The multi-faceted RNA molecule: Characterization and Function in the Regulation of Gene Expression

Mats Ensterö

The multi-faceted RNA molecule: Characterization and Function in the Regulation of Gene Expression

Mats Ensterö

Stockholm University

ISBN 978-91-7155-587-8

Printed in Sweden by Universitetsservice US-AB,
Stockholm 2008
Distributor: Stockholm University Library

To my sun, Eliah

Abstract

In this thesis I have studied the RNA molecule and its function and characteristics in the regulation of gene expression. I have focused on two events that are important for regulation of the transcriptome: Translational regulation through micro RNAs; and RNA editing through adenosine deaminations.

Micro RNAs (miRNAs) are ~22 nucleotides long RNA molecules that by semi complementarity bind to untranslated regions of a target messenger RNA (mRNA). The interaction manifests through an RNA/protein complex and act mainly by repressing translation of the target mRNA. I have shown that a pre-cursor miRNA molecule have significantly different information content of sequential composition of the two arms of the pre-cursor hairpin. I have also shown that sequential composition differs between species.

Selective adenosine to inosine (A-to-I) RNA editing is a co-transcriptional process whereby highly specific adenosines in a (pre-)messenger transcript are deaminated to inosines. The deamination is carried out by the ADAR family of proteins and require a specific sequential and structural landscape for target recognition. Only a handful messenger substrates have been found to be site selectively edited in mammals. Still, most of these editing events have an impact on neurotransmission in the brain.

In order to find novel substrates for A-to-I editing, an experimental setup was made to extract RNA targets of the ADAR2 enzyme. In concert with this experimental approach, I have constructed a computational screen to predict specific positions prone for A-to-I editing.

Further, I have analyzed editing in the mouse brain at four different developmental stages by 454 amplicon sequencing™. With high resolution, data is presented supporting a general developmental regulation of A-to-I editing. I also show that data of editing events are coupled on single RNA transcripts, suggesting an A-to-I editing mechanism that involve ADAR dimers to act in concert.

List of papers included in this thesis

The thesis is based on the following articles, which will be referred to by their Roman numerals in the text.

I. Ohlson J, **Ensterö M**, Sjöberg BM, Öhman M. 2005. A method to find tissue-specific novel sites of selective adenosine deamination. *Nucleic Acids Res 33*:e167.

II. Gorodkin J, Havgaard JH, **Ensterö M**, Sawera M, Jensen P, Öhman M, Fredholm M. 2006. MicroRNA sequence motifs reveal asymmetry between the stem arms. *Comput Biol Chem 30*:249-254.

III. **Ensterö M**, Åkerborg Ö, Lundin D, Wang B, Furey T.S, Öhman. M Lagergren J. 2008. A computational screen for site selective A-to-I editing.
Manuscript.

IV. **Ensterö M,** Daniel C, Wahlstedt H, Öhman M. 2008. An in-depth survey of A- to-I editing implies a general developmental regulation and coupling of edited sites.
Manuscript.

Contents

Abbreviations

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
A	Adenosine
C	Cytosine
G	Guanosine
T	Thymin
U	Uridine
GluR	Glutamate receptor
5-HT_{2C}	Serotonin receptor 2C
ds	Double stranded
GABA	gamma-aminobutyric acid
nt	nucleotide(s)
bp	base pair(s)
ADAR	Adenosine deaminase acting on RNA
dsRBM	Double stranded RNA binding motif

Introduction

The multi-faceted RNA molecule: Characterization and function in the regulation of gene expression.

RNA molecules was early in the history of molecular biology firmly introduced in the central dogma as the messenger molecule between the coding DNA and the interpreted protein. Deviants from this dogma was the transfer RNA and ribosomal RNA (tRNA and rRNA, respectively) that are actively involved in the protein synthesis. That RNA has more divergent tasks in the cellular machinery became obvious with the discoveries of the catalytic RNAs of self-splicing group 1 introns and RNase P (Kruger et al., 1982) (Guerrier-Takada et al., 1983), respectively. In the last 15 years the concept of functional non-coding RNA has grown in its significance not only in an increasing number of different species of RNA but also the impact of the regulating capacities they possess. Hence, recent years have exposed numerous RNAs with other capabilities than a temporal information carrier mediating the DNA code for peptide synthesis. RNA has been shown to function both as an essential catalytic macromolecule as well as a regulatory molecule addressing sequence specific interactions that affect gene expression (Nissen et al., 2000) (Lee et al., 2001) (Kishore et al., 2006).

In this thesis I will address two types of regulatory events where RNA plays a major role.

In a number of family members, one of most the prominent example of non coding RNA is microRNAs (miRNAs). The miRNA interacts within a protein complex with messenger RNAs, preferably in their 3' UTR regions and thereby repress translation (Filipowicz et. al., 2008).

The genetic code can also be fine-tuned by regulating the nucleotide composition of the messenger RNA. In site selective A-to-I editing, a deaminase enzyme targets specific adenosines within pre-mRNA fold back structures. Hence, the translation of messengers with single base substitutions can thereby increase the variety of the proteome.

miRNA

MicroRNAs (miRNAs) mainly function in translational inhibition often by repetitive binding to the 3' UTR. The miRNA act as the guide RNA within a protein complex, ribo-nucleoprotein particles (RNPs). Here, the 5' portion (2-8 nucleotides) of the miRNA representing the "seed" sequence, act as a guide to miRNA recognition elements (MREs). The mechanisms of how the miRNP interaction with MRE:s influence regulation of gene expression is still surprisingly obscure but different ideas are reviewed in (Filipowicz et al., 2008) and (Wu et al., 2008).
Although not called miRNAs from start, the phenomena of repression of gene expression was discovered in *C. elegans* where the gene lin-4 was shown to timely regulate the expression of the protein lin-14 (Ambros et al., 1989). The regulatory function is now known to occur both at different developmental stages and during tissue specific differentiation. The realization that lin-4 acted as a small antisense RNA with complementary regions in the 3' UTR of lin-14 mRNA was discovered later (Lee et al., 1993). It took until the beginning of this decade for this class of RNA to formally explode in new discoveries (Lau et al., 2001) (Lee et al., 2001). Today, it is believed that more than a third of the human genes have target sequences for miRNAs (Lewis et al., 2005).

Drosha

MiRNAs are expressed via different processing steps where the primary miRNA transcripts (pri-miRNAs) first is recognized by the nuclear Drosha protein. Drosha cleaves the primary transcript into a shorter precursor miRNA (pre-miRNA) which is exported to the cytoplasm for further processing. The cytoplasmic Dicer trims the precursor down to a duplex of ~22 base pairs in length. One of the strands in the duplex is then incorporated into an RNP complex that suppresses target expression where the miRNA specifies the target recognition by the 5' antisense seed sequence, see Figure 1.
The complete picture of the miRNA biogenesis is however still not fully understood. The different proteins for different organisms involved in miRNA biogenesis is presented in Table 1.

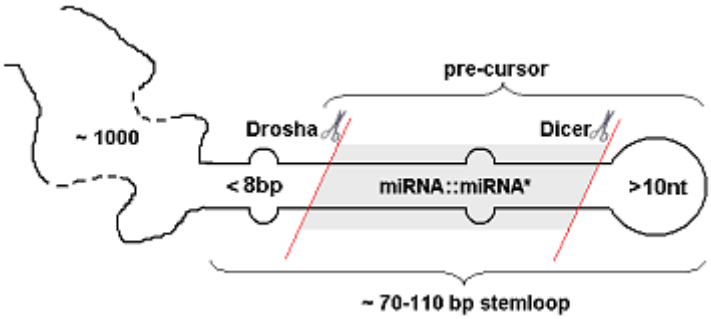


Figure 1.
The nucleolytic cleavages of Drosha and Dicer is here shown to produce the miRNA::miRNA* duplex. Studies have shown that the apex loop contains >10 unpaired nucleotides for optimal processing by Drosha/Pasha. A terminal stem region of maximum 8 base pairs is preferred by Drosha.

Vertebrate expression of miRNAs is known to originate from independent polymerase II transcripts that are initially processed as pri-miRNA molecules (Lee et al., 2004). Pri-miRNAs are ~1000 nucleotides long consisting of a signalling sub-structure generally called miRNA stem loop or miRNA hairpin.

Mapping by directional cloning of the 5'- and 3'-ends shows that the pre-miRNA has a 2 nucleotides 3'-end overhang (Basyuk et al., 2003). This is an RNase III characteristic. The nuclease responsible for the nuclear processing was not known until 2003. The pri-miRNA is recognized by the RNase III endonuclease Drosha probably via the internal apex loop (Zeng et al., 2005) (Lee et al., 2003). Drosha has recently been shown to be associated with a variety of other proteins in the so-called microprocessor complex (Denli et al., 2004) (Gregory et al., 2004). Since Drosha also has essential functions in the rRNA biogenesis the major part of these interactions might be specific also for rRNA processing events. However, the Pasha (**pa**rtner of Dro**sha**) protein is supposedly prone for miRNA genesis. The mammalian homolog to Pasha is DGCR8 (**D**i**G**eorge syndrome **c**hromosomal **r**egion **8**). DGCR8, located in the DiGeorge syndrome locus, has one WW- and one double stranded RNA binding motif (Shiohama et al., 2003). The WW motif is thought to mediate protein-protein interactions via proline rich motifs of an interacting partner. A proline rich motif in the N-terminal domain of Drosha is supposedly the interacting region. The function of this partnership is speculative but is thought to address correct Drosha cleavage of the miRNA hairpin since Drosha alone showed unspecific nuclease activity on an RNA construct (Gregory et al., 2004).

Organism	Protein	Function	ref
Plant (*A. thaliana*)	Dicer-like 1	miRNA biogenesis	Xie et. al., 2004
	Dicer-like 2	siRNA biogenesis	Xie et. al., 2004
	Dicer-like 3	siRNA directing heterochromatin formation	Xie et. al., 2004
	AGO1	Core component of RISC, ”slicer”	Vaucheret et. al., 2004.
C.elegans	Drosha	Nuclear endonuclease, initializes trimming of the primary transcript	Denli et. al., 2004
	Pasha	Partner of Drosha, co-ordinates Drosha cleavage	Denli et. al., 2004
	DCR-1	Cytoplasmic endonuclease, trims the precursor to the miRNA::miRNA* duplex	Tabara et al., 2002
D. melanogaster	Drosha	Nuclear endonuclease, initializes trimming of the primarytranscript	Denli et. al., 2004
	Pasha	Partner of Drosha, co-ordinates Drosha cleavage	Denli et. al., 2004
	Dicer-1	miRNA biogenesis	Bernstein et. al., 2001
	Dicer-2	siRNA biogenesis, interacts with R2D2	Pham et. al., 2004.
	R2D2	Involved in siRNA mature strand selection, interacts with Dicer-2	Liu et. al., 2003
	R3D1-L	Possibly involved in miRNA biogenesis, essential interaction with Dicer-1	Jiang et. al., 2005.
	AGO1	Core component RISC, elusive function	Okamura et. al., 2004
	AGO2	Core component RISC, “slicer”	Meister et. al., 2004.
H. sapiens	Drosha	Nuclear endonuclease, initializes trimming of the primary transcript	Lee et. al., 2003
	Pasha/DGCR8	Partner of Drosha, co-ordinates Drosha cleavage	Denli et. al., 2004 Han et. al., 2004
	Dicer	Cytoplasmic endonuclease, trims the precursor to the miRNA::miRNA* duplex	Hutvagner et. al., 2001
	AGO1	Core component RISC, elusive function in miRNA biogenesis	Meister et. al., 2004
	AGO2	Core component RISC, “slicer”	Meister et. al., 2004.

Table 1.
A compilation of some of the core proteins in the miRNA biogenesis. Adapted from Tang, 2005 (tang et al., 2005).

The miRNAs mir-21, 27a, 30a, and 31 were tested for secondary structural preferences by Drosha (Zeng et al., 2005). Here, a loop size of a minimum of 10 nucleotides seemed necessary for Drosha interaction and was, at least in the test set, sequence independent. Interestingly, all loops and some additional structural elements were found to be mispredicted by previous folding algorithms. The stem terminus of the hairpin has a preference of 8 base pairs for correct processing by Drosha, see Figure 1. An 18 base pairs extension abolished pre-miRNA expression. Further, de-stabilizing the region between the pre-miRNA and the hairpin termini severely affects the mature miRNA expression, (Zeng et al., 2005). Mutational analyses of this region indicate that it is the structural features rather than sequential that determine correct hairpin processing.

Having features addressing optimal interaction with Drosha, leaves an RNaseIII characteristic of 2 nucleotides 3’ overhang, about 2 helical turns from the apex loop (Lee et al., 2003) (Zeng et al., 2004).

The end-product, pre-miRNA, of the micro processing complex leaves a signature through the 3’ overhang to exportin-5 for shuttling to the cytoplasm (Yi et al., 2003) and to Dicer for cytoplasmic processing.

Dicer

The Drosha RNase III cleavage creating a 2 nucleotide 3' overhang, directs further maturation in some crucial aspects. One is the recognition by Exportin-5, and secondly it leaves a canonical substrate for the Dicer class III RNase III through its PAZ domain. Dicer was the first enzyme shown to be involved in the let-7 biogenesis and later crucial for miRNA/RNAi gene supression, (Hutvagner et al., 2001). Depending on the species, Dicer is represented by either one or two proteins, see Table 1. Let-7 is an abundant phylogenetically conserved miRNA known to silence regulatory genes during early larva development in *C. elegans* (Reinhart et al., 2000) (Pasquinelli et al., 2000).

The PAZ domain has been shown to be actively engaged in the interaction with the 2 nt 3'-overhangs in a sequence independent manner (Ma et al., 2004). The PAZ-domain interacts predominantly with the first 7 base(pairs) of the RNA strand in the 3' -> 5' orientation. The specific cleavage of the precursor, executed by Dicer and directed by the PAZ-domain, is believed to be a result of an intramolecular dimer, positioning one of the Dicer constituent catalytic cleavage sites to generate the miRNA::miRNA* (star) intermediate (Zhang et al., 2004a), see Figure 1. This further explained the difference between Dicer and a bacterial RNase III that does not dimerize thus leaving specific cleavage distributions around 10bp.

In the siRNA maturation pathway (that shares many mechanisms with the miRNA maturation) Dicer-2 has been shown to act in coordination with the R2D2 protein in *D. melanogaster* for distinct orientation and correct loading of the siRNA to the **R**NA **I**nduced **S**ilencing **C**omplex (RISC) and Ago2 (Tomari et al., 2004). Based on homology, a possible counterpart for the miRNA biogenesis is the R3D1-L protein that has been shown to interact with Dicer-1 and enhances miRNA expression in vitro (Jiang et al., 2005). Also, R3D1-L is required for normal fly development. The suggestion here is that R3D1-L take on the same function in miRNA biogenesis.

RISC

The end product of Dicer cleavage, miRNA::miRNA* duplex, is readily recognized by the multiprotein complex RISC. Key components of the RISC are the members of the Argonaute protein family – Ago1 and Ago2. Ago1 is thought to be prone for the miRNA pathway (non-cleaving RISC) (Okamura et al., 2004) and Ago2 has been shown to be the actual "slicer" in siRNA silencing (Liu et al., 2004). The

function of Ago2 is still elusive in the miRNA context since it does not induce cleavage and degradation of the targeted transcript. The Ago2 enzyme also has a PAZ-domain that can interact with the 2 nt 3'-end overhang (ma et al., 2004). RISC is thought to contain a helicase component which is presumed to be involved in the selection of the functional mature miRNA in the miRNA::miRNA (Tomari et al., 2004). However, unwinding by a RISC, containing helicase, is uncertain since it also co-immunoprecipitates with Dicer. Regardless, the choice is directed toward the strand that has the least stable 5' end in the duplex. This mechanism of stand selection also holds for siRNA biogenesis (Krol et al., 2004) (Khvorova et al., 2003) (Schwarz et al., 2003). Hence, the mature miRNA can be encoded in either of the 2 arms separated by the apex loop in the precursor. The seed sequence of the mature miRNA is probably presented as target bait by RISC (Bartel, 2004).

Plant – Animal distinctions in miRNA biogenesis

The plant biogenesis of miRNAs is different in several aspects, even speculated to be of independent evolutionary origin (Bartel, 2004). First of all, plants lack any Drosha homologs. The endonucleolytic intermediate processing steps to produce a mature miRNA is believed to be due to a Dicer like protein, DCL1. DCL1 has mantled both Drosha and Dicer nucleolytic processing in the nucleus (Kurihara et al., 2004). Hence, the metazoan processing steps selectively acting in the nucleus and cytoplasm by Drosha and Dicer respectively is in plants coordinated by DCL1 alone. Consequently, in contrast to animal biogenesis, there are low levels of miRNA precursors since the pre-miRNA is such a transient intermediate (Reinhart et al., 2002). The precursor molecules are in addition predicted to be substantially larger then the metazoan counterparts (Reinhart et al., 2002). The processed precursor is transported to the cytoplasm by an Exportin-5 homolog, HASTY (Bollman et al., 2003) (Lund et al., 2004).

The cytoplasmic maturation is however in many aspects shared between the plant and animal kingdom. Also here, the miRNA::miRNA* intermediate duplex is unwound and the strand with the least stable 5'-end in the duplex is incorporated into the RISC (Krol et al., 2004). Also, plant miRNA generally show siRNA-like complementarity (with few if any mismatches) to their targets (Rhoades et al., 2002) (Bartel et al., 2003). Accordingly, miRNAs in plants degrade of mRNA targets rather than acting in translational repression (Llave et al., 2002) (Tang et al., 2003). Although the reason is not clear, it is known that plant miRNAs generally are complementary to coding regions of their targets, while animal miRNAs targets 3' UTRs.

A presentation of the general pathways in the biogenesis of miRNA in metazoan and plants are presented in Figure 2.

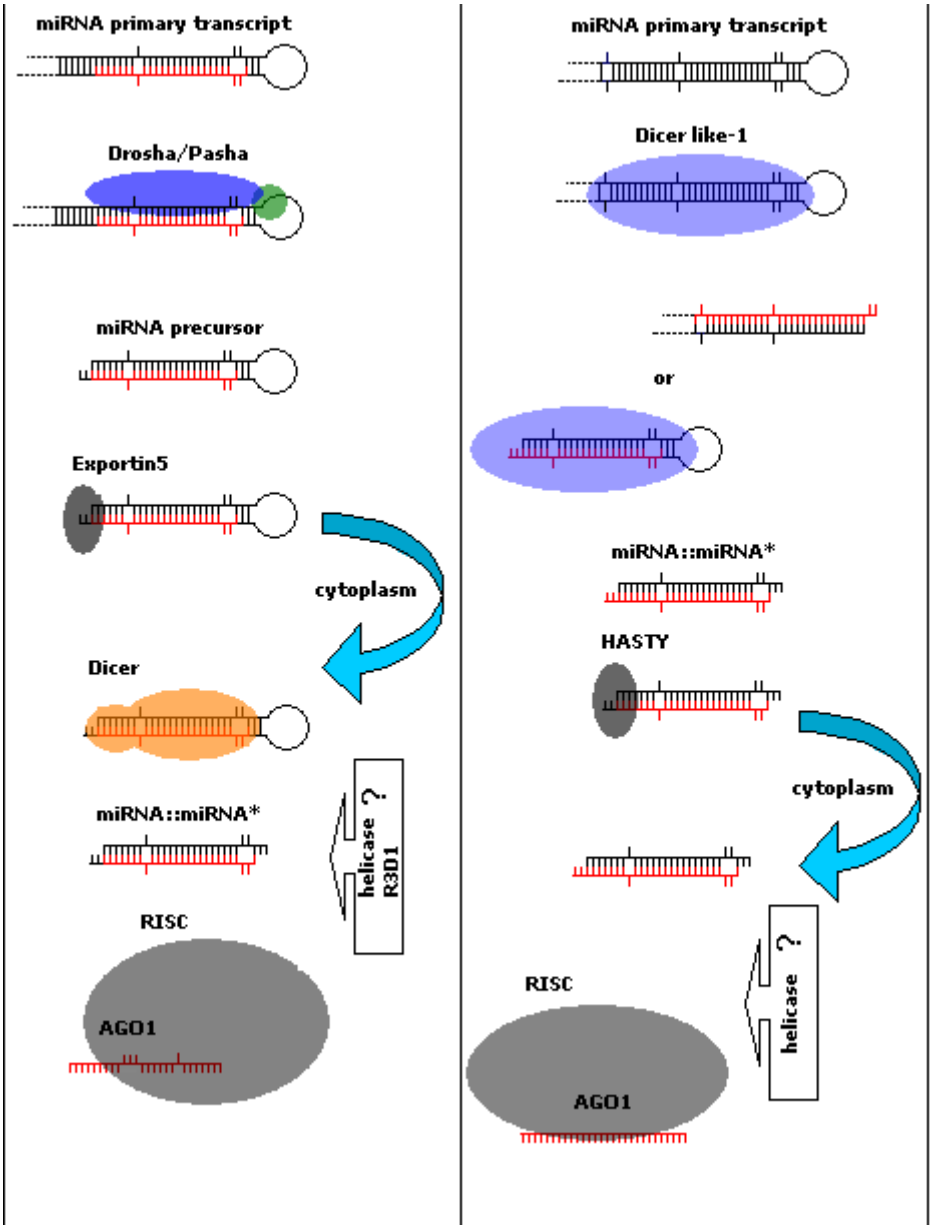


Figure2.
The differencies and similarities of microRNA biogenesis between metazoa and plants. The metazoan primary transcript is recognized by Drosha/Pasha complex via an apex loop and a terminal stem structure of a precursor miRNA. Transported to the cytoplasm and further processed by Dicer to a miRNA::miRNA* duplex. The duplex in unwound by a helicase assymetrically either in a complex with Dicer or RISC. Plants lack a Drosha homolog and the cleavage process producing a miRNA::miRNA* duplex is nuclear by Dicer homologs. Also, the precursor structures are generally much larger and could consequently produce mature duplexes either from the terminal of the hairpin or from near the apex loop. So in plants, the miRNA::miRNA* duplex, not the precursor, is transported to the cytoplasm for the same assymetric strand selection that result in the mature miRNA that are loaded onto RISC.

Bioinformatics

The efforts within the miRNA field have been focused on finding novel species of miRNAs accompanied with more recent screens to find targets for the miRNAs. This has been a successful joint expedition of both experimental (Lagos-Quintana et al., 2001) (Ambros et al., 2003) (Lee et al., 2001) (Lau et al., 2001) (Kim et al., 2004) and (Suh et al., 2004) and bioinformatics approaches (Lim et al., 2003) (Lai et al., 2003) (Bonnet et al., 2004) and (Wang et al., 2004). The strength of the experimental approach has been to selectively extract the miRNAs that are either tissue specific (Kim et al., 2004) (Suh et al., 2004) or involved in the timely regulation of target genes (Krichevski et al., 2003) (Reinhart et al., 2002) while computational screens have a more general agenda of miRNA disclosures. In pursuing the computational quest of finding novel miRNA species the focus has so far been on comparative genomics, with 2 to 4 organisms, and subsequent filtering in regard to both sequential and structural consensus features. The screens, "MiRscan" (Lim et al., 2003) and "MiRseeker" (Lai et al., 2003) are prime examples of this. The general idea is to first extract conserved non-coding regions from related organisms and subsequently fold these in windows of the general length of a pre-miRNA. Accordingly, they are scored for miRNA characteristics. This is based on sequence and structural features compiled from bona fide expressed miRNAs.

As mentioned, plants show few, if any, mismatches in the miRNA/MRE interaction. Therefore, plant target prediction screens are more straightforward in identity-like approaches by comparative genomics (Rhoades et al., 2002) (Jones-Rhoades et al., 2004). Vertebrate screens for miRNA targets are however generally less obvious since the miRNA/target duplex only involves comprehensive base pairing to the seed sequence (<8 nt) (Lagos-Quintana et al., 2003). An often used strategy has been to use comparative genomics to extract conserved sub-regions of 3' UTRs and then let a search algorithm find putative targets (Lewis et al., 2003) (Enright et al., 2003) (Stark et al., 2003) (Kiriakidou et al., 2004). The selection is based on two criteria: 1) highly conserved and non-mismatched ~7 first base pairs and 2) duplex energy formation characteristics based on a training set of bona fide miRNAs. Another target prediction screen utilizes solely the expected hybridization properties of the miRNA/target duplex (Rehmsmeier et al., 2004).

Editing

RNA editing was introduced as an RNA modifying mechanism in 1986 (Benne et al., 1986). This post-transcriptional modification insert or delete uridines within a pre-messenger *RNA* (pre-mRNA). RNA editing is now the collective term for alterations of nucleotides in a transcript that result in a discrepancy between the RNA and the genomic template *DNA*. For nuclear encoded messenger RNAs (mRNAs) two types of editing has been described: cytidine to uridine (C-to-U) (Chen et al., 1987) and adenosine to inosoine (A-to-I) editing (Bass et al., 1988).
A-to-I editing was first introduced as a concept in 1988 when an antisense RNA failed to block translation of a target transcript. The reason was that most *adenosines* in the antisense RNA had been deaminated to inosine*s* hence disrupting the anticipated hybridization properties to the target (Bass et al., 1988). This phenomena, was first called "unwinding/modifying activity", later disclosed in mammals as the function of the dsRAD protein (ADAR1) (Polson et al., Bass et al., 1994), RED1 (ADAR2) (Melcher et al., 1996b) and RED2 (ADAR3) (Melcher et al., 1996a). This type of abundant editing has later been characterized as *hyper* editing in contrast to *site selective* editing.
In response to the heading of this page: A-to-I editing is a phylogenetically conserved post-transcriptional processing event that converts adenosines to inosines by a hydrolytic deamination by the ADAR family of proteins.

A-to-I editing – Phenotyping species

Although not many site selective targets have been discovered, A-to-I editing has been detected in a variety of metazoan species where deficiencies in constitutive editing show phenotype defects. In vertebrates like, *D. melanogaster* there is one ADAR allele but several several isoforms due to distinct promoter signals and alternative processing of the transcript (Palladino et al., 2000a) (Keegan et al., 2005). Here, ADAR null mutants show extreme deficits in neurological function (Palladino et al., 2000b). *C. elegans* have two ADAR homologs, adr-1 and adr-2 where the expression of adr-1 is exclusively confined to the nervous system in adult worms (Tonkin et al., 2002). In chemotaxis assays, *C. elegans* show abnormalities in behavior in homozygous deletions of the two ADAR enzymes (Tonkin et al., 2002). For vertebrates, the lack of ADAR(s) show severe deficiencies in neurophysiology (Higuchi et al., 2000) (Brusa et al., 1995) leads to embryonic lethality due to tissue apoptosis (Hartner et al., 2004) (Wang et al., 2000). ADAR1 seems the most essential, where even ADAR1$^{+/-}$ heterozygotes are

lethal in mice (Wang et al., 2000). However, in retrospect of current models where ADARs are believed to dimerize, this could be an effect of non-canonical dimers formed by the products of the null allele and the wild-type allele respectively. ADAR2$^{+/-}$ however, show no abnormal phenotype and are viable whereas complete knock outs are lethal: mice die within 3 weeks of age while suffering from epileptic seizures. Deficiencies in ADAR expression have been connected to several abnormal phenotypes. Dyschromatosis symmetrica heriditaria (DSH) is a skin disease that have been linked to genomic polymorphisms in ADAR1 alleles (Zhang et al., 2004b). Also, reduced editing efficiency of ADAR2 is implicated both in Epilepsy and amytrophic lateral sclerosis (ALS) where the regulation in Ca^{2+} ion flux is impaired in a glutamate receptor (see below) (Kwak et al., 2005) (Brusa et al., 1995). A link between different editing patterns of the serotonin receptor 2c (see also below) and depression and suicide have been shown (Niswender et al., 2001) (Gurevich et al., 2002).
Looking more at mammals specifically, constitutive editing is found in various tissues and cell lines (Wagner et al., 1990). ADAR1 is more uniformly expressed and was found in all tissues tested (O'Connell et al., 1995). ADAR2 is found primarily in nervous tissues but can also be detected in lung, heart, testis and kidney (Melcher et al., 1996a) (Rueter et al., 1999). ADAR3, which in contrast to the other family members, is expressed in selective brain tissues only (Melcher et al., 1996a). ADAR2 and ADAR3, although specific for the brain, have a differentiated expression pattern looking at various brain tissues (Barbon et al., 2003). In brain, ADAR1 show near homogenous levels of mRNAs while ADAR2 is most prominent in the caudate nucleus, thalamus and cerebellum. ADAR3 is mostly expressed in amygdala and corpus callosum (Barbon et al., 2003). It is worth noting that ADAR3 have no known substrates and no measureable enzymatic activity. Although, endogenously expressed even in a regulated tissue specific manner the function of this family member must be seen as something of a mystery.

ADARs – Description of goods

Focusing on the mammalian system, there are as previously mentioned three ADAR family members. The common domain for all three members, is the highly conserved deaminase domain covering the large part of the C-terminus. They also have double stranded RNA binding motifs (dsRBMs). ADAR1 has three and ADAR2/3 has two dsRBMs. The final common trait is the nuclear localization signal (NLS) (Kim et al., 1994) (O'Connell et al., 1995) (Melcher et al., red2 et al.,

1996). This conclude the similarities. Full length ADAR1 has in addition one nuclear export signal (NES) (Poulsen et al., 2001) and two Z-DNA binding domains (Herbert et al., 1997). ADAR1 has two isoforms that come from an alternate use of two different initiation codons (George et al., 1999). The two isoforms is usually termed p110 and p150. The use of the upstream promoter that results in the p150 version is interferon inducible (Patterson et al., 1995) (George et al., 1999). Interferon production is induced by the immune defense system in response to infectious agents like viruses. Purposely, the p150 form includes the N-terminal part where the NES resides. The function of ADAR1 in the cytoplasm is described in the substrates section. The ADAR2 genomic loci express several different isoforms (Lai et al., 1997) (Gerber et al., 1997). Intriguingly, one alternative transcript results from a feedback mechanism where ADARs target the Adar2 transcript mimicking an (AI) 3' splice site di-nucleotide (Rueter et al., 1999). The result is a 47 nucleotide insertion that creates a frame shift that leads to a truncated protein. ADAR3 is the black sheep in the family. It has a single stranded RNA binding motif (ssRBM) and lack any detectable enzymatic activity. It also inhibits constitutive editing by the other members (Chen et al., 2000) (Sergeeva et al., 2007).

An interesting aspect of domain composition and function relates to the specificity of the ADARs in target recognition. I have briefly discussed RNA target traits that promote ADAR acceptance. The dsRBMs of the ADAR enzymes are obviously one part of the recognition of target RNAs. However, maybe more interesting is that the catalytic domain seem to have a dominant role in substrate recognition (Wong et al., 2001). Here, a chimeric construct was made with interchanged deaminase domains between ADAR1 and ADAR2. Even with exchanged catalytic domains, they kept the substrate specificity respectively. Looking at the dsRBMs of ADAR2, they seem to have overlapping but distinct binding specificities to the target RNA (stephens et al., 2004) (Stefl et al., 2006). In a construct containing the GluR-B Q/R fold back structure (Stephens et al., 2004) show that the two dsRBMs of ADAR2 (subscript I and II) have an overlapping dsRBM/RNA interface. However, RBM_{II} binding to the foldback RNA is severely affected by a modified nucleotide, 19 base pairs from the edited site. RBM_{I} is affected by the same modification of a nucleotide situated 13 base pairs from the edited site. In contrast, at the R/G fold back RNA in the same transcript, (Stefl et al., 2006) place the $dsRBM_{II}$ directly over the edited site and RBM_{I} is shown to interact with the downstream pentaloop. Interestingly, substituting the RBMs on both ADAR1 and ADAR2 to those of another RNA binding protein, PKR, showed significantly different binding properties to the RNA compared to the wild-type composition (Liu et al., 2000) (Stephens et al., 2004). Although, dsRBMs are expected to

adopt the same conserved αββββα fold, these result suggest that distinct amino acid sequences rather than RNA properties direct the correct positioning of the protein interface. The dsRBMs of ADAR1 have also been screened for functional properties. Of the three dsRBMs of ADAR1, the most important seem to be $dsRBM_{III}$ followed by $dsRBM_{I}$ while $dsRBM_{II}$ seem dispensable (Lai et al., 1995) (Liu et al., 1996).

Both ADAR1 and ADAR2 have been shown to localize to the nucleolus (Desterro et al., 2003) (Sansam et al., 2003). Hence, ADAR1 has a tri-partie compartmentalization: nucleolus, nucleoplasm and cytoplasm. ADAR2 was shown to shuttle rapidly between different nuclear loci (Sansam et al., 2003), probably in response to its *modus operandi*. The belief is that the nucleolus function as a storage room in where dimerization, thus catalytic activity, is hindered by the high stoichiometric ratio of rRNA/ADAR, (see also dimerization section).

The crystal structure of the catalytic domain was solved in 2005 (Macbeth et al., 2005). The most striking discovery was that inositol hexakisphosphate IP_6 was present in the active core. This molecule is also present in ADAT1 which is an adenosine deaminase acting on tRNA and also related to the deaminases acting on mRNAs (Maas et al., 2000). IP_6 was shown not to be part of the catalytic centre but rather maintaining the structural properties essential for the catalyses. Interestingly, ADAT1 is here speculated to be the evolutionary link between ADATs and ADARs since the other members of ADATs do not require IP_6 (Macbeth et al., 2005). Domain composition of deaminase family members are presented in Figure 4.

ADARs - Dimerization

When issued, the dimerization as a constitutive property of ADARs raised some controversy (Jaikaran et al., 2002). Although recent results leaves little space for a monomeric ADAR operation there is still some dispute if the RNA is required for dimerization or not (Gallo et al., 2003) (Valente et al., 2007). Another controversy, concerns heterodimerization which is not normally believed to occur (Cho et al., 2003) albeit ADAR1/ADAR2 dimers have been suggested to exist in astrocytoma cell lines where the non-canonical dimer is thought to infer the malignant phenotype due to reduced editing activity of ADAR2 in this conformation (Cenci et al., 2008). Also, as mentioned, the nucleolus is believed to function as a storage room for both ADAR1 and ADAR2 (Desterro et al., 2003) (Sansam et al., 2003). Regarding dimerization, a recent paper have shown that both homodimers and heterodimers exist in the nucleolus (Chilibeck et al., 2006).

In their studies, they use FRET analyses with recombinant ADARs with either CFP or YFP tagged to the N-terminal. FRET, energy transfer signals vary with r^6 (Förster, 1948) hence, detected signals of the fusion proteins is a very strong indication of proximity. Albeit, the question of heterodimerization and the function of such interaction is still open.

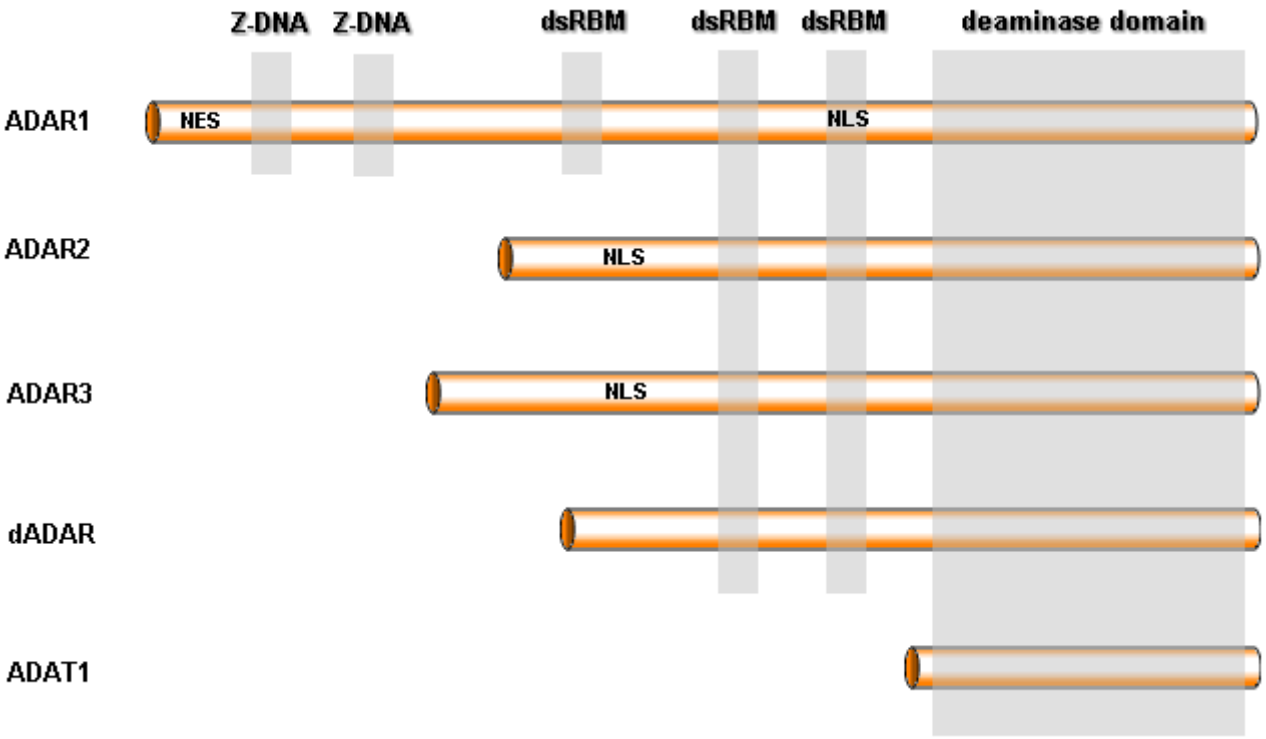


Figure 4.
Functional composition of different domains of selected members of the deaminase family of proteins. ADAR1-3 is the mammal deaminases acting on double stranded RNA. dADAR is the *D. melanogaster* homolog and ADAT1 is a deaminase targeting tRNA and suggested to be the evolutionary link between ADATs and ADARs. NES – Nuclear export signal. NLS – nuclear localization signal. dsRBM – double stranded RNA binding motif. Z-DNA – Z-DNA binding domain.

