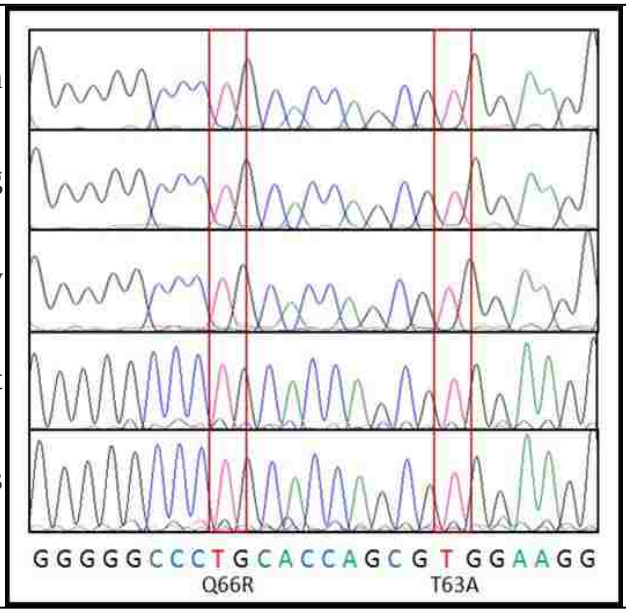


Figure 15: C1QL1 editing analysis of diverse tissues
Sequence tracks (reverse shown) of C1QL1 amplified from various human tissue samples. No evidence of editing is apparent from this analysis.

4.4.2 A distinct RNA fold supports zebrafish C1QL1 editing

The C1QL1 exon 1 sequence is strongly conserved between mammalian species (Figures 14 and 16) which suggests that in addition to the human and mouse gene, also the rat, cow and dog C1QL1 RNA is likely subject to editing. However, we noticed that the predicted secondary structure supporting editing in mammalian C1QL1 is not conserved within the zebrafish (*Danio rerio*) orthologue (Figure 16). The editing site complementary sequence (ECS) within human exon 1, located 50 nucleotides upstream of the recoding editing sites, is not conserved in any of the non-mammal sequences including zebrafish. However, in zebrafish, another RNA fold of similar strength is formed with sequences downstream of the recoding sites within exon 1. Indeed, when we analyze RNAs isolated from adult and four day post-fertilization zebrafish specimens, we readily detect editing at the Q74R site (equivalent to human Q66R) at about 50% in adult and 33% in hatchlings (Figure 16). The distinct RNA fold predicted for the zebrafish sequence is supported by the fact that two additional adenosines located on the opposite

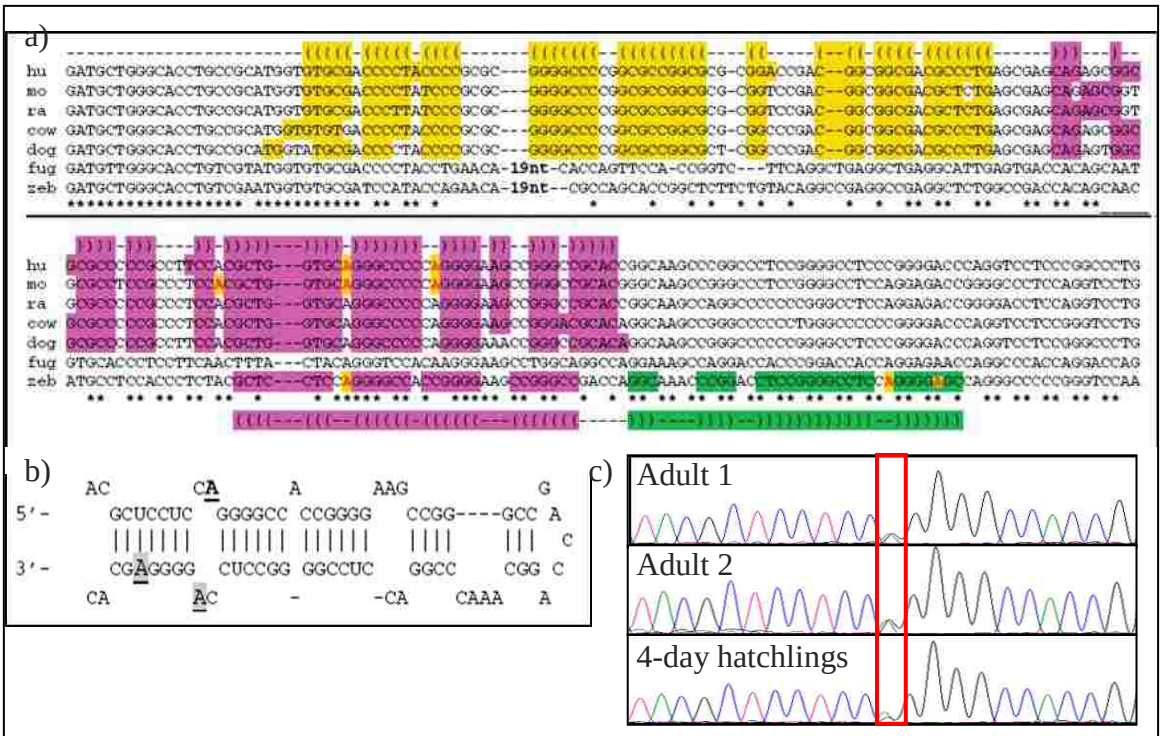


Figure 16: Clustal W (1.81) alignment of vertebrate C1QL1 sequences and RNA secondary structures that mediate editing

a) Base-paired nucleotides in the mammalian structure are depicted in purple for the sequence surrounding the recoding editing sites and in yellow for the upstream, complementary region. Nucleotides involved in base-pairing in the zebrafish RNA structure are shown in purple (region around the recoding editing site) and green (downstream complementary region). Adenosines that were experimentally shown to be subject to editing are shown in red font with yellow shading. Nucleotides identical across all displayed species are indicated by a star below the alignment; base-pairing nucleotides are further highlighted through bracket notation. b) Zebrafish RNA secondary prediction (Mfold). The edited adenosine corresponding to the Q66R site is highlighted in bold and underlined. Adenosines edited on the opposite strand of the duplex are highlighted in bold and gray shade. c) C1QL1 sequence tracks of cDNA derived from two adult (top) and 4-day old hatchlings (bottom) zebrafish. The edited adenosine corresponding to Q74R is boxed. gDNA sequence track showed absence of SNP at the corresponding position, validating it as a true editing site (not shown). Figure adapted from Sie & Maas, 2009.

site of the predicted duplex also undergo editing. In contrast, human C1QL1 does not show any evidence of editing at the downstream adenosines.

4.4.3 C1QL1 editing is not conserved in other species

While the predicted RNA folding patterns of most mammalian sequences are similar if not identical to each other, those of evolutionarily more distant species like *X. tropicalis*, *Tetraodon nigroviridis*, *Takifugu rubripes* (fugu), *Gasterosteus aculeatus* (stickleback), *Oryzias latipes* (medaka) and *Danio rerio* (zebrafish) differ considerably from that of mammals and amongst each other (comparison of sequence homology between different species and homo sapiens C1QL1 see Table 10).

These differences can partly be explained by sequence insertions of up to 15 nucleotides upstream of the editing target site in addition to sequence diversification (see as example fugu and zebrafish in sequence alignment of Figure 16). Of the 10 mammalian sequences analyzed by Mfold, only the predicted structure of opossum differs considerably from that of human. We analyzed C1QL1 from various tissues of both *M. domesticus* and *X. tropicalis*, whose secondary structure predictions are different from both human and zebrafish (Figure 17a), and found that none of the sequence tracks showed detectable editing levels (Figure 17b).

% Identity to Homo sapiens	*aa level*	*nt level (coding sequence exon1)*	*nt level (coding sequence only)*
Macaca mulatta	*99%*	*98%*	*98%*
Mus musculus	*98%*	*97%*	*93%*
Rattus norvegicus	*98%*	*97%*	*92%*
Bos Taurus	*98%*	*94%*	*93%*
Canis famiiaris	*98%*	*94%*	*94%*
Sus scrofa	*96%*	*94%*	*94%*
Monodelphis domestica	*93%*	*84%*	*84%*
Xenopus tropicalis	*82%*	*71%*	*73%*
Tetraodon nigroviridis	*79%*	*ND**	*ND**
Danio rerio	*75%*	*75%*	*74%*

*ND = not determined due to lack of nucleotide sequence

Table 10: Identity of C1QL1 sequences among species

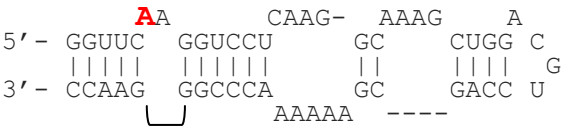
Identity of amino acid and nucleotide sequences, respectively, of C1QL1 in ten species compared to homo sapiens C1QL1.

Figure 17: Analysis of *X. tropicalis* and *M. domesticus* C1QL1 editing

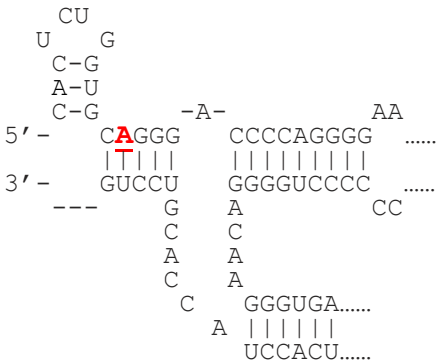
a) Predicted secondary structures of X. tropicalis *and* M. domesticus *encompassing the homologous region that is edited in mammals and zebrafish.*

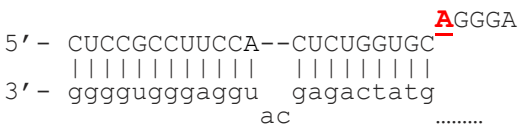
b) Representative sequence tracks of X. tropicalis *and* M. domesticus *shows no evidence of editing of C1QL1 in these species.*

a) *X. tropicalis*

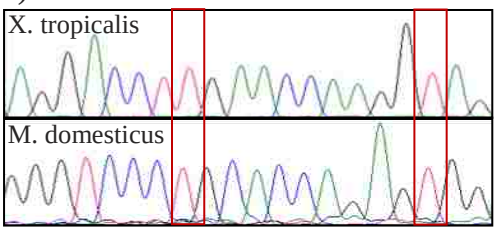


M. domesticus (two alternative structures shown)





b)



4.5 Conclusions

Our findings validating mammalian C1QL1 as a bona fide A-to-I RNA editing target further highlights the effectiveness of our bioinformatics search strategy as applied to the subset of human mRNAs with non-synonymous A-to-G discrepancies chosen from the SNP database. Four of the top five highest scoring sites prove to be *in vivo* editing targets, whereas none of the tested sites with lower scores (an additional 64 positions tested) show detectable editing.

Editing of C1QL1 occurs in human, mouse and zebrafish, where it is facilitated by a different RNA fold. Editing may also occur in other mammalian species, due to a conserved RNA fold. However, editing in this target probably does not extend throughout the animal kingdom, as analysis of C1QL1 of opossum and Xenopus has shown. Further investigation of additional species would be interesting, in particular those with similar RNA secondary structures to that of zebrafish. However, the limited number of species with annotated C1QL1 sequences currently prevents such an undertaking.

The family of C1Q-domain proteins includes important signaling molecules with roles in inflammation, adaptive immunity and energy homeostasis (Ghai et al., 2007). The physiological function of C1QL1 has not been elucidated, but it is expressed highest within the brain, particularly in the brainstem parts of the cerebellum (Iijima et al., 2010), and was suggested to be especially important for neurons involved in coordination and regulation of motor control (Berube et al., 1999). Furthermore, it may be part of a neuroprotective immune response and is expressed from glia cells (Glanzer et al., 2007). One study revealed upregulation of C1QL1 in response to kainic acid induced seizures (Hunsberger et al., 2005). All four C1QL family members are secreted, form homo- and

heteromers and oligomerize to hexamers and high-molecular weight complexes (Iijima et al., 2010). Members of the closely related Cbln family were shown to act as trans-neuronal regulators of synaptic integrity by stabilizing synaptic contacts and controlling functional synaptic plasticity by regulating the postsynaptic endocytosis pathway of AMPA receptors (Yuzaki, 2008). Reminiscent of such a role, the C1QL family members may also act as neuronal cytokines. It is speculated that the degree of multimerization and heteromer formation may allow for the activation of different receptors and functions (Iijima et al., 2010). The T63A and Q66R amino acid substitutions may impact protein oligomerization as they are situated immediately prior to a collagen-like trimerization domain (Figure 8). In other C1Q-domain proteins, such as the hormone adiponectin, this also coincides with a region of protease-mediated processing (Waki et al., 2005). Future studies will show if the amino acid substitutions caused by RNA editing may alter post-translational processing of C1QL1, or if it affects other properties of the protein *in vivo*.

It remains to be elucidated why editing of C1QL1 occurs to such a high level in zebrafish. High levels of editing often point toward a functional consequence for the ensuing protein variants. It is especially intriguing that the same codon in mammals and zebrafish is targeted by ADARs, yet the double-stranded structures that support editing are vastly different from each other. This suggests that the editing sites evolved independently of each other, which again strongly implies a functional impact on the encoded protein. Due to the current incomplete picture of C1QL1 function, possible consequences will have to be assessed at a later date.

5 Consequences of RNA editing on FLNA-protein interactions

5.1 Abstract

Filamin A (FLNA) organizes cytoskeletal F-actin into crosslinked networks and tethers them to the cell membrane. It is indispensable for maintaining and remodeling the cytoskeleton to effect changes in cell shape and migration, whereby it serves as a molecular scaffold by interacting with various proteins. FLNA has a N-terminal actin-binding domain followed by 24 Immunoglobulin (Ig)-like repeats. The 24th repeat serves as a dimerization domain. FLNA is edited in its 22nd C-terminal repeat, The codon change alters an uncharged Glutamine (Q) to a positively charged, bulkier Arginine (R). The two protein isoforms (FLNA and FLNA-Q/R, respectively) are produced in one cell from the same allele. Editing in FLNA is highly conserved and regulated in a tissue-specific manner, suggesting that the amino acid change produces a functionally distinct protein isoform. Here we tested the ability of the two isoforms to bind known interaction partners in a yeast-two-hybrid assay to determine if editing alters binding affinity to any of them.

5.2 Introduction

Filamins were discovered as a family of non-muscle actin-binding proteins in the 1970s. The filamin family encodes three isoforms in mammals, Filamin A, B and C (FLNA, FLNB and FLNC). They consist of an amino-terminal actin-binding domain (ABD) composed of two tandem calponin homology domains (CHD1 and CHD2), followed by a long rod of 24 repeated, anti-parallel β-strands that adopt an immunoglobulin-like fold (Nakamura et al., 2007). The long rod is interrupted by two flexible hinges between repeats 15 and 16 and 23 and 24, respectively (Feng & Walsh, 2004). Dimerization occurs through the last carboxyl-terminal repeat, mediating the formation of a V-shaped structure. These high-molecular weight cytoplasmic dimers serve as structural proteins that link cortical actin filaments into a dynamic three-dimensional structure (Nakamura et al., 2007). The actin cytoskeleton is not only essential for the maintenance of cell shape and motility, but also for the integration of cell signals that initiate and propagate alterations in the cytoskeleton. FLNA has been found to interact with a multiplicity of transmembrane and peripheral membrane proteins, of which more than 30 have been identified (Ohta et al., 1999; Robertson, 2005; Ohta et al., 2006; Zhu et al., 2007). Furthermore, either as a full-length protein or in its cleaved form, it has also been shown to colocalize with transcription factors and nuclear receptors (Popowicz et al., 2006).

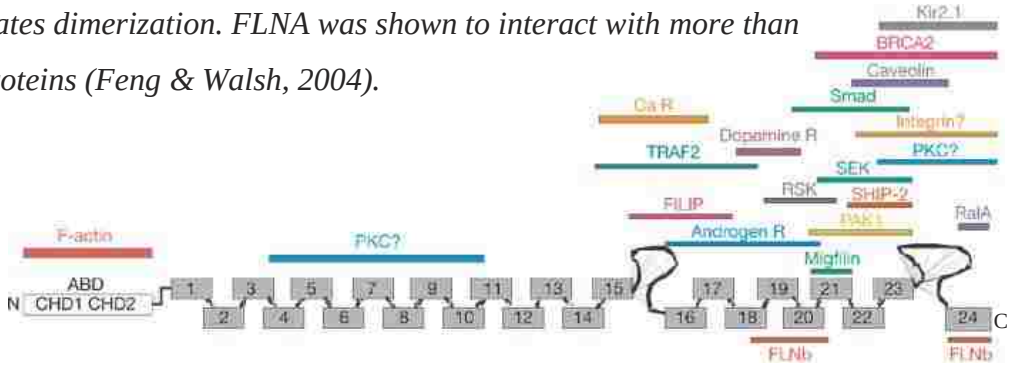
Genetic mutations in FLNA and FLNB have been shown to cause a wide range of human X-linked diseases. Null-mutations cause dysfunctional neuronal migration, which manifest themselves as periventricular heterotopias characterized by a partial failure of neuronal migration into the cerebral cortex with consequent formation of ectopically

placed neuronal nodules in the ventricular and subventricular zones (Guerrini & Carrozzo, 2001; Sheen et al., 2002; Sheen et al., 2005; Sarkisian et al., 2008). Missense-mutations in FLNA are connected with the spectrum disorder otopalatodigital syndrome which includes otopalatodigital syndromes types 1 and 2, frontometaphyseal dysplasia, and Melnick-Needles syndrome (Robertson, 2005). The disorders are a phenotypically heterogeneous group of conditions characterized by a skeletal dysplasia and variable anomalies in the brain, craniofacial structures, cardiac, genitourinary and gastrointestinal systems (Robertson, 2005). More recently, some mutations in FLNA have also been associated with the connective tissue disorder Ehlers-Danlos syndrome, in which affected individuals present joint and skin hyperextensibility and vascular problems (Sheen et al., 2005).

The ~96 amino acid long repeats forming the antiparallel, partially overlapping β-strands are highly homologous. However, β-strands 16-24 have been shown to pack more tightly than those in the N-terminal portion of the protein (Nakamura et al., 2007), some of them apparently folding onto neighboring repeats in a way that may regulate binding to other proteins (Lad et al., 2007). FLNA mRNA was shown to be subjected to A-to-I RNA editing, changing a glutamine to an arginine codon (Q2341R) in its repeat 22 at amino acid 2341 (Levanon et al., 2005). No functional consequences of this amino acid change have yet been reported. Several proteins are known to interact with the C-terminus of FLNA (Feng & Walsh, 2004) (Figure 18), shown either by a yeast-two-hybrid or similar protein binding assay. Here we assessed whether the altered amino acid in FLNA impacts its interaction with one or more of its known binding partners. We selected proteins for analysis based on reports in the literature that indicate that repeat 22

Figure 18: FLNA domain structure and binding proteins

Each subunit of the FLNA homodimer has a C-terminal actin-binding domain composed of two CHD domains followed by 24 antiparallel β-sheets interrupted by two hinge-regions. The N-terminal 24th repeat mediates dimerization. FLNA was shown to interact with more than 30 proteins (Feng & Walsh, 2004).



of FLNA might be required for that specific interaction. Since FLNA acts as a molecular scaffold, we expect that alterations in its binding capability will have significant impact on the functional capacity of the resulting signaling complexes. To test this hypothesis, proteins that interact with fragments of FLNA that include repeat 22 were chosen for analysis in the Matchmaker™ GAL4 Two-Hybrid System 3 (Clontech). The reported interaction proteins were genetically fused to the GAL4-DNA binding domain as bait, and FLNA or FLNA-Q/R was fused to the GAL4-activation domain as prey. Interaction between the fusion proteins was assessed using β-galactosidase assays.

5.3 Materials and Methods

5.3.1 Cloning Procedures

Cloning procedures are outlined in Appendix A, primer sequences can be found in Appendix B. Briefly, the following gene fragments were ligated into the corresponding restriction enzyme sites of the indicated vectors:

Gene	Vector	5' restriction site	3' restriction site	Expressed fragment
FLNA	pGADT7	Sfi1	Xho1	C-term. 476 aa
Integrinβ1	pGBKT7	BamH1	Pst1	C-term. 47 aa
Presinilin	pGBKT7	BamH1	Pst1	aa 259-407
mGluR5a	pGBKT7	BamH1	Pst1	aa 827-933
mGluR7b	pGBKT7	BamH1	Pst1	aa 900-922
mGluR8a	pGBKT7	BamH1	Pst1	aa 855-908
P73a	pGBKT7	Nde1	EcoR1	aa 464-636
P2Y2	pGBKT7	BamH1	Pst1	aa 318-362
Calcitonin Receptor	pGBKT7	BamH1	EcoR1	aa 390-474
BRCA2	pGBKT7	BamH1	Pst1	aa 187-354
Smad5	pGBKT7	Nde1	BamH1	Insertion mutation, non-functional
SEK1	pGBKT7	Nde1	BamH1	Cloning not successful
FLNA Q/R	pGADT7	Retrieved by site-directed mutagenesis of FLNA		
Integrinβ1-Y788E	pGBKT7	Retrieved by site-directed mutagenesis of Integrinβ1		

5.3.2 Yeast Transformation

2x50ml of YPDA was inoculated with three ~3mm large colonies each and incubated at 30˚C for 18 hours, shaking, to OD>1.2. The overnight cultures were diluted to OD 0.25 into 150-250ml YPDA and incubated at 30˚C for 3 hr, shaking, to OD ~0.5. The cell suspension was distributed into 50ml falcon tubes and centrifuged at 1000g for 5 minutes at room temperature. The supernatant was discarded and cells were pooled and resuspended in 2x25ml sterile TE. Cells were centrifuged again for 5 minutes at 1000g,

the supernatant decanted, and the pellet resuspended in 2x1.5ml sterile 1xTE/LiAc solution. In 1.5ml eppendorf tubes, 200ng of pGBKT7-bait plasmid was mixed with 100ng pGADT7-fusion protein plasmid and 100μg herring testes carrier DNA. 100μl yeast competent cells and 600μl sterile PEG/LiAc solution were added and vortexed to mix. The cells were incubated at 30˚C for 30 minutes. 70μl DMSO was added and mixed by inversion, followed by a 15 minute heat-shock at 42˚C. Cells were chilled on ice for 2 minutes, centrifuged briefly, the supernatant removed, and resuspended in 500μl 1xTE. 100–200μl transformed cells were plated on SD –Leu/-Trp plates and incubated at 30˚C for 2-3 days.

5.3.3 β-Galactosidase Colony Lift Assay

Single colonies from the transformation were streaked out on selection medium (SD –Leu/-Trp). The plates were incubated at 30°C for 2-3 days. Whatman filter paper was placed onto the plates and carefully flattened to pick up yeast patches, removed and submerged into LN_2 for 10 seconds to break open the cells. With the cells facing upward, the paper was then placed onto a filter paper pre-soaked in Z-buffer/X-Gal solution. Color was allowed to develop for 4 hours. Frequent photographs were taken to monitor color development over time.

5.3.4 ONPG Assay

Fresh (<3 weeks old) colonies of single-colony purified yeast transformants were submerged into 1ml SD-Trp/-Leu selection medium, vigorously vortexed to disperse the cells and transferred to a 5ml culture. Cultures were incubated at 30°C for 16-20 hours on a rotating wheel. 1ml of overnight culture was transferred to 4ml YPD medium and

incubated for 3-5 h at 30°C until cells were in mid-log phase (OD_{600} = 0.5-0.8). All FLNA-expressing yeast cells had slowed growth and were thus used directly at OD_{600} = 0.5-0.8 from the overnight culture. OD of each culture was recorded before harvest. 3x1.5ml of cultures were centrifuged at maximum speed for 30 seconds. Supernatants were removed and cells resuspended in 1.5ml Z buffer. After centrifugation, pellets were resuspended in 200µl Z buffer (concentration factor is 1.5/0.2 = 7.5 fold). 3x62.5µl of cell suspension was distributed to fresh microcentrifuge tubes. Cells were lysed by three freeze/thaw cycles of 5-10 minutes at -150°C followed by 2 minutes at 37°C. 437.5µl Z-buffer/β-mercaptoethanol was added to reactions and blanks, which contained 62.5µl Z-buffer. Time was recorded and 100µl ONPG in Z buffer was added to each reaction and blank. The reaction was allowed to proceed at 30°C until yellow color developed, anywhere between 30 minutes and 4 h. To stop the reaction, 250µl 1M Na_2CO_3 was added. Elapsed time for each reaction was recorded. Tubes were centrifuged for 10 minutes at maximum speed to pellet cell debris and supernatants were transferred to cuvettes. The spectrophotometer was calibrated against the blank at A_{420} and OD_{420} of samples were measured relative to the blank. β-galactosidase units were calculated according to the Yeast Protocols Handbook (Clontech):

β-galactosidase units = $1{,}000 \times OD_{420} / (t \times V \times OD_{600})$

t = elapsed time in minutes until reaction was stopped

V = 0.0625 ml x 7.5 (concentraction factor)

OD_{600} = A_{600} of 1ml culture (input)

5.4 Results and Discussion

5.4.1 Colony-lift assay

To assess differences in binding between FLNA and FLNA-Q/R protein variants we used the Matchmaker™ GAL4 Two-Hybrid System 3 of Clontech according to the manufacturer's instructions. Briefly, the phenotype of *S. cerevisiae* strain Y187 (phenotype: Matα, ura3-52, his3-200, ade2-101, trp1-901, leu2-3, 112, gal4Δ, met-, gal80Δ, URA3::$GAL1_{UAS}$-$GAL1_{TATA}$-lacZ) was verified by assessing growth on selection plates (-Ura, -His, -Leu, -Trp, -Leu/-Trp) before using them for transformation. As a control, each plasmid was transformed together with an empty vector contributing either Trp or Leu resistance, respectively, so that all transformants could be selected on SD –Leu/-Trp plates. These controls were necessary to determine if the single fusion proteins were able to activate expression of the reporter gene by themselves. Additional controls included a vector that constitutively expresses LacZ (pCL1) and can thus be used for verification of the β-Galactosidase assay, as well as GAL4 fusion proteins that show interaction and activate lacZ transcription when co-expressed (pGADT7/T-antigen and pGBKT7/p53) or fusion proteins that do not interact and are used as negative controls (pGADT7/T-antigen and pGBKT7/Lam). A table of the plasmid combinations for the co-transformations is outlined in Appendix A.

All transformants that express the FLNA fusion protein or its edited variant showed considerably slower growth than other transformants. This reduced growth resulted in noticeably smaller colonies on the transformation plates. On every plate with such small colonies, several (<1%) larger colonies also appeared. To test whether these larger colonies represented transformants that had lost expression of the Gal4-AD-FLNA

fusion protein, a subset of them were also single-colony purified and tested in the colony-lift assay (see below), namely mGluR5a, mGluR8a, p73a, P2Y2, BRCA2, and CTR, co-transformed with either Gal4-AD-FLNA or Gal4-AD-FLNA-Q/R, respectively. The yeast transformants originating from these larger colonies were unable to activate the reporter gene, in contrast to the transformants purified from small colonies. The large colonies therefore represent revertants that may have lost the ability to express the FLNA fusion proteins, which would be consistent with the lack of reporter gene activation. Forthwith, all subsequent transformants were single-colony purified from small colonies for testing in yeast-two-hybrid assays. However, the finding that some cells may be able to lose the ability to express the FLNA fusion proteins was noted.

A β-galactosidase colony-lift assay was used to assess activation of the reporter gene lacZ. On every plate, all three controls (pCL1, pGADT7-T-antigen + pGBKT7-p53, and pGADT7-T-antigen + pGBKT7-Lam) were present at least once, and each experiment was done in triplicate. Colonies of double transformants spread out in $1cm^2$ patches did not grow at comparable rates; FLNA co-transformation slowed yeast growth considerably, and thus these strains had to be incubated 1-1.5 days longer to attain a similarly dense cell lawn.

The positive controls provided with the kit developed a strong signal every time, while the negative controls did not, indicating lack of interaction between Lam and T-antigen. Most control transformations with one fusion protein and a complementing empty vector also showed no signal development, indicating that the fusion-proteins by themselves are unable to activate gene transcription of the reporter gene. However, the negative controls expressing only pGADT7-FLNA or pGADT7-FLNA-Q/R alone

showed color development, indicating that these fusion proteins are able to activate transcription on their own, without interacting with a pGBKT7-bait fusion protein. Interestingly, background activation as seen by pGADT7-FLNA or pGADT7-FLNA-Q/R is almost entirely abrogated when a GAL4-DNA-binding fusion protein that does not interact with FLNA is co-expressed (i.e. pGBKT7-Integrinβ1-Y788E, mutant form of integrinβ that cannot bind FLNA). It seems that the BD/fusion protein outcompetes binding to the promoter region that AD/FLNA may bind to and so eliminates background activation of gene expression.

All ten interaction partners that were tested in this manner activated the reporter gene to a similar extent when expressed with either pGADT7-FLNA or pGADT7-FLNA Q/R, indicating that the proteins interact with both edited and unedited versions of FLNA. However, the colony-lift assay can only be used as a qualitative measure of protein-protein interactions and does not provide a quantitative assessment of binding strength. In order to detect more subtle differences in binding, quanitative assays such as the liquid ONPG assay or pull-down experiments have to be applied. We therefore decided to assess protein-protein binding strength using the ONPG assay with the yeast transformants already at hand.

