DOCTOR OF PHILOSOPHY

A novel miRNA cluster within the circadian clock gene NPAS2 and the implications of rs1811399, an autism enriched single nucleotide polymorphism

Jones, Dylan

Award date:
2014

Awarding institution:
Bangor University

A novel miRNA cluster within the circadian clock gene *NPAS2* and the implications of rs1811399, an autism enriched single nucleotide polymorphism.

Thesis submitted in accordance with the requirements of Bangor University for the degree of Doctor in Philosophy

Dylan Wyn Jones

Bangor University, UK

School of Biological Sciences

December 2014

ACKNOWLEDGMENTS

I would like to extend my deepest thanks to Dr Thomas Caspari for taking on this difficult project and entrusting me to deliver the result. I am grateful for all his work and insight into the scientific process. I would also care to extend my gratitude to Dr Natalia Harrison who trained me during the first few months in the intricacies of tissue culture. I would also like to commend Dr Brad Nicholas and Dr Dawn Wimpory for their invaluable insights and assistance throughout the project. Without their prior work in the field I would not have had these three years. Thanks also go to the Knowledge Economy Skills Scholarship (KESS) and Autism Cymru for funding this research project.

****

Buaswn yn hoffi diolch, a cyflwyno, yr gwaith yma I Bethan Davies-Jones fy annwyl wraig. Dros yr tair blynedd ddweutha rwyt wedi bod yn gefn i mi ac yn ysbridoliaeth. Nid wyf yn deud gormod wrth ddweud na hebdda ti ni buaswn wedi cyraedd yr lle rwyf wedi.

I fy nheulu am yr holl cefnogaeth drost y blynyddoedd. I Mam a Dad am fy magu ac I Endaf, fy mhrawd. Hefyd mewn côf o Nain a Taid Rhos-y-bôl a Taid Tŷ Croes. Effallai nid ydych yma i weld diwedd fy nhaith drwyr brif ysgol ond mae'r cofion melus yn ein cadw ni yn fynd.

****

Last but not least, to Hefin and Esther. By you my interest in science was noticed and nurtured, you encouraged me when others sought to dissuade. Who knew that nights looking at the sky and hungry trips to the zoo could lead here? To Siân; may you find the same sense of wonder in life that I did, and never lose it.

SUMMARY

A novel miRNA cluster within the circadian clock gene NPAS2 and the implications of rs1811399, an autism enriched single nucleotide polymorphism.

This PhD investigated the role of the rs181399 single nucleotide polymorphism (A or C) in the maturation of a previously unknown microRNA. microRNA are potent regulators of gene expression and their expression is finely controlled. Canonically the microRNA processing machinery recognise hairpin-loops of single stranded RNA as a substrate for processing. Mutations, such as rs1811399 can disturb the hairpin leading to reduction in expression.

The work presented in this thesis demonstrates that a novel microRNA cluster is located within intron 1 of NPAS2 which is independently transcribed of its host gene. Secondly, plasmid constructs were used to integrate versions of the novel microRNA hairpin containing either an A allele or a C into cell lines. Utilising this method the impact of the C allele was noted to be deleterious to microRNA maturation.

RNA protection assays have demonstrated that both precursor microRNA and mature microRNA is constitutively expressed within a multitude of tissue types, seemingly independently of its host gene.

The potential impact of the C allele on genetic regulation was analysed bioinformatically by analysing potential gene targets and the pathways they participate in.

Table of Contents

List of Figures

List of Tables

1 Introduction: A novel miRNA cluster within the circadian clock gene *NPAS2* and the implications of rs1811399, an autism enriched single nucleotide polymorphism.

This body of work has been commissioned for the sole purpose of elucidating the impact of the rs1811399 SNP. We will hope to establish whether or not it interfere with miRNA biogenesis. On the way to answering this question it will also be possible to answer some tangential questions:

- Does *NPAS2* host a miRNA gene?
- Is *NPAS2* the host of a miRNA cluster?
- Is the expression of any miRNA dependant on the expression of the host gene?
- What is the expression profile of the miRNA?

The answers to these questions will illuminate a relatively poorly understood field as very little has been published on the influence of SNP on a miRNA. Hopefully we will also be able to contribute to the debate of the host gene-miRNA relationship as this too is a field of much controversy.

1.1The Circadian Clock

1.1.1 Core Circadian Clock

The circadian rhythm machinery has evolved to ensure the synchronicity of life with regards to the Earth's 24h cycle. There are several theories why a circadian clock has evolved using light as the main "zeitgeber". One theory is based on the idea that early single cell organisms needed to avoid the UV in sun light which would have damaged their genetic information

(Ouyang *et al*, 1998). Another theory is linked with the uptake of mitochondria into early cells which would require coordination between DNA synthesis and mitochondrial activity. Mitochondria are the main source of free oxygen radicals which damage DNA, and the clock may allow cells to restrict DNA synthesis to phases of low oxygen production e.g. during the night. This theory is supported by experiments from Chen and McKnight (2007) in which they show that a more simpler clock, the metabolic clock of *Saccharomyces cerevisae,* limits DNA replication to periods of low oxidative stress (as noted by Tu *et al*, 2005). A hypothesis that mammalian cells are under a similar control is supported by the work of Unsal-Kacmaz and colleagues (2005) who note that the TIMELESS protein tightly regulates a cell's progression in the cell cycle.

According to Oster (2006) the core clock in mammals is located within the suprachiasmatic nucleus (SCN), a region of the hypothalamus, with the largest stimuli, or zeitgeber, being the light and dark cycle which acts upon the SCN via glutamate and pituitary adenylate cyclise-activating peptide. Its activity is regulated by the cyclical nature of genes expressed and repressed within the cells of the SCN (see Figure 1.1).

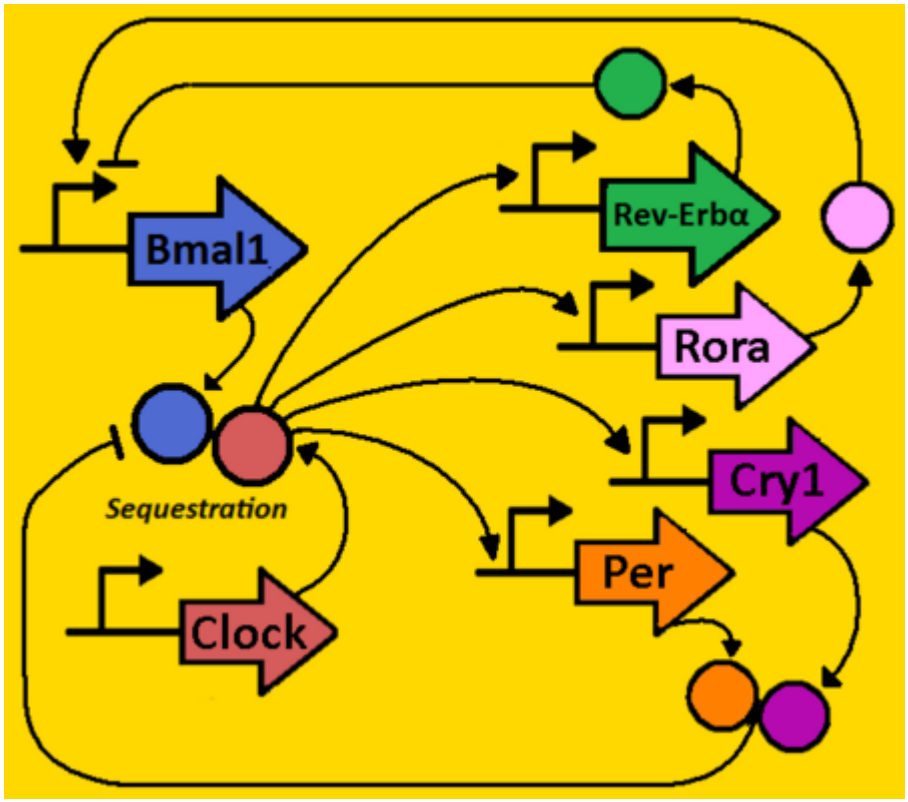


Figure 1.1: A schematic overview of the mammalian circadian clock. A schematic of the core circadian clock in humans reproduced from (http://2008.igem.org/Team:Michigan/Project Accessed 17/11/13)

The negative feedback cycle of the circadian clock consists of several steps: Step 1) The BMAL/CLOCK heterodimer facilitate the expression of the *PER* and *CRY* genes thus allowing for levels of PER and CRY proteins to build up. Step 2) PER and CRY form a heterodimer to prevent their phosphorylation by CK1ε and ubiquitinylation. Step 3) In conjunction with CK1ε the PER-CRY heterodimer translocates into the nucleus and interferes with CLOCK-BMAL mediated transcription of clock-controlled genes such as *Vasopressin*. Oster (2006) then states that this leg of the cycle occurs during the night. The positive feedback cycle consists of the following steps: Step 1) BMAL is translated and forms a heterodimer with CLOCK. Step 2) The BMAL-CLOCK heterodimer binds to the E-box upstream regulatory domain of the *PER, CRY, RORα, REV-ERBα* and other clock-controlled

genes such as *vasopressin* and *prokineticin2* which go on to control/establish the circadian rhythm in other regions of the brain or tissues. The PER and CRY proteins are, as established, the negative regulators of the clock whilst RORα and REV-ERBα act as a stabilising force upon transcription of the *BMAL* gene (according to Jetten, Kurebayashi & Ueda (2001) whilst the gradual decay of PER-CRY over time is enough to restart the cycle. The presence of light will increase levels of compounds such as melatonin which binds to and activates RORα which in turn increases expression of the *BMAL* gene).

1.1.2 Peripheral clock

Circa 10% of all the genes expressed within a majority of tissues are expressed in a circadian pattern (Akhtar *et al*, 2002), with a paper by Storch *et al* (2002) going further to identify these genes as coding for proteins involved in rate limiting steps in many important metabolic pathways such as glycolysis and fatty acid metabolism. As these are not a part of the core clock they are referred to as a peripheral clock (Lamia *et al*, 2008).

An example of a peripheral clock gene would be the D-element binding protein which is upregulated in the presence of BMAL and activates the transcription of hepatic enzymes such as steroid 15-hydroxylase and cournarin-7-hydroxylase (Lavery & Schibler, 1993).

Further evidence for the importance of a peripheral clock can be seen in the vasculature of mammals such as mice in which much work has been done on the circadian control of aortic tissue gene expression. A paper published by Rudic *et al* (2005), for example, identify 307 genes in mouse aorta that exhibit circadian oscillations. A separate review article (Reily, Westgate & FitzGerlad, 2007) identifies these genes as being important for protein folding, metabolism and protein breakdown.

1.1.3 *NPAS2*

The *NPAS2* gene covers a large genomic locus of 176.68kb and encodes for a final transcript of 4007bp and 824 amino acid residues (Ensembl).

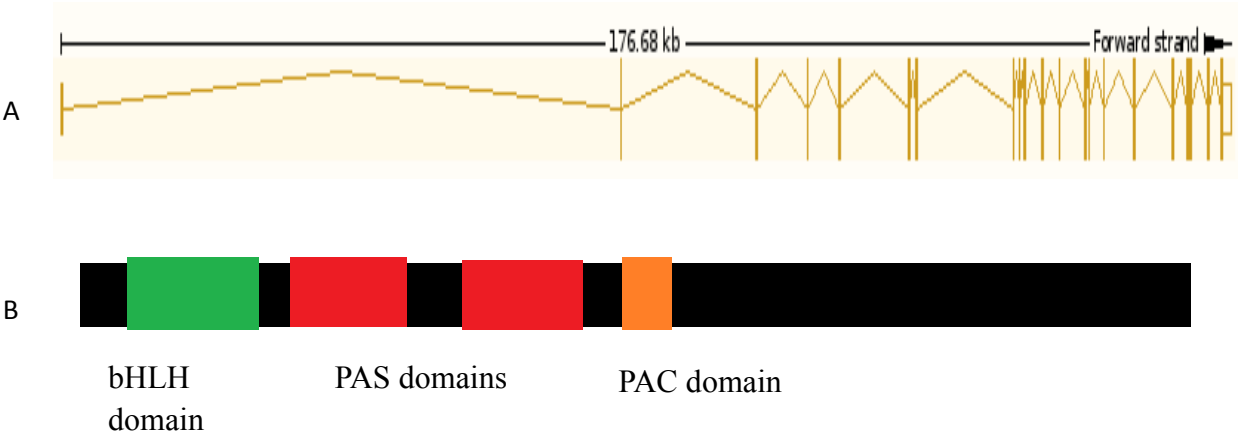


Figure 1.2: a) Organisation of the *NPAS2* gene. Exons are visible as vertical lines, introns as horizontal lines and the read direction is highlighted with the black arrow. This image represents the *NPAS2* transcript associated with the circadian clock. There are 12 other alternative transcripts which may or may not be biologically functional. Image downloaded from Ensembl (Accessed 17/11/13).

b) Domain organisation of translated NPAS2. The green square represents the bHLH domain, red the two PAS domains and orange the PAC domain.

The genes encodes for the NPAS2 protein or neural-PAS domain protein 2 (PAS being the acronym for Period, aryl-hydrocarbon and singleminded) and is a basic helix-loop helix-PAS (bHLH) transcription factor that is preferentially expressed within the mammalian forebrain and is known to have a role in memory development and the circadian clock (Franken *et al*, 2006; Gilles-Gonzalez & Gonzalez, 2004).

According to Gonzalez & Gonzalez (2004) the PAS domain of the protein consists of two slightly conserved domains which only exhibit 12% homology with other members of the PAS family. It is however noted in the literature that all these domains are around 130 amino acid residues long and that proteins can include as many as six PAS domains. Whilst there is great variation in the structure of the PAS domains, their functions are mostly similar in that they respond to environmental stimuli and regulate gene expression accordingly (Rutter et *al*, 2001); however it should be pointed out that not all PAS domain proteins are transcription factors with Dunham *et al* (2003) noting that a PAS protein in *Bradyrhizobium japonicum* is responsible for oxygen sensing.

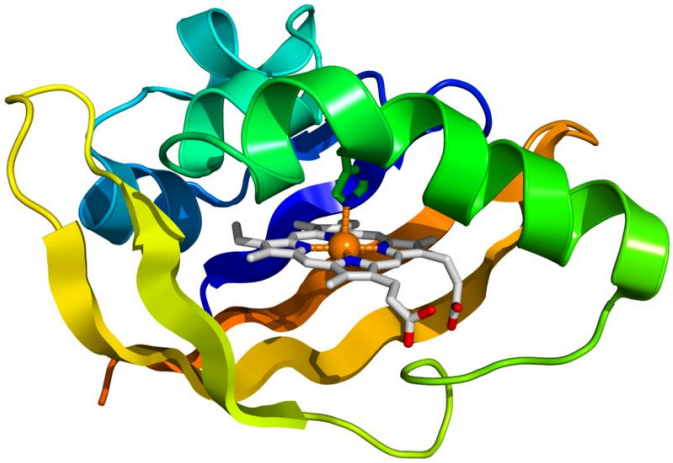


Figure 1.3: Crystallographic structure of the PAS domain of the bacterial oxygen sensor protein FixL. This image reproduced from the Protein Data Bank which demonstrates the two alpha helices integral to the function of the proteins.

Rutter *et al*, (2001) describes that structurally the bHLH motif contains two alpha-helices joined together by a "loop" which is described as the bonding of two carbon atoms not currently engaged in an alpha helix or beta pleated sheet (refer to Figure 1.3). bHLH proteins are a group of proteins conserved across nature with the literature noting over 240 examples of bHLH proteins across many species including *Drosophila*, *Homo sapiens* and *Xenopus*.

They are known to play important roles in neural development and uptake of phosphates (Gonzalez & Gonzalez, 2004). Cavadini *et al* (2007) note that bHLH-motif containing proteins are required to form a heterodimer before they are able to bind to DNA at a site known as the E-box (a pallendromic sequence of CACGTG). The heterodimer partner for NPAS2 is BMAL1 according to Rutter *et al* (2001).

To complement the bHLH motif in DNA binding the NPAS2 protein also employs a heme group as co-factor (Dioum *et al*, 2002). The two heme groups in NPAS2 are postulated to play an important role in stabilising the bHLH motif and linking its ability to bind to DNA with the redox status of the cell. NPAS2 is able to assess the redox state of the cell as heme is able to bind carbon monoxide. If the redox state of the cell is strongly reducing then NPAS2 is able to bind to DNA (Dioum *et al*, 2002)

1.1.4 Roles of NPAS2

The role of NPAS2 in the circadian rhythm is well established. The protein is involved in the activation of transcription of *PER1*, *PER2*, *CRY1* and *CRY2* genes. These proteins then act as a negative regulator of the *NPAS2* gene, preventing further expression of the gene leading to eventual depletion of the cellular pool of NPAS2 protein via proteosomal degredation (Garfield & Schibler, 2007). This auto-regulatory cycle of NPAS2 expression is linked with sleep homeostasis as demonstrated in experiments in *NPAS2*-/- mice (Franken *et al*, 2006).

The role of *NPAS2* in memory was highlighted in an investigation carried out involving mice which had had their *NPAS2* gene altered to remove the bHLH domain coding exon and instead produced an otherwise structurally faithful protein fused to the reporter protein LacZ (Garcia *et al*, 2000). The paper continues by noting the localisation of NPAS2 within the brain of the mice via use of beta-galactoside (LacZ) reporter. They noticed that the protein was located in areas associated with memory (i.e the mesolimbic pathway) and they also

noted that when submitted to a battery of memory tests the *NPAS2*-/- mice demonstrated no impaired short term memory but a deficient long term memory in contextual situations e.g. fear.

1.2 Autism: A Brief introduction.

Autism has a catalogued history dating to its naming by Dr Leo Kanner. Dr Kanner used the term infantile autism to describe eleven children who exhibited symptoms including: lack of social skills and a refusal to break from routine (Kanner 1968). However according to Kuhn & Cahn (2004) the term autism had been previously used to describe a collection of symptoms within schizophrenic patients by Dr Eugen Bleuler in 1910 who was a Swiss psychiatrist from whom Dr Hans Asperger borrowed the word in 1938 to describe the condition which later came to be named Asperger's syndrome which exhibited symptoms similar to Dr Bleuler's patients.

Since the 1970s it was apparent that autism and Asperger's were not isolated syndromes but indeed part of a wider spectrum of conditions all exhibiting similar symptoms with the Diagnostic and Statistical Manual of Mental Disorders (DSM)-4 listing three pervasive developmental disorders which includes autism, Asperger's and PDD-NOS (pervasive developmental disorder not otherwise specified) (Johnson & Myers, 2007).

It is estimated by Johnson & Myers (2007) that the prevalence of Autism Spectrum Disorders (ASD) in the European/American population is approximately 6 per 1000. We may also note that an important epidemiological fact of autism is its strong male disposition with a male to female ratio ranging from 2:1 to 6.5:1 (Johnson & Myers, 2007), with the usually quoted figure being 4:1. However when one focuses on severe autism the ratio leaps to 15:1 (Johnson & Myers, 2007).

Unfortunately for both patients and clinicians, autism's aetiology is not simple with many contributing factors (Muhle, Trentacoste & Rapin, 2004).

A review by Muhle, Trentacoste & Rapin (2004) surmises over 30 years of twin studies and notices that concordance rates between monozygotic twins and dizygotic twins for severe autism to be 60% and 0% respectively. The rates for ASD however increase to 92% and 10% respectively. These statistics demonstrate the roles for genetic factors in autism but not a single gene. Epigenetic and environmental factors can contribute to the condition. Whilst the idea of a genetic cause of autism can raise hopes of a therapy for the disorder or potential screening, the vast spectrum of the disorder results in dozens if not hundreds of candidate genes or loci being identified as significant (Muhle, Trentacoste & Rapin, 2004).

Recent efforts have focused on the role regulatory networks (Figure 1.4) have on leading to the phenotype. For example Wimpory *et al* (2002) and Hu *et al* (2010) who, respectively, questioned the roles of the circadian clock and miRNA regulatory networks in autism.

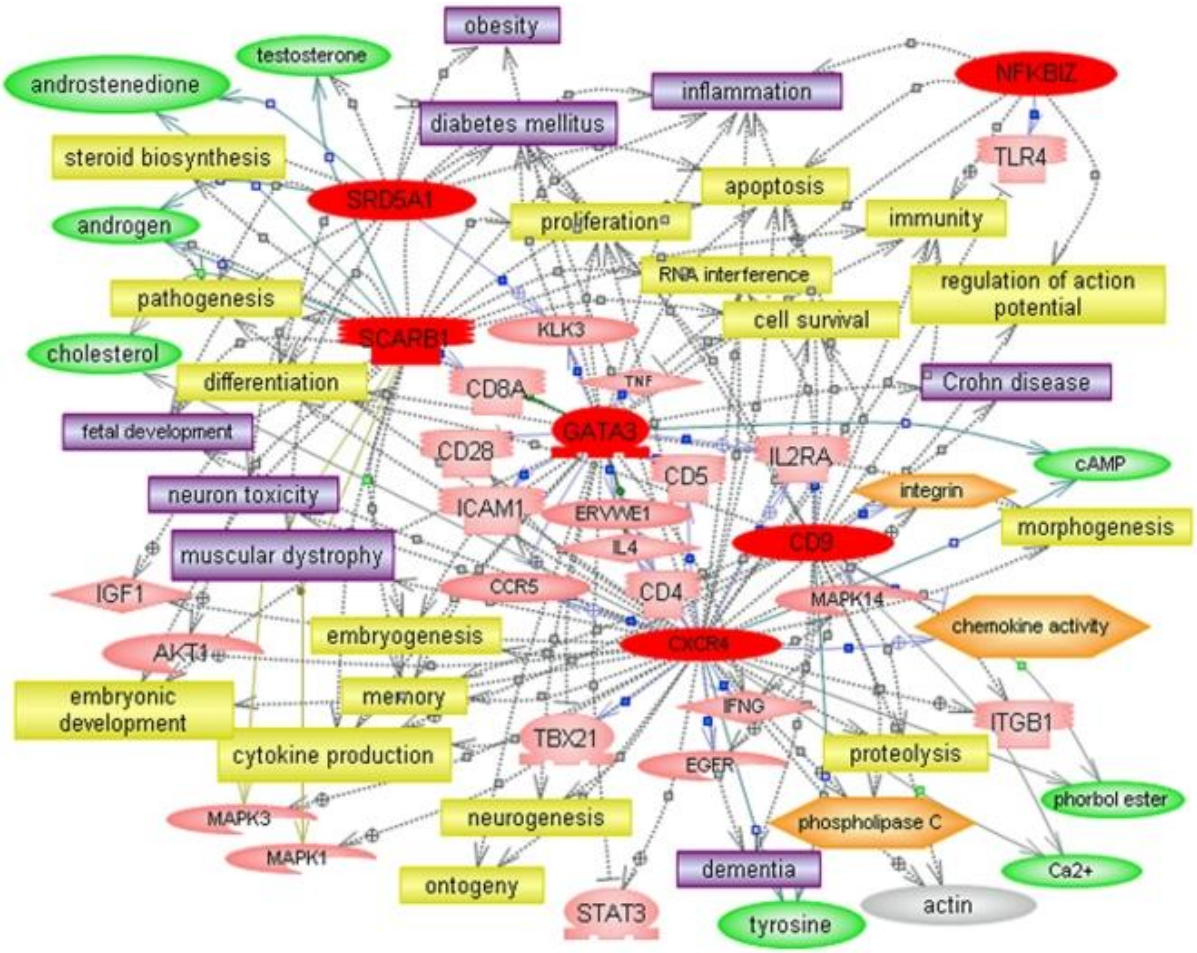


Figure 1.4: A network of related gene pathways with altered expression levels in autism. Genes in this figure have had their expression levels in autistic patients quantified using qPCR. Yellow denotes cellular processes, pink are genes identified as being part of the pathways, red are up-regulated genes, purple are overt disorders, green are small molecules. This diagram demonstrates the complicated nature of autism and the varying aetiologies which can give rise to the phenotype. It should be noted that incidence rates of all the conditions noted above within the autistic population is higher (Hu *et al*, 2009).

1.2.1 Timing in Autism

Among many of the named aetiological factors attributed to developing autism, the role of the circadian clock is perhaps the least understood. It has been hypothesised that one of the

facets of autism is the patient's inability to synchronise their timing during conversations (Wimpory, Nicholas & Nash, 2002). This timing hypothesis is further strengthened when the importance of timing in healthy babies becomes apparent in their interactions between parent and child during, for example, pre-verbal games during which turn taking is established. These rhythms which begin at 1/900ms rapidly quicken to 1/500ms by the age of two months. Studies that adversely affected these rhythms (such as utilising video-links) caused distress to the young children and led to transient-autistic like behaviour (eye avoidance, inability to take part in events etc) which soon vanished under resumption of live socializing (Kubicek, 1980). The extent of improper timing within autism is further demonstrated by an apparent inability to rapidly shift attention between various objects, admittedly this phenomena has also been linked to improper neuronal plasticity with glutamate being essential for correct plasticity but its absorption and metabolism do not fall under circadian control (Paul *et al,* 2008*).*

An experiment carried out by Konopka & Benzer (1971) utilising mutagenesis of a *Drosophila melanogaster per* gene during which three mutants were formed: short period, long period or arrhythmic period flies. These mutations not only caused discrepancies within the flies' circadian rhythms but also affected the fruit fly's courtship song. Thus the first link between social communications and the circadian clock were drawn.

Many of the findings in autism correlate with a potential circadian cause when viewed in this context. Hu *et al* (2009) conducted a survey on lymphoblastic cell lines extracted from 116 autistic patients who had been scored using the Autism Diagnostic Interview-Revised test (ADI-R) to identify such a link. Using the questionnaire she was able to segregate the populations into three categories: severe, mild and savant. From this study came an interesting result, a set of 15 genes (down-regulated: *AANAT, BHLBH2, CRY1, NPAS2, PER3, RORA, NFIL3, CLOCK, PRGDS, PER1, CREM, BHLHB*3 and *DPYD.* Up-regulated:

CRY1 and *NR1D1*) that regulate the circadian rhythm were found to be strongly linked (a false discovery rate (FDR) of <5%) with severe autism.

Prior to this influential work being published, it had been noted by Nicholas *et al* (2008) that the circadian clock genes may be implicated in autism and had identified several mutations at single nucleotides within genes (single nucleotide polymorphisms or SNPs). However none of these SNPs were to be found in the exonic regions and were instead to be found in the intronic sequences of the genes. Such examples include rs1861972, rs1861973, rs1811399 and rs885747 in *Engrailed 2* (*EN2*) a gene which encodes for a transcription factor important in neural development, *NPAS2* and *PER1* (Nicholas *et al*, 2008), respectively. These four SNPs (rs1861972 and rs1861973 in *EN2*) were the subject of an *in silico* survey in an attempt to identify the significance of these mutations in causing the pathophysiology of autism (Nicholas *et al*, 2008).

Rs1811399 has been demonstrated to be significantly associated ($P<0.05$) with autism however the exact mechanism of its interaction was unclear as the SNP does not interfere with protein coding due to its location within intron 1 of the *NPAS2* gene (Nicholas *et al* 2007).

The SNP itself maps to co-ordinates 2:101479014 (Ensembl release 71) henceforth referred to as Ensembl) and is 42137 base pairs from the initiation of the intron and 42180 base pairs from the beginning of exon 2 and is noted with the ambiguity code 'M' denoting the allele can either be C or A (Figure 1.5). The C allele is recorded by Ensembl as being the ancestral allele. The ancestral allele is defined as an allele which is shared between human and chimpanzee (Sepncer *et al*, 2006).

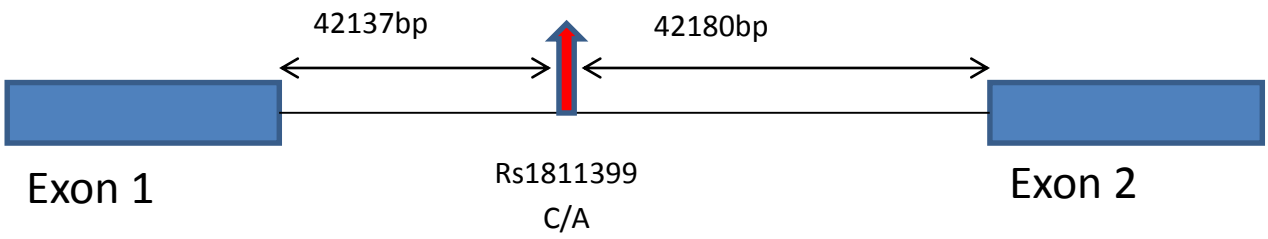


Figure 1.5 Location of rs1811399 in *NPAS2*.

1.3 Regulation of gene expression.

Not every cell requires every protein all the time or indeed at all (Filipowicz *et al*, 2005). Several regulatory mechanisms have evolved to tightly control protein expression (Filipowicz *et al*, 2005). These can be divided into two categories: post-transcriptional and post-translational (Filipowicz *et al*, 2005, Xu *et al*, 2012). Whilst post-translational regulation is beyond the scope of this thesis it is sufficient to note that it involves a variety of chemical modifications of proteins to alter their activity or mark them for degradation (Xu *et al*, 2012). Post-transcriptional regulation involves processes that occur to the mRNA of a gene and will be the focus of this thesis.

RNA interference is a process by which gene expression is regulated in a post-transcriptional mechanism and was experimentally demonstrated in 1998 by Fire *et al* who showed that injecting double stranded RNA (dsRNA) molecules into the nematode *Caenorhabditis elegans* could induce gene silencing. Only dsRNA induced this silencing and anti-sense RNA was not sufficient. Prior to this publication several laboratories reported phenomena that they could not explain: for example Romano & Macino (1992) observed in *Neurospora* gene silencing when a homologous sequence of DNA was inserted into the fungus. They named the process quelling. A similar result was recorded two years previously by Napoli, Lemieux and Jorgensen who introduced a chimeric chalcone synthase gene into petunia flowers in an attempt to gain a darker coloured petal; instead the investigators noted the petals of the

transfected plants were white which implied that both the integrated gene and the endogenous gene were being silenced by an unknown mechanism they termed co-suppression. Melo and Fire shared the Nobel prize in 2006 for their achievement.

There are two pathways by which gene silencing can occur via RNA molecules: small interfering RNA (siRNA) and micro RNA (miRNA). siRNA is modulated by the presence of endogenous (or exogenous) dsRNA identical to that described by Fire *et al* (1998) whilst miRNA are dependent on segments of RNA which contain secondary structural hairpins which then undergo post-transcriptional modification into small ~20 base pairs (bp) long RNA before they can induce silencing. Both pathways utilise the same end point protein complex, known as RISC, to induce silencing (Engels & Hutvagner 2006).

1.3.1 Biogenesis of microRNA

miRNA genes are widely dispersed throughout the human genome and can be found between genes (intergenic) or within protein coding genes (intragenic) (Altuvia *et al*, 2005). Of the intragenic miRNA genes approximately 40% of them are located within an intron with the remainder located within exons often on the opposite strand to the coding region (Rodriguez *et al*, 2004). It was assumed that microRNA were located within protein coding genes to utilise the host's promoter (Baskerville & Bartel, 2005) later however it was discovered that some 35% of intronic miRNA possess their own upstream promoter regions and are therefore potentially capable of their own transcription (Rodriguez *et al*, 2004).

Zhou *et al* (2007) notes that most miRNA are transcribed by RNA Polymerase II and have proximal to their 5' end a region where PolII and its accessory transcription factors can bind. This is accessorized in some cases with CpG islands (27%), PolIII promoters (1%) or CT repeats (100%) which all facilitate transcription.

miRNA that are co-expressed with their host gene often have a co-suppression role (e.g. miR-26b) (Zhu *et al*, 2012). miR-26b is hosted within the *CTDSP2* gene (which encodes for a protein which negatively regulates RNA polymerase II) and post-transcriptionally represses the expression of the gene (Zhu *et al*, 2012). A second example would be miR-126 which is hosted within and down regulates the product of the *EGFL7* gene, which is required for vascular differentiation (Musiyenko, Bitko & Barik 2008). Conversely it has also been demonstrated that hosted miRNA can *improve* the expression and function of their host genes by down regulating antagonistic elements of various pathways e.g. miR-338 which is located with the *AATK* gene (induces apoptosis in myeloid cells) and down regulates an AATK antagonist; NOVA (Barik 2008).

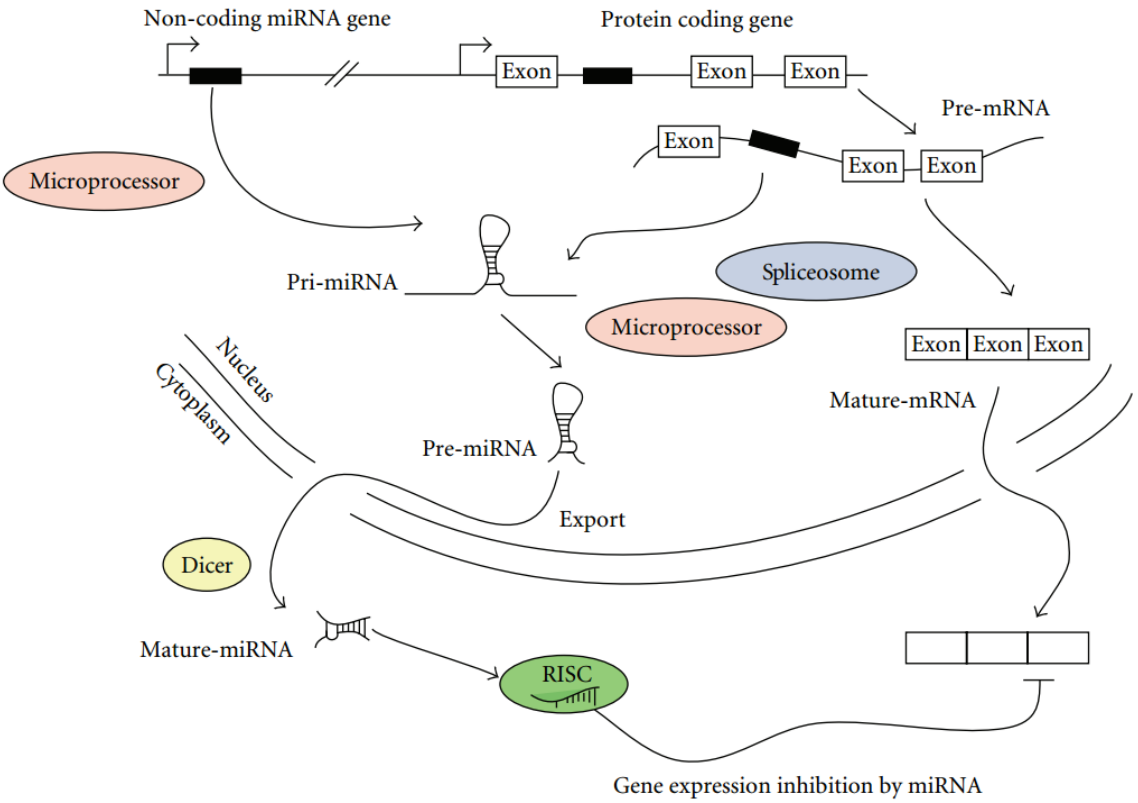


Figure 1.6: miRNA biogenesis: Image reproduced from Shomron & Levy (2009). From this schematic we can see that there are two pathways for production of miRNAs; one

consists of a gene specific for the miRNA whilst the second involves the extraction of the pre-mRNA from intronic or otherwise non-coding sections of DNA.

RNA polymerase II transcribes the microRNA gene, which at this initial stage can be several thousand kilobases long (Gregory, Chendrimada & Shiekhattar 2005). Initially it was assumed that it was RNA polymerase III which was responsible for the transcription of the interfering RNA due to its role in transcribing other non-coding RNA such as tRNA, however once a large amount of microRNA sequences became available it was apparent that several contained poly-T sequences which act as RNAPol III termination sequences (Lee *et al*, 2002). Once the transcription of the primary miRNA transcript (pri-miRNA) is completed it is capped as in a similar method to protein coding genes (Cai, Hagedorn, & Cullen 2004). A characteristic of these pri-miRNA transcripts is the prevalence of RNA hairpin loop secondary structures which form due to non-Watson Crick base pairing events (Sun *et al*, 2012). It is these RNA hairpin loops that contain the mature miRNA sequence (Figure 1.5) (Sun *et al*, 2012).

Considering the size of these pri-miRNA transcripts it isn't surprising to note that each transcript might contain several mature miRNA forms. One of these miRNA clusters is the mir-17-92 cluster (Figure 1.7). Mir-17-92 has a length of 800bp and is found within intron 3 of the gene *C13orf25* and encodes for 6 mature miRNA and another potential 6 miR*, all of these are transcribed from a single promoter (Olive, Jiang & He 2010).

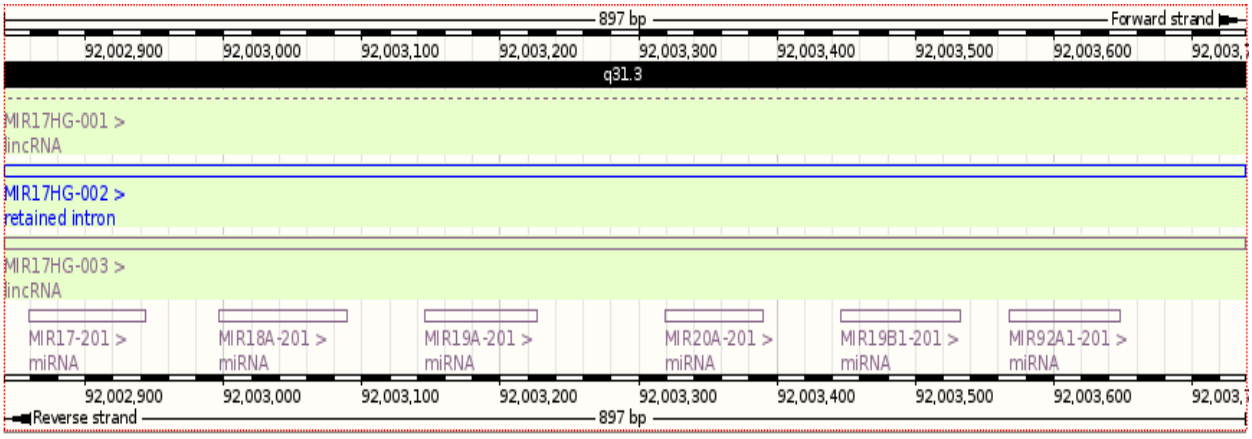


Figure 1.7: The miRNA cluster in *C13orf25* gene. This figure reproduced from *Ensembl* (Accessed 17/11/13) demonstrates the structure of a miRNA gene cluster. All these miRNA genes (open boxes) are in the same orientation and are located close together. Some miRNA clusters can have >3kb between miRNA genes and are still said to be of the same cluster.

Once capped the pri-miRNA transcript is processed by a protein complex called the microprocessor (Figure 1.8) (Han *et al*, 2006). A constituent of this microprocessor is DROSHA, a Class II RNase III like enzyme which selectively cleaves dsRNA hairpins (Han *et al*, 2006).

In humans DROSHA is a 159kDa protein and contains two RNAse III domains which cleave dsRNA in a magnesium/manganese dependant reaction (the metal ions stabilize RNA secondary structure) leaving a dsRNA hairpin of approximately ~80bp (Han *et al*, 2006). Whilst DROSHA is known to have a dsRNA binding domain it only weakly binds to its RNA substrate and as such a second protein which in humans is DGCR8 is required (Han *et al*, 2006).

Figure 1.8: Structure of the DROSHA protein. The RNAse III domains are coloured red (amino acids 876-1056 and 1107-1233 respectively) whilst the dsRNA binding domain (amino acids 1260-1334) is in green. The functionally essential domains are all located on the C-terminal side of the protein with only proline and arginine rich domains of no known functional significance located within the N-terminus.

DGCR8 contains two dsRNA binding domains and requires heme as a cofactor to facilitate binding to the RNA molecule (Barr *et al*, 2011). The heme group however is not always required for a variant that is N-terminal deficient (where the heme binding domain is located) is sufficient for forming a complex with DROSHA (Han *et al*, 2006). Han *et al* (2006) have demonstrated that DGCR8 binds at the ssRNA-dsRNA junction of the pri-miRNA and measures ~11bp from the junction up the stem (Figure 1.9).

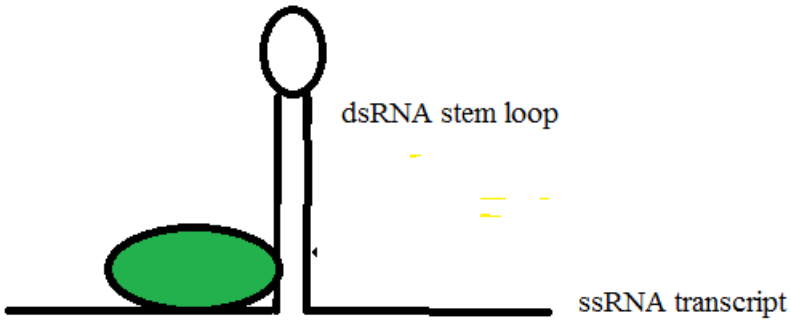


Figure 1.9 Schematic of DGCR8 binding to RNA. The green oval represents DGCR8 and its binding location to RNA during miRNA biogenesis. Its presence at the ssRNA-dsRNA junction allows the microprocessor complex to delineate the terminus of the stem loop RNA (dsRNA).

Han *et al* (2006) have identified the canonical mechanism of DROSHA processing. Pri-miRNA hairpins often have > 3 helical turns within their 3D structure. DROSHA recognises these turns and cleaves the hairpin at 2 helical turns away from the terminal loop and 1 helical turn from the basal sections (Han *et al*, 2006). The terminal loop is not required for processing of pri-miRNA into pre-miRNA but the basal sections is essential (where the hairpin joins the remainder of the transcript). This has led to the ssRNA-dsRNA junction anchoring model of DROSHA activity. This model states that the cleavage location is ~11bp away from the ssRNA-dsRNA junction which corresponds to the locations of the helical turns and the area which DROSHA has marked for cutting (Figure 1.9) (Han *et al*, 2006).

Characteristic of a DROSHA processed RNA hairpin (now called pre-miRNA) is the presence of a 2nt 3' overhang which is essential for its recognition by the exportin complex (Figure 1.10). The pre-miRNA exporting complex contains two proteins: EXPORTIN-5 and RAN-GTP (Kohler & Hurt 2007). RAN-GTP is a member of the Ras superfamily of GTP binding proteins and is responsible through the RanGTP cycle for carrying the pre-miRNA across the nuclear envelope (Kohler & Hurt 2007). EXPORTIN-5 is understood to bind to the

dsRNA helix of the pre-miRNA allowing it to traverse the nuclear envelope without undergoing enzymatic degradation (Kohler & Hurt 2007).