Protein composition and the requirements for dimerization has been studied for the *D. Melanogaster* single dADAR, which in structure is most similar to the mammalian ADAR2. Minimum requirements for dimerization of dADAR is the N-terminal part and the first dsRBM corresponding to amino acids 1-133 (Gallo et al., 2003). A similar result for human ADAR2 has been shown (Poulsen et al., 2006). Here, they show that mutations in $dsRBM_I$ lowers the affinity for the dimerization interface while mutations in $dsRBM_{II}$ do not have the same effect.

A-to-I editing – the RNA

Site selective editing targets single adenosines within an imperfect RNA foldback structure while *hyper* editing indiscriminately edits multiple adenosines within an almost completely duplexed structure. The term hyper editing is sometimes used interchangeable with *promiscuous* editing. The properties that make an RNA prone for *site selective* editing is still not fully understood but the consensus belief is that internal mismatches and bulges constitute a recognizable landscape for the ADAR selectivity (Källman et al., 2003) (Dawson et al., 2004) (Stephens et al., 2004).

The foldback imperfect structure is often composed of an upstream partly exonic element folded to a trailing intron element although all combinations are seen, see Figure 5. The complementary intronic element is called editing complementary sequence (ECS). The loop region of this fold back structure could range from a small penta loop to thousands of nucleotides. Besides the general preferred structural features of the foldback duplex, there is a bias in the nucleotide frequency of adjacent positions of an edited site. There is a clear deficit in guanosines 5' to an edited site. The reverse holds 3' to an edited adenosine, where there is a preference for a guanosine closely followed by a uridine (Polson et al., 1994) (Lehmann et al., 2000) (Ensterö et al., unpublished). There is also a preference toward edited A:s in a A-C mismatch bulge although A:s in a A-U base pair are also edited but to a lesser degree (Wong et al., 2001). However, an editing event targeting an A in an A-G mismatch bulge is never seen. Sequence and structural determinants for ADAR recognition have been studied for the separate RNA targets of the glutamate receptor b (GluR-B), R/G and Q/R sites (Stefl et al., 2006) and (Stephens et al., 2004) respectively. Also, the edited transcript of Adar2 itself have been screened for structural and sequential preferences of ADARs (Dawson et al., 2004). Although interesting in detail, preferences cannot be considered consensus but rather show structural and sequential determinants that are critical for ADAR on those specific substrates. However, certain general things are clear. The specificity for ADAR/RNA interaction is intrinsically dual: Structural and sequential composition of the RNA and properties of the ADARs determined by the amino acid composition.

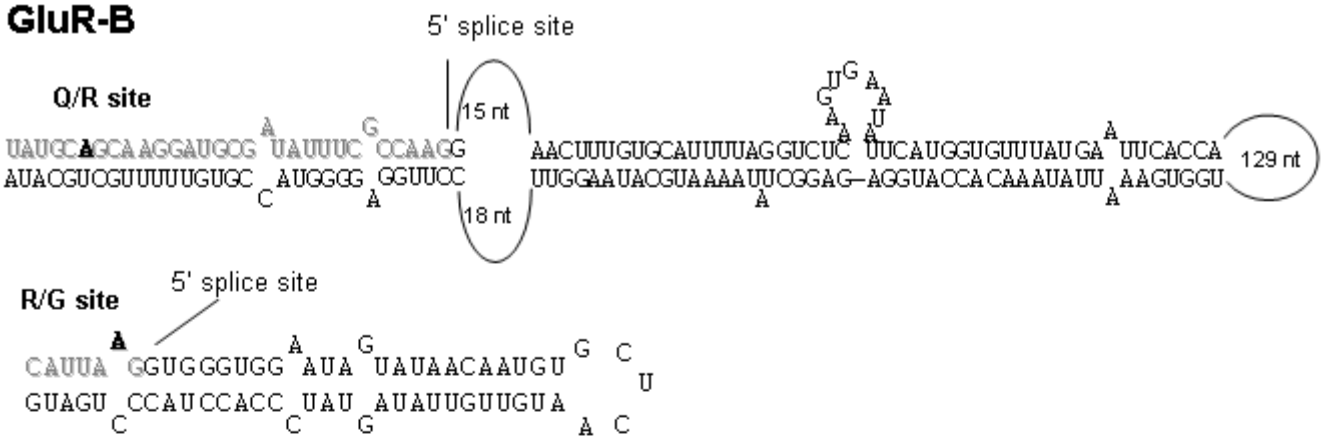
A-to-I editing - substrates

The most prominent examples of A-to-I editing comes from transcripts coding for various ligand or voltage gated transmembrane proteins in the central nervous system. For a near complete list of re-coding editing events see Table 2.

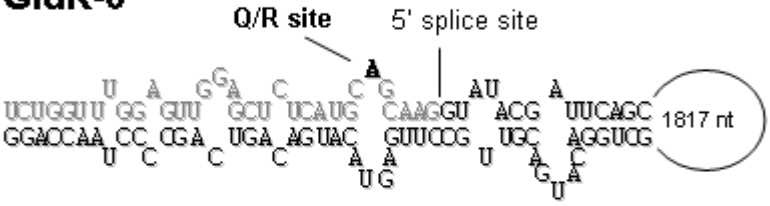
The glutamate receptors (GluRs), divided into AMPA, NMDA and kainite (or non-NMDA) are ion channels responding to the ligand binding of glutamate which is the major neurotransmitter in mammals. The AMPA receptor is a heteromer consisting of the four subunits A, B, C, D. The kainite receptor is mainly a 4 unit heteromer of the subunits 5, 6, 7, KA1 and KA2.

The transcripts of the AMPA subunits B, C, and D are subjected to A-to-I editing. Editing of the GluR-B transcript has been shown to be essential to the organism. (Brusa et al., 1995) (Seeburg et al., 1998).

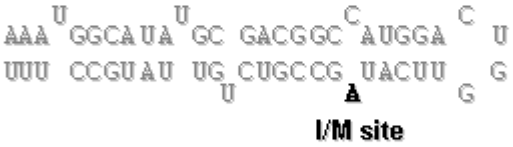
GluR-B



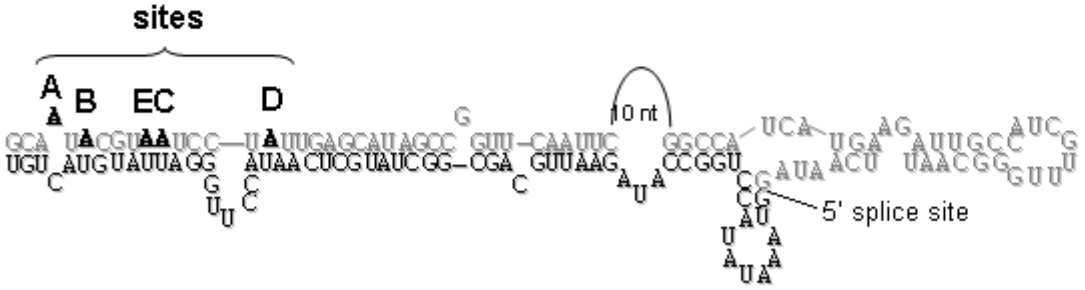
GluR-6



Gabra-3



5-HT$_{2C}$R



Adar2

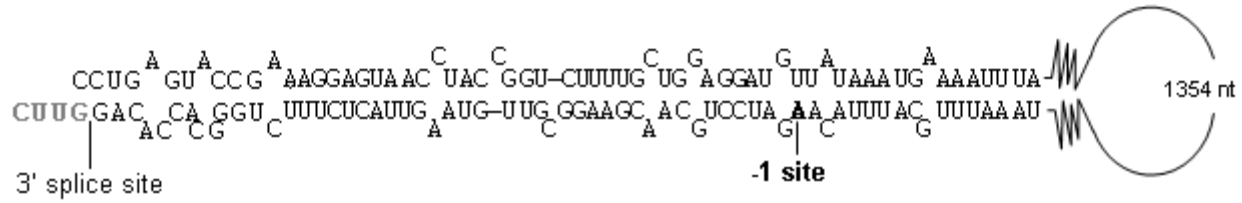


Figure 5.
A selected set of known ADAR target RNAs. The AMPA type GluR-B, glutamate receptor B sites Q/R and R/G. The Q/R site is the "site of sites". Edited to near 100% through development. GluR-6, glutamate receptor subunit 6 of the kainite kainate. GABRA3, $GABA_A$ receptor subunit $\alpha 3$. 5-HT_{2C}R, the serotonin receptor 2c. Edited sites that show high degree of tissue specific pattern and also targeted by both ADAR1 and ADAR2. Adar2, the pre-messenger structure of the heavily edited region that result in the AI di-nucleotide at position -1, mimicking a 3' splice site.

The pre-mRNA of GluR-B has two sites of re-coding A-to-I editing, the Q/R and the R/G site. At the Q/R site, editing of the CAG codon (glutamine = Q) result in the functional CGG (arginine = R). This site is also edited close to 100% through out development (Seeburg et al., 1998). The functional consequence of the arginine substitution is a severe reduction of Ca^{2+} permeability (Sommer et al., 1991) (Geiger et al., 1995). The R/G sites of the GluR-C and –D are edited to a lesser degree and are not as essential for viability as the Q/R site. The R- and G-forms of the receptors show different rates to recover from desensitization (Lomeli et al., 1994).

The kainite receptors are also subjected to editing. The GluR-5 and GluR-6 subunits both have a Q/R substitution site. Subunit 6 has additional sites where an isoleucine is changed to valine and tyrosine to a cysteine (Köhler et al., 1993) (Herb et al., 1996). The functional consequence of these editing events also involves Ca_{2+} permeability (Hollmann et al., 1994).

The serotonin receptors, 5-HT_{1-7} are, exept for 5-HT_3, G-protein coupled receptors with a functional affinity for the neurotransmitter serotonin. Serotonin binding triggers the activity and release of other transmitter substances such as glutamate, dopamine and gamma-aminobutyric acid (GABA). The 5-HT_{2C}, receptor have 5 editing sites, termed A, B, E, C and D, in a small region spanning only 13 nucleotides close to a splice site of the pre-messenger transcript (Burns et al., 1997) (Niswender et al., 1998). Since, the edited sites are in such a close proximity, different combination of triplet compositions can result in a variety of functionally distinct receptor properties (Niswender et al., 1999).

Potassium ion channels are present in virtually all phyla and are found in most cell types in metazoa. The subfamily K_v1 also called the shaker related subfamily have a member $K_v1.1$ or KCNA1 that is A-to-I edited (Hoopengardner et al., 2003) (Bhalla et al., 2004). The KCNA1 ion channel regulate K^+ flow in response to transmembrane currents changing the potential across the membrane. Here, an isoleucine to valine substitution give a 20 fold increase in the recovery rate from fast inactivation. Within mollusks like the squid, additional sites in $K_v1.1$ and another subfamily member (K_v2) show extensive editing (Patton et al., 1997) (Rosenthal et al., 2002). Interestingly, this is possibly due to a self regulatory adaptation to water temperature.

gene	alias	Re-coding	% edited [1]	Specificity[2]	reference
glur-b	gria2	Q/R	100	ADAR2	(Higuchi et al., 1993) (Seeburg et al., 1998) (Barbon et al., 2003)
		R/G	70	ADAR1:**ADAR2**	(Lomeli et al., 1994)
glur-c	gria3	R/G	90	ADAR1	(Lomeli et al., 1994)
glur-d	gria4	R/G	45	ADAR1:**ADAR2**	(Lomeli et al., 1994) (Higuchi et al., 2000)
glur-5	grik1	Q/R	60	ADAR2	(Sommer et al., 1991) (Higuchi et al., 2000) (Barbon et al., 2003)
glur-6	grik2	Q/R	80	ADAR1:**ADAR2**	(Sommer et al., 1991)
		I/V	70	ADAR1:**ADAR2**	(Higuchi et al., 2000)
		Y/C	80	ADAR2	(Köhler et al., 1993) (Barbon et al., 2003)
gabra3		I/M	90	ADAR1:**ADAR2**	(Ohlson et al., 2007)
5-ht2c	htr2c	A [3]	80	ADAR1	(Burns et al., 1997)
		B	70	ADAR1:ADAR2	(Liu et al., 1999)
		E	4	n/a	
		C	25	ADAR1:ADAR2	
		D	60	ADAR2	
cyfip2		K/E	75	n/a	(Levanon et al., 2005)
kcna1	k_v1.1	I/V	25-45	ADAR2	(Bhalla et al., 2004)
blcap	bc10		28-60	n/a	(levanon et al., 2005) (Cutterbuck et al., 2005)
igfbp7		Q/R, R/G	n/a	n/a	(Levanon et al., 2005)
flna		Q/R	40	n/a	(Levanon et al., 2005)
lustr1 [4]	gpr107	H/R	58	n/a	(Athanasiadis et al., 2004)
		Q/R	29		
kiaa0500 [4]		Q/R	27	n/a	(Ahanasiadis et al., 2004)
Ednrb [5]		Q/R	n/a	ADAR1:**ADAR2**	(Tanoue et al., 2005)

Table 2.
A-to-I re-coding sites in mammalian transcripts. 1) Editing frequencies are a pooled consensus of mammal adult editing, both from references and (Ensterö et al., 2008) un-published data. 2) Where applicable, this column specifies the major targeting editing enzyme of the ADAR family. Both ADAR1 and ADAR2 are annotated if the specificity is overlapping. Bold, if overlapping but preferred by one of the ADARs. 3) Different re-coding possibilities due to dual editing events in the same codon. 4) Human specific editing of ALU regions. 5) Human specific editing in a disease phenotype.

A very recent discovery from our own laboratory is the A-to-I editing of the transcript coding for the $GABA_A$ receptor subunit $\alpha 3$, or Gabra3 (Ohlson et al., 2007). $GABA_A$ receptors, are ligand gated Cl^- channels reacting to the binding of GABA which is the major inhibitory neurotransmitter in the brain. On the amino acid level, the editing event leads to an isoleucine to methionone change. Preliminary data propose the editing to affect receptor assembly (Daniel et al., unpublished).

Bladder cancer associated protein (BLCAP or BC10), cytoplasmic FMR1 interacting protein 2 (CYFIP2), Insulin-like growth factor binding protein 7 (IGFBP7) and filamin A (FLNA) were all detected by the computational approaches described in that section (Levanon et al., 2005) (Clutterbuck et al., 2005). BLCAP, has an unknown function but is down regulated during bladder cancer progression (Gromova et al., 2002) yet it is mainly expressed in brain tissue and B-cells (Su

et al., 2004). CYFIP2 is expressed in brain tissue, white blood cells and kidney (Su et al., 2004). The A-to-I editing events were only found in the cerebellum while no editing were found in liver (Levanon et al., 2005). Interestingly, CYFIP2 is a p53 inducible protein (Saller et al., 1999). The tumor suppressor gene p53 has previously been found to also be subjected to A-to-I editing in intronic and 3' UTR ALU elements (Athanasiadis et al., 2004). A link between cancer progression and editing has been proposed earlier where aberrant expression patterns of all three ADAR members could be associated with the proliferation of different cancers (Paz et al., 2007) (Maas et al., 2001) (Cenci et al., 2008). Based on homology with IGFBP5, the editing in IGFBP7 is thought to regulate the stability of di-partie comlex with insulin growth factor (Levanon et al., 2005). In FLNA, the edited adenosine reside in a transcript region coding for an immunoglobulin-like domain of the protein (Levanon et al., 2005). This domain has been shown to interact with integrin beta (Travis et al., 2004) and GTPase Rac1 (Ohta et al., 1999). Also here, based on homology with related solved structures (ABP120 from *D. melonogaster* and gamma filamin), the putative result of the amino acid substitution is a modified interface to the interacting proteins.

Editing in the endothelin receptor B was detected during a mutational screen in patients suffering from Hirschsprung disease (Tanoue et al., 2002). Besides the Q/R codon change, they see a possible pattern between editing and alternative splice variants that are not translated. This editing event has not been confirmed in other mammals.

Editing of ALU elements residing in coding parts has been found in the computational screens. A-to-I editing was found in LUSTR1 and kiaa0500 (Athanasiadis et al., 2004). No function has been suggested for these events.

Several editied viral RNAs has been studied and the most well characterized are the hepatitis delta virus (HDV) (Polson et al., 1996), measle virus measle (Cattaneo et al., 1988), polyoma virus (Kumar et al., 1997) and the recent human herpes virus 8 (HHV8) (Gandy et al., 2007). The viral RNA species that have been found to be selectively edited are the amber/W site in HDV and the K12 open reading frame (ORF) in HHV8. Cytoplasmic A-to-I editing involve the interferon inducible ADAR1 p150. An up-regulation of p150 can be seen in acute inflammations resulting in a cellular interferon immunoresponse (Poulsen et al., 2001) (Yang et al., 2003). Although the general belief is that ADAR1 is part of the cellular defense mechanism for the intrusion of exogenous RNA, editing of the amber/W site changes an amber stop codon to a functional tryptophan *essential* for the viral life cycle (Polson et al., 1996) (Chang et al., 1991). A very recent finding of the ADAR1 editing is the K12 transcript in HHV8, coding for up to three versions of the kaposin protein as well as a miRNA. Here, editing suppress the

tumorigenic potential of the of the ORF. Editing specifically targets the miRNA sequence in the seed part thereby possibly creating a dud miRNA. In general, as in hyperedited measle and polyoma viruses, the functional effect of editing as a defense mechanism is still elusive.

Coming as no surprise to anyone, miRNAs has also been found to be edited. (Luciano et al., 2004) (Blow et al., 2006) (Yang et al., 2006). MicroRNAs have been found to be edited both in the cytoplasm (Luciano et al., 2004) and in the nucleus (Yang et al., 2006). Interestingly, TUDOR-SN has been shown to enhance site specific cleavage of inosine containing dsRNAs (Scadden, 2005) (Scadden et al., 2005). An interference between the Drosha and ADAR machinaries has also been implicated in edited pri-miRNAs (Yang et al., 2006).

Bioinformatics

The concentration of inosine in poly (A) transcripts can not be explained by the only handful of known targets of A-to-I editing (Paul et al., 1998). Together with the fact that inosine levels followed the expression pattern of ADARs (Paul et al., 1998), this left room for more ADAR substrates to be found.

Since then, several computational attempts have been made with the aim to discover novel editing sites. (Athanasiadis et al., 2004) (Blow et al., 2004) (Clutterbuck et al., 2005) (Levanon et al., 2004) (Levanon et al., 2005).

Common ingredients to characterize candidate A-to-I editing substrates have been to base the search on features of the known edited sites. The computational attempts involve in general filters according to Table 3. These screens have the following features of a candidate A-to-I editing event: Since editing acts on the post-transcriptional level, a comparison between an mRNA that has been subjected to editing and the genomic template would yield an A to G discrepancy at the edited position. Consequently, alignments between expressed sequences and the template DNA give a set of possible editing events at A/G mismatch positions. Known target regions of mRNAs often contain additional sites with deaminated adenosines. Hence, A clustering of A/G discrepancies within a limited region is more likely to have been targeted by ADARs than single A/G discrepancies located at distances not normally coherent with bona fide editing events. An A/G discrepancy could originate from genomic A/G polymorphisms within the species. Discrepancies passing a filter to exclude such genomic purine polymorphisms (i.e., SNP database), strengthen the candidate A/G discrepancy as a true editing event. The target RNA is known to adopt an imperfect RNA foldback structure with non-branched helical features.

Hence, a cluster of A/G discrepancies, not of genomic origin, in such a predicted structure further single out true editing events. The RNA target structure of most known editing events reside in regions, highly conserved in sequence and structure. A candidate A/G discrepancy, not of genomic origin, clustered with others, in an imperfect foldback structure, that also showed a high degree of sequence conservation, are very strong signs of a candidate editing event that originates from ADAR targeting.

	Clutterbuck [1)]	Blow [2)]	Levanon [3)]	Levanon [4)]	Athanasiadis [5)]
A/G	x	x	x	x	x
cluster	x	x		x	x
SNP filter	x	x		x	x
stem	x			x	
conservation	x		x		

Table 3.
How recent approaches to find novel editing events have used filters based on the properties of bona fide editing sites. A/G: A genomic adenosine found to be a guanosine at the transcript level. Cluster: A target of ADARs often show multiple inosines within the limited fold back structure. SNP filter: Discard all nucleotide ambiguities that originate from genomic polymorphisms. Stem: Search for predicted stems that fulfill a reasonable good duplex for ADARs to target. Conservation: Most of the known targets are highly conserved, a candidate region should also.
1) (Clutterbuck et al., 2005) 2) (Blow et al., 2004) 3) (Levanon et al., 2004) 4) (Levanon et al., 2005) 5) (Athanasiadis et al., 2004)

The results were unanimously indicating A-to-I editing as a ubiquitous mechanism with numerous targets in the pre-spliced transcriptome. The numerous sites characterized as hyper edited was dominantly localized to ALU repeats/inverted repeats of untranslated regions although some derived from exonized ALU sequences (Athanasiadis et al., 2004). In summary, the outcome with respect to re-coding sites increased the present repertoire with ~50%.

A slightly different approach was conducted in Drosophila, where highly conserved amino acid regions between fly species revealed discrepancies that were deduced to be a result of A-to-I editing (Hoopengardner et al., 2003). The human homolog, KCNA1 were later disclosed (Bhalla et al., 2004).

Present investigation

Paper I.

Asymmetries in the processing of a miRNA::miRNA* duplex regarding the thermodynamic properties of respective duplex termini, led us to investigate this issue in a sequential context. We calculated the information content of a reduced set of well annotated miRNAs from vertebrates, invertebrates and plants. We could show that: Vertebrates have a characteristic sequential motif of a 5' miRNA. Invertebrates show the reverse, a 3' characteristic pattern. Plants have characteristic motifs on both arms. In addition, we used ALLR score to compute if the seen motifs also differ significantly. In Figure 6, is the specific motif seen for mature miRNA from the 5' arm and the corresponding logo for the 3' arm sequential signature in the precursor context

Paper II.

By co-immunoprecipitation, we claim that it is possible to extract novel ADAR RNA targets. Here, a specific anti-ADAR antibody is used pull down the ADAR/RNA complex with sepharose A beads. The motivation was dual: Firstly, a previous study showed that ADAR2 binds more distinctly to site selective substrates than to almost completely duplexed RNA even within the same RNA molecule. Secondly, ADAR2 also seem prone for binds selectively edited *and* un-edited substrates with the same affinity. Consequently, the assumption was that we expected both a bias towards bona fide targets rather than un-specific binding to random dsRNA and also that the ADAR/RNA interaction would be more consistent. The pulled down RNA was subsequently hybridized to three consecutive genomic micro arrays to identify the genomic origin (i.e., gene). The enrichment analyses were made in comparison with the signal to noise from an identical and parallel experiment with pre-immune sera. In addition, we used mouse as a model organism mainly due to the fact that the mouse genome contains very little of the ubiquitously A-to-I edited repetitive elements that are present in humans. We could finally conclude a list of candidate editing targets (genes) based on the level of enrichment of the three micro arrays.

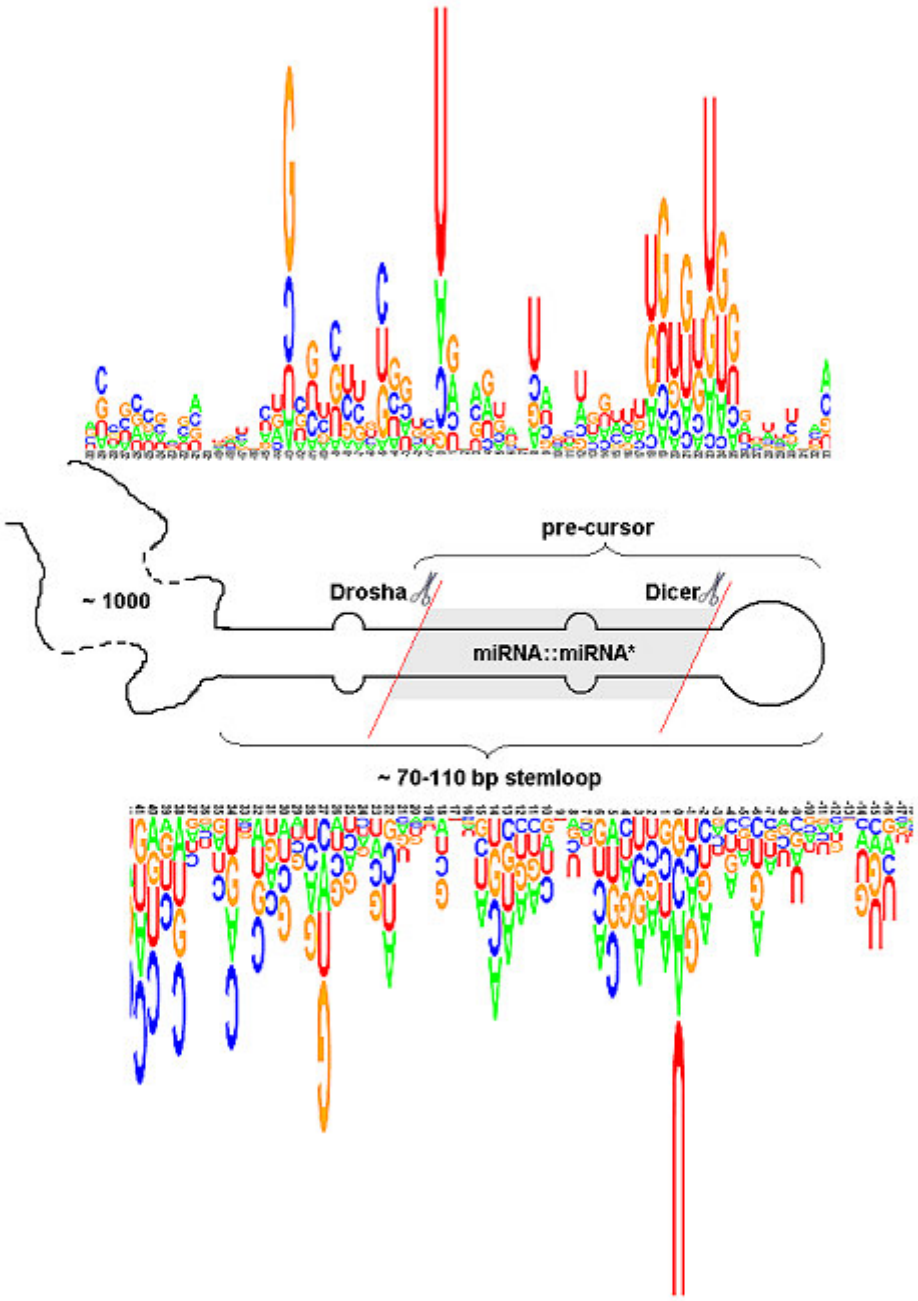


Figure 6.
The human sequence logos of the sequential composition of respective stem arm in the cases where the mature miRNA resides from the 5′ or 3′ side of the miRNA::miRNA* duplex. A significant signature is seem in a mature miRNA of the 5′ arm.

Paper III

We have constructed a computational screen to detect novel A-to-I edited sites. One part is focusing on the detection of inverted repeats within a highly conserved genomic region with the possibility to create a fold back duplex structure. We call this the explorative screen. The evaluation of conservation p We further evaluated the explorative screen for enrichment of A/G discrepancies in an alignment between expressed and genomic data. We significantly show that A/G mismatches are overrepresented in these conserved stems. In addition, we extended the explorative screen by including a site-scoring scheme based on features from the known editing sites. Based on the scoring, we concluded a final list of candidates for novel targets of editing that were experimentally tested through the 454 amplicon sequencing™. Although we can not detect any signs of editing at the predicted positions, we detect other "micro-editing" events within the same region. The lack of discrepancies of the predicted positions could be due to several things: The previous set of bona fide editing events is in fact a near complete assembly of ADAR targets; there is a very limited number of more site selective re-coding events to find. We could also be looking either in the wrong tissue and/or in a developmental stage where editing is regulated. Interestingly, we have shown that ADAR is present at these regions by the disclosure of A/G discrepancies that can not be explained by either sequencing or alignment errors. The other explanation is in line with previous suggestions that many genes are subjected to editing but with very low efficiency (Maas et al., 2003). If so, our screen, indicate that the low-efficiency or micro-editing is a real phenomona.

Paper IV

By 454 amplicon sequencing we sequenced most of the known ADAR targets with high resolution. In addition, we distinguished between four different developmental stages in order to detect timely regulation of A-to-I editing. In the experimental part, RNA from mouse brain was isolated from embryo day 15 and 19 and post natal day 2 and 19. For mice, day 19 is considered an adult. The subsequent 454 sequencing gave us in average 650 reads per developmental stage. Here, one read correspond to one transcript. This unprecedented resolution in compiling editing frequencies through out developmentally has never been presented before. In general, we could see developmental regulated pattern of increased editing essentially for all substrates but the GluR-B Q/R transcript(s) which seem edited close to 100% at all times. Also, due to the large sample size and the possibility to examine an individual transcript, we statistically evaluated

coupling of edited positions. We could see a pattern of coupled sites for the edited Adar2 and GluR-6 pre-mRNA. The apparent pattern of having "hot-spots" of edited A:s every ~12:th nucleotide was concluded to also be significantly coupled. Our interpretation of this phenomena is that multiple ADAR dimers bind in register and, if possible, deaminate adenosines synchronously. Also, consecutive binding of ADARs are coordinated from a principal binding site with high frequency of editing.

Future studies

Sensitizing the screen for novel editing events

In retrospect, the computational setup to detect novel A-to-I editing substrates (paper III) can be fine-tuned to be more sensitive. As is, we first of all did not curate the predicted stems other than the computational parameter cut-offs. As also seen in paper III, there is no big difference between energy as a function of the number of nucleotides of a stem from our candidates compared to the corresponding numbers for a folded random sequence of equal length. Secondly, we only applied our extended screen with the site scoring scheme on candidate editing sites that was located in a stem region above the conservation score 75. A future approach should score everything. Thirdly, I would like to include annotations in the genbank flat file of "translational discrepancy" in the extended scoring scheme. Lastly, I would like to apply the refined computational setup to scan a member of the plant kingdom for post transcriptional modifications of the RNA.

Extending the use of coupled editing to create a full model of the ADAR/RNA interface.

In our experimental and statistical approach to detect coupled patterns between edited sites on the pre-messenger of Adar2, we could for practical reasons only amplify one of the duplex strands. Here, we choose the strand with the -1 site that result in the alternative 3' splice site. The other strand is also heavily subjected to editing. An in-vitro assay with a construct with a significantly reduced loop insertion could, with high resolution sequencing, reveal additional data to include in a compilation of coupled sites. A similar idea on selected ALU repeat/inverted repeat fold back structures that are edited, would also strengthen these results.

Acknowledgements

My supervisor, a.k.a "bossen". More than six years ago it took you somewhere in the time span of nano and pico seconds to say "yes" when I stood on your door step wanting to pursue a computational degree project – hope you haven't regretted extending that time span…I have absolutely loved being in your group in the midst of biology and binary data.

Father Holger and Mother Kerstin. First for obvious reasons and also given the pre-natal certainty I was an XX specimen, I'm extremely happy you didn't stick with "Åsa"….Secondly for being the most generous and sacrificing parents imaginable.

My Stockholm family, Blörn, Fritjof, Stickan, Jens, Morfar, Malin, Tore osv. Both for always having a profound feeling of unity and for having such a compassionate and sensitive feeling for unsatisfactory levels of blood in the alcohol.

My Ludvika family – Christer, Nina, Olle, Sara, Maja.

My Haverö family – Kerstin, Tage, Bengt.

Olivia - After two and a half years you finally learned that what I do is "booooooring"….well it's an improvement over "Uhhh?" However, you gave me the finest gift possible…

Cissi - part of my first years at Molbio making them a joy both in and off curriculum. I will always be sad thinking we cannot pursue our deep friendship.

Bitte - my former co-supervisor. Also professor and head of department, all for a reason. You always seem to have a direct connection to "facit" when one talk to you.

Mormor Stina - you should have been here…seen my son and probably cried.

Boj – You would have been proud.

Bosse – For the generous and sincere interest in how his extended family are.

My new family – always kind and helpful, and the only ones that understands and nods in sympathy when I complain about Olivia's clothes on the floor…:)

Johan – There was not a living soul in the department that DIDN'T know you wore CK underwear. Good luck with the frogs.

Kicki, Nadja, Viktoria - old school, hope we have many years of friendship a head.

Helene, Danne – new school, sharp, sharp minds.

Beskow – Always fun to kid around with, but have to learn how to make a decent cup of coffee and make sure it is served on time.

Seger – Cheer up for god sake!

Gunnel – I will steal a cigarette from you at my dissertation party…

Pat – Great researcher but no humor unfortunately.

Lundin – I probably would have a different and less solid thesis if you hadn't been at the department.

Kerstin – Always a smile and a genuine interest in how my son is doing.

Lasse – Just call them C/D guide RNAs, then you don't have to bother about the location :)

Eva, Annika – Thanks for all the interesting discussions we had in the beginning.

Jacob, Jan – For introducing me to Perl and for all the help during visits in Copenhagen.

Neus, Uli, Ylva, Sara, Pärra, Petra, David, Solveig, Anna, Linus, Anu, Mari-Ann, Micke, Widad, Margareta, Shiva, Britta, Rula, Josefin, Masson and all other present and former co-workers at MolBio for making the department such a joyful place to conduct research at.

References:

Ambros V. 1989. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. *Cell 57*:49-57.

Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D. 2003. MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr Biol 13*:807-818.

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.

Barbon A, Vallini I, La Via L, Marchina E, Barlati S. 2003. Glutamate receptor RNA editing: a molecular analysis of GluR2, GluR5 and GluR6 in human brain tissues and in NT2 cells following in vitro neural differentiation. *Brain Res Mol Brain Res 117*:168-178.

Bartel B, Bartel DP. 2003. MicroRNAs: at the root of plant development? *Plant Physiol 132*:709-717.

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

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

Bass BL, Weintraub H. 1988. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell 55*:1089-1098.

Basyuk E, Suavet F, Doglio A, Bordonne R, Bertrand E. 2003. Human let-7 stem-loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res 31*:6593-6597.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. 1986. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell 46*:819-826.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature 409*:363-366.

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, Enright AJ, Dicks E, Futreal PA, Wooster R, Stratton MR. 2006. RNA editing of human microRNAs. *Genome Biol 7*:R27.

Bollman KM, Aukerman MJ, Park MY, Hunter C, Berardini TZ, Poethig RS. 2003. HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development 130*:1493-1504.

Bonnet E, Wuyts J, Rouze P, Van de Peer Y. 2004. Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci U S A 101*:11511-11516.

Brusa R, Zimmermann F, Koh DS, Feldmeyer D, Gass P, Seeburg PH, Sprengel R. 1995. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science 270*:1677-1680.

Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature 387*:303-308.

Cattaneo R, Schmid A, Eschle D, Baczko K, ter Meulen V, Billeter MA. 1988. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell 55*:255-265.