5.4.2 Liquid ONPG assay

Growth of FLNA-expressing yeast in liquid culture was slowed from a doubling time of approximately 2h (controls) to 4-5h, consistent with slower growth of colonies on agar plates. The cause for this growth reduction is unknown. Due to this slowed growth, such cultures were used directly from overnight incubation, while all others were re-inoculated in the morning into fresh medium (1:5 dilution, see “5.3.4: ONPG Assay”).

Typically, an amount of cells corresponding to an OD_{600} of 1 was used as input (in triplicates), except for the provided positive controls, which were used at OD_{600} of 0.25 (pCL1) and 0.5 (p53 and T-antigen), respectively. For most experiments, a total of 9 data-points per transformant were obtained (three time-points in triplicates), which allowed us to assess the statistical significance of the results.

As expected, the positive control reactions pCL1 and p53:T-antigen elicited a strong signal, whereas the negative control reactions Lam:T-antigen and FLNA:Integrinβ−Y788E showed no color development at all (not shown). The negative control reactions with one fusion protein alone also did not activate reporter gene expression, with the exception that both FLNA and FLNA-Q/R alone again gave positive signal, correlating with the findings of the colony-lift assay.

The β-galactosidase units produced by the transformants were calculated according to the manufacturer's instructions. To analyze whether FLNA Q/R interactions varied from those of FLNA, we calculated the ratio of β-galactosidase units produced by the interaction between FLNA-Q/R and a test protein over units produced by the interaction between FLNA and the same test protein:

$$\frac{\text{β-galactosidase units (FLNA-Q/R:test protein)}}{\text{β -galactosidase units (FLNA:test protein)}}$$

All tested interaction partners elicited a stronger signal when interacting with FLNA-Q/R compared to FLNA (Figure 19). Most importantly, FLNA-Q/R alone also elicited a stronger signal than its unedited variant, a finding, though intriguing, that renders analysis of the results virtually impossible (see "5.5: Conclusions"). This stronger

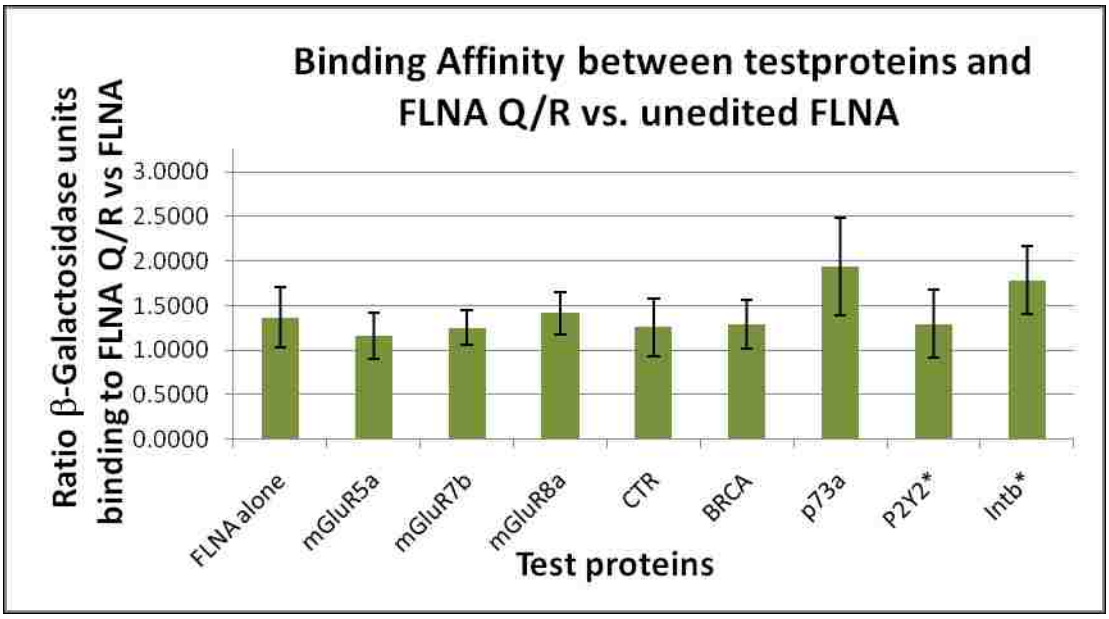


Figure 19: Ratio of binding affinity between test proteins and FLNA-Q/R over FLNA, respectively

Shown are the ratios of β-galactosidase units produced by interaction of a test protein with FLNA-Q/R and FLNA, respectively, with standard error bars. Note higher reporter gene activation in FLNA-Q/R expressing cells alone.
**only one experiment was performed.*

reporter gene activation in FLNA-Q/R expressing cells is statistically significant (t-test, p = 0.001).

The ONPG assay is prone to a lot of variability that results in high standard deviations within the same experiment and renders inter-experimental comparison difficult. As previously mentioned, FLNA-expressing cells grow much more slowly than any of the other transformants, which necessitated adjustments in the experimental outline. Since yeast cells grow slowly in the presence of the GAL4-AD-FLNA fusion protein, it is conceivable that they may produce less of this protein, for example by inactivating FLNA expression (which may have occurred in several of the large colony revertants, as discussed above). Secondly, loss of cell material during the numerous

centrifugation and resuspension steps is very likely. Furthermore, cells may not completely lyse during the freeze-thaw cycles. These factors most likely all contribute to the high standard-deviation seen in the ONPG assay.

A one-sample t-test was performed to compare the experimental samples to the FLNA-only control (n=5, mean=1.3729, standard deviation = 0.2768). Two test proteins elicited a significantly lower signal than the control, namely mGluR7b (p = 0.004) and the calcitonin receptor (p = 0.01), and two elicited a significantly higher signal than the negative control, i.e. p73alpha (p = 0.023) and Integrinβ (p = 0.009) (Table 11).

Test protein	N	Mean	Standard deviation	Standard error mean	t-value	p-value
mGluR5a	3	1.1555	0.1556	0.8983	-2.42	0.137
mGluR7b	9	1.2648	0.8233	0.0274	-3.937	0.004
mGluR8a	10	1.4343	0.0948	0.0299	2.051	0.071
CTR	8	1.1804	0.1549	0.0548	-3.515	0.01
BRCA2	8	1.2788	0.2886	0.0667	-1.411	0.201
p73alpha	4	1.8409	0.2155	0.1078	4.343	0.023
P2Y2	2	1.3381	0.1075	0.076	-.458	0.727
Integrinβ	3	1.7852	0.0667	0.0385	10.71	0.009

Table 11: Statistical significance of results (one-sample t-test)

Ratios of signal intensities elicited by interaction of test proteins with FLNA-Q/R over FLNA were compared to signal elicited by FLNA-Q/R over FLNA alone. Test proteins with significantly lower and higher values are highlighted in red and green, respectively. Confidence level ≥95%.

5.5 Conclusions

The 24 Ig-like repeats of FLNA show high sequence homology, and one could expect that a single amino acid change in a single repeat might not significantly impact protein function. For many interaction partners of FLNA it remains to be elucidated whether the binding is restricted to a specific part or repeat of FLNA. However, binding was shown to occur in a specific manner with individual repeats in certain cases (for example Lad et al., 2008). Furthermore, specific repeats have been shown regulate binding of proteins by competitive interaction with neighboring repeats (Lad et al., 2007), and missense mutations in FLNA can cause otopalatodigital spectrum disorders (Robertson et al., 2003). We thus expect that single amino acid changes can interrupt or increase specific binding to interaction partners. Consequently, the yeast-two-hybrid method to assess differences in interaction between FLNA and FLNA-Q/R and selected proteins was a valid experimental approach, considering our current knowledge on FLNA function.

None of the tested proteins show evidence of dramatic alterations of binding to the edited FLNA. This does not exclude the possibility that interactions are in fact altered *in vivo*. The apparent weaker color development in FLNA and FLNA-Q/R transformants compared to the positive controls may be due to either weaker interaction between the fusion proteins and/or the fact that FLNA and FLNA-Q/R reduce growth of the yeast cells, resulting in smaller colonies and thus reduced absolute lacZ expression.

Four of the tested proteins show significantly different alterations in β-galactosidase production when compared to the ratio of signal elicited by the negative controls. The fact that the negative controls themselves activate reporter gene

transcription and do so differentially (i.e. FLNA-Q/R elicits stronger signal than unedited FLNA) renders proper experimental evaluation and a reliable conclusion difficult. The underlying cause of the reporter gene activation by FLNA alone is unknown. If we assumed that this background activation was negated by the presence of a tested interaction partner, the results would mean that all test proteins interact more strongly with the edited version of FLNA. This assumption is supported by the fact that co-expression of the mutated Integrinβ (mutated at a single amino acid that abrogates binding to FLNA) abolishes reporter gene activation completely (data not shown). On the other hand, why should all binding proteins interact more strongly with FLNA-Q/R than FLNA?

The answer may not lie with the tested proteins, but instead with the assay system that was used. If FLNA were to generally increase transcription or translation by, for example, interacting with a yeast protein, we could partially resolve these apparently irreconcilable differences. Expression of one or several genes would be increased by the presence of FLNA and FLNA-Q/R, respectively, but lack of reporter gene transcription in conjunction with the mutated integrinβ-Y788E would remain silent. However, FLNA and FLNA-Q/R alone also increase reporter gene activation, and therefore must directly impact the GAL4 promoter, possibly by direct DNA binding. Therefore, reporter gene activation with the test proteins may be influenced by several factors: reporter gene activation by FLNA or FLNA-Q/R alone (‘background’), boosting of transcription or translation through an unknown factor, and finally the interaction between test proteins and FLNA isoforms. Consequently, it is impossible to use the yeast-two-hybrid system to test whether proteins interact more or less strongly with the edited FLNA. However, the

finding that FLNA-Q/R is able to elicit reporter gene activation more strongly than FLNA is interesting in itself. Whether or not the assumed underlying interaction with a host factor is physiologically relevant for higher organisms is another question.

While the yeast-two-hybrid method is excellent for analyzing protein-protein interactions, the chosen approach is limited by the subset of potential interaction partners that are analyzed. An alternative, unbiased method is the expression of TAP (Tandem Affinity Purification)-tagged FLNA and edited FLNA in FLNA-deficient cells, purification of the FLNA isoforms together with the respective protein binding partners using tandem affinity purification and subsequent analysis of the purified protein complexes by silver-staining and mass-spectrometry. Such an approach allows probing of the protein binding landscape of the two FLNA isoforms without restriction to known interaction partners. To that end, appropriate cloning vectors (pCeMM-NTAP(GS) and pCeMM-CTAP(GS)), specifically designed for TAP from mammalian cells (Burckstummer et al., 2006), were selected for cloning of the C-termini of FLNA and edited FLNA. FLNA-deficient mammalian cells were generously provided by Dr. Stossel (Brigham & Women's Hospital, Harvard Medical School). The vectors and cells were shared with Dr. M. Jantsch (University of Vienna, Austria) for use in his investigations. The results would be able to answer our question of whether editing of FLNA modifies its protein binding ability.

6 Consequences of RNA editing on IGFBP7 function

6.1 Abstract

Recoding of protein-coding sequences quite often has critical consequences on protein stability, localization, and/or functional properties. In fact, it is necessary for the proper development and physiological functions of higher eukaryotes, underscoring the importance of editing in higher organisms. The production of several protein isoforms in the same cell increases proteome diversity from a limited number of genes. After identifying previously unknown A-to-I RNA recoding events in the first part of my dissertation, my goal was to analyze functional impacts of editing on the alternative protein variants. In Chapter 2, we validated two A-to-I RNA editing sites in the Insulin-like growth factor binding protein 7 (IGFBP7) gene, both leading to amino acid changes when edited. IGFBP7 encodes a secreted protein that modulates the interactions of a cell with its surroundings. Its expression is silenced in several cancers; notably, the restoration of transcription and IGFBP7 protein synthesis in melanoma cells triggers apoptosis. However, IGFBP7 can also promote angiogenesis and stimulate proliferation of fibroblast cells. Besides these, IGFBP7 is important for other cellular functions as well, as it is elevated in the blood of insulin-resistant diabetes patients and inhibits estrogen production in granulose cells, diversifying its biological functions further. Such complexity may be explained at least in part by RNA editing. Here we characterize IGFPB7 editing patterns in different tissues and investigate possible consequences that editing may have on IGFBP7 function.

6.2 Introduction

Differential editing of multiple sites within the same target has been shown to regulate protein function as well as RNA splicing, protein trafficking, and stability (Burns et al., 1997). Importantly, characterized cases of recoding events revealed that editing usually affects highly conserved amino acid residues and dramatically alters protein function. It is thus of interest to get better insight into the editing patterns of the two sites in IGFBP7, especially since both seem to be edited in a significant proportion of mRNA sequences. We hypothesized that editing at either site or at both sites together produce protein variants with distinct properties. For the same reason, we were interested in elucidating the patterns of editing in different tissues to investigate the possibility of cell-type specific editing, as shown for other editing targets (He et al., 2011). In addition, we sought to determine if editing also occurs in mouse tissue, and thus may be conserved in mammals. We analyzed a number of specimens to characterize editing at the two sites in IGFBP7. To reveal editing patterns, individual clones of amplicons of different origin were sequenced as described previously. This strategy allowed us to determine whether or not editing in IGFBP7 is tissue-specific and subject to coordinated or differential regulation.

One distinguishing feature of IGFPB7 is its apparent versatility. It has been isolated on multiple occasions due to one of many characteristic properties and bears several names. IGFBP7 was initially identified as a gene downregulated in meningiomas, and named meningioma-associated cDNA (Mac25) (Murphy et al., 1993). The secreted protein was later independently purified from the human bladder carcinoma cell line EJ-1 and tentatively termed tumor-derived adhesion factor (TAF) (Akaogi et al., 1994), as well

as from cultured fibroblast cells as prostacyclin-stimulating factor (PSF) (Yamauchi et al., 1994). It was later proposed to be renamed angiomodulin (AGM) for its involvement in the formation of new capillary vessels by vascular endothelial cells (Akaogi et al., 1996a).

Since the early days of its characterization, IGFBP7 has been implicated in various forms of cancers, often as a putative tumor suppressor (Chen et al., 2007; Lin et al., 2007) with functions in apoptosis and senescence (Sprenger et al., 2002; Wilson et al., 2002; Mutaguchi et al., 2003; Ruan et al., 2007; Wajapeyee et al., 2008), yet also as a promoter (Adachi et al., 2001; van Beijnum et al., 2006; Pen et al., 2008) or blocker (Tamura et al., 2009) of angiogenesis, and it is overexpressed in circulating endothelial cells (CECs) of metastatic cancer patients (Smirnov et al., 2006). Conflicting reports on its growth stimulatory (Akaogi et al., 1996b) and inhibitory (Wilson et al., 2002) functions underline its complexity. Furthermore, increased levels of IGFBP7 have been detected in the blood of insulin resistant diabetic patients (Lopez-Bermejo et al., 2006) and it interacts with chemokines (Nagakubo et al., 2003). Recently, it was also shown to inhibit estrogen production in granulosa cells (Tamura et al., 2007). Thus, IGFBP7 is involved in a multitude of functions, and we hypothesized this vast functional range may in part be facilitated by editing, which allows the production of four protein isoforms from the same allele.

Similarities between other IGFBPs (IGFBP1 to 6) and IGFPB7 are restricted to the N-terminal cysteine-rich domain (CRD), which is contained within a single exon (Collet & Candy, 1998). IGFBP7 shares only ~20% identity in the CRD with most of the 12 cysteine residues conserved. However, no significant sequence similarity is observed

outside of this domain. CRDs can be found in various proteins associated with the extracellular matrix (ECM), with vastly different functions. The CRD probably arose from exon shuffling, while the other domains of IGFBP7 have distinct structures which are unrelated to those of the rest of the IGFBP family members (Collet & Candy, 1998).

Beside the N-terminal CRD, IGFPB7 has a Kazal-type trypsin inhibitor domain, and a single copy of an Immunoglobulin-like type C repeat in its C-terminal half (Collet & Candy, 1998). The two editing sites change the codons at positions R78 and K95 of the full-length protein, changing R to G and K to R, respectively, and are contained within the CRD. Interestingly, IGFBP7 is proteolytically cleaved after K97, which results in a two-chain form comprised of amino acids 27-97 (8kDa) and 98-282 (25kDa) that are

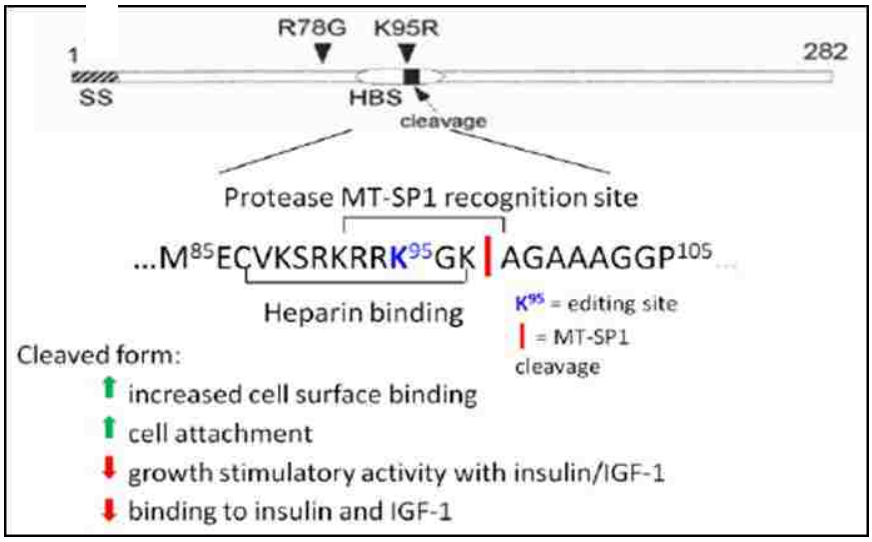


Figure 20: IGFBP7 editing and cleavage sites

Schematic representation of IGFBP7 protein with signal sequence (ss), editing sites, heparin binding site (HBS) and cleavage site indicated. Part of the protein sequence bound by heparin and cleaved by MT-SP1 is enlarged. Consequences of cleavage are shown.

cross-linked by disulfide bridges (Sato et al., 1999; Ahmed et al., 2003; Ahmed et al., 2006). Amino acids 1-26 are a signal sequence that is cleaved during post-translational processing. Proteolytic processing of IGFBP7 modulates its biological activity: intact IGFBP7 stimulates growth of DLD-1 colon carcinoma cells in synergy with insulin/IGF-I, but cleaved IGFBP7 completely abolishes the synergistic growth-stimulatory activity, possibly due to the loss of its insulin/IGF-I binding activity (Ahmed et al., 2003) (Figure 20). At the same time, the heparin-binding activity of IGFBP7 is decreased by proteolysis. Cleaved IGFBP7 appears to bind to syndecan-1 more efficiently than the intact protein (Ahmed et al., 2003). It is supposed that the cleavage induces a conformational change that masks the heparin-binding sequence, while exposing a different binding site to syndecan-1. Thus, the intact and cleaved forms of IGFBP7 have different biological activities. As editing occurs directly within this biologically relevant site, we hypothesized that editing has functional consequences for the ensuing protein isoforms, specifically with regard to proteolytic processing and/or heparin binding affinity. While the elicited amino acid change may seem conservative, other recoding events of the same nature such as in NEIL1 have been shown to dramatically alter protein function (Yeo et al., 2010).

The type II transmembrane protein MT-SP1 (membrane-type serine proteinase matriptase) was identified as the specific protease for IGFBP7 in ovarian clear cell adenocarcinoma (OVISE) and gastric carcinoma (MKN-45) cells, which produce the cleaved IGFBP7 at high levels (Ahmed et al., 2006). MT-SP1 contains an extracellular protease domain with a preferred cleavage sequence [P4-(Arg/Lys) P3-(X) P2-(Ser) P1-(Arg) P1’-(Ala)] and [P4-(X) P3-(Arg/Lys) P2-(Ser) P1-(Arg) P1’-(Ala)] (Takeuchi et al.,

2000). In comparison, IGFBP7 contains amino acids RKGKA (RRGKA in edited proteins) at positions P4 through P1' (Figure 20). A-to-I RNA editing at the second position results in an amino acid change from K to R at P3 of the cleavage recognition site – a conservative change which seems to be readily accepted by the matriptase, i.e. both sequences should be cleaved by MT-SP1. However, direct comparison of the protein isoforms might reveal a preferential proteolysis of one versus the other.

While a substantial body of research has accumulated over the past years about IGFBP7, due to its significance in cancer and disease, no one has specifically compared the functional roles of the different protein isoforms created by RNA editing. However, editing in this target may directly contribute to the large functional repertoire that IGFBP7 seems to possess. Elucidating the effects editing might have on proteolytic processing of the resulting protein variants directly addresses this question as evidence reported in the literature suggests distinct biological activities of cleaved and uncleaved IGFBP7.

6.3 Materials and Methods

6.3.1 PCR analysis

The reaction mixes contained 400nM of each of the primers (human: IGF7D and IGF2U (cDNA) or IGF7D and gIGFU (gDNA); mouse: mIGF12D and mIGF13U (cDNA) or mIGF12D and gIGFU (gDNA), primer sequences see Appendix B), 2µl Phire™ Hot Start DNA Polymerase (NEB), 0.4mM dNTP mix (Invitrogen), 1µl cDNA, and Phire polymerase buffer provided by the manufacturer in a total volume of 100µl (2x50µl). The reactions were carried out in an Eppendorf Mastercycler. Human: 98˚C for 2 minutes followed by 35 cycles of 98˚C for 10s, 71˚C for 5s, and 72˚C for 10s, followed by a final step of 72°C for 1 minute. Mouse: 98˚C for 2 minutes followed by 35 cycles of 98˚C for 10s, and 72˚C for 15s (12s for amplification of genomic DNA), followed by a final step of 72°C for 1 minute. Reaction products were analyzed on a 2% agarose gel.

6.3.2 Subcloning

IGFBP7 amplicons from select tissues were re-amplified using primers IGF8D-Eco and IGF9U-Kpn (Appendix B). The amplicons were restricted with 40U KpnI (NEB) for 3 hours at 37ºC, subsequently purified by phenol-chloroform extraction, ethanol-precipitated and restricted with 40U EcoRI (NEB) for 3 hours at 37ºC. Again, the cut fragments were purified by phenol-chloroform extraction and precipitated with ethanol, and then subjected to DNA gel electrophoresis on a 2.5% agarose gel. The bands of the expected size were excised, purified using the QIAEX II Gel Extraction Kit (QIAGEN) and then ligated into a pBluescript SK II vector also cut with EcoRI and KpnI and purified as described for the amplicons. The ligation reactions contained vector and insert

in a 1:8 molar ratio and 0.5µl T4 ligase (Invitrogen) in a total volume of 10µl and was incubated at room temperature 4 hours to over-night. Z-competent DH5α cells were transformed with 5-10µl of the ligation (Appendix A). The transformed cells were plated on LB containing ampicillin. Individual recombinant clones were used to inoculate liquid LB-Amp and the purified plasmids (QIAprep Spin Miniprep Kit, QIAGEN) were sequenced (Geneway, CA).

6.3.3 Cell culture

HEK293 cells were maintained in 1xMEM (Cellgro) containing 10% FCS and 1x antibiotic/antimycotic solution. Cells were passaged at least twice a week.

6.3.4 Cloning of IGFBP7

Full-length IGFBP7 coding sequence in the pCMV6-XL4 expression vector was purchased from Origene. Three pre-edited isoforms were created by site-directed mutagenesis, changing codon 78 from arginine (AGG) to glycine (GGG) and/or codon 95 from lysine (AAG) to arginine (AGG). IGFBP7 coding sequences of the four editing isoforms were C-terminally tagged with the HA sequence by PCR amplification and corresponding primer sequences (Appendix B), digested with XhoI and XbaI and ligated into the XhoI/XbaI sites of a pCI-neo vector. DNA sequencing analysis confirmed the fidelity of the constructs. For primer sequences see Appendix B.

6.3.5 Stable cell lines

Transfection of the four IGFBP7 isoforms in pCI-neo into HEK293 cells was performd using XtremeGene 9 transfection reagent from Roche according to the manufacturer’s instructions. Reagent to DNA ratio was 9:1. Control cells were produced

by transfecting cells with an empty pCI-neo vector. Stable transfectants were obtained after selection in 500µg/ml G418 for 3 weeks and were maintained in 300µg/ml G418 thereafter.

6.3.6 IGFBP7 protein purification

Culture medium of HEK293 cell lines stably expressing IGFBP7 isoforms was collected, 1/20V 20xTBS and 1/250V EZview Red Anti-HA Affinity Gel (corresponding to 1/500V of bead volume) (Sigma) was added and incubated on a rotating wheel for 1h at 4°C. Beads were washed three times with excess cold 1xTBS. Bound IGFBP7 protein was eluted with elution buffer (1xTBS, 0.05% SDS, 100µg/ml HA peptide) using 5 times the bead volume at 38.5°C for 2h with frequent vortexing. Eluates were diluted 1:5 with adjustment buffer (56.875mM Tris, 28.75mM NaCl, 0.01% tween 20, pH9) to convert the elution buffer into the MT-SP1 assay buffer. Diluted samples were ultrafiltrated using Amicon Ultra-0.5 Centrifugal filters (Millipore) in a table top centrifuge for 15 minutes at 14'000g at 4°C.

6.3.7 MT-SP1 proteolytic cleavage assay

Recombinant human matriptase (MT-SP1, MW = 26kDa) encoded by the ST14 gene was purchased from R&D systems. 100ng/µl stock solutions were prepared in sterile 50mM Tris, 10% glycerol, pH 8.0 and stored at -20°C. Relative concentrations of the purified IGFBP7 isoforms were estimated from Western blots and input for the proteolysis reactions were approximated accordingly. Total reaction volume was adjusted to 39µl with 1x assay buffer (50mM Tris, 50mM NaCL, 0.01% Tween 20, pH9.0). 10% were removed as negative controls before adding 0.9µl 0.5ng/µl MT-SP1 (final

concentration = 12.5pg/μl or 325nM). Proteolysis reactions took place in 200μl tubes at 25°C. At indicated timepoints, 4μl aliquots were removed, mixed with SDS sample buffer and boiled for 5 minutes at 90°C to stop the reaction. Rate of proteolytic cleavage was assessed via SDS PAGE (12% gel, standard protocol, Sambrook laboratory manual) and Western Blotting analysis.

6.3.8 Western blot

Standard protocols were used to separate proteins according to their molecular weight on a 12% SDS polyacrylamide gel and transfer them onto 0.2μm nitrocellulose membranes. Membranes were blocked for two hours at room temperature in 2% milk in PBS/0.05% Tween 20. Primary antibody (rabbit anti-HA, Clontech) was diluted 1:300 in blocking buffer and incubated with the membrane over night at 4°C. Blots were washed with PBS/0.05% Tween 20 and incubated with the secondary HR-conjugated anti-rabbit antibody at 1:25,000. Blots were washed again before application of ECL Plus (Amersham) and exposure to film for 5-30 minutes.

6.4 Results and Discussion

6.4.1 Editing of IGFBP7 is conserved in mammals

Editing of IGFBP7 was analyzed in mouse cortex and cerebellum. Editing at both sites occur to a much higher degree than was found in human: 95% at the R78G site and 85% at the K95R site. The reason for these high editing levels in mouse brain tissue is unknown, but could reflect species-specific editing of this target. Unlike in the editing target C1QL1, editing occurred to comparable levels in both cortex and cerebellum (Table 12). Analysis of the genomic DNA did not reveal occurrence of a SNP at either site.

Sample ID	Tissue	R78G % editing	K95R % editing
Mouse 1	Cortex	95	85
Mouse 1	Cerebellum	90	85
Mouse 2	Cortex	90	85
Mouse 2	Cerebellum	90	85

Table 12: Editing levels in mouse IGFBP7

Levels were estimated from sequence tracks obtained from cortex and cerebellum of two adult mice. Sequence tracks not shown.