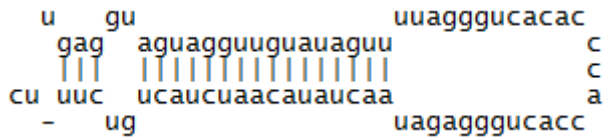


Figure 1.10: An example of a DROSHA processed pre-miRNA hairpin loop: Human miR let7a pre-miRNA. The processed hairpin has a 2nt overlap on its 3' arm. This is used by the EXPORTIN complex as a recognition signal. This image has been made in Notepad using the sequence for the hairpin precursor found in MirBase v19.

Upon export into the cytoplasm the pre-miRNA is processed by the DICER enzyme. DICER is the second RNAse III enzyme involved in the miRNA biogenesis pathway and cleaves the hairpin via an ATP and magnesium dependant mechanism (Park *et al*, 2011). The recognition site for DICER cleavage is not as well understood as that of DROSHA but it is known that the terminal hairpin loop structure is essential (Park *et al*, 2011). It has also been demonstrated to selectively bind to the 3' arm of the hairpin and then to cleave away the hairpin loop thus leaving an imperfectly Watson-Crick base paired RNA duplex which consequently breaks apart into two separate RNA strands either of which can be used as a mature miRNA (Carthew and Sontheimer, 2009).

1.3.2 The RNA Induced Silencing Complex

The mature miRNA strands lack any functional ability on their own and requires a protein complex known as the **R**NA **I**nduced **S**ilencing **C**omplex (RISC) to facilitate gene silencing (Kawamata and Tomari, 2010). In humans this complex contains Argonaute 2 (AGO2), Vasa Intronic Gene (VIG), Fragile-X mental retardation protein (FMRP) and Tudor staphylococcal nuclease (T-SN).

The Argonaute class of proteins are a family of evolutionary conserved proteins which encompass two families: Ago and PIWI both of which bind to small single stranded RNA molecules. Both classes have three domains in common (Kawamata and Tomari, 2010):-

PIWI: Magnesium dependant endonuclease domain.

MID: Responsible for anchoring the small RNA to the Ago protein via a 5' phosphate found on the guide strand.

PAZ: Secures the 3' end of the guide RNA strand to the protein.

T-SN is a member of the highly conserved Tudor proteins, these proteins are distinguished by the presence of a Tudor domain (Ying and Chen, 2012). T-SN is known to bind to dsRNA facilitating the targeting of the miRNA to target mRNA (Ying and Chen, 2012).

VIG is a known phospho-protein and has been demonstrated to be phosphorylated by Protein Kinase C (Ivanov *et al,* 2005). Its exact role within the RISC is unknown, however, a homolog of VIG (PAI-RBP-1) has been demonstrated to bind to AU-rich areas of RNA and has been demonstrated to regulate activity of the *plasminogen activator inhibitor* gene (Heaton *et al*, 2001).

FMRP is a ribonucleoprotein that binds to ssRNA via the presence of a KH-domain (Musco *et al*, 1997). The protein is known to associate with ribosomes and is demonstrated to have a role in nucleocytoplasmic shuttling of RNA (Eberhart *et al*, 1996). FRMP has been demonstrated to be a negative regulator of translation (Laggerbauer *et al*, 2001). FMRP has been shown to inhibit translation of mRNA both *in vitrio* and *in vivo*, the mechanism by which it regulates translation is thought to be due to its binding to mRNA and preventing the formation of a translation initiation complex (Laggerbauer *et al*, 2001). The role of FMRP within the RISC has been demonstrated by Muddashetty *et al* (2011) to be that of facilitating the interaction of the AGO-miRNA complex and the target mRNA.

1.3.3 Mechanism of Action of the RISC

After the loading of the miRNA target strand, the RISC complex recognises a target mRNA via base pairing between the seed sequence of the miRNA (nucleotides 2-8) and a complementary region of at least 7 nucleotides in the messenger RNA (Lewis *et al*, 2005). Each miRNA seems to be capable of regulating hundreds of mRNA species (Wu, Fan and Belasco, 2006).

Recent research has elucidated our understanding of miRNA mediated post-transcriptional repression and the variations between species; *Arabidopsis* for example employ direct cleavage of the mRNA species in its miRNA silencing whilst *Homo sapiens* silencing relies upon poly(A) de-adenylation resulting in mRNA destabilization (Figure 1.11) (Wu, Fan and Belasco, 2006).

Polyadenylation is as essential feature of eukaryotic mRNA expression. Guhaniyogi & Brewer (2010) note that within mammals' polyadenylation allows an mRNA transcript to escape the nucleus and not be enzymatically digested within the cytoplasm. It is also a

consensus site for the binding of Poly(A)-binding protein which recruits translation initiating proteins (Siddiqui *et al*, 2007).

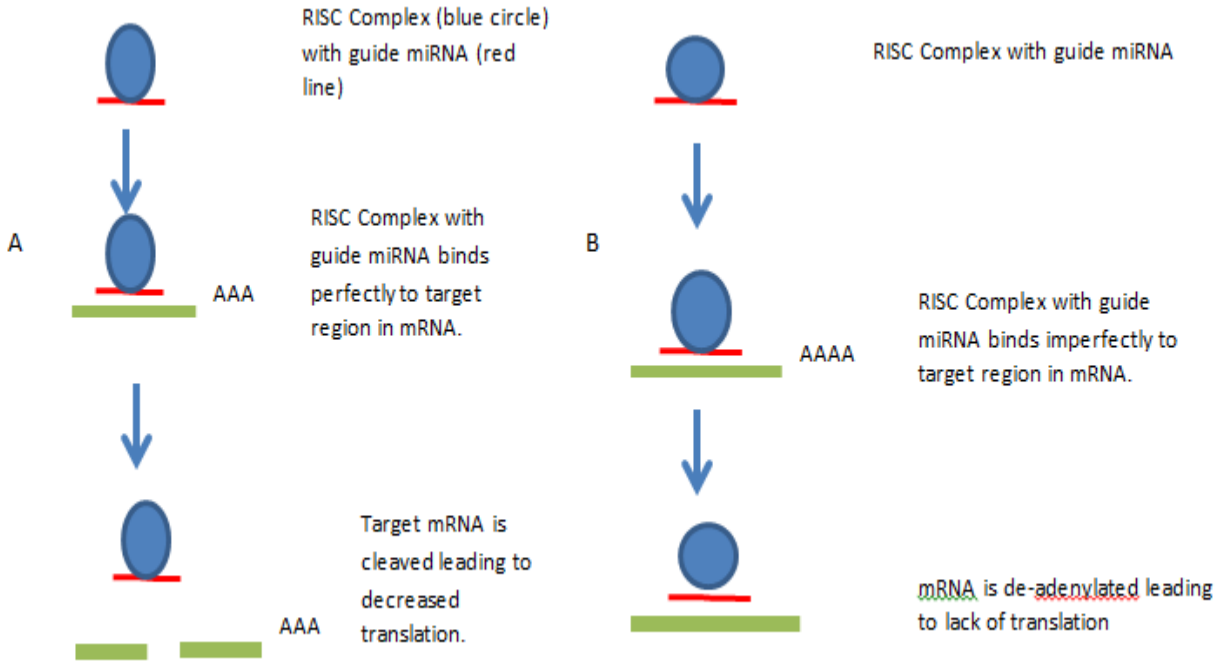


Figure 1.11: Schematic demonstrating the cross-species variability in miRNA induced gene silencing. A) Demonstrates the miRNA induced mRNA degradation in Arabidopsis. B) miRNA induced mRNA degradation in Human.

A second factor believed to be important in the human repression pathway is the conformational changes inflicted upon the mRNA species by the process of de-adenylation which prevents ribosomal binding and thus translation (Guo *et al*, 2010).

1.4 Circadian clock genes as miRNA hosts.

As of June 2014 no evidence exists for a core circadian clock gene existing as a miRNA host. This is not to say they cannot be. Below I will introduce *NPAS2*, the host gene for the rs1811399 SNP and putative miRNA.

1.4.1 miRNA have a role in rhythmicity.

The role that miRNA play in regulating the circadian cycle is recognised and some members of the network have been identified.

Garfield e*t al* (2009) identify the liver specific miR-122 as being integral for regulation of the clock within this tissue. miR-122 demonstrates a five-fold variance in expression dependant on the time of day and its expression is demonstrated to be reliant on REV-ERBα, a circadian clock protein. miR-122 has been shown to have complimentary recognition sites within the 3' UTR of many genes that are expressed in a circadian fashion, with several verified (*PPARβ, SMARCD1* and *HIST1h1C*) (Garfield *et al*, 2009). Of the gene products listed within the paper, several have roles within chromatin remodelling such as *HIST1h1C*, which could aid the circadian clock by allowing fine control over chromatin remodelling. Several others are directly linked with liver metabolism especially that of fatty acid metabolism (*glycogen synthase 1)*.

Two other miRNA species; miR-152 and miR-494, have been demonstrated to directly regulate the core circadian clock gene *BMAL1*. Shende *et al* (2011) have identified that the two miRNA are expressed in a circadian fashion and are out of phase with the BMAL1 protein. They are known to down-regulate expression of *BMAL1* mRNA but it is not known if they are directly controlled by *PERIOD* gene products. miRNA-132 has been demonstrated to control the negative arm of the circadian clock (Alvarez-Saavedra *et al*, 2011). This miRNA is induced in the SCN in a light dependant manner via the MAPK pathway. Other genes are expressed by light and one of these is *MECP2* which they have demonstrated to be a potent transcriptional activator of *PERIOD1* and *PERIOD2* (Alvarez-Saavedra *et al*, 2011). Intriguingly miR-132 has been shown to target *MECP2*, a transcription repressor protein, and down regulate its expression thus limiting translation of the PERIOD proteins.

1.4.2 *NPAS2* as a miRNA host

Hinske *et al* (2010) notes that the archetypal miRNA host gene are on average larger than non-miRNA host genes with the average host gene length measuring 84871bp against genes not hosting miRNA being 83747.5bp. They further stipulate that the intron numbers of these host genes is larger (13 as opposed to 10.5) and with significantly larger five 5' introns. The paper also notes that there is a preponderance of miRNA hosted within regulatory, neurogenic, metabolism and genes in other signalling pathways such as MAPK.

If we consult Figure 1.2 above we can conclude that the *NPAS2* gene is sufficiently large to qualify as a putative host. We can also see that it has a large number of introns with the initial 5' introns being larger than the remainder. The final point is of interest as *NPAS2* is, as described above, part of signalling networks (circadian and oxygen sensing), plays a part within gene regulation (circadian clock) and is involved in neurogenic processes such as memory formation.

1.5 The Evolution and Sequence Diversification of microRNA.

1.5.1 The role of mutations within miRNA

miRNA are strongly conserved regulators of gene expression (Lehnert *et al*, 2011). As such it is possible to take a miRNA sequence from a basal species such as *C.elegans* and search for homologs in higher species (Wheeler *et al*, 2009). This has been the practise in detecting miRNA across the phyla and has proved effective in detecting miRNA genes.

However genomes change over time and mutations arise; beneficial (Wheeler *et al*, 2009), negative (He *et al*, 2012) and neutral (Cuperus *et al*, 2011) mutations are described. On average every time a copy of the human genome is passed from parent to child it accumulates

between 100 and 200 novel mutations, a rate of 1 mutation per 30 million bases (Xue *et al*, 2009). These novel mutations can be identified as follows:

- Microsatellites: Also known as short sequencing repeats for which Whittaker *et al* (2003) define as sequences of DNA usually between 1bp-6bp which can be repeated in a single locus up to a 100 times. The majority of intra-species variation in microsatellite numbers results from replication slippage leading to an insertion or deletion event (Whittaker *et al*, 2003).

- Copy number variation (CNV): Lee *et al* (2007) define CNVs as alteration to the genome which results in individuals having variance in the number of copies of sections of DNA. It is estimated that up to 12% of human genomic DNA is in actuality comprised of CNVs.

- Single nucleotide polymorphisms (SNP): Barreiro *et al* (2008) give the definition of SNP as a genomic locus which between two members of a population has a single letter of DNA difference. For example person A has the sequence of ATA**T**AT at a specific locus whilst person B has the sequence ATA**C**AT (variation underlined).

The three examples described above account for a significant percentage of intra-species variation but the list should not be treated as exhaustive.

1.5.2 rs1811399: A SNP in intron 1 of *NPAS2*

Rs1811399 is a single nucleotide polymorphism (SNP) within the human genome at which the DNA nucleotide can either be adenine (A) or cytosine (C) and is located at co-ordinates chr2:100845196 within the UCSC hg19 database or at 2:101479014 within the Ensembl database. These co-ordinates locate the SNP within the first intron of the *NPAS2* gene.

As Fig.1.12 demonstrates; rs1811399 is located within intron 1. The mutation would therefore not interfere with the coding sequence for the final protein. Its location within the intron however could lead to dysregulation of downstream exons, a similar event can be seen in *P53* where rs12947788T interferes with the expression of exons 8,9,10,11 and 12 (Sailaja *et al*, 2012)

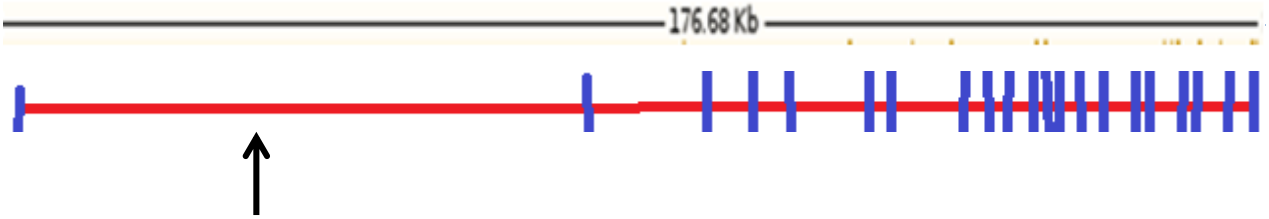
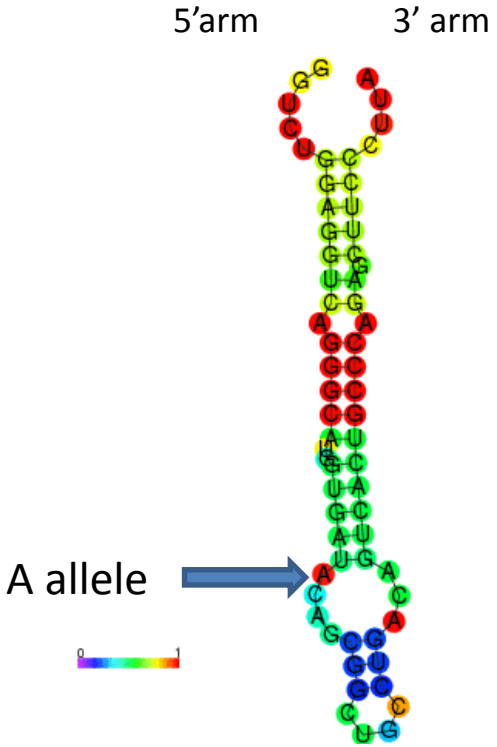


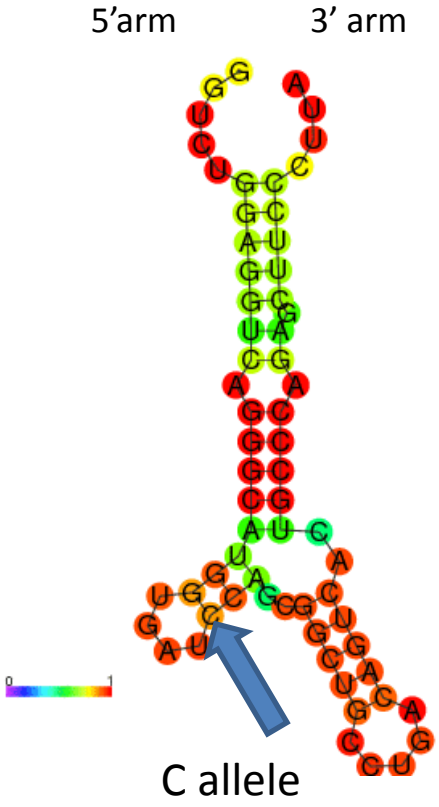
Figure 1.12: Schematic representing the intron-exon structure of the human *NPAS2* gene with an arrow highlighting the approximate location of the rs1811399 SNP (Blue lines denote coding exons and red lines are intronic regions). Being located in the intron the mutation would not be expected to interfere with protein structure/function.

Upon its initial discovery it was not apparent as to how this SNP could be contributing to an autism phenotype. It was believed that the C/A SNP (rs1811399) interfered with miRNA processing. When a sequence proximal to rs1811399A was modelled for RNA folding a hairpin was present that would allow for DICER processing (See Figure 1.13 A). Upon substitution of the A allele for the C the hairpin structure vanishes (See Figure 1.13 B). Based on the role secondary structure plays in miRNA maturation this would negatively influence biogenesis of the miRNA (Park *et al*, 2011).

A



B



Hairpin Sequence: 5’-

GGTCTGGAGGTCAGGGCATGGTGAT**M**CAGCGGCTGCCTGACAGTCACTGCCCAGAG

CTTCCCTTA-3’

Figure 1.13: RNA folding structures of the rs1811399 locus. Each nucleotide is colour coded to reflect the probability that the nucleotide is base-paired to its partner. Purple is 0 (no base pairing) and red is 1 (certain base pairing). A) This panel demonstrates the presence of the A allele at the rs1811399 location and the stereotypical pre-miRNA hairpin loop that is formed. B) This panel demonstrates the ancestral (C) allele. It is apparent that a forked loop replaces the terminal hairpin. The sequence for the hairpin is given at the bottom, ambiguity code M (highlighted in red) is where the A>C substitution occurs. Produced using Vienna RNAfold on 17/11/13.

The working hypothesis is that a cytosine nucleotide at rs1811399 interferes with maturation of a putative miRNA, possibly by interfering with DICER activity. This could result in no mature miRNA being produced leading to deregulation of downstream genes. Alternatively it could simply be a less effective substrate leading to less mature form being produced.

1.5.3 Evolution of miRNA

Whilst it is true that miRNA demonstrate high phylogenetic conservation it is also not the case that all miRNA exist unchanged in all life. *Cnidarians*, a phyla consisting of jellyfish, anemones and corals, have approximately half the number of miRNA species as do bilateral invertebrates (Grimson *et al*, 2008). This number is then greatly increased in vertebrates (Grimson *et al*, 2008). Current understanding identifies this increase in the number of miRNA species within an organism with increased morphological complexity (Grimson *et al*, 2008).

The mechanisms by which miRNA genes evolve are comparatively well understood and a summary will be given.

The first mechanism for miRNA propagation is the duplication of an existing miRNA gene. Evidence for the duplication of miRNA genes is extensive with there being several overarching miRNA families (e.g. Let7), which in humans contains 11 miRNA species, each exhibiting strong sequence homology with each other (Nozawa *et al*, 2010).

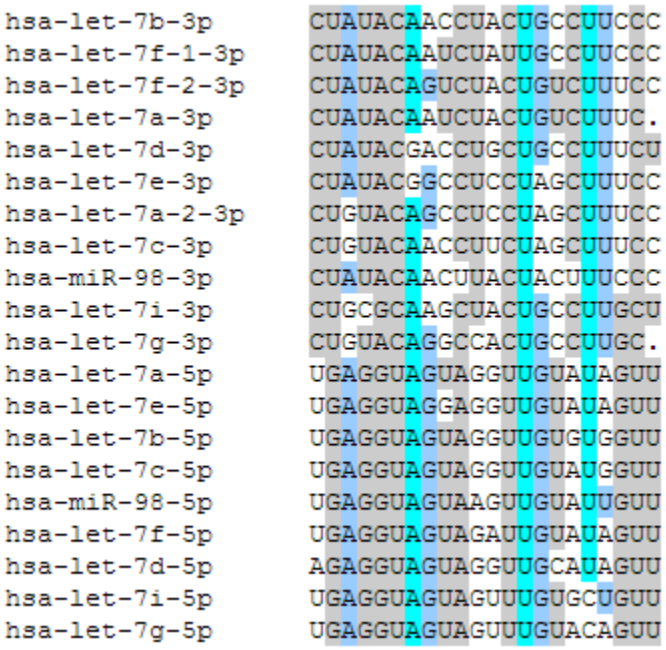


Figure 1.14: ClustalIW alignment of human Let7 miRNA genes.

This is expected to be a key feature in the evolution of the mammalian miRNA repertoire with a top end estimate of 151 novel miRNA families having evolved within this manner (Meunier *et al*¸ 2013). Gene duplication has been understood since the 1970s (Ohno, 1970) to be an evolutionary driver and occur more often than not from unequal crossing over of genetic material during meiosis. An accepted belief (Zhang, 2003) is that now the organism has two (or more) copies of the same gene it is able to accumulate mutations with greater tolerance within one copy of the gene thus, under positive evolutionary selection, can accumulate new functions. This last statement can account for the minor sequence differences exhibited within miRNA families. Within *Arabidopsis* one mechanism by which new miRNA families can arise is by the duplication of a protein coding gene which is inverted before

integration resulting in a strand of RNA which is anti-sense to the coding mRNA (Allen *et al*, 2004).

The *de novo* acquisition of new miRNA families is the second mechanism and requires greater explanation.

Introns are regarded as the prime location for spontaneous miRNA formation. Primarily this is due to the sheer amount of genetic material present within introns thus requiring specific point mutations to create a hairpin structure which can be recognised by the miRNA processing pathways (Campo-Paysaa *et al*, 2011). As an example Pederson (2010) has demonstrated that certain miRNA have derived from tRNA genes. It has been described that in mammals, plants and birds intron based miRNA tend to be younger (Cuperus *et al*, 2011; Li *et al*, 2009 and Meunier *et al*, 2012).

Hertel *et al* (2006) and Smalheiser &Torvik (2005) have conclusively demonstrated that transposable elements can form mature miRNA. Yuan *et al* (2011) have demonstrated that 278 human miRNA genes are directly derived from DNA transposons or retrotransposons, again within introns or inter-genic regions. Junctions between such elements are also fertile ground for miRNA evolution (Zhang *et al*, 2009).

1.5.4 Sequence diversification

After a duplication or *de novo* evolution event, diversification of a miRNA gene sequence can set in. This takes the form of single nucleotide polymorphisms within the microRNA gene which can directly impact on either the mature form or the precursor. The functional unit of a mature miRNA is the seed sequence. This is a sequence of nucleotides from position 2 to position 7 and is responsible for complimentary base pairing of RISC to target mRNA. A

single nucleotide polymorphism within this region can have dramatic effects with regards to target recognition and binding (see Figure 1.15 B) (Duan, Pek & Jin 2007).

Other forms of mutation which occur within the precursor have an effect on the mature form. One of these is the altering of the hairpin so that DICER processing is altered, leading to seed shifting. The seed shifting phenomena is tied in to entities referred to as isomiRs (Figure 1.15C). IsomiRs are mature miRNA that are produced from the same hairpin, but have varying sequences (Landgraf *et al*, 2007). It is noted that the majority of these isomiR have the same 5' sequence but have varying 3' termini. This is understood to be due to irregularities within processing and is not expected to influence the mature miRNA. However some evidence supports that they may have phenotypic impact within the fruit fly (Fernandez-Valverde, Taft & Mattick, 2010). Aberrant processing however, can also lead to 5' isomiR and the seed shifting phenomena (Figure 1.15 D). 5' isomiR are much less common with cells preferring the expression of one isomiR over another however it is interesting to note that this preference varies with species. For example *D.melanogaster* miR-281 has a different seed sequence than that of *Tribolium castaneum* (Red flour beetle) miR-281 which can be demonstrated to have undergone a seed shift (Marco *et al*, 2012).

SNPs within miRNA genes can also make them susceptible to RNA editing (Sun *et al*, 2009).

RNA editing is a well-established phenomenon in eukaryotes in which bases on the mRNA are modified so that the sequence differs from that of the genomic DNA. Within mammals there are two predominant forms of RNA editing: C to U and A to I (Brennicke, Marchfelder & Binder, 2006) .

C to U RNA editing was first described in the case of the *apoB* mRNA. APOB-100 is a protein that plays a role within cholesterol and triglyceride transport and is produced within the liver for incorporation within low and very low density lipoproteins (Powell *et al*, 1987).

APOB-48 is a smaller isoform of APOB-100 and is produced within the intestines to aid in transportation of lipid across membranes (Powell *et al,* 1987). APOB-48 corresponds to the N-terminus of APOB-100 and it was long thought of as a product of alternative splicing (Powell *et al*, 1987). In 1987 however, *APOB* in humans was sequenced as was the cDNA of APOB-48 and 100. Intriguingly the mRNA for hepatic APOB-100 and intestinal APOB-48 was identical bar a single base pair substitution at nucleotide 6,666; genomic DNA and the hepatic cDNA had a cytidine nucleotide, whereas the intestinal cDNA contained a thymidine (Powell *et al*, 1987). The implication of this substitution is the appearance of an in frame stop codon within the intestinal cDNA, thus resulting in a truncated protein variant (Powell *et al,* 1987).

The protein complex responsible for this is referred to as the editsome. The catalytic enzyme is APOBEC-1. APOBEC-1 is a 27kDa, zinc dependant cytidine deaminase which deaminases the cytidine base into a uridine base. In primates it is expressed in the intestines, whilst in rodents and other mammals it has wider expression pattern (Hadjiagapiou *et al*, 1994). A study of its amino acid and cDNA sequence however led to the discovery of several homologous cytidine deaminases with a wider expression profile (Brennicke, Marchfelder & Binder, 2006).

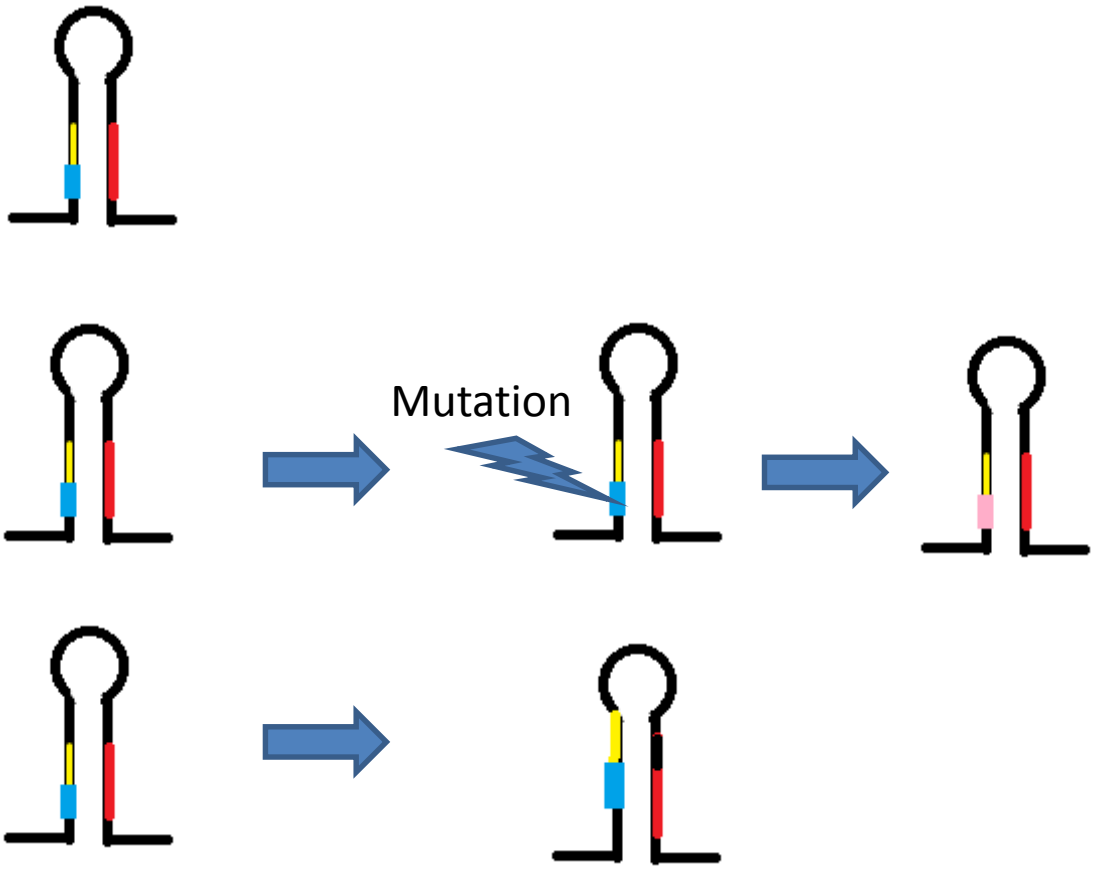
Adenosine deaminases that act on RNA, or ADAR, are the family of enzymes responsible for the second form of RNA editing which occurs within eukaryotes (Kim *et al*, 1996). A to I RNA editing is achieved when an ADAR enzyme (of which there are 2) deaminases adenosine at C6 to form inosine by using water as a nucleophile with requirements for magnesium and zinc as cofactors (Carter, 1998).

It is known that RNA hairpins are the preferred substrate for A to I RNA editing, thus a mutation might create an editsome consensus sequence that would then alter the precursor

(Luciano *et al*, 2004). This mechanism has been demonstrated within human miR-22 (Luciano *et al*, 2004) (Figure 1.14 E).

It was first described in the *C.elegans* (Griffiths-Jones *et al*, 2011) but has since been found in other phyla that arm switching leads to miRNA sequence diversity. As explained, once DICER processes the miRNA duplex, one strand is selected for use as a mature miRNA whilst for a long time the other was assumed to be discarded. What we now know is the second strand forms a second species of miRNA, miR 3'/5' depending on which arm of the pre-miRNA hairpin it is expressed, which can be detected in much smaller quantities than the primary miRNA (Khvorova, Reynolds & Jayasena, 2003). The selection process of which strand to use is largely unknown however, the thermodynamic stability of the dsRNA duplex plays a role (Khvorova, Reynolds & Jayasena, 2003). Of note however is that certain miRNA species switch between arms dependant on the development of the organism (Ro *et al*, 2007). Griffiths-Jones *et al* (2011) also hint at a possible sequence dependant "tuning mechanism" that controls which arm is expressed. It should be noted that arm switching can lead to a permanent change in which mature miRNA is expressed, thus leading to diversification of expressed miRNA pool (de Wit *et al*, 2009) (Figure 1.15 F).

A final mechanism by which sequence divergence can occur is hairpin shifting. This occurs when a mutation causes the hairpin structure, not the original sequence, to move up- or down-stream. Whilst the hairpin motif is kept, and can thus be processed, the original sequences are lost and whole new sequences are expressed as mature miRNA (See Figure 1.15 G).



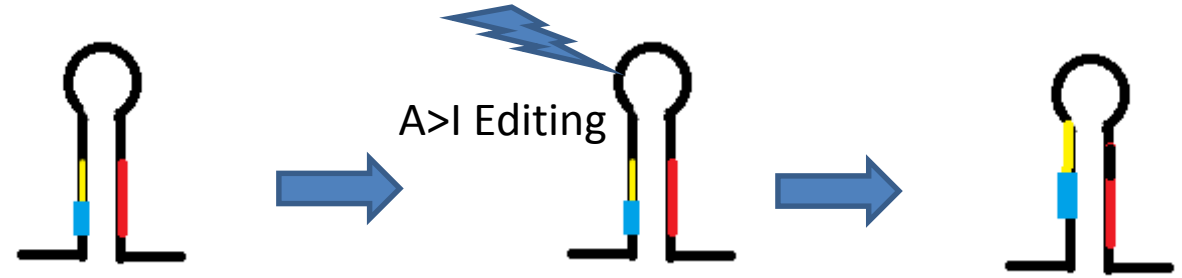
A) Ancestral miRNA

B) Direct mutation within seed site.

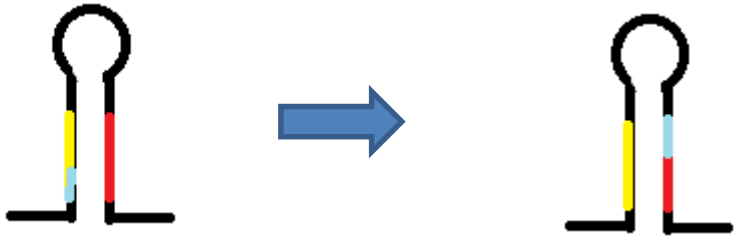
C) 5’ seed shift

```
miR-140       uaccacaggguagaaccacgg---
iso-140-1     -accacaggguagaaccacgga--
iso-140-2     uaccacaggguagaaccacgga--
iso-140-3     ---cacaggguagaaccacggaca
```

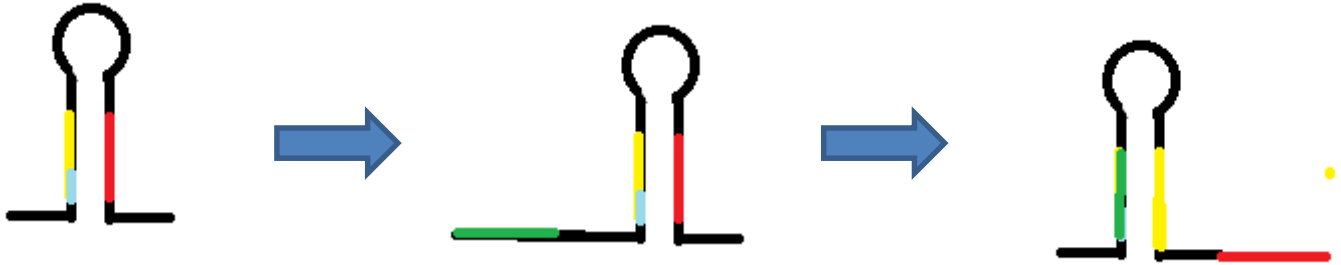
D) Alignment of several hsa-miR-140 isomiR.



E) RNA Editing



F) Arm switching



G) Hairpin switching

Figure 1.15: Series of panels depicting miRNA evolution. A) Schematic representation of the prototype miRNA precursor hairpin. Primary mature miRNA strand is highlighted in yellow with the seed region in blue. miR* strand is highlighted in red. B)Demonstrates a direct mutation within the seed site which then alters the specificity of the miRNA's binding. C) Example of a seed shift event within a prototype miRNA hairpin. It is evident that the primary strand has moved "up" the hairpin (yellow) and this has had a knock on effect on the location of the seed region, thus changing its sequence. This phenomenon usually occurs due to a mutation within the precursor which alters miRNA processing via DICER or DROSHA. D) Example of human miR-140 which has been revealed by deep sequencing to exhibit extensive seed shifting. The red box highlights the seed sequence for each isomiR with the canonical sequence listed first. E) RNA editing occurs via the interaction of an editsome with a suitable substrate. RNA editing can have multiple effects but in this schematic we have demonstrated it affecting the seed location. It can also affect the seed sequence directly by changing one of the nucleotides located there or can cause hairpin shifting or arm shifting. F) Arm shifting is a mechanism by which a precursor miRNA gene switches from which arm the mature form emanates from. This results in a mature miRNA with a completely different sequence. G) Hairpin shifting is a phenomenon where the hairpin motif migrates across a genomic region; this does not entail the movement of genetic sequence. As demonstrated within the schematic, the hairpin has moved upstream from its previous position, integrating a novel stretch of nucleotides into its hairpin. This has resulted in the previous 5' arm sequence now residing on the 3' arm and what was once on the 3'arm being moved out of the hairpin.

1.5.5 Change in expression patterns.

It is understood that young miRNA have low expression levels when compared to ancient miRNA (Chen & Rajewsky, 2007); for example primate specific miRNA have a 30 fold lower average expression level than ancient miRNAs (Meunier *et al*, 2012). It is expected that, should the miRNA survive purifying selection pressures its expression level will increase to similar levels of ancient miRNA (Meunier *et al*, 2012).

Meunier *et al* (2012) have demonstrated within mammals that miRNA genes tend to evolve with their target genes and in primates, a great number of the young miRNA species are preferentially expressed within two types of tissue: cerebellum and cortex. These new miRNA tend to target neuron specific genes. This second conclusion was reached on the basis that Zhang *et al* (2005) sequenced the cDNA of several nervous tissue genes and noted very large 3'UTR within mammals compared to other vertebrates, the region targeted by miRNA, with multiple target sites for old and new miRNA genes (Meunier *et al*, 2012).

As introns are the likely birth place for new miRNA, it would be expected that they would share the same promoter, indeed this was the prevailing opinion for some time (Rodriguez *et al*, 2004). Recent work however has identified that many intronic miRNA are expressed from their own promoters (Zhou *et al*, 2007). This result gives added impetus to findings of He *et al* (2012), who has demonstrated that younger, intronic miRNA tend to demonstrate lower levels of expression with their host gene. The hypothesis being that younger forms are initially expressed by chance from "junk" transcripts, with weak promoters. Splicing events cause the host mRNA to be degraded after excision of the miRNA. If the miRNA by chance exhibits a positive influence in conjunction with the host gene, overtime the miRNA will then become embedded within the transcript of the host gene. He *et al* (2012) cites human miR-338 which is co-expressed with *AATK* and is known to down-regulate *AATK* antagonists such

as *NOVA*. They postulate that long term survival of new miRNA is directly related to its ability to embed within its host gene's transcription unit.

1.6 Summary

The circadian clock allows for the temporal control of gene expression in eukaryotic life (Oster, 2005). Most control is exercised via the binding of circadian expressed transcription factors binding to specific recognition sites in promoter regions such as E-box consensus sequences (Akhtar *et al*, 2002). This pre-transcriptional regulation of gene expression is predicted to influence approximately 10% of all gene expression within mammals (Akhtar *et al*, 2002).

A further mechanism of regulating gene expression is post-transcriptionally via regulatory RNA. One form of regulatory RNA is known as microRNA (miRNA). miRNA are expressed from a intra- or inter-genic DNA and are processed in a two step mechanism by the Drosha and Dicer RNase enzymes to form a single stranded RNA molecule approximately 20 nucleotides long (Zhou *et al*, 2007). The mature 20 nucleotide piece of RNA can then be loaded onto a protein complex known as the RNA Induced Silencing Complex (RISC) which downregulate mRNA species via poly-A de-tailing (Wu, Fan & Belasco 2006). miRNA have been shown to complement the activity of the genes in which they are hosted by downregulating any genes that are antagonistic to their host's expression.

There is currently no evidence within the literature to suggest that any circadian clock gene hosts a miRNA gene. Were a circadian clock gene, however, to host a miRNA gene it is possible that the gene that evolved within that specific locus would make use of the circadian gene's oscillating pattern of expression. This would provide the circadian clock with a further mechanism for controlling gene expression within an oscillating pattern. Of note is the fact that some miRNA species have been identified, such as miR-152 and miR-494 (Shende *et al*,

2011), that oscillate in a circadian fashion and have been demonstrate to regulate clock controlled genes.

Sequence variation has been shown to be integral in driving miRNA evolution but has also been implicated in disruption of expression of established miRNA (Duan, Pak and Jin 2007). Given that sequence polymorphisms are extremely common within the human genome and many have been associated with disease it is possible that single nucleotide polymorphisms may influence the risk of disease by interfering with miRNA expression or activity. Given that genome wide association studies often identify SNPs that have no functional bearing on gene function or expression as being associated with disease, it is tempting to theorise that several of these mutations may be in previously un-described miRNA genes.

B.Nicholas *et al* (2007) linked the rs1811399 SNP, located within intron 1 of the *NPAS2* gene with autism at confidence levels of p=0.018. The genomic locus surrounding the SNP does not appear to play a role in transcription of downstream exons but is co-located within a region predicted to form a hairpin loop that is the required substrate for the miRNA processing proteins. The role of dysregulated miRNA within autism is established as is the role of the circadian clock. It may be the case that a previously un-documented miRNA exists within a circadian clock gene, possibly in order to exhibit similar spatio-temporal expression patterns. If rs1811399 influences the maturation of a novel miRNA it would provide a mechanism by which an intronic SNP could lead to increased risk of developing a disease.

1.7 Working hypothesis

Nicholas *et al* (2007) conducted a screen of circadian clock genes and demonstrated that rs1811399C is associated with autism at a confidence level of p=0.018. It is predicted that rs1811399C impacts upon the secondary structure of a novel miRNA gene impairing its biogenesis (Fig.1.13).

This body of work has been commissioned for the purpose of elucidating the impact of the rs1811399 SNP. We hope to establish whether or not it interferes with miRNA biogenesis. The following will also be addressed:

- Does *NPAS2* host a miRNA gene?
- Is *NPAS2* the host of a miRNA cluster?
- Is the expression of any miRNA dependant on the expression of the host gene?
- What is the expression profile of the miRNA?

The answers to these questions will illuminate a relatively poorly understood field as very little has been published on the influence of SNP on a miRNA biogenesis.

2. Materials and Methods.

2.1 Cell culture Methods.

The following cell lines were used during the course of the project:

a. HeLa: An epithelial cell line extracted from the cervix of a 49 year old lady. The cell line is adherent and requires media supplemented with 5% foetal bovine serum. For the rs1811399 SNP the cell line was heterozygous A|C.

b. SH-SY5Y: A neuroblastoma cell line extracted from the bone marrow of a 4 year old girl. The cell line is adherent and requires media supplemented with 5% foetal bovine serum. For the rs1811399 SNP the cell line was homozygous A|A.

c. HEK-293: A human embryonic kidney cell line. The cell line is adherent and requires media supplemented with 5% foetal bovine serum. For the rs1811399 SNP the cell line was heterozygous A|C.

d. HI2162: A lymphoblastic cell line extracted from an autistic patient. The cell line exists in suspension and requires media supplemented with 5% foetal bovine serum. The cells are homozygous A|A for the rs1811399 SNP.

e. HI2437: A lymphoblastic cell line extracted from an autistic patient. The cell line exists in suspension and requires media supplemented with 5% foetal bovine serum. The cells are homozygous A|A for the rs1811399 SNP.

f. HI2577: A lymphoblastic cell line extracted from an autistic patient. The cell line exists in suspension and requires media supplemented with 5% foetal bovine serum. The cells are heterozygous A|C for the rs1811399 SNP.

g. MEF: An embryonic mouse fibroblast cell line. The cell line is adherent in nature and requires media supplemented with 5% foetal bovine serum. The cell line did not contain an rs1811399 locus.

h. DT40: A chicken lymphoblastic cell line. The cell line exists in suspension and requires media supplemented with 10% foetal bovine serum and 1% chicken serum. The cell line did not contain an rs1811399 locus.

i. EB-176-J: A chimpanzee lymphoblastic cell line. The cell line exists in suspension and requires media supplemented with 5% foetal bovine serum. The cells are heterozygous A|C for the rs1811399 SNP.

2.1.1 Adherent cell line maintenance.

Adherent cell lines (HeLa, SH-SY5Y, MEF and HEK-293) were cultured on 60mm culture plates in 10ml of DMEM media supplemented with 10% FBS (Gibco) and 1% penicillin-streptomycin (Gibco) in a humidified 37°C incubator with 5% CO_2. Once cultures reached confluency (~$9x10^6$ cells) the media was removed and cells were trypsinised using a trypsin/EDTA solution (0.25% Trypsin, 0.2% EDTA) in order to detach the cells from the cell culture vessel. After incubation at 37°C cells were re-suspended in 10ml of fresh media and sub-cultured (or passaged) at a density of $2.5x10^6$ cells/ml into fresh vessels. Upon reaching 25 passages a cell line was disposed of and replaced with a cell line from an earlier generation.