Cenci C, Barzotti R, Galeano F, Corbelli S, Rota R, Massimi L, Di Rocco C, O'Connell MA, Gallo A. 2008. Down-regulation of RNA editing in pediatric

astrocytomas: ADAR2 EDITING ACTIVITY INHIBITS CELL MIGRATION AND PROLIFERATION. *J Biol Chem 283*:7251-7260.
Chang FL, Chen PJ, Tu SJ, Wang CJ, Chen DS. 1991. The large form of hepatitis delta antigen is crucial for assembly of hepatitis delta virus. *Proc Natl Acad Sci U S A 88*:8490-8494.
Chen CX, Cho DS, Wang Q, Lai F, Carter KC, Nishikura K. 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 SH, Habib G, Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M, et al. 1987. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science 238*:363-366.
Chilibeck KA, Wu T, Liang C, Schellenberg MJ, Gesner EM, Lynch JM, MacMillan AM. 2006. FRET analysis of in vivo dimerization by RNA-editing enzymes. *J Biol Chem 281*:16530-16535.
Cho DS, Yang W, Lee JT, Shiekhattar R, Murray JM, Nishikura K. 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.
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.
Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature 432*:231-235.
Desterro JM, Keegan LP, Lafarga M, Berciano MT, O'Connell M, Carmo-Fonseca M. 2003. Dynamic association of RNA-editing enzymes with the nucleolus. *J Cell Sci 116*:1805-1818.
Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in Drosophila. *Genome Biol 5*:R1.
Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet 9*:102-114.
Förster T. 1948. Intermolecular energy migration and fluorescence. *Ann Physik 2*:55-75.
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.
Gandy SZ, Linnstaedt SD, Muralidhar S, Cashman KA, Rosenthal LJ, Casey JL. 2007. RNA editing of the human herpesvirus 8 kaposin transcript eliminates its transforming activity and is induced during lytic replication. *J Virol 81*:13544-13551.
Geiger JR, Melcher T, Koh DS, Sakmann B, Seeburg PH, Jonas P, Monyer H. 1995. Relative abundance of subunit mRNAs determines gating and Ca2+ permeability of AMPA receptors in principal neurons and interneurons in rat CNS. *Neuron 15*:193-204.
George CX, Samuel CE. 1999. 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.
Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R. 2004. The Microprocessor complex mediates the genesis of microRNAs. *Nature 432*:235-240.

Gromova I, Gromov P, Celis JE. 2002. bc10: A novel human bladder cancer-associated protein with a conserved genomic structure downregulated in invasive cancer. *Int J Cancer 98*:539-546.
Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell 35*:849-857.
Gurevich I, Tamir H, Arango V, Dwork AJ, Mann JJ, Schmauss C. 2002. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron 34*:349-356.
Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev 18*:3016-3027.
Hartner JC, Schmittwolf C, Kispert A, Muller AM, Higuchi M, Seeburg PH. 2004. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J Biol Chem 279*:4894-4902.
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.
Herbert A, Alfken J, Kim YG, Mian IS, Nishikura K, Rich A. 1997. A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proc Natl Acad Sci U S A 94*:8421-8426.
Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, Feldmeyer D, Sprengel R, Seeburg PH. 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature 406*:78-81.
Higuchi M, Single FN, Kohler M, Sommer B, Sprengel R, Seeburg PH. 1993. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell 75*:1361-1370.
Hollmann M, Heinemann S. 1994. Cloned glutamate receptors. *Annu Rev Neurosci 17*:31-108.
Hoopengardner B, Bhalla T, Staber C, Reenan R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science 301*:832-836.
Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science 293*:834-838.
Jaikaran DC, Collins CH, MacMillan AM. 2002. Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. *J Biol Chem 277*:37624-37629.
Jiang F, Ye X, Liu X, Fincher L, McKearin D, Liu Q. 2005. Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. *Genes Dev 19*:1674-1679.
Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell 14*:787-799.
Kallman AM, Sahlin M, Ohman M. 2003. ADAR2 A-->I editing: site selectivity and editing efficiency are separate events. *Nucleic Acids Res 31*:4874-4881.
Keegan LP, Brindle J, Gallo A, Leroy A, Reenan RA, O'Connell MA. 2005. Tuning of RNA editing by ADAR is required in Drosophila. *Embo J 24*:2183-2193.
Khvorova A, Reynolds A, Jayasena SD. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell 115*:209-216.
Kim J, Krichevsky A, Grad Y, Hayes GD, Kosik KS, Church GM, Ruvkun G. 2004. Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc Natl Acad Sci U S A 101*:360-365.
Kim U, Wang Y, Sanford T, Zeng Y, Nishikura K. 1994. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc Natl Acad Sci U S A 91*:11457-11461.
Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. 2004. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev 18*:1165-1178.

Kishore S, Stamm S. 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science 311*:230-232.
Klaue Y, Kallman AM, Bonin M, Nellen W, Ohman M. 2003. Biochemical analysis and scanning force microscopy reveal productive and nonproductive ADAR2 binding to RNA substrates. *RNA 9*:839-846.
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.
Krichevsky AM, King KS, Donahue CP, Khrapko K, Kosik KS. 2003. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA 9*:1274-1281.
Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ. 2004. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem 279*:42230-42239.
Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell 31*:147-157.
Kumar M, Carmichael GG. 1997. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc Natl Acad Sci U S A 94*:3542-3547.
Kurihara Y, Watanabe Y. 2004. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A 101*:12753-12758.
Kwak S, Kawahara Y. 2005. Deficient RNA editing of GluR2 and neuronal death in amyotropic lateral sclerosis. *J Mol Med 83*:110-120.
Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science 294*:853-858.
Lai EC. 2003. microRNAs: runts of the genome assert themselves. *Curr Biol 13*:R925-936.
Lai F, Chen CX, Carter KC, Nishikura K. 1997. Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Mol Cell Biol 17*:2413-2424.
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.
Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science 294*:858-862.
Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science 294*:862-864.
Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*:843-854.
Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature 425*:415-419.
Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *Embo J 23*:4051-4060.
Lehmann KA, Bass BL. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry 39*:12875-12884.
Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. 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, Shemesh R, Djinovic-Carugo K, Rechavi G, Jantsch MF, Eisenberg E. 2005. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res 33*:1162-1168.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell 120*:15-20.
Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. *Cell 115*:787-798.
Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev 17*:991-1008.
Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science 305*:1437-1441.
Liu Q, Rand TA, Kalidas S, Du F, Kim HE, Smith DP, Wang X. 2003. R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science 301*:1921-1925.
Liu Y, Emeson RB, Samuel CE. 1999. Serotonin-2C receptor pre-mRNA editing in rat brain and in vitro by splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase ADAR1. *J Biol Chem 274*:18351-18358.
Liu Y, George CX, Patterson JB, Samuel CE. 1997. Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase. *J Biol Chem 272*:4419-4428.
Liu Y, Lei M, Samuel CE. 2000. Chimeric double-stranded RNA-specific adenosine deaminase ADAR1 proteins reveal functional selectivity of double-stranded RNA-binding domains from ADAR1 and protein kinase PKR. *Proc Natl Acad Sci U S A 97*:12541-12546.
Liu Y, Samuel CE. 1996. Mechanism of interferon action: functionally distinct RNA-binding and catalytic domains in the interferon-inducible, double-stranded RNA-specific adenosine deaminase. *J Virol 70*:1961-1968.
Llave C, Xie Z, Kasschau KD, Carrington JC. 2002. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science 297*:2053-2056.
Lomeli H, Mosbacher J, Melcher T, Hoger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science 266*:1709-1713.
Luciano DJ, Mirsky H, Vendetti NJ, Maas S. 2004. RNA editing of a miRNA precursor. *RNA 10*:1174-1177.
Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. 2004. Nuclear export of microRNA precursors. *Science 303*:95-98.
Ma JB, Ye K, Patel DJ. 2004. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature 429*:318-322.
Maas S, Kim YG, Rich A. 2000. Sequence, genomic organization and functional expression of the murine tRNA-specific adenosine deaminase ADAT1. *Gene 243*:59-66.
Maas S, Patt S, Schrey M, Rich A. 2001. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc Natl Acad Sci U S A 98*:14687-14692.
Maas S, Rich A, Nishikura K. 2003. A-to-I RNA editing: recent news and residual mysteries. *J Biol Chem 278*:1391-1394.
Macbeth MR, Schubert HL, Vandemark AP, Lingam AT, Hill CP, Bass BL. 2005. Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science 309*:1534-1539.
Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell 15*:185-197.
Melcher T, Maas S, Herb A, Sprengel R, Higuchi M, Seeburg PH. 1996a. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J Biol Chem 271*:31795-31798.

Melcher T, Maas S, Herb A, Sprengel R, Seeburg PH, Higuchi M. 1996b. A mammalian RNA editing enzyme. *Nature 379*:460-464.
Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science 289*:920-930.
Niswender CM, Copeland SC, Herrick-Davis K, Emeson RB, Sanders-Bush E. 1999. RNA editing of the human serotonin 5-hydroxytryptamine 2C receptor silences constitutive activity. *J Biol Chem 274*:9472-9478.
Niswender CM, Herrick-Davis K, Dilley GE, Meltzer HY, Overholser JC, Stockmeier CA, Emeson RB, Sanders-Bush E. 2001. RNA editing of the human serotonin 5-HT2C receptor. alterations in suicide and implications for serotonergic pharmacotherapy. *Neuropsychopharmacology 24*:478-491.
Niswender CM, Sanders-Bush E, Emeson RB. 1998. Identification and characterization of RNA editing events within the 5-HT2C receptor. *Ann N Y Acad Sci 861*:38-48.
O'Connell MA, Krause S, Higuchi M, Hsuan JJ, Totty NF, Jenny A, Keller W. 1995. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Mol Cell Biol 15*:1389-1397.
Ohlson J, Pedersen JS, Haussler D, Ohman M. 2007. Editing modifies the GABA(A) receptor subunit alpha3. *RNA 13*:698-703.
Ohman M, Kallman AM, Bass BL. 2000. In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. *RNA 6*:687-697.
Ohta Y, Suzuki N, Nakamura S, Hartwig JH, Stossel TP. 1999. The small GTPase RalA targets filamin to induce filopodia. *Proc Natl Acad Sci U S A 96*:2122-2128.
Okamura K, Ishizuka A, Siomi H, Siomi MC. 2004. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev 18*:1655-1666.
Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. 2000a. dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. *RNA 6*:1004-1018.
Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. 2000b. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell 102*:437-449.
Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature 408*:86-89.
Patterson JB, Thomis DC, Hans SL, Samuel CE. 1995. Mechanism of interferon action: double-stranded RNA-specific adenosine deaminase from human cells is inducible by alpha and gamma interferons. *Virology 210*:508-511.
Patton DE, Silva T, Bezanilla F. 1997. RNA editing generates a diverse array of transcripts encoding squid Kv2 K+ channels with altered functional properties. *Neuron 19*:711-722.
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.
Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, Barbash ZS, Adamsky K, Safran M, Hirschberg A, Krupsky M, Ben-Dov I, Cazacu S, Mikkelsen T, Brodie C, Eisenberg E, Rechavi G. 2007. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res 17*:1586-1595.
Pham JW, Pellino JL, Lee YS, Carthew RW, Sontheimer EJ. 2004. A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in Drosophila. *Cell 117*:83-94.
Polson AG, Bass BL. 1994. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *Embo J 13*:5701-5711.
Polson AG, Bass BL, Casey JL. 1996. RNA editing of hepatitis delta virus antigenome by dsRNA-adenosine deaminase. *Nature 380*:454-456.

Poulsen H, Nilsson J, Damgaard CK, Egebjerg J, Kjems J. 2001. CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain. *Mol Cell Biol 21*:7862-7871.
Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA 10*:1507-1517.
Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature 403*:901-906.
Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. 2002. MicroRNAs in plants. *Genes Dev 16*:1616-1626.
Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. 2002. Prediction of plant microRNA targets. *Cell 110*:513-520.
Rhodes A, James W. 1990. Inhibition of human immunodeficiency virus replication in cell culture by endogenously synthesized antisense RNA. *J Gen Virol 71 (Pt 9)*:1965-1974.
Rosenthal JJ, Bezanilla F. 2002. Extensive editing of mRNAs for the squid delayed rectifier K+ channel regulates subunit tetramerization. *Neuron 34*:743-757.
Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature 399*:75-80.
Saller E, Tom E, Brunori M, Otter M, Estreicher A, Mack DH, Iggo R. 1999. Increased apoptosis induction by 121F mutant p53. *Embo J 18*:4424-4437.
Sansam CL, Wells KS, Emeson RB. 2003. Modulation of RNA editing by functional nucleolar sequestration of ADAR2. *Proc Natl Acad Sci U S A 100*:14018-14023.
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, O'Connell MA. 2005. Cleavage of dsRNAs hyper-edited by ADARs occurs at preferred editing sites. *Nucleic Acids Res 33*:5954-5964.
Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell 115*:199-208.
Seeburg PH, Higuchi M, Sprengel R. 1998. RNA editing of brain glutamate receptor channels: mechanism and physiology. *Brain Res Brain Res Rev 26*:217-229.
Sergeeva OA, Amberger BT, Haas HL. 2007. Editing of AMPA and serotonin 2C receptors in individual central neurons, controlling wakefulness. *Cell Mol Neurobiol 27*:669-680.
Shiohama A, Sasaki T, Noda S, Minoshima S, Shimizu N. 2003. Molecular cloning and expression analysis of a novel gene DGCR8 located in the DiGeorge syndrome chromosomal region. *Biochem Biophys Res Commun 304*:184-190.
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.
Stark A, Brennecke J, Russell RB, Cohen SM. 2003. Identification of Drosophila MicroRNA targets. *PLoS Biol 1*:E60.
Stefl R, Xu M, Skrisovska L, Emeson RB, Allain FH. 2006. Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure 14*:345-355.
Stephens OM, Haudenschild BL, Beal PA. 2004. The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. *Chem Biol 11*:1239-1250.
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A 101*:6062-6067.
Suh MR, Lee Y, Kim JY, Kim SK, Moon SH, Lee JY, Cha KY, Chung HM, Yoon HS, Moon SY, Kim VN, Kim KS. 2004. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol 270*:488-498.

Tabara H, Yigit E, Siomi H, Mello CC. 2002. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans. *Cell 109*:861-871.
Tang G. 2005. siRNA and miRNA: an insight into RISCs. *Trends Biochem Sci 30*:106-114.
Tang G, Reinhart BJ, Bartel DP, Zamore PD. 2003. A biochemical framework for RNA silencing in plants. *Genes Dev 17*:49-63.
Tanoue A, Koshimizu TA, Tsuchiya M, Ishii K, Osawa M, Saeki M, Tsujimoto G. 2002. Two novel transcripts for human endothelin B receptor produced by RNA editing/alternative splicing from a single gene. *J Biol Chem 277*:33205-33212.
Tomari Y, Du T, Haley B, Schwarz DS, Bennett R, Cook HA, Koppetsch BS, Theurkauf WE, Zamore PD. 2004. RISC assembly defects in the Drosophila RNAi mutant armitage. *Cell 116*:831-841.
Tonkin LA, Saccomanno L, Morse DP, Brodigan T, Krause M, Bass BL. 2002. RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. *Embo J 21*:6025-6035.
Travis MA, van der Flier A, Kammerer RA, Mould AP, Sonnenberg A, Humphries MJ. 2004. Interaction of filamin A with the integrin beta 7 cytoplasmic domain: role of alternative splicing and phosphorylation. *FEBS Lett 569*:185-190.
Wagner RW, Yoo C, Wrabetz L, Kamholz J, Buchhalter J, Hassan NF, Khalili K, Kim SU, Perussia B, McMorris FA, et al. 1990. Double-stranded RNA unwinding and modifying activity is detected ubiquitously in primary tissues and cell lines. *Mol Cell Biol 10*:5586-5590.
Valente L, Nishikura K. 2007. RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. *J Biol Chem 282*:16054-16061.
Wang Q, Khillan J, Gadue P, Nishikura K. 2000. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science 290*:1765-1768.
Wang XJ, Reyes JL, Chua NH, Gaasterland T. 2004b. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol 5*:R65.
Vaucheret H, Vazquez F, Crete P, Bartel DP. 2004. The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes Dev 18*:1187-1197.
Wong SK, Sato S, Lazinski DW. 2001. Substrate recognition by ADAR1 and ADAR2. *RNA 7*:846-858.
Wu L, Belasco JG. 2008. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol Cell 29*:1-7.
Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol 2*:E104.
Yang JH, Luo X, Nie Y, Su Y, Zhao Q, Kabir K, Zhang D, Rabinovici R. 2003. Widespread inosine-containing mRNA in lymphocytes regulated by ADAR1 in response to inflammation. *Immunology 109*:15-23.
Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K. 2006. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol 13*:13-21.
Yi R, Qin Y, Macara IG, Cullen BR. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev 17*:3011-3016.
Zeng Y, Cullen BR. 2005. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem 280*:27595-27603.
Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. 2004a. Single processing center models for human Dicer and bacterial RNase III. *Cell 118*:57-68.

Zhang XJ, He PP, Li M, He CD, Yan KL, Cui Y, Yang S, Zhang KY, Gao M, Chen JJ, Li CR, Jin L, Chen HD, Xu SJ, Huang W. 2004b. Seven novel mutations of the ADAR gene in Chinese families and sporadic patients with dyschromatosis symmetrica hereditaria (DSH). *Hum Mutat 23*:629-630.

Paper I

Published online October 27, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 19 ***e167***
doi:10.1093/nar/gni169

A method to find tissue-specific novel sites of selective adenosine deamination

Johan Ohlson, Mats Ensterö, Britt-Marie Sjöberg and Marie Öhman*

Department of Molecular Biology and Functional Genomics, Stockholm University, S-106 91 Stockholm, Sweden

Received June 14, 2005; Revised September 26, 2005; Accepted October 10, 2005

ABSTRACT

Site-selective adenosine (A) to inosine (I) RNA editing by the ADAR enzymes has been found in a variety of metazoan from fly to human. Here we describe a method to detect novel site-selective A to I editing that can be used on various tissues as well as species. We have shown previously that there is a preference for ADAR2-binding to selectively edited sites over non-specific interactions with random sequences of double-stranded RNA. The method utilizes immunoprecipitation (IP) of intrinsic RNA–protein complexes to extract substrates subjected to site-selective editing *in vivo*, in combination with microarray analyses of the captured RNAs. We show that known single sites of A to I editing can be detected after IP using an antibody against the ADAR2 protein. The RNA substrates were verified by RT–PCR, RNase protection and microarray. Using this method it is possible to uniquely identify novel single sites of selective A to I editing.

INTRODUCTION

Adenosine to inosine (A to I) RNA editing is known to change the sequence of specific pre-mRNAs in metazoans from fly to human. ADAR2, a member of the ADAR (adenosine deaminase that acts on RNA) family, deaminates A to I selectively within double-stranded RNAs (dsRNAs) interrupted by bulges, mismatches or loops [reviewed in (1)]. ADAR editing with low selectivity can also occur on completely dsRNA. This is a type of hyper-editing that has been found within introns and untranslated regions (UTRs) of mRNAs, preferentially in repetitive Alu sequences (2–6). Only a few site-selective ADAR substrates have been detected. In mammals, most selectively edited sites targeted by ADARs have been found in pre-mRNAs expressed in the central nervous system. The most prominent sites of selective editing are in mRNA coding for several subunits of the AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole) glutamate receptor (GluR). Editing of subunit B (GluR-B) results in altered receptor properties, changing receptor permeability to Ca^{2+} and the ability to recover after desensitization (7–9). In exon 11 the Q/R site is edited to nearly 100% giving rise to a codon change from glutamine (Q) to arginine (R). In exon 13 the edited R/G site causes an arginine (R) to glycine (G) codon change that is developmentally regulated. The dsRNA structure required for ADAR editing at these sites is formed by an inverted repeat located in the downstream intron [review by (10)]. Another prominent substrate for site-selective A to I editing is the transcript of the serotonin receptor 5-HT_{2C}. Transcripts encoding the 2C receptor subtype undergo A to I editing at 5 sites: A, B, C′, C and D situated in close proximity to each other (11). Editing alters the coding potential of the second intracellular loop, reducing the efficiency of the interaction between the receptor and the G protein. Most of the selectively edited sites have been found fortuitously as A to G changes when comparing cDNA with genomic sequence, since inosine is seen as guanosine in the process of reverse transcription. However, a significant amount of inosine has been found within the poly(A) fraction of cellular RNA in mouse brain (12).

Co-immunoprecipitation is a powerful tool to precipitate-specific protein complexes. Further, it has been widely used to study RNA–protein interactions. One example is the identification of target RNA for the Nova protein in mouse brain using an ultraviolet cross-linking and immunoprecipitation assay (13). In another more general approach to identify mRNA–protein complexes (mRNPs) called ribonomics, RNA targets were detected using antibodies to RNA-binding proteins followed by genomic arrays (14).

We have shown previously that ADAR binds more preferentially to selectively edited sites than to random sequences of dsRNA (15). Moreover, ADAR2 was shown to bind with a similar affinity to an editing substrate as to the product (16). Based on this knowledge we have developed a method to find novel ADAR substrates by extracting intrinsic ADAR2–RNA substrate complexes from mouse brain by co-immunoprecipitations using an anti-ADAR2 antibody. The specificity of this method has been verified by the detection

*To whom correspondence should be addressed. Tel: +46 8 16 44 51; Fax: +46 8 16 64 88; Email: marie.ohman@molbio.su.se

of known site-selectively edited substrates using RT–PCR, RNase protection and genomic microarray analyses. We present a powerful method with the potential to find novel sites of selective editing in different tissues and organisms.

MATERIALS AND METHODS

Isolation of RNA–protein complex from mouse brain

Three mouse brains were homogenized in HBSS [1× Hank's solution (HBSS GIBCO no. 14185-045)] and 1 M HEPES (pH 7.3) using a glass grinder. The suspension was washed in cold 1× HBSS and the pellet was frozen in liquid nitrogen. The pellet was resuspended in PXL [1× D-phosphate-buffered solution (PBS) (GIBCO no. 14200-67), 0.1% SDS, 0.5% deoxycholate and 0.5% NP-40] and ribonucleoside vanadyl complex (Sigma) on ice. The suspension was sonicated and treated with DNase I RQ1 (SIGMA). After centrifugation at 10 000 *g* for 20 min, 4°C, the supernatant was used for the immunoprecipitation (IP).

Immunoprecipitation of RNA–ADAR2 complexes

Anti-human ADAR2 antibody was made from recombinant histidine tagged human ADAR2 (hADAR2) protein, kindly provided by professor Brenda Bass' laboratory. The hADAR2 protein was concentrated using a centricon YM30 (Millipore) run out on 8% SDS–PAGE gel. The band corresponding to hADAR2 was excised and immunized four times into rabbits (Agrisera; Umeå Sweden). The serum was checked for immuno-reactivity and supplemented with 0.05% sodium azide.

To reduce non-specific binding prior to use in IPs the Sepharose A beads were incubated with tRNA (100 μg/ml) and BSA (100 μg/ml) in 1× PBS, washed once in 1× PBS and resuspended in 1 vol of 1× PBS and 0.05% NaN_3. The cell lysis extract from one mouse brain was pre-cleared with 50 μl of Sepharose A stock for 30 min at 4°C with rotation. The pre-cleared lysate was incubated with anti-ADAR2 polyclonal antibody or pre-immune serum for 2 h at 4°C with rotation. The lysate-antibody was mixed with 50 μl of prepared Sepharose A stock and incubated for 1 h at 4°C with rotation. The bead–antibody-lysate complex was rinsed three times in wash buffer containing 1× PBS, $MgCl_2$ (2 mM), EDTA (15 mM), NP-40 (1%) and Tween-20 (0.5%) including 1 protease Inhibitor Cocktail tablet/10 ml buffer (Roche) and rinsed once in 1× PBS, and eluted in 1× PBS plus 1% SDS at 65°C for 10 min.

Verification of ADAR2-binding using western blot

The IP eluate (10 μl) was boiled in SDS for 10 min prior to fractionation by electrophoresis on a 4–15% pre-made SDS–PAGE gel (BioRad) and transferred to a PVDF membrane by electroblotting. Anti-hADAR2 was used as primary antibody and anti-rabbit/HRP (DakoCytomation) was used as secondary antibody. The blots were developed using Amershams ECL plus Western Blotting Detection System and developed in a LAS 1000 system (Fujifilm).

Preparation of RNA after immunoprecipitation

The protein fraction was removed from the protein–RNA eluate after the IP by adding 1.8 mg of proteinase K (Roche) and incubated at 37°C for 15 min prior to a phenol/chloroform extraction and precipitation. The RNA was purified using RNeasy according to the manufacturer's instruction (Qiagen).

Microarray preparation

Preparation of labeled cRNA from the immunoprecipitated RNA was done according to Affymetrix Two-Cycle Target Labeling Assay. Labeled cRNA from nine mouse brains were hybridized to each Mouse Genome 430A 2.0 Array (Affymetrix). Scanning was performed after adding streptavidin-phycoerythrin Biotinylated anti-streptavidin antibody (SAPE) according to standard protocols Affymetrix Inc. (Santa Clara, CA).

Verification of known ADAR2 substrates using RT–PCR

The reverse transcription reactions were done with the Sensiscript RT kit (Qiagen) using hexanucleotide mix (Roche). A radioactive PCR using *taq* polymerase from Qiagen was performed for 25 cycles. Primers mGluRB-R/G-R (5′-GGGGAGTTCTATATTCTACGGC-3′), mGluRB-Q/R-R (5′-GACACCATGAATATCCACTTGAGACC-3′) and serotonin-R (5′-GGCCTTAGTCCGCGAATTGAACCGGC-3′) were radioactively labeled by T4 polykinase (invitrogen) using [γ-^{32}P]ATP (NEN Perkin Elmer). The following non-radioactive primers were also used in the different PCRs: mGluRB-R/G-F (5′-CCCACATTTCTGGCCCTTGTGCC-3′), mGluRB-Q/R-F (5′-TTTGCCTACATTGGGGTCAGTG-3′) and serotonin-F (5′-GTCCATCATGCACCTCTGCG-3′). The result was shown on a native 5% PAGE gel. As negative controls the acidic ribosomal protein P0 (ARPP P0) and GluR-A were amplified using primers ARPP P0-F (5′-GCACTGGAAGTCCAACTACTTC-3′), ARPP P0-R (5′-TGAGGTCCTCCTTGGTGAACAC-3′), mGluRA-F (5′-CCAGAGCTGGTGCTGGTCAGCTCTCG-3′) and mGluRA-R (5′-GAAGTATATACGACCACTGTCATC-3′). All primers were labeled with [γ-^{32}P]ATP as described above. For sequencing the R/G site, primer mGluRB-R/G-seq (5′-GGGCCAGTTCTCAAACTTCTCTGGCCCC-3′) was used.

Verification of known ADAR2 substrates using RNase protection

The RNase protection assay was done using Ribonuclease Protection Assay kit (RPA III no. 1414) from Ambion. Template RNA was immunoprecipitated from five mouse brains. To make the probe, the GluR-B was amplified by PCR using the mGluRB-R/G-F and mGluRB-R/G-R primers on genomic DNA from N2 cells, and the PCR product was ligated into the pGEM-T Easy vector (Promega). The vector (insert) was cut with HpaI (10 U, Invitrogen) and a uniformly labeled mGluRB-R/G probe was transcribed using SP6 RNA polymerase (30 U, invitrogen) in the buffer supplied by the provider in the presence of [α-^{32}P]UTP (NEN Perkin Elmer). The 225 nt long radioactive probe (GTTAACTCTTTGTATTCCTATTTTGTTGTTTGTTTATTTTTTAGTGGAGTCACATTCAAGACACTGTATTTGTTTGTTGTGGATGTGAGTACATTGCCGTAGAATATAGAACTCCCCA) is complementary to 118 nt of the GluR situated 698 nt downstream of the R/G site. The probe was purified on a 8% PAGE plus 7 M Urea gel. The assay was performed according to the manufacturer's instructions (Ambion).

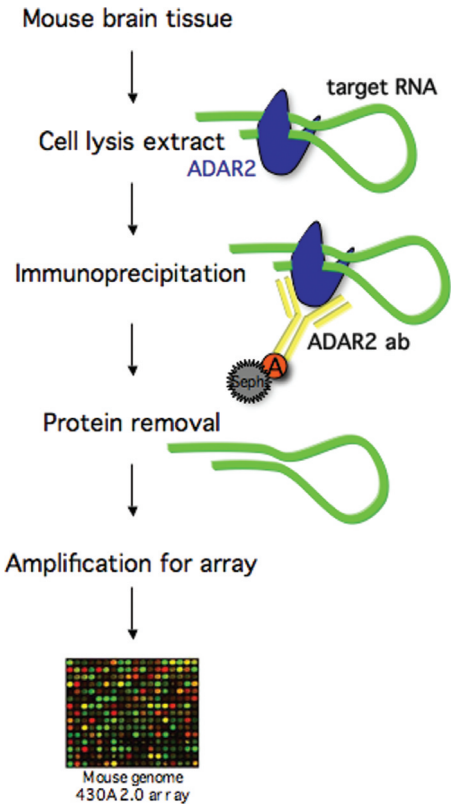


Figure 1. Illustration of the IP-array method to find novel substrates for A to I editing. Cell lysis extract was prepared from mouse brain. The extract was immunoprecipitated using an ADAR2-specific polyclonal antibody. Target RNA was extracted from the mRNP complexes upon protein removal. The RNA was amplified, labeled and further hybridized to a mouse genomic oligo array.

RESULTS

Specific enrichment of targets for site-selective editing

A method was developed to detect novel site-selective A to I editing *in vivo* (Figure 1). To identify ADAR2 associated mRNAs, cell lysate from mouse brain was incubated with anti-human ADAR2 polyclonal antibody. ADAR2–RNA complexes were pulled down using protein A–Sepharose beads. The co-purified pre-mRNAs were identified by probing of microarrays after removal of the proteins. ADAR proteins are known to bind tightly to dsRNA of any sequence (16,17,18). However, from previous studies we know that ADAR2 preferentially binds single sites of selective editing over a random sequence of completely dsRNA (15). This might be due to a higher affinity to site-selectively edited substrates. We therefore hypothesized that this method would specifically enrich RNA transcripts subjected to single sites of selective editing.

ADAR2 co-immunoprecipitation using mouse brain

Using this method it is important to retain intact RNA–protein complexes. Therefore, the cell extracts were treated with a ribonucleoside vanadyl complex to prevent RNA degradation prior to being used as load in the IP. DNA was also removed before further extractions to minimize non-specific

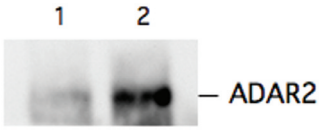


Figure 2. Western blot analysis using anti-human ADAR2 antibody on IP eluates. Three mouse brains were used for ADAR2-specific and pre-immune serum IP, respectively. One-tenth of the IP eluate was used for western blot. Lane 1 represents the pre-immune serum IP and lane 2 shows the amount of ADAR2 in the ADAR2-specific IP.

background. The specificity of the RNA–protein interaction was optimized by washing the immunoprecipitate three times in 1× PBS, $MgCl_2$ (2 mM), EDTA (15 mM), NP-40 (1%) and Tween-20 (0.5%) in presence of protease inhibitor. After SDS treatment the specificity of the IP for ADAR2 was determined by western blot (Figure 2). An enrichment of ADAR2 was seen when the anti-ADAR2 antibody was used in the IP compared with precipitation using pre-immune serum.

Specific amplification of known A to I editing substrates

GluR-B is a transcript that is A to I edited site-selectively at two sites (Q/R and R/G) within the coding sequence [reviewed in (10)]. Although some other receptor subunits are subjected to editing, no editing has been detected in the subunit A (GluR-A) transcript. Another well-known substrate for A to I editing is the transcript of the serotonin receptor 5-HT_{2C}. This transcript has been shown to be site-selectively edited at several sites (A, B, C, C′ and D) (11). The specificity of the IP for these known RNA targets was analyzed by semi-quantitative RT–PCR (Figure 3a and b). An enrichment of target substrates was observed in the ADAR2 IP when primers specific for the edited sites in the GluR-B and 5-HT_{2C} transcripts were used (Figure 3a). The pre-immune IP did not show an enrichment of target RNA. When primers specific for GluR-A were used for amplification no product could be detected during the 25 cycles of PCR considered to give a semi-quantitative product (Figure 3a). During an extended PCR to 30 cycles a product of equal amount could be detected in the ADAR2 and pre-immune IP (data not shown). As an additional negative control primers specific for the mRNA of the ribosomal phosphoprotein P0 that is not edited were used for amplification. No enrichment of this product could be detected as the level of transcripts appears to be equal in the ADAR2 and pre-immune IP elutes (Figure 3b). Editing at the Q/R and R/G sites of the target RNAs was verified by sequencing a population from RT–PCR (Figure 3c). Although a mixed population of edited and non-edited products was seen at the R/G site the Q/R site was edited to 100%. These data are in line with previous results showing the extent of GluR-B editing in the mammalian brain [reviewed in (10)]. From the sequencing analysis we can also verify that both pre-mRNA and mRNA of the GluR-B transcript is present in the specific IP. The GluR-B R/G site was amplified using primers specific for the pre-mRNA while the Q/R site was amplified from primers situated in the exons, giving a product from the spliced mRNA. The specificity for the GluR-B transcript in the ADAR2 IP was also verified by an RNase protection assay detecting an RNA from the ADAR2 IP but not from pre-immune IP (Figure 3d). Our data confirm that an RNA that

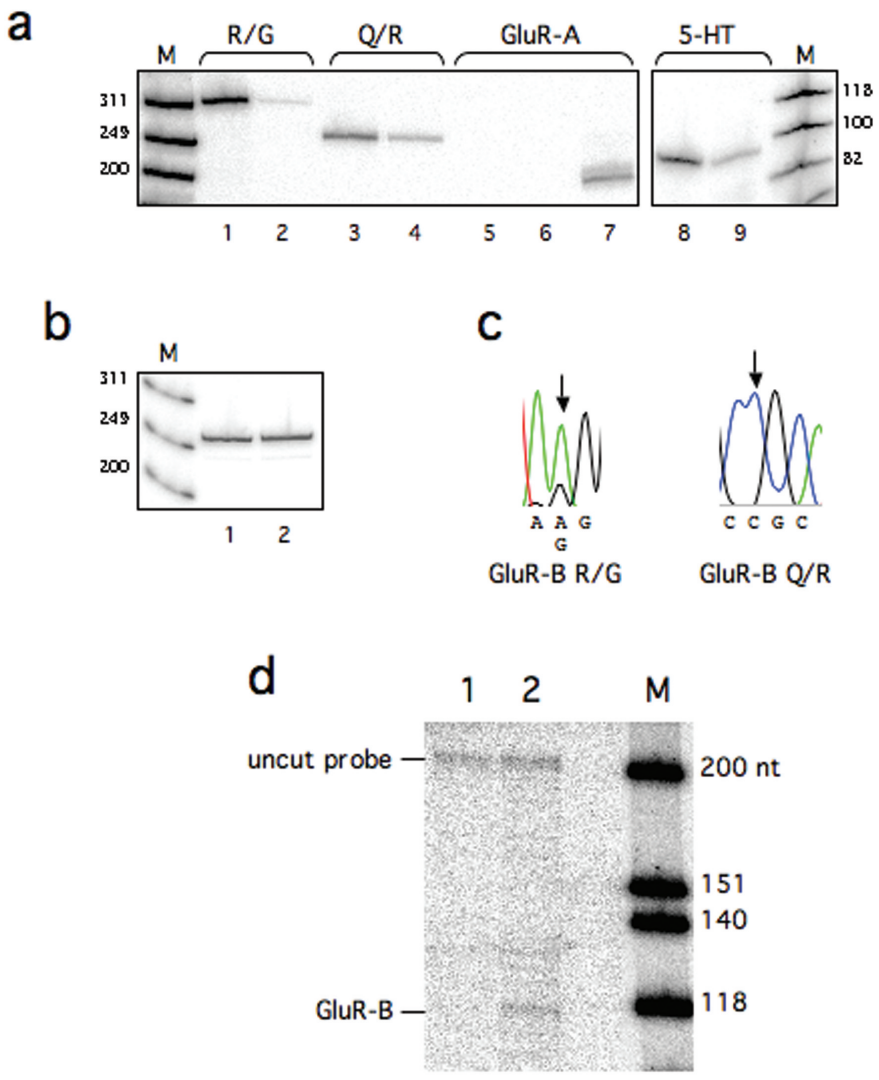


Figure 3. Detection of known substrates for A to I editing using ADAR2-specific IP. (**a**) Semi-quantitative RT–PCR on GluR-B at the R/G and Q/R site, GluR-A and the serotonin receptor (5-HT_{2C}) using radioactively labeled primers. Lane 1 shows the amplification of the R/G site from an ADAR2-specific IP, with an estimated size of 314 bp. Lane 2 shows a product amplified from the R/G site from an IP using pre-immune serum. Lane 3 shows the amplification of the Q/R site from an ADAR2-specific IP, with an estimated size of 253 bp. Lane 4 shows product amplification of the Q/R site from an IP using pre-immune serum. RT–PCR on GluR-A, lacking sites for A to I editing, shows no detectable amplification from an ADAR2-specific IP (lane 5), or a pre-immune serum IP (lane 6). Product using total RNA is shown in lane 7 and the estimated size is 203 bp. RT–PCR on 5-HT_{2C} shows the amplification from an ADAR2-specific IP (lane 8), the estimated size is 94 bp. Lane 9 shows a product amplified from the 5-HT_{2C} transcript from an IP using pre-immune serum. Lane M is a size marker with bands of sizes as indicated. (**b**) RT–PCR on the ribosomal phosphoprotein P0, lacking sites for A to I editing. No enrichment could be detected in the ADAR2 IP (lane 1) compared with the pre-immune serum IP (lane 2). The estimated size is 265 bp. Lane M is a size marker with bands of sizes as indicated. (**c**) The product from the RT–PCR-specific for the R/G and the Q/R sites were DNA sequenced to determine the editing efficiency. At the R/G site a forward primer was used to give a dual A and G peak at the R/G site. At the Q/R site a reverse primer was used so that an edited site is a C in the sequence. Edited nucleotides are indicated with an arrow in the chromatogram. (**d**) An RNase protection assay was used to confirm the enrichment of GluR-B in the presence of anti-ADAR2 antibody. A 225 nt long α-^{32}P-labeled probe, 118 nt complementary downstream of the R/G site, was hybridized to RNA from an IP using pre-immune serum (lane 1) and to RNA from an IP using anti-ADAR2 antibody (lane 2).

is edited can be specifically enriched from a mammalian brain tissue using an anti-ADAR2 antibody in IPs.