6.4.2 Tissue-specific editing patterns of IGFBP7

The editing levels at the two sites of IGFBP7 were estimated from sequence tracks of 8 different human tissues. The wide range of editing from 6-57% for R78G and 30-85% for K95R is a strong indication of tissue-specific regulation of editing in IGFBP7 (Figure 21). Since each sample is derived from a different donor, this variability could also be due to inter-personal differences rather than tissue-specificity. However, Li et al. determined IGFBP7 editing levels by deep sequencing, obtaining the same results for

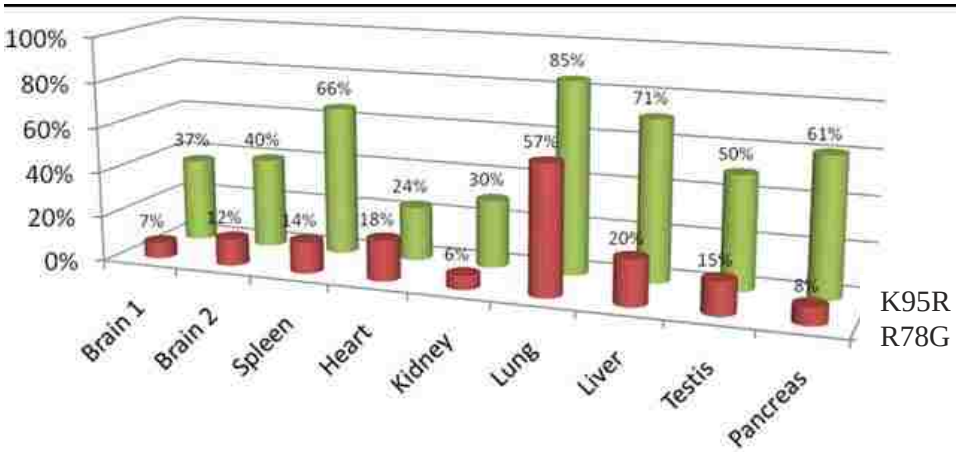


Figure 21: Editing at R78G and K95R across different human tissues

Editing levels were estimated from sequence tracks. Red bars (front): editing at R78G, green bars (back): K95R. Editing percentages are indicated above bars.

kidney samples (7% at R78G, 30% at K95R). Editing at K95R is almost always higher than editing at R78G, with the exception of the heart tissue. The differences do not correlate with observed expression levels of IGFBP7 (GNF expression atlas, not shown).

6.4.3 Tissue-specific, site-independent editing in IGFBP7

To determine the distribution of the four possible transcript versions that can arise due to editing, we subcloned the amplicons from brain, heart and spleen and sequenced at least 94 individual clones per sample. As shown in Figure 22, editing leads to a distinct distribution of transcript variants in the three tissues. Chi-square analysis of the data using a two-way classification reveals no significant difference between the editing levels of the two brain samples. Conversely, when comparing the data from brain with heart or spleen, respectively, editing levels are significantly different (brain: $p<0.01$; spleen: $p<0.025$). Comparison between heart and spleen also shows significantly different editing levels, with a p-value <0.001. These results therefore point toward a tissue-specific

regulation of editing in IGFBP7, which could enable a cell to perform specialized functions associated with its specific environment.

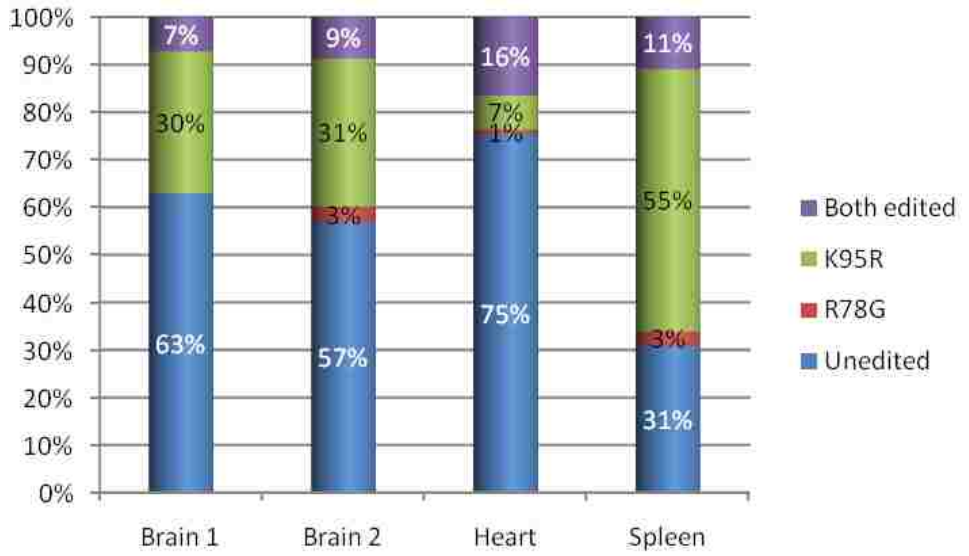


Figure 22: Distribution of transcript variants determined from different tissues
PCR products obtained from the indicated tissues were cloned and single clones isolated and sequenced, allowing detailed analysis of transcript versions present in different tissues. Exact percentages of editing are indicated on bars.

Enstero et al. showed that adenosines located on the same side of a RNA double helix are often edited together, which is referred to as coupling (Enstero et al., 2009). Since RNA double helices contain about 10-12 nucleotides per turn, the R78G and K95R sites, which are distanced from each other by 51 nucleotides, may therefore be coupled. To test this, a Pearson chi-square analysis was performed on the data. Even though the R78G site is edited almost exclusively together with the K95R site, coupling between the two sites is statistically significant only in the first brain ($p<0.01$) and the heart samples ($p<< 0.001$). For the second brain and the spleen sample, the percentage of transcripts edited at both positions simultaneously is due to chance ($p>>0.05$). This is not surprising,

as coupling is generally disrupted by bulges in the RNA helix, of which there are five between R78G and K95R. Therefore, the two sites are edited independently of each other. An unknown mechanism may ensure mainly production of transcripts edited at both sites rather than R78G alone, which would explain the statistically significant coupling seen in the first brain and the heart samples.

6.4.4 Proteolytic cleavage of IGFBP7 isoforms

Initially, we used *in vitro* transcription and translation and radioactive labeling with ^{35}S-Met of IGFBP7 (TNT system, Promega) to produce the respective protein isoforms with an apparent mass of 33kDa. Proteolytic cleavage with MT-SP1 (R&D) yielded the expected fragment of 8kDa (N-terminus, labeled portion). Low signal-to-noise ratio made quantification of the band intensities over a time-course of incubation with MT-SP1 difficult, but these initial experiments indicated that the K95R form seems to be cleaved at a slower rate (not shown). To circumvent the low labeling problem (IGFBP7 only contains 3 methionines at the N-terminus) which results in a high background, we instead decided to tag the protein isoforms with the hemagglutinin (HA) peptide sequence. Expression of these HA-tagged isoforms in cells enabled us to purify them with an anti-HA matrix and analyze them via Western blotting using an anti-HA antibody.

Expression of proteins in cell culture has both advantages and disadvantages over expression *in vitro*. For example, post-translational modifications of the proteins such as glycosylation and phosphorylation occur *in vivo*. On the other hand, the editing competent secondary structure of IGFBP7 is formed entirely by exon 1 and, consequently, transcription of the open reading frame of IGFBP7 also produces RNA

structures amenable to editing. This could result in mixed populations of IGFPB7 isoforms, even with only one type of expression plasmid present in a specific stable cell line. We therefore assessed editing levels in the transcripts of the recombinant isoforms using a plasmid-specific forward primer located in the 5'-UTR that spans an intron and an IGFBP7-specific reverse primer. Even though HEK293 cells typically show low ADAR activity, some editing (up to 20%) occurs in the originally unedited and single edited versions of IGFBP7 (data not shown). The editing levels were not negligible, but low enough to proceed with the proteolysis assay, keeping in mind that the IGFBP7 isoforms thus assessed are not entirely pure.

In vitro proteolysis of IGFBP7-HA isoforms by MT-SP1 results in two cleavage products, a 27kDa C-terminal and 8kDa N-terminal part, respectively. Only the 27kDa product can be detected with the anti-HA antibody on a Western blot, as the HA-tag is C-terminal. The bands of cleaved and uncleaved protein were quantified using ImageJ software. For analysis of cleavage efficiency, the ratios of signal from cleaved (27kDa) versus uncleaved (35kDa) product were plotted against time and linear regression analysis was executed (Figure 23). The cleavage efficiencies of the unedited ('wild-type', or wt) and the R78G isoforms appear to be similar, but in most experiments are markedly different from those of the K95R and double edited variants, respectively. The amino acid change elicited by editing at the K95R position lies directly within the protease recognition site, and in all experiments, this K95R isoform is cleaved less efficiently than the others. The double-edited version (with amino acid changes at both positions 78 and 95) seems to be cleaved at a rate somewhere in between that of K95R and the other two.

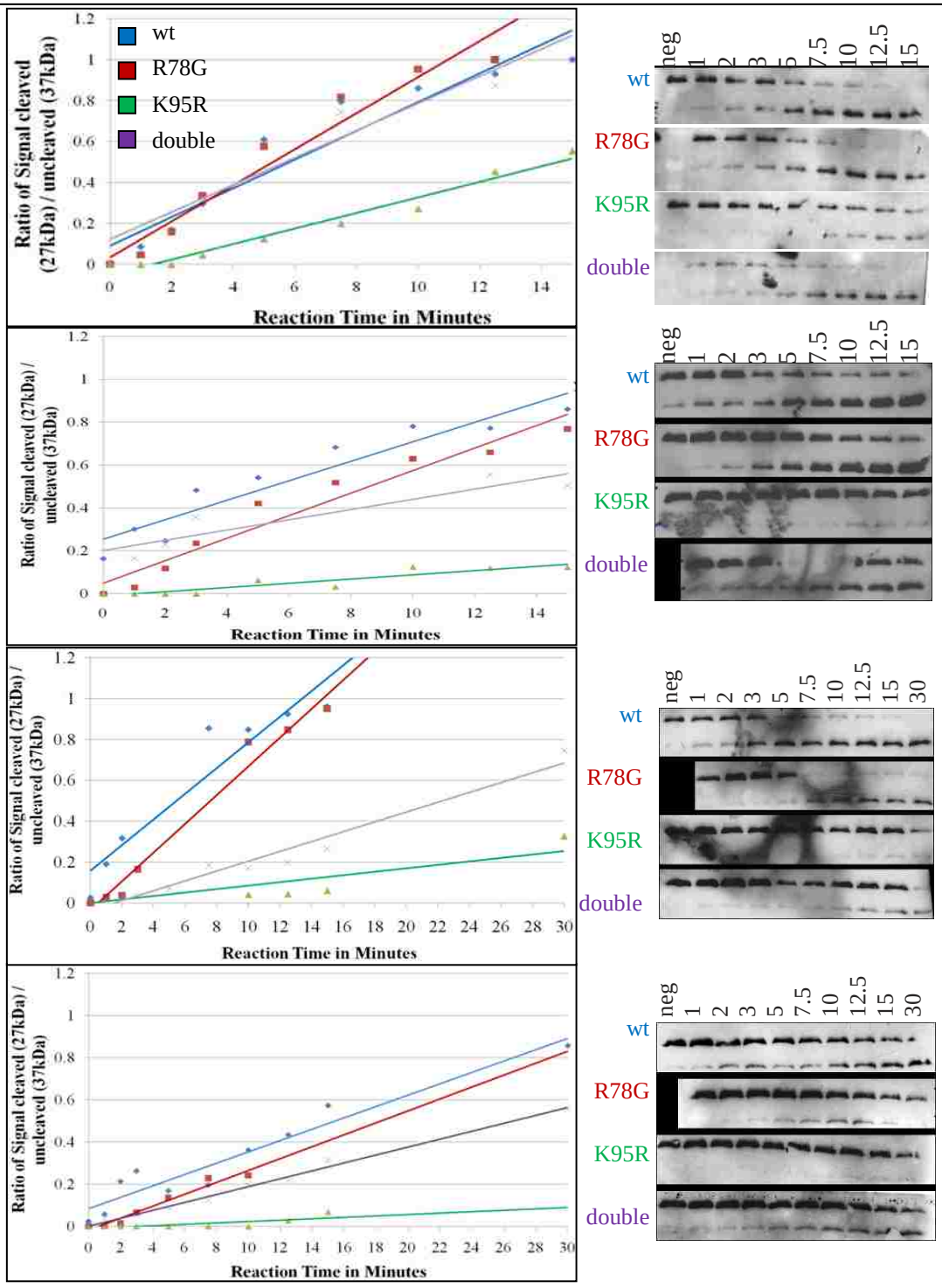


Figure 23: Proteolysis of IGFBP7 isoforms

Four independent experiments are shown, corresponding Western blots on the right.

Graphs display the ratio of 27kDa over the 37kDa IGFPB7 for each isoform.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Wild-type	0.0702x + 0.0904 (0.9152)	0.0454x + 0.2544 (0.9027)	0.0629x + 0.1574 (0.8995)	0.0269x + 0.0835 (0.9287)
R78G	0.0883x + 0.0333 (0.952)	0.0526x + 0.0485 (0.9446)	0.0704x + 0.0347 (0.9782)	0.0283x + 0.0176 (0.9558)
K95R	0.0381x + 0.0538 (0.9696)	0.0099x + 0.0105 (0.086)	0.0085x (0.7327)	0.0033x + 0.0103 (0.5936)
double	0.0664x + 0.1223 (0.9181)	0.0239 + 0.2002 (0.8521)	0.024x + 0.0364 (0.9616)	0.0188 + 0.0002 (0.9913)

Table 13: Linear regression analysis

Equations of the linear regression analysis of four IGFBP7 proteolysis reactions. Equations are given in the form of $y = mx + q$ (R^2), m=slope, q=y-axis intersection. R^2=correlation coefficient.

The slopes of linear regression curves are indicators of cleavage efficiency (Table 13), which is a function of substrate and protease quantity. Due to the small amount of material used for these experiments, substrate input could not be standardized between experiments, as it is below the detection limit of regular protein quantification methods. Substrate input was instead estimated from the relative signal intensities from Western blots with different substrate loading volumes of each IGFBP7 isoform. Therefore, they varied with every experiment, leading to different cleavage efficiencies each time. Furthermore, very dark bands may be saturated with signal and can thus lead to underestimation of actual quantity.

This last detriment is mitigated to some degree by taking the ratio of two bands of different intensities, one of them usually not being saturated. To better account for

	1	2	3	4
R78G	74%	116%	110%	87%
K95R	73%	97%	89%	92%
double	54%	112%	82%	108%

Table 14: Substrate input

Substrate input levels relative to wild-type, estimated from Western blot signals (ImageJ).

differences in substrate input within one experiment, the signal intensities per lane were averaged for each isoform to serve as estimate of relative input compared to the wild type. Taking the average is preferable than basing the analysis on single lanes due to local irregularities that occur around each band from background signal or disparities in protein transfer during blotting. Lanes with very high background or where one band was not blotted properly were excluded. Such analysis shows that in experiment 1, substrate input of R78G and K95R was about 73%, and input of the double-edited isoform only 54% of that of wild-type. Less substrate with equal amounts of protease would result in faster cleavage. Therefore, the discrepancy in input material may be the cause for a cleavage rate of the double-edited IGFBP7 equal to that of wild-type. Input levels of isoforms compared to wild-type for all experiments are shown in Table 14.

The differences in input levels were mostly minor and within the range of +/-18%, with the exception of experiment 1 as discussed above. The fact that the R78G variant is cleaved at a rate comparable to that of wild-type in all four experiments, yet shows variability of input levels akin to the other two isoforms, validates the differences in cleavage rate we observe for both the K95R and the double-edited variant. However, reduction of input by 50% seems to have the noticeable effect as seen for the double-edited variant in experiment 1.

The results of the proteolytic cleavage assay show that editing at the K95R codon has a significant effect on cleavage efficiency of the resulting protein isoform. Cleavage rate of the edited K95R variant is reduced at least two-fold compared to that of the unedited and R78G edited variants. *In vivo*, this reduction in cleavage could have a major impact on the function of IGFPB7.

6.5 Conclusions

IGFBP7 has been shown to be involved in diverse biological functions, from apoptosis, to inhibition or stimulation of growth and angiogenesis, to stimulation of prostacyclin production in endothelial cells and diabetes. It can associate with type IV collagen (Akaogi et al., 1996a; Akaogi et al., 1996b) and bind IGFs and insulin (Akaogi et al., 1996b; Yamanaka et al., 1997). There is only one modification known to date, the N-glycosylation of asparagine 171 (Oh et al., 1996), and thus regulation of its many activities through post-translational modification seems limited. Five alternative splice variants are suggested by sequence annotations, but the functionality of these variants remains to be evaluated. Besides alternative splicing, transcript modification by A-to-I RNA editing leading to recoding might allow expanded functionality of this transcript.

Our analysis of editing patterns in samples of human origin shows a tissue-specific distribution (Figure 21). The obtained editing levels correspond well to those found by a recent deep-sequencing analysis (Li et al., 2009). Detailed analysis of transcript variants by cloning of amplicons and sequencing of single clones revealed that the predominant variant in brain and heart is the unedited version, followed by the K95R and double-edited variants. In spleen, IGFBP7 transcripts are mainly edited at the K95R position, followed by the unedited version and the double-edited variant, respectively (Figure 22). There is therefore a tissue-specific distribution of transcript variants, which possibly gives rise to a tissue-specific distribution of the corresponding protein isoforms. Such tissue-specificity is indicative of regulated editing, especially as editing levels in spleen and lung significantly surpass those seen in brain, which displays the highest levels of editing in known and characterized targets. The nature of this regulation remains

to be investigated. As editing appears to be regulated, it is likely that the resulting transcript variants occupy different functions, which might support the functional diversity of the IGFBP7 protein.

Proteolytic cleavage at K97 of the full-length IGFBP7 results in a two-chain form with modulated biological functions (Ahmed et al., 2003). The cleaved form binds more strongly to cell surfaces and induces cell attachment, and also shows decreased binding to IGF and insulin and reduced growth stimulation of fibroblasts. Editing changes a lysine (K) to an arginine (R) codon at codon 95. Even though K/R represents a conservative amino acid change, it was shown that the same change dramatically impacts lesion specificity of a DNA repair enzyme (Yeo et al., 2010). Our results show that editing at the K95R site results in significant reduction of cleavage efficiency by the specific MT-SP1 protease. Such a reduction might be even more relevant *in vivo*, where other interactions might compete with recognition and cleavage of IGFBP7 by MT-SP1, for example binding to heparin (see below), chemokines, and/or IGFs or insulin. Decreased affinity of the K95R IGFBP7 isoform for MT-SP1 might strongly favor binding to these or other factors and thus shift its biological activity. Different tissues may require a specific distribution of IGFBP7 isoforms to fulfill distinct functions, hence the tissue-specific editing patterns.

The intermediate cleavage efficiency displayed by the double-edited isoform is unexpected, as a behavior similar to that of K95R would have been expected. There is only limited protein structure information available, as only the structure of IGFBP5 has been elucidated (Kalus et al., 1998; Zeslawski et al., 2001). The N-terminal domain (which includes both R78G and K95R) of IGFBP7 is only 57% similar (20% identical) to

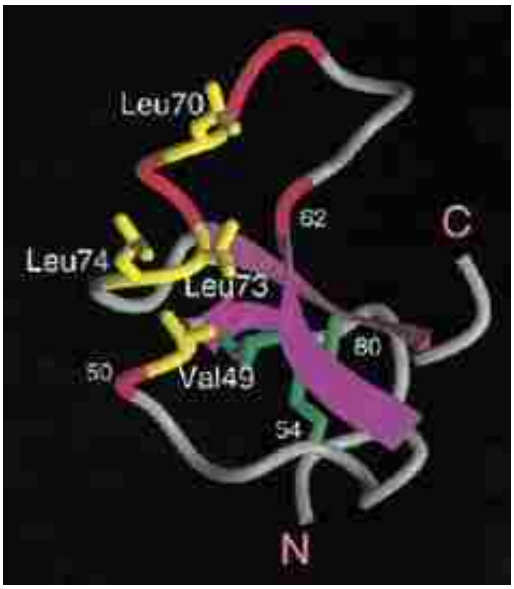


Figure 24: Structure of IGFBP5

The structure prediction of the N-terminus of IGFBP5 is shown. The loop between amino acid 50 and 54 corresponds to the homologous part containing R78G in IGFBP7. The loop between amino acid 62 and Leu70 in IGFBP5 corresponds to the homologous region containing K95R in IGFPB7. Figure adapted from (Kalus et al., 1998).

that of IGFBP1 to 6, which has to be taken into account when aligning IGFBP7 to the structure of IGFBP5. Alignment of the homologous regions of IGFBP7 to this structure shows that both R78G and K95R are in loop-regions, separated by a β-strand (Figure 24). Interestingly, IGFBP5 binds IGF-I with a domain homologous to one close to the K95R site in IGFBP7 (Zeslawski et al., 2001). Cleavage has been shown to change affinity of IGFPB7 for IGF-I and insulin, but it is unknown whether the amino acid change at K95R has an effect on binding affinity, in either cleaved or intact form. Since R78G and K95R are presumably in distinct regions of the protein, both in loop-regions, it is unclear how the amino acid change at R78G in double-edited isoforms could increase MT-SP1 affinity to its recognition site and thus alleviate the reduced proteolytic cleavage seen in K95R single-edited isoforms. It has been shown that the sequence of a gene can tune the speed of translation, which is important for protein folding (Cannarozzi et al., 2010; Fredrick & Ibba, 2010). It is possible that the nucleotide change at codon 78 might alter translation efficiency, which in turn might modify protein folding. A change in protein folding could

impact the local IGFBP7 domain structure such as to change the proteolytic cleavage rate occurring in a loop opposite to the R78G site. In this manner the distant editing sites at codon R78G might be able to reduce the effect of the K95R amino acid substitution on proteolytic cleavage efficiency.

Additional experiments that are likely to provide important insights include assessing other possible consequences of editing on IGFBP7 function. Heparin and heparan sulfates on cell surfaces bind to growth factors, cytokines, enzymes, and inhibitors, thereby modulating their biological activities. It was observed that cell adhesion activities of IGFBP7 are inhibited by heparin (Akaogi et al., 1996a) and K_{89}SRKRRKGK_{97} was identified as the heparin binding site on IGFBP7 (Sato et al., 1999). Thus, the cell binding capacity of IGFBP7 seems to be attributable to a heparin-binding site. IGFBP7 was shown to bind to IGF-I and IGF-II, albeit with lower affinity (5-6 and 20-25-fold lower, respectively) than other IGFBPs, such as IGFBP3 (Oh et al., 1996). However, IGFBP7, IGF, and insulin synergistically stimulate the growth of mouse fibroblasts, indicating that the interaction between IGFBP7, IGF and insulin is biologically relevant (Akaogi et al., 1996b). Both binding of IGFBP7 to cell surfaces and IGFs and insulin could be affected by the amino acid changes introduced by editing. Investigaing this possibility in experiments will assess the binding of IGFBP7 isoforms in cleaved and uncleaved form to cell surfaces. Addition of heparin to the IGFBP7 isoforms will determine whether this cell attachment is dependent on the heparin binding site or not. Finally, edited IGFBP7 isoforms in conjunction with cleavage might have altered interaction with IGF and insulin, which subsequently might lead to changes in stimulation of cell proliferation.

7 Regulation of editing in a highly predicted ADAR target

7.1 Abstract

SerpinA3 emerged as highest-scoring candidate in the biocomputational screening using the REDS program due to a predicted 90bp double-stranded RNA structure (Chapter 3). Since double-stranded RNA structures of 50bp or more have been shown to be promiscuously edited, the lack of editing in this transcript suggests the existence of a novel form of editing regulation. Using minigene constructs comprising parts of the SerpinA3 gene we investigated the effects of splicing and the presence of putative splice-enhancer binding sites on editing efficiency. We show that editing occurs in pre-mRNA, but is absent in fully processed mRNA. Our results further show that the inosines, and neither the strong double-stranded RNA fold nor the A-to-G changes, are responsible for degrading or otherwise obscuring the edited transcripts from detection by our system.

7.2 Introduction

The serine protease inhibitor alpha-1 antichymotrypsin (SerpinA3) has been implicated in the pathology of a number of human diseases including chronic obstructive pulmonary disease, Parkinson's and Alzheimer's diseases, stroke, cystic fibrosis, cerebral haemorrhage and multiple system atrophy (for review see Ye & Goldsmith, 2001; Devlin & Bottomley, 2005; Baker et al., 2007) . SerpinA3 is a secreted acute phase protein and is upregulated by cytokines. At sites of inflammation it binds to and controls target proteinases. Its structure consists of three β-sheets, eight alpha helices and a 23 amino acid long active center loop (RCL), which presents itself as a pseudosubstrate for the target proteinase. After binding, the protease cleaves the P1-P1' peptide bond in the RCL, and a dramatic conformational transition of the serpin protein results in the irreversible insertion of the loop into a β-sheet, which inactivates the protease. The altered conformation of SerpinA3 bound to its target is then recognized by hepatic receptors and cleared from the circulation.

The activity of SerpinA3 relies on its metastability in the native conformation, which allows a dramatic conformational change upon binding to the cognate protease and formation of a stable complex. This metastability depends on a kinetically trapped intermediate fold of the native SerpinA3, which is vulnerable to dysfunction by mutations. Mutations can cause aberrant conformational transitions that bring about the detention of the serpin within the cell, resulting in serpinopathies that manifest themselves in gain of function due to intracellular accumulation or loss of function due to lack of circulating serpin (for review see Lomas, 2005; Davies & Lomas, 2008).

SerpinA3 has been detected in senile plaques and surrounding astrocytes, and in complex with toxic beta-amyloid peptide in brains of Alzheimer's patients (Baker et al., 2007).

SerpinA3 emerged as the highest scoring candidate in the REDS ranked list of putative editing targets (Chapter 3, Table 7a). The pre-mRNA earns its high rank due to a predicted 90 bp completely complementary foldback structure surrounding the putative editing site. In fact, double-stranded structures of 50bp in length or more are often promiscuously edited (Polson & Bass, 1994; Bass, 2002) and the longest completely complementary dsRNA of known site-specific editing targets is 17bp in the case of GluR-B. However, editing of SerpinA3 could not be detected in RNA analyzed from human brain or spleen (Chapter 4). Although the complementary sequences of SerpinA3 transcripts are separated by ~900 nucleotides, editing at other known recoding sites requires folding of sequences even further apart from each other: the Q/R recoding sites of GluR-5 and GluR-6 involve an intronic ECS that is ~1700 nucleotides away from their exonic complementary sequences (Herb et al., 1996). The distance between the editing site and the ECS is thus not likely to be the only reason for the lack or low level of editing in SerpinA3.

To assess whether lack of editing in brain and spleen was due to tissue-specific editing as reported for other targets (He et al., 2011), relevant parts of SerpinA3 transcripts were analyzed in six human tissues (heart, pancreas, liver, lung, kidney and testis). No editing was observed in any of these sequence tracks, providing a unique opportunity to extend our knowledge of editing regulation.

The intronic ECS is embedded in a L2 element. Interestingly, the ECS is present only in human and chimpanzee, but not in other primates such as marmoset, orangutan or

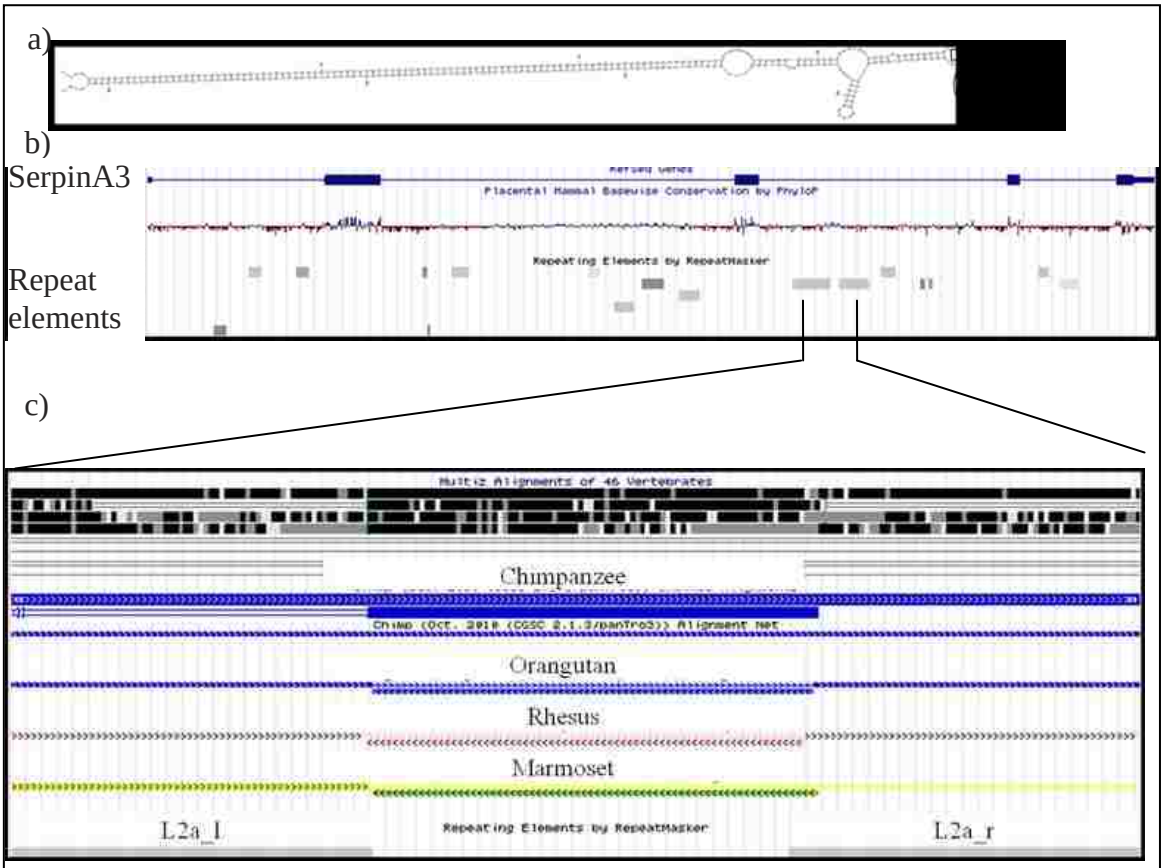


Figure 25: Genetic elements in SerpinA3

a) Predicted RNA secondary structure formed between putative exonic target and intronic ECS (Zuker, 2003; Costa, 2010). b) A schematic representation of the SerpinA3 gene structure with repeat elements indicated at the bottom. c) Magnification of ECS and adjoining L2 element (L2a_l and L2a_r, forming the left and right arms of one L2 element) with conservation in chimpanzee but not orangutan, rhesus and marmoset, as evidenced by the gap in the alignment. The sequence aligned by the program in the opposite orientation is the exonic target sequence. b and c are snapshots taken from the UCSC Genome Browser (Forster & Symons, 1987; Kent et al., 2002).

rhesus, and its insertion must thus be a recent evolutionary event (Figure 25). At the time of the experiments, another L2 element located further downstream in intron 3 was annotated in the UCSC genome browser (human genome build hg18, March 2006). Since its orientation was opposite to that reported for the L2 element surrounding the ECS, it

seemed prudent to investigate whether these elements prevent formation of the editing competent structure by folding onto each other and favoring a competitive RNA fold. (Note that this second oppositely oriented L2 element is not annotated in the latest human genome build hg19 from February 2009). We analyzed whether such sequence elements could possibly prevent editing of the predicted 90bp region *in vivo* using a number of minigene expression constructs. These minigenes consist of parts of the SerpinA3 gene including the 90 nucleotide target sequence and the ECS. They were co-transfected with an ADAR-expression construct into HeLa cells. Specific genetic elements, such as splice and putative splice enhancer sites, were further manipulated to analyze whether editing and splicing compete with each other for the same target, which may prevent editing.