Cell numbers were counted using an improved Neubauer haemocytometer.

2.1.2 Suspension cell line maintenance.

Suspension cell lines (HI2162, HI2437, HI2577, EB-176J and DT40) were cultured in T75 flasks in 10ml of RMPI-1640 media supplemented with 10% FBS (Gibco) and 1% penicillin-

streptomycin (Gibco) in a humidified 37°C incubator with 5% CO_2. DT40 cell lines required further supplementation by 1% chicken serum (Gibco).

Suspension cell lines were sub-cultured at cell densities of ~$8x10^6$ cells/ml into fresh flasks at an initial seeding density of $2x10^6$ cells/ml.

2.1.3 Cryo-storage of cells.

Cells were aliquotted at initial seeding densities (see 2.1.1 and 2.1.2) in media supplemented with 10% DMSO (Gibco). Cells were gradually frozen before storage at -70°C.

In order to awaken frozen cells, vials were gradually warmed up prior to being placed in pre-warmed media.

2.1.4 Serum shock.

Serum shock was conducted in order to synchronise a cell line's circadian gene expression. Cells were grown to confluence before serum enriched media was replaced with serum free media. Cells were incubated in serum free media for 12 hours before the media was replaced with serum rich media (DMEM and 50% FBS). When serum free media was replaced with serum rich media corresponds with time point 0h.

2.1.5 Drug treatment of human cell lines.

Cells were grown to confluency and their media removed and cells rinsed with PBS (Gibco). Media supplemented with the following drug concentrations was then added:

- Camptothecin 4μM final concentration.
- Cisplatin 5μg/ml final concentration.
- Gemcitabine 100nM final concentration.

Samples were taken at the stated time points up to 24 hours after introduction of the cytotoxic drugs.

2.1.6 Temperature Shock.

Exponentially growing cells were inoculated at 2.2×10^6 cells mL into 60mm tissue culture dishes containing 10 mL of supplemented culture medium. The cultures were cultured at 37 °C in a humidified 5% CO2 incubator for 24 hours and then transferred to an incubator at 32°C. After 24 hours at the dishes were removed from the incubator and RNA extraction undertaken.

2.2 Molecular Biology Methods.

2.2.1 Polymerase chain reaction (PCR).

PCR was performed for the specific amplification of DNA used to generate DNA for use in constructs as well as for the screening of cDNA libraries. PCR was performed in a final volume of 50 µl containing: template DNA (10-100ng), 1 x reaction buffer, 200µM of each dNTP, 1µM of each primer and 1U GoTaq-polymerase (NEB). The reaction was run in a thermocycler using the following program; After an initial denaturation step (95 °C, 3 min) 35 cycles of denaturation (95 °C, 30 sec), primer annealing (50-60 °C, 30 sec) and fragment extension (72 °C, ~1 min/kb) were followed by a final extension (72 °C, 5 min) before the end of the PCR.

For proof reading PCR (e.g. for studying miRNA sequences) we used Phusion DNA polymerase (NEB) to assure exact amplification.

2.2.2 Reverse transcription.

RNA was isolated as described before and samples were digested with DNase to remove contaminating DNA. 1µg RNA was digested with 2 units DNase in a final volume of 10 µl (in 1 x RT buffer). Samples were incubated at room temperature for 15 min before 1 µl STOP mix was added and incubation was continued for 15min at 70 °C to inactivate the DNase. These samples were used to reverse transcribe the RNA pool with oligo dT primers and/or random hexamer primers into cDNA. Reverse transcription was carried out using the Optimax First Strand cDNA synthesis kit as per the manufacturer's instructions: Mix 1µg of RNA, 1µl of primers and RNase-free water (up to 20µl) in a PCR tube. Incubation for 10 min at 65 °C and for 5 min at 4 °C was followed with the addition of the following: 1µl of 10mM dNTP mix, 1µl of RNaseOUT, 4µl of 10× RT buffer, 4µl of 0.1 M DTT, 8µl of 25mM MgCl2, and 1µl of Reverse Transcriptase. A final incubation for 60 min at 42 °C and for 5 min at 85 °C to inactive of the reaction. RNAse treatment was then undertaken to prevent RNA contamination for further PCR.

2.2.3 Plasmid DNA isolation from Escherichia coli cells.

Plasmid DNA from E. coli was extracted from overnight cultures using the DNA extraction mini- or midi-prep kit (Qiagen) depending on the amount of DNA required.

2.2.4 Genomic DNA isolation from Human Cell lines.

Genomic DNA from cell lines was extracted from confluent cell lines using the Fermentas Genomic DNA extraction kit as per the manufacturer's instructions.

Concentration of genomic DNA was measured spectrophotometrically using wavelengths of 260 and 320nm. The concentration can then be calculated using the following equation:

Concentration (µg/ml) = (A_{260} reading – A_{320} reading) × dilution factor × 50µg/ml

2.2.5 RNA isolation from Human Cell lines.

Total RNA was extracted from confluent cell lines using the Ambion miRVana RNA isolation kit as per the manufacturer's instructions. Total RNA was eluted in a final volume of 50µl of DEPC-treated water.

Concentration of RNA was measured spectrophotometrically using a wavelength of 260nm. The concentration can then be calculated using the following equation: Concentration (µg/ml) =40(Extinction coefficient of RNA) x A260

2.2.6 Small RNA enrichment of total RNA pool.

The eluate from the total RNA extraction methodology above contains RNA of varying molecular weights. Large and abundant RNA species can bias the RT-PCR reaction and result in less amplification of smaller RNA species.

Total RNA was produced using the miRVana isolation kit (Ambion) and underwent further purification steps using the miRVana isolation kit (Ambion) as per the manufacturer's instructions. This resulted in a pool of RNA molecules <250 nucleotides.

2.2.7 Poly(A) cloning of small RNA.

1µg of small RNAs extracted as above, 10µl of 5× E-PAP Buffer, 5µl of 25mM MnCl2, 5µl of 10mM ATP, 1µl(2 U) of E. coli Poly(A) Polymerase I and RNase-free water (up to 50µl) were mixed in RNAse free reaction tubes. The reaction was incubated at 37 °C for 1 hour prior to purification using the miRVana isolation column kit. Reverse transcription was then undertaken as described above with 1µg of purified tailed-RNA and using the miRTQ primer as opposed to random hexamer/poly-A primers.

Further cloning and PCR was undertaken as above with a gene specific primer and a primer specific to the miRTQ region.

2.2.6 DNA restriction enzyme digest.

Restriction enzymes were used to generate and verify plasmid constructs. 10 µg of DNA was cut with the appropriate amount of restriction enzyme (1U). Most digests were incubated at 37 °C for 2 hours. Double digests were performed when possible in the appropriate buffer.

The DNA was purified between successive digestions using the gel extraction.

2.2.7 Agarose Gel Electrophoresis.

Agarose gel electrophoresis was used to analyse DNA according to its molecular weight. To analyse plasmids were mixed with 1/10 vol of 10 x loading dye and run on a 1% (w/v) agarose (w/v). The gel was prepared with 1 x TAE buffer (40mM Tris, 1% (v/v) acetic acid, 1mM EDTA, pH 8.0), which was also used to run the gel at 120 V for 45 minutes. For estimation of the size of DNA in the applied sample, a marker with defined fragment sizes and DNA amounts was run in parallel. Bands were visualized by UV-light as the gel contained ethidium bromide (1.5mg/L).

miRNA PCR products were visualised and extracted from a 2% (w/v) agarose gel.

2.2.8 DNA extraction from agarose gel.

To extract DNA from agarose gels for further experiments, the required band was cut from the Ethidium bromide stained gel under UV-light with a scalpel. The gel slice was weighed and DNA was extracted using a Genelute gel extraction kit (Sigma), following the manufacturer's instructions. DNA was eluted from the filter column with 25µl H2O.

2.2.9 DNA dephosphorylation.

To prevent re-ligation of the vector backbone in ligation reactions, linearized plasmid DNA was 5'-dephosphorylated prior to ligation. Shrimp alkaline phosphatase was used (Roche) for de-phosphorylation according to the instructions of the manufacturer.

2.2.10 Ligation.

Vector backbones and the desired DNA fragments were ligated using T4-ligase (Roche). For a standard reaction 50ng of dephosphorylated vector DNA was incubated with a threefold molar excess of insert, 1x buffer and 1U of T4 ligase at 16 °C over-night.

2.2.11 Heat shock transformation of *Escherichia coli.*

A single fresh colony of *E. coli* TOP10 cells was inoculated into 50 ml LB medium and grown over night at 37 °C with shaking (250 rpm). 5ml of the starter culture was used to inoculate 500 ml of LB medium, which was grown at 37 °C, 220 rpm to an OD600 of ~0.4.

After cooling on ice the cells were harvested by centrifugation (30 min, 4 °C, 2500 x g), the pellet was re-suspended in 250ml ice cold 100mM CaCl2, centrifuged again and then re-suspended in 50ml 100mM CaCl2. After another centrifugation, cells were re-suspended in 5 ml 100 mM CaCl2, 20% (w/v) glycerol and 50 µl aliquots were transferred into pre-chilled, sterile reaction tubes. Cells were frozen immediately in liquid nitrogen and stored at -70 °C until use.

50µl of chemically competent TOP10 cells was thawed on ice and gently mixed with 10ng of DNA in chilled reaction tubes. After incubation on ice for 10 min, cells were incubated at 42 °C for 1 min, suspended in 900 µl SOC-medium and incubated at 37 °C for 45 min to

recover. Afterwards cells were plated on selective media (LB + antibiotics) and incubated at 37 °C over-night. Transformed colonies were picked the following day for sub-culturing.

2.2.12 PCR-mediated site directed mutatagenesis.

PCR primers were designed that incorporated the required mutation for both the sense and anti-sense strands. PCR is conducted using Phusion polymerase (NEB) and results in two amplicons, one with the mutation at the 3' end and the second amplicon with the mutation at the 5' end. Both amplicons have a region of homology with each other, conferred by the mutagenesis primers. The two amplicons can then be fused together and amplified using Phusion polymerase (NEB) as previously described.

2.2.13 RNase Protection Assay.

Using PCR reaction described above, a DNA amplicon of the target RNA region is produced. The DNA amplicon is then fused to a T7 promoter region using T4 ligase (Roche) as above. Once fused to the T7 promoter, the DNA is incubated with T7 polymerase (Ambion) under the following conditions: 2µl 10x transcription buffer, 1µl 10mM ATP, 1µl 10mM GTP, 1µl 10mM UTP, 10 units of T7 polymerase, 8.5µl DEPC-H_2O and 2.5µl of 10mM radiolabelled-CTP. The reaction is incubated for 30 minutes at 37°C before DNase I is introduced to remove the template.

The radiolabelled RNA probe is then purified on a 15% denaturing acrylamide gel (21g urea, 12.5ml DEPC-H_2O, 18.75ml acrylamide solution, 2.5ml 10xTBE, 400µl 10% APS and 40µl TEMED) running at 180V for 1.5 hours. The probe is excised using a sample and placed in 300µl of Probe Elution Buffer (Ambion).

The eluted probe is hybridised to the target RNA via over-night incubation at 42°C using the ribonuclease protection kit (Ambion) before the mix is digested with RNase A/T1 for 45

minutes at 37°C. The digestion reaction is terminated by addition of 225µl of RNAse inactivation solution and 225µl of 100% ethanol. The solution is then precipitated for 2 hours at -20°C prior to centrifugation of 12000rpm at 4°C for 1.5 hours. The resulting pellet was then re-suspended in 10µl of gel loading buffer (Ambion) and rand on a 15% denaturing acrylamide gel.

Imaging the reaction was conducted via exposure to chemi-luminescence film.

2.2.14 *In vitro* Dicer assay.

An un-radiolabelled RNA probe was produced and purified as above.

HeLa cells were grown to confluency and were lysed with 500µl of subcellular fractionation buffer (250mM Sucrose, 20mM HEPES (pH7.4), 10mM KCl, 1.5mM $MgCl_2$, 1mM EDTA, 1mM EGTA, make up to 50ml with sterile water), the plates were scraped and the lysate collected and passed through a 25Ga needle. After incubation on ice for 10 minutes the nuclear fraction was extracted with centrifugation at 3000rpm for 5 minutes. The remaining supernatant is the cytosolic fraction. Protein yield can be calculated using the Bradford assay.

Prior to the assay the protein extract is treated with RNase 1 (Ambion) for 1 hour and then inactivated by RNase A/TI inactivation solution (Ambion).

10µg of cytoplasmic protein is mixed with 60ng of RNA probe in 50ul of Dicer assay buffer (20mM Tris-HCl pH7.5, 250mM NaCl and 2.5mM $MgCl_2$) and incubated for 1 hour at 37°C before running on a 2% agarose gel.

2.2.15 *In vitro* RNA editing.

10µg of nuclear protein extract and 100ng of RNA probe are produced as described above. These are incubated together in a final volume of 50µl of RNA editing buffer (2U RNasin, 0.25mM DTT, 50mM EDTA, 50mM KCl, 10% glycerol and 10mM HEPES (pH7.9)). The solution is then incubated for 3 hours at 30°C prior to poly(A) small RNA cloning noted above.

2.2.16 Electrophoretic Mobility Shift Assay (Panomics).

The electrophoretic mobility shift assay can demonstrate protein binding to specific regions of DNA via the retardation of DNA on a gel. Using PCR DNA probes of a 150b.p region centred on rs1811399 were created, one for each allele whilst a nuclear extract was made from HeLa cells.

A reaction buffer was prepared as per the manufacturer's instructions (3.6ml 50% glycerol, 360ul 1M HEPES (pH 7.9), 120µl 1M Tris-HCl (pH 8.0), 60µl 0.5M EDTA (pH 8.0), 150µl 100mM DTT and 10.7ml H_2O) and 12µl of this buffer was incubated with 3µl of BSA (1µg/µl), 2µl of 0.5µg/µl Poly(dI-dC), 3µl of nuclear extract (1µg/µl) and 3µl H_2O. This mixture was incubated at room temperature for 20 minutes prior to the addition of the DNA probe (1µg) and was further incubated at room temperature for 20 minutes.

The mixture was then ran on a 2% agarose gel and imaged with ethidium bromide staining.

2.2.17 TOPO-TA Cloning.

TOPO-TA cloning (Invitrogen) was used to make cDNA libraries for sequencing of miRNA sequences.

The vector used for the cloning was pCR-DNA3 with topoisomerase-I covalently bound to an overhanging thymidine nucleotide located on the linearized vector.

Taq polymerase has a template-independent terminal transferase function that places an adenine nucleotide on the 3' end of all PCR products. The unbound hydroxyl group on the 5' end of the PCR amplicon then attacks the covalent bond between the topoisomerase-I and overhanging thymidine residue resulting in instant ligation of the amplicon to the vector.

4µl of DNA amplicon produced as described above was incubated with 1µl of linearized vector and 1µl of salt solution (200mM NaCl, 10mM $MgCl_2$) at room temperature for 5 minutes. Upon completion of the incubation the vector was transformed into TOP-10 competent *E.coli* as previously described.

Sequencing of inserts can be undertaken using universal M13 primers.

2.2.18 Primers.

NPAS2 F-TGGGAACCTCAGGCTATGAC R-AGTCTGCAGCCAGATCCACT

CLOCK F-CCAGAAGGGGAACATTCAGA R-TGGCTCCTTTGGGTCTATTG

PER1 F-AGGTACCTGGAGAGCTGCAA R-TTCTTGGTCCCCACAGAGAC

PER2 F-TCCAGTGGACATGAGACCAA R-CGCTACTGCAGCCACTTGTA

CRY1 F-CAGGTTGTAGCAGCAGTGGA R-GACTAGGACGTTTCCCACCA

rs1811399 F-CTTTTCTAGTCTACTGAGGAAGG R-CAAATCAAGGGCTGGTATTAAC

nmiR-1273 F-GGCATGAGAATCGCCTGAAC R-GAGATGGAGTCTCGCTCTG

pre-miR-122 F-CAATGGTGGAATGTGGAGGT R-CATTTATCGAGGGAAGGATT

β-ACTIN F-CGTCATACTCCTGCTTGCTGATCC

R-GAGCGCGGGTACAGCTTCACC

Promoter construct F-AAGCTTACACAAGCTTACCATGACC

R- CTCATGCCCCCATAACGAAAGGGAACACACAGCAAGTGTTT

GAPDH F-AACCTGCCAAATATGATGAC R-ATACCAGGAAATGAGCTTGA

HPRT1 F-TGACACTGGCAAAACAATGCA R-GGTCCTTTTCACCAGCAAGCT

miRTQ

CGAATTCTAGAGCTCGAGGCAGGCGACATGGCTGGCTAGTTAAGCTTGGTACCG

AGCTCGGATCCACTAGTCCTTTTTTTTTTTTTTTTTTTTTTT

1273a GGGCGACAAAGCAAGACTCTTTCTT

92a AGGTTGGGATCGGTTGCAATGCT

191 CAACGGAATCCCAAAAGCAGCTG

106b TAAAGTGCTGACAGTGCAGAT

21 TAGCTTATCAGACTGATGTTGA

miR6725-3p GGGAAGCTCTGGGCAGTGACTG

miR6725-5p CATGCCCTGACCTCCAGACCTG

miR1273h GGTTCAGGCGATTCTCATGCCT

Let7a TGAGGTAGTAGGTTGTATAG

RTQ-UNIr CGAATTCTAGAGCTCGAGGCAGG

2.3 Plasmid constructs.

The following plasmids were used during this work:

a. pBSK+/KS-: A standard vectors for the sub-cloning of PCR products to facilitate their specific excision with restriction enzymes. An ampicillin resistance cassette and a multiple cloning site which is located in the LacZ gene allow for blue/whiteselection of positive transformants on plates with X-Gal and IPTG. The vector was kindly provided by J.Eykelenbloom of University of Galway.

b. pJE28: A vector designed and provided by J.Eykelenbloom of University of Galway. The vector contains sequences of homology with the chicken *ovalbumin* locus which allows for integration of genetic material. The vector contains a puromycin cassette for selection in DT40 cell lines.

c. pcDNA3.1 puro+: A vector designed for the expression of genetic material in eukaryotic cell lines. Expression is driven by a promoter extracted from cytomegalovirus (CMV). A vector with the puromycin resistance cassette was kindly provided by T.Dantes and J.Eykelenbloom of University of Galway.

d. psi-RNA: A vector for the expression of short-hairpin RNA sequences. Expression of shRNA is driven by a h7sk promoter and eukaryotic cell selection is undertaken via zeocin selection. The vector was purchased from Invivogen.

e. pCR3.1-TOPO: A vector that allows for rapid integration of *Taq* polymerase PCR products. This vector was used in the cloning of small RNA cDNA libraries and sequencing. An ampicillin resistance cassette and a cloning site which is located in the LacZ gene allow for blue/white selection of positive transformants on plates containing X-Gal.

2.4 Bioinformatics.

2.4.1 Sequence, Structure and Conservation (SSC) profiler.

SSCProfiler was developed by Oulas *et al* (2009) to detect putative miRNA sequences within genomic DNA regions. The server was hosted by the Computational Biology Lab at the Institute of Molecular Biology and Biotechnology of Heraklion University and can be found at http://mirna.imbb.forth.gr/SSCprofiler.html.

SSCProfiler utilizes a probabilistic approach based upon Profile Hidden Markov Models (HMM) mathematics to identify miRNA precursor hairpin loops. The software package parses genomic DNA between two stated co-ordinates (UCSC hg17) less than 1kb apart into 104nt segments and moves across the genomic DNA in 11nt windows. The software then filters its results based on eight categories: number of hairpins, number of nucleotides situated in bulges, number of nucleotides situated in loops, asymmetry, number of nucleotides located within bulges and loops, length of hairpin, minimum free energy and conservation of nucleotides.

2.4.2 *Vienna* RNAFold.

The RNAFold server was used to predict the secondary structures of single stranded RNA molecules. The software was located at: http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi

The algorithm developed by Zuker *et al* (1981) utilises the total free energy given the nucleotide sequence, the final structure of an RNA molecule is the secondary structure that consists the minimum of free energy. RNA secondary structures can be categorised as loops and of duplexed RNA nucleotides external to the loop. As described by Zuker *et al* (1981): "The loop-based model categorises the free energy F(s) of an RNA secondary structure s as the sum of the contributing free energies FL of the loops L contained in s. According to the

chosen energy parameter set and a given temperature (defaults to 37 °C) the secondary structure s that minimizes F(s) is computed."

The software could compute structures for RNA molecules up-to 7.5kb, unless otherwise stated the sequence size used for this project was ~150 nt.

2.4.3 *In silico* Drosha processing.

For *in silico* prediction of Drosha processing a software package developed by Helvik *et al* (2007) was utilised. The software can be located at: https://demo1.interagon.com/miRNA/

The software runs a support vector machine (SVM) algorithm that has been designed to consider the following variables in identifying Drosha processing sites; Precursor length and loop size, distance from the 5' processing site to the loop start, nucleotide occurrences and frequencies at each position in the 24 nt regions of the precursor 5' and 3' arms, total number and identity of each base-pair in the 24 nt at the precursor base, nucleotide occurrences at each position in the 50 nt of the 5' and 3' flanking regions outside the precursor.

The algorithm can distinguish between random hairpins and those of pri-miRNA in ~80% of cases and in those miRNA shown to be a true positive the algorithm correctly identifies that cutting site in ~90% of the cases.

A sequence of <108nt was required by the software for analysis. Sequences were selected based upon their secondary structures as described using the RNAFold service described above.

Scoring is based on a positive predictive value (PPV) system with a value ranging from -0.5-1.0 demonstrating a favourable likelihood of a true result.

2.4.4 *In silico* Dicer processing.

In silico Dicer is software designed by Tippman *et al* (2005) to determine Dicer cutting sites in RNA hairpins.

The program can be located at http://bibiserv.techfak.uni-bielefeld.de/insilicodicer/webstart.html

The software compares inputted sequence against known miRNA sequences and estimates Dicer cutting sites based on sequence conservation between the query sequence and known miRNA molecules. The software also compares the primary and secondary structure of the RNA sequence against known targets of *Arabidopsis* RNase III enzymes to increase result confidence.

2.4.5 *Ensembl* Genome Browser.

Ensembl is a repository of DNA sequences of multiple species. It was possible using the browser to identify SNPs and assess their population distribution using data published by, amongst others, the 1000Genomes project.

Given that the database was continually updated during this project, efforts were made to ensure that only the most relevant database was utilised (version 71).

For the purposes of this work the following functions of the browser were utilised:

- Genomic context
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phylogenetic context

All searches were conducted for SNP rs1811399.

2.4.6 *UCSC* Genome Browser.

Similar to *Ensembl* above, the University of California, Santa Cruz (UCSC) browser allows access to genome information.

For the purpose of this work the following non-default tracks were utilised on the browser:

- Switchgear TSS: This identifies the location of transcription start sites (TSS) throughout the human genome based on experimental evidence.
- TFBS Conserved: This track contains the score of transcription factor binding sites conserved across human/mouse and rat. A binding site is conserved across the species when the score meets the threshold for its binding matrix across all 3 species. The Transfac Matrix Database is responsible for collating all the requisite scores.
- CSHL Small RNA-seq: The Cold Spring Harbor Lab (CSHL) small RNA track demonstrate short total RNA sequencing data from various cell lines that can be individually selected.
- ENC TF Binding: Demonstrate ChIP-seq evidence for transcription factor binding in a specified locus.

Data was exported directly from the website using their Table Browser functionality.

2.4.7 TargetScan.

TargetScan was designed and maintained by the Bartel laboratory (Bartel *et al*, 2005)

(TargetScan (version 6.2) was accessed via http://www.targetscan.org/ .

TargetScan is a tool for predicting potential mRNA targets for mature miRNA. Targets are

identified via base-pairing sequence homology of the miRNA's seed region and a gene's 3'UTR. Predictions are scored on the predicted efficacy of targeting as calculated using the context+ scores of the sites. The context+ score is calculated via: location in 3'UTR of target site, local AU nucleotide makeup, site-type scoring and 3' pairing of nucleotides from outside the seed sequence.

In order to input seed sequences of novel miRNA the TargetScan Custom (http://www.targetscan.org/vert_50/seedmatch.html) program was utilised. Seed sequences of 7-8nt can be inputted into the software and targets detected using the same methodology as the regular TargetScan.

2.4.8 gProfiler.

gProfiler allows the construction of protein interaction networks and can be accessed at http://biit.cs.ut.ee/gprofiler/ .

gProfiler makes use of publicly available data produced by the Gene Ontology network who have ascertained the relationship between many proteins and the pathways in which they have a role (via KEGG annotations). Inputting a variety of genes into the software will allow the program to deduce any relationship the expressed proteins will have with each other, this includes: member of protein complexes, transcription factor required to drive expression of target gene or a shared localisation pattern.

Gene information can be inputted into gProfiler from a variety of sources, however gene data extracted from the TargetScan software noted above has been used for this thesis.

2.4.9 ImageJ.

In the absence of quantitative data, it was possible to utilise ImageJ to semi-quantitatively analysis PCR products ran on an agarose gel. ImageJ is an open source software package produced by the National Institutes for Health (NIH) and version 1.48 was downloaded from: http://imagej.nih.gov/ij/download.html .

Briefly, agarose lanes can be defined using a selection tool and the background subtracted reducing the image to the PCR amplicons. The intensity of these bands is then reported by the software package allowing relative semi-quantification to be conducted.

3. *NPAS2* is inducible in a wide variety of cell lines.

The work presented below has identified the mechanisms of *NPAS2* transcriptional induction. Brief serum starvation induces circadian clock oscillation within HeLa and SH-SY5Y cell lines. Heat and DNA damage is also demonstrated to induce expression of *NPAS2*. This will be important in establishing the relationship between the host gene (*NPAS2*) and the miRNA.

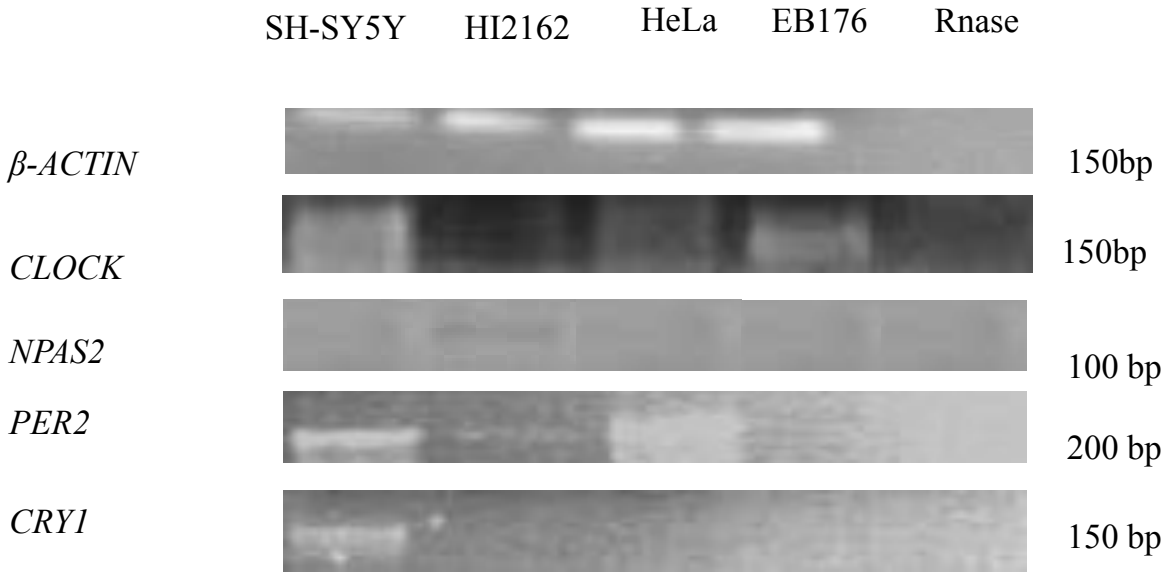
The experiments herein are also designed to identify the *differences* in expression of the circadian clock genes within various tissue types.

3.1 Expression of Core Circadian Clock genes in asynchronous cells

When cells are removed from an organism they lose the ability to synchronise their core clock (Olofsson & McDonald, 2010). They are then referred to as being asynchronous. Each cell lines had 1µg of total cellular RNA extracted and reverse transcribed to cDNA in order to assess the expression of the clock genes. The negative control had Rnase A added to remove the RNA before reverse transcription. The Rnase reaction was performed to ensure that the PCR fragment amplified is a genuine product of RNA reverse transcription, each primer pair was also designed to cross exon-exon boundaries to ensure only a spliced mRNA product would be amplified.

Fig.3.1 below demonstrates that *CLOCK* is expressed across 3 cell lines (HI2162 being the exception) with *NPAS2* being absent in the 4 cell lines. *PER2* is expressed in both SH-SY5Y and HeLA and weakly in HI2162 and absent in EB176. *CRY1* was only detectable in SH-SY5Y cell line.

A



B

	Cell Line			
Gene	SH-SY5Y	HI2162	HeLa	EB176
β-*Actin*	26881.01	27797.29	30993.98	31001.12
CLOCK	1355.487	101.87	150.02	2020.326
NPAS2	15.254	19.254	20.177	19.645
PER2	1126.284	97.216	1301.897	0
CRY1	988.564	0	0	0

Figure 3.1: Expression of core circadian clock genes within four cell lines plus Rnase A treated control. A) RNA was extracted from each of the cell lines and reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: CLOCK-F+CLOCK-R, B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R, PER2-F+PER2-R and CRY1-F+CRY1-R. For the negative control the RNA was treated with RNase A prior to reverse transcription. Panel B demonstrates the relative intensity of each PCR band. In the absence of quantitative data, ImageJ was used to provide data on the relative amounts of product in each gel

based on the intensity of each band when stained with ethidium bromide and viewed under UV.

Not all cell lines expressed the circadian clock genes, which either implies the absence of expression of these genes within those cell lines or the number of cells expressing these genes at the time of RNA extraction was too low thus limiting the amount of the specific miRNA within the pool of total RNA.

According to the literature (Baskerville *et al*, 2005) miRNA are often co-expressed with their host gene. In asynchronous cells it was attempted to reset the phase of the circadian cycle in order to detect both the host gene (*NPAS2*) and the miRNA.

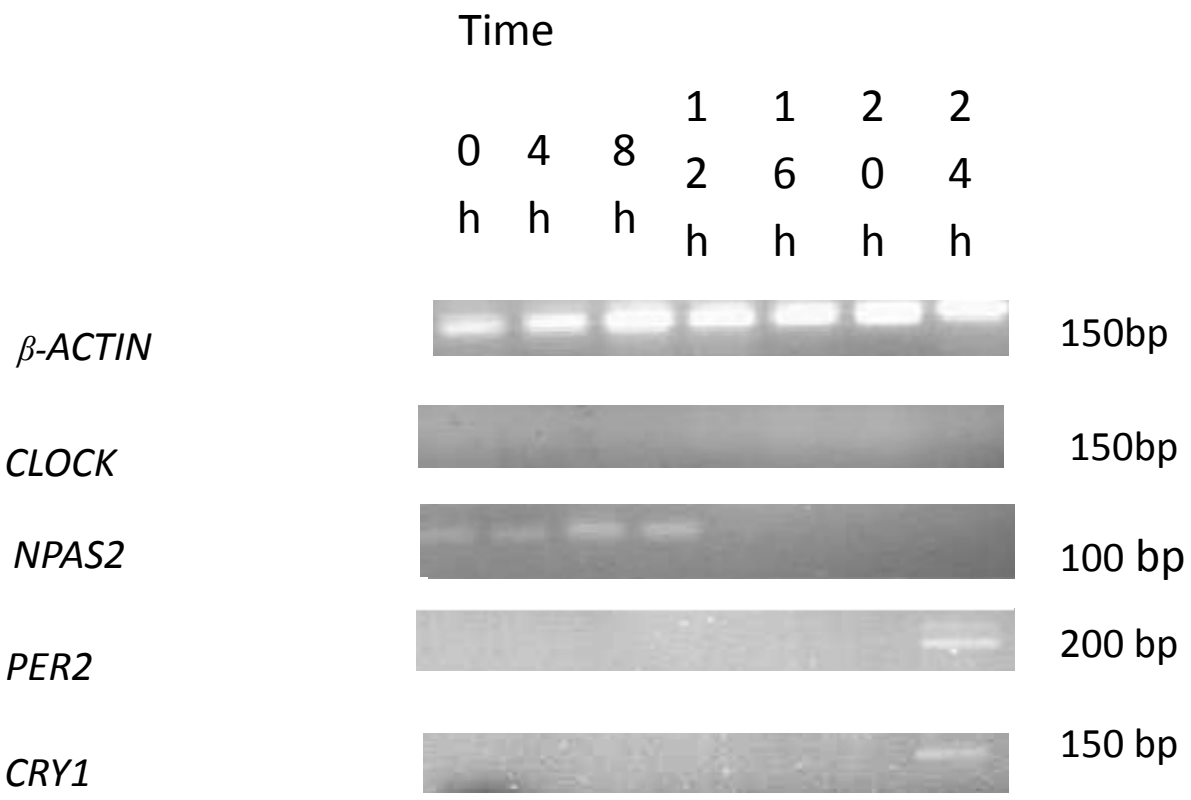
3.2 Serum starvation of cell lines

In order to circumvent the asynchronous nature of normal tissue culture conditions there are several mechanisms for synchronising the circadian cycle. In a whole organism the mechanisms for synchronisation would include light entrainment or nutritional uptake, as the cells have no light sensitive regions, a modified version of the second is required. Briefly cells are incubated within a protein free media before being subjected to a high protein environment. The resultant serum shock then induces expression of circadian genes (Balsalobre *et al*, 1998).

Figures 3.2 and 3.3 demonstrate that serum shock is able to synchronise the circadian clock in HeLa and SH-SY5Y cell lines. Two cell lines were tested demonstrated an initiation of expression of the core circadian clock. Initial expression of the positive transcription clock genes (*CLOCK* and/or *NPAS2*) and the silencing of the repressing arm (*PER2* and *CRY1*). NPAS2 or CLOCK proteins are required to activate the transcription of *PER2* and *CRY1*.

Figures 3.2 and 3.3 demonstrate that *NPAS2* and *CLOCK* expression begins at 4h whilst *PER2* and *CRY1* only begin around 24h.

A



B

	Time (h)						
Gene	0	4	8	12	16	20	24
β-*Actin*	30186.46	30543.92	30980.92	30482.13	30832.04	31001.2	30997.15
CLOCK	0	0	0	0	0	0	0
NPAS2	145.647	288.877	620.033	650.489	0	0	0
PER2	150.632	0	0	0	0	0	557.366
CRY1	0	0	0	0	0	0	575.654

Figure 3.2: HeLa Serum Shock circadian clock expression profile. A) HeLa cells were grown to 80% confluency and incubated in zero serum media for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media (50% FBS). RNA was extracted at each of the time points and reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: CLOCK-F+CLOCK-R, B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R, PER2-F+PER2-R and CRY1-F+CRY1-R. Panel B demonstrates the relative intensity of each PCR band. In the absence of quantitative data, ImageJ was used to provide data on the relative amounts of product in each gel based on the intensity of each band when stained with ethidium bromide and viewed under UV.

After the serum shock *NPAS2* was detectable for 12 hours. After this point its expression declined. *PER2* and *CRY1* became detectable 24h after the initiation of the experiment. This observation is consistent with the model of *PER2* and *CRY1* expression being dependant on the initial expression of genes such as *NPAS2* or *CLOCK* which constitute part of the positive arm of the circadian clock.

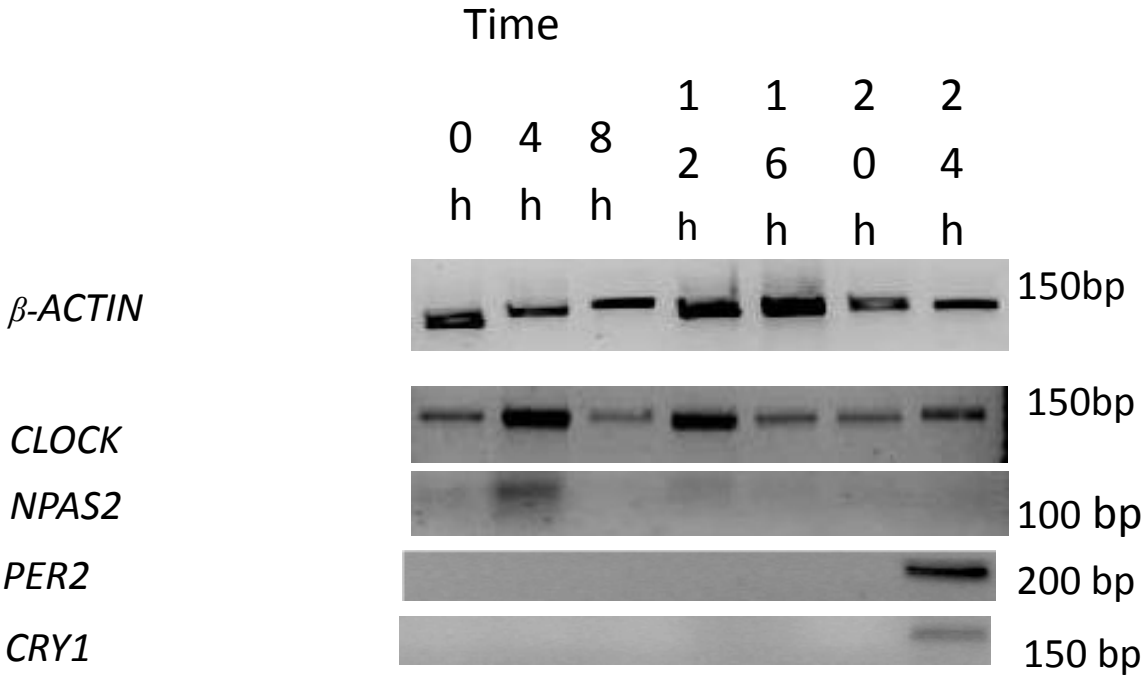


Figure 3.3: SH-SY5Y Cell line circadian clock expression profile. SH-SY5Y cells were grown to 80% confluency and incubated in zero serum media for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media (50% FBS). RNA was extracted at each of the time points and reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: CLOCK-F+CLOCK-R, B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R, PER2-F+PER2-R and CRY1-F+CRY1-R.

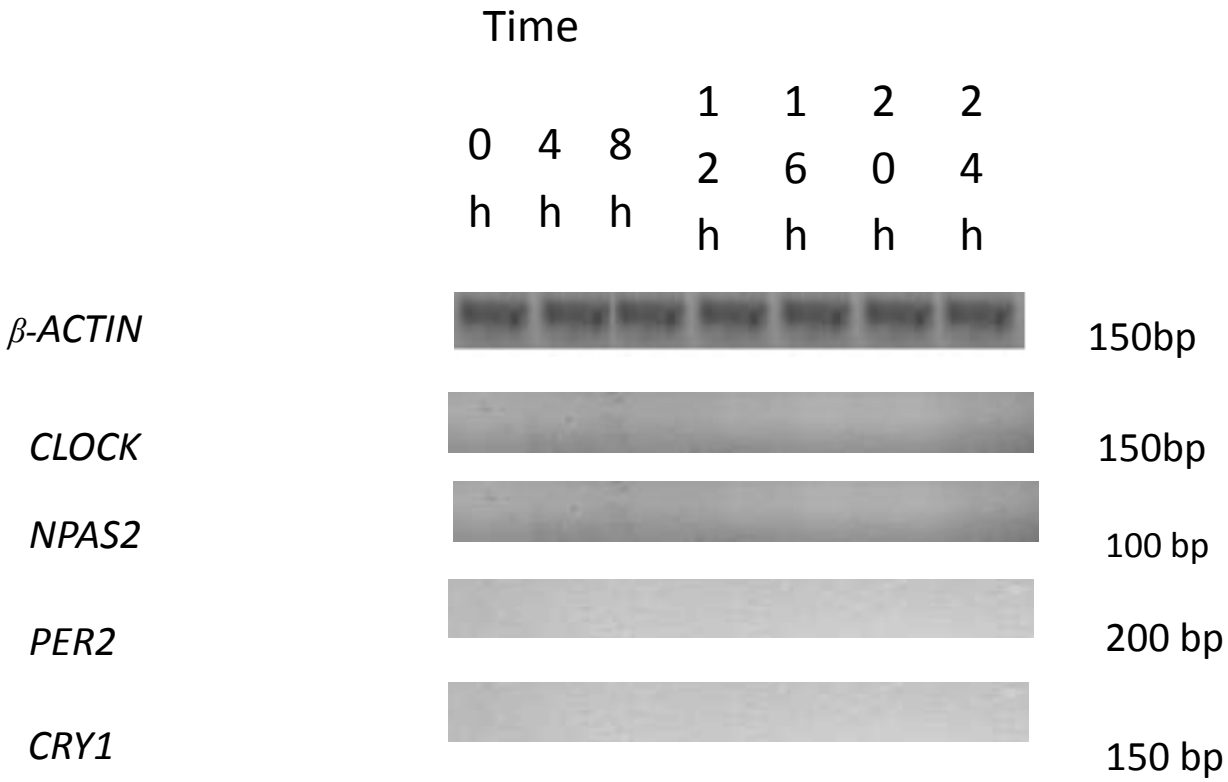


Figure 3.4: **Lymphoblastic cell line circadian clock expression profile. It was not possible to induce circadian clock expression in HI2162 lymphoblastic cell lines. HI2162 cells were grown to 80% confluency and incubated in zero serum media for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media**

(50% FBS). RNA was extracted at each of the time points and reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: CLOCK-F+CLOCK-R, B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R, PER2-F+PER2-R and CRY1-F+CRY1-R. The genes were present within genomic DNA as it was possible to clone them using standard PCR but were not transcribed into mRNA.

3.3 High temperature induces expression of circadian clock.

Serum shock is not the only mechanism of inducing circadian expression. Several other techniques are described within the literature including heat. The involvement of heat in regulation of circadian clock oscillation has been established as far back as 1996 (Rensing & Monnerjahn, 1996) who noted the circadian rhythmicity of heat shock proteins within the prokaryote *Synechocystis*. Although no evidence was reported of a similar effect of heat in human cells, it is well established that biochemical reactions respond to temperature to compensate their kinetics (Brown and Webb, 1948).

As the core human body temperature is 37^0C, any temperature above can induce a heat shock response (Abravaya, Phillips and Morimoto, 1991). For this experiment cells were incubated for 1 hour at 42^0C prior to restoration to 37^0C . Total RNA was harvested at the specified time points upon return to 37^0C..

Fig.3.5 below demonstrates the potential role of heat in inducing the expression of *NPAS2* and *PER2*.