Detection of A to I editing targets using microarray

After protein removal from the IP using proteinase K treatment and phenol/chloroform extraction the RNA was amplified and labeled according to Two-Cycle Target Labeling assay (Affymetrix). The cRNA was hybridized to a mouse genome array 430A 2.0 (Affymetrix) to detect enriched ADAR2–RNA targets compared with IPs using pre-immune serum. Three arrays from three independent target extractions were done. The diagram in Figure 4 illustrates the extent of enrichment of the 200 genes that are significantly enriched in all three arrays. The GluR-B transcript was significantly amplified in the three arrays and is indicated in red. Other known A to I substrates enriched in the specific IPs are specified in Table 1. Although present, the GluR-A transcript did not show an increase in the microarray (Table 1). This is in agreement with the presented data from RT–PCR and RNase protection. These results indicate that the method specifically amplifies known selectively edited targets that can be detected by microarray.

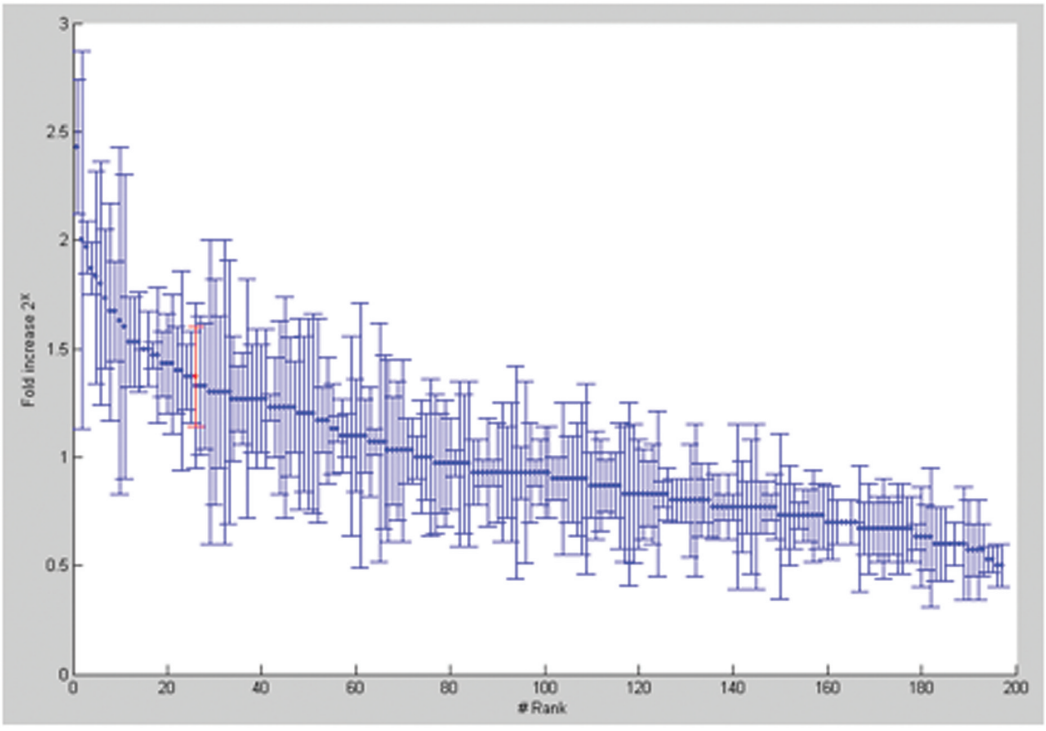


Figure 4. Genes enriched in ADAR2 IP compared with pre-immune serum IP. 200 genes were significantly increased in all three different arrays. The mean value for the three arrays is shown as fold increase 2^x. The GluR-B, marked in red, is ranked 25 of the 200 enriched genes.

Table 1. Enriched known editing targets and non-edited transcripts verified by microarray

Transcript	Mean (2^x-fold increase)	SD
Known editing targets		
GluR-B	1.37	0.23
5-HT_{2C}	1.03	1.63
Ednrb	0.97	1.02
Igfbp	0.50	0.10
Blcap	0.33	0.12
ADAR2	0.33	1.20
Non-edited transcripts		
GluR-A	−0.17	0.11
ARPP P0	−0.50	0.20

The following abbreviations are used: GluR-B, Glutamate receptor subunit B; 5-HT_{2C}, serotonin receptor subtype 2C; Ednrb, Endothelin receptor type B; Igfbp, insulin-like growth factor-binding protein 2; Blcap, bladder cancer associated protein; ADAR2, adenosine deaminase that acts on RNA 2; GluR-A, Glutamate receptor subunit A; ARPP P0, acidic ribosomal phophoprotein P0.

DISCUSSION

During the past decade several methods have been developed to find new ADAR substrates. By computational analysis a vast amount of edited sites have been detected in 5′- and 3′-UTRs within Alu repetitive elements that are hyper-edited at multiple sites, but very few sites were found in coding sequences (4–6). Although editing of Alu repeats might be important, no function has so far been proposed.

We have developed a method to detect single sites of A to I editing *in vivo* and have chosen mouse brain in our initial experiments. The mouse genome contains fewer repetitive elements than the human genome and lacks the Alu repeats. By choosing ADAR2 and mouse material we can focus on single sites of editing in coding sequences, with the potential of creating alternative isoforms of the protein. Mouse is therefore a good model organism to avoid extensive A to I hyper-editing of non-coding sequence.

Most dsRNA-binding proteins (dsRBPs) interact with the RNA by sequence-specific structural features rather than base-specific interactions [reviewed in (19)]. A dsRNA-binding motif makes at least two structure-specific interactions with the RNA double-helix. These interactions have been proven to occur in a sequence-independent manner (20,21). However, it has been proposed by us and others that the mismatch opposing the R/G site in GluR-B serves as a structural feature in concert with the neighboring nucleotides to direct site-selective editing (18,22). Further, studies on other dsRBPs indicate that there are regions in the RNA-binding motif that interact with RNA loop structures in the vicinity of the helical structure (23–26). These results are in line with our previous result indicating that the ADAR2 enzyme discriminate between a completely dsRNA structure and a selectively edited substrate interrupted by bulges and loops, possibly with a slower off rate on the latter sites (15). To minimize the background binding to dsRNA we exclude any form of cross-linking between RNA and protein prior to the IP.

Using our approach we can collect potential ADAR substrates *in vivo* and enrich for selectively edited sites. Using microarray analysis as the method to detect potential targets allows us to tolerate a certain amount of background but also to detect products of relatively low abundance since the material is amplified prior to the array. However it should be noted that the microarray is limited in its detection of enriched transcripts. Table 1 shows the enrichment of known edited

substrates. Most of the known transcripts subjected to A to I editing show a significant enrichment in the microarray after the ADAR2-specific IP. However, in order to get a better statistical value on the microarray an increased number of independent array analyses are required. The enrichment of edited substrates in the specific IP was verified by semi-quantitative RT–PCR on a selective set of RNAs using primers specific for the GluR-B R/G site, the GluR-B Q/R site and the A–D sites in the 5-HT_{2C} transcript. Using this technique we could detect an enrichment of RNA containing all of these sites but not for GluR-A and ARPP P0 transcripts that are not edited. We are therefore confident that edited substrates indeed are enriched in the specific IP. From sequencing analysis of the PCR products from the amplified edited transcripts we can detect both pre-mRNA and mRNA. Detection of spliced transcripts indicate that splicing has occurred subsequent to binding during the IP. Since ADAR2 has been shown to bind to the inosine containing product with almost the same affinity as to the substrate (16), we expect that edited as well as non-edited A to I substrates are extracted using this assay. Approximately 200 genes showed a significant increase in all three arrays compared with microarrays based on RNA form an IP using pre-immune serum (Figure 4). We are using computational analysis to identify the position of editing sites in the candidate genes. When a computational search on an entire genome is used as the sole method to identify A to I editing it is hard to detect single sites of selective editing in the background of single nucleotide polymorphisms, sequencing errors and mis-alignments. Since we utilize the candidates identified in the experimental setup as the input the computational search can be more general. Three main criteria are used to get a high score on editing probability: (i) A/G mismatches between genomic and cDNA sequence, (ii) phylogenetic conservation of the A/G mismatch between mammals and (iii) inverted repeats with acceptance of mismatches and internal loops (M. Ensterö, B.-M. Sjöberg and M. Öhman, manuscript in preparation). Each criterium is scored individually and high score candidates are verified experimentally. This unique combination of experimental and bioinformatical analysis has the potential to detect novel sites of selective editing that have previously been foreseen using the methods separately. We have detected several new candidates of A to I editing substrates in mouse brain using this strategy (J. Ohlson, M. Ensterö, B.-M. Sjöberg and M. Öhman, manuscript in preparation).

Our approach has numerous applications, it can be used to find novel editing substrates in different tissues as well as to identify editing discrepancies between different species. It is also possible to apply this method on other ADAR protein family members like ADAR1 but also ADAR3, so far without known targets, as well as on other dsRBPs. A to I editing is an essential event for normal brain function (27). Several diseases with altered brain functions have been shown to have an effect on specific sites of editing (28,29). Our method has a potential to give a more general overview of the editing events in a normal brain compared with a diseased one.

ACKNOWLEDGEMENTS

We thank Lars Wieslander for helpful discussions and comments on the manuscript. We are grateful to Ann-Kristin Östlund Farrants and Patrick Asp and the Wennergren Institute, WGI, Stockholm University for technical assistance. We also thank the Affymetrix core facility at the Karolinska Institute, Novum. This work was supported by grants from Wallenberg consortium North. Funding to pay the Open Access publication charges for this article was provided by Wallenberg consortium North.

Conflict of interest statement. None declared.

REFERENCES

1. Bass,B.L. (2002) RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.*, **71**, 817–846.
2. Morse,D.P. and Bass,B.L. (1999) Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)+ RNA. *Proc. Natl Acad. Sci. USA*, **96**, 6048–6053.
3. Morse,D.P., Aruscavage,P.J. and Bass,B.L. (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. USA*, **99**, 7906–7911.
4. Levanon,E.Y., Eisenberg,E., Yelin,R., Nemzer,S., Hallegger,M., Shemesh,R., Fligelman,Z.Y., Shoshan,A., Pollock,S.R., Sztybel,D. *et al.* (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.*, **22**, 1001–1005.
5. Blow,M., Futreal,P.A., Wooster,R. and Stratton,M.R. (2004) A survey of RNA editing in human brain. *Genome Res.*, **14**, 2379–2387.
6. Athanasiadis,A., Rich,A. and Maas,S. (2004) Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.*, **2**, e391.
7. Hume,R.I., Dingledine,R. and Heinemann,S.F. (1991) Identification of a site in glutamate receptor subunits that controls calcium permeability. *Science*, **253**, 1028–1031.
8. Burnashev,N., Monyer,H., Seeburg,P.H. and Sakmann,B. (1992) Divalent ion permeability of AMPA receptor channels is dominated by the edited form of a single subunit. *Neuron*, **8**, 189–198.
9. Lomeli,H., Mosbacher,J., Melcher,T., Hoger,T., Geiger,J.R., Kuner,T., Monyer,H., Higuchi,M., Bach,A. and Seeburg,P.H. (1994) Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science*, **266**, 1709–1713.
10. Seeburg,P.H., Higuchi,M. and Sprengel,R. (1998) RNA editing of brain glutamate receptor channels: mechanism and physiology. *Brain Res. Brain. Res. Rev.*, **26**, 217–229.
11. Burns,C.M., Chu,H., Rueter,S.M., Hutchinson,L.K., Canton,H., Sanders-Bush,E. and Emeson,R.B. (1997) Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature*, **387**, 303–308.
12. Paul,M.S. and Bass,B.L. (1998) Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J.*, **17**, 1120–1127.
13. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
14. Tenenbaum,S.A., Lager,P.J., Carson,C.C. and Keene,J.D. (2002) Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods*, **26**, 191–198.
15. Klaue,Y., Källman,A.M., Bonin,M., Nellen,W. and Öhman,M. (2003) Biochemical analysis and scanning force microscopy reveal productive and non-productive ADAR2 binding to RNA substrates. *RNA*, **9**, 839–846.
16. Öhman,M., Källman,A.M. and Bass,B.L. (2000) *In vitro* analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. *RNA*, **6**, 687–697.
17. Lai,F., Drakas,R. and 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.
18. Stephens,O.M., Yi-Brunozzi,H.Y. and Beal,P.A. (2000) Analysis of the RNA-editing reaction of ADAR2 with structural and fluorescent analogues of the GluR-B R/G editing site. *Biochemistry*, **39**, 12243–12251.
19. Fierro-Monti,I. and Mathews,M.B. (2000) Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem. Sci.*, **25**, 241–246.

20. Nanduri,S., Carpick,B.W., Yang,Y., Williams,B.R. and Qin,J. (1998) Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation. *EMBO J.*, **17**, 5458–5465.
21. Ryter,J.M. and Schultz,S.C. (1998) Molecular basis of double-stranded RNA–protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.*, **17**, 7505–7513.
22. Källman,A.M., Sahlin,M. and Öhman,M. (2003) ADAR2 A→I editing: site selectivity and editing efficiency are separate events. *Nucleic Acids Res.*, **31**, 4874–4881.
23. Chanfreau,G., Buckle,M. and Jacquier,A. (2000) Recognition of a conserved class of RNA tetraloops by *Saccharomyces cerevisiae* RNase III. *Proc. Natl Acad. Sci. USA*, **97**, 3142–3147.
24. Ramos,A., Grunert,S., Adams,J., Micklem,D.R., Proctor,M.R., Freund,S., Bycroft,M., St Johnston,D. and Varani,G. (2000) RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J.*, **19**, 997–1009.
25. Nagel,R. and Ares,M.,Jr (2000) Substrate recognition by a eukaryotic RNase III: the double-stranded RNA-binding domain of Rnt1p selectively binds RNA containing a 5′-AGNN-3′ tetraloop. *RNA*, **6**, 1142–1156.
26. Leulliot,N., Quevillon-Cheruel,S., Graille,M., Van Tilbeurgh,H., Leeper,T.C., Godin,K.S., Edwards,T.E., Sigurdsson,S.T., Rozenkrants,N., Nagel,R.J. *et al.* (2004) A new alpha-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III. *EMBO J.*, **23**, 2468–2477.
27. Higuchi,M., Maas,S., Single,F.N., Hartner,J., Rozov,A., Burnashev,N., Feldmeyer,D., Sprengel,R. and Seeburg,P.H. (2000) Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature*, **406**, 78–81.
28. Akbarian,S., Smith,M.A. and Jones,E.G. (1995) Editing for an AMPA receptor subunit RNA in prefrontal cortex and striatum in Alzheimer's disease, Huntington's disease and schizophrenia. *Brain Res.*, **699**, 297–304.
29. Sodhi,M.S., Burnet,P.W., Makoff,A.J., Kerwin,R.W. and Harrison,P.J. (2001) RNA editing of the 5-HT(2C) receptor is reduced in schizophrenia. *Mol. Psychiatry*, **6**, 373–379.

Paper II

Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 249–254
www.elsevier.com/locate/compbiolchem

MicroRNA sequence motifs reveal asymmetry between the stem arms

J. Gorodkin [a,*,1], J.H. Havgaard [a,1], M. Ensterö [b], M. Sawera [a], P. Jensen [a], M. Öhman [b], M. Fredholm [a]

[a] *Division of Genetics and Bioinformatics, IBHV and Center for Bioinformatics, The Royal Veterinary and Agricultural University, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*
[b] *Department of Molecular Biology and Functional Genomics, University of Stockholm, SE-106 91 Stockholm, Sweden*

Received 17 April 2006; accepted 24 April 2006

Abstract

The processing of micro RNAs (miRNAs) from their stemloop precursor have revealed asymmetry in the processing of the mature and its star sequence. Furthermore, the miRNA processing system between organism differ. To assess this at the sequence level we have investigated mature miRNAs in their genomic contexts. We have compared profiles of mature miRNAs within their genomic context of the 5′ and 3′ stemloop precursor arms and we find asymmetry between mature sequences of the 5′ and 3′ stemloop precursor arms. The main observation is that vertebrate organisms have a characteristic motif on the 5′ arm which is in contrast to the 3′ arm motif which mainly show the conserved U at the position of the mature start. Also the vertebrate 5′ arm motif show a semi-conserved G 13 nucleotides upstream from the first position. We compared the 5′ and 3′ arm profiles using the average log likelihood ratio (ALLR) score, as defined by Wang and Stormo (2003) [Wang T., Stormo, G.D., 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 2369–2380.] and computing a *p*-value we find that the two profiles differs significantly in their 3′ end where the 5′ arm motif (in contrast to the 3′ arm motif) has a semi-conserved GU rich region. Similar findings are also obtained for other organisms, such as fly, worm and plants. The observed similarities and differences between closely and distantly related organisms are discussed and related to current knowledge of miRNA processing.

Keywords: miRNA; Mature miRNA organization; 5′ and 3′ hairpin arms; Sequence logos; Sequence profiles

1. Introduction

Mature miRNA forms a ∼ 22 nucleotide RNA duplex together with its star sequence, miRNA*, and is processed out in an asymmetric fashion from its stemloop precursor structure (reviewed by Bartel, 2004).

The asymmetry results from the two (metazoan) processing steps conducted by the nuclear Drosha and the cytoplasmic Dicer. Both these RNase III endonucleases act on different precursor signals. Drosha is thought to interact with the hairpin apex loop and cuts the hairpin near the terminal base, thus defining one end of the mature miRNA (Zeng and Cullen, 2004; Lee et al., 2003). The Drosha processing of a hairpin structure is further coordinated by Pasha which have two RNA binding motifs, a homolog to the mammalian DGCR8 (DiGeorge syndrome chromosomal region 8) (Denli et al., 2004; Gregory et al., 2004). Dicer is acting via its PAZ-domain which is known to interact with the 2 nucleotide 3′ overhang (Ma et al., 2004). Dicer then cuts away the loop subsequently defining the miRNA::miRNA* duplex.

Dicer has been shown to associate with a variety of different proteins including another highly conserved group—the Argonaute family which also share the PAZ-domain (Hammond et al., 2001). The RISC (RNA induced silencing complex) is a multi-protein complex and the understanding of the biogenesis scheme from miRNA::miRNA* duplex to final single stranded mature miRNA is not fully delineated. The Ago2 has been shown to be the actual slicer within the RISC (Meister et al., 2004). While siRNA specifically degrade their target through Ago2 miRNAs is believed to mainly interact with the first ∼ 7 nucleotides hence diversifying the target repertoire. Another important RISC component for fine tuning strand selection is R2D2 (*Drosophila melanogaster*) which can sensor the different thermodynamic inequalities for accurate strand incorporation (Tomari et al., 2004).

It has been observed that miRNAs are less stable in the 5′ end than in their 3′ end (5′ end of the star sequence)

* Corresponding author. Tel: +45 3528 3578; fax: +45 3528 3042.
E-mail address: gorodkin@bioinf.kvl.dk (J. Gorodkin).
[1] These authors contributed equally to this work.

1476-9271/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2006.04.006

Table 1
The table gives an overview of the miRNAs used as well as the distribution of the left and right matures

Org	Genome	DB	Hair	U-hair	5′arm	3′arm	Mat	$5'_{red}$arm	$3'_{red}$arm
hsa	NCBI35	332	332	332	203	191	394	122	119
mml	MMUL0.1	71	63	62	44	23	67	32	20
mmu	NCBIM34	270	276	267	159	148	307	101	101
rno	RGSC3.4	234	228	228	138	116	254	92	88
gga	WASHUC1	144	144	144	88	64	152	42	40
dre	WTSIZv5	372	335	310	149	198	347	49	45
fru	FUGU4	131	132	130	68	65	133	36	39
tni	TETRAODON7	131	142	131	72	70	142	35	39
dme	BDGP4.0	78	78	78	34	51	85	29	29
dps	DPSE2.0	73	73	73	28	46	74	25	28
cbr	cb25.agp8	79	82	79	26	56	82	22	37
cel	WS140	114	114	113	41	75	116	30	55
ath	Refseq[a]	117	117	117	62	57	119	28	21
osa	TIGR3.0	178	124	123	62	62	124	20	21

Org, the organism; genome, the release (see text for details); DB, the number of hairpins in the miRNA database; hair, the number of hairpins with genome coordinates; U-hair, the number unique hairpins with genome coordinates; 5′ arm (3′ arm, respectively), the number mature miRNAs on the 5′ arm of (3′ arm, respectively) of the hairpin; mat, the number of mature sequences with genome coordinates; $5'_{red}$arm ($3'_{red}$arm, respectively), the number of mature miRNAs with genome coordinates on the 5′ arm (3′ arm, respectively) left after similarity reduction (see text for details).

[a] Consist of GenBank accessions: NC_003070.5, NC_003071.3, NC_003074.4, NC_003075.3, NC_003076.4.

(Schwarz et al., 2003; Khvorova et al., 2003; Krol et al., 2004) and that the molecular processing machinery can sensor this (Tomari et al., 2004). Also, recent findings for intronic miRNAs in zebrafish suggest a non-canonical asymmetry in the process of strand selection acting concurrently with thermodynamical properties (Lin et al., 2005). Here, we further investigate this asymmetry and show that the organization in the genomic sequence context is asymmetric with respect to the mature sequence in the 5′ and 3′ arms of the stemloop precursor. This organization is similar for related organisms, but different for distantly related organisms.

2. Materials and methods

2.1. Data

Organisms represented in mirBASE version 8.0 (Griffiths-Jones et al., 2005) was extracted in their genomic contexts and only miRNA hairpins with genome gff coordinates was used. All coordinates were checked by comparing the extracted sequence and the sequence in the registry. Hairpins for which genomic coordinates were not given were ignored. One hairpin from *cel* was removed as it had no mature sequence annotated. For each organism the sequence data was divided into two sets one containing the mature sequences on the 5′ arm in the stemloop precursor and one where the mature sequences are on the 3′ arm in the stemloop precursor. For stemloops containing mature sequences on both the 5′ and 3′ arms, the mature sequences were used in their respective contexts. The number of such cases is in general low.

Furthermore, we made similarity reduced sets by grouping the sequences into families by the nucleotides 2–8 of the mature sequences, using only one sequence from each family (Lewis et al., 2005). Only organisms with at least 20 sequences left for both the 5′ and 3′ arms were included in the data set. These are: *Arabidopsis Thaliana* (*ath* (Arabidopsis Genome Initiative, 2000)), *Caenorhabditis briggsae* (*cbr* (C. elegans Sequencing Consortium, 2006)), *Caenorhabditis elegans* (*cel* (C. elegans Sequencing Consortium, 1998)), *Drosophila melanogaster* (*dme* (Celniker et al., 2002)), *Drosophila pseudoobscura* (*dps* (Richards et al., 2005)), *Danio rerio* (*dre* (The Zebrafish Sequencing Group, 2006)), *Fugu rubripes* (*fru* (Aparicio et al., 2002)), *Gallus Gallus* (*gga* (Int. Chicken Genome Sequencing Consortium, 2004)), *Homo Sapiens* (*hsa* (Int. Human Genome Sequencing Consortium, 2004)), *Macaca mulatta* (*mml* (HGSC at Baylor College of Medicine, 2006)), *Mus Musculus* (*mmu* (Mouse Genome Sequencing Consortium, 2002)), *Oryza sativa* (*osa* (Yuan et al., 2003)), *Rattus norvegicus* (*rno* (Rat Genome Sequencing Project Consortium, 2004)) and *Tetraodon nigroviridis* (*tni* (Jaillon et al., 2004)). The miRNA sequences ("hairpins" as in mirBASE) were then matched with their genomic context, and whole segments typically of 3000 nucleotides were extracted. The details of the data are listed in Table 1.

2.2. Sequence profiles

To construct sequence profiles the miRNAs along with their surrounding genomic context, were aligned by the start of their mature sequence. For each of the considered organisms this was done for the 5′ and 3′ arm mature sequences, respectively. Next, sequence logos (Schneider and Stephens, 1990) were generated by computing the relative entropy as by Gorodkin et al. (1997), with nucleotide frequencies computed for each position of the aligned sequence. Briefly, the information content for each position in the alignment is defined as $I = \sum_l q_l \log_2 q_l/p_l$, where l belong to the set of nucleotides. The fraction q_l is the observed nucleotide distributions, whereas the fraction p_l is the expected (background) nucleotide frequencies drawn from the miRNA hairpin excluding the mature sequence. For each position in the logo the correspond to the information content I, and the height of the letter l is the portion $q_l I$. When $q_l < p_l$ letter l is displayed upside down.

2.3. Comparing distributions

To compare the significance of the difference between 5′ and 3′ arm motifs the corresponding weight matrices (profiles) from the sequence logos were stored for computing the average log-likelihood ratio (ALLR) score as defined by Wang and Stormo (2003). This measure can be used to distinguish two corresponding columns from each weight matrix. It is the joint probability of observing the data generated by one distribution given the likelihood ratio of the other distribution. The ALLR score is a log-likelihood test of how one data set fits into another and vice versa. It is the average of the two log-likelihood ratios. When the data sets are unrelated the ALLR is expected to be negative. For details, see Wang and Stormo (2003).

Here, we compute the ALLR score for the two profiles (5′ arm and 3′ arm, respectively) when aligning them up- and downstream from the beginning of the mature sequence. For this fixed alignment we compute the ALLR score across several different regions. When comparing the two profiles, an ALLR score is computed over the corresponding regions of the two profiles. One of the nice features of the ALLR score is that it takes into account that the profiles compared can be made from different numbers of sequences.

Empirical *p*-values for significance of the obtained ALLR scores are computed in a given region by keeping the columns of a window from the 5′ arm (3′ arm, respectively) fixed while shuffling the columns (100 times) in the 3′ arm window (5′ arm, respectively) and for each shuffling, computing the ALLR score. The rank of the true ALLR score gives the empirical *p*-value.

3. Results

The data sets for the organisms considered here are shown in Table 1, where we observe the following: for the non-reduced data sets of the organisms *has, mml, mmu, rno, gga* that there seems to be a slight over representation of 5′ arm mature miRNAs. In contrast for fly and worm the over representation seems to be for the 3′ arm miRNAs. For plants, the number on both arms appears to be the same. The latter have also been noticed in by others (Bartel and Bartel, 2003).

However, here we focus on the similarity reduced sets of mature miRNAs in the 5′ and 3′ arms and unless mentioned otherwise we refer to this set. Results similar to those presented for reduced sets are obtained on the full non-reduced 5′ and 3′ arm data sets (not shown). For each organism we constructed profiles of the 5′ and 3′ arms with the precursor in the genomic context as described in Section 2. The profiles can be represented by sequence logos (Schneider and Stephens, 1990) as shown in Fig. 1 for the human case. The corresponding profiles for the organisms listed in Table 1 are shown in the supplementary material Figure S1. Note that the first position of the mature miRNA is indicated as position zero in the logos.

By inspection, we observe that all organism profiles show different characteristics between their 5′ and 3′ arm motifs. For the vertebrate organisms (*hsa, mml, mmu, rno, gga, dre, fru, tni*) we observe that they have a characteristic motif on the 5′ arm. In contrast, the 3′ arm motif essentially only displays the well

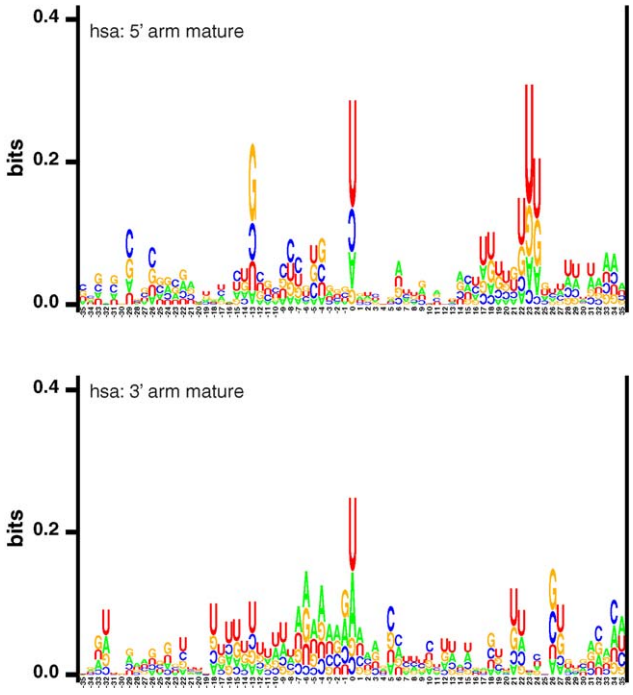


Fig. 1. The sequence logos of the 5′ (top) and 3′ (bottom) arms of the human miRNAs in their genomic context. Position zero corresponds to mature sequence start. The 5′ arm logo was generated from 122 sequences and the 3′ arm logo from 119 sequences. Letter sizes are shown according to their frequencies. (Upside-down is less than the expected frequency.).

known conserved U at the start of the mature sequence. For the invertebrates organisms flies and worms (*dme, dps, cel, cbr*) the 3′ arm motif is more characteristic showing a highly conserved U at the mature start. For plants (*ath* and *osa*) both the 5′ and the 3′ arm motifs show characteristic, but different motifs, the 5′ arm motif having a strongly conserved U at the mature and the 3′ arm a conserved C at the mature end. However, as there are relatively few plant sequences in the reduced sets, more sequences will are likely to provide more information.

For the characteristic vertebrate 5′ arm motif it contains the well known U conservation at the beginning of the mature sequence (position zero in logos) and a GU rich region in the 3′ end (of the 5′ arm) around positions 18–25. Interestingly, the 5′ arm motif also contains an upstream semi-conserved G at position −13. For the invertebrate organisms (*dme, dps, cel, cbr*), the 3′ arm motif seems characteristic. Fly seems to have more conserved positions in the neighborhood of the mature start in particular a semi-conserved U at position −9. (See Figure S1 in the supplementary material for details.)

To compute which parts of the 5′ and 3′ motifs are similar and different, we utilized a sliding window across the two profiles and computed the ALLR score and an empirical *p*-value at corresponding positions (Section 2). Window sizes from 6 to 14 were utilized all providing the same information with different resolutions. In Fig. 2 we show the scan on human for window size 7. It is in particular notable that around positions 15–20 the ALLR score drops significantly while the *p*-value at the same time is getting close to one. This indicates that the 3′ end of the two types of mature sequences differs (low ALLR score) and

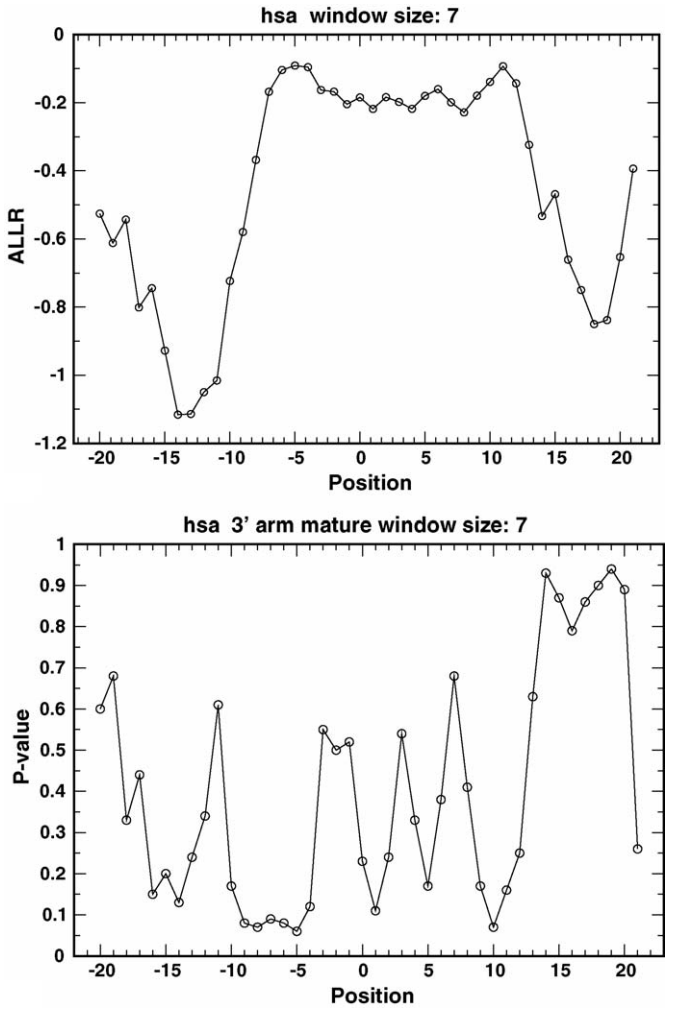


Fig. 2. Profiles of ALLR scores (top) and p-values (bottom) for human using a window size of 7 nucleotides across the two profiles, that are assumed aligned to the corresponding regions, see Wang and Stormo (2003) for details. For each position, the three neighboring nucleotides on both sides were used to compute the ALLR score. The p-values for each of the sliding windows are computed empirically by shuffling the columns (100 times) in each of the windows of the 5′ arm motif while fixing 3′ arm motif, see Section 2 for details. An almost identical plot is obtained by shuffling the 3′ arm motif while fixing the 5′ arm motif (not shown). The profiles in the top row are computed by keeping the columns of the 5′ arm motif fixed while shuffling the columns of the 3′ arm motif (corresponding positions).

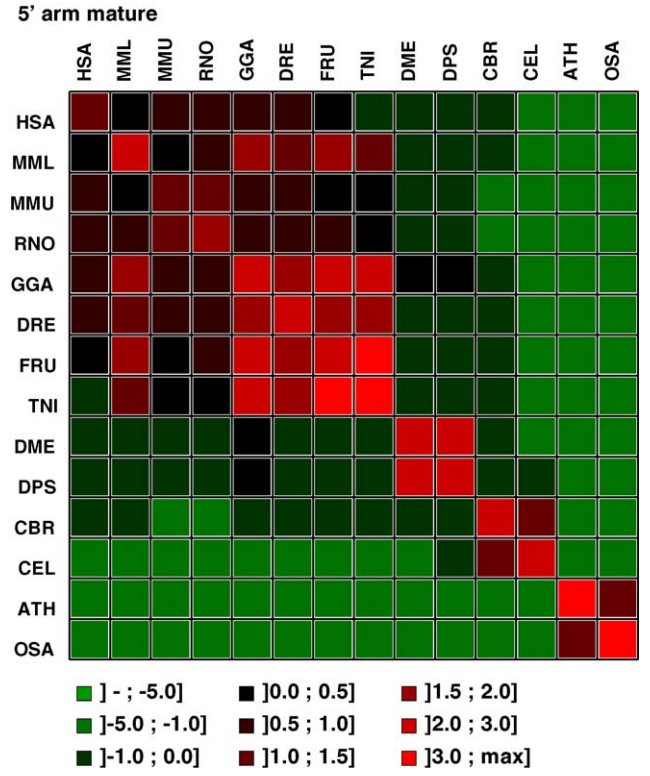


Fig. 3. Pairwise comparisons for the organisms listed in Table 1. The ALLR score was computed of the region spanning from the mature start position and 25 nucleotides downstream.

also differs significantly as the p-value is high. Similar type of observations are obtained for the other organisms listed in Table 1, however the curves do in some instances vary differently upstream from the 3′ end of the compared mature regions (data not shown). In few cases the p-value signal on positions 15–20 is not so strong, and the p-value drops to 0.55 in a case (*cel*). Also the peak might be shifted towards position zero.

In contrast to the difference observed between the 3′ ends of the 5′ and 3′ arm mature sequences, we for the 5′ ends observe that the ALLR score, although negative it is only slightly negative and the p-value is close to zero. This indicates that the ALLR score in this region is not significantly different from what would be expected when comparing with shuffled sequences. Hence no conclusion can drawn about similarity between the 5′ ends of the mature sequences in the 5′ and 3′ stem arms.

We also compared the 5′ arm (3′ arm, respectively) among the organisms computing the ALLR score. We compared the region covering the entire mature sequence starting at position 0 ending 22–29 nucleotides downstream. For each comparison we find that the related organisms have a relatively high score among themselves and score lower with more distant organisms. An example using a region spanning 24 nucleotides downstream is shown in Fig. 3. Note that the ALLR score of a profile against itself is exactly the information content of the profile within the considered region. The only less inconsistent pattern is the *mml* comparison. This is likely to be due to the relatively few sequences compared to the many sequences for the close related organism *hsa, mmu, rno*. This is also likely to be reflected in the higher score of *mml* against itself than *hsa, mmu, rno* against themselves.