Lack of editing in SerpinA3 provides us with an opportunity to dissect how editing is regulated, a vastly unexplored area of research. Even though editing of specific targets has been shown to be regulated temporally and spatially, as well as by certain environmental stimuli, it has never been shown that editing is actively prevented from taking place in putative targets. Such a finding would substantially add to our comprehension of regulation of editing. Surprisingly, we show that SerpinA3 pre-mRNA is promiscuously edited, even though the spliced mRNA product does not show evidence of editing. Using the minigene constructs we show that the strong double-stranded structure formed between the 90 nucleotide target sequence and the ECS is not responsible for removing the edited transcripts. Our results demonstrate that possible deleterious promiscuous editing of the SerpinA3 transcript is detected and eliminated, possibly through a yet unidentified surveillance mechanism.

7.3 Materials and Methods

7.3.1 SerpinA3 minigene construct

All sequence elements from the SerpinA3 gene were amplified using Hot Start Pfu Polymerase (Invitrogen) according to the manufacturer's instructions. Restriction enzyme recognition sites were added to primer sequences as indicated and the resulting amplicons were subjected to restriction digestion according to the manufacturer's instructions (restriction enzymes from NEB). For a graphic illustration of construct variants, refer to Figure 26. For primer sequences see Appendix B.

Amplification of SerpinA3 sequences for the full-length wild-type construct occurred in several steps. Initially, the first half, between exon 3 and about 300 nucleotides into intron 3, was amplified with SPI5D-EcoR1 and SPI8U-Nsi1 and primers SPI7D-Nsi1 and SPI6U-Kpn1 were used to amplify the rest of the sequence through the end of the 90 bp editing site complementary sequence (ECS). The two restricted amplicons were ligated into pCI vector cleaved with EcoR1 and Kpn1 in a three-way ligation, yielding construct SPI_02. Primers SPI13D and SPI14U-Xma1 were then used to amplify the rest of intron 3 and exon 4 and ligated into SPI_02 restricted with Stu1 and Xma1, yielding construct SPI_04. An internal Stu1 restriction site about 50 nucleotides upstream of the 90 nucleotide intronic ECS facilitated this strategy. SPI14U-Xma1 anneals with exon 4.

To replace parts of intron 3 in the vectors with unrelated sequences, sections of the upstream intron 2 were used. Primers SPI15DmutDAge1 and SPI16UmutUAge1 enabled introduction of an Age1 restriction site 55 nucleotides into intron 3 in construct SPI_02 via site-directed mutagenesis. Vector SPI_02 thus mutated was cut with Age1

and PvuII. An internal PvuII restriction site immediately 5' to the ECS allowed precise replacement of intronic sequence between exon 3 and ECS. A part of intron 2 was amplified with primers SPI17D-Age1 and SPI30U-PvuII and inserted into the restricted vector, yielding construct SPI_03, which have intronic sequence disparate from wildtype between exon 3 and the ECS.

Construct SPI_05 required addition of the rest of intron 3 and exon 4 to SPI_03. The PvuII restriction site located before the ECS was mutated to SacII (primers SPI43D-mutPvu-Sac and SPI44U-mutPvu-Sac) in construct SPI_03, as the intronic sequence between the ECS and exon4 contains three additional PvuII sites and makes use of this enzyme impossible. The mutated construct and the amplicon derived with SPI5-SacII and SPI14U-XmaI were cut with the enzymes and ligated to yield SPI_05.

For construct SPI_06, the sequence between the ECS and exon 4 was exchanged using primers SPI25D-KpnI and SPI26U-Sal1 to amplify parts of intron 2, and SPI27D-Sal1 and SPI14U-Xma1 to amplify the 3'-end of intron 3 (including branch-point) and parts of exon 4. The two amplicons were restricted with the respective enzymes and ligated into vector SPI_02 cut with KpnI and XmaI in a three-way ligation.

Similarly, to exchange the part of the L2 element directly downstream of the ECS, intron 2 sequence was amplified using primers SPI25D-Kpn1 and SPI31U-Sal1. Parts of intron 3 and exon 4 were amplified with SPI32D-Sal1 and SPI14U-Xma1. Both amplicons were restricted with the respective enzymes and ligated into vector SPI_02 cleaved with KpnI and XmaI in a three-way ligation to yield SPI_07.

Construct SPI_08 was generated in two steps to exchange only the L2 element located in the 3'-half of intron 3 (annotated as L2 on the negative strand in hg18, March

2006). First, SPI34D-Kpn1 and SPI26U-Sal1 were used to amplify part of intron 2, SPI27D-Sal1 and SPI14U-Xma1 to amplify part of intron 3 and exon 4. These amplicons were ligated into SPI_02 in a three-way ligation. The resulting construct was cleaved with KpnI and StuI, which cuts directly before the ECS, and an amplicon containing sequence from the ECS to the downstream L2 (using primers SPI13D and SPI33U-Kpn1) was inserted.

Finally, a short construct was generated as positive control (SPI_01), whereby exonic sequence amplified with SPI3D-Nhe1 and SPI4U-EcoR1 as well as the ECS amplified with SPI5D-EcoR1 and SPI6U-Kpn1 were cloned into the pCI vector. This constructs lacks most of the sequence separating exonic target and intronic ECS, leaving less than 50 nucleotides in between.

7.3.2 SerpinA3 minigene site-directed mutagenesis

7.3.2.1 Splice-site mutants

For mutation primers see Appendix B. Splice sites were mutated in two rounds. First, primers SPI35D-mut5'(2) and SPI36U-mut5'(2) were used to mutate the 'Gg' (last nucleotide of exon, first nucleotide of intron) to a 'TA' dinucleotide. Splice acceptor sites were mutated using primers SPI37D-mut3' and SPI38U-mut3'. Since the first mutations did not eliminate splicing completely (not shown), the splice sites were further mutated. SPI47D-mut5'(3) and SPI48U-mut5'(3) changed two additional nucleotides, the canonical intronic T and a G of a 'GT' dinucleotide in the exon that could serve as alternative splice site (see Figure 28, page 153). Similarly, primers SPI39D-mut3'(2) and SPI40U-mut3'(2) mutated the 3'-splice acceptor sites, changing a total of four guanosines

at the splice-acceptor present in the canonical ‘AG’ dinucleotide conformation (Figure 28).

7.3.2.2 Splice-enhancer mutants

Splice-enhancers may bind to the 90 nucleotide target sequence and prevent formation of the 90bp dsRNA. Putative splice-enhancer sites were determined as outlined in “7.4.4: Splice-enhancers” (Figure 29). The two most promising potential recognition sites were subsequently mutated. Due to possible annealing of the mutagenesis primers to both exonic target and ECS in the SPI_04 construct, primers were designed according to a protocol by Liu and Naismith that facilitates multi-site mutagenesis (Liu & Naismith, 2008). The protocol allows newly synthesized DNA to be used as template, thus greatly enhancing the efficiency of the site-directed mutagenesis reaction and increasing the probability of mutated clones. Briefly, the two primers annealing to one target site are only partially overlapping, which facilitates their reannealing to newly amplified DNA by allowing them to bridge the nick. According to Liu and Naismith, the non-overlapping sequence of the primers are designed to have an annealing temperature 5-10˚C higher than either of the overlapping (complementary) primer sequences. 12 amplification cycles consisted of 95°C for 1 minute, Tm (non-overlapping) -5°C for 1 minute and 68°C for 14 minutes. The PCR cycles were finished with an annealing step at Tm (complementary) -5˚C for 1 minute and an extension step at 72°C for 30 minutes. Primer sequences SPI51D-mutSRp40.1, SPI52U-mutSRp40.1 (mutagenesis of first SRp40 binding site) and SPI53D-mutSRp40.2, SPI54U-mutSRp40.2 and SPI55D-mutSRp40.2-ECS (mutagenesis of second SRp40 binding site) are given in Appendix B.

7.3.2.3 ECS mutants

To destroy the 90bp double-stranded structure, the ECS was mutated at a total of 11 nucleotides. Four rounds of site-directed mutagenesis were required, primer sequences can be found in Appendix B and a graphic depiction of the engineered mutants are shown in Figure 31, page 160.

7.3.2.4 A-to-G mutants

Adenosines subjected to A-to-I RNA editing in the 90 nucleotide target sequence were mutated to guanosines to mimic editing (Figure 32). Twelve adenosines were shown to be edited to a substantial amount. Since simultaneous mutations in the complementary ECS were not desired, I used two previously generated plasmids that were mutated in the ECS (Figure 31a, mutants 2 and 3) to specifically target the mutagenesis primers only to the exonic 90 nucleotide sequence. Four sequential mutagenesis reactions were necessary to mutagenize 12 adenosines in total.

7.3.3 Co-transfection experiments

HeLa cells were transfected with expression plasmids using SuperFect (Qiagen) or X-treme Gene HP (Roche) as described by the manufacturers. The SerpinA3 expression constructs were either co-transfected with an ADAR2 expression vector or with an empty pCI vector as control. RNA was extracted from transfected cells using Trizol reagent (Invitrogen) as suggested by the manufacturer. RNA was treated up to three times with DNase to remove residual plasmid DNA. To this end, RNA was dissolved in 24 μl DEPC-treated water, supplemented with 3 μl 10xDNase buffer, 2 μl DNase I (4U, NEB) and 1 μl RNasin (40U, Invitrogen) and incubated for 1 hour at 37˚C. DNase-treated RNA was phenol-chloroform extracted and precipitated with ethanol in

between treatments. Purified DNase-treated RNA was then reverse transcribed using Superscript III Reverse Transcriptase (Invitrogen) and specific reverse primers SP14U and SP16U at a concentration 50nM each. PCR was performed on the cDNA using construct-specific primers and the amplicons were purified as described previously and sequenced (Geneway).

7.4 Results and Discussion

7.4.1 Genetic elements of SerpinA3

To assess the impact of certain genetic elements on editing in SerpinA3, we cloned parts of the gene into an expression vector. We focused on exon 3, intron 3, and exon 4 for the experiments, as the intron contains both the ECS and genetic elements that have the potential to interfere with the predicted folding of the SerpinA3 transcript. Three partial repeat elements belonging to the LINE family were annotated for this intron. Two L2 elements flank the putative ECS and are located on the negative strand, and a third partial L2 element is located about 200 nucleotides upstream of exon 4 on the positive

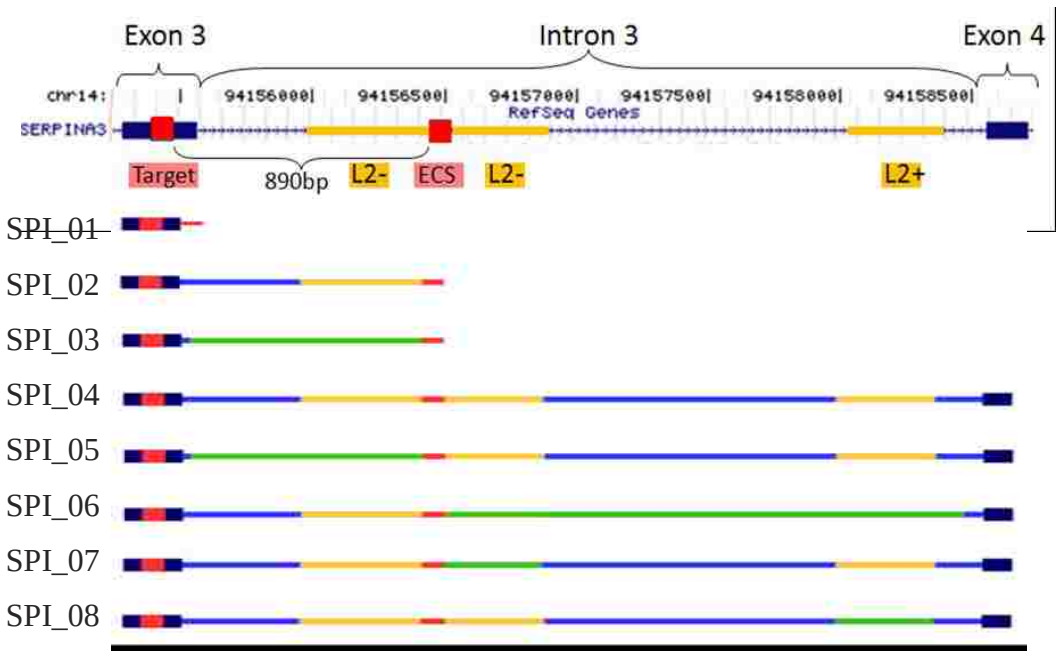


Figure 26: SerpinA3 constructs

Part of the SerpinA3 gene between exons 3 and 4 with genetic elements indicated and the 8 different constructs derived from the wild-type sequence. Constructs are color coded like genetic elements, green parts indicate exchange of sequence with unrelated sequence derived from intron 2 from same gene.

strand (Figure 26). This third L2 element was removed from the sequence in a newer version of the database (hg19, release date February 2009). All three L2 elements are incomplete fragments that do not correspond to the same part of the consensus L2 sequence and alignment using blastn did not reveal sufficient sequence similarity to suggest formation of a strong secondary structure between these elements. In fact, the two L2 elements flanking the ECS are one contiguous partial L2 element interrupted by the insertion of the ECS in recent evolutionary history (see Figure 26 and "7.2: Introduction"). Nevertheless, we decided to analyze whether these elements or any other sequence present in intron 3 may preclude editing by, for example, preventing the formation of the 90bp double-stranded structure. To that end, constructs SPI_01-SPI_08 were cloned, with intronic information exchanged with unrelated sequence of equal length and GC-content, derived from intron 2 of SerpinA3 (Figure 26).

7.4.2 SerpinA3 short transcripts are edited

Parts of the SerpinA3 sequence from exon 3 to exon 4 were expressed in a cell-based system (Figure 26). Even with simultaneous overexpression of ADAR2, mRNA was not edited at detectable levels in the target sequence using Sanger sequencing (Figure 27, threshold set at 10%). Exchanging parts of the intron with unrelated sequences of equal lengths also did not increase editing to a detectable level. However, transcripts from shortened vectors comprising only sequence from exon 3 to the ECS were edited at multiple adenosines to high levels (Figure 27). Indeed, editing levels in the 'short' transcripts are comparable to those seen in promiscuously edited Alu repeat elements (Athanasiadis et al., 2004). It has been shown that editing and splicing are coordinated in some editing targets with intronic ECSs such that editing occurs before splicing

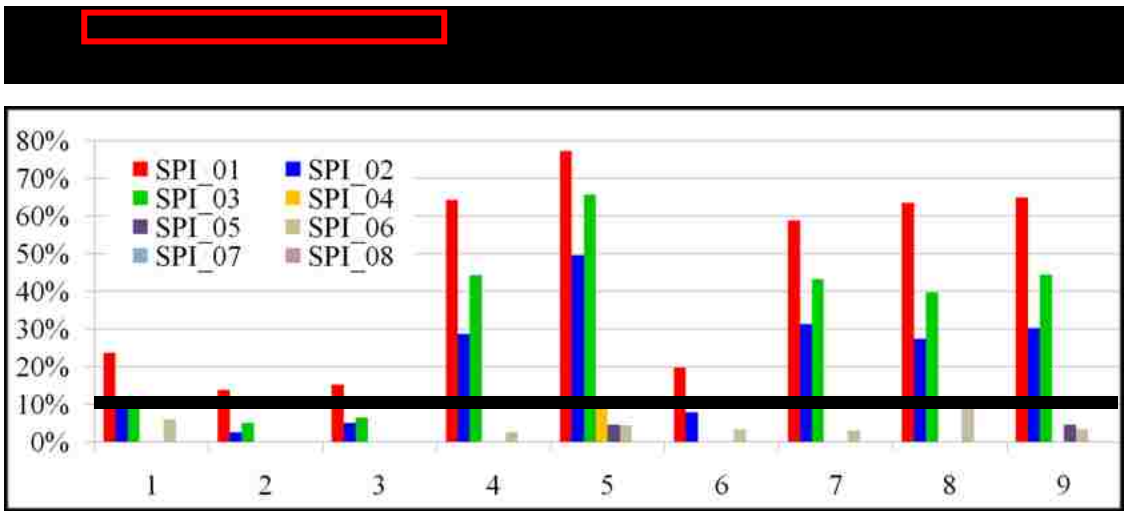


Figure 27: Editing in SerpinA3 transcripts derived from eight expression constructs

Editing in transcripts from eight expression constructs outlined in Figure 26 was evaluated at the first 9 adenosines of the exonic target sequence (adenosines within red box of 90bp dsRNA shown above graph). Evaluation of the rest of the sequence was difficult for the highly edited constructs SPI_01 - _03 due to shifted sequence peaks. No editing above a 10% threshold was detected for the other transcripts further downstream. Dotted line = 10% threshold.

(Laurencikiene et al., 2006; Ryman et al., 2007). It remains to be shown whether this coordination occurs on a more global scale. It is possible that in the case of SerpinA3, or indeed during transcription of most genes, such coordination does not occur and splicing removes the intron before the RNA forms a secondary structure amenable to editing. We therefore decided to destroy the splice sites through site-directed mutagenesis in order to prevent splicing in transcripts SPI_04-SPI_08.

7.4.3 Splice-inefficient transcripts are edited to a low extent

Expression vectors SPI_04 to SPI_08 were mutated at splice-donor and acceptor sites as shown in Figure 28a. Splice inefficiency was shown by PCR amplification of

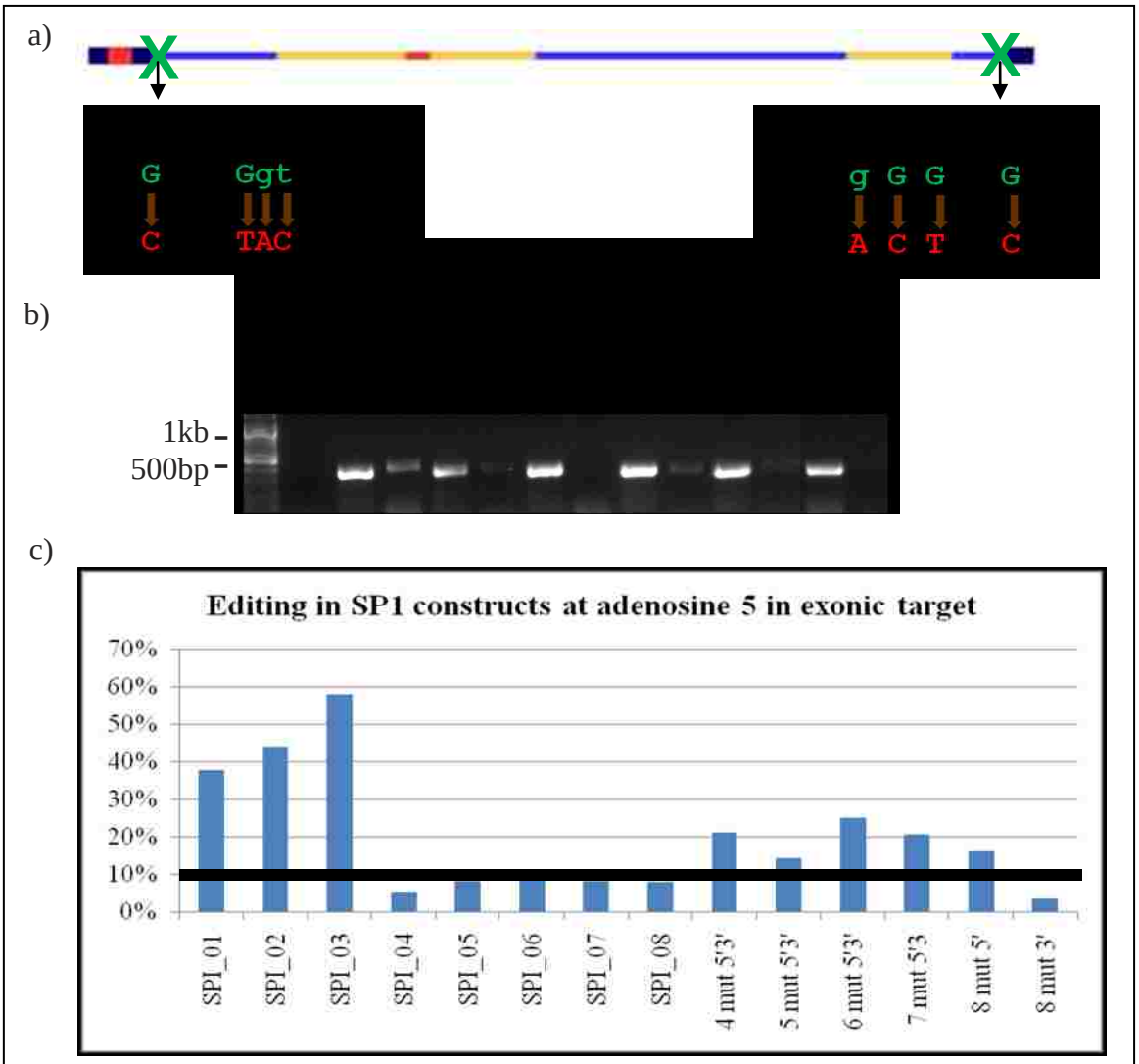


Figure 28: Editing in SerpinA3 transcripts and splice-site mutants

a) Splice sites were mutated in constructs SPI_04 to SPI_08 through site-directed mutagenesis as indicated below graphic (upper sequence = wild-type, lower sequence = mutants). b) Reduction of splice-efficiency was verified by PCR amplification across exon/exon boundaries using primers SPI23D and SPI14U. c) Editing in transcripts from eight expression constructs outlined in Figure 26 and their corresponding splice-site mutants was evaluated at the 5th adenosine of the exonic target sequence. Dotted line = 10% threshold.

spliced transcripts across exon/exon boundaries, using the appropriate primers (Figure 28b). As can be seen in the gel electrophoresis, splice mutants still exhibit some degree of

splicing, despite the substantial mutagenesis at four positions at each splice site. Some amplicons appear to differ in size from those derived from the corresponding non-mutated parent constructs. Sequencing revealed the use of alternative splice donor and acceptor sites, possibly facilitated by the minor spliceosome, which uses recognition sites different from that of the major spliceosome (Will & Luhrmann, 2005). Nevertheless, splicing is at least greatly reduced or almost undetectable in the splice mutants. Yet, when we analyzed editing levels in splice-deficient mutants, only the fifth adenosine in the target sequence showed editing levels reaching up to 25% (SPI_06 mutated at both 5'-donor -and 3'-acceptor sites, Figure 28).

7.4.4 Splice-enhancers

Splicing in SerpinA3 appears to be extraordinarily efficient. Four point-mutations at each splice-site did not suffice to abolish splicing in most constructs. The majority of splice-sites require the action of splice-enhancers for an efficient reaction together with SR-proteins that bind to the RNA at recognition sites and promote assembly of the spliceosome at the splice site (Long & Caceres, 2009). If splice-enhancers were to bind to the exonic target sequence, formation of the 90bp double-stranded RNA structure could be prevented. The recognition sites of splice enhancers are relatively short and ill-defined (Chasin, 2007). Still, web-based algorithms allow the search for putative splice-enhancer on a given sequence (Cartegni et al., 2003; Smith et al., 2006; Desmet et al., 2009). Additionally, SR-proteins bind single-stranded RNA, and thus their binding sequences must be displayed on single-stranded RNA loops (Shepard & Hertel, 2009). In order to predict true splice-enhancer sites within the exonic target sequence, web-derived potential splice-enhancer sites were aligned to the probability that a particular nucleotide

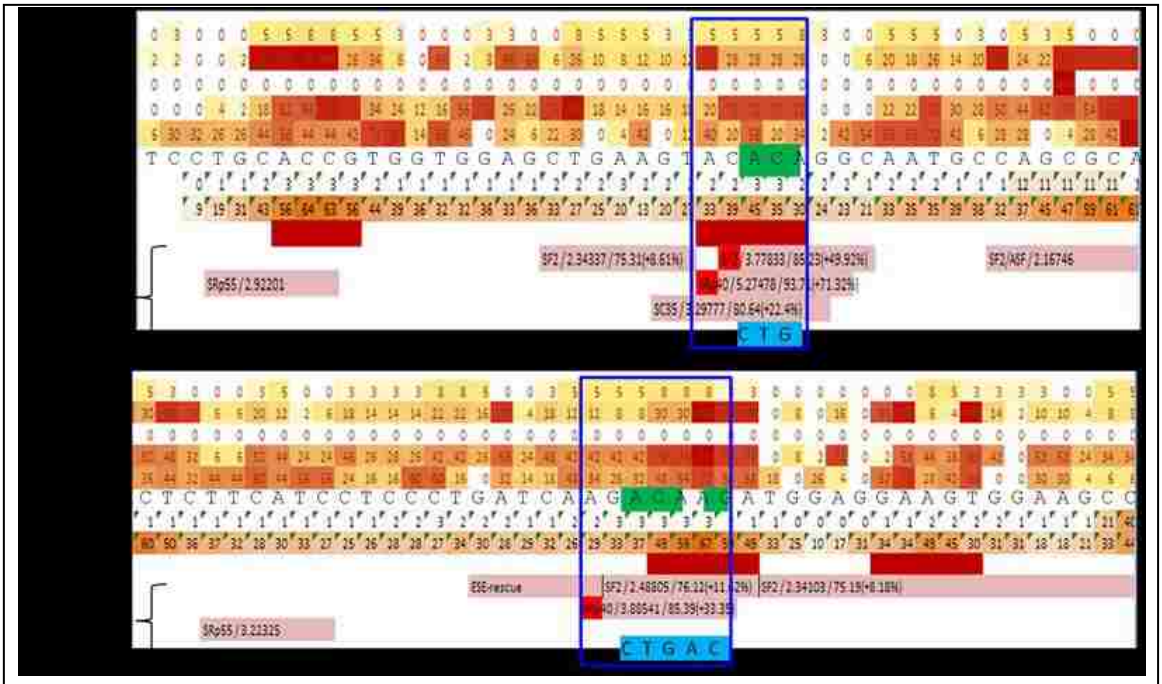


Figure 29: Putative splice-enhancer sites

The ss-count (Mfold) for each nucleotide of SerpinA3 sequences from four species were aligned and color-coded (white = 0%, dark red = 100%). The human sequence without ECS was also analyzed in this manner (H. w/o ECS). Ss-counts were averaged across the three sequences without ECS (H. w/o ECS, orangutan, macaque) or the two with ECS (human and chimp) and likewise color-coded. The nucleotides which are likely to be single-stranded are therefore orange to red. The human sequence shown was then submitted to the ESE Finder 3.0 and Human Splicing Finder Version 2.4.1 (Cartegni et al., 2003). The sites were aligned to the sequence and strong ESE candidates overlapping with single-stranded areas were selected for analysis (blue boxes). Nucleotides in green were mutated to sequence in blue.

be single-stranded (ss-count), as predicted by Mfold (Zuker, 2003). Mfold determines this probability based on folding of up to 50 lowest-energy secondary structures of a given RNA sequence (Zuker, 2003). Since human and chimpanzee both have very low or zero probability of single-stranded nucleotides within the 90 nucleotide target sequence, SerpinA3 sequences from orangutan and macaque, as well as human RNA sequence

lacking the ECS (i.e. with the two L2-arms joined together) were analyzed, assuming that SR-protein binding sites in primates are relatively conserved. Figure 29 shows a heat-map of the percentage of RNA folds in which a given nucleotide is in single-stranded configuration from 0% (white) to 100% (dark red) of the five analyzed sequences. Binning calculated the average of single-strandedness along 5 nucleotides of the two sequences with ECS (human and chimp) or the three sequences without ECS (H. w/o ECS, macaque, orangutan). Color-coding highlights the nucleotides which are more likely to be single-stranded in these closely related species. The human SerpinA3 sequence was then submitted to ESEfinder 3.0 (Cartegni et al., 2003; Smith et al., 2006), and Human Splicing Finder Version 2.4.1 (Desmet et al., 2009), web-based tools that predict potential exonic splice enhancer sites and score them according to their strength. Two strong ESE candidate sites that overlap with areas that are likely to be single-stranded were chosen for analysis. They were mutated in a way as to destroy their consensus binding sites for the SRp40 and SF2/ASF splice-enhancer proteins.