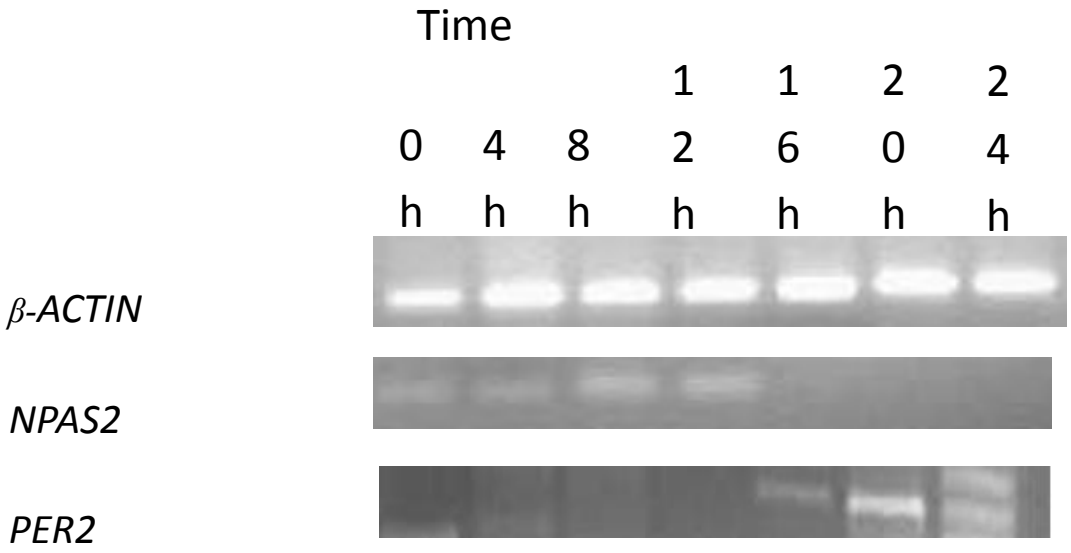


Figure 3.5: Heat induction of circadian clock within HeLa cells. For this experiment HeLa cells grown to 80% confluency and were incubated at 42^0C for 1 hour before RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R.

If we compare the results in this figure with those in Fig. 3.1 (asynchronous), we see an induction in the circadian clock genes. Of note is the appearance of two bands for *PER2* at 24h post induction. Recent work within the group has identified several smaller circadian clock protein isoforms that are expressed in response to stress.

Several core circadian clock genes (*PER2*, *CLOCK* and *BMAL1*) have been demonstrated to have heat shock element consensus sequences within their promoter regions with the *PER2* sequence being directly proximal to the E-box transcription factor sequence (Tamaru *et al*,

2011). Given the presence of a heat shock element in the 5' UTR promoter, this seems to also be the case for *NPAS2* as Fig3.6 below demonstrates.

chr2:101,396,546-101,397,485

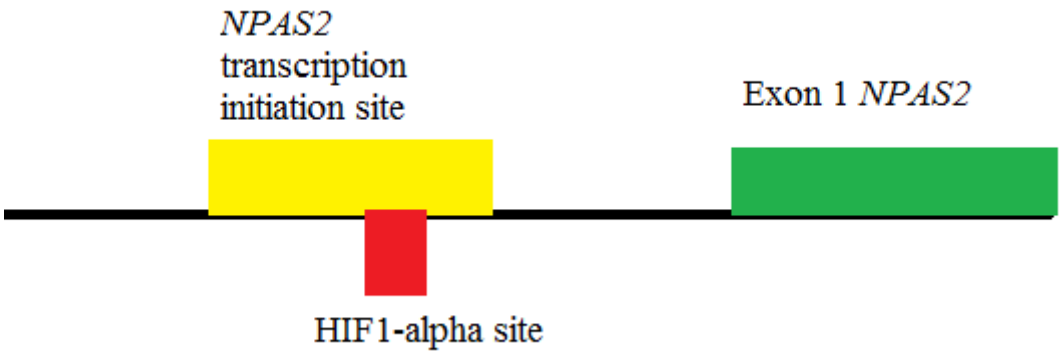


Figure 3.6: Map of the 5' upstream promoter region of the circadian clock gene *NPAS2*. The region mapping to chr2:101,397,288-101,397,543 (256bp) has the canonical sequence of a heat shock element (nnCnnGAAnnTCCn) which will allow the binding of HIF-1, a heat inducible transcription factor.

The finding that heat can stimulate the expression of *NPAS2* is of importance as it is known that NPAS2 has a role in DNA damage repair whilst heat has been demonstrated to cause DNA damage *in vivo* (Hoffman *et al*, 2008; Purschke *et al*, 2010).

3.4 Low temperatures do not induce circadian expression

It has been noted within the literature that low temperatures can dampen the oscillation of *in situ* circadian clocks (Hopfer & Sunderman, 1988). This phenomenon is much more pronounced within ectothermal animals such as the ruin lizard (*Podarcis sicula*) that are dependent on outside temperatures to regulate their core body temperature (Magnone *et al*, 2005). Endothermic animals that can self-regulate body temperature are able to limit the

impact of this phenomenon, but if a rat becomes hypothermic it can take upwards of 80h for normal rhythms to be restored (Hopfer & Sunderman, 1988).

For this experiment cells were incubated for 1 hour at 30^0C prior to restoration to 37^0C. Total RNA was harvested at the specified time points upon return to 370C.

In contrast to the findings of Fig.3.5 above, it does not seem that low temperatures can induce expression of certain circadian clock genes (Fig.3.7).

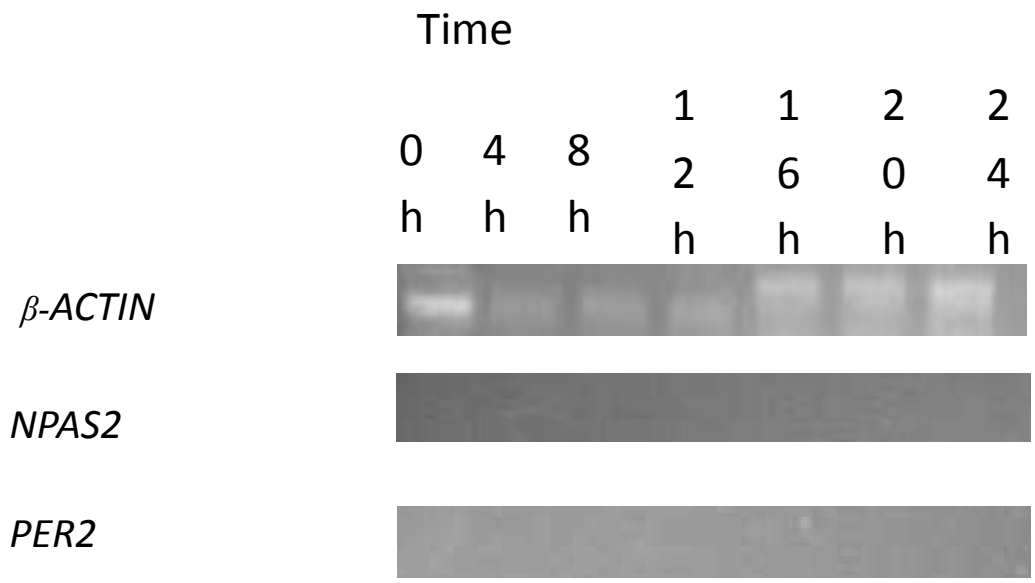


Figure 3.7: Influence of low temperature (30 degrees Celsius) on circadian clock expression within HeLa cell lines. HeLa cells were grown to 80% confluency and incubated at 30^0C for 1 hour. After incubation total RNA was extracted and reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R. The above figure was produced using data from HeLa cells as they proved more resilient to the prolonged exposure to low temperature required. Whilst the genes were not expressed there does seem to be some down-regulation of their expression.

Having explored the effects of temperature and nutrition on the clock it was possible to investigate the influence of other forms of stress on induction.

3.5 DNA damage induces *NPAS2* transcription.

A specific mutation within *NPAS2* at position 394 in which threonine replaces alanine (A394T) significantly increases the risk of developing non-Hodgkins lymphoma (Zhu *et al*, 2007). The polymorphism causes a structural change which causes improper DNA binding leading to immune dysregulation. In a further paper the role of *NPAS2* in the DNA damage repair pathway is identified (Hoffman et al 2008).

MCF-7 cell lines were treated with siRNA against the *NPAS2* and then exposed to mutagenic chemicals. In comparison with a control group, there was a significant variation in cell populations at various stages in the cell cycle upon down regulation of *NPAS2* as opposed to normal cells indicating a checkpoint defect. Hoffman *et al* (2008) also discovered that not only does *NPAS2* have a role in cell cycle checkpoints, but that it is also involved in DNA damage repair. A comet assay, which detects broken DNA, revealed an increase in damaged DNA in cells lacking *NPAS2* upon treatment with a mutagen (Hoffman *et al*, 2008).

To replicate the findings HeLa cells were treated with camptothecin and gemcitabine before RNA extraction. The cytotoxic agents were incorporated into the medium and cells incubated until removed at the stated time points for RNA extraction.

Figures 3.8 and 3.9 demonstrate how DNA damage can induce expression of two circadian clock genes within cultured cells.

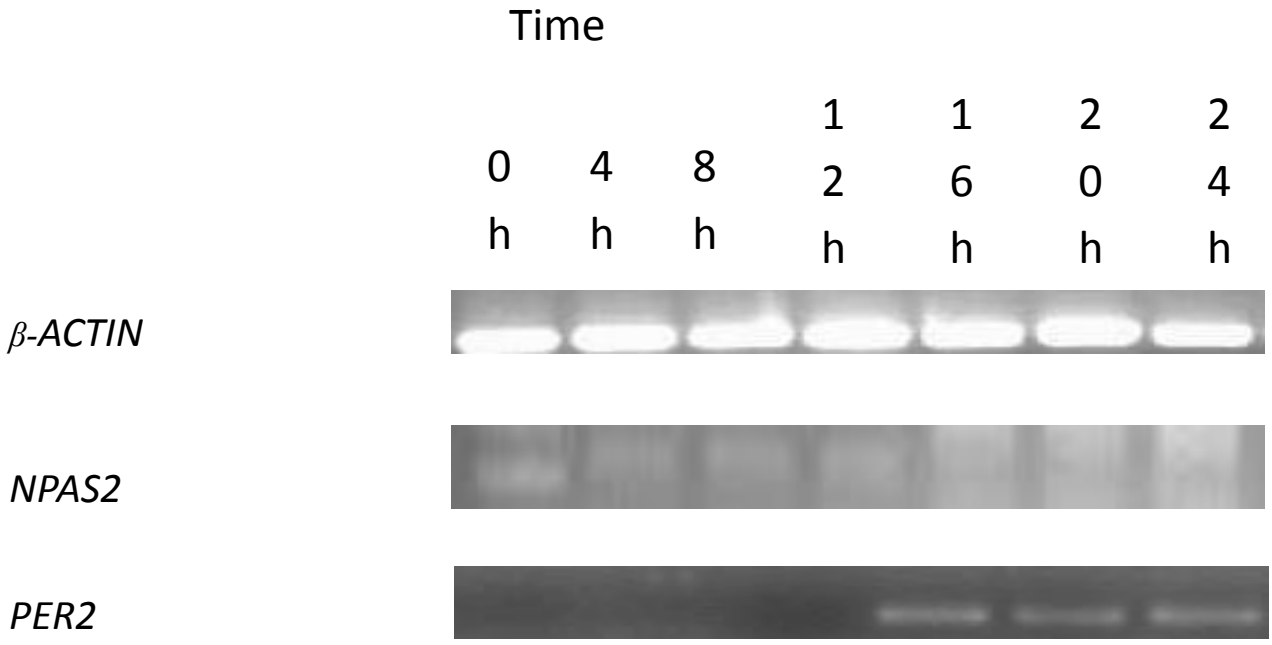


Figure 3.8: Treatment of HeLa cells with 4uM camptothecin (CPT) until RNA extraction. 4uM of CPT was included within the standard media of 80% confluent HeLa cells. At the stated time points the media was removed, the cells rinsed with PBS and the total RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R. After incubation with CPT *PER2* expression is induced by 16h. This is not matched by a decrease in *NPAS2* expression which is expected given the increase in *PER2*. *NPAS2* is induced within 4h and remains expressed over the 24h investigated. Indeed multiple bands of *NPAS2* appear in the PCR which are visible in the figure. These inducible isoforms of *NPAS2* and *PER2* are consistent with other findings within the group.

Camptothecin is a potent agent of DNA damage and acts by stabilizing the DNA-topoisomerase I complex (Efferth *et al*, 2007). This stable complex forms an obstacle for advancing replication forks which break when they collide with the immobilised complex (Pommier, 2006)

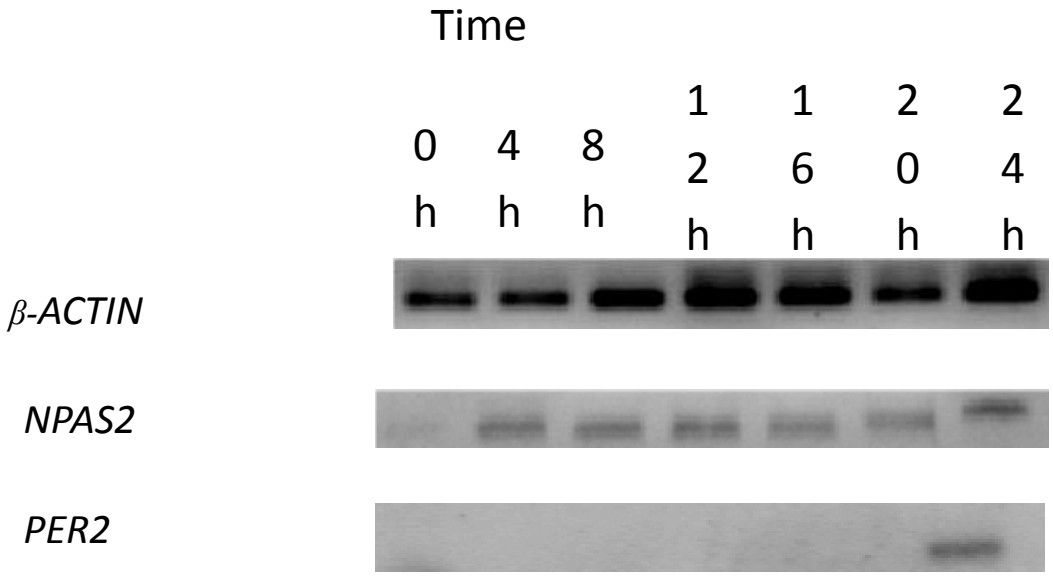


Figure 3.9: Treatment of HeLa cells with 100nM Gemcitabine until RNA extraction. 100nM of Gemcitabine was included within the standard media of 80% confluent HeLa cells. At the stated time points the media was removed, the cells rinsed with PBS and the total RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R. A similar result to that of CPT above (Fig, 3.8) including the expression of *NPAS2* isoforms.

Gemcitabine is a nucleoside analogue which impacts upon DNA synthesis via its recognition by DNA polymerase as a cytidine nucleotide. Once incorporated however no other nucleotide

can be attached thus terminating DNA synthesis and inducing cell death in cells which are deficient for repair (Cerqueira, Fernandes and Ramos, 2007).

NPAS2 has been implicated with the DNA damage response (Hoffman *et al* 2008). The above two figures demonstrate that *NPAS2* is induced in response to DNA damage and is maintained whilst *PER2* expression is also induced.

3.6 siRNA attenuates *NPAS2* expression.

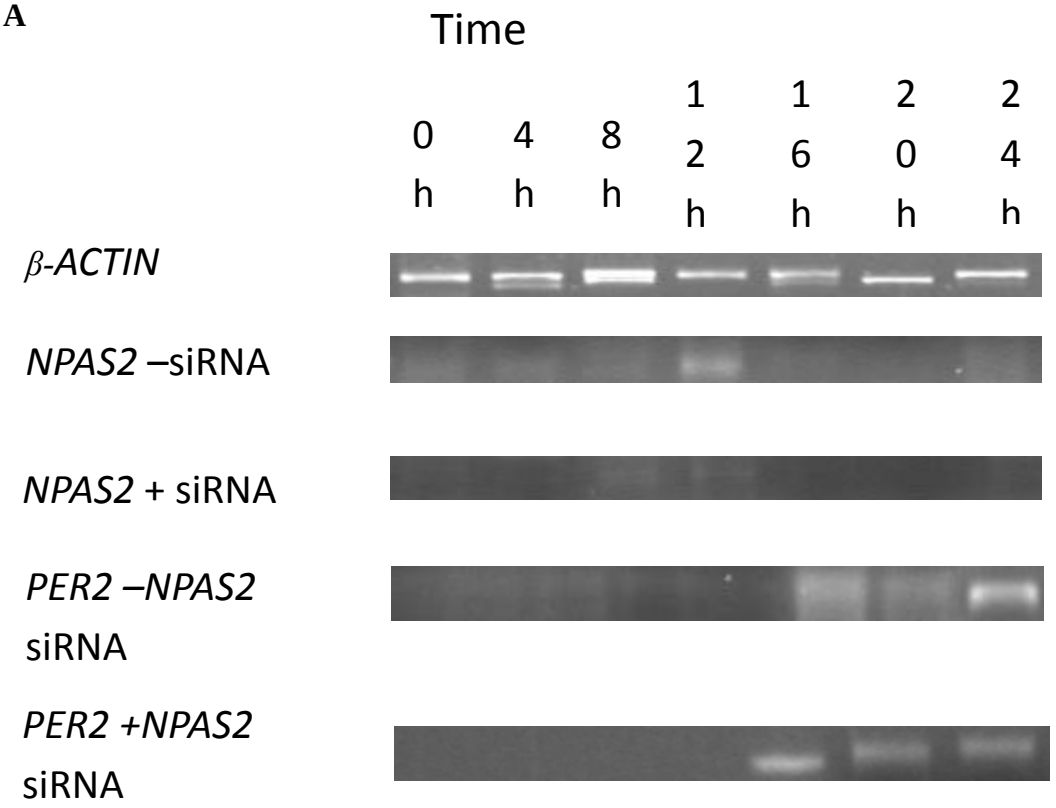
siRNA are artificially generated interfering RNA duplexes which have been designed to complimentary bind to regions of a gene's mRNA. Once bound the siRNA then allows for translational repression and degradation of mRNA transcript (Bartlett and Davis, 2006).

Cells will be treated with a siRNA before being serum shocked to synchronise its circadian clock gene expression.

HeLa cells were subjected to serum shock for synchronous induction of circadian rhythm genes. In order to produce the siRNA molecule primers NPAS2-F and NPAS2-R were used to clone a region complementary to the mRNA, this sequence was then cloned into a siRNA expressing plasmid called psiRNA. The plasmid was introduced into HeLa using LyoVec chemical transfection. Prior to transfection 2x106 cells were seeded onto a plate and the LyoVec-vector conjugate introduced into the media.

Fig 3.10 below demonstrates that silencing the expression of *NPAS2* using siRNA does not attenuate the expression of *PER2.*

A



B

	Time (h)						
Gene	0	4	8	12	16	20	24
β-*Actin*	2550.145	2510.24	2773.991	2507.564	2557.347	2607.498	2672.241
NPAS2 -siRNA	149.146	198.154	304.414	606.561	0	0	0
NPAS2 +siRNA	0	0	79.588	81.158	0	0	0
PER2 -NPAS2 siRNA	0	0	0	0	603.549	678.354	1057.458
PER2 +NPAS2 siRNA	0	0	0	0	780.158	705.157	997.487

Figure 3.10: siRNA treatment of HeLa. 24h after transfection with siRNA. In the plus siRNA samples *NPAS2* expression is attenuated but with no apparent impact on *PER2* expression. This implies that normal circadian function is continued even with only minimal *NPAS2*. Panel B demonstrates the relative intensity of each PCR band. In the

absence of quantitative data, ImageJ was used to provide data on the relative amounts of product in each gel based on the intensity of each band when illuminated with UV.

The siRNA used was not wholly effective in down-regulating *NPAS2* completely but there was a decrease in level. The minimum levels of *NPAS2* however were sufficient, in conjunction with *CLOCK* to facilitate a circadian clock.

As several non-human cell lines will be used in the course of the work it is important to understand the clock within them.

3.7 Ascertaining *NPAS2* and *PER2* expression in other vertebrate models.

As covered in the introduction, the circadian clock is ubiquitous in life. It should be possible to induce the expression of the circadian clock in a similar fashion. There are three species represented in this set of experiments: mouse, chimpanzee and chicken. Each of the three cell lines will be subjected to serum shock as in the human cells and their RNA extracted and reverse transcribed.

Figures 3.11, 3.12 and 3.13 demonstrate attempts at inducing circadian cycle synchronicity in EB176, MEF3T3 and DT40 cell lines respectively.

The EB176JC cell line is a lymphoblastic cell line extracted from a chimpanzee which has been transformed by the Epstein-Barr virus. Cells were grown to 80% confluency and incubated in zero serum media for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media (50% FBS).At the stated time points the media was removed, the cells rinsed with PBS and the total RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R.

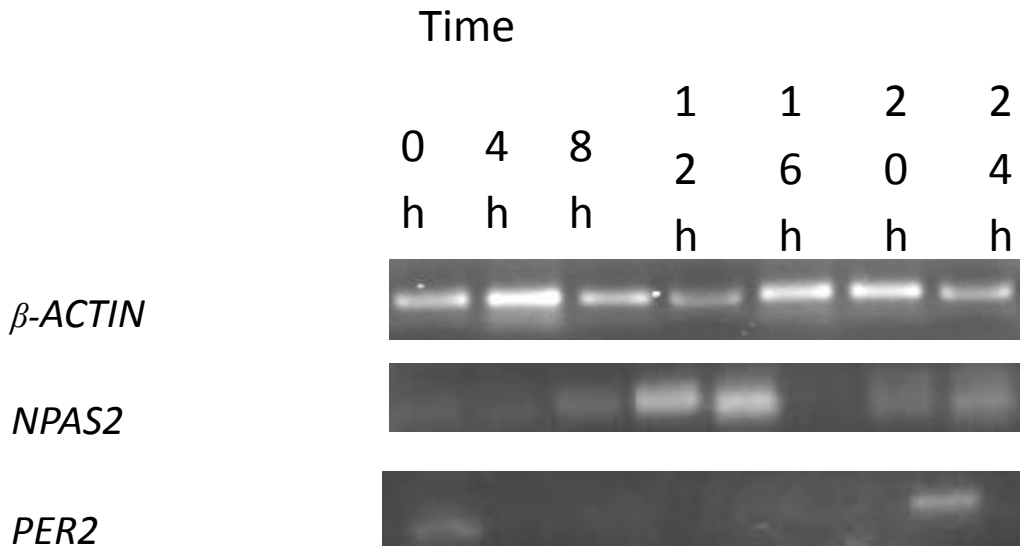


Figure 3.11: Serum Shock of EB176JC (*Pan troglodytes)* cell line. The circadian clock is inducible as in human cells with the exception that levels of *NPAS2* expression persist for longer.

MEF 3T3 is a fibroblastic cell line extracted from a mouse embryo. Cells were grown to 80% confluency and incubated in zero serum media for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media (50% FBS).At the stated time points the media was removed, the cells rinsed with PBS and the total RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, NPAS2-F+NPAS2-R and PER2-F+PER2-R.

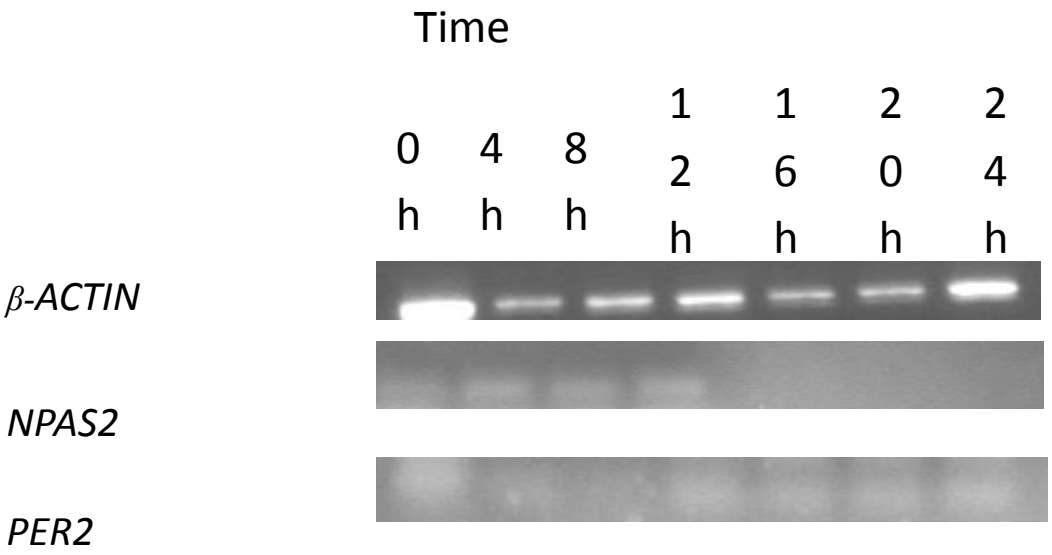


Figure 3.12: Serum shock of MEF 3T3 (*Mus musculus)* cell line. *NPAS2* is inducible and expression persists for 12h. *PER2* however seems constitutively expressed at low levels. This could be an artefact of the peculiar nature of the cells such as high levels of chromosome instability (Todaro & Green, 1963). Conceivably this could have placed *PER2* under a different promoter.

DT40 cells are b-lymphoblastic cells extracted from the White-leghorn species of chicken. DT40 demonstrates substantial in vivo genetic recombination and is therefore a model vertebrate system. Cells were grown to 80% confluency and incubated in zero serum media (including no chicken serum) for 12h prior to the experiment. At point 0h the serum free media was replaced with serum rich media (50% FBS).At the stated time points the media was removed, the cells rinsed with PBS and the total RNA was extracted. Once extracted the RNA was reverse transcribed using random hexamer primers. The cDNA was then subjected to PCR using the following primers: B-ACTIN-F+B-ACTIN-R, gNPAS2-F+gNPAS2-R and gCLOCK-F+gCLOCK-R.

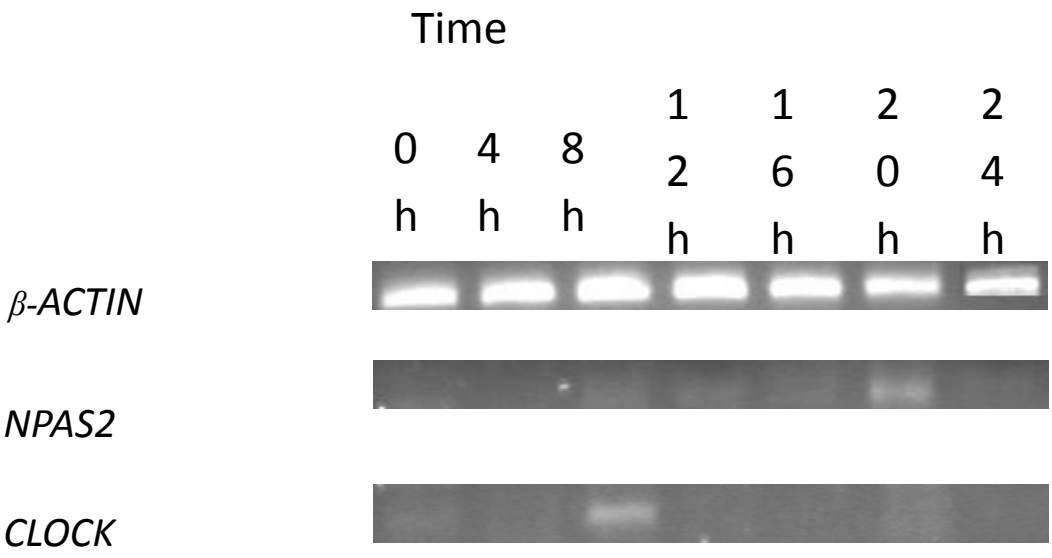


Figure 3.13: Serum shock of DT40 (*Gallus gallus*) cell line. Unfortunately the sequence for *PER2* or *CRY1* had not been deduced for chicken at the time of the experiment and only *NPAS2* and *CLOCK* were utilised. The cell line however proved impossible to induce.

Of the three non-human vertebrate cell lines tested it proved relatively easy to induce circadian rhythmicity within the mammalian cell lines. The DT40 clones however proved recalcitrant to serum shocking (Fig. 3.14). DT40 cell lines were tested from three different sources and no clone expressed rhythmicity via serum shocking.

4 Identification of potential novel microRNA cluster within intron 1 of *NPAS2*.

Initially it was thought that introns had no function, hence the sobriquet "junk DNA". Attitudes have changed since several useful functions for introns have been discovered, including as hosts for microRNA (Ambros *et al*, 2003). Initially it was understood that intronic miRNA shared the same promoter as the host gene (Baskerville and Bartel, 2005); however more recent work has established that ~35% of all intronic miRNA have their own promoter regions (Monteys *et al*, 2010). These promoter regions have much in common with regular promoter regions including transcription factor binding regions, RNApol II recognition sites and DNA methylation sites (Monteys *et al*, 2010).

The aim of this chapter is to correlate the transcription of the host gene with the expression profile of the novel miRNA identified *in silico* by B.Nicholas *et al* (2008). The same author also described the appearance ~3.2kb upstream of the rs1811399 of a sequence which resembled a novel member of the miR-1273 family (Figure 4.2). Fig4.1 below demonstrates the relative locations of both proposed miRNA hairpins.

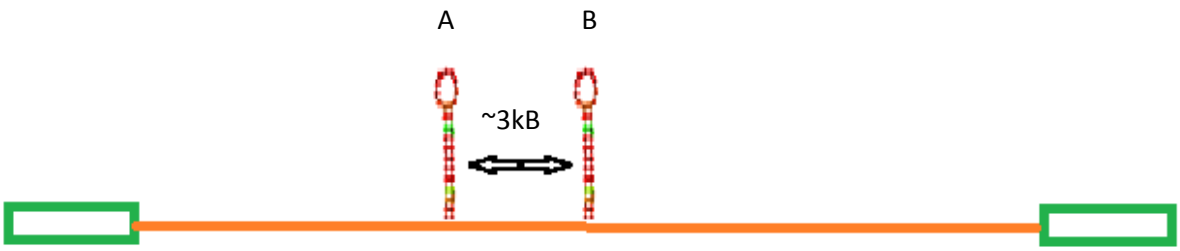


Figure 4.1: Schematic representation of *NPAS2* intron 1. The intron (orange line) contains at least two detected novel miRNA. The haipin labelled A represents the novel

miR-1273 whilst hairpin B is that of rs1811399. Both these hairpins are approximately 3Kb away from each other.

4.1 Identification of novel miRNA within intron 1 of *NPAS2*

The miR-1273 family was first detected by a mass screening of human embryonic stem cells (Morin *et al*, 2008). There are 7 separate mature miRNA within the miR-1273 family: miR-1273a, miR-1273d, miR-1273e, miR-1273f, miR-1273g, miR-4430 and miR-4459.

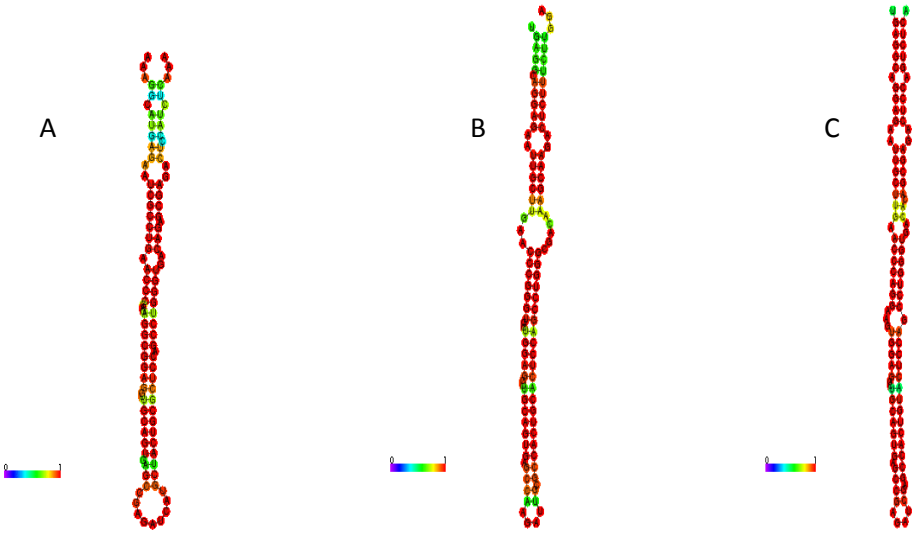


Figure 4.2: The DNA in region chr2:101477345-101477628 is predicted to form a miRNA hairpin similar to the miR1273 family. This figure demonstrates the predicted RNA folding structures of A) novel miR 1273 clone, B) hsa-miR 1273 and C) hsa-miR1273-e. This figure was produced using Vienna RNAfold. Scoring for each nucleotide (as denoted by the colour scheme) is based on the probability of each nucleotide base-pairing at this location as calculated using the McCaskill algorithm (McCaskill 2004), a score of 1 is more likely to base-pair with its opposite nucleotide.

miR 1273a is found on chromosome 8 while miR-1273e is located on chromosome 17 based on a BLAT search.

Separate miRNA can be said to be a part of a specific miRNA family based on its sequence. Each member of a miRNA family has high sequence homology with each other as demonstrated in Fig.4.3.

A

```
hsa-miR-1273e       UUGCUUGAACCCA---GG---AAGU-GGA-- 22
hsa-miR-1273d       ------GAACCCAUGAGGUUGAGGCUGCAGU 25
hsa-miR-1273        -GGG--CGACAAAGCAAG---ACUCUUUCUU 25
hsa-miR-1273c       --GG--CGACAAAACGAG---ACCCUGUC-- 22
novel-var           ------AGGCAUGAGAAU---CGCCUGAACC 22
                          ..*. .   .     .      .
```

B

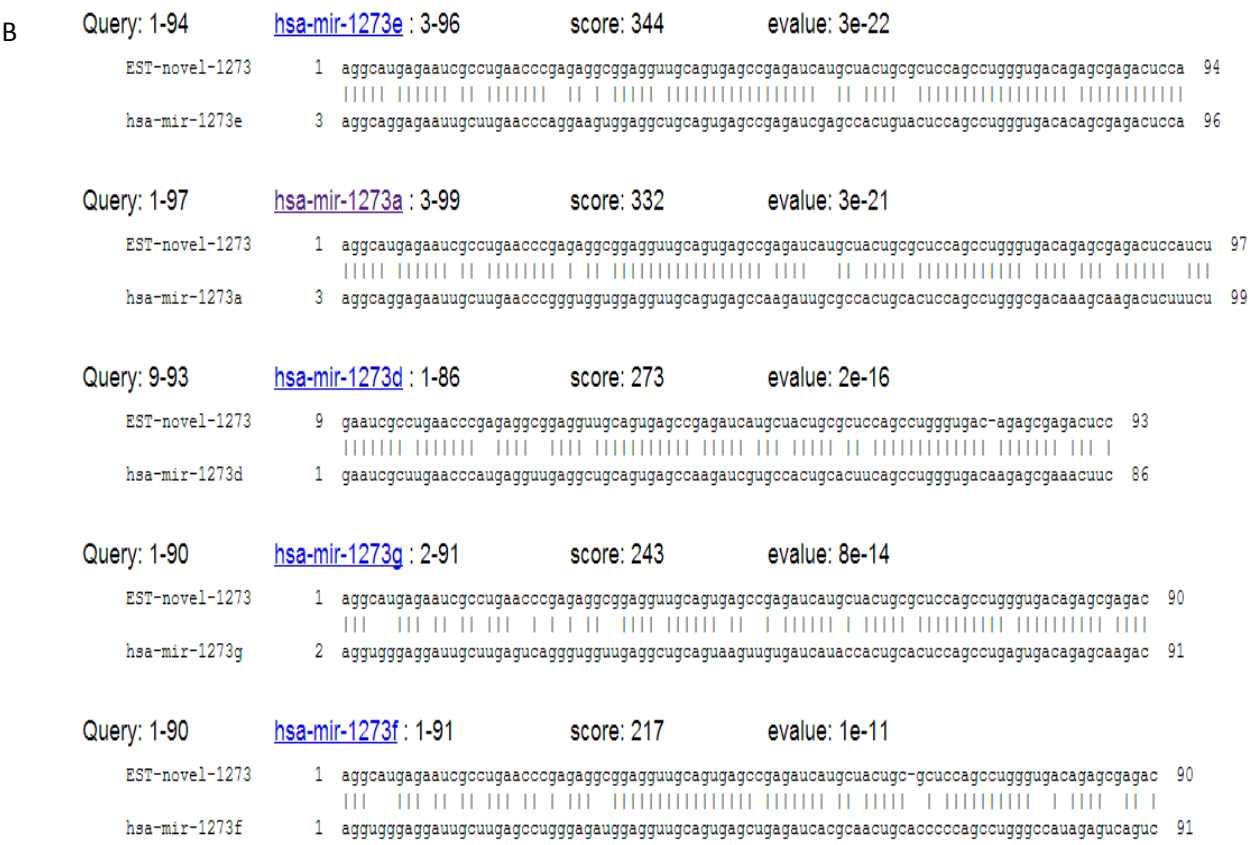


Figure 4.3: Alignment of known miR1273 family members against novel variant. A) Seed sequences are first 5-8 nt on 5' side. As evident, sequence conformity within the

mature form is low within the miR1273 family. This is in line with our understanding of miRNA evolution of sequence duplication and then diversification. Period and asterisks under the blat alignment denote locations of conserved nucleotides (asterisks) or locations containing nucleotides of broadly similar properties (periods). B) This series of alignments demonstrates the sequence homology between the hairpin sequences of all miR1273 homologue and the sequence of the novel homologue. The expect value (E-value) for each alignment is also given. The E-value is a parameter that describes the number of results that would appear by chance when searching the human genome database. It decreases exponentially the as the confidence within the search increases, so the lower the E-value, or the closer it is to zero, the more significant the match is.

The novel variant of miR-1273 identified above has a high E-value which identifies it as part of the family.

The appearance of two miRNA (novel miR-1273 and the rs1811399 miRNA) so close to each other implies (Fig.4.1) the presence of a cluster that either requires *NPAS2* for its spatiotemporal expression or has simply evolved within the locus by chance.

Tools exist to allow the genome to be parsed and analysed for hairpin loops which could form putative miRNAs, one such tool is the SSCprofiler (Oulas *et al*, 2009).

By entering a set of co-ordinates centred on rs1811399 and setting the HMM score to 3 (scores of 1-3 increase sensitivity) we are provided with two potential results (Figure 4.4). A HMM score as shown in Fig.4.4 below provides a sensitivity of 85.96% and a specificity of 88.02% for the proposed hairpins.

chr	nucStart	nucEnd	strand	Max_Expression	Max_Exp_location	HMMScore	Cell_Line	2ry Structure
2	100937492	100937595	top_strand	7.0	100937508	5.1	HeLa	2ry Structure
2	100937865	100937968	top_strand	9.5	100937935	5.0	HeLa	2ry Structure

Figure 4.4: SSCProfiler results for 1kb region of genomic DNA centred on rs1811399. Note, co-ordinates within this figure vary to those given in text due to SSCProfiler operating on an older version of the UCSC database (hg17). This thesis is written with Ensembl v71 or UCSC hg19, which are now synchronised with regards to nomenclature.

The first putative miRNA identified is the rs18113999 miRNA. It is given the genomic co-ordinates chr2:101,478,974-101,479,077 and is shown to have the genomic sequence of:

AGAAGGCTGTGGTCAGGTCTGGAGGTCAGGGCATGGTGAT**A>C**CAGCGGCTGCC

TGACAGTCACTGCCCAGAGCTTCCCTTACCATAACCTTCCTCAGTAGACTAG

The second miRNA predicted to lie within the 1kb region on the positive strand is given the genomic coordinates chr2:101,479,337-101,479,443; it is predicted to lay 270 base pairs downstream of the rs1811399 miRNA. The sequence provided is as follows:

ACGTATGTGAGTGACCCTCTTCTTACCCTGCCTTTTGGCAGCATTAATAACAGCT

ATGTATGGA**R**(G>A)CTGTTAGTGAAATGGTGTTGTCGGGTCAGGAAGGAGTAAG

CT

Of note is the presence of a SNP within this proposed miRNA (rs74734518). The G allele is predominant (98%) whilst there are no recorded homozygous A, only heterozygous G|A (2% according to Ensembl). This would imply that the A allele is deleterious to some extent, or a recent mutation. The impact of rs74734518 on predicted RNA folding does not produce a dramatic de-stabilisation of the hairpin loop similar to that created by rs1811399 (Figure 4.5).

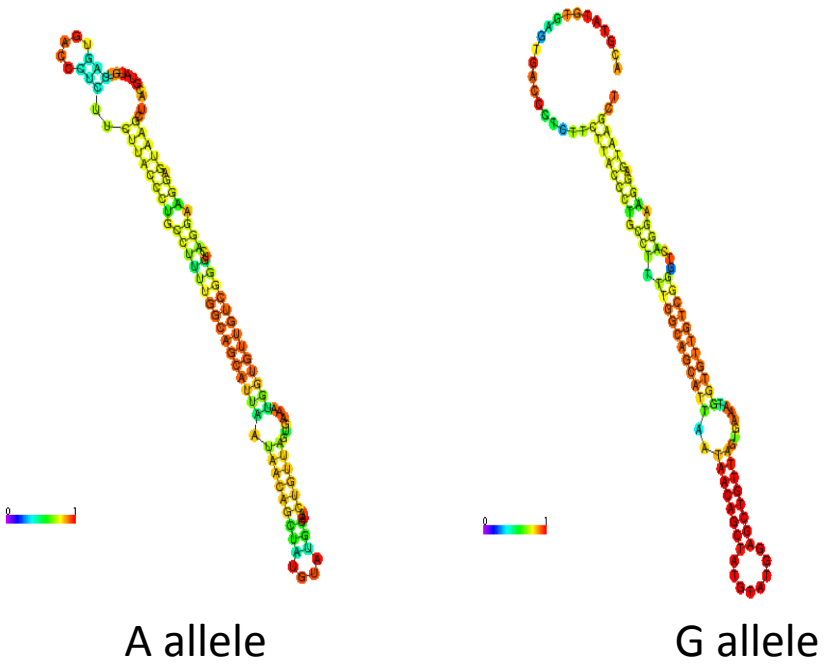


Figure 4.5: Hairpin loop from chr2: 100937865-100937968 which is predicted by SSCProfiler. The locus has a SNP (rs74734518) A>G with the following effects on the hairpin. This figure was produced using Vienna RNAfold.