4. Discussion

We did find that there is an asymmetry (difference) between miRNA sequence motifs when the mature sequence is located in the 5′ and 3′ arms of the stemloop precursor. A key question is, whether the 5′ and 3′ arm of the mature miRNA sequences are processed in the same way. Given the asymmetry observed here, this could be possible if, for example, the 3′ arm of the mature miRNA* sequence contains the same features as the mature 5′ arm sequence. However, recent studies have provided biochem-

ical verification showing that a mature miRNA is less stable at its 5′ end than its 3′ end (5′ end of star sequence) (Schwarz et al., 2003; Khvorova et al., 2003; Krol et al., 2004). This shows that the star sequences of the 3′ arm cannot have the same properties as the mature sequences on the 5′ arm and vice versa. These observations suggest that the miRNA processing machinery not only acts in an asymmetric fashion with respect to the mature and its star sequence, as showed by Tomari et al. (2004), but is also asymmetric with respect to processing 5′ and 3′ arm sequences. For vertebrate organisms, the upstream conserved G of the 5′ arm motif is a candidate for playing a role in the asymmetry.

We also observed that profiles among the organisms somewhat differs and that more closely related organisms have a higher score among themselves than more distantly related organisms. The vertebrate profiles seems to share very similar profiles of the 5′ arm motif, but interestingly fish appears to differ on the 3′ arm motif with and A-dominated signal at position 13 in the logos. However, in particular the plant appears to a specific feature, namely semi-conserved C in 3′ end of both the 5′ and 3′ arm mature sequences.

In agreement with this, the plant species are well known to have major differences in the biogenesis. They lack the processing of Drosha which instead is mediated through Dicer-like endonucleases— and more specifically DCL1 (Dicer-like protein 1) (Papp et al., 2003). DCL1 acts, in contrast to metazoan homologs, in the nucleus as the first processing steps of the pre-miRNAs which are further categorized differently from the metazoan intermediates. It is both more variable in size and have a high turn-over rate most likely from a coupled processing in the nucleus from the DCL endonuclease, resulting in a temporary precursor intermediate (Reinhart et al., 2002). Moreover, plant miRNA::target interaction is also more precise and shows a near-perfect complementarity (Rhoades et al., 2002). No obvious conservation of any miRNA gene and lack of Drosha homologs between the animal and plant kingdoms even propose an independent origin of this mechanism, as suggested by Bartel (2004).

Even though there is no experimental results concerning the different organization of 5′ and 3′ arm mature sequences between fly/worm and amniotic deployment, related distinctions have been observed. Note, that the RISC complex have only been studied in detail for fly (Tomari and Zamore, 2005) and similar studies might reveal variation in processing, for example, between human and fly. A related difference in the RISC complex with respect to RNAi have been observed, where Argonaute 2 is the only slicer in human that provides a fully functional RISC complex (Liu et al., 2004). Mammals do not have an endogenous siRNA expression in contrast to the lower eukaryotic species (reviewed by Bartel, 2004). The mammalian miRNA biogenesis has evolved to a state of fine tuning the processing steps solely with miRNA expression at hand. The other clustered groups have different silencing pathways (DNA methylation, siRNAs) relying in general on the same set of processing machinery hence, signals for biogenesis properties had to evolve coordinately in contrast to the mammalian way where miRNA associated proteins and miRNA signals have evolved synchronously. Another aspect is also target substrates. Although many miRNA families seem to be evolutionary conserved, which also is a trait distinguishing them from siRNA, there is a rising number of mammal specific miRNAs, for example the mir-196 involved in regulating expression from the HOX-gene clusters (Yekta et al., 2004).

Our observations also indicate differences between worm and fly and it has been suggested that they could have different RNAi pathways (Zamore, 2002). In fact, recent findings (Lin et al., 2005) show that the location of the mature sequence in intronic miRNAs in zebrafish are crucial for proper processing. Here, it is suggested that Dicer promotes asymmetry in strand selection possibly due to sequence bias within the apex loop. Hence, our observations are not in conflict with the current knowledge of miRNA processing, but contribute further to the possibility of variations in the miRNA processing machinery.

Acknowledgements

Thanks to Gary D. Stormo and Ting Wang for many long discussions and suggestions to comparing the two types of motifs. Thanks to Anders Krogh on comments on an early version of the manuscript. This work was supported by Danish Research Councils (SJVF/STVF) and the Danish Center for Scientific Computation. MÖ and ME were supported by Wallenberg Consortium North.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2006.04.006.

References

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M., Roach, J., Oh, T., Ho, I., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S., Clark, M., Edwards, Y., Doggett, N., Zharkikh, A., Tavtigian, S., Pruss, D., Barnstead, M., Evans, C., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297, 1301–1310.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Bartel, B., Bartel, D.P., 2003. MicroRNAs: at the root of plant development? Plant Physiol. 132 (2), 709–717.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

Celniker, S., Wheeler, D., Kronmiller, B., Carlson, J., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S., Frise, E., Hodgson, A., George, R., Hoskins, R., Laverty, T., Muzny, D., Nelson, C., Pacleb, J., Park, S., Pfeiffer, B., Richards, S., Sodergren, E., Svirskas, R., Tabor, P., Wan, K., Stapleton, M., Sutton, G., Venter, C., Weinstock, G., Scherer, S., Myers, E., Gibbs, R., Rubin, G., 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol. 3 (RESEARCH0079).

C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.

C. elegans Sequencing Consortium, 2006. http://www.sanger.ac.uk/pub/wormbase/cbriggsae/cb25.agp8.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., Hannon, G.J., 2004. Processing of primary microRNAs by the microprocessor complex. Nature 432, 231–235.

Gorodkin, J., Heyer, L.J., Brunak, S., Stormo, G.D., 1997. Displaying the information contents of structural RNA alignments: the structure logos. CABIOS 13, 583–586, http://www.cbs.dtu.dk/gorodkin/appl/slogo.html.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., Shiekhattar, R., 2004. The microprocessor complex mediates the genesis of micrornas. Nature 432, 235–240.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A., 2005. Rfam: annotating non-coding rnas in complete genomes. Nucl. Acids Res. 1, D121–D124.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., Hannon, G.J., 2001. Argonaute2, a link between genetic and biochemical analyses of RNAi. Science 293, 1146–1150.

HGSC at Baylor College of Medicine, 2006. http://www.ensembl.org/pub/current_macaca_mulatta/data/fasta/dna.

Int. Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.

Int. Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K., McEwan, P., Bosak, S., Kellis, M., Volff, J., Guigo, R., Zody, M., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E., Weissenbach, J., Roest, C.H., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

Khvorova, A., Reynolds, A., Jayasena, S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. Cell 115, 209–216.

Krol, J., Sobczak, K., Wilczynska, U., Drath, M., Jasinska, A., Kaczynska, D., Krzyzosiak, W.J., 2004. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. J. Biol. Chem. 279, 42230–42239.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., Kim, V.N., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425 (6956), 415–419.

Lewis, B.P., Burge, C.B., Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell 120, 15–20.

Lin, S.-L., Chang, D., Ying, S.-Y., 2005. Asymmetry of intronic pre-miRNA structures in functional RISC assembly. Gene. 356, 32–38.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., Hannon, G.J., 2004. Argonaute2 is the catalytic engine of mammalian RNAi. Science 305, 1437–1441.

Ma, J.-B., Ye, K., Patel, D.J., 2004. Structural basis for overhangspecific small interfering RNA recognition by the PAZ domain. Nature 429, 318–322.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., Tuschl, T., 2004. Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing. Mol. Cell 10, 544–550.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Papp, I., Mette, M.F., Aufsatz, W., Daxinger, L., Schauer, S.E., Ray, A., van der Winden, J., Matzke, M., Matzke, A.J., 2003. Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. Plant Physiol. 132, 1382–1390.

Rat Genome Sequencing Project Consortium, 2004. Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature 428, 493–521.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., Bartel, D.P., 2002. MicroRNAs in plants. Genes Dev. 16 (13), 1616–1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., Bartel, D.P., 2002. Prediction of plant microRNA targets. Cell 110 (4), 513–520.

Richards, S., Liu, Y., Bettencourt, B., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M., Chen, R., Meisel, R., Couronne, O., Hua, S., Smith, M., Zhang, P., Liu, J., Bussemaker, H., van, B.M., Howells, S., Scherer, S., Sodergren, E., Matthews, B., Crosby, M., Schroeder, A., Ortiz-Barrientos, D., Rives, C., Metzker, M., Muzny, D., Scott, G., Steffen, D., Wheeler, D., Worley, K., Havlak, P., Durbin, K., Egan, A., Gill, R., Hume, J., Morgan, M., Miner, G., Hamilton, C., Huang, Y., Waldron, L., Verduzco, D., Clerc-Blankenburg, K., Dubchak, I., Noor, M., Anderson, W., White, K., Clark, A., Schaeffer, S., Gelbart, W., Weinstock, G., Gibbs, R., 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15, 1–18.

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display concensus sequences. Nucl. Acids Res. 18, 6097–6100.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., Zamore, P.D., 2003. Asymmetry in the assembly of the RNAi enzyme complex. Cell 115, 199–208.

The Zebrafish Sequencing Group, 2006. http://www.ensembl.org/pub/current_danio_rerio.

Tomari, Y., Matranga, C., Haley, B., Martinez, N., Zamore, P.D., 2004. A protein sensor for siRNA asymmetry. Science 306, 1377–1380.

Tomari, Y., Zamore, P.D., 2005. Perspective: machines for RNAi. Genes Dev. 19, 517–529.

Wang, T., Stormo, G.D., 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 2369–2380.

Yekta, S., Shih, I.H., Bartel, D.P., 2004. Microrna-directed cleavage of hoxb8 mrna. Science 304, 594–596.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J., Buell, C., 2003. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res. 31, 229–233.

Zamore, P.D., 2002. Ancient pathways programmed by small RNAs. Science 296, 1265–1269.

Zeng, Y., Cullen, B.R., 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Res. 32 (16).

Supplementary material:

MicroRNA sequence motifs reveal asymmetry between the stem arms

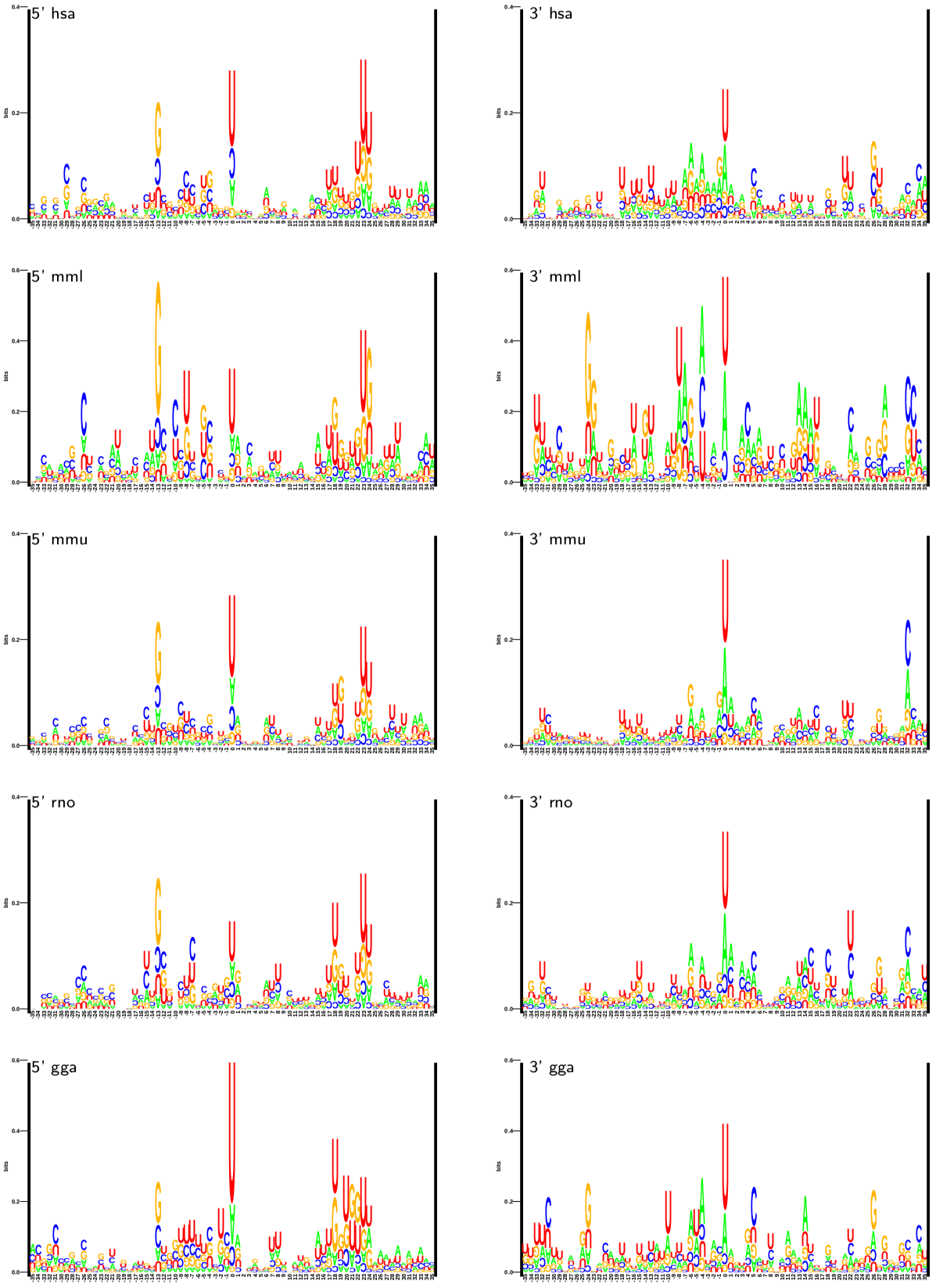
Jan Gorodkin[1], Jakob H. Havgaard[1], Mats Ensterö[2], Milena Sawera[1], Peter Jensen[1], Marie Öhman[2] and Merete Fredholm[1]

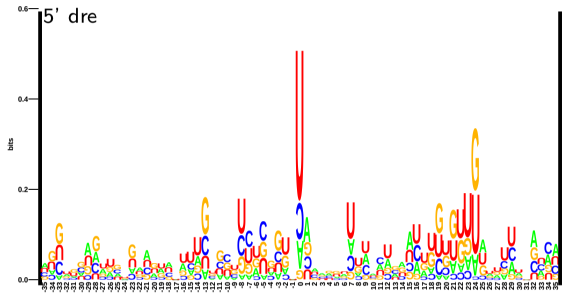
[1]Division of Genetics and Bioinformatics, IBHV and Center for Bioinformatics, The Royal Veterinary and Agricultural University, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

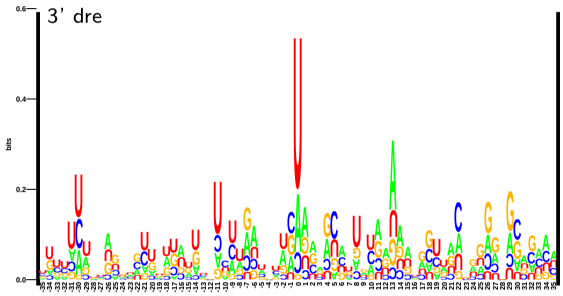
[2]Department of Molecular Biology & Functional Genomics, University of Stockholm, SE-106 91 Stockholm, Sweden

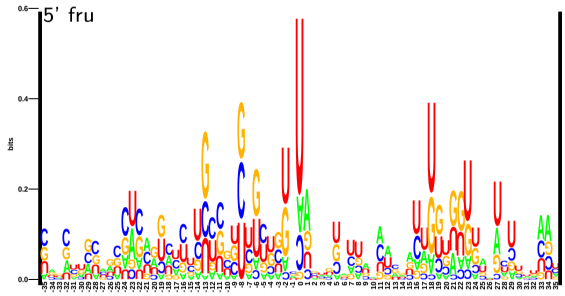
1 Sequence profiles of 5’ and 3’ arms

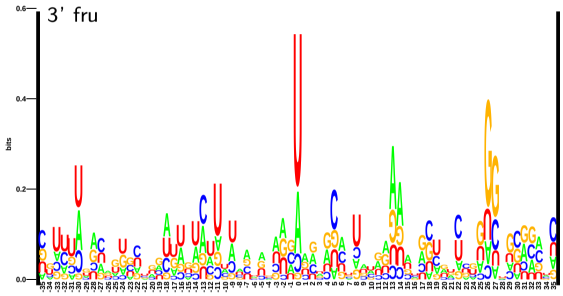
Figure S1: The sequence profiles various organisms. Left column represent 5’ arm motifs. Right column represent 3’ arm motifs. The scales are arbitray.

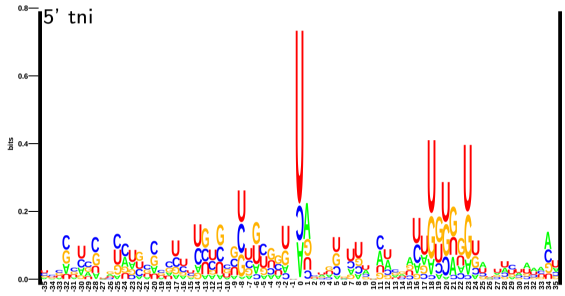


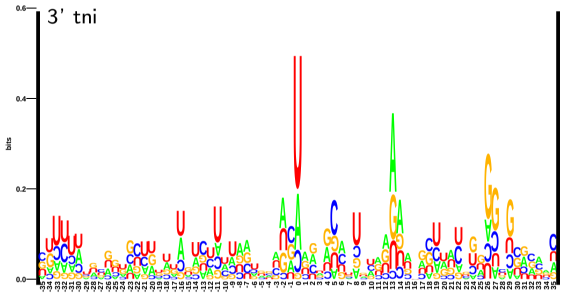

5' dre

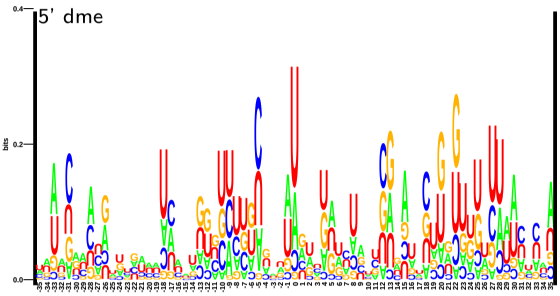

3' dre

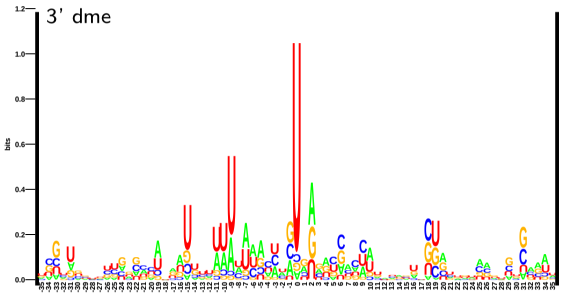

5' fru

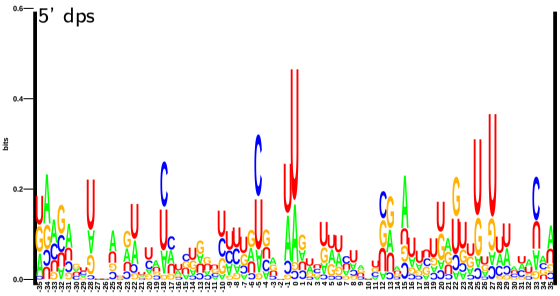

3' fru

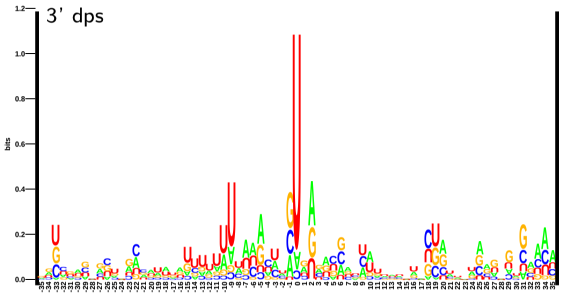

5' tni

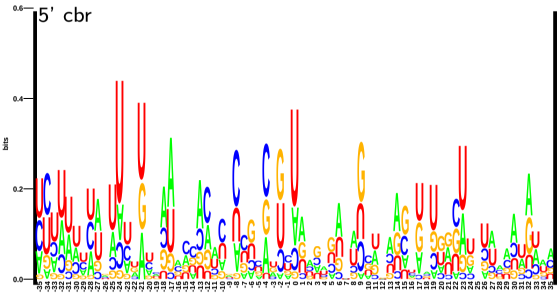

3' tni

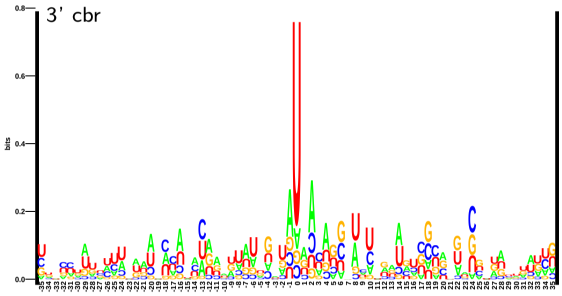

5' dme

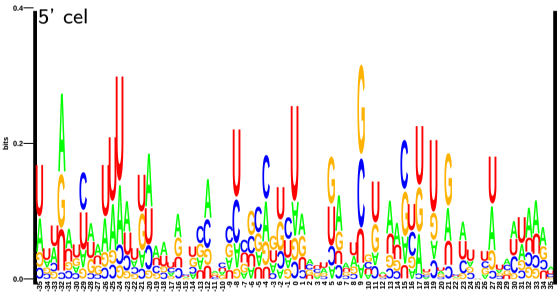

3' dme

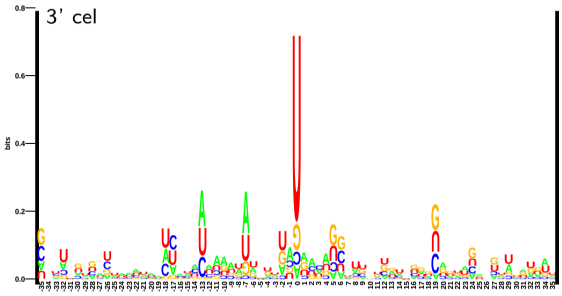

5' dps

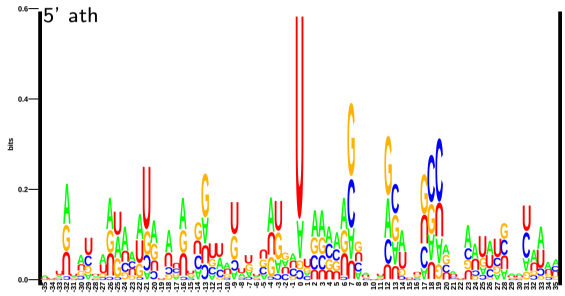

3' dps

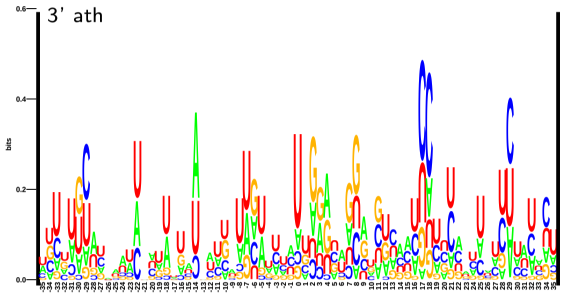

5' cbr

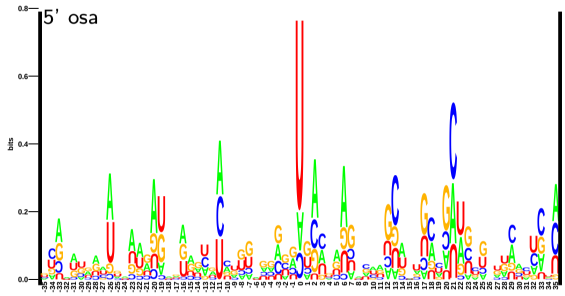

3' cbr

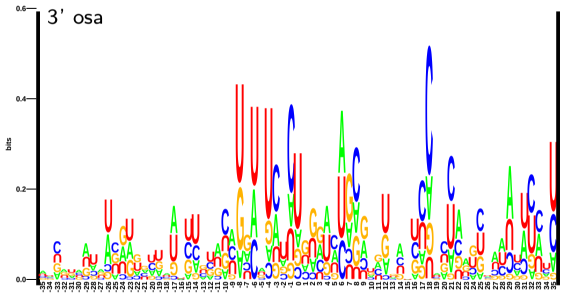

5' cel


3' cel


5' ath


3' ath


5' osa


3' osa

Paper III

A computational screen for site selective A-to-I editing

Mats Ensterö * Örjan Åkerborg * Daniel Lundin Bei Wang
Terrence S. Furey Marie Öhman Jens Lagergren

April 23, 2008

Abstract

Several bioinformatic approaches have previously been used to find novel sites of ADAR mediated A-to-I editing in human. These studies have discovered thousands of genes hyper-edited in their non-coding regions, but very few substrates that are site-selectively edited. Known substrates suggest, however, that site selective A-to-I editing is particularly important for normal brain development in mammals. We have compiled a screen that enables identification of new sites of selective editing, primarily in coding sequences. To avoid hyper-edited repeat regions, we have applied our screen to the *alu*-free mouse genome. Choosing mouse also facilitates the experimental verification and enable us to analyze the extent of editing at early developmental stages of the brain. First we constructed an explorative screen based on RNA structure and genomic sequence conservation. We evaluated the explorative screen by means of enrichment of *A-G mismatch*, that is, the discrepancy between the genomic template and the expressed sequence having an A and an I (read as G), respectively, in corresponding positions. The enrichment implicate A-to-I editing as an important mechanism in fine tuning proteome diversity. We extended the explorative screen by including a specific scoring scheme based on characteristics for known A-to-I edited sites. The extent of editing in the candidate genes was verified using total RNA from mouse brain and 454 sequencing. Editing with low efficiency was verified at several sites within the regions that were predicted.

Introduction

The eukaryote cellular machinery has been shown to contain several alternative processing mechanisms acting on RNA. On the pre-mRNA level alternative splicing is a well-known mechanism altering the transcript. This type of alternative processing is particularly important in the nervous system, where it helps determining the properties of many types of neurons [Li et al., 2007]. Although RNA editing has received less attention it is known to fine-tune messenger RNA composition by changing single nucleotides. The most common enzymes that perform editing in mammals are the ADAR (*adenosine deaminase that acts on RNA*) proteins. The ADAR enzymes ADAR1 and ADAR2, convert adenosines to inosines (A-to-I) within double stranded RNA by a hydrolytic deamination (reviewed in [Maydanovych and Beal, 2006]). Since inosine is interpreted as guanosine (G) by the splicing and translational machineries, ADAR editing effectively results in an A-to-G change that may alter the amino acid sequence encoded by the substrate. There are two types of A-to-I edited sites, (1) hyper-edited sites which are abundant in non-coding and untranslated regions of long, almost completely double stranded, stem loop structures [Morse and Bass, 1999, Levanon et al., 2004] and (2) selectively edited sites which consists of imperfect stem loop structures, often formed by an exon and a trailing intron sequence. Site selective editing is believed to be a rare event where the known examples have mainly been found in genes involved in neurotransmission.

The known substrates for site selective editing typically have a functional significance due to non-synonymous alteration of a codon. Both strands of a substrate stem often show high conservation of sequence

*These authors contributed equally to this work

as well as structure in species from man to chicken [Aruscavage and Bass, 2000, Hoopengardner et al., 2003, Ohlson et al., 2007]. Imperfections in form of bulges and internal mismatches are important structural features for site selective editing [Bass, 2002]. Even though only a handful substrates have been identified, editing has been proven to be important for the function of the developing brain in both invertebrates [Palladino et al., 2000] and vertebrates [Hartner et al., 2004, Higuchi et al., 2000, Wang et al., 2000].

Our explorative screen for selectively edited sites has two components, RNA structure prediction and sequence conservation. We use StemPrediction to predict stems among genomic sequences containing sequence pairs being approximately reverse complementary. To extract duplexes found in conserved regions an in-house conservation measure is applied to multiple alignments of 17 vertebrate genomes [Margulies et al., 2003]. We use an alignment between genomic data and an expressed sequence database [Boguski et al., 1993] in order to extract A-G mismatches. Our explorative screen is evaluated based on enrichment of A-G mismatch for highly conserved stems. The explorative screen is then extended to include a specific 6-bit scoring scheme based on characteristics for known A-to-I edited sites. Interestingly, a stand-alone application of the explorative screen recently lead to the discovery of editing of the GABA-A receptor, subunit alpha 3 (Gabra3) [Ohlson et al., 2007, Pedersen et al., 2006].

Similar ideas have been used previously to construct computational screens with the same purpose [Levanon et al., 2004, Athanasiadis et al., 2004, Blow et al., 2004, Clutterbuck et al., 2005]. The hallmarks of these prior screens have been the A-G discrepancy and the clustering of adjacent discrepancies. Less used components involve conservation (often mouse/human) and prediction of target RNA foldback structures. These studies have mainly led to the discovery of thousands of hyper-edited substrates where the editing events arise from inverted repetitive elements such as Alu sequences.

Our aim has been to find single sites of selective editing having the potential of re-coding the open reading frame. To do this we have focused on finding stem-loop structures that contain A-G discrepancies and are conserved in sequence between species. Our screen also benefits from the features of the 6-bit scoring scheme. The result of applying our extended screen to the mouse genome gives a substantial number of novel putative substrates of which 45 have been experimentally validated. 38 comes from our combined explorative and extended screen and an additional 7 comes from the explorative screen alone. In the latter, we look for possible editing events within 7 highly conserved stem regions without any conditional A-G mismatch.

Results

Here we describe in more detail the components in, the evaluation of, and the results of applying, our screens. The process is illustrated with a flow chart presented in Figure 1.

BLASTZ

We used BLASTZ [Schwartz et al., 2003] to extract sequences from the Mm8 assembly [Karolchik et al., 2003] containing reverse complementary pairs of subsequences, reasoning that these are likely to form RNA duplexes. In a previous study [Ohlson et al., 2005], an ADAR2/RNA co-immunoprecipitation strategy was applied to a microarray in order to identify ADAR2 substrates. The study evaluated 11,827 well annotated mouse genes. Our BLASTZ search was restricted to genomic regions which: (1) are bound by any of these 11,827 genes and (2) are alignable with at least 10 of the 16 species in the multiple sequence alignment (MSA) [Karolchik et al., 2003] containing the mouse genome aligned to 16 other vertebrates. We used BLASTZ parameters as described in Methods. We noted that the amount of result was very sensitive to certain parameters but decided to use the default ones. The total number of sequence pairs extracted with BLASTZ was 53,729,218, around 5.000 per gene on average.

StemPrediction

We then used *StemPrediction* (see Methods) to filter this large sequence collection for pairs of sequences exhibiting characteristics of known ADAR substrates. A key parameter was the MAX_ENERGY cut-off

corresponding to minimum free energy for the stems. We avoided a strict cut-off, since the free energy for known ADAR substrates are often moderately low (see Figure 2). On the other hand, an overly liberal cut-off will inevitably result in a vast amount of noise sequence pairs to analyse in the next step. Based on these considerations $MAX\_ENERGY = -15$ was chosen. When inspecting the result we find it unlikely that a looser cut-off would yield any addition of interesting predictions. The energy values for the retrieved stems ranged between extremes -15 and -1382. The empirical distribution is shown in Figure 3. The total number of retrieved stems was 2,919,511. Of those, 55%, or 1,611,913, had a predicted *stem loop* longer than 10,000 nucleotides. These were removed, since all the confirmed substrates have stem loops shorter than 5,000 nucleotides, which left us with 1,307,598 *candidate stems* to analyse further.

Mm8 conservation labelling

Using the above mentioned mouse vs. 16 vertebrates MSA, we scored each Mm8 site/nucleotide according to its level of conservation (see Methods). This MSA contains a collection of mouse sequences, each aligned to as many of the other 16 genomes as possible. Each mouse site included in an alignment containing at least 10 out of the 17 species were given a positive *conservation score* while all other positions were given a conservation score of zero. The conservation score for these *10-aligned* sites is the sum of the *parsimony score* and the *tree score* (see Methods), both computed relative to a window of k nucleotides upstream and k nucleotides downstream of s. We found $k = 10$ to be suitable, i.e., the conservation score for s depends on the sites in a window of width 21 surrounding s.

The number of sites on the mouse genome that was given a positive conservation score was 58,192,830, i.e, around 2% of the mouse genome, and the values ranged from just above zero to 110. An *area* with conservation score c is a set of contiguous sites, with at least one site scoring c or higher, delimited by 50 consecutive sites all having a score below c. The number of sites and areas which have conservation scores within various conservation score intervals is shown in Table 1. In Table 1 is also shown the number of sites and areas which have a conservation score within these intervals and overlap a gene.

The idea behind using the parsimony score and the tree score is that the former should capture absolute conservation, i.e., its value will be high for sites in which very few mutations have occurred, while the latter will capture conservation in the mouse and human part of the tree which relates the aligned species (see Figure 4). That is, a site in which several substitutions have occurred in some small subtree distant from mouse, but where no substitutions has occurred elsewhere, will have a high tree score value. In Figure 7 is shown an alignment of a genomic region overlapping GluR-B R/G. The window containing the first 21 nucleotides, surrounded by a green box in the figure, contains five substitutions altogether. All these have occurred in Tetrafish resulting in a high tree score for this window.

In Figure 5 is shown a section of mouse chromosome 3 overlapping the positions for already recognized ADAR substrates GluR-B Q/R and GluR-B R/G sites respectively. It is clear that the genome positions for these two substrates score high.

Stem conservation scoring

Using the Mm8 conservation labelling, we scored each candidate stem according to its level of conservation. We expect that ADAR substrates should be highly conserved in terms of structure, at least in areas close to the edited site. Typically, it is the bases in the helical regions of the ADAR substrates whose identity is conserved whereas bases in nonhelical regions are not, although their nonhelical state is maintained [Aruscavage and Bass, 2000]. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing; Daniel and Öhman, unpublished). We therefore required a high conservation score on both stem arms of the putative substrates. Consequently, the candidate stem was given the conservation score of its *lowest* scoring stem arm, and each stem arm was, in turn, given the score of the highest scored site on that arm. The number of candidate stems which have conservation scores above various thresholds is shown in the rightmost column of Table 1.

A-G mismatcher

We used two databases, *mouse EST* [Boguski et al., 1993] and *mouse SNP* [Sherry et al., 2001] to extract A-G mismatches. The genomic sequences used in the alignment corresponded to the CDS boundaries, also called *genomic mRNA* below. An A in genomic data and a G at the same position in EST data for an *individual*, is a trait of A-to-I editing. However, the sequences in the databases correspond to many individuals, so an A-G mismatch may be caused by an SNP. For this reason we used the mouse SNP database to remove known SNPs from our predicted A-G mismatches. Nevertheless, it has previously been shown that over one hundred SNPs in human are actually somatic changes most likely due to A-to-I editing [Eisenberg et al., 2005]. Therefore, A to G SNPs verified by sequencing of ESTs were not excluded from the screen. A total of 142,136 A-G mismatches were filtered out and of those 32,948 were rejected due to concurrent hits in the SNP database. Thus, 109,188 high quality A-G mismatches were detected.

The number of genes containing a certain number of A-G mismatches ranged between the extremes 0 and 420 according to the distribution in Figure 6. 10841 genes contain at least one A-G mismatch. The gene Spna2 contains the highest number of A-G mismatches, 420.

A-G mismatch enrichment analysis

When each stem had been given a conservation score, we evaluated our collection of candidate stems for A-G mismatch enrichment. We partitioned the spectrum of conservation scores into sections < 50, $50-60$, $60-70$, $70-80$, $80-90$ and ≥ 90. We expect, if conservation score and A-G mismatches are indeed both ADAR substrate characteristics, that A-G mismatches will be enriched among candidate stems with high conservation score. We evaluated this using a null hypothesis according to which an A-G mismatch is independent of A-to-I editing. Since we view editing as the only possible explanation for dependence between A-G mismatch and conservation and in order to get a computable p-value, we extend the null hypothesis to include independence between A-G mismatch and conservation. In Table 2 is shown absolute numbers and relative frequencies of A-G mismatch for various conservation scores. The frequency of A-G mismatch among stems with conservation score < 50 and $80-90$ are 0.128 and 0.273, respectively (see rightmost column of Table 2). The probability of having such a high discrepancy assuming that A-G mismatch is a random phenomenon with the same distribution in both ranges of conservation score is $< 10^{-85}$, p-value calculated with Hoeffding's bound [Hoeffding, 1956]. We conclude that there is a correlation between editing and high conservation score. In both ranges, some of the A-G mismatches could be attributed to random phenomena independent of editing, and we assume that the fraction of randomly occurring A-G mismatches is the same in both ranges. This fraction can be no larger than 0.128, implicating that the fraction of edited stems in the $80-90$ range is at least 0.145, corresponding to 483 stems.