Sequence track analysis of amplified transcripts derived from these constructs did not show evidence of editing in the 90 nucleotide target sequence (data not shown). Additionally, splicing was not inhibited as estimated from the PCR amplification (data not shown). Despite careful analysis and computational predictions it is possible that the predicted and mutated ESE recognition sites may not represent true ESE binding sites.

7.4.5 Pre-messenger RNA is edited promiscuously

For amplification of transcripts from splice-mutant minigene constructs, a reverse primer annealing to the 3’-end of exon 3 was used, and thus both spliced and unspliced transcripts were amplified. The forward primer, spanning a plasmid-specific intron,

ensures that only transcripts from expression constructs were amplified. We hypothesized that differences in editing in spliced and unspliced transcripts could explain the findings obtained thus far. If editing levels were higher in unspliced transcripts, slightly higher levels of editing in splice-deficient transcripts could be expected, as pre-mRNA may accumulate more than from splice-competent transcripts. To address this question, we amplified pre-mRNA specific sequences using a reverse primer annealing to the intron of transcripts from select constructs and their corresponding splice-inefficient variants (Figure 30a). To our surprise, all transcripts, regardless of expression construct, were edited to a high level at multiple adenosines within the 90 nucleotide exonic target sequence. The high percentage of edited pre-mRNA obscured the sequence track

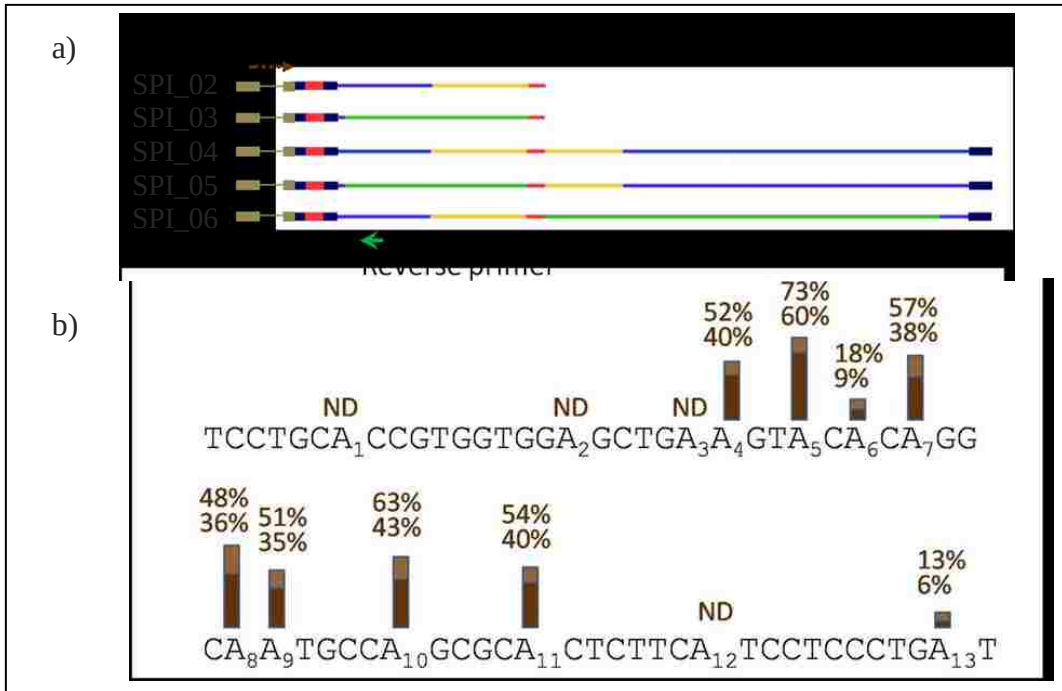


Figure 30: Editing in pre-mRNA

a) Constructs and primers used for analysis of editing in pre-mRNA.

b) Portion of 90 nucleotide target sequence analyzed for editing in pre-mRNA. Bars indicate range of editing observed at the specific adenosine.

analysis. The bars in Figure 30b depict the range of editing levels in the pre-mRNA variants at a specific adenosine in the 90 nucleotide target sequence. Editing extents are independent of the transcript variant (i.e. the original construct).

While pre-mRNAs of SerpinA3 transcripts from expression plasmids are promiscuously edited, mature mRNAs show low to no evidence of editing. Several possible scenarios could explain this observation. First, edited pre-mRNA and/or mRNA may be degraded or otherwise made unavailable for amplification. Tudor-SN is an RNase that specifically recognizes and cleaves inosine-containing RNA and has been shown to associate with the U5 snRNP of the spliceosome (Yang et al., 2007). It is tempting to speculate that Tudor-SN recognizes and cleaves edited SerpinA3 RNA, leading to its directed destruction. Only unedited transcripts and/or those with few inosines would persist. Alternatively, formation of the strong 90bp double-stranded structure could trigger removal of such folded transcripts from the system. If the last scenario were true, one would expect to see no editing at all in any mRNA. However, editing levels of up to 10% and 25% were observed in certain SerpinA3 transcript variants and splice-inefficient mutants, respectively. Consequently, a double-stranded structure must have been present for these transcripts to be edited. Still, it is important to determine whether edited transcripts are cleared form the detectable pool of transcripts due to editing or due to a strong double-stranded secondary structure.

7.4.6 Editing, not double-strandedness, causes transcripts to be cleared

To determine whether double-stranded structures degrade or otherwise make SerpinA3 pre-mRNA undetectable, the intronic ECS was mutated at eleven positions to disrupt the perfect complementarity between it and its target sequence. Four sequential

site-directed mutagenesis reactions were performed on plasmid SPI_04, yielding expression constructs with consecutively smaller extents of complete complementarity (Figure 31a). While editing levels are lower than in SPI_04-transcripts, as expected due to the successively smaller sections of complementarity, editing is detected in pre-mRNA (Figure 31b) but not in processed mRNA (not shown) of these ECS mutants. We conclude that the strong double-stranded structure by itself is not required for reducing the pool of edited pre-mRNA.

7.4.7 Editing triggers degradation of pre-mRNA

As can be deduced from the editing levels seen in pre-mRNA, most pre-mRNAs are edited (the fifth adenosine is edited up to 75%). However, editing levels are low to undetectable in PCR products amplified with a primer that amplifies both mRNA and pre-mRNA alike. This observation is significant in so far as it implies that the edited pre-mRNA pool is substantially smaller than the corresponding unedited spliced mRNA pool. Therefore, two fates exist for a pre-mRNA: either it gets edited and rapidly degraded, or it remains unedited, matures and is relatively stable.

There are two possibilities for how the edited pre-mRNA is recognized: either by recognition of A-to-G changes or by recognition of inosines. To distinguish between these possibilities, adenosines that are highly edited in pre-mRNA were mutated to guanosines in expression constructs (Figure 32b). Since mutagenesis primers could anneal to two identical sequences in the wild-type minigene (SPI_04), we chose instead to mutate the ECS mutants 1 and 2. This provided for some sequence specificity of the primers, which were used to mutate 12 adenosines in four sequential mutagenesis

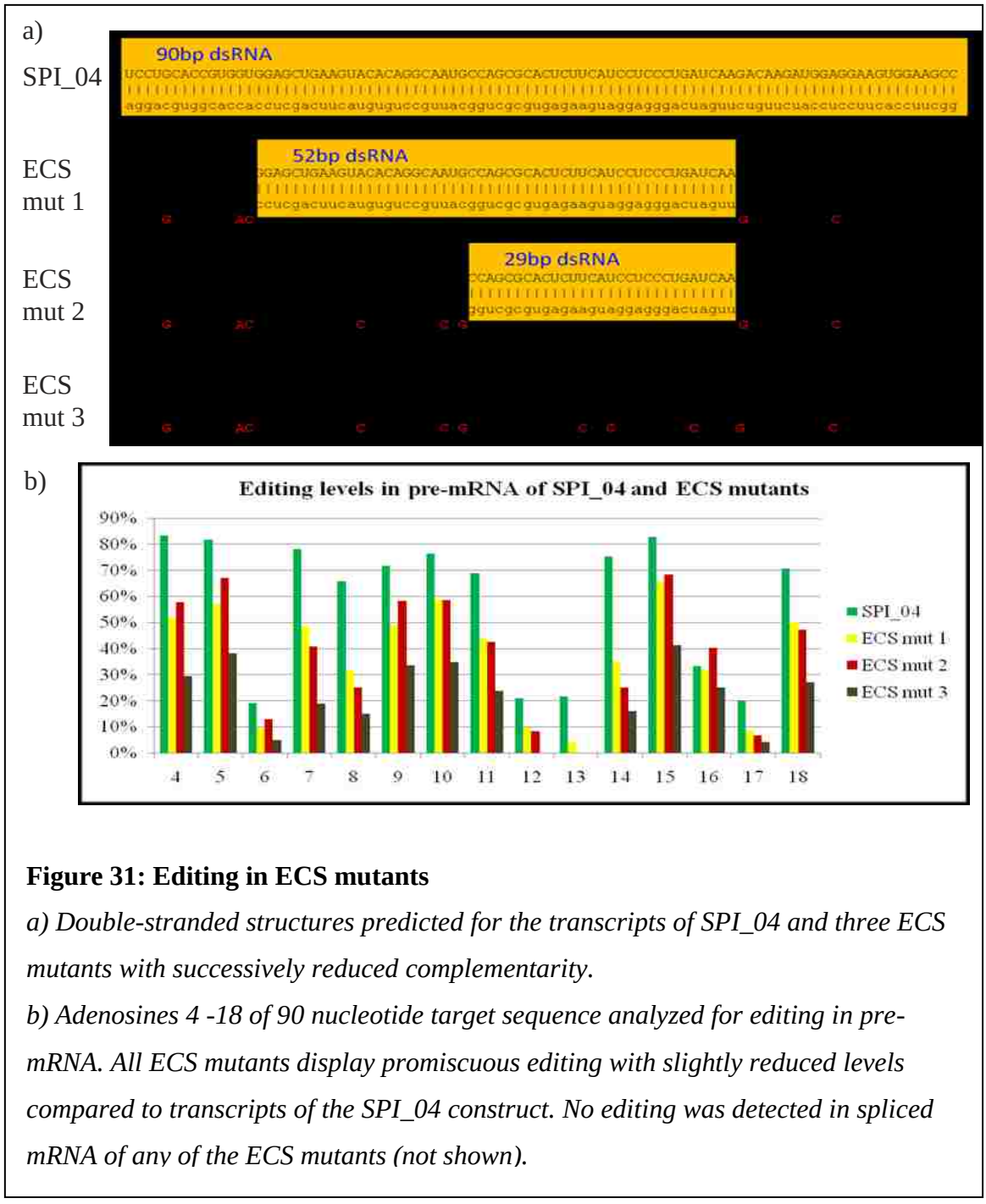


Figure 31: Editing in ECS mutants

a) Double-stranded structures predicted for the transcripts of SPI_04 and three ECS mutants with successively reduced complementarity.

b) Adenosines 4 -18 of 90 nucleotide target sequence analyzed for editing in pre-mRNA. All ECS mutants display promiscuous editing with slightly reduced levels compared to transcripts of the SPI_04 construct. No editing was detected in spliced mRNA of any of the ECS mutants (not shown).

reactions. If the inosines were specifically recognized, these mutants should be immune to the mechanism that rids the system of edited pre-mRNA and therefore accumulate mature RNA faster than editing-competent transcripts (Figure 32a). If, on the other hand, the sequence changes from A to G trigger such a mechanism, the mRNA of the A/G

mutants would be targeted for destruction just like the edited portion of pre-mRNA in editing-competent transcripts. Indeed, there may be no accumulation of mRNA from A/G mutants in that case.

The qRT-PCR analysis planned initially to quantify pre-mRNA and mRNA levels proved to be problematic, possibly due to an inhibitory effect of the strong double-stranded structure that can form between exon and ECS. Similarly, residual pre-mRNA in the cDNA preparation might also anneal to the 90 nucleotide target sequence. Additionally, such annealing might result in fluorescence to give a false-positive signal. A semi-quantitative PCR reaction was performed instead of the qRT-PCR. Preliminary results showed no signficiant differences in amplification among wild-type (SPI_04), ECS mutant and corresponding A/G mutant, neither in unspliced nor spliced transcripts. It is interesting that A/G mutant mature mRNA can be amplified, indicating that the nucleotide changes do not appear to influence pre-mRNA processing. While these preliminary results are not conclusive, they at least show that splicing is not inhibited by the presence of guanosines in place of adenosines in the 90 nucleotide target sequence.

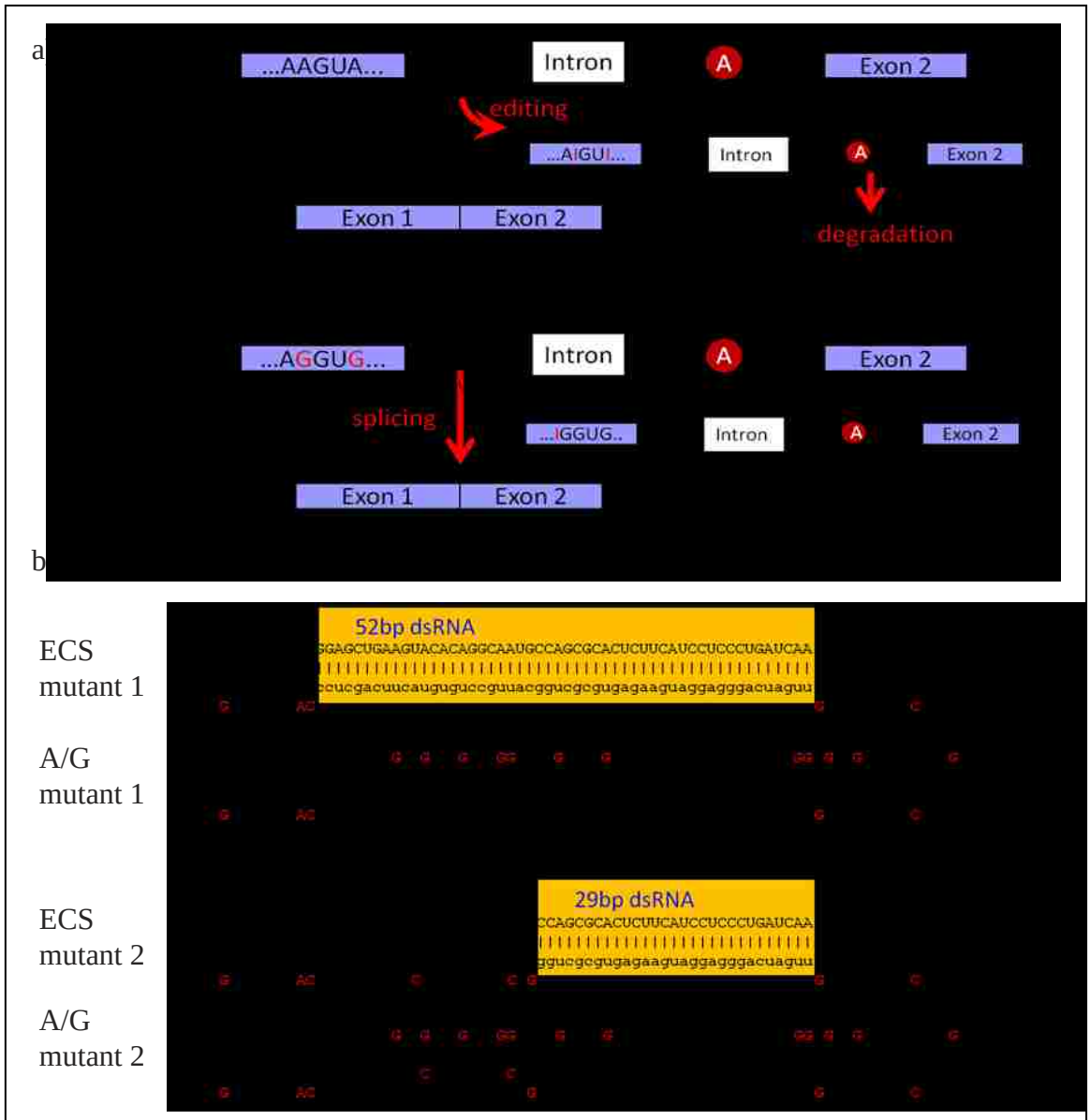


Figure 32: A/G mutants

a) Model demonstrating how editing, splicing and elimination of edited pre-mRNA might alter the equilibrium of unspliced and mature mRNA. Top of the schematic shows that wild-type pre-mRNA is primarily edited and thus somehow eliminated. At the bottom, this equilibrium might be changed in A/G mutants with no editing in pre-mRNA and thus a higher rate of splicing.

b) ECS mutants 1 and 2 were mutated in the 90 nucleotide target sequence at adenosines that are highly edited in the cell culture experiments. Sequence changes are highlighted in red.

Therefore, it is likely that the inosines themselves are responsible for triggering the elimination of edited pre-mRNA. Tudor-SN is an inosine-specific RNase and was shown to target edited pri- and pre-miRNAs for degradation. We hypothesize that Tudor-SN, which is associated with the U5 snRNP and functions in spliceosome assembly and pre-mRNA splicing, recognizes hyper-edited SerpinA3 pre-mRNA and initiates its degradation. Inhibition of Tudor-SN *in vivo* with the specific inhibitor Thymidine 3',5'-diphosphate would lead to the accumulation of edited pre-mRNA if Tudor-SN was indeed involved in the degradation of edited SerpinA3 pre-mRNA. This could be assess by determining the editing levels in amplicons obtained using an exon3-specific reverse primer, which leads to amplification of both pre-mRNA and mRNA.

7.5 Conclusions

SerpinA3 inhibits cognate proteases by the insertion of a reactive center loop into their active sites and the formation of an irreversible, stable complex (Devlin & Bottomley, 2005). This activity requires a metastable native conformation, which allows it to undergo a large conformational change upon binding. Such metastability is easily disrupted by mutations, which manifest themselves as devastating diseases termed serpinopathies (Crowther et al., 2004; Lomas, 2005; Davies & Lomas, 2008).

SerpinA3 pre-mRNA is predicted to form a 90bp completely complementary structure between a 90 nucleotide exonic sequence and an intronic ECS. For this reason, the REDS program listed SerpinA3 as highest-ranking candidate. Its lack of editing in mRNA from several human tissues led us to question the reasons why this presumably 'perfect' ADAR target is not edited. We show here that while fully mature mRNA does not have detectable levels of editing, SerpinA3 pre-mRNA is indeed promiscuously edited. A mechanism that we hypothesize to involve Tudor-SN appears to ensure that edited SerpinA3 pre-mRNA does not mature into functional mRNA. The potential existence of a surveillance mechanism in this transcript is not surprising, considering the devastating effect promiscuous editing and thus a host of recoding events would have on the functionality of SerpinA3. The nature of this surveillance mechanism and whether it acts on a more global scale remains to be elucidated. The insertion of the ECS into the L2 element appears to be a recent event in evolutionary terms, as it is present in human and chimpanzee, but not in macaque and orangutan. This mechanism must have been in place when this insertion happened, in order to prevent promiscuous editing, and allow the insertion to persist.

When highly edited adenosines were mutated to guanosines in minigene constructs, transcripts were still processed to fully mature mRNA. This indicates that the surveillance mechanism specifically recognizes inosines, and not the nucleotide changes. Tudor-SN is a RNase that specifically recognizes and cleaves inosine-containing RNA. It has been shown to associate with the U5 snRNP of the spliceosome. We hypothesize that Tudor-SN is involved in recognition and degradation of promiscuously edited pre-mRNA as part of this surveillance system. This mechanism might act in a manner similar to the one shown to function in the miRNA pathway (see Chapter 1 and Yang et al., 2006). Indeed this mechanism might act in several instances where promiscuous editing occurs in exonic sequences, as evidence of editing is absent in many of the high-ranking candidates predicted by the REDS program.

8 Conclusions

8.1 Bioinformatics prediction of A-to-I RNA editing recoding sites in the era of RNA deep-sequencing

Recent technological advances have made in-depth analyses of transcriptomes available to a larger research community through relatively affordable high-throughput sequencing. Will the brute force of vast amounts of sequence data reduce the need for bioinformatics-driven predictive models? Recent deep-sequencing analyses successfully identified many more editing targets than previously known (Li et al., 2009; Enstero et al., 2010). High-throughput analyses today usually only allow sequencing of short stretches of transcripts, limiting their efficacy to specific sites within in the transcriptome. Thus, analyses to date rely on computational screening methods to arrive at a candidate list of putative editing targets. Therefore, at least for the moment, computational screening and high-throughput sequencing go hand-in-hand. Consequently, the experimental data can only be as good as the initial bioinformatics screening.

Due to the increasing information content in publicly available sequence databases, recoding editing events with high penetrance and wide tissue distribution should be readily detectable through a screen such as REDS. Recent bioinformatics-driven screens from us and others, and a few deep-sequencing analyses, indicate that few high-level editing sites exist, most of which have now been identified. However, it is increasingly clear that editing is regulated in a target-, tissue-, time-, and/or stimuli-dependent manner. The challenge now is to identify recoding editing sites that are of consequence and possibly highly regulated. Again, a bioinformatics approach with high predictive force is indispensable to pinpoint such elusive targets. A program with a high data flow-rate, that allows analysis of several databases within a short period of time,

adjustment of user-defined screening settings, and integration of new insights into ADAR target preferences in combination with high-throughput sequencing could potentially enable the ambitious search for targets that are edited only intermittently. The incorporation of nearest-neighbor analysis, optional analysis of the degree of conservation, and information about the tissue-specificity of mRNAs with annotated G-discrepancy might be sufficient refinements of the REDS algorithm to support such an endeavor.

8.2 Consequences of editing on protein function

Functional analysis of the effects of recoding events on the affected protein variants is an important part of assessing the impact of editing on transcriptome and proteome diversity. The need to characterize consequences of editing on protein function is growing with the discovery of more targets such as C1QL1, IGFBP7, FLNA, CDK13, and COPA, to name a few. In order to conceive valid assays that test relevant functions, each target requires a streamlined study that integrates knowledge of protein function, domains, interaction partners, post-translational modifications and processing. Integration of available information can enable a well-informed experimental approach to allow testing the functionality of the different protein isoforms. Another way to elucidate the impact of editing on protein function is the creation of transgenic animals that are either incompetent in editing a certain target (by removal of the ECS for example), or that can only express the edited variant. This allows analyzing effects in an organismal context, enabling researchers to assess effects on various tissues and at different developmental time-points. While successful in some cases (for example for the GluR-B Q/R site), it is expensive and time-consuming and does not guarantee success in determining the

underlying molecular bases of observations. In conclusion, analysis of functional consequences for the ensuing protein variants upon editing may require vastly different experimental analysis methods, rationalized upon the distinct features of the affected protein. Much research is needed on this aspect of A-to-I RNA editing.

8.3 Unexplored RNA world

Much of the A-to-I RNA editing field has so far focused on the identification and characterization of recoding events. Due to the increasing awareness of the importance of non-coding RNAs in the past few years, new efforts have been made in trying to elucidate the consequences of editing on non-coding RNAs such as miRNAs. Indeed, functional RNAs are not, as long presumed, the solitary remnants of the primordial RNA world in which self-replicating, catalytic RNA reigned. Instead, RNA has continued to play a central role throughout evolution. In fact, non-protein coding portions of genomes, the majority of which is pervasively transcribed from yeast to human (Kapranov et al., 2002; Miura et al., 2006; Kapranov et al., 2007), increase with increasing complexity of organisms, while the quantity of sequences coding for proteins remains comparatively similar (Lander et al., 2001; Stein et al., 2003; International Human Genome Sequencing Consortium, 2004). Although co-transcriptional modifications such as editing and alternative splicing considerably expand proteome diversity from a limited number of genes, they can only partially explain the elaborateness of metabolic, developmental, architectural, and cognitive systems in higher organisms. The true force underlying organismal complexity may be more related to the sophistication of regulatory networks encrypted in non-coding sequences. Therefore, it has been proposed that the complexity

of higher organisms is directly related to the information stored in the non-protein coding part of a genome (Taft et al., 2007).

Even though ncRNAs are starting to receive the attention they deserve only of late, their involvement in many regulatory networks is already apparent, such as regulating the epigenetic state of a cell and controlling gene expression on transcriptional and translational levels in both *cis* and *trans* (Costa, 2010; Mattick, 2010; Tsai et al., 2010). The combinatorial control of complex networks governed largely by regulatory RNAs seems to be particularly important for the highly complex human nervous system. Consequently, ncRNAs are now subject of intense research efforts, aimed at trying to elucidate their various roles in cellular functions. A critical aspect of how a cell controls the actions of such ncRNAs is its ability to modulate and regulate them both on a transcriptional level and through posttranscriptional regulation and modification. RNA interacting partners in general and RNA modifying enzymes in particular may assume the critical responsibility of directing and modulating functional RNAs. ADARs have been shown to edit and as a result profoundly impact the regulation and function of miRNAs, but we are just beginning to explore their regulatory power on other RNAs. Alu elements appear to be evolutionarily important for primates, especially in conjunction with A-to-I RNA editing (Eisenberg et al., 2005b). This notion has been supported by the finding that editing can lead to the exonization of an Alu element by changing an AA dinucleotide into an AG 3' acceptor splice site (Lev-Maor et al., 2007). The sophisticated regulation of neurons mediated by RNAs may in part explain the difference between human and other species. Incidentally, inosine-levels are highest in brain (Paul & Bass, 1998). In order to comprehend how regulatory RNAs exert their functions, it is important to elucidate how

these RNAs are governed by their own superimposed regulatory networks – attacking this uncharted territory promises to break new ground in a young research field.

9 Appendix A

9.1 Computational methods

9.1.1 Databases and Bioinformatics procedure for scoring and ranking candidates for RNA editing, SNP study

Annotations for human SNPs from the dbSNP database build 125 (Sherry et al. 2001) were downloaded using the UCSC genome table browser (Kuhn et al. 2007). For subsequent analysis of candidate genes the UCSC human genome browser (assembly May 2004) was used.

Cross-species conservation was analyzed on two levels. Initially, conservation was evaluated for all 554 candidate genes using the UCSC genome browser conservation track, which is based on the phastCons program designed to identify conserved elements in multiple aligned sequences (Siepel et al. 2005). PhastCons is based on a phylogenetic hidden Markov model (phylo-HMM), a type of statistical model that considers both the process by which nucleotide substitutions occur at each site in a genome and how this process changes from one site to the next (Siepel et al. 2005). PhastCons produces a continuous valued ‘‘conservation score’’ for each base of the reference genome. The conservation score at each base in the reference genome is defined as the posterior probability that the corresponding alignment column was generated by the conserved state (rather than the nonconserved state) of the phylo-HMM, given the model parameters and the multiple alignment. Therefore, the scores range between 0 and 1, corresponding to 0%– 100% conservation.

All 554 candidate genes were grouped into five bins according to the PhastCons score covering the region of 15 nucleotides (nt) upstream of and 15 nt downstream from each candidate site for editing. The bins were: high (H), for conservation of higher than

90%; high-medium (HM) for conservation between 75% and 90%; medium (M) for conservation of 50%–75%; medium-low (ML) for conservation of 25%–50%; and low (L) for conservation <25%. Only candidates from the H and HM bins were used for further analysis.

The second level of cross-species conservation taken into consideration was the conservation of the potentially edited adenosine. Candidates where only human and mouse homologous carry an adenosine at the predicted editing position, but not the rat counterpart (and/or chicken if available for the gene), were eliminated from further analysis even if previously grouped into the H or HM bin. This two-step evaluation of cross-species conservation is based on the data from known editing sites where the sequence surrounding the editing site as well as the edited adenosine itself are conserved to a higher degree than the general conservation of exonic, coding sequences, since in addition to encoding amino acids, the sequences also participate in forming a functional RNA structure.

293 of the 554 candidate sites remained for further analysis, whereas 261 entries were filtered out at this step. Next, evidence for *in silico* editing was analyzed for each of the 293 sites using the BLASTN program (NCBI). To this end 30nt upstream of and 30 nt downstream from the predicted sites were successively blasted against the nr (NCBI) and the human EST databases (NCBI) and the percentage of sequences that carry a G instead of an A at the predicted site was recorded. For 204 candidates in silico editing was equal to or higher than 1%, whereas for 89 entries no evidence of editing was detected in silico.