If we take the regulatory region upstream of the novel miR1273 clone as the defining edge of our presumptive miRNA cluster (chr2:101,473,826-101,484,225) and work downstream in 1kb sections, using both SSCProfiler and a homology search (BLASTing against known miRNA sequences), we arrive at a total of at least five predicted miRNA precursors. These are:

- chr2:101,475,627-101,475,726 (novel miR-1273)

  AGGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGAG

  ATCATGCTACTGCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTCAA

- chr2:101,475,817-101,475,910

  CCTTTATTCCAGAAAATATGCTTCAGCCCTGGGAATTGAAAGTGAGGAAA

  ACAAGTCAAACCCAGAGCTCATAGAATAGTGGGATAGATGGGCA

- chr2:101,477,368-101,477,460 (a second novel miR-1273 family gene located within EST T59368)

  GTTTTGAGACAGGGTCTTGCTCTTTTGCCCAAGCTGGAGTACAGTGGCTC
  ATTGCAGCCTGGAACTCCAGGGCTCAAGCGATCCTCTCACCTC
- chr2:101,478,974-101,479,077 (rs1811399 containing hairpin)

  AGAAGGCTGTGGTCAGGTCTGGAGGTCAGGGCATGGTGATCCAGCGGCTG
  CCTGACAGTCACTGCCCAGAGCTTCCCTTACCATAACCTTCCTCAGTAGAC
  TAG
- chr2:101,479,347-101,479,450

  AGGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGA
  GATCATGCTACTGCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTCA
  A

Of these all have correct hairpin structures (Figure 4.6) and all have practicable DROSHA processing sites (Figure 4.7)

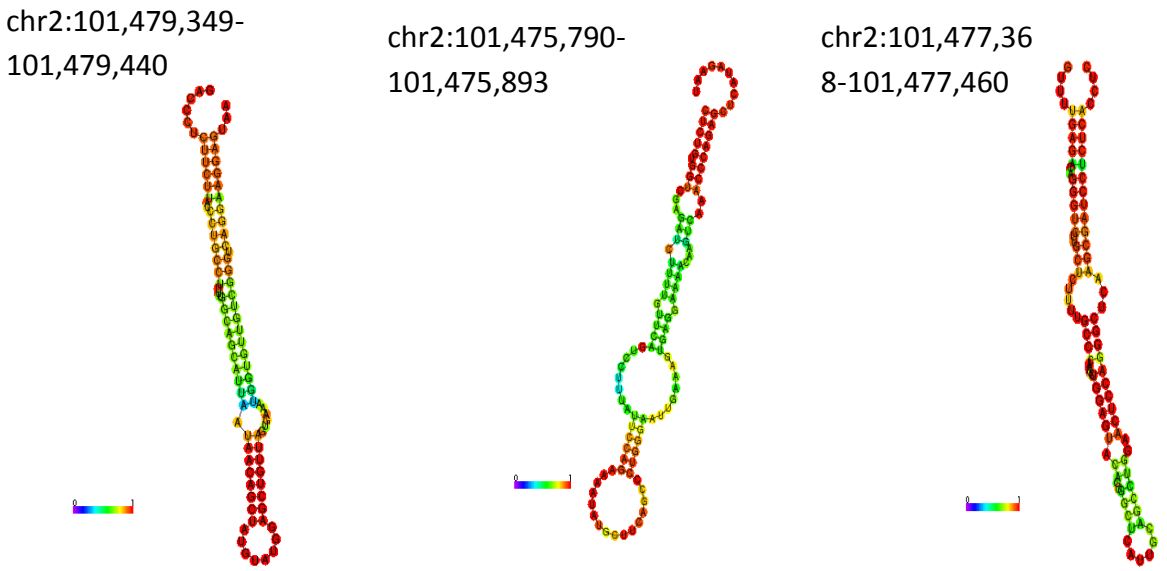
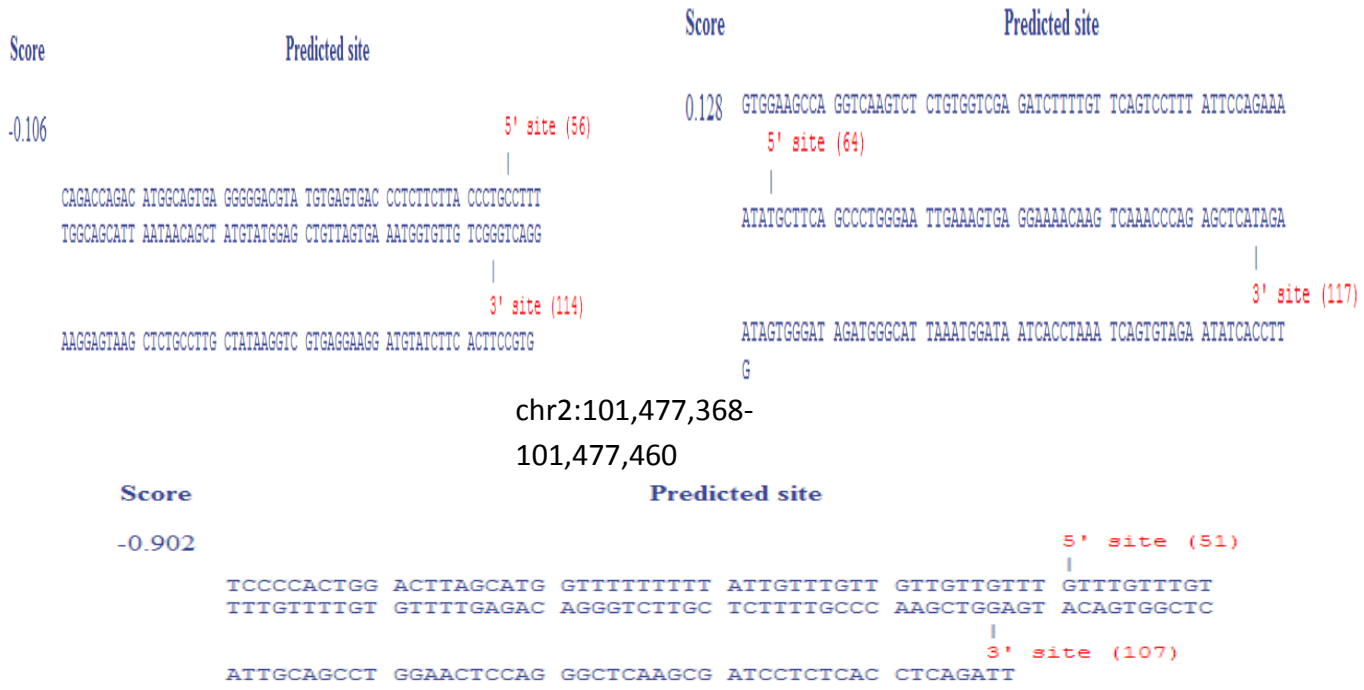


Figure 4.6: Hairpin loop structures of three previously un-described miRNA precursors located within intron 1 of *NPAS2*. Each has the characteristic terminal loop required for processing. This figure was produced using Vienna RNAfold.

chr2:101,479,349-101,479,440

chr2:101,475,790-101,475,893



chr2:101,477,368-101,477,460

Figure 4.7: DROSHA cutting site prediction. Using support vector machinery algorithms (Helvik *et al*, 2007), it is possible to calculate where DROSHA will cut on the hairpin. The prediction is based on distance from terminal loop and from the basal level.

The fact that each of these sequences forms a hairpin, has a predicted DROSHA cutting site and has been calculated (via SSCProfiler) as hosting probable miRNA strengthens the case for a miRNA cluster. It is therefore important to experimentally validate as many of them as possible. However, caution must be used with reference to the third hairpin (chr2:101,477,368-101,477,460) given the low PPV associated with the result. The PPV for the rs1811399 hairpin was described as 0.38 in Nicholas *et al* (2008).

4.2 Expression of precursor is not tissue specific

It is known that several miRNA are expressed in a tissue specific manner and can play a key developmental and functional role within that tissue (Wen & Friedman, 2012). miR-122 is an example of such as miRNA and so far has only been isolated from hepatic tissue, where it has a role in fatty acid metabolism (Wen & Friedman, 2012).

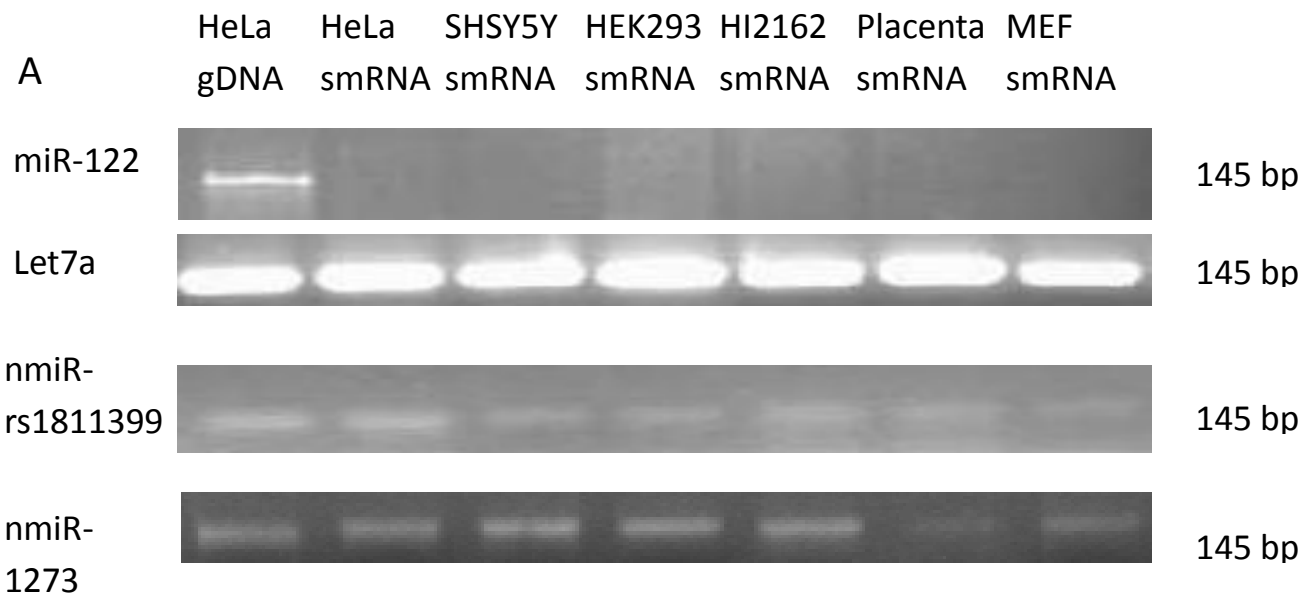
A second example of a tissue specific miRNA would be miR-9. miR-9 has been demonstrated to be an important neuron determinant. Shibata *et al* (2008) expands on this by describing the homology of miR-9 between invertebrates and vertebrates as 100%. It is preferentially expressed in the mammalian cortex, where it has been shown to regulate the FOXG protein, which is a potent repressor of neural cell differentiation (Shibata *et al*, 2008). FOXG protein is essential for repressing the formation of Cajal–Retzius cells (Shibata *et al*, 2008), which are amongst the earliest neural cell to mature and is responsible for providing the REELIN network for further cell migration and maturation (Bar, Lambert & Goffinet 2000).

Having established a precedent for the tissue specificity of microRNA, it was only logical to assess whether our two novel miRNA; nmiR-1273 and nmiR-rs1811399 exhibited similar pattern of specificity. To assess this, a selection of cell lines were grown to confluency and the small RNA pool was extracted via column elution. This small RNA fraction was then reverse transcribed to produce cDNA which was then subjected to PCR using hairpin specific primers.

To assess for genomic DNA contamination a test PCR using miR-122 (which is only expressed within the liver) primers was carried out on tissue RNA and one genomic DNA as a positive control. For the reasons specified in the above a positive band was found only in the genomic DNA pool due to its lack of transcription within tissues other than liver. We can

conclude that genomic DNA contamination at the sensitivity of this assay is minimal. As a positive test of cDNA pool Let7a was chosen. Let7a is ubiquitously expressed in human tissues and as such provides an excellent positive control to test for a positive reverse transcription reaction.

A



B

	Cell line						
Gene	HeLa gDNA	HeLa smRNA	SHYS5Y	HEK293	HI2162	Placenta	MEF
Let7a	13000.911	12509.11	12505.24	12059.53	12105.26	12001.97	12435.36
rs1811399	2087.285	2175.65	2009.788	2078.203	2080.007	1999.711	1980.083
nmIR1273	2103.215	2150.017	2127.398	2199.077	2203.179	1589.071	1998.278

Figure 4.8: As demonstrated by this PCR reaction the precursor of the novel rs1811399 and nmiR-1273 locus miRNA did not exhibit any tissue specificity. A) Each of the cell lines was grown to confluency prior to total RNA extraction. The total RNA pool was then enriched so that only small RNA molecules (<500nt) remained. The smRNA was then reverse transcribed using random hexamer primers. The primers used to test the smRNA cDNA pool were MIR122-F+MIR122-R, LET7-F+LET7-R, '399-F+'399-R and N1273-F+N1273-R. mIR-122 was used as a negative control due to its absence in most

tissue types whilst Let-7a was a positive control due to its ubiquitous expression. This follows an established pattern in the miR-1273 family of miRNA which have an extensive expression range across tissues. B) ImageJ semi-quantitative analysis of band intensity for miRNA precursor expression.

From Fig.4.8A and B, we can demonstrate that both microRNAs in intron 1 of *NPAS2* exhibit similar expression profiles with nmiR-1273 expressed weakly in the placental cell line.

4.3 Expression of mature forms has identical expression profile as precursor

It is not uncommon for miRNA to exhibit a universal expression for their precursor molecules but to have tissue specific maturation (Yu *et al*, 2012). The next process will be assessing the cells for the mature product.

Evidence for this tissue specific maturation is available in the case of miR-9 which as is expressed within neurones (Yu *et al*, 2012). Its precursor however is expressed within Schwann cells; the cells chiefly responsible for the production of myelin, but not its mature form. This implies a post-transcriptional regulation of the maturation miRNA precursors (Thomson *et al*, 2006).

A manner by which miRNA maturation is regulated is by the activity of RNA binding proteins such as hnRNP A1 which is known to play a role in up-regulating expression of miR-18 (Guil & Cáceres, 2007). miR-18 is a part of the miR-17-92 cluster of miRNA and should therefore be expected to be expressed at a similar level to the other miRNA within the cluster. This however is not the case (Guil & Cáceres, 2007). The hnRNP A1 protein has been shown to preferentially bind to a consensus region within the pre-miR-18 and to facilitate the recruitment of DROSHA protein, thereby promoting the maturation of the

precursor. The expression of miR-18 is therefore several times greater than that of the remainder of the cluster (Guil & Cáceres, 2007).

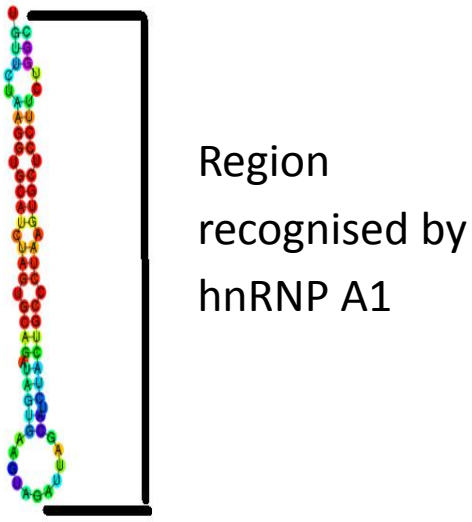


Figure 4.9: RNA hairpin of pre-miR-18. The area highlighted with black lines corresponds to the region of hairpin selectively associated with the hnRNP A1 protein. Not only does the protein encourage the recruitment of DROSHA but it might also have a role in preventing the formation of inhibitory secondary structures and also have a role in exporting the Drosha'ed fragment from the nucleus into the cytoplasm. This figure was produced using Vienna RNAfold.

RNA protection assays allow us to identify the presence of small single stranded RNA molecules by hybdrizing them to a radiolabelled complimentary RNA probe. The hybridized mix is then digested by an RNAse A and RNAse T1 mix (a mix is used to maximise RNA degredation) to remove all the single stranded RNA leaving only the duplexed RNA which is protected from cleavage.

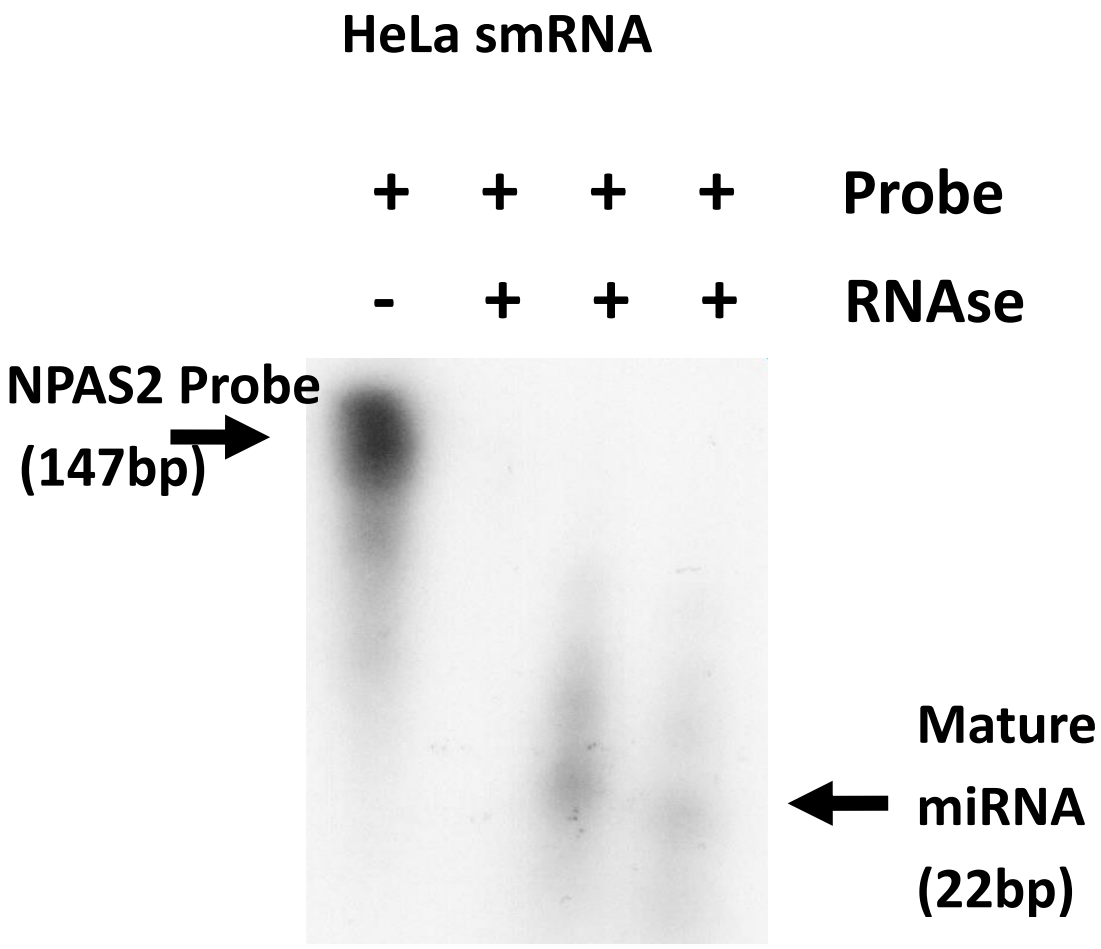


Figure 4.10: The above figure demonstrates a typical gel of a successful protection assay. Note the probe is significantly larger than the target miRNA and as such in a probe only control well runs much slower. To demonstrate its single stranded nature in the second well we have included the RNAse A/T1 mix which has completely digested the ssRNA probe. The third and fourth well contains smRNA extracted from HeLa which was duplexed with probe before enzymatic treatment.

The protection assay was conducted on asynchronous RNA extracted from 4 cell lines and 1 control RNA provided in the Ambion kit.

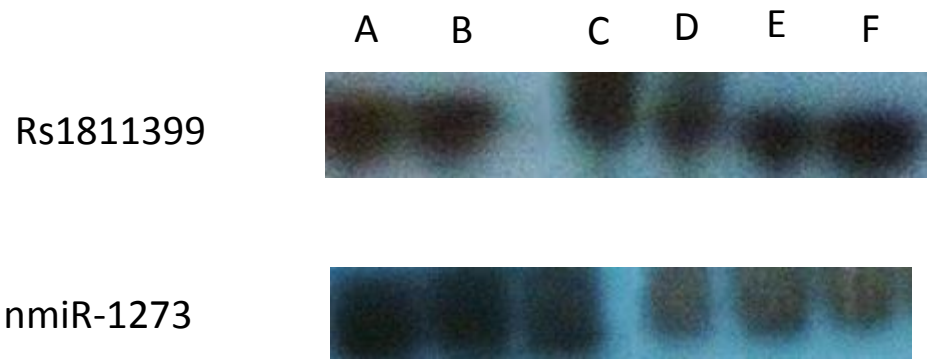


Figure 4.11: Autoradiograph of mature forms of two novel miRNA. Probes were produced using the primers '399-F+'399-R and N1273-F+N1273-R. A T7 promoter region was then ligated onto the DNA for transcription by T7 polymerase. Detection was made by the presence of radiolabelled cytidine within the RNA probe. The following asynchronous RNA pools were used: A) HeLa including miR-16 control (miR-16 is ubiquitously expressed; therefore probes complimentary to miR-16 can be used as a positive control) B) HeLa (Human cervical epithelium) C) HI2162 (human lymphoblastic) D) SH-SY5Y (human neuroblastoma) E) HEK293 (human embryonic kidney) F) Placental RNA.

As evident from these two autoradiograph films, there seems to be an identical expression pattern of the mature forms; implying that maturation of the miRNA is allowed to progress in all cell types investigated (Fig. 4.11). The functional implication of this on the miRNA is that the mature miRNA fulfil some essential housekeeping duty within human cells.

It is also possible from the data in Fig.4.11 to conclude that the product detected for rs1811399 miRNA and novel miR-1273 are the same size as a mature miRNA. This can be concluded from the fact that it runs at the same level as the miR-16 control which is designed to give a 22bp duplexed product.

The RNA used in this experiment was from asynchronous cells and the same RNA was used to conduct the experiment described in Figure 3.1. As the miRNA is expressed regardless of expression of its host gene, do factors which induce the host gene influence its expression?

4.4 Expression is constant regardless of exogenous factors.

Due to the implicit role miRNA play in regulating several functions of cellular behaviour such as activating responses to genotoxic stress, regulating response to energy levels and regulating cell division, one would expect several miRNA‘s expression to alternate dependant on the conditions prevalent in the local environment.

In Humans, heat and cold shock are demonstrated regulators of miRNA expression. miR-1, miR-21 and miR-24 are all up-regulated in response to heat shock whilst cold shock domain proteins have been demonstrated to remodel pre-miR let7g, thus regulating its expression (Yin, Wang & Kukreja, 2008: Mayr *et al*, 2012). The mechanism for this temperature control is thought to be a synergistic combination of conformational change brought about by changes in local kinetic energy within the RNA secondary structure and global transcription factor change (Mayr *et al*, 2012).

The phenomenon of conformational change is well established with regards to protein amino acid structure when exposed to heat (Voellmy & Boellmann, 2007). There is also a body of evidence which suggests that mRNA conformation is essential to facilitate translation (Kozak 2005). Conformation can for example prevent capping of mRNA due to increased secondary structures (Furuichi and Shatkin, 2000).

A bioinformatics survey of the pre-miR of nmiR173 demonstrates the structural variance which varying temperatures can induce.

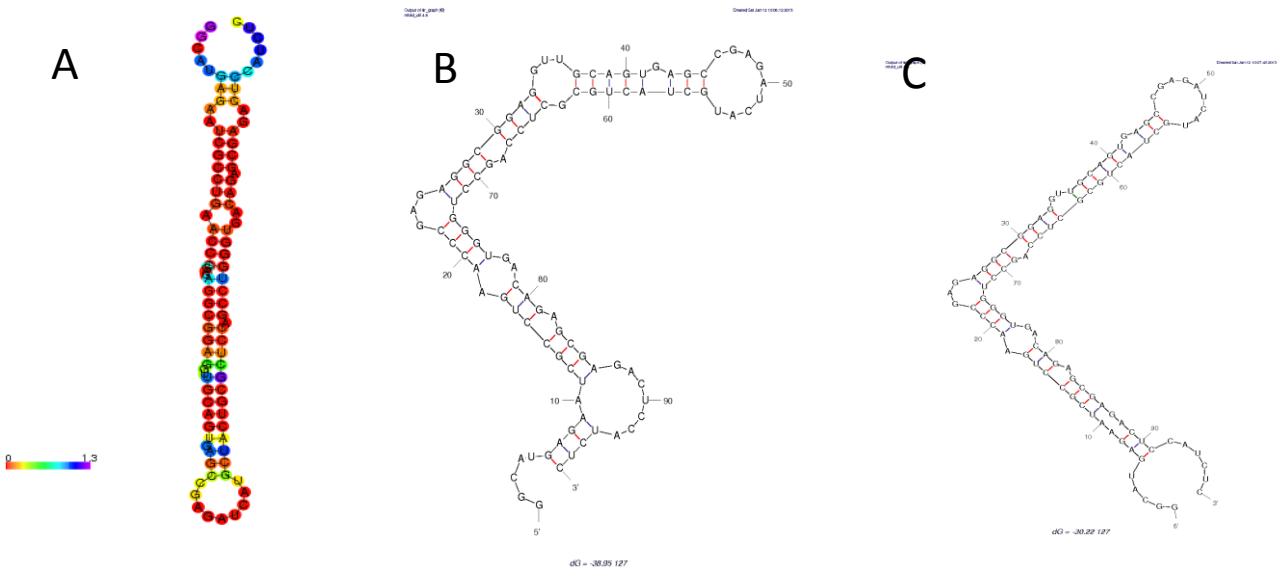


Figure 4.12: Thermodynamic stability of novel pre-miR-1273. Panel A demonstrates the predicted structure of the hairpin at 37 degrees Celsius (using Vienna RNAfold) whilst B and C show the predicted structure (using mfold) at 30 and 42 degrees Celsius respectively.

It is apparent that if we follow the expected model of miRNA biogenesis such conformational variances would hinder the expression of the mature form.

HeLa cells were cultured as before in both heat shock and hypothermic conditions (42^0C for 1 hour for heat shock, 30^0C for 1 hour for cold shock), RNA was extracted and reverse transcribed for analysis.

A

Total RNA enriched until only smRNA remain

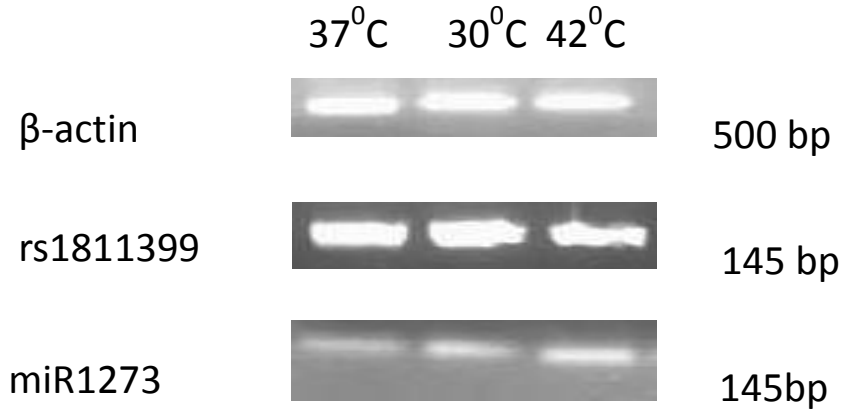
smRNA incubated with Poly(A) polymerase

Linker molecule attached to Poly(A) tail

Linked RNA is reverse transcribed

cDNA is PCRed using linker specific primer (light blue) and a miRNA specific primer (Dark blue)

B



C

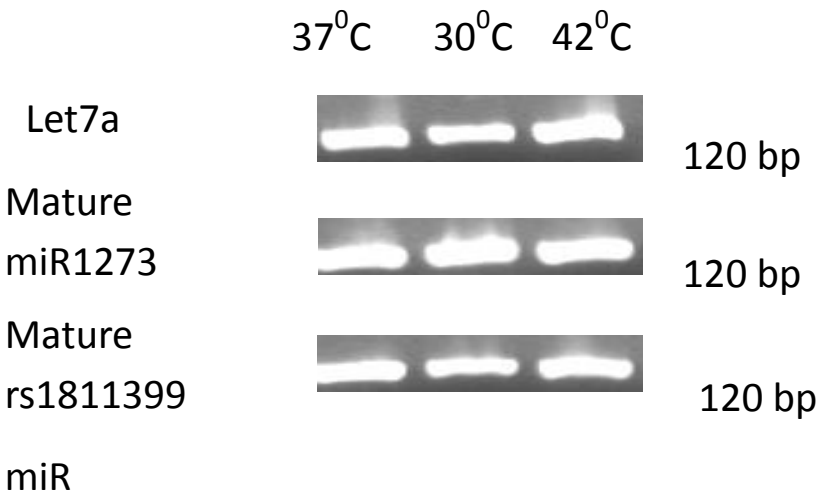


Figure 4.13: miRNA response to temperature variation. HeLa cells were grown to confluency prior to temperature shock. One plate was incubated at 37^0C for an hour prior to extraction of total RNA whilst a second plate was incubated at 30^0C or 42^0C for an hour. The total RNA pool was then enriched so that only small RNA molecules (<500nt) remained. The smRNA was then reverse transcribed using random hexamer primers. (A) Poly(A) cloning of mature miRNA. A pool of RNA enriched for small RNA (<500nt) is poly-adenylated before linker ligation. This ensures the molecule is of sufficient size in order to clone once a reverse transcription reaction has occurred. The linker sequence allows its use as a primer for probing the cDNA pool to detect mature miRNA. (B) demonstrates the response of the precursor miRNA hairpin to temperature. Primers used to probe the cDNA pool were β-ACTIN-F+β-ACTIN-R, '399-F+'399-R and N1273-F+N1273-R. (C) Demonstrates the response of the mature form as detected by Poly(A) linker PCR. It is evident that temperature does not influence either the expression of the precursor molecule or its maturation into its final form. The primers used for (C) were RTQ-UNI+mLET7, RTQ-UNI+m399 and RTQ-UNI+mn1273.

As shown in Fig.4.13 we can demonstrate that neither heat nor cold shock has an appreciable effect on the expression of the pre-miR in either case, again furthering the notion that the key targeted genes of these miRNA are of routine, housekeeping genes.

The fact that the mature form of both miRNA is expressed would imply that within the structure of the precursor molecule lays a consensus sequence for chaperone proteins (Mayr *et al*, 2012). However as these sequences are poorly understood it is unfeasible to ascertain which sequences these would be.

DNA damage was a further method of induction and its influence should be investigated.

4.5 Expression of miRNA in response to DNA damage.

As previously described, cells were incubated in the presence of genotoxic stress in order to induce DNA damage. If the novel miRNA being investigated have a role in modulating the DNA damage response there should be a difference in expression patterns.

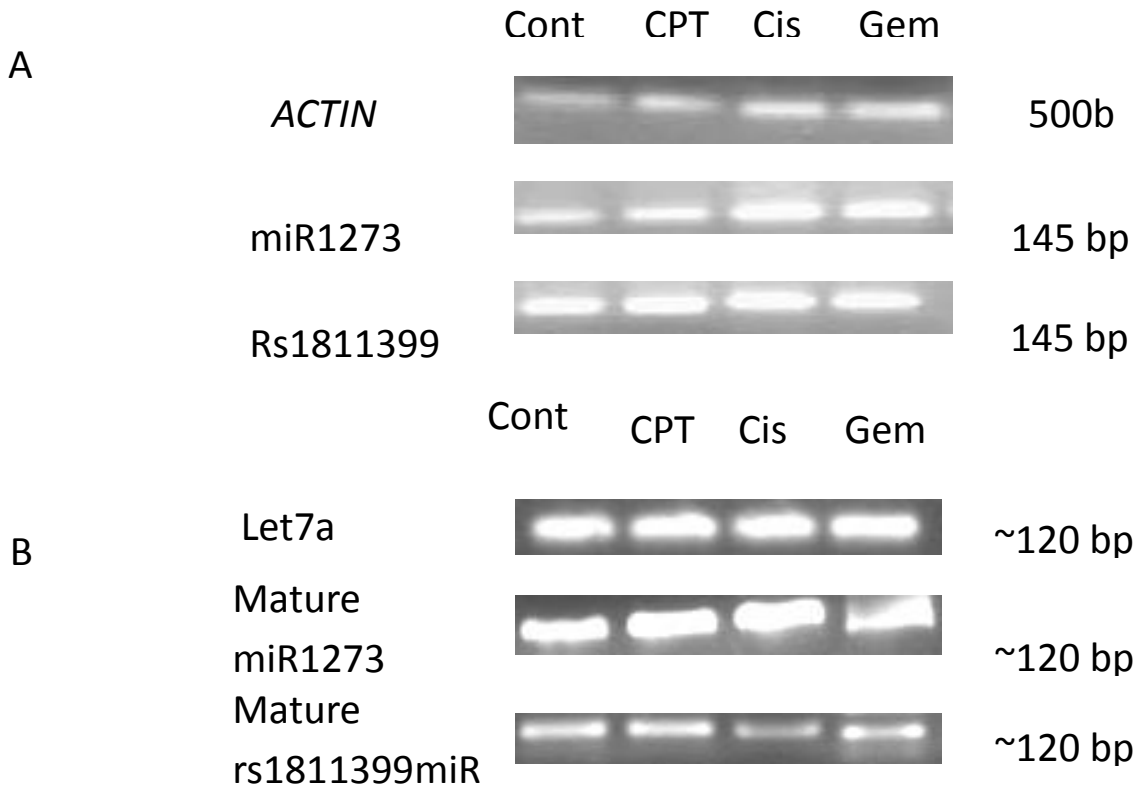


Figure 4.14: miRNA response to DNA damaging agents. 4uM of CPT and 100nM of Gemcitabine was included within the standard media of 80% confluent HeLa cells. After incubation with the drugs for 24h, the cells were rinsed with PBS and the total RNA was extracted. For the data in (B) the small RNA pool was poly-adenylated prior to linker ligation. Finally it was reverse transcribed into cDNA. (A) Demonstrates the response of the precursor miRNA hairpin to DNA damaging agents. Primers used to probe the cDNA pool were β-ACTIN-F+β-ACTIN-R, '399-F+'399-R and N1273-F+N1273-R. (B) Demonstrates the response of the mature form as detected by Poly(A)

linker PCR. The primers used for (B) were RTQ-UNI+mLET7, RTQ-UNI+m399 and RTQ-UNI+mn1273.

Figure 4.14 above identifies that there is no significant difference in expression between the precursor and mature forms of the miRNA in the presence of DNA damaging drugs. This would imply that the mature form of the miRNA is responsible for maintaining some essential housekeeping genes not involved within the DNA damage pathways.

4.6 Cell density's impact on miRNA expression.

Mori *et al* (2014) have demonstrated that cell-cell contact is essential for the activation of expression of many miRNA. It is therefore important to know if the novel miRNA within *NPAS2* fits into this category.

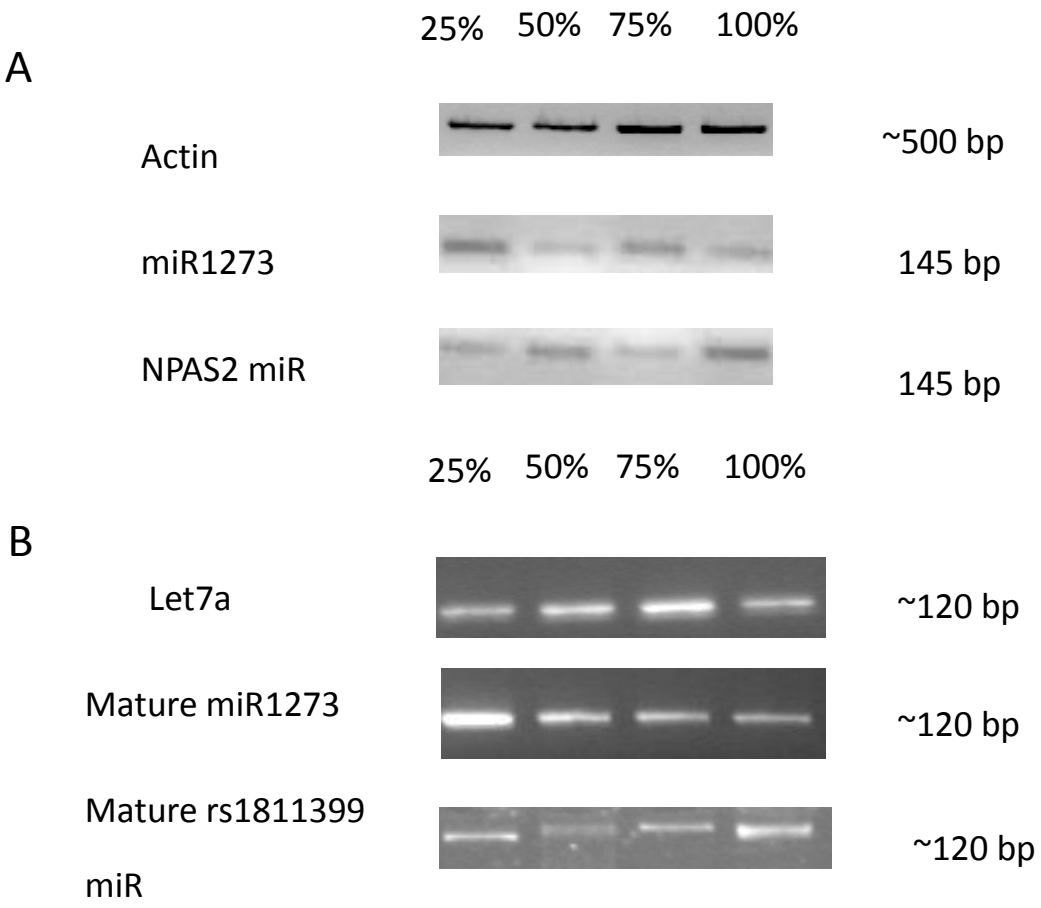


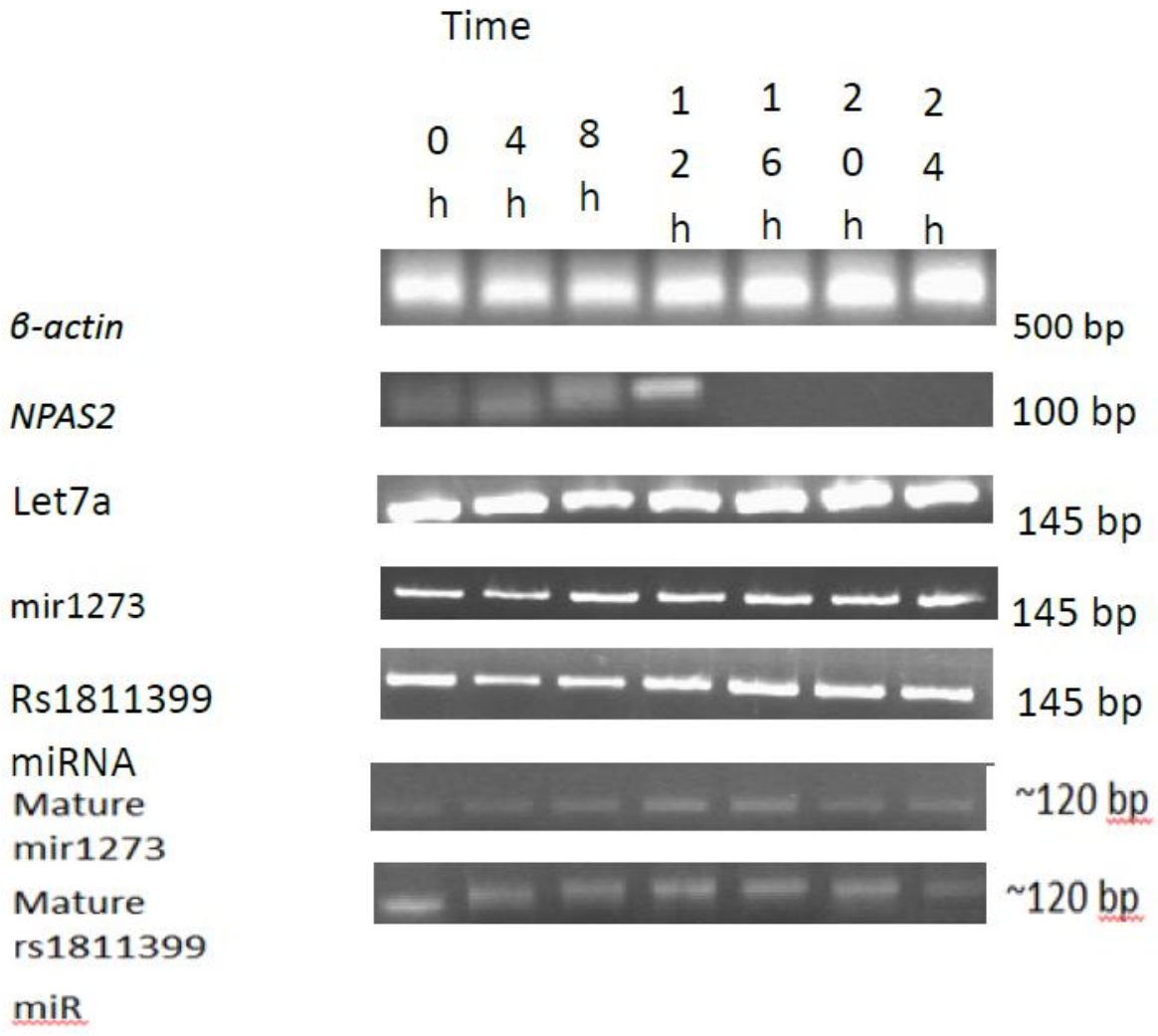
Figure 4.15: miRNA response to cell-cell contact. HeLa cells were grown to varying density before their RNA was extracted and analysed for expression. At the stated confluency total RNA was extracted. Once extracted the RNA either instantly reverse transcribed using random hexamer primers or using poly-adenylation linker PCR. (A) Response of the precursor miRNA hairpin to cell density. Primers used to probe the cDNA pool were β-ACTIN-F+β-ACTIN-R, '399-F+'399-R and N1273-F+N1273-R. (B) Demonstrates the response of the mature form as detected by Poly(A) linker PCR to cell density. The primers used for (B) were RTQ-UNI+mLET7, RTQ-UNI+m399 and RTQ-UNI+mn1273.

Figure 4.15 identifies that the novel variants within *NPAS2* are not linked to cell density. This supports the idea that the miRNA are involved in essential cellular processes.

4.7 Expression of novel miRNA does not require transcription of host-gene.

The expression of many miRNA is linked to the expression of its host gene. Baskerville & Bartel (2005) identified that miRNA are dependent on their host gene for their promoter regions. As the novel miRNA exhibited expression within asynchronous cells (Fig.4.8) it is reasonable to assess their expression in cycling cells. As before, cells were submitted to a serum shock and RNA was extracted at the noted time intervals for reverse transcription.