We further used the results from the IP assay presented in [Ohlson et al., 2005]. In [Ohlson et al., 2005] an anti ADAR2 antibody was used to bind ADAR2 in complex with RNA targets. The RNA was subsequently used for microarray analysis using Affymetrix mouse chip 430A. The aim was to identify novel selectively edited ADAR2 substrates and afterwards use a computational approach to find the edited sites. Three biological replicates and control experiments were conducted. Using standard approach, microarray p-values were calculated for the probability that the differential hybridization between arrays stained with samples corresponding to the ADAR2 specific and the non-specific antibody, respectively, occurred purely by chance. We partitioned the spectrum of microarray p-values into sections 10^{-1}, $10^{-5}-10^{-1}$, $10^{-10}-10^{-5}$, $\leq 10^{-10}$. The conservation scores and microarray p-values are presented in Table 2. Comparing the $\leq 10^{-10}$ and 10^{-1} microarray p-value ranges in a way similar to how the two conservation ranges was compared above, yields an estimate that the former interval has a fraction of 0.055 edited stems, corresponding to 798 stems.

Thus, it seems from the enrichment that the conservation score is a better predictor of ADAR editing than the microarray p-value. Based on this we decided to pick candidates for experimental validation using a scoring scheme with the conservation score as the main component. The microarray p-value was not used.

Site ranking using the extended screen

Our screen utilizes a site ranking scheme as a filtering step to narrow down the number of candidates. As a pre-filter we evaluated a conservation cut-off score for a candidate to enter the site ranking scheme. Hence, we compiled the conservation scores for the predicted stems of the positive controls see Table 6. To include half of them (8/15), we would have had set the cut-off to 75 or above. The curated assessment of this cut-off was consequently set to ≥ 75. By applying the explorative screen and a conservation cut-off of 75 and among the thereby obtained stems selecting those with A-G mismatches, we obtained 2,524 stems from the 53,729,218 sequences with complementary sub sequences retrieved with BLASTZ. To achieve a further reduction we added the site scoring criteria based on common features among known ADAR substrates. We used a bit-scoring scheme in which a candidate stem could have a maximum score of 6.

The first two bits were used to credit conservation even further. All 2,524 had a score ≥ 75 but we decided to score ≥ 80 and also ≥ 90. The reasoning to promote the conservation further is that the 3 top scored regarding conservation (GluR-B: R/G 96, Q/R 85 and Gabra3: I/M 85) also are edited close to 100%. Assuming the editing frequency to be a quality marker for the conservation trait to improve this screens specificity, we decided to add 2 bits in total for highly conserved stems. These bits are called *cons_80* and *cons_90*, respectively, in Table 3. The third bit specifically scores whether an A-G mutation has occurred. This bit is called *A-G mutation*. A stem has an A-G mutation if it has an A-G mismatch, and if mouse and the species close to mouse have an A in the A-G mismatch site in the alignment, while species in some subtree distant from mouse have a G. In Figure 7 is shown a 17 species alignment for the Mm8 genome segment harboring the GluR-B R/G site. The edited site is shown as a red column in the alignment and we note specifically in that column the A-G mutation occurring in tetrafish. The fourth bit was used to further reward distinct A-G mismatches in both stem arms. The underlying reasoning is that if the occurrence of A-G mismatches in stem arms is independent of how we pair stem arms into stems, the probability of having A-G mismatches in both stem arms is significantly lower than the probability of having an A-G mismatch in only one stem arm. This bit is called *A-G_both*. The data in EST libraries is often suffering from sequencing errors. To analyze if the identified A-G mismatches were identified as changes in entered protein sequences we downloaded all available, mRNA and protein sequences from the Entrez gene site [Maglott et al., 2007]. If amino acid changes appeared in the protein sequence due to A to G changes it was scored as *annotated_aa_change*. It has previously been shown that there is a sequence bias in the vicinity of an edited adenosine [Lehmann and Bass, 2000]. Hence, we used algorithms for calculating information content [Schneider and Stephens, 1990] to sort out if and how to score a nearest neighbour distribution of an edited site (see Methods). The calculation shows the following: the upstream and downstream neighbour sites had information contents of 0.43 bits and 0.50 bits, respectively. The upstream site had no guanosines at all in the 24 sequences used and the downstream site had a *preferred* guanosine (0.27 bits). The slightly higher information content of the downstream neighbour and also the presence of a downstream G in most of the known ADAR subtrates, motivated us to score a downstream guanosine of a candidate editing site, (*ds_G*).

We compiled a list with all candidates having a site score of ≥ 3 (see Table 5). This list contains 53 sites where 4 of them are known ADAR target sites. Due to RNA preparation procedures for sequential verification we had to reduce the number of sites being experimentally validated, since the region of interest in the mouse genome could not be un-ambiguously amplified. Hence, 38 sites were experimentally tested.

454 results

Validation was performed using amplicon analysis by 454 sequencing, see Methods. We aligned the collection of sequences retrieved for each of the 38 validated genes from 454 sequencing of the adult mouse, post natal day 21. The number of sequences retrieved for our 38 candidate genes, and thus the number of alignment rows, ranged from 46 to 1267. All alignments showing any sign of poor quality was discarded. A total of 175 positions was found where a genomic A was replaced by a G in at least one of the sequences. In most cases, the fraction of G in any such position was very low and. None of the sites correspond to the predicted positions of the A/G discrepancy. The candidate site in which the highest fraction was found was

a synonymous site at Elavl2. The alignment corresponding to this site contained 625 sequences out of which 15 (2.4%) showed a A-G replacement in the site in question. In a total of 10 sites the A-G replacement frequency was larger than 1%.

To certify that these replacements were in fact due to editing we performed a number of tests. We manually checked that the sequences were in fact unique to the corresponding genomic position. We further considered 454 sequencing errors. The 454 output contains a *phred score* for each position indicating the risk of erroneous sequencing for the position in question. In most cases the phred score was reported to be between 20 and 30 corresponding to 1% and 0.1% risk respectively. Using the phred scores we calculated p-values for the event that all A-G replacements would have been due to sequencing errors. In 40 out of 175 cases the p-value was found to be < 0.0001. In 12 cases the p-value was found to be $< 10^{-9}$.

Discussion

We have compiled an explorative screen for selectively A-to-I edited sites, based on two components, RNA stem structure and conservation of the corresponding sequence. For the stem structure, we use a free energy threshold, while the conservation score is used to rank stems.

An assay was designed for our explorative screen that tests whether highly conserved stems are enriched for positions with an A-G mismatch between the genomic and the transcribed sequence. The result of the evaluation is that A-G mismatches are significantly enriched in highly ranked stems. Comparing stems in the $80-90$ conservation score range with those in the < 50 range yields an estimate of 483 edited stems in the former. The same type of comparison between each of the three intervals $70-80$, $60-70$, and $50-60$ and the interval < 50 yields an estimate of 18074 edited stems in the combined conservation score range $50-80$. These values are surprisingly high, and it is reasonable to believe that ADAR editing is the biological phenomenon explaining these high numbers. It is noticeable that the conservation score range 90 contains relatively few stems with an A-G mismatch. We find two possible explanations for this: (1) several of the known ADAR substrates are in this range but have been excluded and (2) known functional edited sites often have a G in fish and amphibians that are more distantly related to mammals and this prevents a very high conservation score.

P-values for differential hybridization from the ADAR2/RNA IP assay [Ohlson et al., 2005] was available. In [Ohlson et al., 2005] the array p-values are evaluated solely based on the rank of known ADAR substrates. We investigated sequences having p-values in four different intervals with respect to enrichment of stems having an A-G mismatch. It would have been possible to include also these p-values in the extended screen. We chose not to do so because genes showing no differential hybridization but high conservation are also enriched for A-G mismatch. Although the converse is also true, it is less pronounced. In fact the fraction of edited stems in the $\leq 10^{-10}$ and 10^{-1} microarray p-value ranges is 0.055, that is 798 stems. It is however also interesting to note that except the most highly conserved stems (conservation score ≥ 90) each conservation interval for sequences having p-values 10^{-1} has a higher fraction of sequences having an A-G mismatch, than the corresponding conservation interval for p-value range $10^{-5} - 10^{-1}$. We have not found any plausible and logical explanation of this.

We extended our screen by including several additional components of which A-G mismatch is one. Our extended screen was applied to the mouse orthologs of the known human ADAR substrates. As seen in Table 6, of the ADAR selectively edited sites, 5 are contained in a stem structure that: (1) has an A-G mismatch in mouse as well as human (2) has a free energy below the threshold, and (3) has a conservation score above 75. By restricting ourselves to structures with conservation score above 75, we lose some of the known ADAR substrates but the majority satisfy this requirement.

From the final 53 candidate list, it is worth noting that the R/G and Q/R site of GluR-B, the Gabra3 I/M site and the Kcna1 I/V sites are among the absolute top ranking candidates (see Table 5). This is a strong indication that our screen in total have an intrinsic capacity to detect ADAR targets. Out of the final list of candidates, 45 have been investigated further. We have performed experimental editing tests using amplicon analysis by 454 sequencing, on RNA extracted from the mouse brain. By using the 454 sequencing method the sequence in single transcripts can be analyzed. Altogether we found editing in 175

positions. Although the same position is edited in several transcripts from one candidate, the efficiency is very low. This indicates that the predicted RNA structures are recognized as ADAR editing substrates but that editing of these substrates is down regulated in the samples we have analyzed. In general editing has been shown to be regulated in certain tissues and during the development of the brain (Ensterö et al. manuscript). It is therefore possible that a higher editing efficiency can be detected during other conditions or in other tissues. In all, we have constructed a screen that can detect targets of A-to-I editing. Further, we have experimentally tested our top candidate novel edited regions. Although we cannot detect editing at the predicted positions, we detect other micro-editing events within the same region. The lack of discrepancies of the predicted positions could be due to several things: The previous set of bona fide editing events is in fact a near complete assembly of ADAR targets; there is no more site selective re-coding events to find. These are true editing substrate but ADAR is regulated so that these substrates are not edited in brain tissue and/or in the developmental stages that has been analysed. Interestingly, we indeed show that ADAR is present at these regions by the disclosure of A/G discrepancies that can not be explained by either sequencing or alignment errors. Another explanation is in line with previous suggestions that many genes are subjected to editing but with very low efficiency [Maas et al., 2003]. If so, our screen, indicate that the low-efficiency or micro-editing is extremely prevalent.

Methods

BLASTZ and StemPrediction

We used NCBI gene ID:s to download a complete set of genbank files for the 11,827 unique genes represented on the Affymetrix 430A microarray. Genes that could not be unambiguously mapped to a Genbank entry were discarded. We used BLAT [Kent, 2002] to align the head and tail sequences (100 nucleotides of the 5' and 3'-end of a gene respectively) to the corresponding chromosome. All sequences that could not be completely and uniquely aligned to their corresponding chromosome (NCBI build 36) were also discarded. BLAT was used with default parameters with the exception of $MIN\_IDENTITY$. $MIN\_IDENTITY = 100$ was chosen since we wanted to eliminate incomplete alignments. For each gene, sequences containing each exon and each and their corresponding adjacent introns were extracted. To determine potential stem-loop forming structures, first BLASTZ [Schwartz et al., 2003] was used to align each sequence to the reverse complement of itself, using parameter settings as shown in Table 4. We constructed a custom weight matrix for these alignments, shown in Table 4, that reflects the contribution of each base pairing to the stability of the structure including the non-standard G-U pairing (G-T in DNA sequence). Resulting alignments were further filtered using our StemPrediction software. StemPrediction first determines the lowest energy confirmation of a stem-loop structure formed by the BLASTZ aligned sequences using RNAfold [Hofacker et al., 1994]. Parameter settings (Table 4) allow potential stem-loops to be further filtered based on characteristics of the predicted structure such as the RNAfold determined minimum free energy, the length of the stem, and the number of paired and unpaired bases (bulges) in the stem. Stems from disjoint structures can be joined to create larger structures if stems sequences are within a specified distance of each other. These characteristics of stem-loop structures have been previously shown to be important in RNA editing [Tian et al., 2004, Carlson et al., 2003, Lehmann and Bass, 1999].

A-G mismatcher

Identifiers for the 11,827 genes from the array of the co-immunoprecipitation strategy were used to download the most recent set of gene sequences, including UTR's, and exon coordinate annotations for all transcript isoforms [Benson et al., 2007]. The coordinates gave a complete set of genomic coding sequences. We used two databases as of February 2007, a mouse *EST database* [Boguski et al., 1993] and an *SNP database*, build 126 [Sherry et al., 2001]. The genomic mRNA sequences were aligned with the EST database using BLASTN [Altschul et al., 1997], in order to deduce A-G mismatches between the template DNA and the expressed sequences. To reduce the risk of promoting an A-G mismatch originating from sequencing errors and/or low

quality alignments, we discarded alignments shorter than 100 nucleotides and alignments containing $\geq 20\%$ mismatches. We further used the SNP database to remove A-G mismatches likely to have a polymorphic genomic origin.

Mm8 conservation labeling

Each Mm8 site included in an alignment containing at least 10 of the 17 species was scored according to:

$$cons.score_{window} = pars.score_{window} + tree.score_{window}.$$

The parsimony score for column s is calculated as a sum over the individual values for the columns in the window centered at s:

$$pars.score_{window} = \sum_{col=s-k_{left}}^{s+k_{right}} pars.score_{col} \qquad \begin{array}{l} k_{left} = \#\text{ columns upstream } s \text{ (10 using window size 21)} \\ k_{right} = \#\text{ columns downstream } s \end{array}$$

where $pars.score_{col}$ depends on the minimum number of substitutions needed to explain that column. The details are to be found in [Margulies et al., 2003] where this algorithm is entitled *parsimony-based method for MCS detection*. The calculation is done with respect to the structure of the species tree, see Figure 4, the tree's edge lengths, and a substitution rate matrix (we follow [Margulies et al., 2003] and use the HKY neutral substitution rate matrix [Hasegawa et al., 1985]). When calculating the tree score we consider all columns in the window simultaneously and we observe where in the tree nucleotides deviating from the consensus are found:

$$tree.score_{window} = \sum_{m'=1}^{m} \left\{ \begin{array}{l} \#\text{ rooted subtrees} \\ \text{with } m' \text{ leaves} \end{array} \right\} \cdot \prod_{i=1}^{21} \frac{\binom{m'}{d_i}}{\binom{n}{d_i}} \qquad \begin{array}{l} n = \text{total } \# \text{ leaves} \\ m = \#\text{ leaves in subtree with mutations} \\ d = \text{number of mutations in column } i \\ k = \text{total number of columns} \end{array}$$

This value will be large if all deviating nucleotides are isolated to some small subtree (c.f. the GluR-B example shown in Figure 7 where in the boxed window all deviations are found in tetrafish). In this case the parsimony score will be lowered by the five columns having a substitution but the tree score will be rather high since they are all in the same one-species subtree.

Site scoring scheme

A scoring scheme containing bits cons_80, cons_90, A-G_mutation, A-G_both, annotated_AA_change and ds_G were used. The values of bits cons_80, and cons_90 was retrieved directly from the mm8 conservation labelling output and A-G_mutation, and A-G_both was similarly retrieved directly from the A-G mismatcher output correlated with the mm8 conservation labelling and StemPrediction respectively. In scoring annotated_AA_change we aligned amino acid sequences for a gene with the translated genomic mRNA using DIALIGN [Morgenstern, 1999]. The amino acid sequences was retrieved from NCBI *Entrez gene* [Maglott et al., 2007]. Either protein sequences from the *Entrez protein* or translated sequences from *Entrez nucleotide*. If a position annotated as an A-G mismatch also shows a corresponding amino acid discrepancy this site was scored. To compile sequence bias around an edited site (i.e. bit ds_G) we calculated the information content ± 200 nucleotides from a selected set of 24 edited adenosines from the known substrates.

$$H(l) = -\sum_{n=A}^{T} f(n,l)\log_2 f(n,l) \tag{1}$$

$H(l)$ is the uncertainty (entropy) [Shannon and Weaver, 1949] at position l. n is the 4 nucleotides to be summed over and $f(n,l)$ is the frequency of nucleotide n at l. The total information at position l is: $I(l) = 2 - H(l)$. From the information calculation we decided to bit score a downstream G.

454 amplicon sequencing

RNA was isolated from mouse brains at embryo day 15 and 19 and post natal day 2 and 21 using TRIzol (Invitrogen). Gabra3 had the addition of RNA from the post natal day 2. For the first-strand cDNA synthesis random primers was used. PCR was carried out with primers specific for known edited regions, see Table 1. Fused to the primers were adaptor oligonucleotids that were specific for the following sequencing procedure. Superscript III RT (Invitrogen) was used in all reverse transcription reaction, and FastStart High Fidelity PCR System (Roche) was used in all PCR reactions. To exclude that the samples was contaminated with genomic DNA, RT- controls was also carried out. Amplified PCR products was run on a 1.5% agarose gel and the expected bands was cut out and gel purified. All amplified PCR products from one developmental stage were pooled and the sample from Gabra3 P7 was added to the P2 aliquot and distinguished by 2 nt addition to the primer sequence. In the 454 procedure, the PCR products were immobilized on DNA capture beads. The bead/DNA were emulsified in a water-in-oil mix that contain reagents for amplification. Hence, one bead correspond to one fragment or transcript. The amplified fragments are loaded onto a PicoTiterPlate^{TM}- one bead/well=one read. The plate was then subjected to sequencing reagents using the pyro-sequencing technique (Roche).

Tables and figures

Conservation score	Sites	Gene overlapping sites	Areas	Gene overlapping areas	Stems
90	6713	4874	673	481	438
80-90	76503	59450	4385	3395	3397
70-80	243781	191259	19654	15619	40600
60-70	1299386	70587	1050411	56467	93004
50-60	3348784	2690862	97464	78472	83222
<50	53217663	42298490	N/A	N/A	1086937
Total	58192830	45315522			1307598

Table 1: Number of sites, gene overlapping sites, areas (see text for definition), gene overlapping areas, and predicted stems within various conservation score intervals are shown.

Cons.	$< 10^{-10}$		$10^{-10} - 10^{-5}$		$10^{-5} - 10^{-1}$		10^{-1}		Total	
score	Total	**A-G mm**	Total	**A-G mm**	Total	**A-G mm**	Total	**A-G mm**	Total	**A-G mm**
90	8	**1**	11	**2**	191	**25**	195	**23**	405	**51**
		12.5%		**18.2%**		**13.1%**		**11.8%**		**12.6%**
80-90	29	**13**	182	**58**	1545	**394**	1572	**443**	3328	**908**
		44.8%		**31.9%**		**25.5%**		**28.2%**		**27.3%**
70-80	665	**254**	1547	**483**	17877	**4459**	20238	**5645**	40327	**10841**
		38.2%		**31.2%**		**24.9%**		**27.9%**		**26.9%**
60-70	1314	**429**	3327	**806**	41222	**8386**	46741	**10628**	92604	**20249**
		32.7%		**24.2%**		**20.3%**		**22.7%**		**21.9%**
50-60	1038	**280**	2764	**512**	37164	**6091**	41992	**7568**	82958	**14451**
		27.0%		**18.5%**		**16.4%**		**18.0%**		**17.4%**
<50	11459	**1940**	34829	**4774**	448586	**53283**	589997	**78288**	1084871	**138285**
		16.9%		**13.7%**		**11.9%**		**13.3%**		**12.8%**
Total	14513	**2917**	42660	**6635**	546585	**72638**	700735	**102595**	1304493	**184785**
		20.1%		**15.5%**		**13.3%**		**14.6%**		**14.2%**

Table 2: The number of candidate stems in various conservation score intervals (rows), and microarray p-value intervals (columns) are shown. For each combination of conservation score and chip p-value is tabulated the total number of stems, the number of stems with A-G mismatch (typeset in bold), and the percentage of A-G mismatches (bold).

Site score	Description
Cons_80	The predicted stem has a conservation score of ≥ 80
Cons_90	The predicted stem has a conservation score of ≥ 90
AG_mutation	If a distant sub-tree has a DNA coded G at the position of an A-G mismatch, see also Figure 7.
AG_both	There are A-G mismatches on both stem arms
annotated_aa_change	The A-G mismatch results in an amino-acid discrepancy
ds_G	The nucleotide downstream of the A-G mismatch position is a G

Table 3: Filters used in the candidate scoring process.

	Parameter	Value	Description
BLASTZ	O	150	Gap opening cost
	E	100	Gap extension cost
	K	500	Maximal segment pair (MSP) score
	L	500	Gapped alignment threshold
	W	6	Word size
StemPrediction	MIN_ARM_LENGTH	16	Minimum stem arm length (bases)
	MAX_ENERGY	-15.0	Minimum free energy of the stem
	MAX_BULGE_SIZE	5	Maximum number of unpaired bases on a single strand in the stem
	MAX_BULGE_BASES	7	Maximum number of unpaired bases on both strand in the stem
	MAX_GLUE_DISTANCE	10	Maximum distance for two stems to be glued (joined)
	MAX_FILTER_ENERGY	-15.0	Minimum free energy of the glued stem

BLASTZ weight matrix

	A	C	G	T
A	80	-100	-100	-100
C	-100	120	-100	-100
G	20	-100	120	-100
T	-100	20	-100	80

Table 4: Upper table: Parameters used with BLASTZ and StemPrediction. Lower table: Weight matrix used with BLASTZ.

gene	codon_change	cons_80	cons_90	AG_mutation	AG_both	annotated_aa_change	ds_G	total sum
GluR-B	R:G	1	1	1	0	1	1	5
Adipor1	K:R	1	0	1	1	0	1	4
GluR-B	Q:R	1	0	1	0	1	1	4
Ccnc	Q:R	1	1	0	1	0	1	4
Elavl1	S:G	1	0	1	1	0	1	4
Gabarapl2	syn	1	0	1	1	0	1	4
Cnot2	N:S	1	1	1	1	0	0	4
Tra1	syn	1	0	1	1	0	1	4
Acin1	K:R	1	0	1	1	0	1	4
Eif4a2	syn	1	0	1	1	0	1	4
Gabra3	I:M	1	0	0	1	1	1	4
Eif4e2	K:R	0	0	0	1	1	1	3
Ptpra	Q:R	1	0	0	1	0	1	3
Etv3	Q:R	1	0	0	1	0	1	3
GluR-B	syn	1	1	0	0	0	1	3
GluR-B	I:V	1	1	1	0	0	0	3
Lmo4	K:R	1	0	0	1	0	1	3
Elavl2	K:R	1	0	0	1	0	1	3
Elavl2	syn	1	0	1	1	0	0	3
Stk22c	Q:R	1	0	0	1	0	1	3
Dhx15	Q:R	1	0	0	1	0	1	3
Fzd1	S:G	1	0	1	0	0	1	3
Ywhag	K:R	1	0	0	1	0	1	3
Kcna1	I:V	1	1	0	0	1	0	3
Kcna1	syn	1	0	0	1	0	1	3
Ptn	S:G	0	0	1	1	0	1	3
Arfip2	Q:R	0	0	1	1	0	1	3
Tial1	M:V	1	0	1	1	0	1	3
Gabarapl2	S:G	1	0	0	1	0	1	3
Crsp6	S:G	0	0	0	1	1	1	3
Ets1	syn	1	0	1	1	0	0	3
Atp5b	Q:R	1	0	0	1	0	1	3
Cnot2	M:V	1	1	0	1	0	0	3
Cnot2	Q:R	1	1	0	1	0	0	3
Cnot2	K:E	1	1	0	1	0	0	3
Tra1	K:R	1	0	0	1	0	1	3
Tra1	S:G	1	0	0	1	0	1	3
Cyfip2	Q:R	1	0	0	1	0	1	3
Nmt1	K:R	1	0	0	1	0	1	3
Sox9	K:R	1	0	0	1	0	1	3
Sox9	syn	1	0	0	1	0	1	3
Akt1	R:G	1	0	0	1	0	1	3
Evl	S:G	1	0	0	0	1	1	3
Kns2	N:D	1	0	1	1	0	0	3
Pcbp2	Q:R	1	0	0	1	0	1	3
Ap2m1	K:R	1	0	0	1	0	1	3
Ap2m1	Q:R	1	0	0	1	0	1	3
Actr1a	Q:R	1	0	0	1	0	1	3
Pten	Q:R	1	0	0	1	0	1	3
Hnrph2	Q:R	1	0	0	1	0	1	3
Hnrph2	K:R	1	0	0	1	0	1	3
Timm8a	K:R	1	0	0	1	0	1	3
Ube1x	N:S	1	0	0	1	1	0	3

Table 5: The final list of candidates (53) which are handpicked from all sites having a score ≥ 3 (124).

substrate			A-G mismatcher			
name	entrez gene	codon change	Mm	Hs	StemPrediction [a]	stemConservation [b]
Adar2	Adarb1	intron	n/a	n/a	yes	no
Bc10	Blcap	Y/C	yes	yes	yes	no
	Cyfip2	K/E	yes	yes	yes	no
	Flna	Q/R	no	yes	yes	no
	Ednrb	Q/R	n/a [c]	no [d]	yes	no
	Gabra3	I/M	yes	yes	yes	yes
GluR-B	Gria2	Q/R	yes	yes	yes	yes
		R/G	yes	yes	yes	yes
GluR-C	Gria3	R/G	no	yes	yes	yes
GluR-D	Gria4	R/G	yes	yes	yes	yes
GluR-5	Grik1	Q/R	no	yes	yes	no
GluR-6	Grik2	Q/R	no	no	yes	no
		Y/C	no	no	yes	yes
		I/V	no	no	yes	yes
5-ht2c	Htr2c	I/V_1	no	no	yes	yes
		I/V_2	no	no	yes	yes
		I/V_3	no	no	yes	yes
		N/S	no	no	yes	yes
	Igfbp7	R/G	yes	yes	yes	no
		K/R	yes	yes	yes	no
	Kcna1	I/V	yes	yes	yes	yes

[a] States whether Stemprediction has assigned any stem overlapping an edited position, regardless of the stem ranking (or if it is the correct one).

[b] States whether the stem according to column 4 has a conservation score ≥ 75.

[c] to our knowledge this site has not been confiremed in mouse which is also emphasized by low sequence similarity between the 2 species

[d] A-G mismatcher does not detect the annotated site but finds 2 additional A-G mismatches in the vicinity, inferring an I/M and a D/G codon change respectively.

Table 6: Compilation of the known ADAR substrates with respect to how they are captured by the pipe - noted by a *yes* or a *no* for the respective screen.

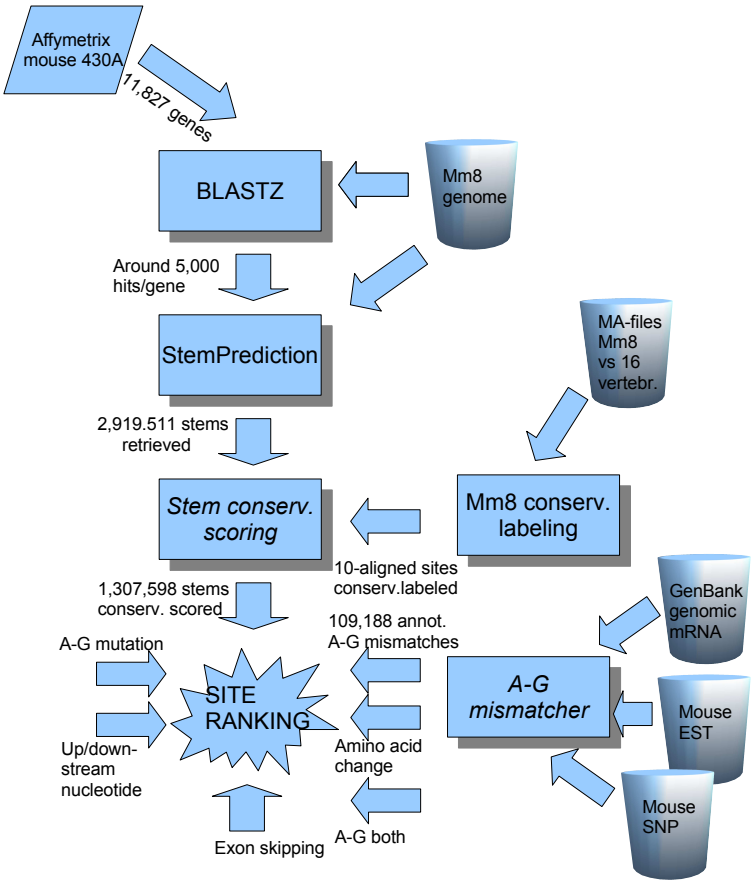


Figure 1: Flowchart describing the process of surveying and assigning potentially RNA A-to-I edited sites.

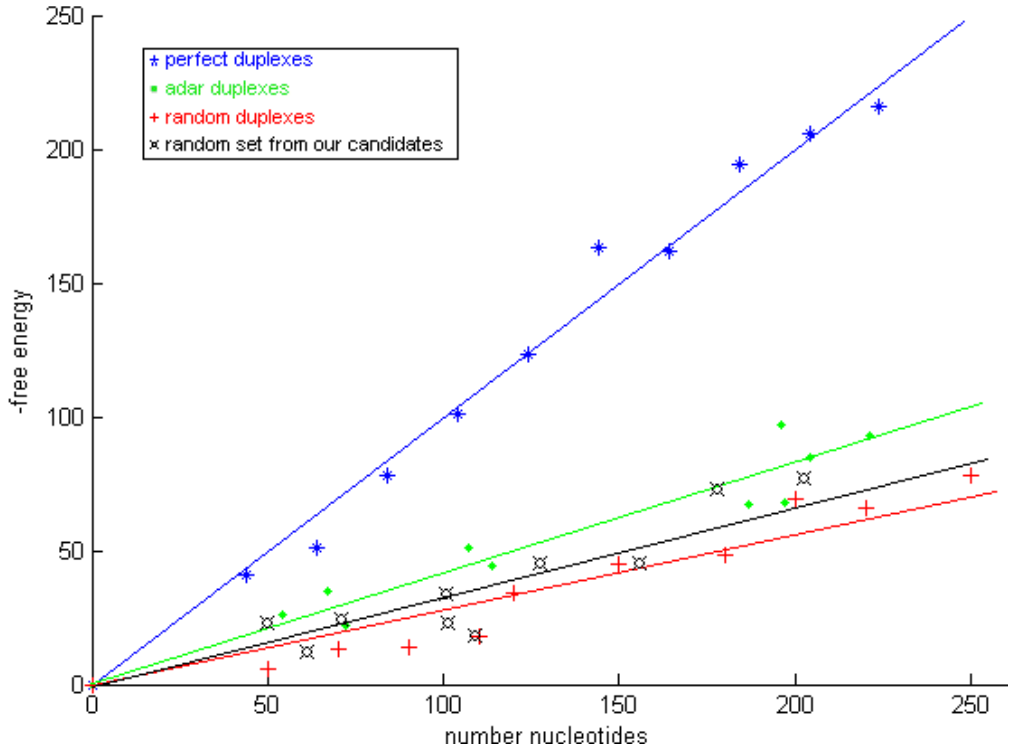


Figure 2: We plot the minimum free energy as a function of duplex length in nucleotides for ten examples of respectively perfect duplexes, known ADAR duplexes, random duplexes and our candidate duplexes. We conclude that the trend is clear in the assumption that we would benefit from not being too strict in assigning parameters to StemPrediction.

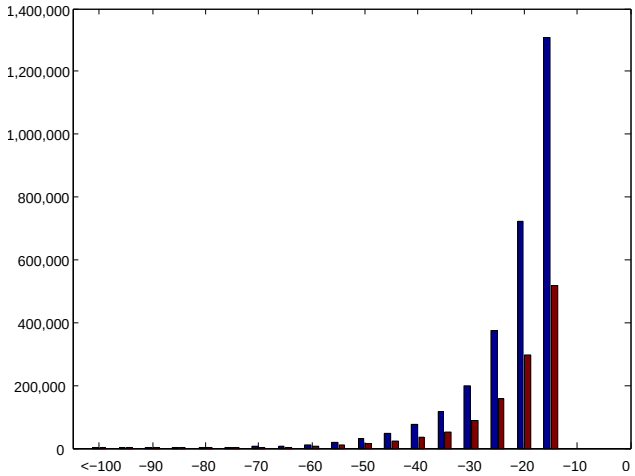


Figure 3: Distribution of free energy for all the 2,919,511 stems retrieved from StemPrediction (blue bars), and the 1,307,598 stems having a predicted stem loop shorter than 10,000 nucleotides (red bars).

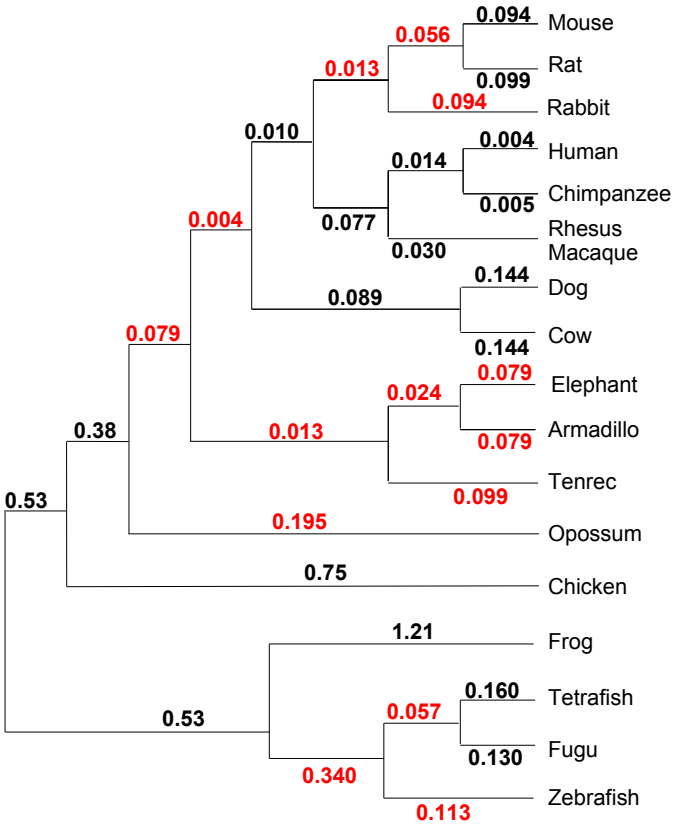


Figure 4: Phylogenetic tree relating the 17 vertebrate species used to evaluate conservation. Numbers on edges represent edge lengths measured in average substitutions per site. Black numbers are estimations made by Adam Siepel using PAML. Red numbers are estimated with the use of TimeTree [Hedges et al., 2006] assuming local molecular clocks.

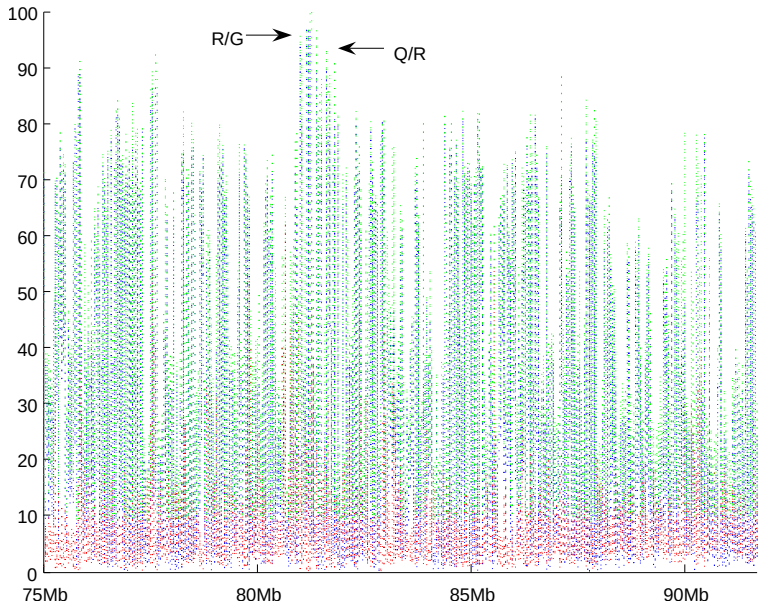


Figure 5: The conservation score distribution for section 75-92 Mb of Mm8 chromosome 3 is shown. The conservation score for a site (shown with the green curve) is the sum of the parsimony score (red curve) and the tree score (blue curve) for that site. Approximate conservation score for the genome positions of GluR-B R/G (conservation score for highest scoring stem arm = 96.5) and GluR-B Q/R (93.4) are specified.