The possibility of a RNA fold-back structure was then investigated for each of the 204 remaining candidate genes. In known cases of RNA editing, the RNA fold-back

structure usually involves the exonic sequence immediately surrounding the edited adenosine and an editing site complementary sequence (ECS), which is often located in the downstream intron in mammalian targets. For fold-back analysis we used the MFOLD program (Zuker 2003) in the batch mode, which allows for the folding of up to 800 nt of RNA sequence. Initially, 700 nt upstream of and 100 nt downstream from, or 100 nt upstream of and 700 nt downstream from, the predicted editing site were run and the resulting secondary structures were inspected for fold-back substructures that included the immediate region surrounding the predicted site. If no distinctive structure was found, additional sequences were folded using MFOLD by selecting z100 nt upstream of and downstream from the predicted site together with up to 600 nt of sequences from another region within the gene and < 2.5 kb upstream of or downstream from the predicted site. This selection is based on known edited genes, where the ECS was found to be located in intronic regions up to a few kilobases away from the exonic editing site. Only those sequences were selected that showed a high degree of conservation according to the PhastCon track of the UCSC human genome browser.

The substructure or substructures covering the sequence region around the predicted editing site that showed the highest doublestranded character for each candidate were then grouped into bins 1–5 based on a calculated structural score (STR). The structural score STR was obtained from values for three different features determined for each evaluated candidate. First, the base-pairing (BP) score was calculated, which corresponds to the number of base pairs present in the structure multiplied by the fraction of nonbase-paired nucleotides [BP = bp(1-bp/nt)]. The value for this feature reflects the

fraction of nucleotides that are base paired in the structure, and also accounts for the total lengths of the structure including base-paired as well as nonbase-paired nucleotides.

Second, the GC content of the base pairs was analyzed (the GC score) by determining the sum of base pair values using a value of 3 for a G/C base pair and a value of 2 for an A/T or a G/T base pair.

Third, a penalty value (IS score) was determined for the length of intervening sequence between the two base-paired regions, as our previous study of intramolecular folding and editing of Alu-element-containing sequences showed that the level of editing decreases with an increasing size of the intervening sequence. The individual IS score bins were: Intervening sequence of >100 nt: penalty reducing score by 10%; >500 nt: 18%; >750 nt: 23%; >1000 nt: 30%; >1250 nt: 38%; >1500 nt: 45%; >1750 nt: 51%; >2000 nt: 60%; and >2500 nt: 80%.

The overall structural score STR follows as:

STR = BP3GC – IS.

Candidate structures with a STR score <100 were placed in bin 5; scores between 100 and 300 in bin 4; scores between 300 and 900 in bin 3; scores between 900 and 1800 in bin 2 and scores >1800 in bin 1.

Our scoring of fold-back structures is uniquely tailored to identify folds that are more likely functional in supporting RNA editing and does not simply select the most thermodynamically stable structures. This is, for example, reflected in that the penalty for intervening sequences between the base-pairing regions is based on the known and characterized editing targets

For each of the molecular features analyzed (identity of -1 and +1 nucleotide; conservation; structure, as described in the Results section) we then computed a comparative score. For each feature I with a value xi we calculated a log-odds score:

$$si(xi) = log2\,(\frac{fi(xi)}{gi(xi)})$$

based on a relative entropy approach (Lim et al. 2003). fi(xi) corresponds to the frequency of the parameter value xi in the reference set of known edited exons, and gi(xi) being the frequency of xi among the sample set of all pre-mRNA sequences in our prefiltered database. Finally, a combined score for each candidate editing site is derived from the sum of the log-odds scores for each analyzed parameter:

$$S = \sum_{i=1...4} si(xi)$$

9.1.2 Computational Methods and Databases used for the REDS study

Human, mouse and zebrafish genomic DNA sequences, mRNA data files, SNP data, as well as table annotations were retrieved using the UCSC genome browser ftp site (assembly March 2006) (Kuhn et al., 2007). REDS consists of three consecutive computational stages (Supplemental Fig. S1). Stage 1 aligns expressed sequences (UCSC mRNA database) from a given species to the corresponding genomic sequence. Coding sequences are translated, allowing for determination of non-synonymous codon positions within ORFs. According to user specifications (type of base difference, coding versus non-coding, synonymous versus non-synonymous), specific types of base discrepancies are mapped and output with chromosome location and position, mRNA accession and

position, gene ID and description and affected amino acid. Previously known RNA editing site are flagged. Stage 2 compares the list of base discrepancies to the species-specific SNP database (Sherry et al., 2001) and all positions that correspond to genomically validated SNPs are filtered out.

The third stage of the computational pipeline evaluates RNA folding characteristics for each of the remaining sites. The user defines several parameters: a first sequence window determines how much of the genomic sequence upstream and downstream of the putative editing site will be analyzed. A second sequence window selects a small gene section surrounding the candidate site, for which the algorithm generates the reverse complement sequence with which to scan window 1 for matches (including G-T wobble base pairs and allowing for single, symmetric mismatches). Furthermore, the cut-off value for how many consecutive base pair matches within the RNA secondary structure are required to pass the filter and finally a minimum value for the length of the intervening sequence between the base-pairing sections are determined.

A multi-segment heuristic combines pairs of base-pairing segments within the sequence that may be part of one composite RNA secondary structure. A score is assigned to this composite structure, whereby base-pairs involving nucleotides within the inner window 2 are weighted more strongly according to a user-defined value. This biases the search for *bona fide* editing targets since base-paired segments of an RNA fold that supports editing include, or are in close vicinity to, the editing site. Finally, the output list of candidate sites is ranked by an overall score. For further information about REDS see Supplementary Documents.

9.2 Molecular Biology materials and methods

9.2.1 RNA isolation from adherent cell culture

Cells grown in monolayer were lysed directly in the dish by adding 100µl TRIzol reagent/cm^2 surface, passing the cells multiple times through a pipette and incubating the samples 5 minutes at RT. 200µl chloroform per 1ml TRIzol were added and tubes shaken vigorously for 15 seconds. After incubation for 3 minutes at RT, samples were centrifuged for 15 minutes at 12,000g at 4°C. The upper aqueous phase was mixed with 0.5ml isopropanol per 1ml TRIzol used and incubated 20 minutes at RT to precipitate RNA. RNA was pelleted by centrifugation for 10 minutes at 12,000g at 4°C. The RNA pellet was washed once with 75% ethanol, centrifuged for 5 minutes at 7,500g at 4°C, air dried and dissolved in DEPC-treated ddH_2O.

9.2.2 Phenol/chloroform extraction and ethanol precipitation of nucleic acids

Phenol/chloroform extraction was performed to purify nucleic acids from protein contaminations after e.g. DNase treatment. An equal volume of phenol/chloroform/ isoamylalcohol (PCI, Invitrogen) was added to the aqueous nucleic acid solution, vortexed for 10 seconds and centrifuged at maximum speed for 5 minutes at RT. The upped aqueous phase was transferred to a fresh tube, mixed with an equal volume of chloroform/isoamylalcohol (24:1), and centrifuged at maximum speed for 5 minutes at RT. The upper aqueous phase was again transferred to a fresh tube and nucleic acid was precipitated by adding 1/10 volume 3M NaOAc, pH 3, and 2.5V 100% EtOH. Usually, nucleic acids were allowed to precipitate for 20 minutes to over night at -20°C. Precipitates were pelleted by centrifugation for 30 minutes at maximum speed at RT in a

tabletop centrifuge. Pellets were washed once with 75% EtOH, centrifuged 5 minutes at maximum speed, the supernatant removed and air dried before being dissolved in ddH_2O (DEPC-treated for RNA).

9.2.3 DNase treatment

To remove contaminating traces of DNA from RNA isolates, RNA samples were treated multiple times with either 3-5 units RQ1 RNase-free DNase (Promega) or 6 units TURBO DNase (Ambion) according to the protocols provided by the manufacturers. The reactions were stopped by Phenol/Chloroform extraction and Ethanol precipitation.

9.2.4 RT-PCR

Tissue specific total RNA was reverse transcribed in a reaction containing 2-10µg of RNA, 250ng random primers, 1mM dNTPs, 10mM DTT, Rnasin, first strand buffer and 150U SuperScript III Reverse Transcriptase (Invitrogen) in 20µl final volume. When later amplification by PCR was expected to be difficult, specific reverse primers were added to the mix at a concentration of 100nM (e.g. for C1QL1 amplifications: *homo sapiens* C1Q13U; *mus musculus*: mC1Q8U; *danio rerio*: drC1Q16U)

For subsequent amplification of any of the other targets, addition of a reverse primer during reverse transcription was not necessary. The reaction was incubated at 55°C for 30 minutes, another 150U reverse transcriptase was added, and the reaction was allowed to proceed for another 30 minutes before it was stopped by incubation at 70°C for 15 minutes.

9.2.5 RNA editing analysis

Human total RNA and gDNA isolated from the same specimen (Biochain) were processed using standard protocols for reverse transcription and PCR. Gene-specific fragments of cDNA as well as genomic regions were amplified by PCR and subjected to dideoxy-sequencing as described previously (Athanasiadis et al., 2004; Gommans et al., 2008). Editing at the predicted positions was analyzed by inspecting the sequence traces for double peaks, with the ratio of the peak heights giving an indication of approximate editing levels. The occurrence of SNPs at candidate positions was excluded by analyzing the specimen-matched gDNA.

9.2.6 Site-directed mutagenesis

Site-specific mutations (up to 5 point mutations at once) were introduced into cloning vectors through site-directed mutagenesis using HotStart PfuTurbo polymerase (Invitrogen) and specific mutagenesis primers (forward and reverse), which generally consisted of approximately 20 nucleotides flanking (the) desired nucleotide change(s) on either side. Reactions contained 1x Pfu reaction buffer, up to 200ng dsDNA template (vector), 300nM of forward and reverse primer each, 200μM dNTP mix, and 2.5U of PfuTurbo DNA polymerase in a total volume of 50μl. Reactions were overlayed with 30μl mineral oil to prevent evaporation. An initial denaturation step at 95°C for 30 seconds was followed by 12-18 cycles of 95°C for 30 seconds, 55°C annealing for 1 minute, and 68°C extension for 1 minute/kb of plasmid length. A final extension occurred at 72°C for 15 minutes. Methylated, non-mutated parental DNA was digested with 10 units DpnI at 37°C for one hour. Products were purified with phenol/chloroform

extraction and precipitated with ethanol. Typically, half or all of the mutagenesis reaction was used to transform Z-competent E. coli cells (see below).

9.2.7 Z-competent cells

Z-competent E. coli were prepared using the Z-Competent E. coli Transformation Buffer Set™ (Zymo Research) according to the manufacturer's instructions, aliquoted at 150µl and stored at -80°C.

9.2.8 Transformation

Z-competent cells were thawed on ice, DNA to be transformed was added, the tubes briefly flicked and incubated on ice for 20 minutes. After a heat shock at 42°C for 45 seconds (alternatively, 37°C for 90 seconds was also successfully used), cells were placed on ice for 2 minutes before addition of 250µl SOC medium. Cells were allowed to recover for one hour at 37°C under gentle agitation before plating up to 200µl of transformants onto LB selection plates (containing either ampicillin or kanamycin).

9.3 Protein biology

9.3.1 Co-transfection experiments

Plasmids were used for transfection of HeLa cells using SuperFect (Qiagen) as described by the manufacturer. The construct was either co-transfected with an ADAR2 expression vector or with an empty pCI vector as control. RNA was extracted from transfected cells using Trizol reagent (Invitrogen) as suggested by the manufacturer. RNA was treated up to three times with DNase to remove residual plasmid DNA. To this end, RNA was dissolved in 24 µl DEPC-treated water, supplemented with 3 µl

10xDNase buffer, 2 μl DNase I (4U, NEB) and 1 μl RNasin (40U, Invitrogen) and incubated for 1 hour at 37˚C. DNase-treated RNA was Phenol-Chloroform extracted and precipitated with ethanol in between treatments. Purified DNase-treated RNA was then reverse transcribed using Superscript III Reverse Transcriptase (Invitrogen) and specific reverse primers SP14U and SP16U at a concentration of 0.8 M each. PCR was performed on the cDNA using construct-specific primers and the amplicons were purified as described previously and sequenced (Geneway).

9.4 Cell biology

9.4.1 Cell culture

HeLa and HEK293 cells were maintained in 1xMEM (Cellgro) containing 10% FCS and 1x antibiotic/antimycotic solution. Cells were trypsinized at least twice a week and never allowed to grow over-confluent.

10 Appendix B

10.1 Media, buffers and solutions

10.1.1 HeLa and HEK293 cell culture medium

500ml 1xMEM (Cellgro)

5.5ml Antibiotic/Antimycotic solution (Cellgro)

50ml Fetal Calf Serum (HyClone)

Assembled under sterile conditions, stored at 4°C.

To select for stably transfected HEK293 cells that express the neomycin resistance gene encoded on the pCI-neo vector, 500μg/ml G418 (from 100mg/ml in 40mM Hepes frozen stock) was added to the medium immediately prior to use. Stable transfectants were subsequently maintained in 300μg/ml G418.

10.1.2 LB Medium

10g Tryptone (BD)

5g Yeast extract (BD)

10g NaCl

15g Bacto Agar (for plates only) (BD)

ddH_2O to 950ml

pH to 7.0 with ~200μl 5N NaOH, ddH_2O to 1l and autoclaved for 20 min., 15 psi, 121°C

10.1.3 LB Ampicillin or Kanamycin

1000x Ampicillin stock solutions were prepared at 50mg/ml in ddH_2O and stored in 1ml aliquots at -20°C. 1000x Kanamycin stock solutions were prepared at 30mg/ml in

ddH_2O and stored in 1ml aliquots at -20°C. Appropriate volumes of Ampicillin and Kanamycin were added fresh to either LB broth or melted LB agar (T<70°C) before use.

10.1.4 SOB Medium

20g Tryptone (BD)

5g Yeast extract (BD)

500mg NaCl

ddH_2O to 950ml

10ml 250mM KCl

pH to 7.0 with ~200µl 5N NaOH

ddH_2O to 1l and autoclaved for 20 minutes at 15 psi, 121°C

10.1.5 SOC Medium

Assembled under sterile conditions:

10ml sterile SOB medium

200µl sterile 1M glucose

100µl 1M $MgCl_2$

10.1.6 40% Glucose

40g Glucose

ddH_2O to 100ml and sterile filtered

10.1.7 YPDA Medium

20g/l Difco peptone (BD)

10g/l Yeast extract (BD)

20g/l Bacto agar (for plates only) (BD)

15ml/l 0.2% adenine hemisulfate

ddH_2O to 950ml and autoclaved for 20 minutes at 15 psi, 121°C

50ml sterile 40% glucose solution added under sterile conditions

10.1.8 Synthetic Dextrose –Trp-Leu dropout medium (SD-Trp-Leu)

26.7g Minimal SD Base (Clontech)

640mg -Leu/-Trp DO Supplement (Clontech)

20g/l Agar (for plates only) (BD)

ddH_2O to 950ml and autoclaved for 20 minutes at 15 psi, 121°C

50ml sterile 40% glucose solution added under sterile conditions

10.2 Solutions yeast two-hybrid

10.2.1 10x TE buffer

100mM Tris-HCl

10mM EDTA

pH adjusted to 7.5 and autoclaved for 20 minutes at 15 psi, 121°C

10.2.2 10x LiAc

1M LiAc

pH adjusted to 7.5 with dilute acetic acid and autoclaved for 20 minutes at 15 psi, 121°C

10.2.3 X-gal

20mg/ml in Dimethylformamide, stored at -20°C in the dark

10.2.4 50% PEG

Prepared with sterile ddH_2O, warmed to 50°C for 30 minutes to dissolve PEG.

10.2.5 Z buffer

16.1g/l	$Na_2HPO_4 \cdot 7H_2O$
5.5g/l	$NaH_2PO_4 \cdot H_2O$
750mg/l	KCl
246mg/l	$MgSO_4 \cdot 7H_2O$

pH adjusted to 7.0 and autoclaved for 20 minutes at 15 psi, 121°C

10.2.6 Z buffer/X-gal solution

100ml	Z buffer
270µl	β-mercaptoethanol
1.67ml	X-gal stock solution

10.2.7 Z buffer with β-mercaptoethanol

100ml	Z buffer
270µl	β-mercaptoethanol

10.2.8 ONPG (o-nitrophenyl β-D-galactopyranoside)

ONPG was prepared fresh by dissolving 4mg/ml in Z buffer 1-2h before use, pH was adjusted to 7.0

10.3 Primer sequences

Length = length of primer in number of nucleotides

% GC = GC content of primer sequence

Tm [°C] = melting temperature, based on nearest neighbor

Tm F. [°C] = melting temperature, according to Finnzyme Tm calculator (for use with Phire polymerase)

Nucleotides in bold and italic = restriction enzyme recognition sites

Nucleotides underlined = mutagenesis (site-directed nucleotide exchanges or insertions)

10.3.1 SNP Screening

Rank	Name	SNP	Primer name	Primer sequence	Length [nt]	% GC	Tm [°C]
2, 4, 7	Insulin-like Growth Factor Binding Protein 7	rs11555284, rs1133243, rs11555293	IGF1D	GGCATGGAGTGCGTGAAGAGC	21	62	59
			IGF2U	CGATGACCTCACAGCTCAAGTAC	23	52	57
			gIGFU	CAGGTGCCCTTGCTGACCTG	20	65	57
			IGF7D	GTGCGGCTGCTGCCCTATGTGC	22	68	63
5	Complement component 1, q, subcomponent-like 1	rs11538450	C1Q13U	TCCGCGTCCTGGGCAATA	19	74	73
			C1Q14D	CCTGAGCGAGCAGAGCGGC	18	61	70
			gC1QU	ACGTCGTCAAACTTGAGTACCTCG	24	50	57
10	Component of Oligomeric Golgi Complex 1, OGC1	rs11544801	OGC1D	CGCTGAAGCGGCTGGATCTGC	21	67	63
			OGC2U	TTTCTAAGAGTAGCTTGATCTGG	23	39	53
			gOGCU	CGCATCTGGCCGATGGTGTCG	21	67	61
			OGC4D	GAGACGCATGGAGCGGAGGAGAT	23	61	62
			OGC3U	CTGAGAGGCTTCCATCGAGCTCCAGATC	28	57	65
12	RARS – arginyl-t-RNA synthetase	rs1059443	RA1D	TCATGCAATTAAGGCTGCATATC	23	39	54
			RA2U	GCCTACATGCATCTCTTTAGC	21	48	53
13	S-adenosylhomocysteine hydrolase	rs11552694	SA1D	GCCTGCCATCAATGTCAATGAC	22	50	54
			SA2U	GATGATGTCAATACAGCCTGTG	22	45	52
29	Zinc Finger Protein 289, ID1 regulated	rs11542792	ID1D	CGTCCATCTGAGCTTCATCAG	21	52	55
			ID2U	TGTCTATCCAAAGATCAGTGCCATG	25	44	57
57	C1Q-like Tumor necrosis factor 5	rs11538245	CQT1D	TGCCAGCGCTATGAGGCCACTC	22	64	62
			CQT2U	TCTCGGAGCGCTTGGCGCTGAAGG	24	67	67

10.3.1 REDS Screening

Rank	Name	mRNA with discrepancy	Primer name	Primer sequence	Length [nt]	% GC	Tm [°C]
n/a	Sorting Nexin 1 Isoform C	AK222793	SNX1D	AGAGCTAGCGCTGAACACAGC	21	57	60
			SNX2U	CATATTGAGTCACCCGAGACTC	22	50	53
			SNX3D	TTGCAAAGAGTCTAGCCATGC	21	48	55
			SNX3D-Kpn	GAGTA***GGTACC***TTGCAAAGAGTCTAGCCATGC	21	48	55
			SNX2U-Eco	G***GAATTC***ATATTGAGTCACCCGAGACTC	22	50	53
			gSNX1D	ccagAGCTAGCGCTGAACACAGC	23	61	63
			gSNX2U	CAGAAGCTTACGCGGACTATGG	22	55	58
n/a	Frizzled 7	AB010881	FZD1D	CGTCTTCAGCGTGCTCTACAC	21	57	57
			FZD3D	CTTCAGCGTGCTCTACACAGTG	22	55	58
			FZD2U	GCTGCTGTGGCTAAGTCTGTGG	22	59	59
			FZD4U	GCTGTGGCTAAGTCTGTGGTAG	22	55	56
n/a	ATPase, H+ transporting, V0 subunit	AK094602	V01D	CGGAGTGATCATCACCATGCTG	22	55	57
			V02U	ATGGCGATGAGCCAGCTTGGATGC	24	58	63
			V03D	CCTGGTTCGTGCCGAAGGGAC	21	67	58
			V04U	GTACCAGATGGTCTCATTCTTCAGCTG	27	48	57
			gV05D	TAAGAGGAGGTGATTCTGACGTGCTG	26	50	60
			gV06U	GACAGCAAAGCATGGTGAGCCTG	23	57	60
n/a	Arylsulfatase E (chondrodysplasia punctata 1)	BC130438	ARSE1D	CTGCCAGCGATGCTCGCTGTACTG	24	63	65
			gARSEU	AAACGCAGAAGTGCAGAACTC	21	48	55
			ARSE2U	GTCAGCTTCACGCCGTCCTCTGC	23	65	63
n/a	Pregnancy Specific β1-Glycoprotein 9	BC005925	PS1D	CTCGACTTGTCCTGCTTCACG	21	57	57
			gPSD	GTGAAAGCAAGCCTAGTTCTCTGAG	25	48	59
			PS2U	CAGTGTCTCAGTTGTTGCAGTC	22	50	54

n/a	Anoctamin-1 (Transmembrane protein 16A)	BC033036	TMP1D	CAAAGACATCGGAATCTGGTAC	22	45	52
			gTMPD	CCACATGACTGAGAGTGTAG	20	50	50
			gTMPU	GCTGCACCATGGCTAGAGGC	20	65	58
n/a	Ribosomal Protein L17	BC066324	L1D	AAAAACGCAG AGAGTAATGC TG	22	41	54
			L2U	CGTAAGGATCATCTCAATGTGGC	23	48	55
n/a	Upstream binding protein 1	BC047235	U1D	AGCAGCAAGCTGCAAGCAGTGC	22	59	63
			U2U	TTATGTGGATGCCATCACTAC	21	43	50
1	Serine Peptidase Inhibitor, Clade A, member 3	BC034554	SPI1D	ATGATGAGTTTGCATCACCTGAC	23	43	54
			SPI2U	GTATGTCGTTCAGGTTATAGTC	22	41	49
			gSPIU	GACATTGGTGAGACCTTGACGTG	23	52	56
4	Unknown Protein	BC027448	UNK1D	GCTTCTGCTTGGTGGTAGATTACTTGC	27	48	60
			UNK2U	GAGAGCAGTTCTGTCTTCTGAGGCC	25	56	61
			gUNK3D	CAAGCTAAGACTTTCCTCTGCTGCTG	26	50	62
12	Homo sapiens PC4 and SFRS1 interacting protein 1, isoform 2	BC040032	PSIP1D	CAGCAATGAGGATGTGACTAAAGCAGT	27	44	60
			PSIP2U	GCTGTTGCTGTTGTCACTACTCCTGC	26	54	61
13	SUMO1/sentrin/SMT3 specific peptidase 2	AK074357	SPP1D	GGCTGTAATAGAAGACCAGGTGGCC	25	56	59
			SPP2U	GAAGTTACAGGACAGACAGAGTTTCCATG	29	45	59
17&18	Chromosome 19 Open Reading Frame 21	BX648389	HY2P1D	GAGCACAAAGCAAGAGGCATCGAAG	25	52	62
			HY2P2U	GAGTAACTGCCCGTGATGCCGGATG	25	60	62
20	Chromosome 12 Open Reading Frame 43	AK225162	HYP1D	TCACATCTGTCCCTGGAGGCCGTG	24	63	62
			HYP2U	AGCTCACCTGACTTCTGCTTCTGGAC	26	54	62
21	Exophilin 5	CR627226	EXPH1D	GTTCCTCCGGCGTTTGATTTCAGTTTC	27	48	60
			EXPH2U	AACTCATTCTCCAGCAGTGAAGCATC	26	46	60
22	Tetraspan NET-5 variant	AK223179	TSP1D	TGTTGGTCATCCTCCTAGCAGAGCTG	26	54	62
			TSP2U	TCGTACTTCTTACCAGTCCGGTGGATG	27	52	60

29	Pantothenate Kinase 2	AF494409	PANK2-1D	CATGGTTTGGACTGGATATCGGTGGAAC	28	50	60
			PANK2-2U	CTAAGATGCTAACCCCTGAGCCAATG	26	50	60
34	Breast Cancer Nuclear Receptor-Binding Auxiliary Protein	AF126008	AKAP1D	GAGAGCAGAGACCTTTGGAGGATTTGAC	28	50	61
			AKAP2U	TGCGCTGCTTCTCCTGCTCAATG	23	57	63
40	General Transcription Factor II	BC099907	GTF21D	ACGGCGATTAAGGAGAGCACCTC	23	57	60
			GTF22U	CTTCCAACTGGCTTGGTTCAGCTG	24	54	60

10.3.2 C1QL1

Primer name	**Primer sequence**	**Length [nt]**	**% GC**	**Tm [°C]**	**Tm F. [°C]**	**Specific Purpose**
C1Q1D	CACCTGCCGCATGGTGTGCGAC	22	68	62		
C1Q2U	CGCGTCCTGGGCAATAGCACTG	22	64	60	74	
C1Q3D	GCCACTATGAGATGCTG	17	53	47		
gC1QU	ACGTCGTCAAACTTGAGTACCTCG	24	50	57	67.5	Amplification of genomic sequence
C1Q4U	GCTGTTGCCGCCGTGTGCTTTGC	23	65	62	79	
C1Q5D	GAAGGCCACTATGAGATGCTG	21	52	54	64	
C1Q6D	GAAGGCCACTATGAGATGCTGGGCACCTGC	30	60	67		
mC1Q7D	GCTGGTGGTGCTCATCCCGGTGCTG	25	68	65		*Mus musculus* specific primer, forward
mC1Q8U	GTTGCTGGCATAGTCGTAGTTCTGGTC	27	52	61		*Mus musculus* specific primer, reverse
mgC1Q9U	TCGTAGTTGTTGCCTAGGTTGGTGAC	26	50	59		*Mus musculus* specific primer, for genomic sequence
mC1Q7D-Eco	G***GAATTC***GCTGGTGGTGCTCATCCCGGTGCTG	25	68	65		Subcloning of *Mus musculus* amplicon, forward
mC1Q8U-Kpn	CTAGA***GGTACC***GTTGCTGGCATAGTCGTAGTTCTGGTC	27	52	61		Subcloning of *Mus musculus* amplicon, reverse
C1Q10D	GACGCCCTGAGCGAGCAGAGC	21	71	63	75	
C1Q11U	AGCTGTGGCCCAGGCGGTGTTGAG	24	67	64	79	
C1Q12D	ATGCTGCTGGTGCTGGTGGTGCTCATC	26	59	66	79	
C1Q13U	TCCGCGTCCTGGGCAATA	19	74		73	*Homo sapiens* specific primer, forward
C1Q14D	CCTGAGCGAGCAGAGCGGC	18	61		70	*Homo sapiens* specific primer, reverse
drC1Q15D	GTGTGCGATCCATACCAGAACAAG	24	50		69	*Danio rerio* specific primer, forward

drC1Q16U	GTCTAGATGCAGAATCACACTGTTCGA	27	44		68.5	*Danio rerio* specific primer, reverse
mdC1Q17D	ACGGTGAACCACTGAGTGAGCAC	23	57		70	*Monodelphis Domesticus* specific primer, forward
mdC1Q18U	GATGTAGACCTCGTCCCCAGCA	22	59		70	*Monodelphis Domesticus* specific primer, reverse
xtC1Q19D	CGTGCAGGAACTGGAGCTGAAC	22	59		70.6	*Xenopus tropicalis* specific primer, forward
xtC1Q20U	TCTGGTCAGCGTCTTGAGCAATG	23	52		71	*Xenopus tropicalis* specific primer, reverse
C1Q21D-mut	CCTTCCACGCTGGTGC*G*GGGCCCCCAGGGGAAG	33	76			Site directed mutagenesis of codon 66 from CAG (Q) to CGG (R) (forward)
C1Q22U-mut	CTTCCCCTGGGGGCCCC*C*GCACCAGCGTGGAAGG	33	76			