A



B

	Time (h)						
Gene	0	4	8	12	16	20	24
β-actin	23662.037	21933.886	22100.381	23111.24	22330.41	22223.04	22000.21
NPAS2	3377.861	4923.369	6159.473	9580.296	0	0	0
Let7a	13100.217	12589.811	13405.276	12059.95	12113.07	14005.28	13435.35
miR-1273	5235.062	4195.477	5335.098	4078.456	4198.812	4117.85	4354.21
rs1811399	5912.073	4708.267	4699.627	4503.254	4777.911	4693.277	4421.588
mat-1273	2005.58	2101.369	2100.397	2378.125	2789.27	2309.288	2273.374
mat-399	2981.101	2890.3	2902.366	3125.745	2977.664	2907.367	2807.354

Figure 4.16: Expression of both precursor and mature forms is non-circadian cycle dependant. A) HeLa cells were subjected to serum shock assay before RNA was extracted at certain time points. RNA was then subjected to RT for precursor detection and Poly(A) linker RT-PCR for mature forms, the unequal addition of adenine to the pool of mature miRNA has led to a slight variation in size for all the miRNA bands. Experiment was repeated with HEK-293, lymphoblastic and SH-SY5Y with identical result. Primers used to probe the cDNA pool were β-ACTIN-F+β-ACTIN-R, '399-F+'399-R and N1273-F+N1273-R. For the detection of the mature form by Poly(A) linker PCR the primers used were RTQ-UNI+mLET7, RTQ-UNI+m399 and RTQ-UNI+mn1273. B) Panel B demonstrates the relative intensity of each band as per the ImageJ semi-quantitative analysis.

Figure 4.16 suggests that the expression profile for the host gene and the novel miRNA was different given that *NPAS2* was not expressed. Whilst the expression of *NPAS2* increases for the first 12h before disappearing, the level of precursor is constant at an average of 4502.138 intensity for miR-1273 and 4816.571 for rs1811399 (in comparison to Let7a at 12958.42). This implied that the miRNA cluster within intron 1 might have an independent mechanism for its expression which should be detectable.

4.8 Role of locus in potential regulation of *NPAS2*

Utilising the UCSC browser it is possible to identify regions of active transcription, promoter binding or other signs of "active" DNA. The browser utilises data produced by the ENCODE project.

The figures presented below will identify if the region has a role in allowing the continuation of transcription for downstream exons. It will be possible to identify any splice sites, transcription factor binding sites, methylation zones or any other marker of active DNA through which the rs1811399 SNP may interfere.

Referring to Figure 4.17 were the region to be one with an essential role within regulation of the *NPAS2* gene it would have regions of transcription factor binding. One mechanism by which the UCSC browser detects potential transcription factor binding is by utilising the Biobase algorithm which calculates a weighted average consensus sequence for each transcription factor based on multiple experiments in human, mouse and rat. With this matrix it is then possible to scan a genome and allocate potential binding sites.

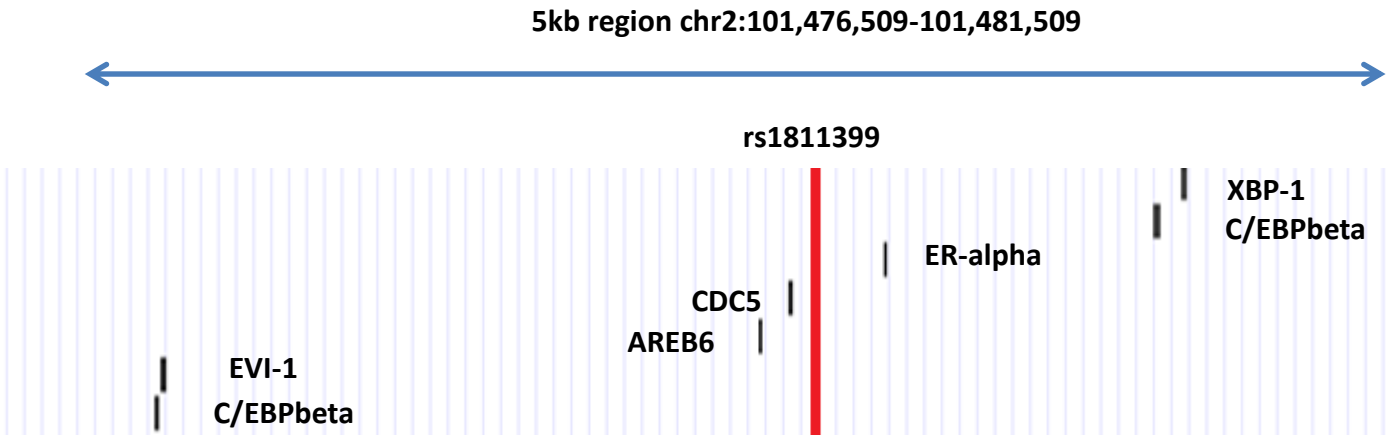


Figure 4.17: Transcription facilitators' prediction. The UCSC uses a mathematical algorithm to detect sequences in the genomic DNA which are conserved transcription factor binding sites. Rs1811399 SNP is denoted by the red line, each transcription factor binding is noted with a black line and labelled. Conserved within the region (chr2:101,476,509-101,481,509) is the recognition sequences for: XBP-1, C/EBPbeta, EVI-1, AREB6, CDC5 and ER-alpha. None of these transcription factors have been demonstrated to bind to the region in ChIP-seq experiments on the database.

Figure 4.17 demonstrates a potential for 7 transcription factor binding sites within this 5kb locus:

- XBP-1: chr2:101,480,119-101,480,135
- ER-alpha: chr2:101,480,036-101,480,054
- C/EBP alpha: chr2:101,479,217-101,479,229
- Cdc5: chr2:101,478,930-101,478,941
- AREB6: chr2:101,478,838-101,478,846
- Evi-1: chr2:101,477,027-101,477,041
- C/EBP beta: chr2:101477009-101477022

Of these predicted transcription factor binding sites, none are associated with any SNP. Of the listed transcription factors, two have been implicated in neurological conditions such as schizophrenia; XBP-1 (Chen *et al*, 2004) and ER-alpha (Weicker *et al*, 2008), AREB6 has been demonstrated to be associated with Alzheimer's disease (Grupe *et al*, 2010)

When correlated with the ChIP-seq data (Lee *et al*, 2012) for the locus none of these predicted transcription factors was found to bind to the locus across many different tissue types. Of all the tissue types and transcription factors assayed by Lee et al (2012), only the transcription factor CTCF in the human lymphoblastic cell line GM13976 was demonstrated to bind within the locus (chr2: 101,477,408-101,477,585). These co-ordinates would place the transcription factor binding site 1.4kb upstream of the SNP. There are no SNPs with which rs1811399 is in linkage with within the site.

When one compares the DNase I hypersensitivity sites described in Figure 4.18 one can detect a broader consensus as the DNase sites are conserved within a further 18 cell line types (out of a total of 125).

4.9 Chromatin State

A further indicator of any potential role the region might have with regards to regulation can be deduced, dependant on the state of chromatin at that location. Chromatin is the collective term for DNA tightly bound to its storage proteins: histones (Dame, 2005). Chromatin can either be described as open for transcription (euchromatin) or closed for transcription (heterochromatin). Neither of these states is absolute and transition between them is dependent on the requirements of the cell.

Ernst *et al* (2011) have mapped the complete status of human chromatin within nine cell types by using a combination of chromatin immune-precipitation and DNA sequencing to identify regions that are actively transcribed within each cell type. Below is data from their publication centred on the rs1811399 locus.

The figure below represents data collated from ChIP-seq experiments and placed on the UCSC browser. This represents an attempt to identify regions of chromatin which exist in varying states with regards to transcription of DNA. This variance in regulation between cell lines may reflect the expression of *NPAS2* across the varying tissues in humans. K562 is a lymphoblastic cell line in which transcription of *NPAS2* is repressed (according to the ENCODE data), this would replicate our findings with HI2162 in which *NPAS2* was not detected (Fig 3.1 and Fig.3.4). The other cell lines are all fibroblastic in nature and express the region and *NPAS2* as observed in Fig 3.1, Fig.3.2 and Fig.3.3.

Intriguingly rs1811399 is located towards the 3' end of a weak transcription enhancer region and the 5' end of a transcribed region within the normal human lung fibroblast (NHLF) cell line. It is noted within the literature (Taka *et al*, 2011) that a SNP within such a region can contribute to altered expression of downstream exons.

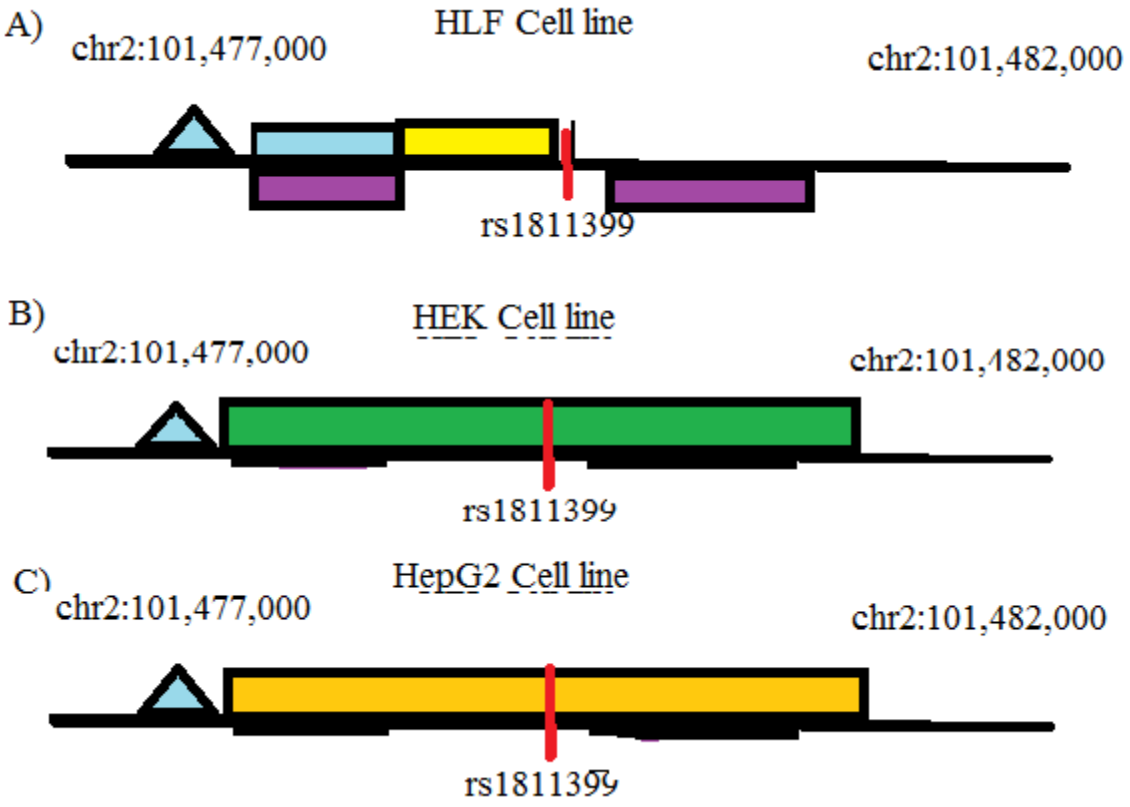


Figure 4.18: Figure demonstrating the correlation between the H3K27me3 mark and predicted chromatin structure of the rs1811399 locus across varying cell lines. The rs1811399 SNP is denoted by the red line whilst the blue triangle represents the novel miR-1273 species. A) The blue block represents a transcription insulator region (chr2:101,477,569-101,478,168) whilst the yellow block represents an area of transcription enhancement (chr2:101,478,169-101,478,768). The purple blocks represent areas of H3K27 methylation that identify regions of low transcription. B) The green block (chr2:101,472,569-101,476,168) within the HEK cell line is an area of active DNA

that is transcribed. C) The orange block (chr2:101,472,569-101,476,168 represents an area of transcriptional activity as in HEK cells.

Intriguingly the above datasets (Fig 4.19) may point to the region having a role within regulating gene expression. An H3K27me3 mark was detected by O'Geen, Echipare and Farnham (2011), which is a histone modification often found within silenced regions of chromatin. Hagarman *et al* (2013) have demonstrated that methylation of DNA, a mark of its transcriptional activity, halts the placing of methyl groups on histones. Areas of high transcription would therefore have low levels of H3K27me3.

4.10 Expressed Sequencing Tags

Perhaps a simple way of noting the transcriptional activity of the locus is by the detection of expressed sequencing tags (EST). ESTs are cDNA copies of short (~500bp) fragments of expressed genes and are indicative of active DNA regions.

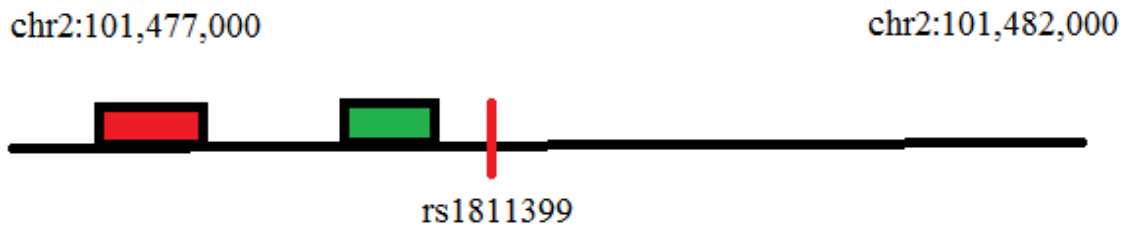


Figure 4.19: rs1811399 locus with known ESTs highlighted. Sequences that have been cloned are present as coloured blocks: T59368 (red block) and BI033160 (green block).

ESTs are registered with several databases such as GenBank. It is therefore possible to search for curated sequences and map them to a region of interest. Using the UCSC browser it was possible to detect 2 recorded EST within 5kb of rs1811399.

EST T59368 is 342 base pairs long and was isolated from RNA extracted from the ovary of a 49 year old female and exhibits 97.4% homology with the reference genome. The residual 2.6% variance is not contiguous with any known SNP and is caused by the insertion of 4 guanine residues at locations 237, 264, 272 and 298. These mutations can be account for by either *de novo* mutations or sequencing error as they do not appear in any SNP catalogue of healthy individuals to date.

T59368 is of interest as this EST contains the precursor for the novel miR-1273 miRNA detected and analysed during the course of this work.

EST BI033160 is a 284bp fragment of which no providence is known bar that it is from an adult human. It has 96.5% homology with the reference genome and appears to be of the reverse strand. rs117623721, rs143817583, rs192325412, rs76376883, rs183699561 and two *de novo* or sequencing errors would account for the sequence variance.

BI033160 may also be of interest as it contains a region of high homology with miR-297 (E-value of $4x10^{-4}$).

4.11 Repeating elements within rs1811399 locus

Repeating elements are sequences of DNA which are found in multiple copies across the genome. They are known to constitute around half of the human genome (Lander *et al*, 2001). Repeating elements are important evolutionary drivers and dozens of protein coding genes have derived from them (Lander *et al*, 2001). Repeat elements are involved with gene expression regulation so it is important to review their location relevant to rs1811399, and whether rs1811399 may reside in one.

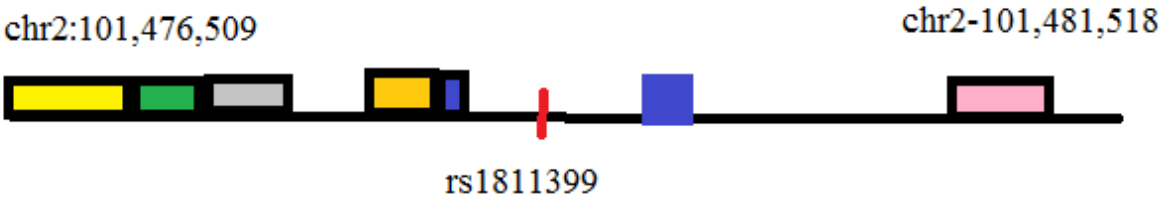


Figure 4.20: Repeating elements within genome locus of rs1811399 within UCSC hg19. Sequencing has identified the following: AluJR: chr2:101477345-101477628 (yellow box), L1MC4: chr2:101476248-101477314 (green box), L2b: chr2:101477645-101478106 (gray box), Trigger11a: chr2:101478136-101478417 (orange box), (TATG)n: chr2:101478431-101478478 (purple box) and chr2:101479780-101479816 (blue box) and Charlie 1b: chr2:101481195-101481655 (pink box).

Rs1811399 does not in itself directly impinge upon any predicted repeating elements as recorded by the UCSC.

Whilst bioinformatics has identified the rs1811399 SNP as not being involved with the regulator machinery in *NPAS2* it is important to test experimentally as much as possible.

4.12 Rs1811399 hairpin locus does not appear to bind transcription factor and its binding ability is not allele dependant.

Introns play an important role in regulating gene expression and are home to many regulatory regions (Scohy *et al*, 2010) and it has recently been discerned that SNP and other mutations can cause aberrant transcription factor binding which influence the expression of downstream exons, such as those in p53 response elements which interfere with p53 binding (Bandele *et al*, 2011). As the rs1811399 SNP has been associated with a phenotype (Nicholas *et al*, 2008) and when combined with the finding that expression of clock genes are disturbed within the

phenotype in question raises the possibility of the mutation impacting on expression (Hu *et al*, 2009).

In order to test if a region of DNA can bind a transcription factor electrophoretic mobility shift assay can be used. This was achieved by producing DNA sequences and incubating them with a protein extract. Once the DNA-protein complex has been formed it can be ran on a gel and if the DNA has bound any protein its progression down the gel will be retarded. As a positive control a region of known transcription factor activity will be incubated with protein extract parallel to the main experiment, this will be the cytomegalo virus promoter (CMV) which is routinely used in human expression vectors to drive expression of a DNA sequence (Barrow, Campo and Ward, 2006)

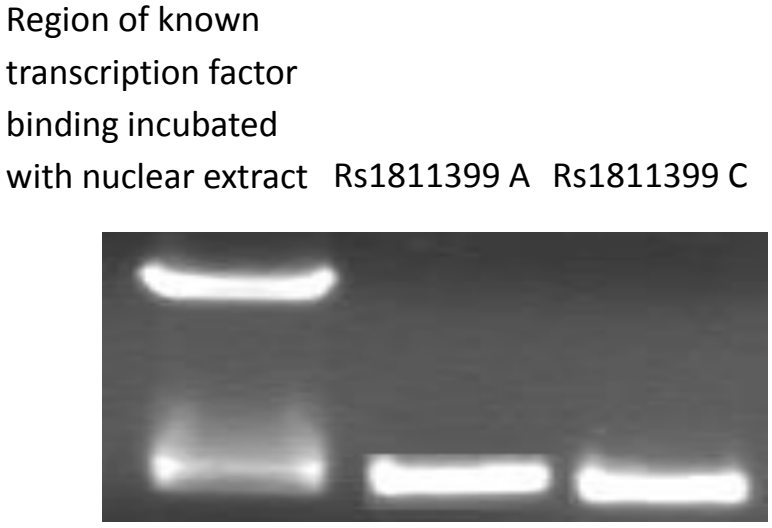


Figure 4.21: Electrophoretic mobility shift assay performed on nucleoprotein agarose separation. For positive control CMV was excised from pcDNA3.1 (Mfe1 and Nhe1 digest) to give a 600bp band. The hairpin region was cloned using fusion-PCR to provide a hairpin per allele. The hairpin was then incubated with nuclear protein extracted from HeLa cells. If the DNA actively binds transcription factor its progress down the agarose gel would be retarded.

Figure 4.22 suggested that the rs1811399 locus does not bind a transcription factor protein regardless of the allele. This informs us that the region might not be part of the transcriptional regulatory machinery of *NPAS2* and whatever influence rs1811399 has on causing a phenotype might have nothing to do with aberrant protein expression.

4.13 Identification of novel transcription start site.

As the above evidence has demonstrated the rs1811399 miRNA and novel 1273 miRNA variants can be expressed without the host gene, *NPAS2*, being transcribed. It should be possible to detect its transcription start site.

Figure 4.22: Upstream region of novel miR1273 (chr2:101,473,696-101,475,303). Novel 1273 genomic locus is identified by the arrow. 600 base pairs upstream is a ~1kb region of established histone3k27 acetylation. Further evidence for this locus being important for some transcriptional activity within the gene was the presence of bound transcription factors as ascertained by ChIP-sequencing collated by the ENCODE group.

Figure 4.23 identifies a potential region of transcriptional regulation in the UCSC browser. Of the transcription factors which have been detected at the locus, all three are of interest when viewed in the context of autism: GATA-2 is implicated in negative regulation of neuronal precursor cells, GR is a hormone induced transcription factor of significance in central nervous system development and c-Fos is involved in neuronal action potentials (Wakil *et al*, 2006; Maletic *et al*, 2007 and Dragunow & Faull, 1989).

PCR primers for a region upstream and downstream of the predicted region were designed in order to experimentally validate the transcriptional activity of the region.

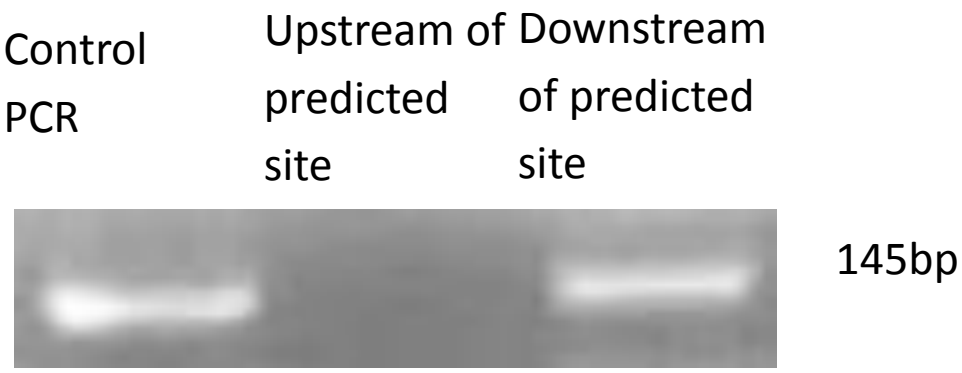


Figure 4.23: PCR was undertaken using primers specific for regions downstream and upstream of the promoter region. Control reaction was undertaken on genomic DNA using primers upstream of the putative promoter whilst the test PCR were undertaken on HeLa cDNA.

Whilst Figure 4.24 implies that only a region downstream of the area is actively transcribed it does not conclusively prove that the region is an active promoter.

4.14 Functional assessment of putative promoter region

A common method of assessing the functionality of any proposed mammalian promoter region is as follows. A plasmid vector, minus mammalian promoter, is cut with restriction enzymes and the sequence of the proposed promoter region is inserted. Downstream of this promoter a second piece of DNA is inserted, for example GFP. It is then possible to note if the initial DNA sequence is sufficient to drive the expression of a downstream gene by transfecting the construct into a human cell line and noting the expression pattern of the gene.

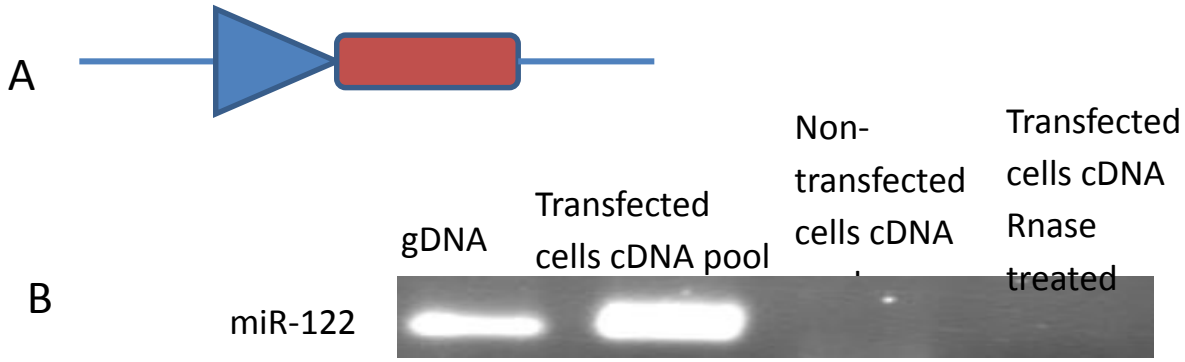


Figure 4.24: A) Schematic of plasmid construct. Putative promoter was cloned into pBluescript (blue triangle). Downstream of new promoter construct was cloned miR-122 from genomic DNA. B) As miR-122 has a hepatic expression pattern it is not detectable in HeLa cDNA as evident in the non-transfected cDNA. When the cells were transfected with the plasmid construct however a band appeared in the PCR. Potential genomic DNA contamination was disproven by the application of an DNAse treatment step to RNA prior to reverse transcription.

Using this system it was possible to demonstrate that chr2:101,473,696-101,475,303 can indeed drive the expression of a gene and may act as a potential transcription start site for the miRNA cluster.

4.15 Sequencing of rs1811399 miRNA and novel miRNA-1273 in intron 1 of *NPAS2*

Without using deep sequencing technology, identifying the sequence of a miRNA is very difficult. Primarily this is due to the short length of a mature miRNA (<25nt). In order to circumvent this, it is possible to increase the size of small RNA using a poly(A) polymerase enzyme to add a string of adenine to each RNA molecule. Reverse transcription can then be carried out and the miRNA are extended in size to 80-120nt. These can then be cloned and sequenced.

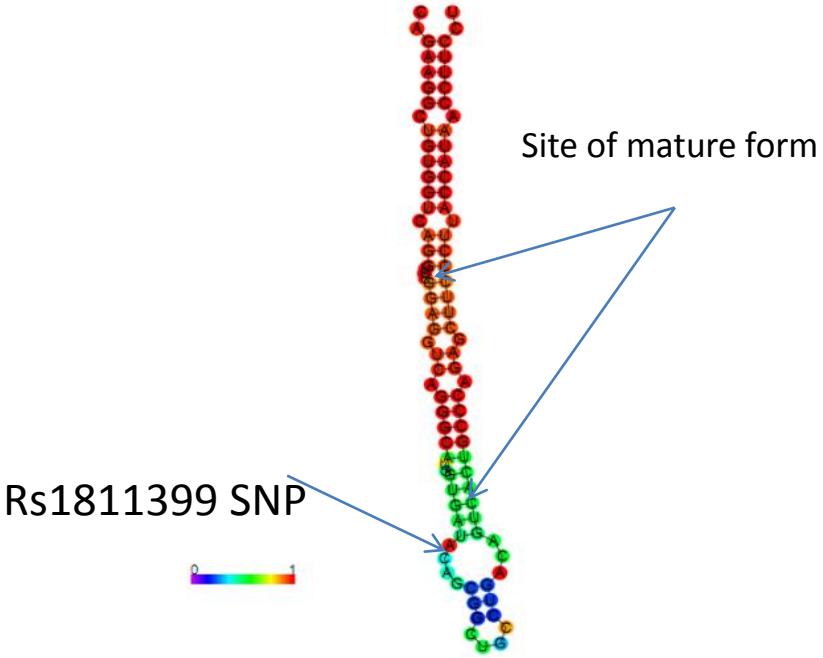


Figure 4.25: Sequence of mature rs1811399 miRNA. Elucidated via sequencing of Poly(A) linker cDNA pool. This figure was produced using Vienna RNAfold.

Figure 4.26 identifies the sequence of a small miRNA mapped to the hairpin precursor. The sequence is on the 3' arm of the hairpin and does not include rs1811399. The significance of

this will be explored later. The sequence of the mature rs1811399 miRNA is CAGUCACUGCCCAGAGCUUCCC .

The same procedure was used to identify the novel miR-1273.

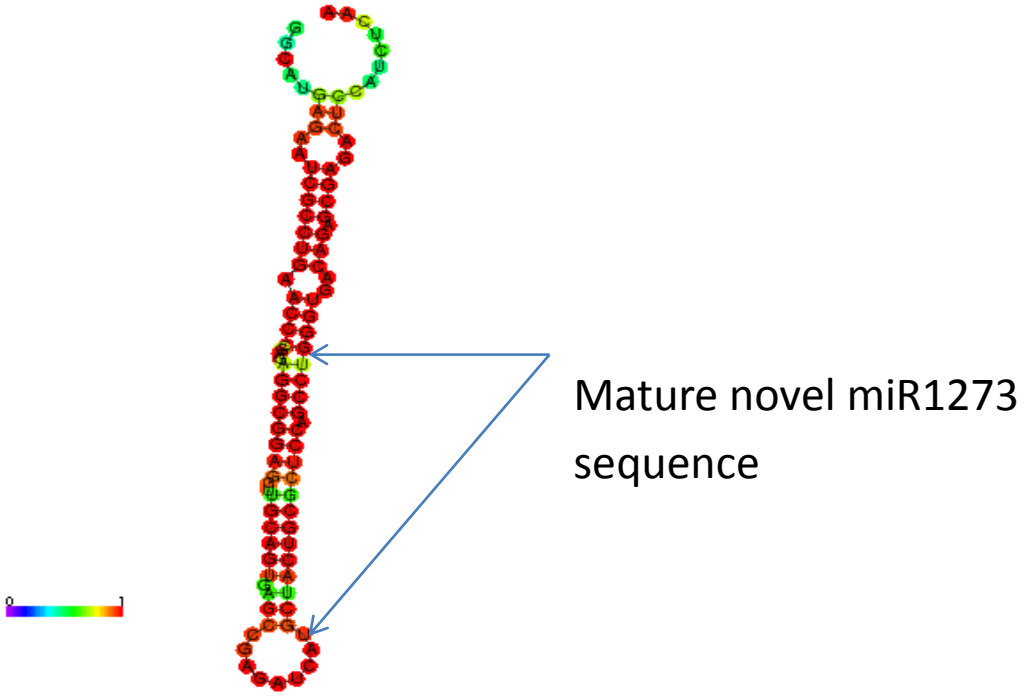


Figure 4.26: Sequence of mature novel miR1273 miRNA. Elucidated via sequencing of Poly(A) linker cDNA pool. This figure was produced using Vienna RNAfold.

The sequence for the mature form of novel miR-1273 was AGGCAUGAGAAUCGCCUGAACC.

With both miRNA sequenced it was possible to identify potential targets. This was important to understand any role the miRNA may have in any pathologies.

4.16 miRNA on both arms of rs1811399 hairpin.

All known miRNA are released from the DICER complex as a dsRNA duplex. One of these arms is then "selected" for use as the guide arm in the silencing complex. This would appear wasteful as half of the potential siRNA is discarded. Recent studies have however identified occasions when a miRNA is utilised from both arms of the dsRNA duplex (Allen *et al*, 2004).

Primers were designed for the region on the opposite arm of the hairpin and RT-PCR was conducted on a poly(A) treated cDNA pool.

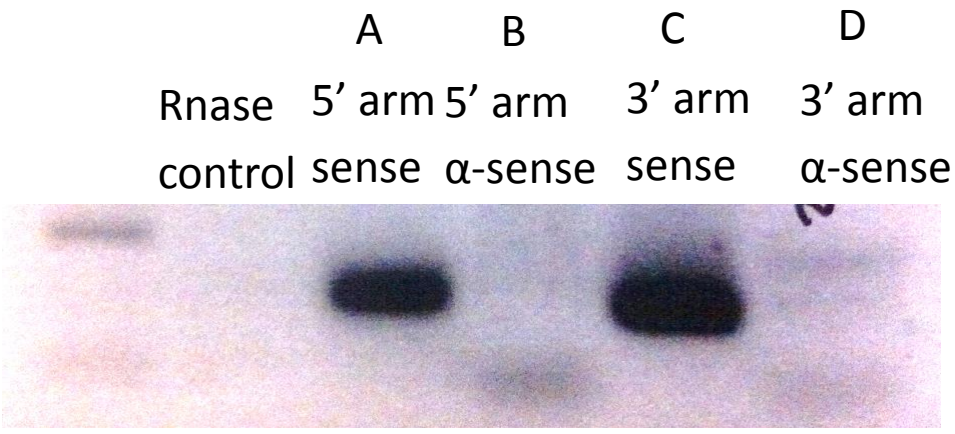


Figure 4.27: Agarose gel image of ootential miR located on opposite arm of rs1811399 miRNA precursor hairpin. HeLa cells were grown to 80% confluency under standard conditions. RNA was extracted from the cells and enriched for small RNA. The small RNA pool was then poly-adenylated prior to linker ligation. The poly-adenylated pool was then probed using RTQ-UNI+m399 (lane A), RTQ-UNI+m-399b (lane B), RTQ-UNI+m399c (lane C) and RTQ-UNI+m399d (lane D).

From the Poly(A) linker PCR methodology used to extract this information we know the sequence of the second miRNA to be: CAGGUCUGGAGGUCAGGGCAUG

The appearance of a second miRNA from the rs1811399 hairpin is interesting. Rajagopalan *et al* (2006) identifies young miRNA as being most likely to express two or more forms of miRNA from a single hairpin. This is believed to be because the function of the mature form(s) has not been fixed into routine processes.

4.17 Target prediction.

By downloading the 3' UTR sequences from GenBank, it is possible to code a program using PERL which will analyse the UTR sequence and search for regions complimentary to a miRNA seed sequence. Compiled code was downloaded from the TargetScan website (Lewis, Burge & Bartel, 2005) and ran on Perl Package Manager.

TargetScan was selected as the primary target searching algorithm as its many citations increase the confidence that can be attributed to the results. Further, it was decided that more emphasis would be given to target genes with more than one target sequence within the 3'UTR. It is noted within the literature that a gene which contains multiple target sites for a single miRNA species is much more likely to be regulated by that miRNA (Fang & Rajewsky, 2011).

Once a set of gene targets had been identified it was possible to assign each gene to a protein interaction network using gProfiler (Reimand, Arak & Vilo, 2011). gProfiler assigned each query gene to a particular network based on established gene ontologies. If a protein was not known to take part in a particular network the software would analyse the protein's sequence and place it with proteins of a similar sequence. Each grouping was awarded a *P*-value, the lower the value the more significant the grouping and the less likely that the association was random. This further analysis allowed the author to investigate further potential downstream implications of the putative miRNA.

The targets which are identified may not be regulated by the miRNA in question as by definition, the program only searches for sequence matches. *In vivo* several processes would regulate if a miRNA regulates a specific gene.

4.17.1 Novel miR-1273 miRNA target candidates.

The seed sequence of the experimentally validated miR1273 homologue was entered into the PERL script and a potential 148 targeted genes were shown. Of the two (*PTGER3* and *USP47*) have more than one complimentary site within their UTR, raising the possibility of them being *bona fide* targets. 57 others have full base pairing between the seed region and one locale within their UTR. Of the remaining 89 only an imprecise match up of 7 nucleotides is evident.

PTGER3 is a member of the prostaglandin receptor family and has been implicated in autism (Nava *et al*, 2013) whilst USP47 is a ubiquitin peptidase and has not been implicated in any specific neurodevelopmental disorder.

Table 4.1 below summarises the pathways that all the proposed targets are involved in.

Pathway in which target gene is active	**Name of target gene**	**P value of grouping**
Nervous system development	*ATP7A,BMI1,ELAVL3,GABRB2,NEURL,ONECUT2,OPRM1,PRICKLE2,PRKCA,PTCH1,SOX4,TIAM1,UBE3A, ARHGEF2,CACNB1,CELSR2,NAV2,PEX5,DCX,LMO4, MYH10,PAX6,PLXNA2,PRKCI,RPS6KA3,SEMA6B,SOX11,TIMP3*	2.10E-02
Central nervous system development	*ATP7A,BMI1,PRKCA,PTCH1,SOX4,UBE3A,NAV2,PEX5,DCX,LMO4,MYH10,PAX6,PLXNA2,RPS6KA3,SOX11,TIMP3*	3.05E-02
Central nervous system neuron differentiation	*ATP7A,PRKCA,SOX4,PEX5,DCX,LMO4,PAX6*	3.24E-02

Table 4.1: Predicted targets of novel miR-1273-1 and the pathways in which they are involved. Majority of targets are seemingly involved with embryonic development, especially central nervous system development.

Intriguingly, several of the proteins targeted by the miRNA might have an impact on the function of circadian rhythmicity. For example; *STC2* (a glycoprotein growth hormone (Chang & Reddel 1998), *CELSR2* (a neuronally expressed growth factor receptor (Vincent, Skaug and Scherer, 2001) and *ZNF346* (a dsRNA binding zinc finger protein Yang, May and

Ito, 1999)) are implicated within pathways with *BHLHE40*. BHLHE40 is a basic loop helix protein which selectively binds to E-box regions and *represses* transcription. If the novel miR-1273-1 were able to influence the activity of this gene, it would allow NPAS2 to bind to the E-box and commence transcription. Intriguingly BHLHE40 is also known to directly bind to ARNTL possibly inhibiting its action. This is of importance as ARNTL forms a heterodimer with NPAS2 which is required for its transcription factor activity.

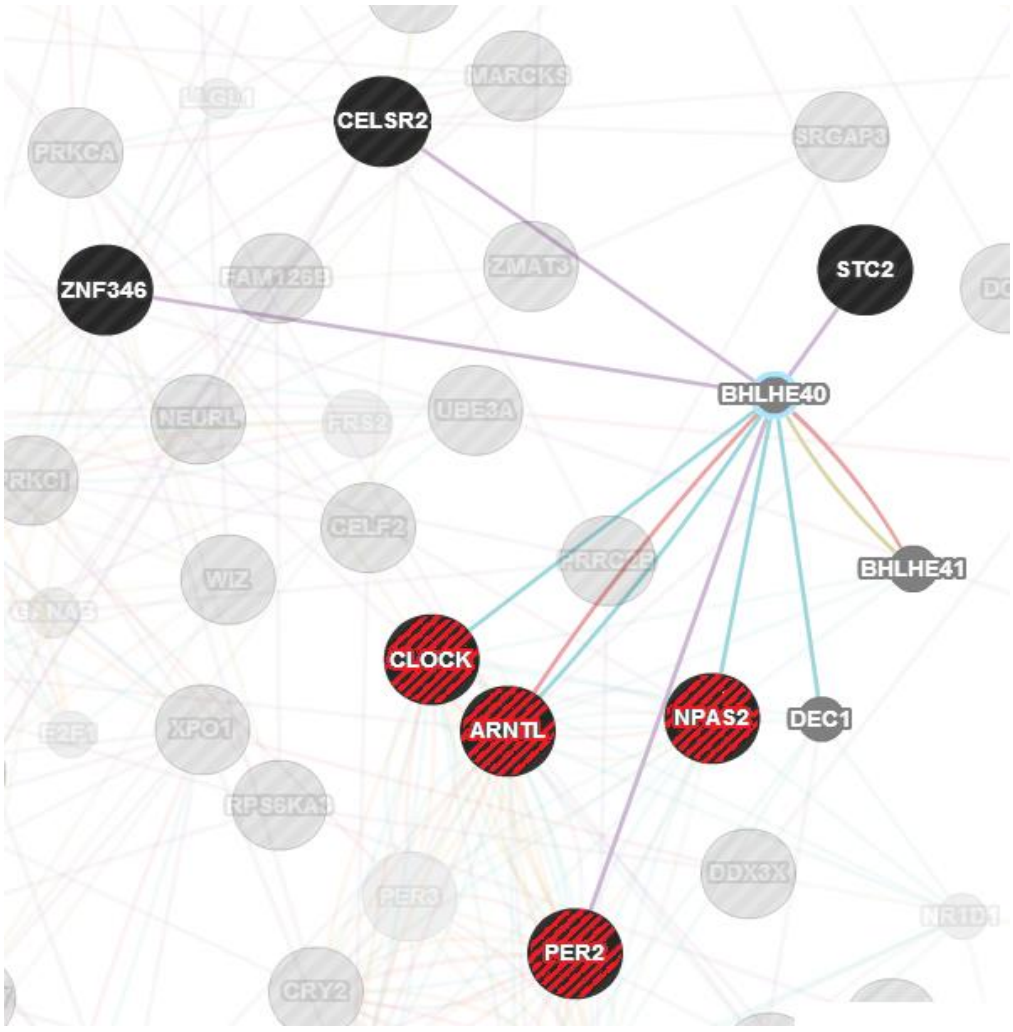


Figure 4.28: Protein interaction network of novel miR-1273-1 targets. Gene names in black are mRNA species that might be putative targets based on *in silico* analysis by novel miR-1273-1, red and black circles are circadian clock genes input as reference, and small grey circles are intermediary proteins.

Novel miR-1273-1 also appears to play a role in regulating a neuronal development circuit.

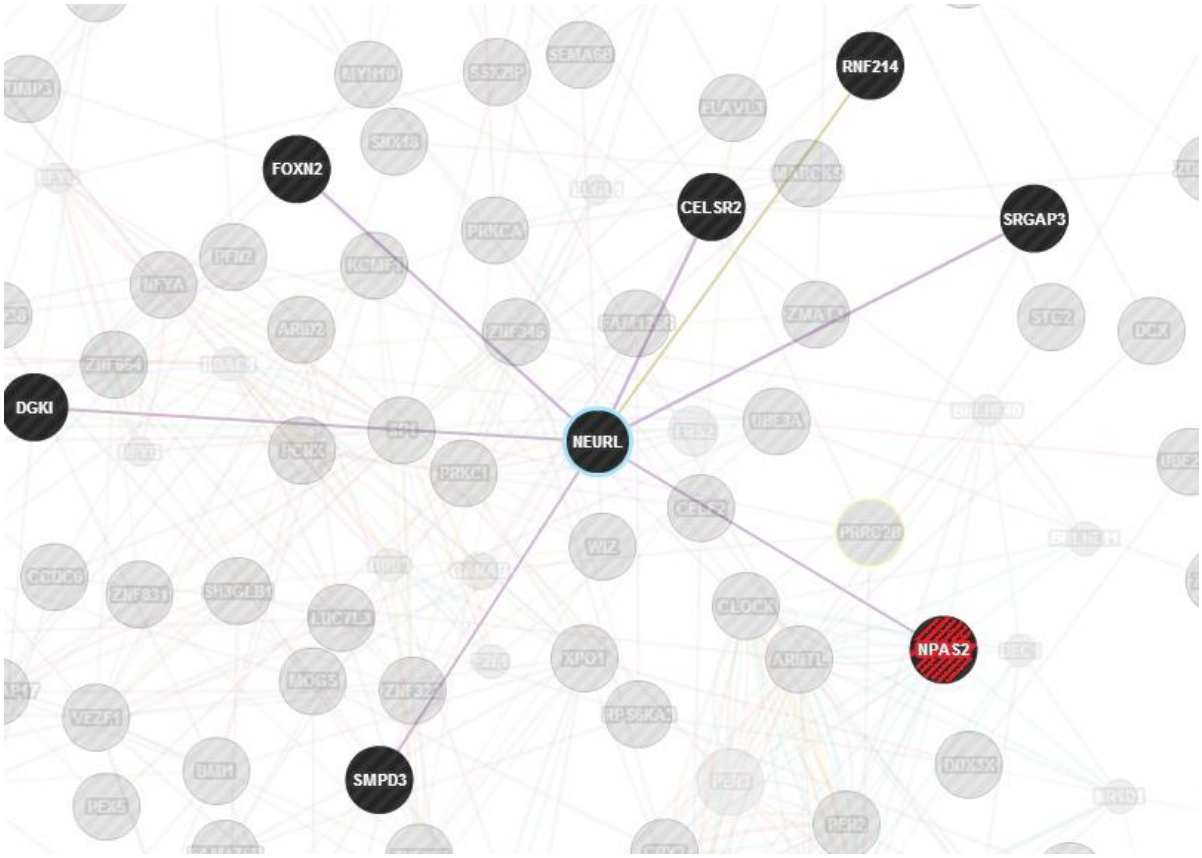


Figure 4.29: Neural development circuit putatively targeted by novel miR-1273 based on *in silico* analysis. Gene names in black are mRNA species putatively targeted by novel miR-1273-1, red and black circles are circadian clock genes input as reference.