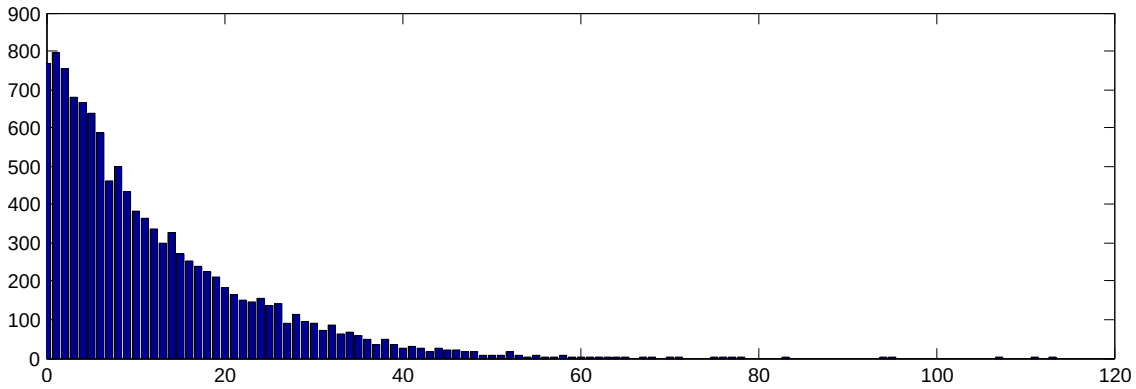


Figure 6: Distribution of the number of genes that overlap a certain number of A-G mismatches. Bars for genes Ubc (which overlap 188 A-G mismatches), Mll5 (231), and Spna2 (420), are not shown.

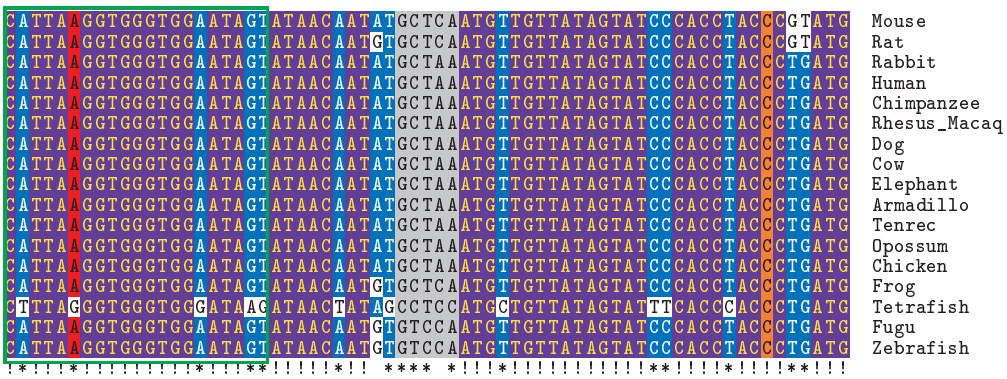


Figure 7: A 17-species alignment, visualized with TeXshade [Beitz, 2000], of the genomic region overlapping GluR-B R/G is shown. The column corresponding to the edited site is shown in red, while the complementary site is shown in orange. The loop is shown in grey. We note: (1) extreme conservation, (2) lost conservation in tetrafish, (3) the G in tetrafish in the edited column. The green rectangle surrounds a 21-column window used as an example in Methods.

References

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389–402.

[Aruscavage and Bass, 2000] Aruscavage, P. J. and Bass, B. L., 2000. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA*, **6**(2):257–69.

[Athanasiadis et al., 2004] Athanasiadis, A., Rich, A., and Maas, S., 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol*, **2**(12):e391.

[Bass, 2002] Bass, B. L., 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem*, **71**:817–46.

[Beitz, 2000] Beitz, E., 2000. TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*, **16**(2):135–139.

[Benson et al., 2007] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L., 2007. GenBank. *Nucleic Acid Res.*, **35**(database issue):D21–5.

[Blow et al., 2004] Blow, M., Futreal, P. A., Wooster, R., and Stratton, M. R., 2004. A survey of RNA editing in human brain. *Genome Res*, **14**(12):2379–87.

[Boguski et al., 1993] Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M., 1993. dbEST–database for "expressed sequence tags". *Nature Genetics*, **4**(4):332–3.

[Carlson et al., 2003] Carlson, C. B., Stephens, O. M., and Beal, P. A., 2003. Recognition of double-stranded rna by proteins and small molecules. *Biolpolymers*, **70**(1):86–102.

[Clutterbuck et al., 2005] Clutterbuck, D. R., Leroy, A., O'Connell, M. A., and Semple, C. A., 2005. A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics*, **21**(11):2590–2595.

[Eisenberg et al., 2005] Eisenberg, E., Adamsky, K., Cohen, L., Amariglio, N., Hirshberg, A., Rechavi, G., and Levanon, E. Y., 2005. Identification of rna editing sites in the snp database. *Nucleic Acids Res*, **33**(14):4612–4617.

[Hartner et al., 2004] Hartner, J. C., Schmittwolf, C., Kispert, A., Muller, A. M., Higuchi, M., and Seeburg, P. H., 2004. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J Biol Chem*, **279**(6):4894–902.

[Hasegawa et al., 1985] Hasegawa, M., Kishino, H., and Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, **22**(2):160–74.

[Hedges et al., 2006] Hedges, S. B., Dudley, J., and Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**(23):2971–2972.

[Higuchi et al., 2000] Higuchi, M., Maas, S., Single, F. N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P. H., 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature*, **406**(6791):78–81.

[Hoeffding, 1956] Hoeffding, W., 1956. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, **27**(3):713–721.

[Hofacker et al., 1994] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P., 1994. Fast folding and comparison of rna secondary structures. *Monatshefte f. Chemie*, **125**:167–188.

[Hoopengardner et al., 2003] Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R., 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science*, **301**(5634):832–836.

[Karolchik et al., 2003] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., *et al.*, 2003. The UCSC genome browser database. *Nucleic Acids Res*, **31**(1):51–54.

[Kent, 2002] Kent, W. J., 2002. BLAT – the BLAST-like alignment tool. *Genome Res*, **12**(4):656–64.

[Lehmann and Bass, 1999] Lehmann, K. A. and Bass, B. L., 1999. The importance of internal loops within rna substrates of adar1. *J Mol Biol*, **291**(1):1–13.

[Lehmann and Bass, 2000] Lehmann, K. A. and Bass, B. L., 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry*, **39**(42):12875–84.

[Levanon et al., 2004] Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., *et al.*, 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol*, **22**(8):1001–1005.

[Li et al., 2007] Li, Q., Lee, J. A., and Black, D. L., 2007. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci*, **8**(11):819–31.

[Maas et al., 2003] Maas, S., Rich, A., and Nishikura, K., 2003. A-to-I RNA editing: recent news and residual mysteries. *J. Biol. Chem*, **278**(3):1391–4.

[Maglott et al., 2007] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T., 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, **35**(Database issue):D26–31.

[Margulies et al., 2003] Margulies, E. H., Blanchette, M., Haussler, D., and Green, E. D., 2003. Identification and characterization of multi-species conserved sequences. *Genome Res*, **13**(12):2507–18.

[Maydanovych and Beal, 2006] Maydanovych, O. and Beal, P. A., 2006. Breaking the central dogma by RNA editing. *Chem Rev*, **106**(8):3397–411.

[Morgenstern, 1999] Morgenstern, B., 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**(3):211–218.

[Morse and Bass, 1999] Morse, D. P. and Bass, B. L., 1999. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proc Natl Acad Sci U S A*, **96**(11):6048–53.

[Ohlson et al., 2005] Ohlson, J., Ensterö, M., Sjöberg, B. M., and Öhman, M., 2005. A method to find tissue-specific novel sites of selective adenosine deamination. *Nucleic Acids Res*, **33**(19):e167.

[Ohlson et al., 2007] Ohlson, J., Pedersen, J. S., Haussler, D., and Öhman, M., 2007. Editing modifies the GABA(A) receptor subunit alpha3. *RNA*, **13**(5):698–703.

[Palladino et al., 2000] Palladino, M. J., Keegan, L. P., O'Connell, M. A., and Reenan, R. A., 2000. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell*, **102**(4):437–49.

[Pedersen et al., 2006] Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D., 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, **2**(4):e33.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M., 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**(20):6097–100.

[Schwartz et al., 2003] Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W., 2003. Human-mouse alignment with BLASTZ. *Genome Res*, **13**(1):103–107.

[Shannon and Weaver, 1949] Shannon, C. E. and Weaver, W., 1949. *The mathematical theory of communication*. The University of Illinois Press, Urbana, Illinois.

[Sherry et al., 2001] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**(1):308–11.

[Tian et al., 2004] Tian, B., Bevilacqua, P. C., Diegelman-Parente, A., and Mathews, M. B., 2004. The double-stranded-rna-binding motif: interference and much more. *Nat Rev Mol Cell Biol*, **5**(12):1013–23.

[Wang et al., 2000] Wang, Q., Khillan, J., Gadue, P., and Nishikura, K., 2000. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science*, **290**(5497):1765–1768.

Paper IV

An in-depth survey of A-to-I editing implies a general developmental regulation and coupling of edited sites

Mats Ensterö, Chammiran Daniel, Helene Wahlstedt and Marie Öhman

Department of Molecular Biology and Functional Genomics, Stockholm University, S-106 91 Stockholm, Sweden

Abstract

ADAR mediated A-to-I editing has been shown to be an important finetuning mechanism for protein diversity. ADAR targets adenosines within RNA duplex structures and converts them to inosines by a hydrolytic deamination. However, the mechanism of substrate recognition for the ADAR enzymes is still largely unknown. In mammals the majority of all selectively edited A-to-I sites have been found in the brain within genes involved in neurotransmission. It has also been shown that editing is important for a normal development of the brain. Here, we analyze changes in the editing pattern of most of the known site selectively A-to-I edited substrates in the developing mouse brain. The coupling between edited adenosines in regions where editing have been shown to cluster is also analyzed. We use the 454 amplicon sequencing™ technique for editing determination and χ^2-test to compile a coupling scheme of edited adenosines. To our knowledge, this is the first time 454 sequencing have been used to determine editing frequency. Using this technique we can study editing of single transcripts in an extent that has previously not been possible. In average, we analyzed 650 single transcripts per edited region for each developmental stage. We show that most selectively edited sites are developmentally regulated with low levels of editing during embryogenesis but increasing gradually with age until adulthood. Further, an extensive coupling scheme show that edited sites are only coupled within specific distances from each other indicating that the helical structure of the substrate is important for how a substrate is recognized by the editing enzyme.

Introduction

Adenosine to inosine (A-to-I) editing was discovered in 1988 when an antisense RNA failed to block translation of a target transcript. The reason for this was that most *adenosines* in the antisense RNA had been deaminated to *inosines* hence disrupting the anticipated hybridization properties of the target (Bass et al., 1988). This type of abundant editing has later been characterized as *hyper*-editing in contrast to *site selective* editing. Site selective editing targets single adenosines within an imperfect RNA foldback structure while *hyper*-editing indiscriminately edits multiple adenosines in longer almost completely duplexed structures. Untranslated regions of human transcripts consisting of ALU repeats/inverted repeats has recently been found to be hyper edited and is believed to correspond to the bulk of editing events in the human transcriptome (Athanasiadis et al., 2004) (Levanon et al., 2004). The catalytic activity responsible for the deamination of adenosines within duplexed RNA was first recognized in 1994 (Polson et al., 1994). Later a family of ADAR (adenosine deaminase that acts on RNA) proteins have been identified, ADAR1-3. However, only ADAR1 and ADAR2 have been shown to have a catalytic activity in mammals (Melcher et al., 1996) (Chen et al., 2000). Inosine is interpreted as a guanosine (G) by the cellular machineries and also by the reverse transcriptase during cDNA synthesis. If the editing event occurs within the coding region of an mRNA it can give rise to amino acid changes and variant functional properties of the final protein. Also intronic editing has the potential to change the translated protein by either disrupt or constitute splice sites. The properties that make an RNA prone for *site selective* editing is still not fully understood but the assumption is that internal mismatches and bulges within a foldback structure are important for ADAR selectivity (Källman et al., 2003) (Dawson et al., 2004) (Stephens et al., 2004). There is also a bias in the nearest neighbor preference to an edited site where the 5' upstream nearest neighbor rarely constitutes a G while the most common nucleotide as the 3' nearest neighbor is G (Lehmann et al., 2000) (Ensterö and Åkerborg, 2008 unpublished). Further, there is a preference for cytosine opposing the targeted adenosine, although this preference seem more pronounced for ADAR1 than ADAR2. (Wong et al., 2001).

Most of the known mammalian editing sites have been found in transcripts coding for proteins involved in neurotransmission. This includes the AMPA glutamate receptor subunits (GluR-B, C and D), the kainate glutamate receptor subunits (GluR-5 and -6) (Higuchi et al., 1993) (Lomeli et al., 1994) (Sommer et al., 1991) and the serotonin receptor 2C (5-HT_{2C}) (burns et al., 1997) (Liu et al., 1999). More recently, Gabra-3 coding for the α3 subunit of the $GABA_A$ receptor and the potassium voltage gated ion channel KCNA1 (K_v1.1) has been shown to be edited (Ohlson et al., 2007)

(Bbhalla et al., 2004). In most of these substrates the editing event have been shown to have functional consequences on the receptor. ADAR2 has also been shown to auto-edit its own pre-messenger RNA (Dawson et al., 2004). In this case a new splice site is created upon editing giving rise to a truncated protein. Editing within transcripts coding for proteins that are not brain specific has also been detected in the bladdercancer associated protein (BLCAP), cytoplasmic FMR1 interacting protein 2 (CYFIP2), and filamin A (FLNA) (Levanon et al., 2005). However the functional consequences of the editing event in these substrates are not known.

ADAR1 and ADAR2 have specific but overlapping specificities for site selective editing. In the serotonin receptor 2C (5-HT_{2C}), the A-site seems prone for ADAR1 editing while the D- and GluR-B Q/R sites are almost exclusively edited by ADAR2 (Burns et al., 1997) (Higuchi et al., 2000). Most of the other sites has the potential be edited by both ADAR1 and ADAR2. It has been suggested that it is the deaminase domain of each ADAR that delegate the target specificity (Wong et al., 2008).
It is known that the Q/R site in GluR-B is highly edited during early embryogenesis and continues to be efficiently edited to almost 100% in the adult brain (Seeburg et al., 1998). On the contrary, the R/G site of GluR-B is inefficiently edited in the embryonic states but increase during the development of the brain (Lomeli et al., 1994) (Bernard et al., 1994) (Barbon et al., 2003). Further, we have recently shown that editing of the Gabra-3 transcript of the $GABA_A$ receptor increase from 50% in the newborn mouse up to close to 100% in the adult mouse brain (Ohlson et al., 2007).

In this study we have used *454 amplicon sequencing*™ to compile a substantiated survey of editing frequencies of most of the known site selectively A-to-I edited substrates in the mammalian brain. The number of reads from the experimental procedure give us a large population of sequenced target RNA:s. Hence, we have a well founded statistical basis to deduce both editing frequencies and the coupling of nearby editing events.

We present a near complete list of editing frequencies in the mouse brain during early embryogenesis up to adulthood with unprecedented resolution. We show that site selective editing in general is developmentally regulated. Most of the edited sites are inefficiently edited during early embryogenesis, editing increase gradually and does not reach a maximum until day 21. The resolution also gives us the opportunity to disclose any coupled events between nearby edited sites. We have chosen 3 target regions where there are multiple editing events within an RNA foldback structure to investigate if there is coupling between edited sites. These targets are the pre-mRNAs of GluR-6 (I/V and Y/C), 5-HT_{2C} (A, B, E, C and D sites) and Adar2 (-28 to +28 sites). Our data indicate that there is a coupling between

edited sites at fixed distances from each other. This result indicates that the ternary structure of the RNA substrate is important for the coupling.

Results

Site selective editing increase during brain development

We wanted to investigate if the editing pattern of selectively edited sites changes during the development of the mouse brain. To analyze editing frequencies, RNA known to be targets for A-to-I editing were isolated and amplified by reverse transcriptase followed by a polymerase chain reaction (RT-PCR). The products were subsequently sequenced according to the *454 amplicon sequencing*™ protocol (Margulies et al., 2005). This sequencing procedure gives rise to a large amount of sequenced data (reads) that can be treated with high statistical significance. One advantage with the use of this technique for editing analyses is that even a small number of edited transcripts can be detected since often over a thousand reads are obtained per substrate. All sequences corresponding to the individual A-to-I editing substrates were aligned and analyzed for A to G changes.

The extent of editing was determined for 4 different developmental stages of the mouse brain: embryonic day 15 and 19, (E15 and E19) as well as postnatal day 2 and 21, (P2 and P21) (Table 1 and Fig. 1). One editing substrate, Gabra3, was also quantified at postnatal day 7, (P7).

There is a general trend that the editing patterns of the substrates show developmental regulation with very low levels of editing during embryogenesis that increase over time. There is however a variation in how gradual the increase in editing is. The Flna Q/R site show less than 7% editing in E15, E19 and P2 but the efficiency increase drastically at the adult P21 stage to 43% editing (Table1 and Fig. 1B). A similar pattern is seen for the I/V site of Kcna1 but with surprisingly low levels of constitutive editing (25%) at P21 compared to what previously has been observed (Fig. 1C) (Bhalla et al., 2004). In contrast to other sites, the GluR-B Q/R and the C and D sites of 5-HT_{2C} are quickly saturated to their adult editing levels with only a moderate increase from P2 to P21 at the D site (Fig. 1G and 1I). The Q/R site has previously been shown to be efficiently edited during early embryogenesis (Seeburg et al., 1998). However, as previously known, editing of the GluR-B R/G site, situated in the same transcript as Q/R, increase gradually during the development (Fig. 1G) (Lomeli et al., 1994). In a similar way the R/G site of GluR-C, Q/R of GluR-5, Cyfip2 K/E and the auto-editing of the Adar2 transcript increase during the development of the brain. The same is true for the B site in the 5-HT_{2C} transcript situated in close proximity to the C and the D sites. Perhaps the most defined gradual increase in editing during the development can be seen at the I/M site of Gabra-3 (Fig. 1F).

Here only 6% of the transcripts are edited at E15 while 92% are edited at P21. An almost 5 fold increase in editing is seen between embryonic day 15 and 19 and a 20% increase between postnatal day 7 and 21.

In summary, our data reveals that most sites of selective editing in the mammalian brain are regulated during brain development. This regulation cannot entirely be explained by the presence of ADAR enzymes during the development since the Q/R site of GluR-B is almost 100% edited at early embryogenesis. Further, there are no indications that the expression of ADAR1 and ADAR2 changes during the development (Wahlstedt and Öhman, unpublished).

Edited sites within the same transcript are coupled in a defined way

By using the 454 sequence method to analyze editing events we have the possibility to statistically analyze each target transcript individually. We wanted to determine if distinct editing events show any combinatorial behavior. That is, if position N in a target transcript is edited, position N' is also edited. Or mutually exclusive so that if N is not edited, N' isn't either. We have chosen regions of transcripts surrounding the -1 site in Adar2, the A to D sites in 5-HT_{2C} and the I/V and Y/C sites in GluR-6 (Table 2). The edited positions reveals that there is an apparent pattern of "hot-spot" editing at a semi-fixed distance from each other. There are 3 positions around +24, the position +10, 3 positions centered by the -2 site and the 2 positions at -28, -27. Position +10 has a spacer distance of approximately 11-14 nucleotides (nt) from the nearby hot-spots and the distance between hot-spot -2 and -27 is 2 x that spacer distance. Between +10 and -27 there is 3 x 12 = 36 nt. Noteworthy is that there are adenosines in between these hot-spots that are not edited. Interestingly, the distance between the GluR-6 I/V and Y/C sites is also 13 nt. Our hypothesis is that there is no coincidence that a spacer distance is preferably ~12 nt between apparent hot-spots. We use a χ^2-test, with significance p=0.05 and 1 degree of freedom to see whether 2 separate sites are edited in a coupled or un-coupled manner. The null hypothesis is that the positions are independently edited (see also Methods). These sites and their sequential context and a schematic χ^2 matrix are shown in Figure 2. Along the green diagonal, there is a pattern where either the two positions are edited (G and G) or not edited (A and A) (Fig. 2). The red diagonal follow entries that correspond to the reverse, A and G or G and A. Hence, if either diagonal is significantly favored, for example: a + d > c + b, they are considered to be positively coupled. A negative coupled pair of positions would significantly favor the elements along the red diagonal. Since we perform multiple statistical measurements of coupled sites, i.e., we calculate 36 different χ^2 values from 36 2x2 matrices in the Adar2 region, we make a Bonferroni correction to the p-value (Bonferroni, 1935).

The motivation for this is that if we assign p to 0.05 we statistically anticipate to falsely reject the null-hypothesis exactly 1 time if we perform a measurement 20 times (1/20=0.05). It should be noted that the usage of a Bonferroni correction of the p-value is considered to be very strict and by its implementation we increase the quality level our results although we might loose in sensitivity (for further details see Methods).

In addition to the χ^2-test, we decided to make cluster analyses. Therefore, mutual editing patterns were compiled into classes where members in a class share a common feature. We use it as a complement to the χ^2-test due to the intrinsic illustratively way we can deduce coupling patterns from the resulting dendrograms, (Fig. 3 and 4). Overall, there are several sites that are positively coupled (Table 2). We consider the coupling to be weak (+) if only one of the tests show coupling, a strongly positive coupling (++) is indicated if both the χ^2-test and the cluster analyses show coupled properties. In general the same positions are coupled at the different developmental stages, making the results of the coupling analysis even stronger.

In the 5-HT_{2C} transcript, the A is coupled to both the B and the C site at all four stages (Table 2 and Fig. 4). The following sites are coupled in at least 2 stages: A to D; B to C; B to D and C to D. The A to B coupling has previously been stated (Liu et al., 1999). Interestingly, the A and D sites have been shown to be preferentially edited by different ADAR enzymes, the A site by ADAR1 and the D site by ADAR2 (Higuchi et al., 2000) (Burns et al., 1997). Strikingly, the E site in 5-HT_{2C} is consistently detached from the other classes in the cluster analyses. At stage P2 it is even clearly negatively coupled to the A, B and D sites and show no coupled properties at the other 3 stages.

There are also 3 sites in the Adar2 region that are coupled in all stages: -1 to 10; 10 to 24 and -1 to 24 (Table2 and Fig. 3). In Adar2, we can in addition to the sites showing coupled behavior in all 4 stages also conclude that coupling occurs between sites -27 and 10; -4 and 23; -2 and -1; -2 and 10; 10 and 23; 23 and 24 as well as 24 and 28 in at least 2 stages. Site 28 is novel and has previously not been annotated. However, only 10% of the transcripts at P21 are edited at this site. Also in this case coupling is seen between sites that are edited by different enzymes; -1 have previously been assigned to ADAR1 and -2 to ADAR2 (Higuchi et al., 2000).

The coupling does not discriminate between sites that are edited by ADAR1 or ADAR2. Rather, in several cases a coupling occurs between sites edited by different enzymes indicating that there is a communication between ADAR1 and ADAR2 on the same transcript.

Coupling of edited sites occurs at defined distances from each other

An interesting question based on the coupling results, is whether there is a consensus of the spacer regions between coupled sites. In order to see such a pattern we calculated all distances between sites that showed coupling (see distance, Table 2). For every possible spacer distance, 1 to 55, we counted the number of coupled sites having that spacer distance in the Adar2 transcript (Fig. 5). There is a clear pattern of preferred spacer distances of coupled editing events. The peaks in the graph are centered on the average number of the total number of coupled sites within the clustered spacer distances. These distances are 1 to 4, 10 to 14, 23 to 26, 36 to 37 and 50 to 51. It is evident that the most common spacer distance, clusters at 23-26 nt with 7 coupled positions in this region. The cluster centered at ~2 nt is probably a result of a different protein/RNA interaction then the other coupled distances, where one active enzyme targets adjacent adenosines. Interestingly, we can see a similar strong correlation of mutual editing at the I/V and Y/C sites in GluR-6, here with a distance of 13 nt from each other. Thus, there is a clear pattern of coupled sites separated by multiples of ~12 in both of these substrates. Therefore, there is no coincidence that adenosines between hot-spots are not edited.

In the 5-HT_{2C} transcript the space between coupled edited positions is somewhat different from Adar2 and GluR-6, simply because the maximum number of nucleotides between the sites that are edited are 12 and we can not detect coupled sites separated by more than that. Nevertheless, the A and D sites which are separated by 12 nucleotides show strong coupling. The A and B sites, separated by 1 nucleotide, are also positively coupled, hence following the pattern seen in the cluster centered at 2 nt in Adar2. However, coupled positions separated by 7 (A and C), 5 (C and D) and 10 (B and D) are also apparent in the 5-HT_{2C} transcript. We therefore conclude that the ADAR enzyme(s) binds in a different way to the 5-HT_{2C} than to the other two substrates investigated, particularly since there are uncoupled and even negatively coupled positions in combinations with the E site.

Discussion

We have used the *454 amplicon sequencing protocol*™ to evaluate single transcript A-to-I editing in brain. In this study, the editing efficiency of most of the known mammalian editing substrates are being analyzed at early embryogenesis up to the adult mouse. The amplified single transcripts are evaluated statistically to deduce editing frequencies for these sites at 4 different developmental stages. In general, we see developmentally regulated editing with increased frequencies over time. For

some of the sites this has previously been observed (Barbon et al., 2003) (Lomeli et al., 1994) (Bernard et al., 1994).

As previously mentioned, ADAR1 and ADAR2 have distinct specificities for editing of some of the sites. However our data indicate no consensus editing pattern between any of these sites implying no consistent connection that separate ADAR1 editing from ADAR2 during development. For example, we have a near constant editing frequency for the ADAR2 edited GluR-B Q/R and 5-HT_{2C} D sites while the GluR-B R/G site, Adar2 -1 site and GluR-6 Y/C increase during the development. The same pattern could be seen for the ADAR1 specific targets. We conclude that the E-site has an altered editing pattern compared to the other sites in 5-HT_{2C} and is edited mechanistically independent from the other sites. Since the A- and D-sites are also coupled although assigned to the two different ADARs, this could imply a non-canonical dimer not seen (or active) on other substrates. To further support the aberrant editing behavior of the 5-HT_{2C} sites, we see spacer distances (5 and 7) not seen at other coupled sites in the Adar2 transcript.

The ADAR enzymes are believed to be non-processive and deaminate, at most, neighboring adenosines (Bass, 2002). Also, as previously observed, ADAR2 associates more strongly (selectively) to an imperfect RNA foldback structure than to a perfect RNA duplex within the same molecule (Klaue et al., 2003). Consequently, two types of patterns could be expected: 1/ Multiple sites that showed a positive coupling. In this case there is a strong positive coupling of sites with a certain distance from each other consistent with n x 12, where n=1,2,3,4 but also for neighboring adenosines. 2/ Few sites with negative coupling. Our result supports both of these patterns and it also correlates with what has previously been reported in footprinting analyses where ADAR2 protects 11-16 base pairs of the R/G stem loop (Öhman et al., 2000).

The A coupling scheme has the potential to resolve some of the ADAR/RNA interactions properties. A dimeric interaction with two active catalytic sites facing the RNA could result in coupled sites with defined intramolecular spacer regions. In support of the current opinion of dimeric function of ADARs, we find coupled sites even leaving room for multi-dimeric binding of a target RNA. A monomer ADAR targeting of the RNA foldback structure would, at most, favor no other adenosines than those fulfilling known preferred criteria (i.e., nearest neighbor etc). Consequently, we would not expect multiple edited sites within a recognized target RNA. Also, there would be no obvious reason for ADAR(s) to target adenosines separated by defined spacer distances.

We hypothesized that ADAR editing of multiple site regions of "hot-spots" would show a pattern of distinct coupled positions since there is an apparent equidistance of edited hot-spots in the Adar2 transcript. Taken together the positively coupled

positions have two distinct spacer regions; 1-2 nucleotides; and multiples of ~12 nucleotides. Coupled positions of the adjacent edited sites are probably a result of an ADAR slipping to neighboring adenosines. As previously suggested, upon ADAR/RNA binding, adjacent adenosines could be sequentially edited before disassociation (Bass, 2002).

We see high degree of editing at site +24 (80%, P21) and decreasing amounts of editing following the hot-spots upstream to -28 (9% P21). Our hypothesis is that the +24 site is the principal editing site attracting a dimer ADAR. Upon the first dimer interaction, consecutive dimers bind in register to deaminate subsequent adenosines *if* fulfilling the optimal sequential/structural environment. This model would explain the coupled positions with defined spacer distances and the fact that we see lesser degree of editing further upstream of +24 (Fig. 6). A preliminary model that would be completed with corresponding data from the editing of the other strand suggest that the pattern we see is consecutive binding of dimers either by dimer/dimer interaction or binding in register for sterical reasons.

This model fits poorly with the editing pattern of the 5-HT_{2c} receptor and the distances between coupled positions therein. In this transcript there are coupled edited positions also distanced by 5, 7 and 10 nt. We find two possible explanation for this. First of all, the region on the 5-HT_{2c} pre-mRNA including the A to D sites, is perturbed by two large asymmetric bulges. The constraints on the recognition criteria and ADAR interface is thereby hard to realize. Non-canonical binding properties could therefore not be excluded. Secondly, a recent study show that in a certain neurons (ar2a) with low expression of ADAR3 but high expression of ADAR1/2 the level of editing is severely decreased at all five serotonin sites (Ssergeeva et al., 2007). The opposite is valid for the GluR-B which is highly editied in absence of ADAR3 expression. Together with the implication that the E-site is negatively coupled to some of the other sites, the only thing that can cause such an alteration is changes in the protein mechanistic behavior since the target remain the same. Our data support the non-canonical editing pattern of the 5-HT2C. We therefore suggest a different, yet not fully understood, mechanism of the editing pattern of the serotonin receptor.

Taken together, we propose synchronized editing event of consecutive dimers that bind in register on substrates with multiple editing sites that fulfill the criteria of being site selectively edited.

Methods

454 amplicon sequencing

RNA was isolated from mouse brains at embryo day 15 and 19 and post natal day 2 and 21 using TRIzol (Invitrogen). For the Gabra3 substrate additional RNA was extracted at post natal day 2. For the first-strand cDNA synthesis random primers were used. PCR was carried out with primers specific for the known edited regions, see Table 1. The sequence of the primers can be provided by the authors upon request. Adaptor deoxyoligonucleotids specific for the sequencing procedure were fused to the primers according to the instructions for *454 amplicon sequencing*™ by the provider (Roche). Superscript III reverse transcriptase (Invitrogen) was used in all reverse transcription reactions, and FastStart High Fidelity PCR System (Roche) was used in all PCR reactions. Amplified PCR products were purified on a 1.5% agarose gel. All amplified PCR products from one developmental stage were pooled to a final concentration of 5 ng/µl per sample. The PCR product from Gabra3 P7 was added to P2 aliquot and distinguished by 2 nt addition to the primer sequence. The products were sequenced using the 454 amplicon sequencing™ technique (Margulies et al., 2005) according to the instructions by the manufacturer (Roche)

Identification of edited transcripts from the 454 sequencing

For the subsequent collection and compilation of the data from the 454 sequencing we used *in-house* scripts. The correct gene tag to a sequence was recognized by the gene specific primer that initiated each sequence. We used DIALIGN 2 (Morgenstern et al., 1999) to create multiple sequence alignments (MSA:s). For each edited and sequenced region we aligned the corresponding data and included genomic data in which the coordinates were known for the expected edited site. For all positions of the known editing events we calculated the proportion between A:s and G:s. For the 5-HT_{2C} and Adar2 edited regions we used a different approach, since multiple editing events in a limited region inferred low quality to the alignment in the same region. To ensure not to include mis-aligned A:s or G:s in the wrong category we used a pattern matching approach. In the sequence spanning the edited region, all positions containing an edited A were classified with a logical OR, i.e., either an A OR G. Hence, the pattern sequence would match the correct 454 sequence regardless of A/G ambiguities. Inherent in Perl, all A:s and G:s could be directly calculated through the matching.

Determination of editing frequencies and coupling

The error estimation of editing frequencies is given in Table 1 and was calculated by:

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

where p is the proportion of G:s (or editing frequency) and n is the sample size, in this case number of reads that were used.

In order to asses patterns of coupling we use a χ^2-test, using an *i* x *j* matrix where *i*=number of rows and *j*=number of columns. The matrix contain a dataset where we statistically look for differences that let us either reject or accept the null-hypothesis of independent events. We use a 2 x 2 matrix with the possible outcomes of two editing events at separate positions: G_1 AND G_2, G_1 AND A_2, A_1 AND G_2 and subsequently A_1 AND A_2 (the subscript assigns the nucleotide to either position 1 or 2) (Fig. 2). The null-hypothesis is that the editing events are independent events. Since we perform multiple measurements of χ^2 we also make a *Bonferroni* correction where we use an adjusted p value p´ where p´ = p / n, and n is the number of measurements and p = 0.05 (bonferroni, 1935).

Cluster analyses were used as a complement to the χ^2-test. We used an Excel add-in, XLSTAT (Xlstat, 2008) to do the cluster analyses.

With the χ^2-test as a starting point we could for example group positions n and n' with positions k and k' if they were categorized as the same class although not directly evident from the χ^2 analyses. Our data can statistically be reduced to binary data since a position qualitatively is either edited or not, i.e., 1 or 0. Therefore we have used Dice coefficients in a similarity matrix (or dissimilarity depending on how we use the data).

The edited region of Adar2 have 9 positions with the potential of being edited., We state 1 or 0 for each position in each sequence of this transcript. The similarity matrix contains elements calculated from D_{ij} = 2ad / (2ad + b + c),. Consequently, "a" is the total number of 1 and 1 (edited and edited) for a position pair. Further, "b" and "c" are 1 and 0 or 0 and 1, while "d" is 0 and 0. The reason for us to choose Dice coefficients in the similarity matrix is seen from the formula where we put twice the weight on agreements (1 and 1) in a position pair, hence the factor 2ad. This is in contrast to the χ^2 analyses where 0 and 0 is weighted equally to 1 and 1, we wanted to emphasize the actual editing events.

Acknowledgement:
We are thankful to Jan-Olov Persson for the statistical input and advice. This work was supported by Wallenberg Consortium North and the Swedish Research Council.

References:

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.

Barbon A, Vallini I, La Via L, Marchina E, Barlati S. 2003. Glutamate receptor RNA editing: a molecular analysis of GluR2, GluR5 and GluR6 in human brain tissues and in NT2 cells following in vitro neural differentiation. *Brain Res Mol Brain Res 117*:168-178.

Bass BL, Weintraub H. 1988. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell 55*:1089-1098.

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

Bernard A, Khrestchatisky M. 1994. Assessing the extent of RNA editing in the TMII regions of GluR5 and GluR6 kainate receptors during rat brain development. *J Neurochem 62*:2057-2060.

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.

Bonferroni, C. E. "Il calcolo delle assicurazioni su gruppi di teste." In Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, pp. 13-60, 1935.

Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature 387*:303-308.

Chen CX, Cho DS, Wang Q, Lai F, Carter KC, Nishikura K. 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.

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.

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.

Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, Feldmeyer D, Sprengel R, Seeburg PH. 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature 406*:78-81.

Higuchi M, Single FN, Kohler M, Sommer B, Sprengel R, Seeburg PH. 1993. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell 75*:1361-1370.

Kallman AM, Sahlin M, Ohman M. 2003. ADAR2 A-->I editing: site selectivity and editing efficiency are separate events. *Nucleic Acids Res 31*:4874-4881.

Klaue Y, Kallman AM, Bonin M, Nellen W, Ohman M. 2003. Biochemical analysis and scanning force microscopy reveal productive and nonproductive ADAR2 binding to RNA substrates. *RNA 9*:839-846.

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

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol 22*:1001-1005.

Liu Y, Emeson RB, Samuel CE. 1999. Serotonin-2C receptor pre-mRNA editing in rat brain and in vitro by splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase ADAR1. *J Biol Chem 274*:18351-18358.

Lomeli H, Mosbacher J, Melcher T, Hoger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science 266*:1709-1713.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*:376-380.

Melcher T, Maas S, Herb A, Sprengel R, Higuchi M, Seeburg PH. 1996. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J Biol Chem 271*:31795-31798.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics 15*:211-218.

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

Ohman M, Kallman AM, Bass BL. 2000. In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. *RNA 6*:687-697.

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

Seeburg PH, Higuchi M, Sprengel R. 1998. RNA editing of brain glutamate receptor channels: mechanism and physiology. *Brain Res Brain Res Rev 26*:217-229.

Sergeeva OA, Amberger BT, Haas HL. 2007. Editing of AMPA and serotonin 2C receptors in individual central neurons, controlling wakefulness. *Cell Mol Neurobiol 27*:669-680.

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.

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

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

XLSTAT. Addinsoft (2008, DEMO). XLSTAT 2008, Data Analysis and Statistics Software for Microsoft Excel. Paris, France.

Figure legends.

Figure 1, A-K
A graphical presentation of the editing frequencies of most of the known Adar2 substrates from the different developmental stages E15, E19, P2 and P21. For abbreviations, se text. For Gabra3 (F), P7 has been included as shown. All figures have the same scale for comparative reasons and due to pixel resolution, exact error estimation should be read from table 1. Also, x-axis spacing is only grouped by developmental stage, not linear in time. Abbreviations used: GluR-B: Glutamate receptor b. Gabra3: gamma-aminobutyric acid A receptor. Adar2: Adenosine deaminase acting on RNA 2. 5-ht2c: Serotonin receptor 2c. Cyfip2: cytoplasmic FMR1 interacting protein 2. Flna: Filamin A. Kcna1: potassium voltage-gated channel, shaker-related subfamily, member 1. Blcap: Bladder cancer associated protein

Figure 2.
A. An example of the 2x2 matrix used in the χ^2-test. a, b, c and d stands for the total number of GG, AG, GA, and AA respectively for the outcome of editing for two different positions. Following the green arrow, we expect appositively coupled signal if the χ^2 value is above 10.22 (after the Bonferroni correction). Negatively coupled position would have an overrepresentation along the red arrow. B-D. The sequential context in which the edited position reside for the GluR-6 Y/C:I/V, 5-ht2c A:B:E:C.D and Adar2 -28 to +28 sites respectively. The purine annotation within square brackets, [ ], is to be read as A OR G at those positions.