10.3.3 IGFBP7

Primer name	Primer sequence	Length [nt]	% GC	Tm [°C]	Specific Purpose
IGF1D	GGCATGGAGTGCGTGAAGAGC	21	62	59	
IGF2U	CGATGACCTCACAGCTCAAGTAC	23	52	57	
gIGFU	CAGGTGCCCTTGCTGACCTG	20	65	57	
IGF3D	GTGGCGGCGCCGGCGGGGGGTACTGC GCGCC	31	87	76	Site directed mutagenesis of codon 78 from AGG (R) to GGG (G) (forward)
IGF4U	GGCGCGCAGTACCCCCCGCCGGCGCC GCCAC	31	87	76	Site directed mutagenesis of codon 78 from AGG (R) to GGG (G) (reverse)
IGF5D	GAAGAGCCGCAAGAGGCGGAGGGGT AAAGCCGGGGCAGCAG	41	68	77	Site directed mutagenesis of codon 95 from AAG (K) to AGG (R) (forward)
IGF6U	CTGCTGCCCCGGCTTTACCCCTCCGCC TCTTGCGGCTCTTC	41	68	77	Site directed mutagenesis of codon 95 from AAG (K) to AGG (R) (reverse)
IGF7D	GTGCGGCTGCTGCCCTATGTGC	22	68	63	
IGF8D-Eco	G***GAATTC***GTGCGGCTGCTGCCCTATGT GC	22	68	63	Subcloning into pBluescript vector
IGF9U-Kpn	CTAGA***GGTACC***CGATGACCTCACAGC TCAAGTAC	23	52	57	Subcloning into pBluescript vector
IGF20D-Xho1	CTAGA***CTCGAG***CCCGCCATGGAGCGG CCGTC	20	80	73	Cloning of full-length IGFBP7 ORF into pCI-neo vector
IGF21U-HA-XbaI	CTGC***TCTAGA***TTAAGCGTAATCTGGTA CGTCGTATGGGTATAGCTCGGCACCTT CACC	48	50	72	Cloning of full-length IGFBP7 ORF into pCI-neo vector, addition of HA sequence to C-terminus (underlined)

10.3.4 SerpinA3

Primer name	Primer sequence	Length [nt]	% GC	Tm [°C]	Tm F. [°C]	Specific Purpose
SPI1D	ATGATGAGTTTGCATCACCTGAC	23	43	54	65	
SPI2U	GTATGTCGTTCAGGTTATAGTC	22	41	49	55	
gSPIU	GACATTGGTGAGACCTTGACGTG	23	52	56	68	
SPI3D-Nhe1	CATA***GCTAGC***ATGATGAGTTTGCATCACCTGAC	23	43	54		
SPI4U-EcoR1	TA***GAATTC***AGAGAGTCTCTCCACCGCTTCAG	23	57	60		
SPI5D-EcoR1	TA***GAATTC***ggcttccacttcctccatcttg	22	55	56		
SPI5D-Sac	TCC***CCGCGG***GGCTTCCACTTCCTCCATCTTG	22	55		68	Clone construct SPI_05
SPI6U-Kpn1	CATA***GGTACC***gctcctgcaccgtggtggagctg	24	71	66		
SPI7D	CACTC***ATGCAT***GGTTTCTCCTGGTCTC	27	52	61	73	Nsi1 site bold and italic
SPI8U	GAGACCAGGAGAAACC***ATGCAT***GAGTG	27	52	61	73	Nsi1 site bold and italic
SPI9U	CACTGCATTCTAGTTGTGGTTTGTC	25	44	56	66	pCI vector primer
SPI10D	GTCCACTCCCAGTTCAATTACAGCTC	26	50	58	69	pCI vector primer
SPI11D	GCAGAGCTCGTTTAGTGAACCGTCAG	26	54	62	72	pCI vector primer
SPI12U	ACCGCTTCAGGGTCTCTGGGAGCAG	25	64	64	77	
SPI13D	TGGCAGAGGGAACAGTCAGTGCAAG	25	56	62	74	
SPI14U-Xma1	CATA***CCCGGG***CCTCAATGCCCAGCTGGAGAAG TATG	26	54	61	73	
SPI15mutD	CTGCTCCCAGAGACCCTGA***A<u>C</u>CGGT***GGAGAGA CTCTCTGG	40	62.5			
SPI16mutU	CCAGAGAGTCTCTCC***ACCGG<u>T</u>***TCAGGGTCTCTG GGAGCAG	40	62.5			
SPI17D-AgeI	CAT***ACCGGT***TCCTGACCCCTGTTCTAGGGGAG	23	61	58		

SPI18U-Pvu2	CATA***CAGCTG***CTGAGGAGAAATGGGACAGCCTTTG	25	52	59		
SPI19Dmut R9	AATTCGGCTTCCACTTCCTCCATCTTGTCTTGATCAGGGAG	41	49			Correct mutation in small construct
SPI20Umut R9	CTCCCTGATCAAGACAAGATGGAGGAAGTGGAAGCCGAATT	41	49			Correct mutation in small construct
SPI21D	CTGCTCCCAGAGACCCTGAAGCGGT	25	64	64		Complement to SPI12U
SPI22D	CACAAGCCACTGTTATATGTTCACATC	27	41	56		
SPI23D	GTCGTGAGGCACTGGGCAGGTGTC	24	67	62		Across exon/exon boundary of vector intron
SPI24U	CTACGTGGCCCTACCTAAGCTGCTC	25	60	62	70.5	
SPI25D-Kpn1	CATA***GGTACC***GAGATAGGGACACAAAGAGTGATCAC	26	46	56		
SPI26U-Xba1	GC***TCTAGA***TCCATTCTTGGTTTTAGGAGATGCTTGG	28	43	59		
SPI26U-Sal1	CTACGC***GTCGAC***TCCATTCTTGGTTTTAGGAGATGCTTGG	28	43	59		
SPI27D-XbaI	GC***TCTAGA***GTAGAGGAAACCAGGGAAGACCATG	25	52	57	68.7	
SPI27D-Sal1	GAATCA***GTCGAC***GTAGAGGAAACCAGGGAAGACCATG	25	52	57		
SPI28D-mutAgeB	CTCCACGTCAAGGTCTCACCGGTGTCCAGGGACAAGGGCATG	42	62			
SPI29U-mutAgeB	CATGCCCTTGTCCCTGGACACCGGTGAGACCTTGACGTGGAG	42	62			
SPI30U-PvuII	CATA***CAGCTG***GAACATATAACAGTGGCTTGTGGGAGG	27	48	58		
SPI31U-Xba1	GC***TCTAGA***CACCTGTTTACCACCTGCCTGTTTC	25	52	58		

SPI31U-Sal1	CTACGC***GTCGAC***CACCTGTTTACCACCTGCCTGTTTC	25	52	58		
SPI32D-Xba1	CG***TCTAGA***GGACACACGGAATTGAGCTGACCG	24	58	60	74.5	
SPI32D-Sal1	GAATAG***GTCGAC***GGACACACGGAATTGAGCTGACCG	24	58	60	75	
SPI33U-Kpn1	CATA***GGTACC***GAGTTGCTCCACGGGGCGGTG	21	71	61	77	
SPI34D-Kpn1	CTAT***GGTACC***CTAGAAAGCCCAAAGGTGGGGAAG	24	54	58	70	
SPI35D-mut5'(2)	GAGAGACTCTCTGGAGTTCA**TA**TGATTCTTCCTGGCCCCCAA	42	50			Site-directed mutagenesis 5' splice site, forward
SPI36U-mut5'(2)	TTGGGGGCCAGGAAGAATCA**TA**TGAACTCCAGAGAGTCTCTC	42	50			Site-directed mutagenesis 5' splice site, reverse
SPI37D-mut3'	CTTTTTTTTTCTCGTTTTCTA**A**AGAGATAGGTGAGCTCTACCTG	44	36			Site-directed mutagenesis 3' splice site, forward
SPI38U-mut3'	CAGGTAGAGCTCACCTATCTCT**T**TAGAAAACGAGAAAAAAAAAG	44	36			Site-directed mutagenesis 3' splice site, reverse
SPI39D-mut Pvu-Kpn	GCCCAGGGAGTGCATGTGGG**TA**C**C**TGTGTGGGGTAGAGAGGAA	43	60			
SPI40U-mut Pvu-Kpn	TTCCTCTCTACCCCACACAG**G**T**A**CCCACATGCACTCCCTGGGC	43	60			
SPI43D-mut Pvu-Sac	GCCACTGTTATATGTTCC**C**GC**G**GGGCTTCCACTTCCTCCA	40	57.5			
SPI44U-mut Pvu-Sac	TGGAGGAAGTGGAAGCCC**C**GC**G**GGAACATATAACAGTGG	40	57.5			
SPI47D-mut5'(3)	GTGGAGAGACTCTCTGGA**C**TTCA**TAC**GATTCTTCCTGGCCCCCAA	45	53			Site-directed mutagenesis 5' splice site, 2nd round, forward

SPI48U-mut5'(3)	TTGGGGGCCAGGAAGAATCGTATGAAGTCCAGAGAGTCTCTCCAC	45	53			Site-directed mutagenesis 5' splice site, 2nd round, reverse
SPI49D-mut3'(3)	CTTTTTTTTTCTCGTTTTCTAAACATATACGTGAGCTCTACCTGCCAAAG	50	36			Site-directed mutagenesis 3' splice site, 2nd round, forward
SPI50U-mut3'(3)	CTTTGGCAGGTAGAGCTCACGTATATGTTTAGAAAACGAGAAAAAAAAAG	50	36			Site-directed mutagenesis 3' splice site, 2nd round, reverse
SPI51D-mutSRp40.1	GAAGTACCTGGGCAATGCCAGCGCACTCTTCATCCTC	Tm complementary = 45˚C Tm non-overlapping = 56˚C				Site-directed mutagenesis putative SRp40.1 site, forward
SPI52U-mutSRp40.1	CATTGCCCAGGTACTTCAGCTCCACCACGGTGCAG	Tm complementary = 45˚C Tm non-overlapping = 55˚C				Site-directed mutagenesis putative SRp40.1 site, reverse
SPI53D-mutSRp40.2	*GATCAAGCTGACATGGAGGAAGTGGAAGCCATGCTGCTC*	Tm complementary = 47˚C Tm non-overlapping = 57˚C				Site-directed mutagenesis putative SRp40.2 site, forward
SPI54U-mutSRp40.2	*CTCCATGTCAGCTTGATCAGGGAGGATGAAGAGTGCGCTG*	Tm complementary = 47˚C Tm non-overlapping = 60˚C				Site-directed mutagenesis putative SRp40.2 site, reverse
SPI55D-mutSRp40.2-ECS	*GATCAAGCTGACATGGAGGAAGTGGAAGCCcagctgctc*	Tm complementary = 47˚C Tm non-overlapping = 58˚C				Site-directed mutagenesis ECS for complementarity to mutated putative SRp40.2 site
SPI56D-mutcorUM3	GAAGAGTCAGGCCAGAGCAGCTGGGCTTCCACTTCCTCCATG					Site-directed mutagenesis to correct mutation in SRp40.2 mutant, forward

SPI57U-mutcorUM3	CATGGAGGAAGTGGAAGCCCAGCTGCTCTGGCCTGACTCTTC					Site-directed mutagenesis to correct mutation in SRp40.2 mutant, reverse
SPI58D	CACCGCCCCGTGGAGCAACTC	21	71		77	
SPI59D	CTTGACTCGCAGACAATGATGGTCC	25	52	59	72	
SPI60U	GGTCTTTGGGGGCCAGGAAGAATC	24	58	58	74	
SPI61U	GAGCTCACCTATCTCTCTGAACTCCAGAG	29	52	63	69.5	
SPI_ECS_mut1D	GAGCAGCTGGGCTTCCACTTCCCCCATCTTGTGTTGATCAGGGAG	45				Site-directed mutagenesis of ECS #1, forward
SPI_ECS_mut1U	CTCCCTGATCAACACAAGATGGGGGAAGTGGAAGCCCAGCTGCTC	45				Site-directed mutagenesis of ECS #1, reverse
SPI_ECS_mut2D	CTGTGTACTTCAGCTCCCACACGGTGGAGGAGCAGAGAGAGTGAG	45				Site-directed mutagenesis of ECS #2, forward
SPI_ECS_mut2U	CTCACTCTCTCTGCTCCTCCACCGTGTGGGAGCTGAAGTACACAG	45				Site-directed mutagenesis of ECS #2, reverse
SPI_ECS_mut3D	ATGAAGAGTGCGCTGGGACTGCCTGTGCACTTCAGCTCCCACACG	45				Site-directed mutagenesis of ECS #3, forward
SPI_ECS_mut3U	CGTGTGGGAGCTGAAGTGCACAGGCAGTCCCAGCGCACTCTTCAT	45				Site-directed mutagenesis of ECS #3, reverse
SPI_ECS_mut4D	CCCATCTTGTGTTGACCAGGGAGGCTGCAGAGTGCGCTGGGAC	43				Site-directed mutagenesis of ECS #4, forward
SPI_ECS_mut4U	GTCCCAGCGCACTCTGCAGCCTCCCTGGTCAACACAAGATGGG	43				Site-directed mutagenesis of ECS #4, reverse
SPImutA-G1D	GCACCGTGGTGGAGCTGAGGTGCACGGGCAATGCCAGCGCACTC	44				Site-directed mutagenesis 1 of adenosines to guanosines in target sequence, forward
SPImutA-G1U	GAGTGCGCTGGCATTGCCCGTGCACCTCAGCTCCACCACGGTGC	44				Site-directed mutagenesis 1 of adenosines to guanosines in target sequence, reverse

SPImutA-G2D	GTGGAGCTGAGGTGCACGGGCGGTGCCGGCGCGCTCTTCATCCTC	45				Site-directed mutagenesis 2 of adenosines to guanosines in target sequence, forward
SPImutA-G2U	GAGGATGAAGAGCGCGCCGGCACCGCCCGTGCACCTCAGCTCCAC	45				Site-directed mutagenesis 2 of adenosines to guanosines in target sequence, reverse
SPImutA-G3D	CATCCTCCCTGATCGGGGCAGGATGGAGGAGGTGGAAGCCATG	43				Site-directed mutagenesis 3 of adenosines to guanosines in target sequence, forward
SPImutA-G3U	CATGGCTTCCACCTCCTCCATCCTGCCCCGATCAGGGAGGATG	43				Site-directed mutagenesis 3 of adenosines to guanosines in target sequence, reverse

10.3.5 FLNA

Name	Primer name	Primer sequence	Length [nt]	% GC	Tm [°C]	Specific Purpose
FLNA	FLN-3D-Sfi	CAT***GGCC***ATGGA***GGCC***CAGGATATGACA GCCCAGGTGACCAGC	27	59	73	
	FLN-4U-Xho	GTACA***CTCGAG***ACTCA GGGCACCACAACGCGGTAG	24	63	62	
	FLN8D	CTGGAGAGAGCTGAAGCTGGAGTGC	25	60	64	
	FLN9U	CAGATACTGAATTCGGCTGGCACTC	25	52	60	
	FLN12D-Xho	CAGT***CTCGAG***CCAGGATATGACAGCCCAG GTGAC	24	58	59	Amplification from pGADT7 for cloning into pCeMM-NTAP(GS)
	FLN13U-Not	CTATCT***GCGGCCGC***TCAGGGCACCACAAC GCGGTAG	22	64	60	
	FLN14D-Pme	AGCAGA***GTTTAAAC***CAGGATATGACAGCC CAGGTGAC	23	57	57	Amplification from pGADT7 for cloning into pCeMM-CTAP(SG)
	FLN15U-FseI	CATA***GGCCGGCC***AGGGCACCACAACGCG GTAG	20	65	58	
Presenilin-1	PS11D-Bam	CATA***GGATCC***GTTGTCCGAAAGGTCCACT TCG	22	55	56	
	PS11U-Pst	CATA***CTGCAG***AAACAGGCTATGGTTGTGT TCCAGTC	26	46	57	
Integrin-β	Inb11D-Bam	CATA***GGATCC***GTAAGCTTTTAATGATAAT TCATGAC	26	27	50	
	Inb11U-Pst	CATA***CTGCAG***CATTGCTACTTTGCATTCAG TGTTGTG	27	41	58	
Metabotropic glutamate receptor 5a	GRM5-1D-Bam(2)	GC***GGATCC***TGGCCAAACCAGAGAGAAAC GTGC				
	GRM5-3U-Pst(2)	AA***CTGCAG***GCTGGGCCAGTCTCCTGTCTT TGTAC	25	56	59	

Metabotropic glutamate receptor 7b	GRM7b-3D-Bam	CTAAC***GGATCC***ACTGTATACCACCAGTAAGAAAGAG	25	40	54	
	GRM7b-4U-Pst	AATAT***CTGCAG***ATACTGTTGGTGGGATAGTGTACCAAG	27			
Metabotropic glutamate receptor 8a	GRM8-3D-Bam(2)	GC***GGATCC***GCTTCAAGGCTGTGGTGACAGCTG	24	58	61	
	GRM8-4D-Bam(2)	GC***GGATCC***CCAACACTTCCTCTACCAAGACAACATA	28	43	58	
	GRM8-5U-Pst	AA***CTGCAG***ATTGTGCCATTTCCCTGTTTCAGATTG	27	41	58	
p73alpha	P73A-5D-Nde (2)	GGAATTC***CATATG***GAGATGAGCAGCAGCCACAGC	20	60	59 (70)	
	P73A-6U-Eco	CAG***GAATTC***TCAGTGGATCTCGGCCTCCGTG	22	64	60 (74)	
P2Y2	P2Y2-1D Bam	CTAAC***GGATCC***TCTACTTCCTGGCTGGGCAG				
	P2Y2-2UPst	AATAT***CTGCAG***CCTACAGCCGAATGTCCTTAGTG				
	P2Y2-3D-Bam	CTAAC***GGATCC***CCCGAGATGCCAAGCCACCCACT	23	65	62	
	P2Y2-4UPst	AATAT***CTGCAG***GCCTAGAGTCCTCACTGCTGCC	22	64	60	
Calcitonin Receptor	CTR-2U-Bam	CTAAC***GGATCC***CGATGCTGTGTTTGCTTCACATTCAAG	27	44	61	
	CTR-3D-Eco(2)	G***GAATTC***ATCTACTGCTTCTGCAACAATGAGGTCC	27	48	61	
Breast Cancer Associated 2	BRCA2-1D-Bam	AA***CTGCAG***GCTCCCGTGGCTGGTAAATCTGA	24	50	59	
	BRCA2-2U-Pst (2)	GC***GGATCC***TGGAGGCCCAACAAAAGAGACTAGAAG	23	57	59	

SEK	SEK1-1D-Nde(3)	GACAACTC***CATATG***CAGGGTAAACGCAAAGCACTGAAG				No amplification
	SEK1-2U-Bam	CTAAC***GGATCC***ATCAATCGACATACATGGGAGAGCTG	26	46	59	
	SEK1-3D	CTTCAGAGAGGGTGACTGTTGGATC	25	52	58	
	SEK1-4U	CACTGATGCCGAAGTCACAGAGC	23	57	59	
Smad5	SMD5-2U-Bam	CTACA***GGATCC***TGTCACTGAGGCATTCCGCATACAC	25	52	60	Vector has insertion before start ATG
	SMD5-3D-Nde	GGTATAC***CATATG***TACAGGAAGGTCTCCGAAGATTTGTGTC	28	46	59	

11 References

Adachi Y, Itoh F, Yamamoto H et al. 2001. Expression of angiomodulin (tumor-derived adhesion factor/mac25) in invading tumor cells correlates with poor prognosis in human colorectal cancer. *Int J Cancer 95*:216-222.

Ahmed S, Jin X, Yagi M et al. 2006. Identification of membrane-bound serine proteinase matriptase as processing enzyme of insulin-like growth factor binding protein-related protein-1 (IGFBP-rP1/angiomodulin/mac25). *Febs J 273*:615-627.

Ahmed S, Yamamoto K, Sato Y et al. 2003. Proteolytic processing of IGFBP-related protein-1 (TAF/angiomodulin/mac25) modulates its biological activity. *Biochem Biophys Res Commun 310*:612-618.

Akaogi K, Okabe Y, Funahashi K et al. 1994. Cell adhesion activity of a 30-kDa major secreted protein from human bladder carcinoma cells. *Biochem Biophys Res Commun 198*:1046-1053.

Akaogi K, Okabe Y, Sato J et al. 1996a. Specific accumulation of tumor-derived adhesion factor in tumor blood vessels and in capillary tube-like structures of cultured vascular endothelial cells. *Proc Natl Acad Sci U S A 93*:8384-8389.

Akaogi K, Sato J, Okabe Y et al. 1996b. Synergistic growth stimulation of mouse fibroblasts by tumor-derived adhesion factor with insulin-like growth factors and insulin. *Cell Growth Differ 7*:1671-1677.

Angov E. 2011. Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnol J 6*:650-659.

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol 2*:e391.

Baker C, Belbin O, Kalsheker N, Morgan K. 2007. SERPINA3 (aka alpha-1-antichymotrypsin). *Front Biosci 12*:2821-2835.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell 116*:281-297.

Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA 11*:241-247.

Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem 71*:817-846.

Bass BL, Weintraub H. 1987. A developmentally regulated activity that unwinds RNA duplexes. *Cell 48*:607-613.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet 3*:370-379.

Beck R, Ravet M, Wieland FT, Cassel D. 2009. The COPI system: molecular mechanisms and function. *FEBS Lett 583*:2701-2709.

Berg KA, Cropper JD, Niswender CM et al. 2001. RNA-editing of the 5-HT(2C) receptor alters agonist-receptor-effector coupling specificity. *Br J Pharmacol 134*:386-392.

Berube NG, Swanson XH, Bertram MJ et al. 1999. Cloning and characterization of CRF, a novel C1q-related factor, expressed in areas of the brain involved in motor function. *Brain Res Mol Brain Res 63*:233-240.

Bhalla T, Rosenthal JJ, Holmgren M, Reenan R. 2004. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat Struct Mol Biol 11*:950-956.

Blow M, Futreal PA, Wooster R, Stratton MR. 2004. A survey of RNA editing in human brain. *Genome Res 14*:2379-2387.

Blow MJ, Grocock RJ, van Dongen S et al. 2006. RNA editing of human microRNAs. *Genome Biol 7*:R27.

Buetow KH, Edmonson MN, Cassidy AB. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet 21*:323-325.

Bundschuh R. 2004. Computational prediction of RNA editing sites. *Bioinformatics 20*:3214-3220.

Burckstummer T, Bennett KL, Preradovic A et al. 2006. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat Methods 3*:1013-1019.

Burns CM, Chu H, Rueter SM et al. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature 387*:303-308.

Cannarozzi G, Schraudolph NN, Faty M et al. 2010. A role for codon order in translation dynamics. *Cell 141*:355-367.

Cartegni L, Wang J, Zhu Z et al. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res 31*:3568-3571.

Chasin LA. 2007. Searching for splicing motifs. *Adv Exp Med Biol 623*:85-106.

Chen CX, Cho DS, Wang Q et al. 2000. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA 6*:755-767.

Chen LL, Carmichael GG. 2008. Gene regulation by SINES and inosines: biological consequences of A-to-I editing of Alu element inverted repeats. *Cell Cycle 7*:3294-3301.

Chen LL, DeCerbo JN, Carmichael GG. 2008. Alu element-mediated gene silencing. *EMBO J 27*:1694-1705.

Chen Y, Pacyna-Gengelbach M, Ye F et al. 2007. Insulin-like growth factor binding protein-related protein 1 (IGFBP-rP1) has potential tumour-suppressive activity in human lung cancer. *J Pathol 211*:431-438.

Cho DS, Yang W, Lee JT et al. 2003. Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. *J Biol Chem 278*:17093-17102.

Clutterbuck DR, Leroy A, O'Connell MA, Semple CA. 2005. A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics 21*:2590-2595.

Collet C, Candy J. 1998. How many insulin-like growth factor binding proteins? *Mol Cell Endocrinol 139*:1-6.

Connolly CM, Dearth AT, Braun RE. 2005. Disruption of murine Tenr results in teratospermia and male infertility. *Dev Biol 278*:13-21.

Costa FF. 2010. Non-coding RNAs: Meet thy masters. *Bioessays 32*:599-608.

Crowther DC, Belorgey D, Miranda E et al. 2004. Practical genetics: alpha-1-antitrypsin deficiency and the serpinopathies. *Eur J Hum Genet 12*:167-172.

David SS, O'Shea VL, Kundu S. 2007. Base-excision repair of oxidative DNA damage. *Nature 447*:941-950.

Davies MJ, Lomas DA. 2008. The molecular aetiology of the serpinopathies. *Int J Biochem Cell Biol 40*:1273-1286.

Dawson TR, Sansam CL, Emeson RB. 2004. Structure and sequence determinants required for the RNA editing of ADAR2 substrates. *J Biol Chem 279*:4941-4951.

de la Pena M, Garcia-Robles I. 2010. Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep 11*:711-716.

Desmet FO, Hamroun D, Lalande M et al. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res 37*:e67.

Desterro JM, Keegan LP, Jaffray E et al. 2005. SUMO-1 modification alters ADAR1 editing activity. *Mol Biol Cell 16*:5115-5126.

Desterro JM, Keegan LP, Lafarga M et al. 2003. Dynamic association of RNA-editing enzymes with the nucleolus. *J Cell Sci 116*:1805-1818.

Devlin GL, Bottomley SP. 2005. A protein family under 'stress' - serpin stability, folding and misfolding. *Front Biosci 10*:288-299.

Djuranovic S, Nahvi A, Green R. 2011. A parsimonious model for gene regulation by miRNAs. *Science 331*:550-553.

Eckmann CR, Neunteufl A, Pfaffstetter L, Jantsch MF. 2001. The human but not the Xenopus RNA-editing enzyme ADAR1 has an atypical nuclear localization signal and displays the characteristics of a shuttling protein. *Mol Biol Cell 12*:1911-1924.

Eggington JM, Greene T, Bass BL. 2011. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun 2*:319.

Eisenberg E, Adamsky K, Cohen L et al. 2005a. Identification of RNA editing sites in the SNP database. *Nucleic Acids Res 33*:4612-4617.

Eisenberg E, Nemzer S, Kinar Y et al. 2005b. Is abundant A-to-I RNA editing primate-specific? *Trends Genet 21*:77-81.

Enstero M, Akerborg O, Lundin D et al. A computational screen for site selective A-to-I editing detects novel sites in neuron specific Hu proteins. *BMC Bioinformatics 11*:6.

Enstero M, Akerborg O, Lundin D et al. 2010. A computational screen for site selective A-to-I editing detects novel sites in neuron specific Hu proteins. *BMC Bioinformatics 11*:6.

Enstero M, Daniel C, Wahlstedt H et al. 2009. Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA. *Nucleic Acids Res 37*:6916-6926.

Even Y, Durieux S, Escande ML et al. 2006. CDC2L5, a Cdk-like kinase with RS domain, interacts with the ASF/SF2-associated protein p32 and affects splicing in vivo. *J Cell Biochem 99*:890-904.

Feng Y, Walsh CA. 2004. The many faces of filamin: a versatile molecular scaffold for cell motility and signalling. *Nat Cell Biol 6*:1034-1038.

Forster AC, Symons RH. 1987. Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell 50*:9-16.

Fredrick K, Ibba M. 2010. How the sequence of a gene can tune its translation. *Cell 141*:227-229.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res 19*:92-105.

Fritz J, Strehblow A, Taschner A et al. 2009. RNA-regulated interaction of transportin-1 and exportin-5 with the double-stranded RNA-binding domain regulates nucleocytoplasmic shuttling of ADAR1. *Mol Cell Biol 29*:1487-1497.

Gallo A, Keegan LP, Ring GM, O'Connell MA. 2003. An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. *EMBO J 22*:3421-3430.

Gan Z, Zhao L, Yang L et al. 2006. RNA editing by ADAR2 is metabolically regulated in pancreatic islets and beta-cells. *J Biol Chem 281*:33386-33394.

George CX, Das S, Samuel CE. 2008. Organization of the mouse RNA-specific adenosine deaminase Adar1 gene 5'-region and demonstration of STAT1-independent, STAT2-dependent transcriptional activation by interferon. *Virology 380*:338-343.

George CX, Samuel CE. 1999a. Characterization of the 5'-flanking region of the human RNA-specific adenosine deaminase ADAR1 gene and identification of an interferon-inducible ADAR1 promoter. *Gene 229*:203-213.

George CX, Samuel CE. 1999b. Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. *Proc Natl Acad Sci U S A 96*:4621-4626.

Gerber A, O'Connell MA, Keller W. 1997. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *RNA 3*:453-463.

Ghai R, Waters P, Roumenina LT et al. 2007. C1q and its growing family. *Immunobiology 212*:253-266.

Glanzer JG, Enose Y, Wang T et al. 2007. Genomic and proteomic microglial profiling: pathways for neuroprotective inflammatory responses following nerve fragment clearance and activation. *J Neurochem 102*:627-645.

Godfried Sie C, Kuchka M. 2011. RNA Editing adds flavor to complexity. *Biochemistry (Moscow) 76*:869-881.

Gommans WM, Dylan E. Dupuis, Jill E. McCane, Nicholas E. Tatalias, Stefan Maas. 2008. Diversifying Exon Code through A-to-I RNA Editing. *RNA and DNA Editing*: Wiley InterScience. pp 1-30.

Gommans WM, Mullen SP, Maas S. 2009. RNA editing: a driving force for adaptive evolution? *Bioessays 31*:1137-1145.

Gommans WM, Tatalias NE, Sie CP et al. 2008. Screening of human SNP database identifies recoding sites of A-to-I RNA editing. *RNA 14*:2074-2085.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res 36*:D154-158.

Guerrini R, Carrozzo R. 2001. Epileptogenic brain malformations: clinical presentation, malformative patterns and indications for genetic testing. *Seizure 10*:532-543; quiz 544-537.