This development network is important as it provides an insight into an expanded role for *NPAS2* in neurodevelopment. Whilst *NPAS2's* role as a transcription factor and clock gene is well known, the fact it hosts miRNA genes allows it to have greater influence on genetic regulation via the influence that novel miR-1273-1 might have on other gene regulation pathways.

4.17.2 Rs1811399 5’ arm miRNA target prediction.

This miRNA’s seed sequence is predicted to target 121 transcripts. 6 of these transcripts have two sites within their 3’ UTR (*SENP5, HIC2, METAP1, MLL, SOX11 and ZFHX4*), 27 have perfect 8nt seed recognition and the remainder have 7nt recognition.

Of the genes with two binding sites within their UTR *SOX11* and *MLL* are implicated in neurodevelopment disorders such as autism or schizophrenia (Lo-Castro *et al*, 2009 and Huang *et al*, 2007).

This miRNA perhaps has the more interesting predicted targets of the cluster. For example it is predicted to regulate DROSHA, DICER, DGCR8 and even AGO protein expression.

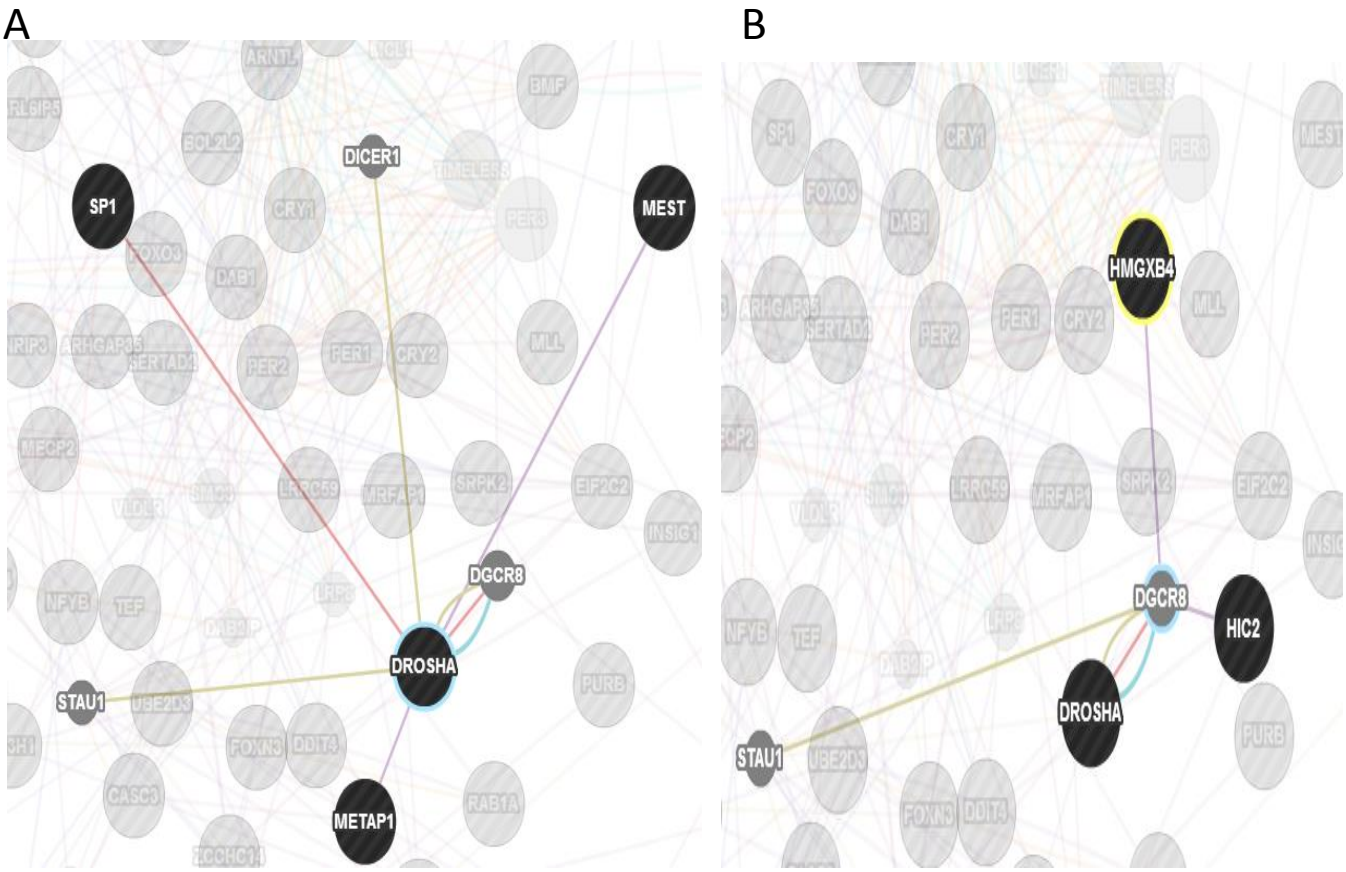


Figure 4.30: Putative targets of Rs1811399 5' include components of the microprocessor complex. Genes in black circles are directly targeted by the miRNA; genes in small grey circles are indirectly targeted (expression of gene is dependant on gene that is directly targeted by the miRNA) . Panel A describes the potential regulatory effect of the miRNA on *DROSHA* as well as the targeting of other genes within the Drosha pathway. Roles for *MEST* and *METAP1* are unknown; however some role as transcription initiators of *DROSHA* may be possible. *METAP1* however also has a metal binding activity which is essential for the microprocessor. *SP1* on the other hand is known to have a direct interaction with Drosha via pulldown experiments (Gunther, Laithier & Brison, 2010). Panel B implies the indirect regulation of *DGCR8*. HIC2 is a zinc ion binding protein and has a role in the repression of transcription in a DNA binding dependant manner (Deltour *et al*, 2010). HMGXB4 is part of a developmental signalling

pathway and has sequence similarity with the HMG-box transcription factor family of which SRY is a member of.

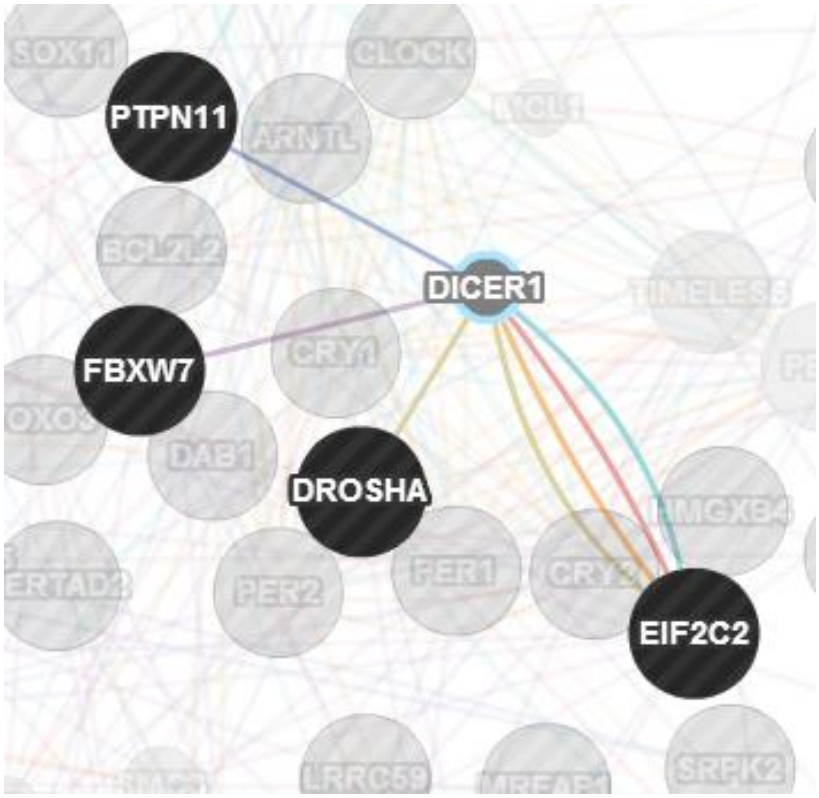


Figure 4.31: 5' arm rs1811399 miRNA's potential role in regulating miRNA processing. Whilst the miRNA might not directly regulate *DICER* mRNA it does appear to regulate *DROSHA* which is responsible for initial processing of the pri-miRNA into a DICER suitable substrate. The miRNA also targets *EIF2C2* which is a synonym for *AGO 2*, a key component of the RISC. *PTPN11* is a key regulator of cranio-facial development and is co-expressed with DICER protein (Johnson *et al*, 2003). The *FBXW7* gene encodes for a protein which again is developmentally linked and highly expressed within brain tissue (Li *et al*, 2002).

The potential role this locus could have on a developmental disorder of the brain could be explained from Fig. 4.31 and Fig, 4.32. Whilst none of the genes listed in the two figures are

implicated in autism at the time of writing. It should be noted that the impact could be amplified as the effects progress down their networks.

4.17.3 Rs1811399 novel miRNA 3' arm.

The 186 genes potentially targeted by this particular miRNA are involved in over 100 different categories of function, vastly more than any of the other cluster miRNA. The true figure of pathways involved is obviously higher as each of these categories is a group of similar processes.

Of the targets; nine have two binding sites within their 3'UTR. They are: *C15orf57*

C18orf34, KCTD10 (constituent of an ubiquitin ligase complex; Wang *et al*, 2005), *KIAA2022* (a protein of unknown function found on the X chromosome implicated in autism; Van Maldergem *et al*, 2013), *ATRX* (a ATPase/helicase implicated in X-linked mental retardation; Leung *et al*, 2013), *KLHL2* (an actin binding protein expressed in the brain; Williams *et al*, 2005), *MKLN1* (inhibitor of cell migration; Adams *et al*, 1998), *CENTG1* (complements the anti-apoptopic effect of nerve growth factor; Cai *et al*, 2012) and *PVRL4* (a cell-cell adhesion molecule; Mühlebach *et al*, 2011).

The presence of two binding sites in *KIAA2022* and *ATRX* is interesting given their association with autism. *KIAA2022* has been demonstrated to reduce neurite outgrowth causing malformed dendrites and axons, although the mechanism by which this occurs is currently unknown (Van Maldergem et al, 2013). *ATRX* was first implicated in X-linked Alpha thalassaemia-mental retardation in 2006 by Gibbons *et al*. Gong *et al* (2006) demonstrated that female relatives of many autism patients who had aberrant X-inactivation had mutations within the *ATRX* gene.

Name of pathway in which targeted genes participate.	**Name of target gene**	**P value of association**
Regulation of synaptic plasticity	*SNCA,VGF,EGR1,MAP1B,NEUROD2,RELN,CNTN2*	1.71E-02
Behaviour	*SNCA,ADAM17,EGR1,ETV1,GIGYF2,GNAO1,HMGCR,HOXD9,LEP,NAV2,NEUROD2,NR4A3,RELN,SLC12A5,CNTN2,FOXP2,SCN2A*	1.18E-03
Adult behaviour	*SNCA,GIGYF2,HOXD9,LEP,NR4A3,CNTN2,SCN2A*	2.37E-02
Regulation of inner ear receptor cell differentiation	*HES1,HES5,DLL1*	7.26E-03
Negative regulation of mechanoreceptor differentiation	*HES1,HES5,DLL1*	7.35E-04
Regulation of developmental process	*CDC42SE1,SEMA4F,TCF4,CFL1,CTNNB1,DDIT3,DPYSL2,EGR1,FOXG1,HES1,HES5,HMGCR,LEP,MAP1B,MAPK14,MYST3,NEUROD2,RELN,SMAD4,SMAD5,SP7,SPEN,SUFU,ALOX12,BRWD1,CNTN2,DLL1,FOXP2,SOCS5*	8.45E-03

Gene expression	*AFF4,HNRNPUL1,RRM2,SNCA,TCF4,WIBG,ADAR,ANKRD49,BAZ2A,CTNNB1,DDIT3,DDX3X,DDX58,EDA,EGLN2,EGR1,EID1,EIF5A2,EPC2,ESRRG,ETV1,FOXG1,FXR1,GALNT2,HES1,HES2,HES5,HOXD9,JHDM1D,LEP,MAN1C1,MAPK14,MBNL1,MED29,MNX1,MRPL42,MYST3,NCK1,NDST1,NEUROD2,NR4A3,PAPOLA,PDCD7,RBBP5,RELN,RNF141,SMAD4,SMAD5,SP7,SPEN,SUFU,WTAP,ZBTB34,ZMIZ1,ZNF618,ZXDC,ALOX12,BRWD1,CNTN2,CREBZF,DLL1,FOXP2,GABPA,GMCL1,HNRNPA1,MYT1L*	3.59E-03

Table 4.3: Selection of pathways in which the 3' arm rs1811399 miRNA might be involved within. Whilst the miRNA is also involved in a range of metabolic and synthetic pathways (see appendix 1) these chosen categories can illuminate the potential role this miRNA within the autism phenotype. Several of these target genes are directly involved with foetal development including cranio-facial development (*FOXP2* and *NEUROD2,* for example). Intriguingly FOXP2 has been implicated with severe language and speech disorders (Lai *et al,* 2001), which can be a feature of autism spectrum disorders. Aberrant synpatic plasticity is one of the many hypothesis within the scientific community as to an aetiology for autism and the presence of *REELIN, CNTN2* (an axonal isoform of contactin) and *NEUROD2* might imply a role for this miRNA in regulating this process.

It should be noted that each gene listed here is only a predicted target, experimental validation would need to be carried out on each and every one to be sure of the interaction.

4.18 Phylogenetic conservation

miRNA genes are broadly conserved across phyla, this is mainly due to the routine nature of most tasks expected of them. There are cases however of miRNA evolution, especially in higher eukaryotes (Altuvia *et al*, 2005). This set of experiments aims to ascertain whether this miRNA cluster is widely conserved or is it a relatively young microRNA cluster.

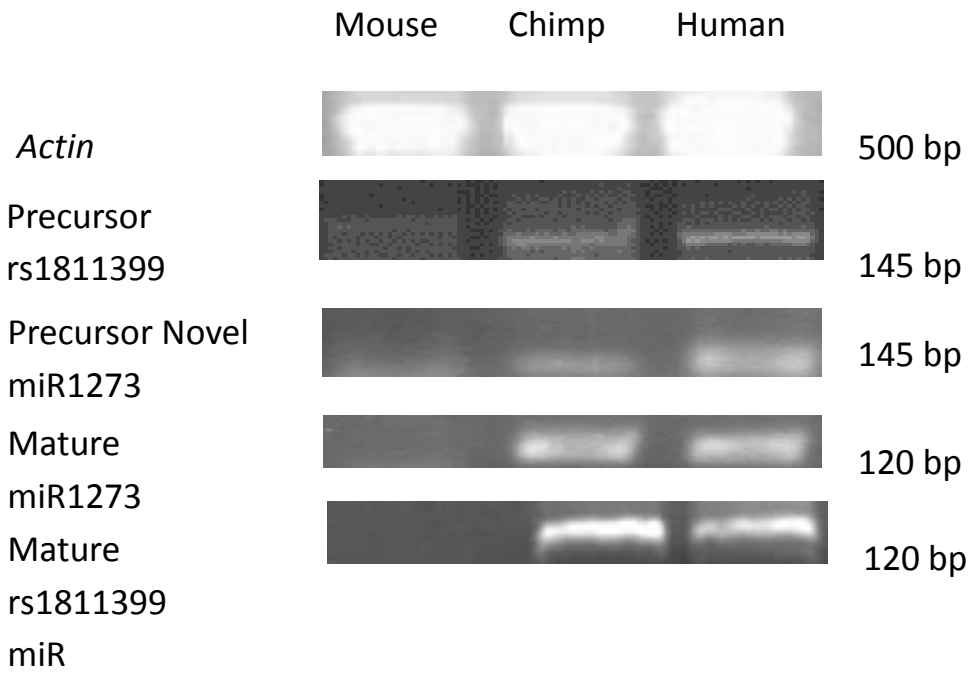


Figure 4.32: Broad conservation of expression (both precursor and mature form) is detectable across primates but not in mouse. Cells were grown to confluency then total RNA was extracted. Once extracted RNA was reverse transcribed using random hexamer primers or using poly-adenylation linker PCR. Primers used to probe the cDNA pool were β-ACTIN-F+β-ACTIN-R, '399-F+'399-R and N1273-F+N1273-R. The mature form was detected by Poly(A) linker PCR using primers: RTQ-UNI+mLET7, RTQ-UNI+m399 and RTQ-UNI+mn1273. The primers noted above were utilised in an attempt to isolate both the precursor and mature forms from across the phyla. Given

the high sequence conservation of miRNA across the species the primers originally designed for human DNA could be expected to identify a conserved region in both mouse and chimpanzee.

The presence of the precursor in human and chimpanzee should not be surprising considering the levels of homology between the two genomes. It also implies a role for the miRNA which may be important in higher primates.

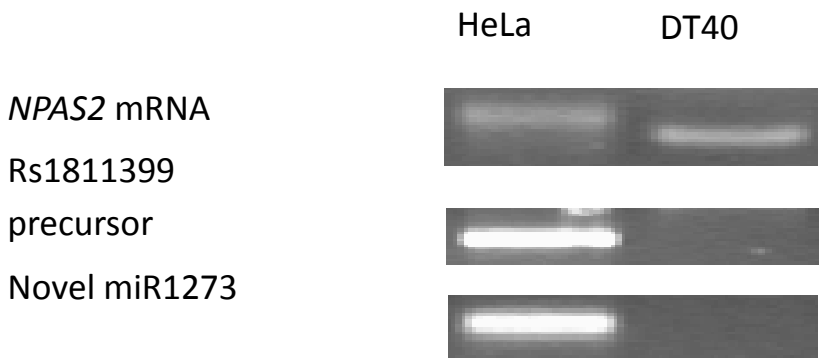


Figure 4.33: Neither miR1273 nor rs1811399 precursor is detectable in chicken DT40 cells.

The absence in chicken of either of the novel miRNA is explained by Fig. 4.35 below.

A

1: n1273locushuman (97 letters) Vs. LATESTGP

Gallus_gallus	0 alignments, 0 hits	[RawResult]
Homo_sapiens	106 alignments, 22 hits	[RawResult]
Mus_musculus	0 alignments, 0 hits	[RawResult]
Pan_troglodytes	104 alignments, 26 hits	[RawResult]

B

Links	Query			Chromosome				Stats			
	Start	*End*	*Ori*	*Name*	*Start*	*End*	*Ori*	*Score*	*E-val*	*%ID*	*Length*
[A] [G] [C]	1	97	+	Chr:2A	101135301	101135397	+	492	9.5e-79	100.00	97

C

1: nrs1811399 (101 letters) Vs. LATESTGP

Gallus_gallus	0 alignments, 0 hits	[RawResult]
Homo_sapiens	1 alignments, 1 hits	[RawResult]
Mus_musculus	0 alignments, 0 hits	[RawResult]
Pan_troglodytes	1 alignments, 1 hits	[RawResult]

D

Links	Query			Chromosome				Stats			
	Start	*End*	*Ori*	*Name*	*Start*	*End*	*Ori*	*Score*	*E-val*	*%ID*	*Length*
[A] [G] [C]	1	101	-	Chr:2A	101138641	101138742	+	487	8.5e-78	99.02	102

E

```
Query:       101 tgtggtcaggtctggaggtcagggcatggtgatcaagcggctgcctgacagtcactgccc 43
                 ||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||
Sbjct: 101138641 tgtggtcaggtctggaggtcagggcatggtgatccagcggctgcctgacagtcactgccc 101138700

Query:        42 agagcttcccttaccataaccttcctcagtagactagaaaag 1
                 |||||||||||||||||||||||||||||||||||||||||
Sbjct: 101138701 agagcttcccttaccataaccttcctcagtagactagaaaag 101138742
```

Figure 4.34: BLAST alignment of two novel miRNA sequences across mouse, human, chimp and chicken genomes. Panel A describes a BLAST alignment performed with the novel miR 1273 sequence which resulted in no hits for mouse or chicken with the top scoring chimp alignment (panel B) being in the same locus as the human gene. Panel C, D and E describe a similar BLAST undertaken with the rs1811399 hairpin. As evident there are no hits in mouse and chicken and only one hit in chimp, again in the *NPAS2*

locus. The one nucleotide discrepancy between the human and chimp sequence can be attributed to the rs1811399 SNP.

Fig 4.36 demonstrate that the rs1811399 miRNA can only be detected within primates. This raises the possibility of the miRNA being a new, primate specific miRNA.

4.19 miRNA cluster is conserved across primates

The detection of the rs1811399 miRNA and miR1273 within the chimpanzee cell line leads us to believe that the cluster may be conserved across the primates. Below is a review of our *in silico* data which hopes to elucidate the relationship.

4.19.1 Genomic location of *NPAS2* across phyla.

Initially genomic co-ordinates of the *NPAS2* gene were extracted from the Ensembl database (Table 4.4). Results proved negative for mouse, dog and chicken.

Table 4.4 below demonstrates the location of *NPAS2* across several species. This was required in order to conduct synteny testing (Fig4.36) to assess the presence of the miRNA cluster within a similar genetic locus across the species. Synteny testing is a mechanism by which a gene or chromosome can be compared across species to detect conservation.

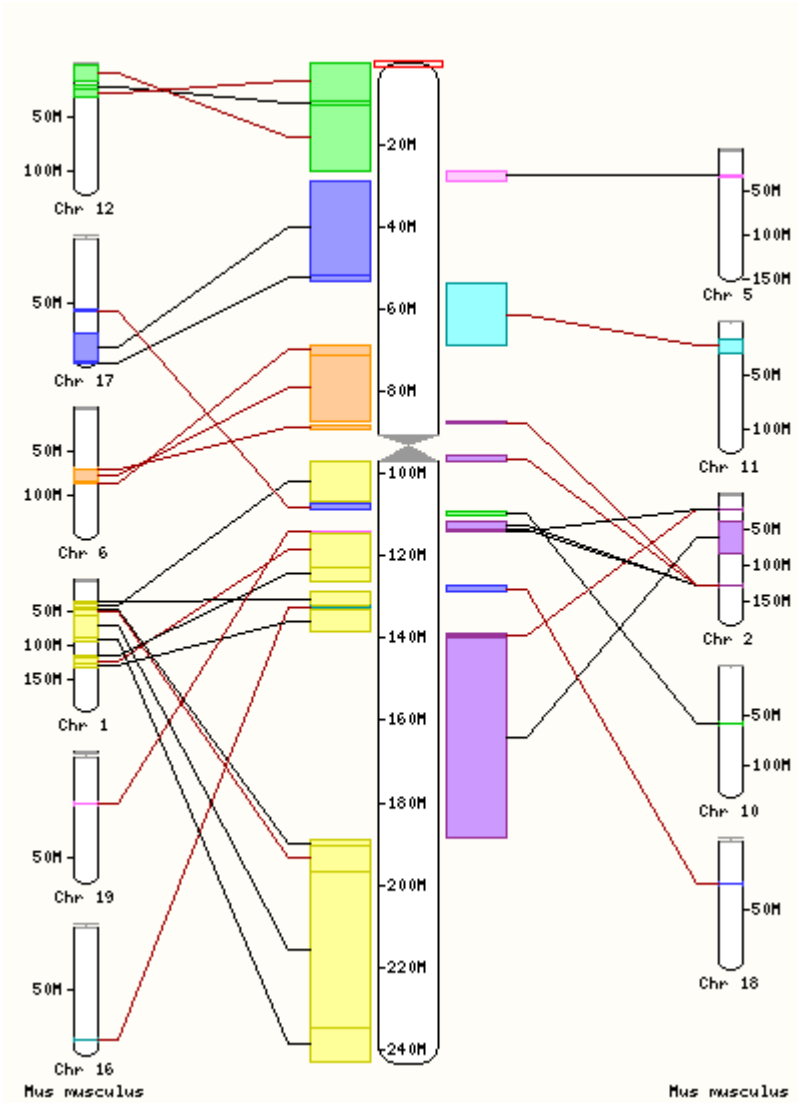
Species	**Genomic location of *NPAS2***
Human	2: 101,436,614-101,613,291
Chimpanzee	2A:101095425-101285563:1
Orangutan	2a:8518713-8702232:1
Gorilla	2a:98256242-98441206:1
Macaque	13:100833971-100928573:1
Mouse	1:39193731-39363234:1
Dog	10:41847885-41998014:-1
Chicken	1:132655000-132723188:1

Table 4.4: Demonstrates the genomic locus of the *NPAS2* gene in several mammals and chicken. This data was extracted so that synteny testing could be conducted. The region that was queried ran from the end of the proposed promoter region down to the end of the rs1811399 miRNA sequence, a region of ~14kb.

Synteny testing was then carried out on positive results (plus mouse) to ensure the loci were conserved in a wider context.

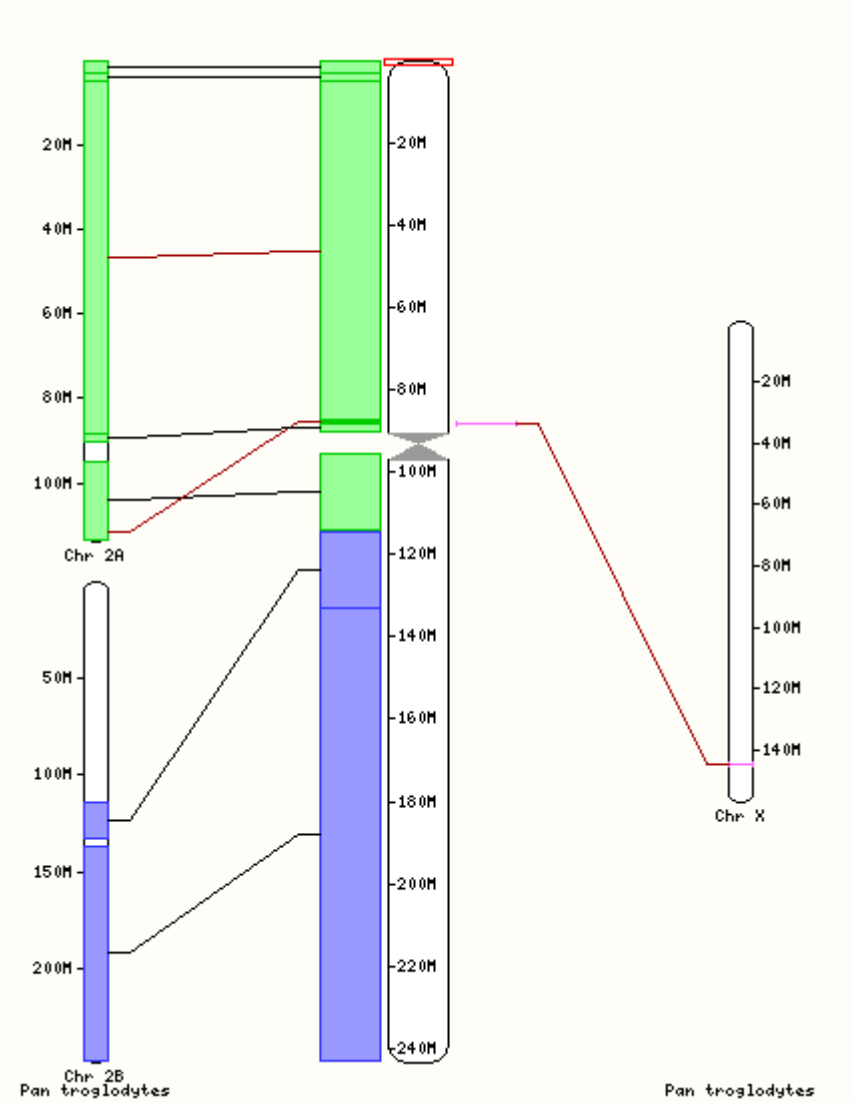
A

Mouse



B

Chimpanzee

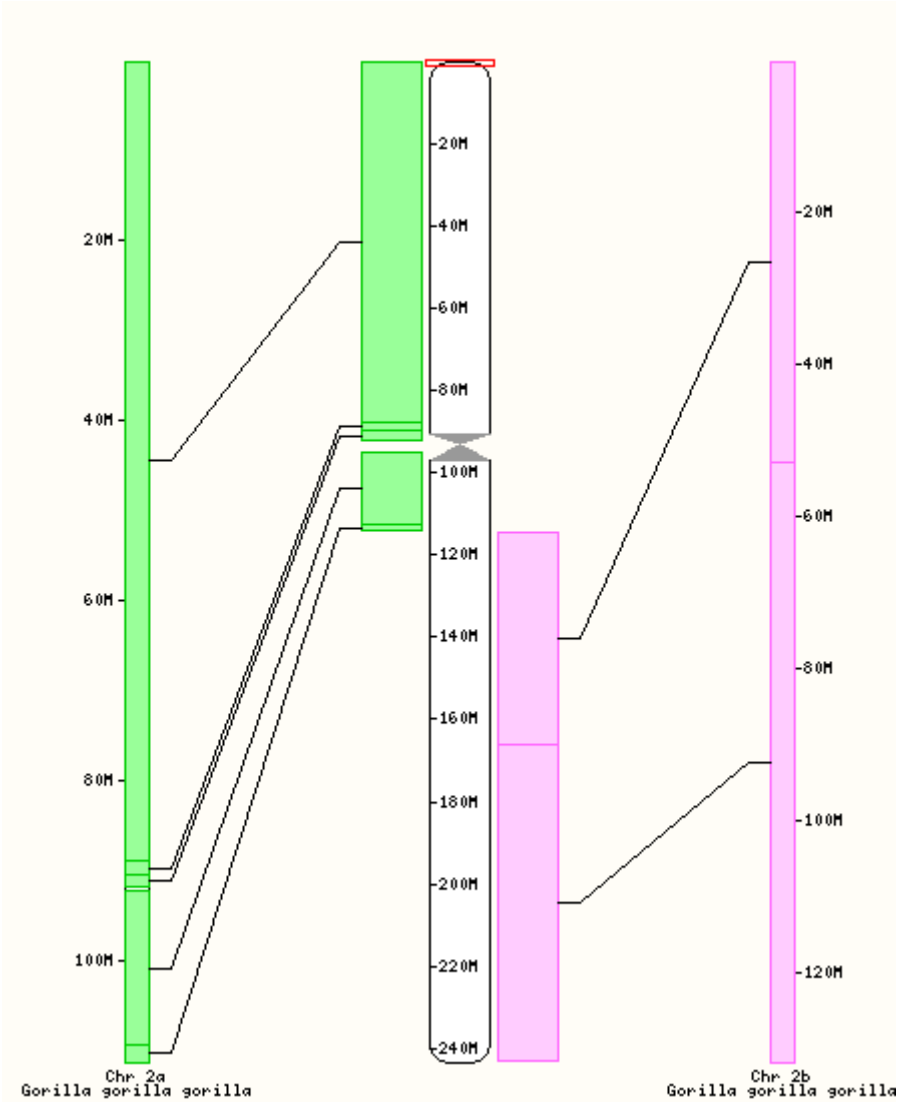


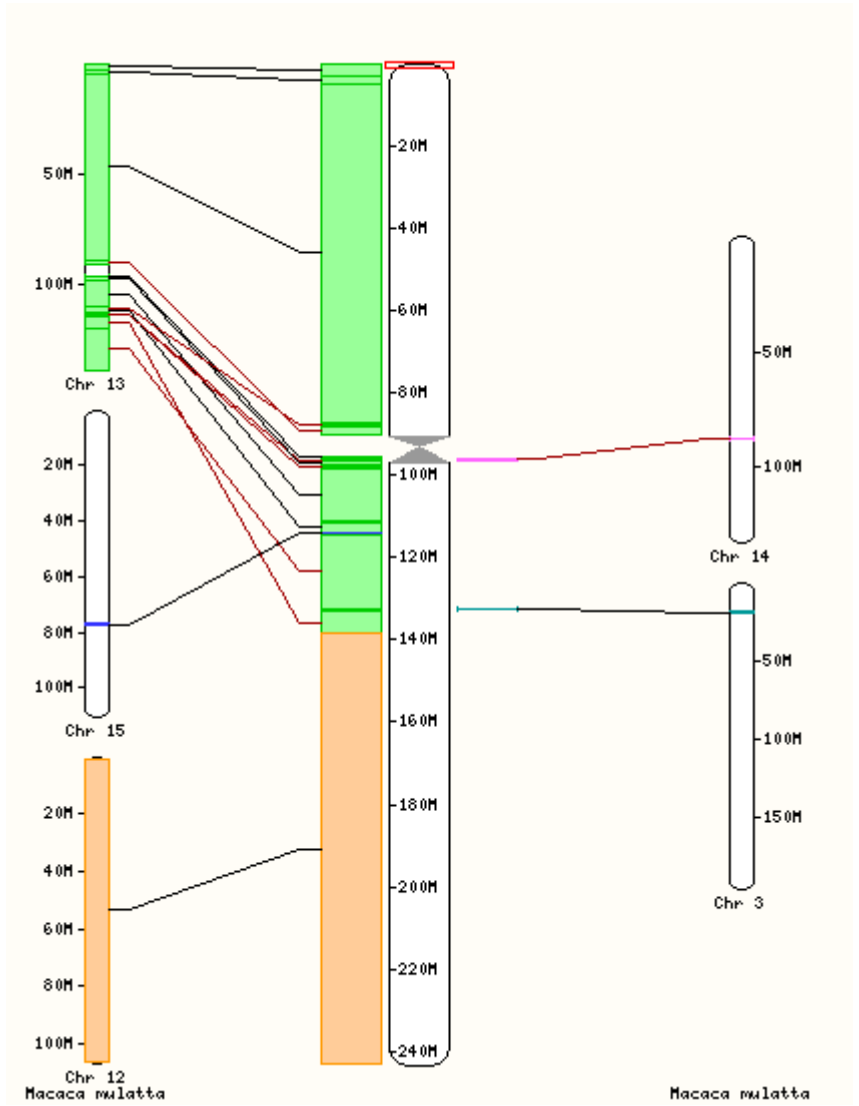
C

D

Gorilla

Macaque





Orang-utan

E

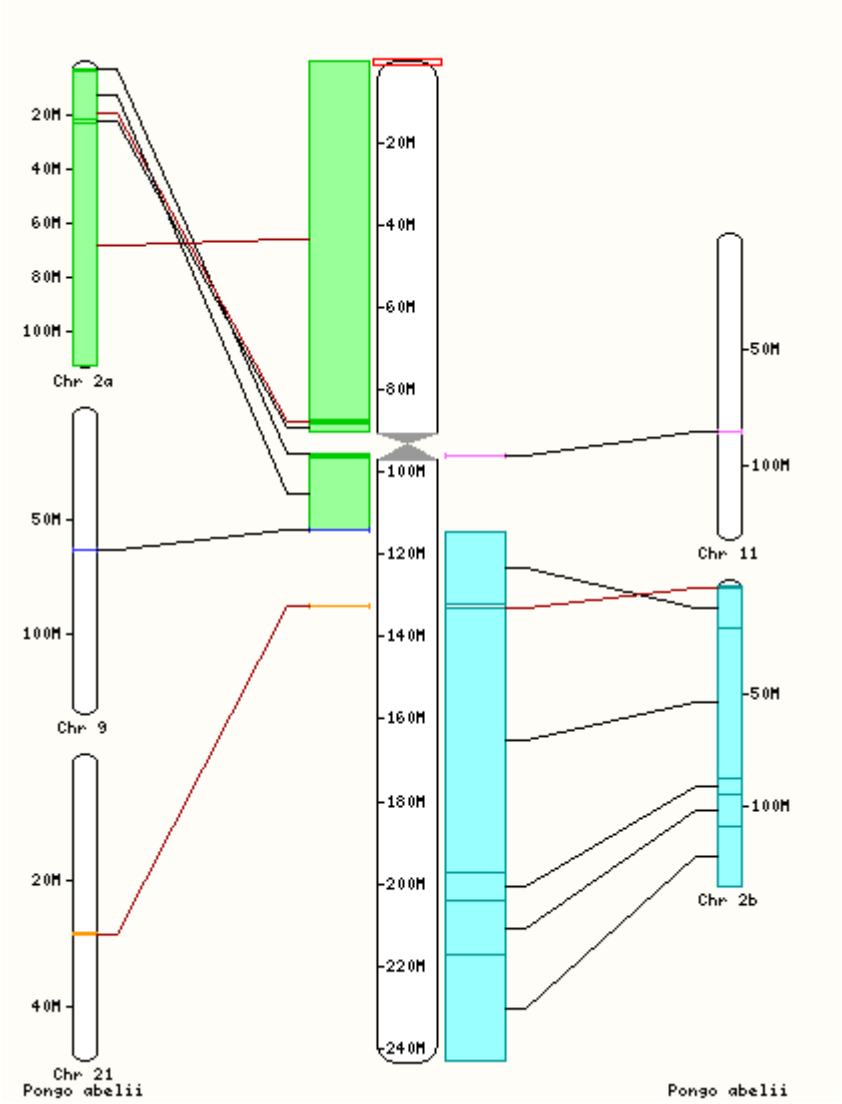

20M
40M
60M
80M
100M
Chr 2a
50M
100M
Chr 9
20M
40M
Chr 21
Pongo abelii
20M
40M
60M
80M
100M
120M
140M
160M
180M
200M
220M
240M
Pongo abelii
50M
100M
Chr 11
50M
100M
Chr 2b

Figure 4.35: Synteny testing carried out using Ensembl data and browser. In each case the human chromosome 2 is situated within the middle of each panel and the corresponding chromosomes which contribute to human chromosome 2 are located on the outsides. The coloured blocks demonstrate the localisation of specific sections of human chromosome 2 onto the respective chromosome of the second species.

The synteny data reveals that human chromosome 2 is a fusion product of two smaller primate chromosomes. This is further revealed by the presence of telomere like repeats close to the fusion site (chr2:114,359,542-114,362,512) demonstrated in Figure 4.37 below.

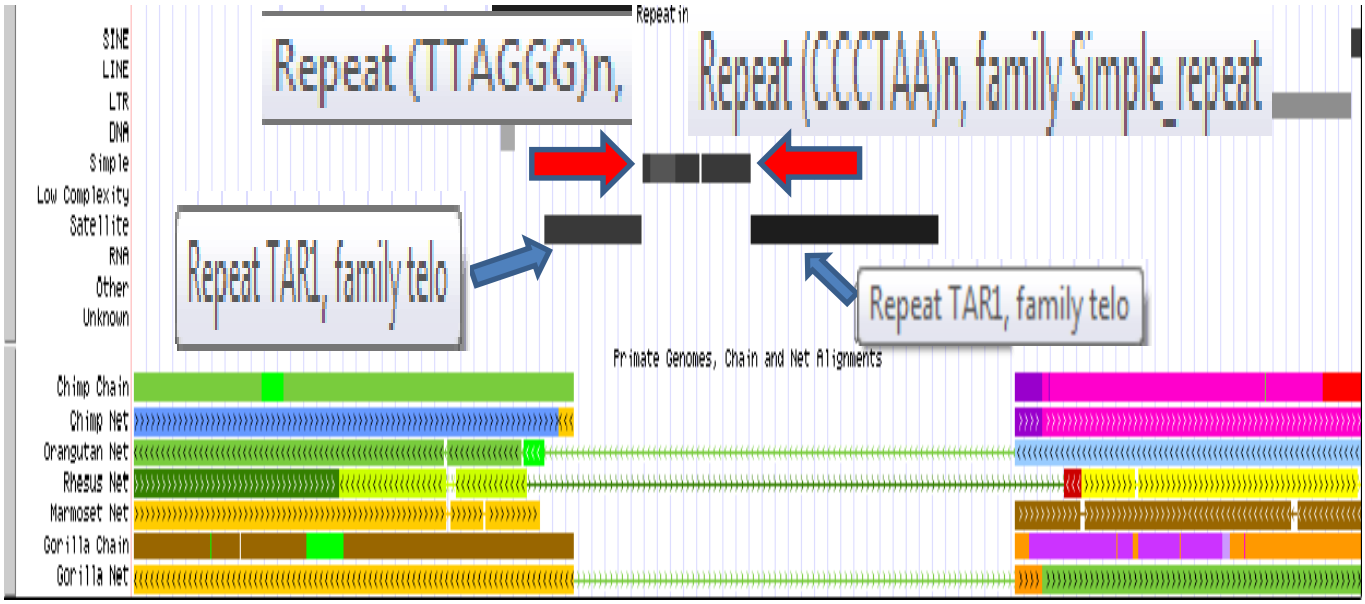


Figure 4.36: Bioinformatic survey of the fusion site within human chromosome 2 chr2:114,359,542-114,362,512. From this data we can see that within human chromosome 2 are two TAR1 repeats which are usually only associated with telomeres. The evidence for this being a fusion site is further strengthened by a region of TTAGGG CCCTAA inverted repeats which cover an area of approximately 800bp. The coloured bars within the bottom of the diagram represent synteny experiments. These

demonstrate the regions flanking the archaic telomere site within human chromosome 2 are homologous to two separate chromosomes in non-human primates.

4.20 Sequence of novel miR-1273 is highly conserved.

A BLAST query of the novel miR1273 locus revealed strong homologies. Below are the alignments of the query primates and the impacts any variants may have on RNA folding.