Figure 3, A-D
The resulting dendrograms from the cluster analyses of Adar2 edited positions with the corresponding clustering of positions (classes) to the right. A – E15, B – E19, C – P2 and D – P21.

Figure 4, A-D
The resulting dendrograms from the cluster analyses of 5-ht2c edited positions with the corresponding clustering of positions (classes) to the right. A – E15, B – E19, C – P2 and D – P21.

Figure 5.
The graph show the result from the compilation of clustering of spacer distances of coupled positions in Adar2 (blue) and 5-ht2c (red). The x-axes show the discontinuous available spacer distances (1-55) and the left y-axes show the number of coupled position pairs having the corresponding distance. The peaks are put in at the spacer distances that tend to group: 1-4, 10-14, 23-26, 36-37 and 50-51. The height of one peak correspond to the total number of coupled positions in that group, right y-axes. For example, the peak centred on 24.2 says that the total number of coupled positions that have spacer distances between 23 and 26 nucleotides are 7.

Figure 6.
3D imaging of the A-form duplex of Adar2 edited region seen from the side (top) and front (bottom). All edited positions are marked except +28. Note that the duplex is idealized where no consideration has been made to the bona fide mismatches or bulges. Everything is base paired. Corresponding editing frequencies in parenthesis.

Table 1. Editing frequencies from stage E15 (embryo week 15), E19, P2 (post natal week 2) and P21. Blcap, Y/C and syn sites, could not be amplified for stages E19 and P2. The same holds for the Q/R and M/V sites of GluR-6 and stage E19. For Gabra3, we also included adult week 7, P7. 1) States upper and lower limits within a confidence interval of 95%. 2) Termed according to {Dawson_2004}, * except +28. 3) Termed according to {Niswender_1999}. GluR-B: Glutamate receptor b. Gabra3: gamma-aminobutyric acid A receptor. Adar2: Adenosine deaminase acting on RNA 2. 5-ht2c: Serotonin receptor 2c. Cyfip2: cytoplasmic FMR1 interacting protein 2. Flna: Filamin A. Kcna1: potassium voltage-gated channel, shaker-related subfamily, member 1. Blcap: Bladder cancer associated protein

E15				
Gene	*Site*	*%editing*	*Confidence* [1)]	*reads*
GluR-B	Q/R	95.2	± 9.1	21
	R/G	3.6	± 6.9	28
Gabra3	I/M	6.4	± 2.0	592
Adar2 [2)]	-28	n/d	-	-
	-27	< 2	-	-
	-4	< 2	-	-
	-2	< 2	-	-
	-1	3.9	± 1.5	567
	+10	5.1	± 1.8	567
	+23	3.1	± 1.4	540
	+24	19.6	± 3.3	540
	+28*	< 2	-	-
5-ht2c [3)]	A	19.3	± 6.0	166
	B	7.8	± 4.0	166
	E	3.6	± 2.8	166
	C	21.1	± 6.2	166
	D	47.9	± 7.6	165
GluR-C	R/G	14.6	± 8.9	58
GluR-5	Q/R	7.5	± 3.2	254
GluR-6	I/V	2.7	± 2.6	148
	Y/C	8.1	± 4.4	148
	Q/R	26.9	± 2.6	1094
	M/V	< 2	-	-
Blcap	Y/C	n/d	-	-
	Syn	n/d	-	-
Flna	Q/R	< 2	-	-
Kcna1	I/V	7	± 9.3	29
Cyfip2	K/E	4	± 1.1	1175

E19				
Gene	*Site*	*%editing*	*Confidence* [1)]	*reads*
GluR-B	Q/R	100	0	120
	R/G	36.5	± 4.2	499
Gabra3	I/M	33.6	± 3.1	901
Adar2 [2)]	-28	2.3	± 1.3	532
	-27	< 2	-	-
	-4	< 2	-	-
	-2	< 2	-	-
	-1	7.5	± 1.8	547
	+10	18.8	± 3.2	547
	+23	12.3	± 2.7	529
	+24	53.1	± 4.2	529
	+28*	2.8	± 1.3	535
5-ht2c [3)]	A	56.3	± 2.6	1359
	B	41.2	± 2.6	1359
	E	2.9	± 0.8	1358
	C	20.8	± 2.1	1358
	D	45.3	± 2.6	1342
GluR-C	R/G	47.7	± 3.1	973
GluR-5	Q/R	31.1	± 2.3	1547
GluR-6	I/V	26.2	± 2.0	1690
	Y/C	47.9	± 2.3	1690
	Q/R	n/d	-	-
	M/V	n/d	-	-
Blcap	Y/C	6.3	± 2.4	365
	Syn	n/d	-	-
Flna	Q/R	5.6	± 0.9	2218
Kcna1	I/V	4.9	± 1.4	876
Cyfip2	K/E	19.5	± 1.4	2738

Table 1 continued.

P2				
Gene	*Site*	*%editing*	*Confidence* [1)]	*reads*
GluR-B	Q/R	100	0	605
	R/G	52.1	± 8.9	121
Gabra3	I/M	53.7	± 3.1	971
Adar2 [2)]	-28	3.8	± 1.3	820
	-27	2.2	± 1.0	820
	-4	< 2	-	-
	-2	7.0	± 1.6	883
	-1	11.6	± 2.1	883
	+10	23.7	± 2.8	885
	+23	12.5	± 2.2	861
	+24	65.4	± 3.2	861
	+28*	2.6	± 1.1	875
5-ht2c [3)]	A	77.3	± 2.8	867
	B	64.4	± 3.2	867
	E	4.5	± 1.4	859
	C	23.5	± 2.8	859
	D	45,4	± 3.3	859
GluR-C	R/G	67.3	± 3.4	738
GluR-5	Q/R	34.8	± 8.1	132
GluR-6	I/V	55.9	± 4.0	592
	Y/C	63.9	± 3.9	590
	Q/R	83.5	± 3.9	358
	M/V	5.6	± 2.4	358
Blcap	Y/C	n/d	-	-
	Syn	n/d	-	
Flna	Q/R	6.8	± 1.7	821
Kcna1	I/V	6.3	± 1.8	702
Cyfip2	K/E	34.9	± 2.2	1776

P7				
Gene	*Site*	*%editing*	*Confidence* [1)]	*reads*
Gabra3	I/M	77.9	± 9.9	68

P21				
Gene	*Site*	*%editing*	*Confidence* [1)]	*reads*
GluR-B	Q/R	100	0	221
	R/G	72.1	± 4.6	358
Gabra3	I/M	92.5	± 2.0	638
Adar2 [2)]	-28	8.8	± 4.8	136
	-27	16.2	± 6.2	136
	-4	11.4	± 5.1	149
	-2	< 2	-	-
	-1	25.9	± 7.1	147
	+10	34.7	± 7.7	147
	+23	30.5	± 7.9	131
	+24	80.9	± 6.7	131
	+28*	10.4	± 4.8	144
5-ht2c [3)]	A	85.2	± 2.5	804
	B	74.6	± 3.0	804
	E	4.2	± 1.4	788
	C	25.6	± 3.0	788
	D	63.5	± 3.3	788
GluR-C	R/G	91.5	± 3.8	212
GluR-5	Q/R	62.5	± 3.7	661
GluR-6	I/V	73.8	± 5.6	190
	Y/C	80.5	± 6.2	191
	Q/R	81.1	± 3.5	477
	M/V	8.2	± 2.5	478
Blcap	Y/C	28.9	± 4.3	425
	Syn	17.2	± 3.6	425
Flna	Q/R	43.2	± 8.6	125
Kcna1	I/V	25.3	± 3.8	516
Cyfip2	K/E	75.0	± 3.0	800

Table 2, A-D
All regions and sites examined by the χ^2-test and cluster analyses to reveal coupled positions and the distances between them. If both the χ^2-test and the cluster analyses show coupling, we assign that as a strong coupled signal (++). If one show coupling we assign the coupling as weak (+). The first columns state the region (gene) and the corresponding sites that were evaluated for coupled properties. The χ^2 column show the calculated χ^2-value. After the Bonferroni correction, column values should be compared to >10.22 for the Adar2 sites and >7.88 for the 5-HT2c
A. The results from E15. * no evaluation for position -28 since there are no edited sites at all in E15. B. results from E19. C P2 and D for P21.

A. E15

gene	site	χ^2	cluster (class)	conclusion	distance (nt)
Adar2	-28 : -27	- *	-		
	-28 : -4	-	-		
	-28 : -2	-	-		
	-28 : -1	-	-		
	-28 : +10	-	-		
	-28 : +23	-	-		
	-28 : +24	-	-		
	-28 : +28	-	-		
	-27 : -4	1.811	-		
	-27 : -2	3.470	-		
	-27 : -1	0.050	-		
	-27 : +10	0.038	-		
	-27 : +23	0.098	-		
	-27 : +24	3.910	-		
	-27 : +28	3.474	-		
	-4 : -2	1.811	-		
	-4 : -1	0.001	-		
	-4 : +10	0	-		
	-4 : +23	15.27	2:3	+ coupled	26
	-4 : +24	0.324	-		
	-4 : +28	1.811	-		
	-2 : -1	22.77	1:3	+ coupled	1
	-2 : +10	20.79	1:3	+ coupled	11
	-2 : +23	0.098	-		
	-2 : +24	3.913	-		
	-2 : +28	3.474	-		
	-1 : +10	10.17	3:3	+ coupled	10
	-1 : +23	0.209	-		
	-1 : +24	13.319	3:3	++ coupled	24
	-1 : +28	0.050	-		
	+10 : +23	25.45	3:3	++ coupled	13
	+10 : +24	23.65	3:3	++ coupled	14
	+10 : +28	0.038	-		
	+23 : +24	1.088	-		
	+23 : +28	0.098	-		
	+24 : +28	3.913	-		
5-ht2c	AB	57.01	1:1	++ coupled	2
	AE	1.312	-		
	AC	12.28	1:3	+ coupled	7
	AD	0.001	-		
	BE	0.453	-		
	BC	5.321	-		
	BD	0.300	-		
	EC	1.104	-		
	ED	0.373	-		
	CD	9.981	3:3	++ coupled	5
GluR-6	I/V : Y/C	9.517	-	coupled	13

B. E19.

gene	site	χ^2	cluster (class)	conclusion	distance (nt)
Adar2	-28 : -27	0.033	-		
	-28 : -4	0	-		
	-28 : -2	3.591	-		
	-28 : -1	5.147	-		
	-28 : +10	4.362	-		
	-28 : +23	4.867	-		
	-28 : +24	10.54	1:4	+ coupled	51
	-28 : +28	0.137	-		
	-27 : -4	0.027	-		
	-27 : -2	7.347	2:2	+ coupled	25
	-27 : -1	4.445	-		
	-27 : +10	2.833	-		
	-27 : +23	0.041	-		
	-27 : +24	0.037	-		
	-27 : +28	0.033	-		
	-4 : -2	0.007	-		
	-4 : -1	1.903	-		
	-4 : +10	0.154	-		
	-4 : +23	0.684	-		
	-4 : +24	0.710	-		
	-4 : +28	0	-		
	-2 : -1	2.798	-		
	-2 : +10	1.374	-		
	-2 : +23	0.849	-		
	-2 : +24	7.806	-		
	-2 : +28	0.071	-		
	-1 : +10	3.004	4:4	+ coupled	10
	-1 : +23	5.975	4:4	+ coupled	23
	-1 : +24	8.110	4:4	+ coupled	24
	-1 : +28	0.017	-		
	+10 : +23	3.004	4:4	+ coupled	13
	+10 : +24	22.74	4:4	++ coupled	14
	+10 : +28	8.015	-		
	+23 : +24	11.81	4:4	++ coupled	1
	+23 : +28	0.159	-		
	+24 : +28	10.54	4:5	+ coupled	4
5-ht2c	AB	652.9	1:1	++ coupled	2
	AE	0.118	-		
	AC	41.11	1:3	+ coupled	7
	AD	7.823	1:1	+ coupled	12
	BE	0.978	-		
	BC	51.78	1:3	+ coupled	5
	BD	67.13	1:1	++ coupled	10
	EC	18.92	2:3	+ coupled	1
	ED	3.033	-		
	CD	34.87	3:1	+ coupled	5
GluR-6	I/V : Y/C	252.7		coupled	

C. P2.

gene	site	χ^2	cluster (class)	conclusion	distance (nt)
Adar2	-28 : -27	0.448	-		
	-28 : -4	3.408	-		
	-28 : -2	2.222	-		
	-28 : -1	1.156	-		
	-28 : +10	2.582	-		
	-28 : +23	20.03	1:4	+ coupled	50
	-28 : +24	0.391	-		
	-28 : +28	0.556	-		
	-27 : -4	0.007	-		
	-27 : -2	0.730	-		
	-27 : -1	1.018	-		
	-27 : +10	15.16	2:4	+ coupled	36
	-27 : +23	2.901	-		
	-27 : +24	3.807	-		
	-27 : +28	0.234	-		
	-4 : -2	16.42	3:4	+ coupled	2
	-4 : -1	18.77	3:4	+ coupled	3
	-4 : +10	12.86	3:4	+ coupled	13
	-4 : +23	2.925	-		
	-4 : +24	0.903	-		
	-4 : +28	6.092	-		
	-2 : -1	102.7	4:4	++ coupled	1
	-2 : +10	42.91	4:4	++ coupled	11
	-2 : +23	0.010	4:4	+ coupled	24
	-2 : +24	31.74	4:4	++ coupled	25
	-2 : +28	2.160	-		
	-1 : +10	6.388	4:4	+ coupled	10
	-1 : +23	3.980	-		
	-1 : +24	15.36	4:4	++ coupled	24
	-1 : +28	1.992	-		
	+10 : +23	24.01	4:4	++ coupled	13
	+10 : +24	53.55	4:4	++ coupled	14
	+10 : +28	7.058	-		
	+23 : +24	12.14	4:4	++ coupled	1
	+23 : +28	0.357	-		
	+24 : +28	10.75	4:3	+ coupled	4
5-ht2c	AB	415.4	1:1	++ coupled	2
	AE	18.28	1:2	neg-coupled	6
	AC	30.95	1:3	+ coupled	7
	AD	11.66	1:1	++ coupled	12
	BE	29.94	1:2	neg-coupled	4
	BC	28.07	1:3	+ coupled	5
	BD	41.88	1:1	++ coupled	10
	EC	2.287	-		
	ED	8.030	2:1	neg-coupled	6
	CD	0.463	-		
GluR-6	I/V : Y/C	98.12	-	coupled	13

D. P21.

gene	site	χ^2	cluster (class)	conclusion	distance (nt)
Adar2	-28 : -27	1.586	-		
	-28 : -4	0.991	-		
	-28 : -2	0.003	-		
	-28 : -1	0.717	-		
	-28 : +10	10.81	1:2	+ coupled	37
	-28 : +23	0.066	-		
	-28 : +24	0.357	-		
	-28 : +28	1.723	1:1	+ coupled	55
	-27 : -4	2.690	-		
	-27 : -2	0.049	-		
	-27 : -1	1.866	2:2	+ coupled	26
	-27 : +10	0.364	2:2	+ coupled	36
	-27 : +23	0.463	-		
	-27 : +24	5.498	2:2	+ coupled	50
	-27 : +28	0.591	-		
	-4 : -2	0.009	-		
	-4 : -1	0.652	-		
	-4 : +10	3.635	-		
	-4 : +23	35.54	3:3	++ coupled	26
	-4 : +24	3.584	-		
	-4 : +28	4.28	-		
	-2 : -1	0.152	-		
	-2 : +10	0.261	-		
	-2 : +23	0.204	-		
	-2 : +24	0.096	-		
	-2 : +28	0.005	-		
	-1 : +10	0.291	2:2	+ coupled	10
	-1 : +23	0.278	-		
	-1 : +24	0.013	2:2	+ coupled	24
	-1 : +28	0	-		
	+10 : +23	0.193	-		
	+10 : +24	0.425	2:2	+ coupled	14
	+10 : +28	0.587	-		
	+23 : +24	1.661	-		
	+23 : +28	2.476	-		
	+24 : +28	1.163	-		
5-ht2c	AB	375.7	1:1	++ coupled	2
	AE	1.989	-		
	AC	8.331	1:3	+ coupled	7
	AD	29.37	1:1	++ coupled	12
	BE	0.477	-		
	BC	3.715	-		
	BD	45.76	1:1	++ coupled	10
	EC	5.761	-		
	ED	3.148	-		
	CD	1.541	-		
GluR-6	I/V : Y/C	132.2	-	coupled	13

Figure 1, A-K

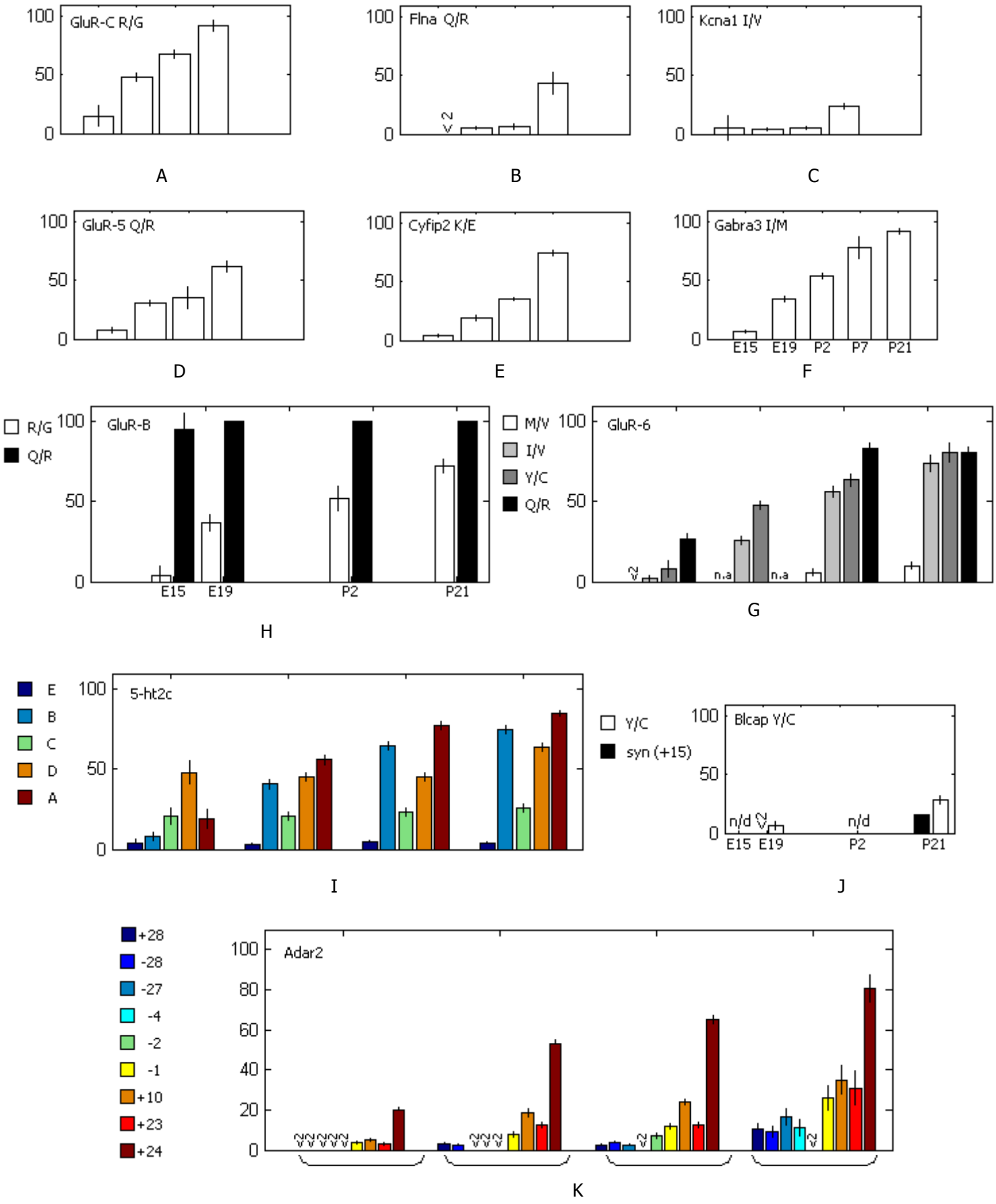
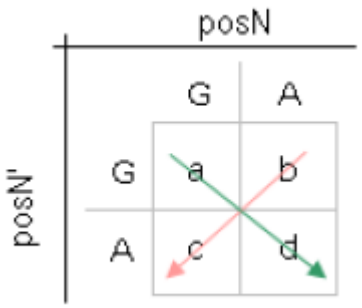

GluR-C R/G
100
50
0
A
Flna Q/R
Q
R
B
Kcna1 I/V
C
GluR-5 Q/R
D
Cyfip2 K/E
E
Gabra3 I/M
E15
E19
P2
P7
P21
F
R/G
Q/R
GluR-B
H
M/V
I/V
Y/C
Q/R
GluR-6
n.a
G
E
B
C
D
A
5-ht2c
I
Y/C
syn (+15)
Blcap Y/C
n/d
J
+28
-28
-27
-4
-2
-1
+10
+23
+24
Adar2
K

Figure 2, A-D

A



B

```
        I/V                 Y/C
TGTAT[AG]TTCTGCTGGCTT[AG]CT
```

C

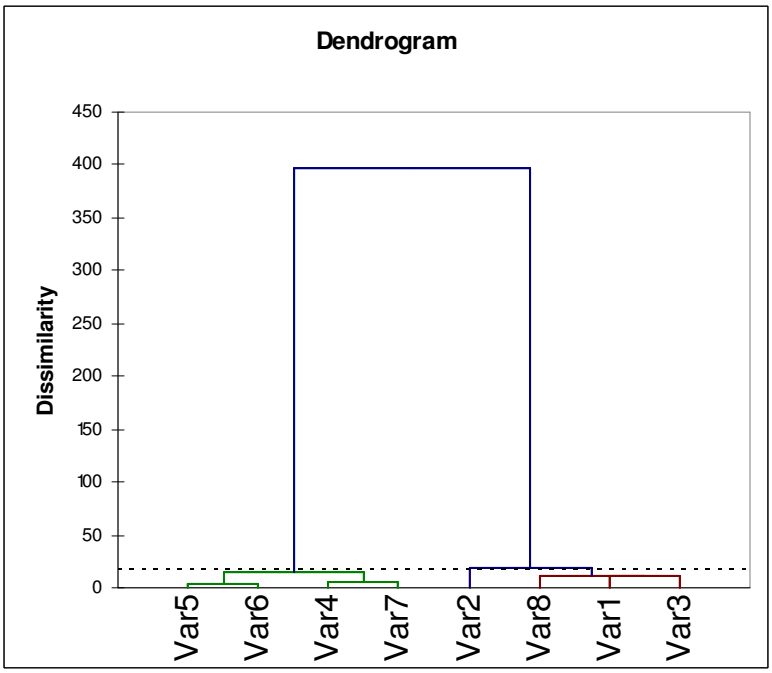
```
  A    B       E   C        D
[AG]T[AG]CGT[AG][AG]TCCT[AG]TTGAGCATAGCCGGT
```

D

```
 -28 -27             -4   -2  -1           +10             +23 +24    +28
[AG][AG]AGGAA..TATTT[AG]C[AG][AG]GATCCTGCA[AG]CGAAG..TTGT[AG][AG]GTT[AG]
```

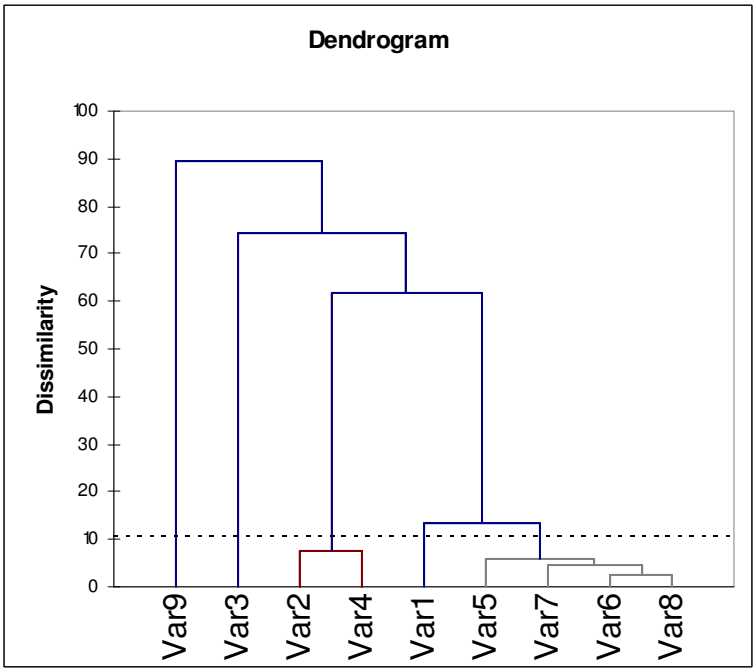
Figure 3, A-D

A. E15



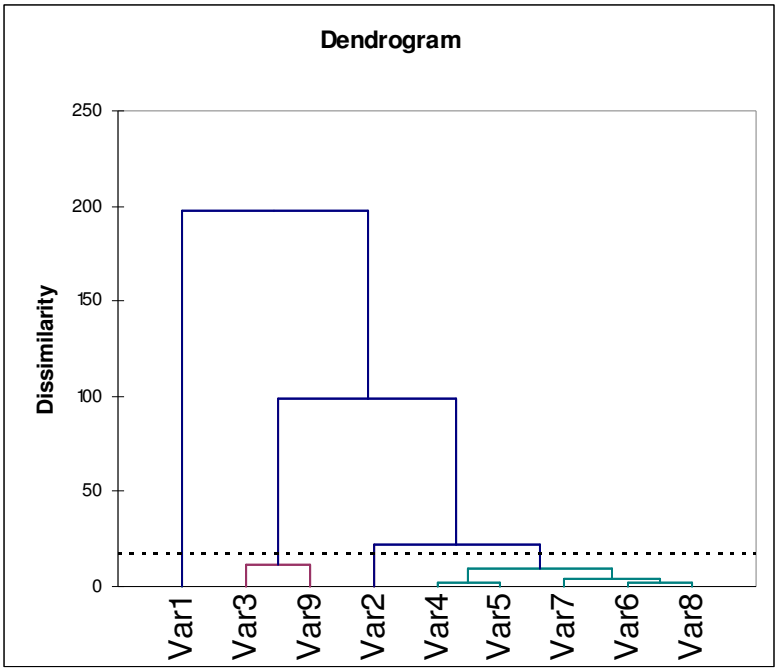
site	class
Var1 / -27	1
Var2 / -4	2
Var3 / -2	1
Var4 / -1	3
Var5 / +10	3
Var6 / +23	3
Var7 / +24	3
Var8 / +28	1

B. E19



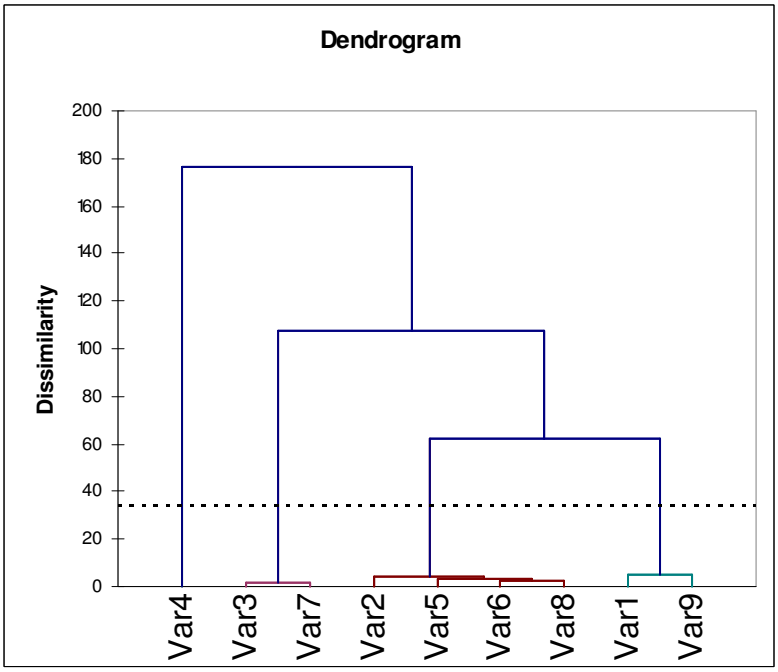
site	class
Var1 / -28	1
Var2 / -27	2
Var3 / -4	3
Var4 / -2	2
Var5 / -1	4
Var6 / +10	4
Var7 / +23	4
Var8 / +24	4
Var9 / +28	5

C. P2



site	class
Var1 / -28	1
Var2 / -27	2
Var3 / -4	3
Var4 / -2	4
Var5 / -1	4
Var6 / +10	4
Var7 / +23	4
Var8 / +24	4
Var9 / +28	3

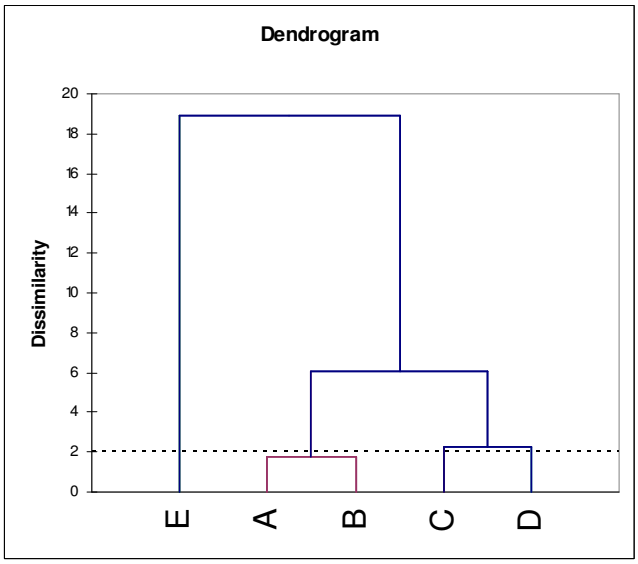
D. P21



site	class
Var1 / -28	1
Var2 / -27	2
Var3 / -4	3
Var4 / -2	4
Var5 / -1	2
Var6 / +10	2
Var7 / +23	3
Var8 / +24	2
Var9 / +28	1

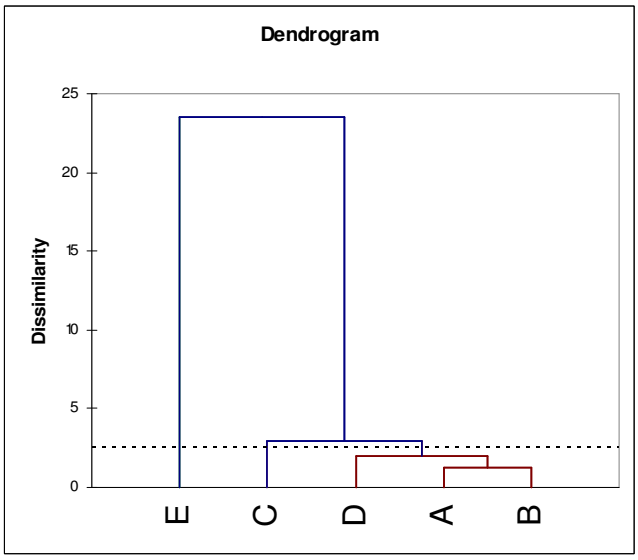
Figure 4, A-D

A. E15



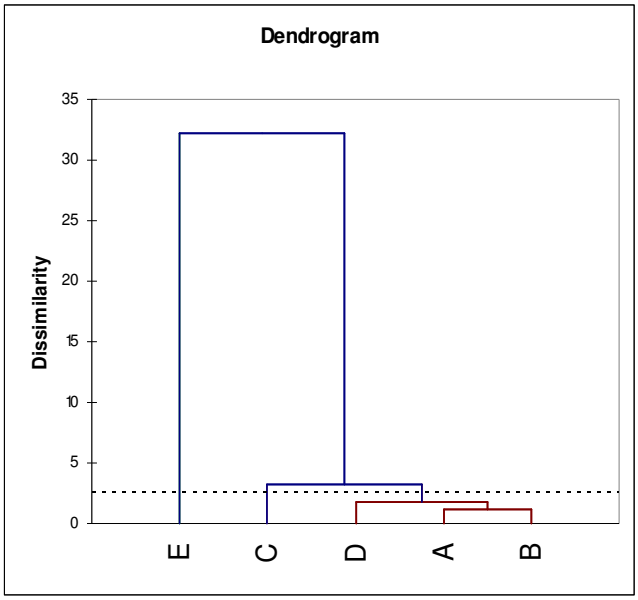
site	class
A	1
B	1
E	2
C	3
D	3

B. E19



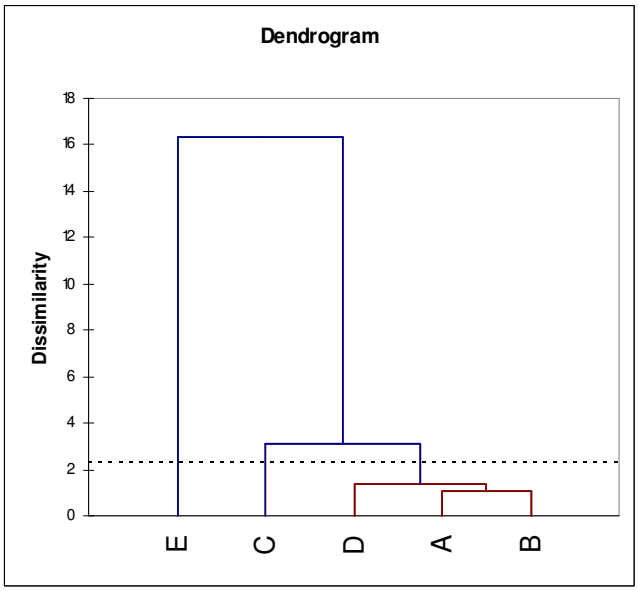
site	class
A	1
B	1
E	2
C	3
D	1

C. P2



site	class
A	1
B	1
E	2
C	3
D	1

D. P21



site	class
A	1
B	1
E	2
C	3
D	1

Figure 5.

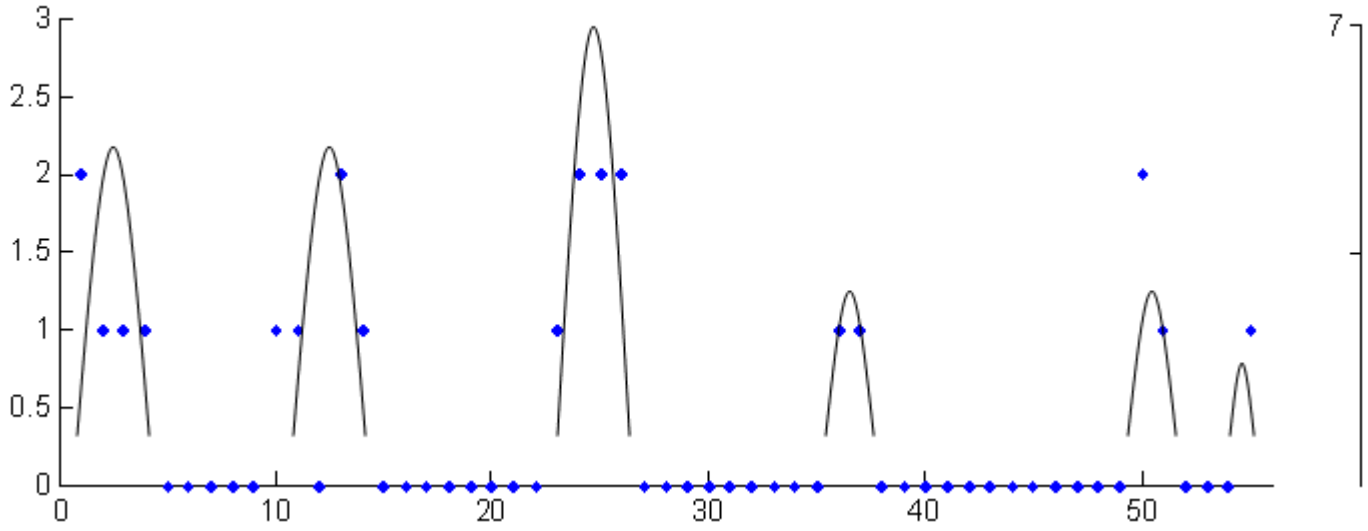


Figure 6.