Hartner JC, Schmittwolf C, Kispert A et al. 2004. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J Biol Chem 279*:4894-4902.

Hartner JC, Walkley CR, Lu J, Orkin SH. 2009. ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. *Nat Immunol 10*:109-115.

Hasler J, Samuelsson T, Strub K. 2007. Useful 'junk': Alu RNAs in the human transcriptome. *Cell Mol Life Sci 64*:1793-1800.

Hasler J, Strub K. 2006. Alu elements as regulators of gene expression. *Nucleic Acids Res 34*:5491-5497.

He T, Wang Q, Feng G et al. 2011. Computational detection and functional analysis of human tissue-specific A-to-I RNA editing. *PLoS One 6*:e18129.

Heale BS, Keegan LP, McGurk L et al. 2009. Editing independent effects of ADARs on the miRNA/siRNA pathways. *EMBO J 28*:3145-3156.

Herb A, Higuchi M, Sprengel R, Seeburg PH. 1996. Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. *Proc Natl Acad Sci U S A 93*:1875-1880.

Herrick-Davis K, Grinde E, Niswender CM. 1999. Serotonin 5-HT2C receptor RNA editing alters receptor basal activity: implications for serotonergic signal transduction. *J Neurochem 73*:1711-1717.

Higuchi M, Maas S, Single FN et al. 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature 406*:78-81.

Hoopengardner B, Bhalla T, Staber C, Reenan R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science 301*:832-836.

Hundley HA, Bass BL. 2010. ADAR editing in double-stranded UTRs and other noncoding RNA sequences. *Trends Biochem Sci 35*:377-383.

Hundley HA, Krauchuk AA, Bass BL. 2008. C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. *RNA 14*:2050-2060.

Hung T, Chang HY. 2010. Long noncoding RNA in genome regulation: Prospects and mechanisms. *RNA Biol 7*.

Hunsberger JG, Bennett AH, Selvanayagam E et al. 2005. Gene profiling the response to kainic acid induced seizures. *Brain Res Mol Brain Res 141*:95-112.

Iijima T, Miura E, Watanabe M, Yuzaki M. 2010. Distinct expression of C1q-like family mRNAs in mouse brain and biochemical characterization of their encoded proteins. *Eur J Neurosci 31*:1606-1615.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature 431*:931-945.

Irizarry K, Kustanovich V, Li C et al. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat Genet 26*:233-236.

Jacobs MM, Fogg RL, Emeson RB, Stanwood GD. 2009. ADAR1 and ADAR2 expression and editing activity during forebrain development. *Dev Neurosci 31*:223-237.

Jeanmougin F, Thompson JD, Gouy M et al. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem Sci 23*:403-405.

Jin Y, Zhang W, Li Q. 2009. Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life 61*:572-578.

Kalus W, Zweckstetter M, Renner C et al. 1998. Structure of the IGF-binding domain of the insulin-like growth factor-binding protein-5 (IGFBP-5): implications for IGF and IGF-I receptor interactions. *EMBO J 17*:6558-6572.

Kapranov P, Cawley SE, Drenkow J et al. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science 296*:916-919.

Kapranov P, Cheng J, Dike S et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science 316*:1484-1488.

Kawahara Y, Megraw M, Kreider E et al. 2008. Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res 36*:5270-5280.

Kawahara Y, Zinshteyn B, Chendrimada TP et al. 2007a. RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep 8*:763-769.

Kawahara Y, Zinshteyn B, Sethupathy P et al. 2007b. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science 315*:1137-1140.

Kent WJ, Sugnet CW, Furey TS et al. 2002. The human genome browser at UCSC. *Genome Res 12*:996-1006.

Kim DD, Kim TT, Walsh T et al. 2004. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res 14*:1719-1725.

Kim U, Nishikura K. 1993. Double-stranded RNA adenosine deaminase as a potential mammalian RNA editing factor. *Semin Cell Biol 4*:285-293.

Kimchi-Sarfaty C, Oh JM, Kim IW et al. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science 315*:525-528.

Kohler M, Burnashev N, Sakmann B, Seeburg PH. 1993. Determinants of Ca2+ permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. *Neuron 10*:491-500.

Kuhn RM, Karolchik D, Zweig AS et al. 2007. The UCSC genome browser database: update 2007. *Nucleic Acids Res 35*:D668-673.

Lad Y, Jiang P, Ruskamo S et al. 2008. Structural basis of the migfilin-filamin interaction and competition with integrin beta tails. *J Biol Chem 283*:35154-35163.

Lad Y, Kiema T, Jiang P et al. 2007. Structure of three tandem filamin domains reveals auto-inhibition of ligand binding. *EMBO J 26*:3993-4004.

Lai F, Drakas R, Nishikura K. 1995. Mutagenic analysis of double-stranded RNA adenosine deaminase, a candidate enzyme for RNA editing of glutamate-gated ion channel transcripts. *J Biol Chem 270*:17098-17105.

Lander ES, Linton LM, Birren B et al. 2001. Initial sequencing and analysis of the human genome. *Nature 409*:860-921.

Laurencikiene J, Kallman AM, Fong N et al. 2006. RNA editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO Rep 7*:303-307.

Lehmann KA, Bass BL. 1999. The importance of internal loops within RNA substrates of ADAR1. *J Mol Biol 291*:1-13.

Lehmann KA, Bass BL. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry 39*:12875-12884.

Lev-Maor G, Sorek R, Levanon EY et al. 2007. RNA-editing-mediated exon evolution. *Genome Biol 8*:R29.

Levanon EY, Eisenberg E, Yelin R et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol 22*:1001-1005.

Levanon EY, Hallegger M, Kinar Y et al. 2005. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res 33*:1162-1168.

Li JB, Levanon EY, Yoon JK et al. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science 324*:1210-1213.

Lin J, Lai M, Huang Q et al. 2007. Methylation patterns of IGFBP7 in colon cancer cell lines are associated with levels of gene expression. *J Pathol 212*:83-90.

Liu H, Naismith JH. 2008. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol 8*:91.

Lomas DA. 2005. Molecular mousetraps, alpha1-antitrypsin deficiency and the serpinopathies. *Clin Med 5*:249-257.

Lomeli H, Mosbacher J, Melcher T et al. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science 266*:1709-1713.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J 417*:15-27.

Lopez-Bermejo A, Khosravi J, Fernandez-Real JM et al. 2006. Insulin resistance is associated with increased serum concentration of IGF-binding protein-related protein 1 (IGFBP-rP1/MAC25). *Diabetes 55*:2333-2339.

Luciano DJ, Mirsky H, Vendetti NJ, Maas S. 2004. RNA editing of a miRNA precursor. *RNA 10*:1174-1177.

Maas S, Godfried Sie CP, Stoev I et al. 2011. Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites. *Biochem Biophys Res Commun*.

Maas S, Gommans WM. 2009. Novel exon of mammalian ADAR2 extends open reading frame. *PLoS ONE 4*:e4225.

Macbeth MR, Lingam AT, Bass BL. 2004. Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. *RNA 10*:1563-1571.

Marion S, Weiner DM, Caron MG. 2004. RNA editing induces variation in desensitization and trafficking of 5-hydroxytryptamine 2c receptor isoforms. *J Biol Chem 279*:2945-2954.

Mattick JS. 2004. RNA regulation: a new genetics? *Nat Rev Genet 5*:316-323.

Mattick JS. 2010. RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays 32*:548-552.

Mattick JS, Mehler MF. 2008. RNA editing, DNA recoding and the evolution of human cognition. *Trends Neurosci 31*:227-233.

Melcher T, Maas S, Herb A et al. 1996. A mammalian RNA editing enzyme. *Nature 379*:460-464.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet 10*:155-159.

Mian IS, Moser MJ, Holley WR, Chatterjee A. 1998. Statistical modelling and phylogenetic analysis of a deaminase domain. *J Comput Biol 5*:57-72.

Miura F, Kawaguchi N, Sese J et al. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A 103*:17846-17851.

Miyoshi K, Miyoshi T, Siomi H. 2010. Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. *Mol Genet Genomics 284*:95-103.

Moller-Krull M, Zemann A, Roos C et al. 2008. Beyond DNA: RNA editing and steps toward Alu exonization in primates. *J Mol Biol 382*:601-609.

Morse DP. 2004. Identification of substrates for adenosine deaminases that act on RNA. *Methods Mol Biol 265*:199-218.

Morse DP, Aruscavage PJ, Bass BL. 2002. RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. *Proc Natl Acad Sci U S A 99*:7906-7911.

Morse DP, Bass BL. 1999. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proc Natl Acad Sci U S A 96*:6048-6053.

Murphy M, Pykett MJ, Harnish P et al. 1993. Identification and characterization of genes differentially expressed in meningiomas. *Cell Growth Differ 4*:715-722.

Mutaguchi K, Yasumoto H, Mita K et al. 2003. Restoration of insulin-like growth factor binding protein-related protein 1 has a tumor-suppressive activity through induction of apoptosis in human prostate cancer. *Cancer Res 63*:7717-7723.

Nagakubo D, Murai T, Tanaka T et al. 2003. A high endothelial venule secretory protein, mac25/angiomodulin, interacts with multiple high endothelial venule-associated molecules including chemokines. *J Immunol 171*:553-561.

Nakamura F, Osborn TM, Hartemink CA et al. 2007. Structural basis of filamin A functions. *J Cell Biol 179*:1011-1025.

Neeley WL, Essigmann JM. 2006. Mechanisms of formation, genotoxicity, and mutation of guanine oxidation products. *Chem Res Toxicol 19*:491-505.

Nickel W, Brugger B, Wieland FT. 2002. Vesicular transport: the core machinery of COPI recruitment and budding. *J Cell Sci 115*:3235-3240.

Nielsen KB, Sorensen S, Cartegni L 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*:416-432.

Nimmich ML, Heidelberg LS, Fisher JL. 2009. RNA editing of the GABA(A) receptor alpha3 subunit alters the functional properties of recombinant receptors. *Neurosci Res 63*:288-293.

O'Connell MA, Krause S, Higuchi M et al. 1995. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Mol Cell Biol 15*:1389-1397.

Oh Y, Nagalla SR, Yamanaka Y et al. 1996. Synthesis and characterization of insulin-like growth factor-binding protein (IGFBP)-7. Recombinant human mac25 protein specifically binds IGF-I and -II. *J Biol Chem 271*:30322-30325.

Ohlson J, Enstero M, Sjoberg BM, Ohman M. 2005. A method to find tissue-specific novel sites of selective adenosine deamination. *Nucleic Acids Res 33*:e167.

Ohlson J, Pedersen JS, Haussler D, Ohman M. 2007. Editing modifies the GABA(A) receptor subunit alpha3. *RNA 13*:698-703.

Ohta Y, Hartwig JH, Stossel TP. 2006. FilGAP, a Rho- and ROCK-regulated GAP for Rac binds filamin A to control actin remodelling. *Nat Cell Biol 8*:803-814.

Ohta Y, Suzuki N, Nakamura S et al. 1999. The small GTPase RalA targets filamin to induce filopodia. *Proc Natl Acad Sci U S A 96*:2122-2128.

Osenberg S, Paz Yaacov N, Safran M et al. 2010. Alu sequences in undifferentiated human embryonic stem cells display high levels of A-to-I RNA editing. *PLoS One 5*:e11173.

Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. 2000. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell 102*:437-449.

Paschen W, Djuricic B. 1995. Regional differences in the extent of RNA editing of the glutamate receptor subunits GluR2 and GluR6 in rat brain. *J Neurosci Methods 56*:21-29.

Paul MS, Bass BL. 1998. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J 17*:1120-1127.

Pen A, Moreno MJ, Durocher Y et al. 2008. Glioblastoma-secreted factors induce IGFBP7 and angiogenesis by modulating Smad-2-dependent TGF-beta signaling. *Oncogene 27*:6834-6844.

Pokharel S, Beal PA. 2006. High-throughput screening for functional adenosine to inosine RNA editing systems. *ACS Chem Biol 1*:761-765.

Polson AG, Bass BL. 1994. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J 13*:5701-5711.

Popowicz GM, Schleicher M, Noegel AA, Holak TA. 2006. Filamins: promiscuous organizers of the cytoskeleton. *Trends Biochem Sci 31*:411-419.

Poulsen H, Nilsson J, Damgaard CK et al. 2001. CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain. *Mol Cell Biol 21*:7862-7871.

Prasanth KV, Prasanth SG, Xuan Z et al. 2005. Regulating gene expression through RNA nuclear retention. *Cell 123*:249-263.

Pullirsch D, Jantsch MF. 2010. Proteome diversification by adenosine to inosine RNA editing. *RNA Biol 7*:205-212.

Robertson SP. 2005. Filamin A: phenotypic diversity. *Curr Opin Genet Dev 15*:301-307.

Robertson SP, Twigg SR, Sutherland-Smith AJ et al. 2003. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. *Nat Genet 33*:487-491.

Rodriguez MS, Desterro JM, Lain S et al. 1999. SUMO-1 modification activates the transcriptional response of p53. *EMBO J 18*:6455-6461.

Ruan W, Xu E, Xu F et al. 2007. IGFBP7 plays a potential tumor suppressor role in colorectal carcinogenesis. *Cancer Biol Ther 6*:354-359.

Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature 399*:75-80.

Rula EY, Lagrange AH, Jacobs MM et al. 2008. Developmental modulation of GABA(A) receptor function by RNA editing. *J Neurosci 28*:6196-6201.

Ryman K, Fong N, Bratt E et al. 2007. The C-terminal domain of RNA Pol II helps ensure that editing precedes splicing of the GluR-B transcript. *RNA 13*:1071-1078.

Sampson DA, Wang M, Matunis MJ. 2001. The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. *J Biol Chem 276*:21664-21669.

Samuel CE. 2011. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology*.

Sarkisian MR, Bartley CM, Rakic P. 2008. Trouble making the first move: interpreting arrested neuronal migration in the cerebral cortex. *Trends Neurosci 31*:54-61.

Sasahara K, Yamaoka T, Moritani M et al. 2000. Molecular cloning and expression analysis of a putative nuclear protein, SR-25. *Biochem Biophys Res Commun 269*:444-450.

Sato J, Hasegawa S, Akaogi K et al. 1999. Identification of cell-binding site of angiomodulin (AGM/TAF/Mac25) that interacts with heparan sulfates on cell surface. *J Cell Biochem 75*:187-195.

Saunders LR, Barber GN. 2003. The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J 17*:961-983.

Scadden AD. 2005. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat Struct Mol Biol 12*:489-496.

Scadden AD. 2007. Inosine-containing dsRNA binds a stress-granule-like complex and downregulates gene expression in trans. *Mol Cell 28*:491-500.

Scadden AD, Smith CW. 1997. A ribonuclease specific for inosine-containing RNA: a potential role in antiviral defence? *EMBO J 16*:2140-2149.

Scadden AD, Smith CW. 2001. Specific cleavage of hyper-edited dsRNAs. *EMBO J 20*:4243-4252.

Schulte JH, Marschall T, Martin M et al. 2010. Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res 38*:5919-5928.

Seeburg PH. 2002. A-to-I editing: new and old sites, functions and speculations. *Neuron 35*:17-20.

Shah SP, Morin RD, Khattra J et al. 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature 461*:809-813.

Sheen VL, Feng Y, Graham D et al. 2002. Filamin A and Filamin B are co-expressed within neurons during periods of neuronal migration and can physically interact. *Hum Mol Genet 11*:2845-2854.

Sheen VL, Jansen A, Chen MH et al. 2005. Filamin A mutations cause periventricular heterotopia with Ehlers-Danlos syndrome. *Neurology 64*:254-262.

Shepard PJ, Hertel KJ. 2009. The SR protein family. *Genome Biol 10*:242.

Sherry ST, Ward MH, Kholodov M et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res 29*:308-311.

Shi X, Kachirskaia I, Yamaguchi H et al. 2007. Modulation of p53 function by SET8-mediated methylation at lysine 382. *Mol Cell 27*:636-646.

Sie CP, Maas S. 2009. Conserved recoding RNA editing of vertebrate C1q-related factor C1QL1. *FEBS Lett*.

Siepel A, Bejerano G, Pedersen JS et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res 15*:1034-1050.

Slezak-Prochazka I, Durmus S, Kroesen BJ, van den Berg A. 2010. MicroRNAs, macrocontrol: regulation of miRNA processing. *RNA 16*:1087-1095.

Smirnov DA, Foulk BW, Doyle GV et al. 2006. Global gene expression profiling of circulating endothelial cells in patients with metastatic carcinomas. *Cancer research 66*:2918-2922.

Smith PJ, Zhang C, Wang J et al. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet 15*:2490-2508.

Sommer B, Kohler M, Sprengel R, Seeburg PH. 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell 67*:11-19.

Sprenger CC, Vail ME, Evans K et al. 2002. Over-expression of insulin-like growth factor binding protein-related protein-1(IGFBP-rP1/mac25) in the M12 prostate cancer cell line alters tumor growth by a delay in G1 and cyclin A associated apoptosis. *Oncogene 21*:140-147.

St Laurent G, 3rd, Wahlestedt C. 2007. Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci 30*:612-621.

Stefl R, Oberstrass FC, Hood JL et al. 2010. The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. *Cell 143*:225-237.

Stein LD, Bao Z, Blasiar D et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol 1*:E45.

Stephens OM, Haudenschild BL, Beal PA. 2004. The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. *Chem Biol 11*:1239-1250.

Strehblow A, Hallegger M, Jantsch MF. 2002. Nucleocytoplasmic distribution of human RNA-editing enzyme ADAR1 is modulated by double-stranded RNA-binding domains, a leucine-rich export signal, and a putative dimerization domain. *Mol Biol Cell 13*:3822-3835.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays 29*:288-299.

Takeuchi T, Harris JL, Huang W et al. 2000. Cellular localization of membrane-type serine protease 1 and identification of protease-activated receptor-2 and single-chain urokinase-type plasminogen activator as substrates. *J Biol Chem 275*:26333-26342.

Tamura K, Hashimoto K, Suzuki K et al. 2009. Insulin-like growth factor binding protein-7 (IGFBP7) blocks vascular endothelial cell growth factor (VEGF)-induced angiogenesis in human vascular endothelial cells. *Eur J Pharmacol 610*:61-67.

Tamura K, Matsushita M, Endo A et al. 2007. Effect of Insulin-Like Growth Factor-Binding Protein 7 on Steroidogenesis in Granulosa Cells Derived from Equine Chorionic Gonadotropin-Primed Immature Rat Ovaries. *Biol Reprod*.

Tonkin LA, Saccomanno L, Morse DP et al. 2002. RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. *EMBO J 21*:6025-6035.

Tsai MC, Manor O, Wan Y et al. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science 329*:689-693.

van Beijnum JR, Dings RP, van der Linden E et al. 2006. Gene expression of tumor angiogenesis dissected: specific targeting of colon cancer angiogenic vasculature. *Blood 108*:2339-2348.

Wahlstedt H, Daniel C, Enstero M, Ohman M. 2009. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res 19*:978-986.

Wajapeyee N, Serra RW, Zhu X et al. 2008. Oncogenic BRAF induces senescence and apoptosis through pathways mediated by the secreted protein IGFBP7. *Cell 132*:363-374.

Waki H, Yamauchi T, Kamon J et al. 2005. Generation of globular fragment of adiponectin by leukocyte elastase secreted by monocytic cell line THP-1. *Endocrinology 146*:790-796.

Wang Q, Khillan J, Gadue P, Nishikura K. 2000. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science 290*:1765-1768.

Will CL, Luhrmann R. 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem 386*:713-724.

Wilson HM, Birnbaum RS, Poot M et al. 2002. Insulin-like growth factor binding protein-related protein 1 inhibits proliferation of MCF-7 breast cancer cells via a senescence-like mechanism. *Cell Growth Differ 13*:205-213.

Wong SK, Sato S, Lazinski DW. 2001. Substrate recognition by ADAR1 and ADAR2. *RNA 7*:846-858.

XuFeng R, Boyer MJ, Shen H et al. 2009. ADAR1 is required for hematopoietic progenitor cell survival via RNA editing. *Proc Natl Acad Sci U S A 106*:17763-17768.

Yamanaka Y, Wilson EM, Rosenfeld RG, Oh Y. 1997. Inhibition of insulin receptor activation by insulin-like growth factor binding proteins. *J Biol Chem 272*:30729-30734.

Yamauchi T, Umeda F, Masakado M et al. 1994. Purification and molecular cloning of prostacyclin-stimulating factor from serum-free conditioned medium of human diploid fibroblast cells. *Biochem J 303 (Pt 2)*:591-598.

Yang J, Valineva T, Hong J et al. 2007. Transcriptional co-activator protein p100 interacts with snRNP proteins and facilitates the assembly of the spliceosome. *Nucleic Acids Res 35*:4485-4494.

Yang W, Chendrimada TP, Wang Q et al. 2006. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol 13*:13-21.

Ye S, Goldsmith EJ. 2001. Serpins and other covalent protease inhibitors. *Curr Opin Struct Biol 11*:740-745.

Yeo J, Goodman RA, Schirle NT et al. 2010. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. *Proc Natl Acad Sci U S A 107*:20715-20719.

Yi-Brunozzi HY, Stephens OM, Beal PA. 2001. Conformational changes that occur during an RNA-editing adenosine deamination reaction. *J Biol Chem 276*:37827-37833.

Yuzaki M. 2008. Cbln and C1q family proteins: new transneuronal cytokines. *Cell Mol Life Sci 65*:1698-1705.

Zeslawski W, Beisel HG, Kamionka M et al. 2001. The interaction of insulin-like growth factor-I with the N-terminal domain of IGFBP-5. *EMBO J 20*:3638-3644.

Zhang Z, Carmichael GG. 2001. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell 106*:465-475.

Zhu TN, He HJ, Kole S et al. 2007. Filamin A-mediated down-regulation of the exchange factor Ras-GRF1 correlates with decreased matrix metalloproteinase-9 expression in human melanoma cells. *J Biol Chem 282*:14816-14826.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res 31*:3406-3415.

12 Vita

Overview of Qualifications

- Experienced in a broad range of laboratory research techniques:
  - Cellular biology (cell culture of HeLa, HEK293, primary chondrocytes; histocytochemistry; transfections) and tissue engineering (human and ovine articular cartilage)
  - Molecular biology (PCR, cloning, mutagenesis)
  - Bioinformatics analysis (UCSC Genome browser, BLAST, Mfold)
  - Protein biochemistry (expression in cell culture, purification, SDS PAGE, Western blotting, *in vitro* proteolysis, *in vitro* transcription/translation, radioblotting)
- Managed high-risk project in start-up biotech company
- Demonstrated effective prioritization of several research projects while efficiently managing numerous other responsibilities in research, teaching, and services
- Accustomed to be accountable for the planning, preparation and completion of projects with a budget
- Experienced working in interdisciplinary, international, and collaborative settings

Education

Ph.D., Molecular Biology, **Lehigh University**, Bethlehem, PA, USA, defended September 30th 2011, accepted October 28th 2011, expected graduation January 2012

M.Sc., Biological Sciences, **Swiss Federal Institute of Technology Zurich (ETH)**, Zurich, Switzerland, 2002

Employment and Work Experience

Research Assistant, **Lehigh University**, Bethlehem, PA, USA 2007 – 2011

- Planned and supervised six undergraduate research projects over the past four years

- Mentored undergraduates to develop hypotheses, plan experiments, and present research

Teaching Assistant, **Lehigh University**, Bethlehem, PA, USA 2007 – 2011

- Research lab course (Science Education Alliance program, HHMI)
- Graduate student research fellow to BDSI, an interdisciplinary research initiative by HHMI
- Recitation courses (Genetics, Cellular and Molecular Biology)

Management Assistant, **AIM International AG**, Lucerne, Switzerland 2006

- Administrative work, travel and meeting arrangements, implementation of filing system

Project Manager (R&D), **Millenium Biologix AG**, Schlieren, Switzerland 2003 –2005

- Tissue engineering of articular cartilage for the generation of implants for joint defects
- Planned and managed pilot animal trial (sheep)
- Part of interdisciplinary/international team to develop an automated cell culture system (ACTES)

Intern, **Veterinary Research Institute (VUVEL)**, Brno, Czech Republic 2002 –2003

- Identification of Mycobacteria in animals with Tuberculosis and Crohn's patients (multiplex PCR)

Tutorial Assistant, **Swiss Federal Institute of Technology**, Switzerland '01 –'02

- Part-time assistant at the Institutes of Plant Physiology and Microbiology

Publications

Godfried Sie, C.P., Kuchka, M. *RNA editing adds flavor to complexity*. Biochemistry (Moscow) 2011; 76(8): 869-881.

Maas, S., **Godfried Sie, C.P.**, Loev, I., et al. *Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing* sites. Biochemical Biophysical Research Communications 2011; 412(3): 407-12.

Sie, C.P. and Maas, S. *Conserved recoding RNA editing of vertebrate C1q-related factor C1QL1*. FEBS Letters 2009; 583: 1171-1174.

Gommans, W.M., Tatalias, N.E., **Sie, C.P.** et al. *Screening of human SNP database identifies recoding sites of A-to-I RNA editing.* RNA 2008; 14(10): 1-12.

Skibbens, R.V., **Sie, C.P.**, Eastman, L. *Role of chromosome segregation genes in BRCA1-dependent lethality.* Cell Cycle 2008; 7(13): 2071-2.

Francioli, S.-E., Martin, I., **Sie, C.P.**, et al. *Growth Factors for Clinical-Scale Expansion of Human Articular Chondrocytes: Relevance for Automated Bioreactor Systems*. Tissue Eng. 2007; 13(6): 1227-1234.

Invited Talks and Seminars

Godfried Sie, C.P. *Understanding the lack of A-to-I RNA editing in a highly predicted ADAR target*.

Department of Biological Sciences Seminar, Lehigh University, PA, April 2011

Godfried Sie, C.P., Kuchka, M. and Maas, S. *Anatomy of a perfect ADAR target that is not edited* in vivo.

Gordon Research Conference on RNA Editing, Galveston, TX, January 2011

Godfried Sie, C.P. *Identification and Characterization of A-to-I RNA Editing Targets*.

Department of Biological Sciences Seminar, Lehigh University, PA, March 2010

Sie, C.P. *The Cellular Cytoskeleton*.

Guest Speaker "Cellular and Molecular Biology, Core I", Lehigh University, PA, March 2008

Conferences and Meetings

Gordon Research Conference on RNA Editing, Galveston, TX — Jan. 2011

Editing and Modification of RNA and DNA. Invited talk and poster presentation.

American Society of Cell Biology 50th Annual Meeting, Philadelphia, PA — Dec.2010

RNA Biology, Philadelphia RNA Club, Philadelphia, PA — Mar. 2009

Gordon Research Conference on RNA Editing, Galveston, TX — Jan. 2009

Roles of RNA and DNA editing and Modification in Cellular Function. Poster presentation.

The Biology of Small RNA, Newark, DE — Apr. 2008

American Society of Cell Biology 46th Annual Meeting, San Diego, CA — Dec. 2006

Bioimaging and Engineered Biosystems, Bethlehem, PA — Sept. 2006

Regenerative Medicine Showcase, Ann Arbor, MI — May 2005

Strategies in Tissue Engineering, Würzburg, Germany — June 2004

Awards and Funding

Fellowships and Funding:

- Sigma Delta Epsilon-Graduate Women in Science (GWIS) fellowship — 2011
- Sigma Xi Grants in Aid of Research (GIAR) — 2011
- Nemes Fellowship, Department of Biological Sciences, Lehigh University — 2009 and 2011
- Lehigh University Forum Student Research Grant — 2010
- CAS Graduate Research Fund, Lehigh University — 2010
- Datatel Scholars Foundation Scholarship — 2007/08 and 2009/10
- Biosystems Dynamics Summer Institute Research Fellowship — 2008
- Medicus Scholarship, Swiss Benevolent Society NY — 2007-2008
- Lehigh University Fellowship award — 2006-2007

Awards:

- Graduate Student Merit Award, Alumni Association of Lehigh University — 2011

Other Skills and Interests

Computer skills:

Microsoft Office, basic level Photoshop, EndNote

Leadership Programs and Workshops:

Iacocca Institute Workshops, Iacocca Institute, Bethlehem PA 2009/2010

Teacher Development Program, Lehigh University, Bethlehem PA 2009/2010

Competence in Project Management, Zurich, Switzerland 2005

Languages:

Fluent in German and English, conversation level French and Spanish, basic level Italian

Services

Student Services Committee to the graduate student body of Lehigh University, Bethlehem, PA, 2011-2012

Graduate Liaison Officer to the Graduate Student Committee at the Department of Biological Sciences, Lehigh University, Bethlehem, PA, 2011

Representative of the Graduate Students of Biological Sciences at the Graduate Student Senate, Lehigh University, Bethlehem, PA, 2010

Founder of the Lehigh Valley German Speaking Group (weekly meetings, special events), June 2009

PJAS Judge at the Pennsylvania Junior Academy of Science regional competition in Easton, PA, 2009/2010/2011

Volunteer for the Lehigh River and Stream Clean-up, Bethlehem, PA, spring 2009

Participant of the Relay for Life Fundraiser, Bethlehem, PA, 2008