A

```
human          GGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGAGATCATGCTACT 60
chimp          GGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGAGATCATGCTACT 60
Otang          GGCATGAGAATCGCCTGAACCCGAGAGGTGGAGGTTGCAGTGAGCCGCGATCATGCTACT 60
gorilla        GGCATGAGAATCGCCTGAACCCAAGAGGCGGAGGTTGCAGTGAGCCGAGATCGTGCTACT 60
               ********************.***** ******************.****.*******

human          GCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTC 97
chimp          GCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTC 97
Otang          GTGCTCCAGCCTGGGTGACAGAGTGAGACTCCATCTC 97
gorilla        GCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTC 97
               * ********************* *************
```

B

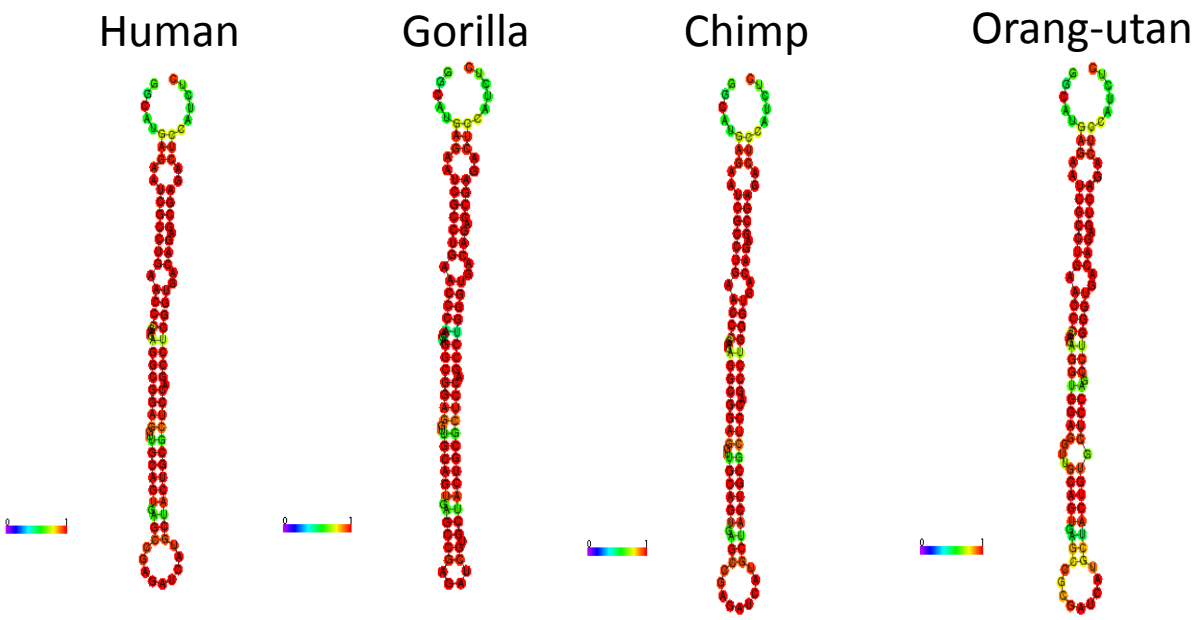


Figure 4.37: Panel A demonstrates the high levels of conservation between primate species for the novel miR-1273 locus. Panel B demonstrates the variations within the RNA structure that each variation could be expected to have on the hairpin. Panel B was produced using Vienna RNAfold.

4.21 rs1811399 sequence is conserved.

A similar experiment was conducted but with the rs1811399 locus.

A)

```
chimp          GGTCAGGTCTGGAGGT-CAGGGCATGGTGATCCAGCGGCTGCCTGACAGTCACTGCCCAG 59
gorilla        GGTCAGGTCTGGAGGT-CAGGGCATGGTGATCCAGCGACTGCCTGACAGTCACTGCCCAG 59
human          GGTCAGGTCTGGAGGT-CAGGGCATGGTGATACAGCGGCTGCCTGACAGTCACTGCCCAG 59
otang          GGTTAGGTCTGGAGGTGCAGGGCATGGTGATCCAGCGGCTGCCTGACAGTCACTGACCAG 60
               *** ************ **************.*****.****************.****

chimp          AGCTTCCCTTACCAT 74
gorilla        AGCTTCCCTTACCAT 74
human          AGCTTCCCTTACCAT 74
otang          AGCTTCCCTTATCAT 75
               *********** ***
```

B)

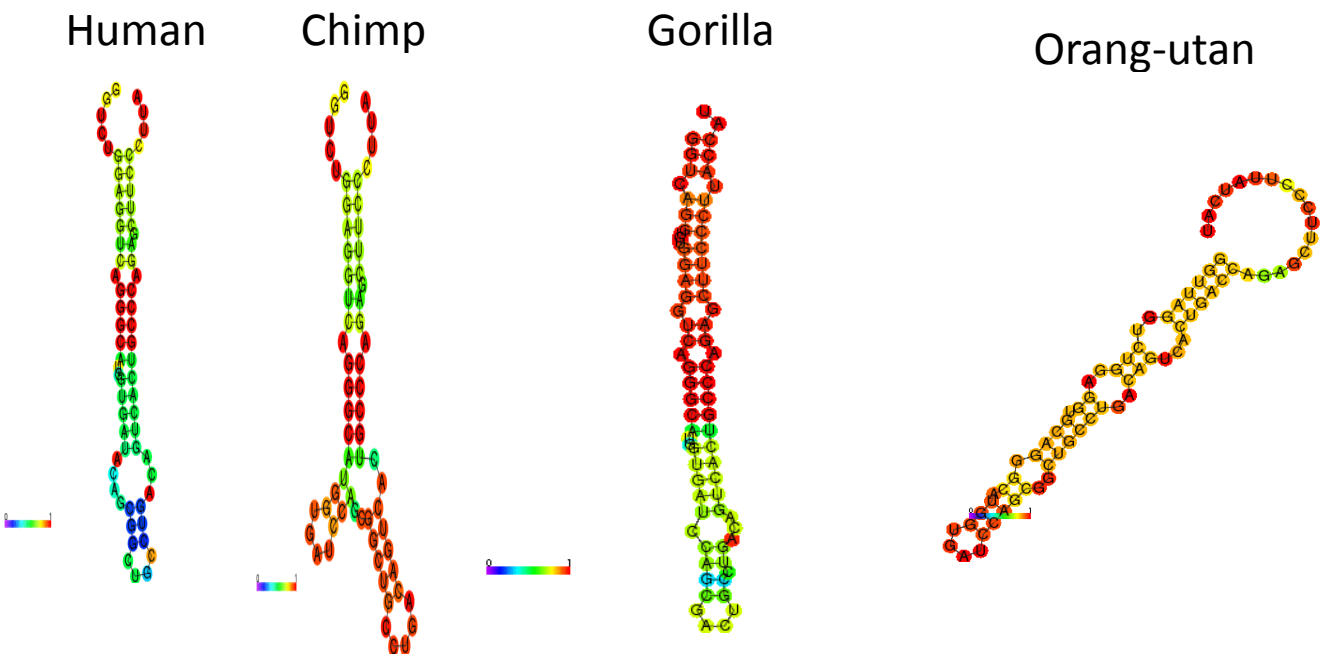


Figure 4.38: Panel A demonstrates alignment of human rs1811399 locus with that of other primates. Chimp is identical to human rs1811399C sequence whilst gorilla has one

extra variation. Orangutan is the species with the lowest conservation of the hairpin. Panel B demonstrates the RNA folding of the locus in each of the species. This figure was produced using Vienna RNAfold.

As the miRNA precursors' sequences are conserved across the primates it should be possible to produce a tree identifying the divergence and evolutionary history of the region.

4.22 Phylogenetic evolution of the two miRNA precursors.

Utilising Dereeper *et al* (2008)'s phylogenetic tree software it is possible to construct an evolutionary pathway for each of the miRNA precursors. This will allows us to deduce when each species diverged from the other in terms of this miRNA cluster.

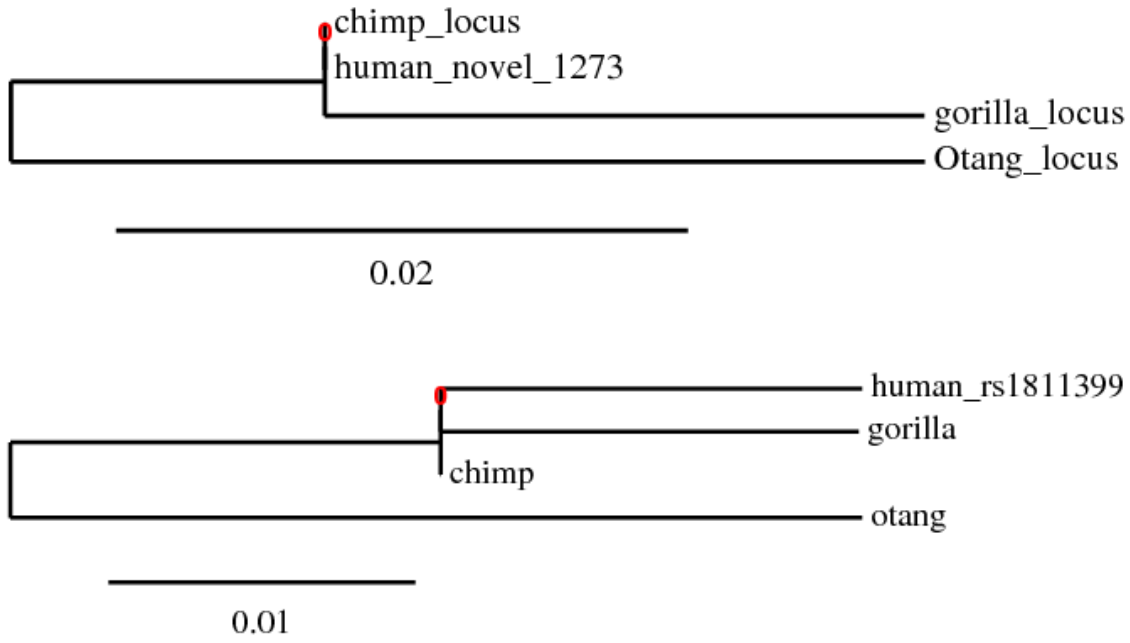


Figure 4.39: From both trees (novel 1273 top tree, rs1811399 bottom tree) we can see that orangutan has diverged from the same path as the three other primates whilst they have followed similar evolutionary paths. This is reflected in the larger branch lengths for the orange-tan line which symbolises a greater rate of genetic diversity.

The above results are in keeping with our understanding of primate evolution.

5. Impact of SNP rs1811399 which may be linked with autism.

Previous work has identified the SNP rs1811399 as a variant possibly linked with the autism phenotype (Nicholas *et al*, 2008). In this study the circadian clock genes (*NPAS2, CLOCK, BMAL, CRY1, CRY2, PER1, PER2, PER3, TIMELESS, CK1* and *DBP*) of 110 autistic patients including their parents were sequenced. Two SNP were found to be significantly associated with autism in *PER1* (rs885747 and rs6416892) and one in *NPAS2* (rs1811399).

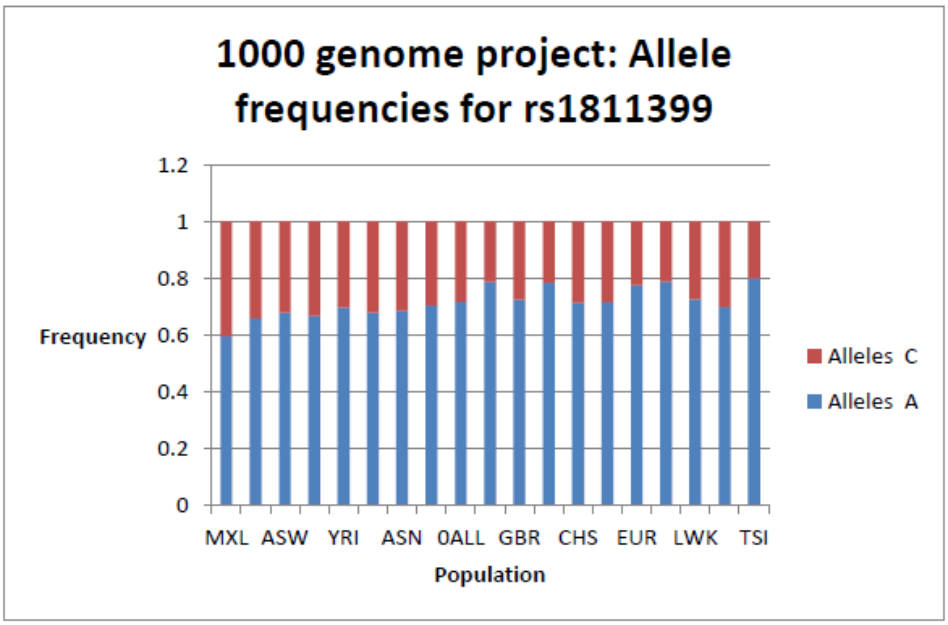
The mechanism by which this single nucleotide polymorphism contributes to the phenotype is currently unknown. Unlike the SNP rs2305160 in *NPAS2*, which causes the substitution of threonine for alanine which causes an increased risk of non-Hodgkins lymphoma and breast cancer (Zhu *et al*, 2007: Yi *et al*, 2010), the SNP rs1811399 is intronic.

5.1 Population statistics.

SNP allele frequencies are known to vary from population to population. The cause of such variance can be due to numerous factors: human migrations, natural disaster resulting in small surviving population or even as a positive adaption to the environment. Xin Yi *et al* (2010) for example note that within high altitude populations (Tibetan) positive selection has introduced and maintained SNP within the *EPAS1* gene which increases red blood cell production. Two key databases will be mined to extract population statistics for the rs1811399 SNP: 1000 genomes project and the HapMap project (Fig 5.1).

Population legend is as follows:

Population Code	Population Description	Super Population Code	Population Code	Population Description	Super Population Code	Population Code	Population Description	Super Population Code
CHB	Han Chinese in Bejing, China	ASN	KHV	Kinh in Ho Chi Minh City, Vietnam	ASN	STU	Sri Lankan Tamil from the UK	SAN
JPT	Japanese in Tokyo, Japan	ASN	CEU	Utah Residents (CEPH) with Northern and	EUR	ITU	Indian Telugu from the UK	SAN
CHS	Southern Han Chinese	ASN		Western European ancestry		BEB	Bengali from Bangladesh	SAN
CDX	Chinese Dai in Xishuangbanna, China	ASN	TSI	Toscani in Italia	EUR	PUR	Puerto Ricans from Puerto Rico	AMR
GWD	Gambian in Western Divisons in The Gambia	AFR	FIN	Finnish in Finland	EUR	CLM	Colombians from Medellin, Colombia	AMR
MSL	Mende in Sierra Leone	AFR	GBR	British in England and Scotland	EUR	GIH	Gujarati Indian from Houston, Texas	SAN
ESN	Esan in Nigera	AFR	IBS	Iberian population in Spain	EUR	PJL	Punjabi from Lahore, Pakistan	SAN
ASW	Americans of African Ancestry in SW USA	AFR	YRI	Yoruba in Ibadan, Nigera	AFR			
ACB	African Carribbeans in Barbados	AFR	LWK	Luhya in Webuye, Kenya	AFR			
MXL	Mexican Ancestry from Los Angeles USA	AMR	PEL	Peruvians from Lima, Peru	AMR			



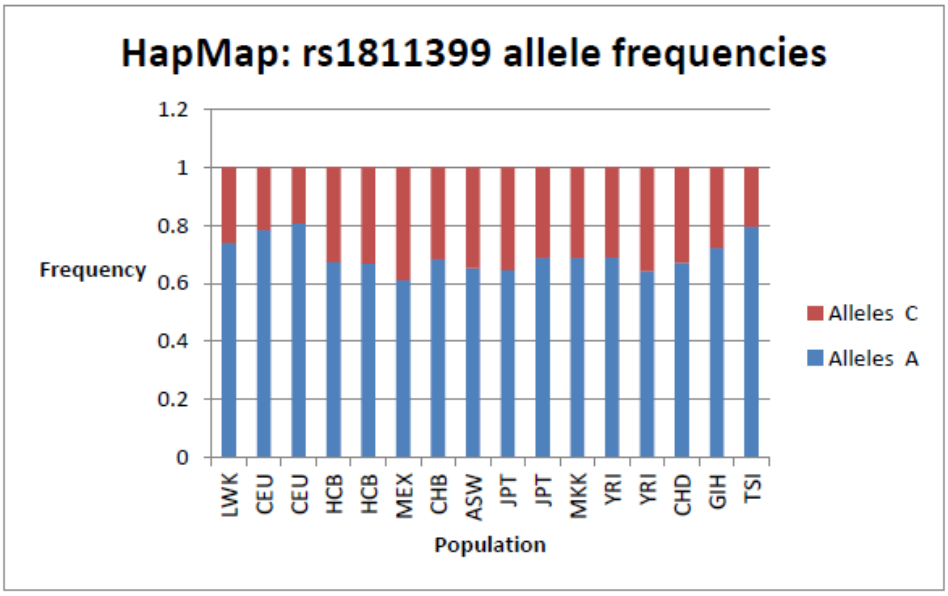
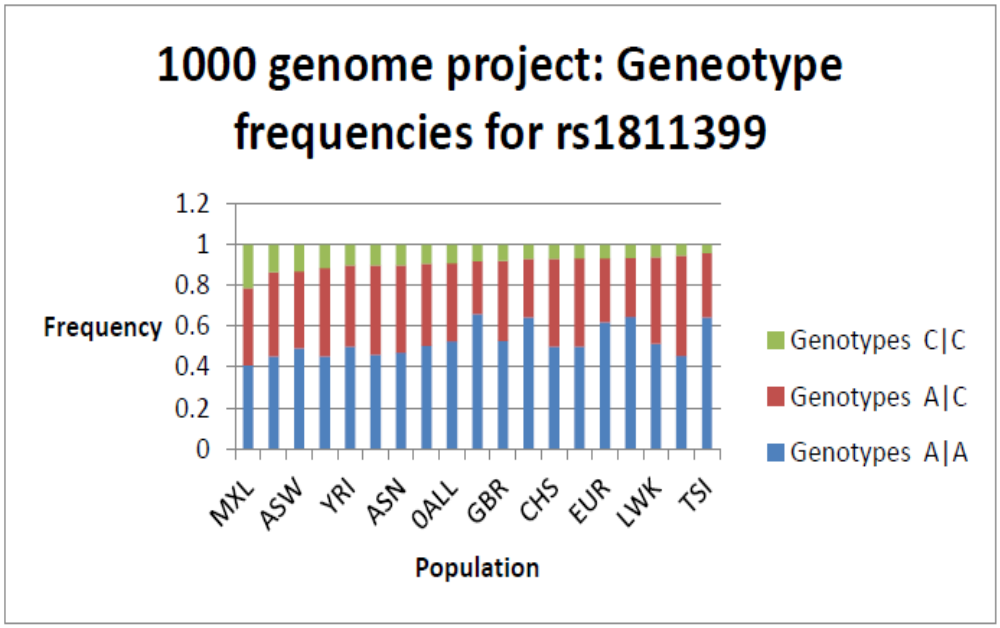


Figure 5.1: Allele frequencies for rs1811399 extracted from two genomic population statistic databases. The databases were 1000 Genomes Project phase 3 (http://www.1000genomes.org/) and HapMap3 (http://hapmap.ncbi.nlm.nih.gov/).

The hypothesis that rs1811399C>A can solely lead to autism is not supported by the allele frequencies. By collating all the data from both surveys one can reach the conclusion that globally approximately 29% of all alleles will be C (Figure 5.1). Because the C allele has a reduced frequency in autistic children (17.89%) compared to their parents (22.94%)

(Nicholas *et al.*, 2007), it is likely that the C allele exerts a protective function. Hence the reduced maturation of the micro RNA caused by the C nucleotide (rs1811399) may have a beneficial effect. Figure 5.2 clearly indicates the bias away from homozygous C|C genotypes in the general population.



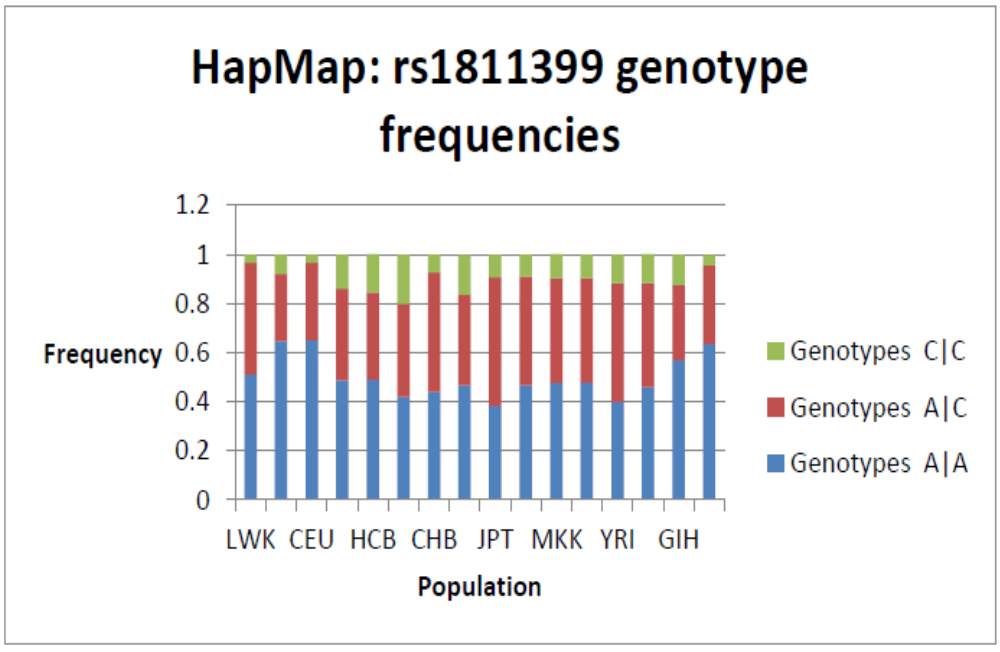


Figure 5.2: Genotype figures for varying populations.

There is also a slight sexual dichotomy with regards to prevalence of the homozygous C genotype: 7.6% of males and 10.2% of females. The significance of this however is doubtful and nowhere near as apparent as the dichotomy in autism incidence rates as referenced in the introduction.

The ancestral allele at the rs1811399 locus was C. The A allele has since replaced the C allele in the majority of cases (Fig.5.1) and cases of homozygous C|C are rare (Fig.5.2). It is postulated that the A allele confers some selective advantage and has begun to replace the C allele. Nicholas *et al* (2008) postulates that the C allele is in itself detrimental and that selection pressure is favouring towards the A allele. That the C allele is also under transmitted from parents to offspring might also be evidence towards negative selection pressure against the C allele.

There exists the possibility that rs1811399 is linked with a second genetic variation (such as SNPs in linkage disequilibrium) which can contribute to its effect. It is possible to identify these linked SNP and perform similar analyses as have been done for rs1811399.

5.2 Linkage disequilibrium.

Linkage disequilibrium is the association of alleles at more than one loci in a non-random manner (Knight, 2009).

Rs1811399 in the general population is known to be in linkage with 23 other SNP within a 50kb locus (Table 5.1). None of these linked SNPs have been directly implicated with any phenotype. Linked SNPs can be given two statistical measures to describe their linkage: D' that is a measure of dependency (a figure of 1 implies you will always find one SNP if you have the second present) and R squared, which demonstrates the closeness of fit of the data with linear regression curve of available data.

Variation	Location	Distance (bp)	r^2	D'
rs1118509	2:101476892	2122	0.849471	1
rs11904563	2:101494480	15466	0.968868	1
rs12472319	2:101487457	8443	0.827519	1
rs12472321	2:101487468	8454	0.827519	1
rs13011414	2:101495624	16610	0.827519	1
rs13032665	2:101483815	4801	0.820203	0.969714
rs13034472	2:101481119	2105	0.849471	1
rs1369481	2:101511959	32945	0.874064	0.955196
rs1435511	2:101478658	356	0.875137	0.999999
rs2043534	2:101480885	1871	0.849471	1
rs2082816	2:101477085	1929	0.849471	1
rs2871389	2:101495174	16160	0.827519	1
rs34873464	2:101486641	7627	0.827519	1
rs4081946	2:101500089	21075	0.830929	0.911553
rs6542996	2:101485827	6813	0.875137	0.999999
rs6718451	2:101493549	14535	0.875137	0.999999
rs6750976	2:101496109	17095	0.827519	1
rs6759386	2:101483372	4358	0.875137	0.999999
rs7564936	2:101491972	12958	0.957981	0.999999
rs7582455	2:101471707	7307	0.875137	0.999999
rs7590391	2:101483103	4089	1	1
rs930309	2:101484534	5520	1	1
rs983287	2:101480401	1387	0.875137	0.999999

Table 5.1: SNPs that are within linkage disequilibrium with rs1811399 within a 50kb locus. D is a measure of dependency, the closer the figure is to 1 the greater the possibility of finding one allele within a sample if the second exists. R squared is the coefficient of determination or simply put how close the data fits the linear regression curve of data. Both R squared and D are different ways of identifying how likely you are to discover one allele at a particular locus if you have another at a different locus.

Of the 23 other SNPs, 6 are registered as occurring within regulatory regions (Table 5.2).

Variation Name	Chr	Position on Chromosome (bp)	Regulatory Feature Stable ID	Regulatory Feature Allele String	Regulatory Feature Consequence Type
rs12472319	2	101487457	ENSR00001544767	C/T	Regulatory_region_variant
rs12472321	2	101487468	ENSR00001544767	C/T	Regulatory_region_variant
rs13011414	2	101495624	ENSR00000595332	C/A	Regulatory_region_variant
rs4081946	2	101500089	ENSR00000595335	A/G	Regulatory_region_variant
rs6750976	2	101496109	ENSR00000595332	C/T	Regulatory_region_variant
rs7564936	2	101491972	ENSR00001544768	T/G	Regulatory_region_variant

Table 5.2: Table demonstrating potential significant linked SNPs. All 6 of these are variant alleles which are located within regulatory regions of the *NPAS2* gene.

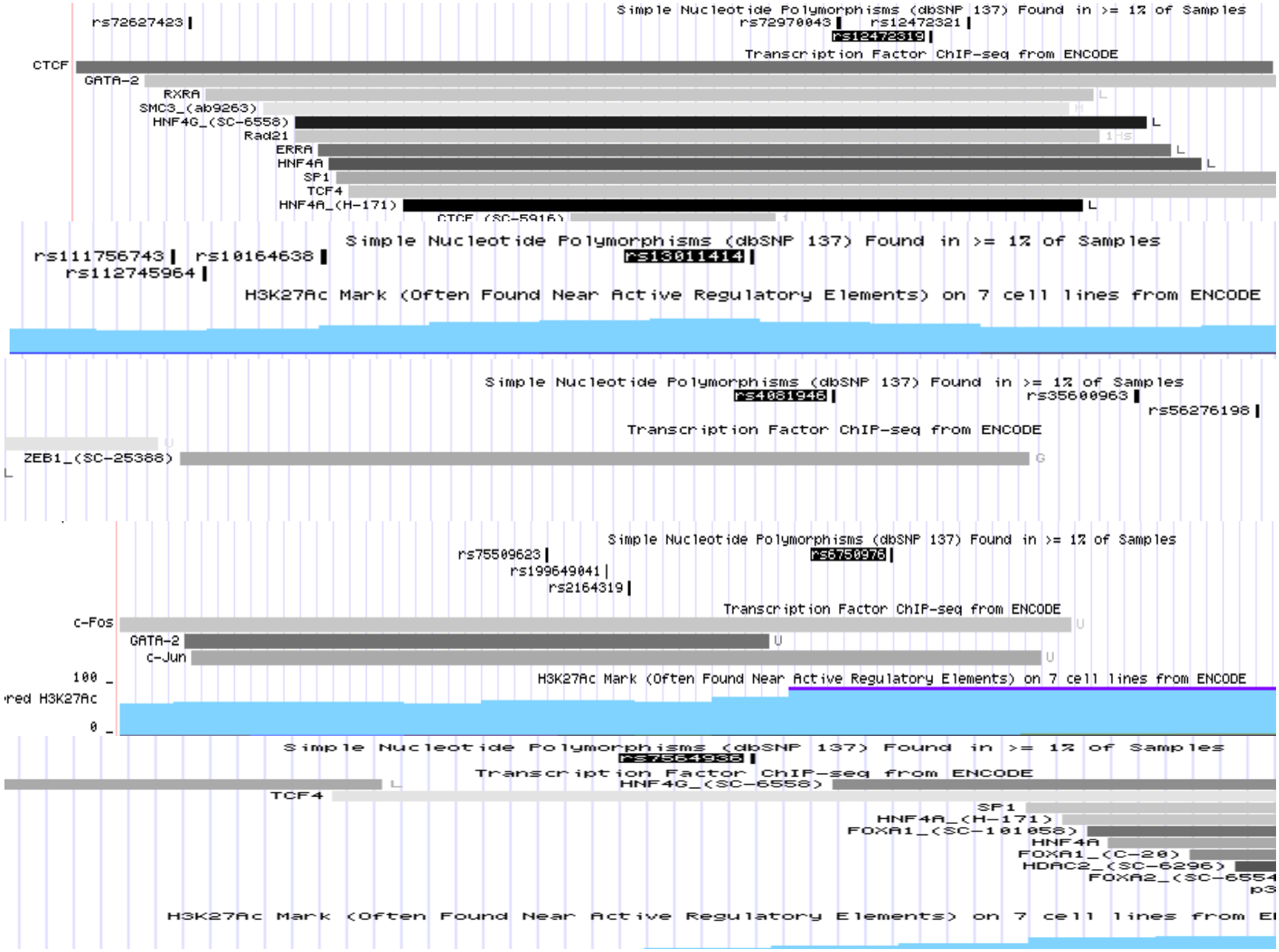

A
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
rs72627423
rs72970043
rs12472321
rs12472319
Transcription Factor ChIP-seq from ENCODE
CTCF
GATA-2
RXRA
SMC3_(ab9263)
HNF4G_(SC-6558)
Rad21
ERRA
HNF4A
SP1
TCF4
HNF4A_(H-171)
B
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
rs111756743
rs10164638
rs13011414
rs112745964
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
C
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
rs4081946
rs35600963
rs56276198
Transcription Factor ChIP-seq from ENCODE
ZEB1_(SC-25388)
D
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
rs75509623
rs6750976
rs199649041
rs2164319
Transcription Factor ChIP-seq from ENCODE
c-Fos
GATA-2
c-Jun
100
red H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
E
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
rs75649365
Transcription Factor ChIP-seq from ENCODE
HNF4G_(SC-6558)
TCF4
SP1
HNF4A_(H-171)
FOXA1_(SC-101058)
HNF4A
FOXA1_(C-20)
HDAC2_(SC-6296)
FOXA2_(SC-6554
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from

Figure 5.3: Genomic context of all linked SNPs. A) rs12472319 and rs12472321 SNP with overlapping transcription factor binding sites. This locus has a particularly dense collection of experimentally proven transcription factor binding sites (Encode ChIP-seq) but neither of the SNPs negatively impacts any of these factor binding sites. B) rs13011414 is embedded in an open chromatin region marked by histone3k027 acetylation. Impacts of SNPs in such loci are difficult to determine. The greatest potential impact it could have on the chromatin state would be if it were part of a CpG island but this is not the case. C) rs4081946 again is overlapping a proven transcription factor binding site but is not predicted to have any influence on the binding. D) rs6750976 is located within the binding sites of two transcription factors and an area of high chromatin open-ness. It does not seem to affect the loci however. E) rs7564936 cannot be demonstrated to interact with the TCF4 binding.

All we can conclude from this exercise is that none of the linked SNPs are directly involved with any aberrant transcription factor binding which might cause a phenotype. Thus our focus must shift to viewing rs1811399 as the contributing factor to a phenotype.

5.3 SNP impact on miRNA processing.

It is established within the literature that SNP can have dramatic impact upon processing of miRNA (Duan, Pack and Jin, 2007). The mechanism by which SNP can influence maturation of a miRNA is twofold: Either Drosha processing at the stem loop-ssRNA junction is influenced (Duan, Pack and Jin, 2007) or Dicer processing is (Sun *et al*, 2009). It has also been reported that the converse is also true and that SNPs in miRNA genes can increase efficiency of biogenesis such as the case of a G>A substitution at nucleotide 4 within miR-510 (Sun *et al*, 2009).

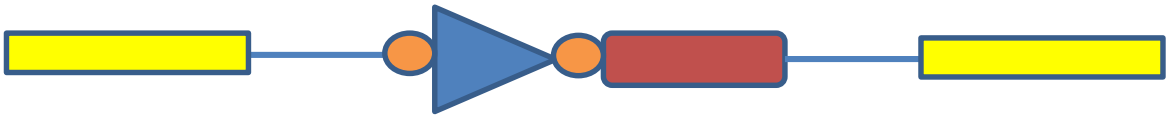
In order to assess the viability of this theory with regards to our novel *NPAS2* rs1811399 miRNA *in vivo* assays will be performed.

5.3.1 *In vitro* analysis of SNP impact on miRNA processing

As the DT40 cell lines had been demonstrated to not express the precursor under investigation, it was decided to use the cells as a model for studying the influence of the SNP *in vivo*.

The hairpin sequence was initially cloned from HeLa genomic DNA before using fusion-PCR to produce two hairpins, one for each allele. Each hairpin was then cloned into the pJE28 plasmid supplied by Dr John Eykelenboom. The pJE28 vector has been designed to integrate genetic material into the chicken *Ovalbumin* gene which is not expressed in DT40. The targeting of the construct to the *Ovalbumin* locus is achieved by the presence in the construct of two targeting arms, which are homologous with regions in the genomic region. In order to drive expression of the hairpin, the human H1 promoter was excised from pSuperior and integrated upstream of the hairpin sequence. In order to control the expression of the hairpins a tetracycline responsive element was implemented into the construct; when co-transfected with a Tet-on plasmid, this will allow for the expression of the precursor to be switched on. Constructs of both alleles were then transfected into wild type DT40 cells using electroporation or chemical transfection and expression of processing was detected using protection assay and PCR.

Figure 5.4 below illustrates the cloning strategy utilised in these experiments.



Yellow: Ovalbumin locus targeting arms.
Orange: Tetracycline response elements
Blue: Human H1 promoter
Red: miRNA hairpin

Figure 5.4: Schematic of cloning strategy. Targeting plasmid pJE28 was kindly provided by Dr John Eykelenboom from the University of Galway. To this plasmid a human H1 promoter region (blue triangle) with two tet-response elements (orange circles) was cloned into the Nhe1 site, fused to this was the rs1811399 hairpin (red rectangle).

The initial strategy called for a tetracycline inducible miRNA however upon transfection it was noted that incubation with tetracycline complete cell death was achieved. Initially it was assumed that the miRNA was inducing this lethality however a serendipitous finding by Dr Ellen Vernon who was working on the same cell line discovered that prior to our receiving the cell line a TET-OFF system had been integrated into the cell and upon induction with tetracycline several important housekeeping genes were silenced. Knowing this, the system was redesigned around the pcDNA3.1 plasmid.

A pcDNA 3.1 puromycin vector was received from colleagues at University of Galway. This plasmid allowed for positive selection of transfected clones using 1mg/ml of puromycin. In order to transfect the cells $1x10^7$ DT40 cells centrifuged and suspended in 1ml of PBS. The construct was then incubated with the suspended cells on ice prior to electro-transfection at 400V. Cells were rested for 24h before plating on 96 wells with puromycin antibiotic for selection of positive clones.

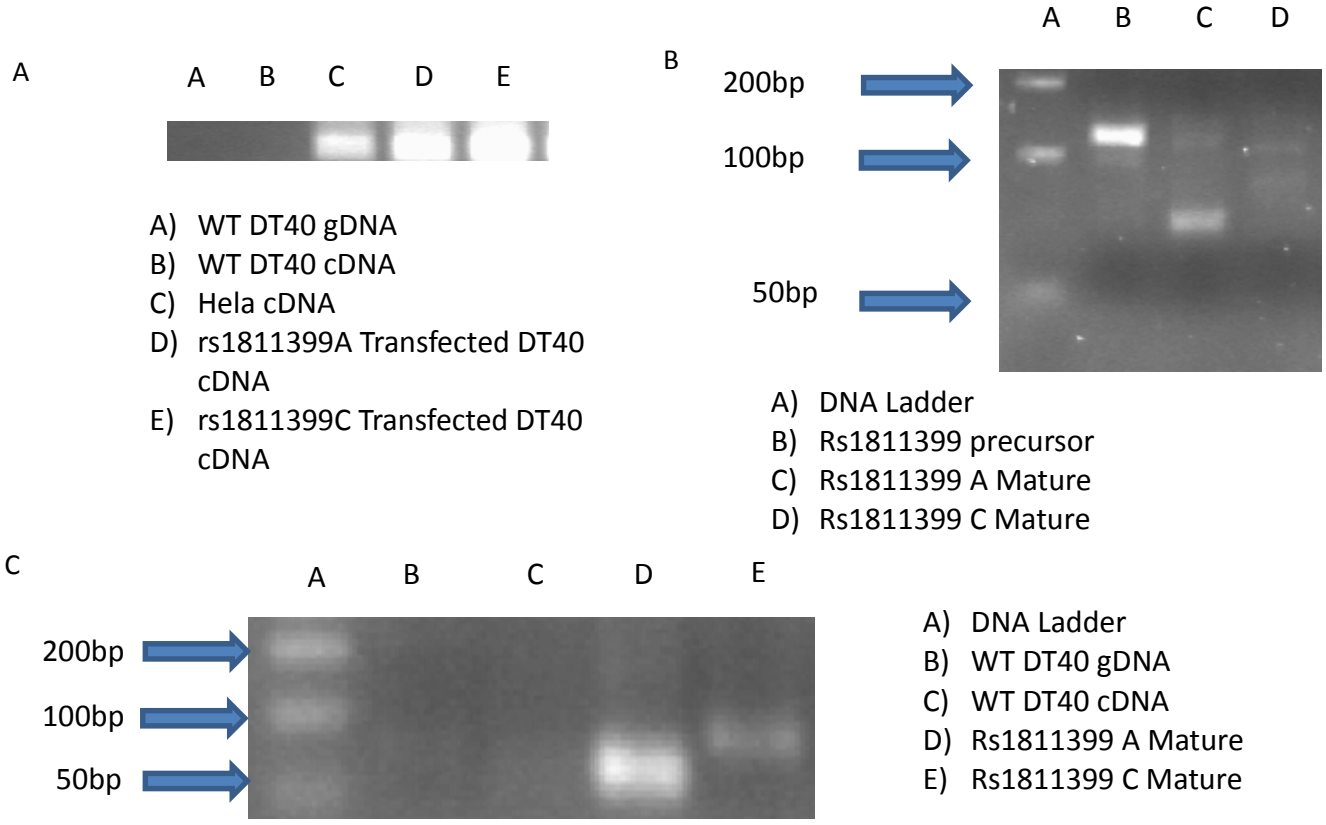


Figure 5.5: DT40 in vitro method of detecting SNP impact on miRNA biogenesis. (A) Demonstrates the absence of any visible precursor in both genomic DNA and cDNA library of un-transfected DT40 cells. After transfection it is possible to isolate the precursor from cDNA (lane D and E panel A). (B) Demonstrates the impact of maturation the rs1811399 SNP has on miRNA biogenesis via Poly(A) linker PCR. (C) demonstrates the maturation of the rs1811399 hairpin in a Poly(A) linker PCR during which 5ug of total RNA was used as starting material as opposed to 1ug, Size variation in (C) lanes D and E could arise due to the incorrect processing of the hairpin. Semi-quantitative image analysis demonstrate relative intensity of 86108.352 for lane D in panel C and an intensity of 9525.087 for the mature form in lane E.

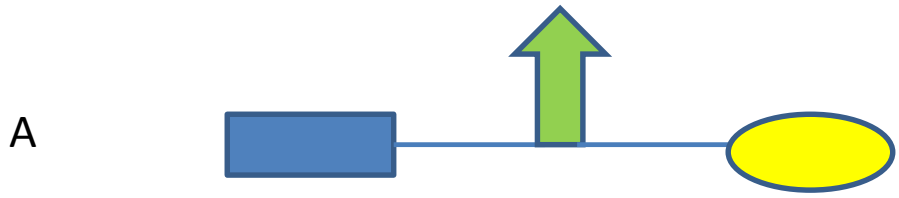
It is suggested in Fig.5.5 that the rs1811399 C appears to interfere with maturation of the mature form *in vivo*. However, a repeat experiment (Fig 5.6 below) suggests that the levels of the mature form present in the C sample might have been too low to be detected.

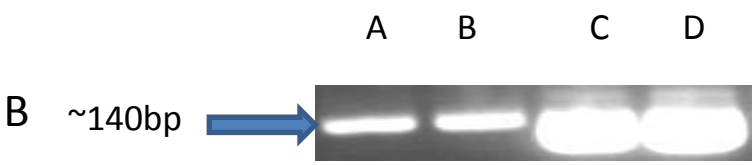
Whilst the miRNA processing machinery is the same in chickens as it is in humans it is important to test this result in human cells.

5.3.2 *In vitro* analysis in human (HeLa) cell lines.

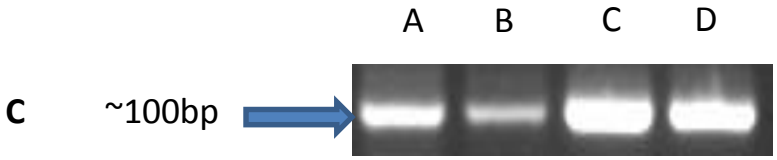
Current experimental studies have gone some way to addressing the mechanism by which the rs1811399 C could be contributing to a disease phenotype. To increase the relevance of the experimentation hitherto carried out it was decided to continue the work established within the DT40 cell line in HeLa cells.

The hairpin sequence was initially cloned from HeLa genomic DNA before using fusion-PCR to produce two hairpins, one for each allele. The psi-RNA plasmid is a commercial product which utilises an h7sk promoter to drive expression of siRNA. The hairpins are cloned into the vector using blunt end cloning. This vector allowed for selective transfection using zeocin to produce stable clones (Fig 5.6A).





A) Let-7a rs1811399 A cDNA
B) Rs1811399 precursor untransfected
C) Rs1811399 precursor in cells transfected with rs1811399A
D) Rs1811399 precursor in cells transfected with rs1811399C



A) Let-7a mature
B) Rs1811399 mature untransfected
C) Rs1811399 mature in cells transfected with rs1811399A
D) Rs1811399 mature in cells transfected with rs1811399C

Figure 5.6: RT-PCR and Poly(A) linker PCR analysis of impact of rs1811399 SNP on maturation within human cell line. (A) A schematic of the psiRNA plasmid used in the human cell lines. The blue box represents the h7sk promoter, green arrow is the precursor hairpin and the yellow oval is the zeocin resistance cassette. (B) Demonstrates the relative difference in amount of precursor that is detectable in cells pre-transfection

(lane B) and post-transfection (lane C and D). Semi-quantitative analysis of the gel was carried out using ImageJ. The relative intensity of the precursor band in untransfected cells was 12352.66 whilst the intensity of bands post-transfection was 21618.246

21703.296 for lanes C and D respectively. This trend is continued in (C) where Poly(A) linker PCR data is shown. Note the relative abundance of PCR product in lane C (rs1811399 A vector) versus product in lane D (rs1811399 C vector). Semi-quantitative analysis of panel C reveals an intensity of 8543.903 for the mature form in un-transfected cells against an intensity reading of 26713.602 and 24687.773 for lanes C and D respectively.

From Fig5.6C it appears that rs1811399 C may be under-processed. Caution should be used when interpreting these results given that the variance may arise due to copy number variations during transfection.

5.4 RNA Editing

Whilst the structural requirements for the editing site varies between the type of editing in question and the enzyme responsible, one factor is universal, the requirement for dsRNA. As we know that the locus surrounding rs1811399 forms a double stranded helix and that a cluster of 9 A to I editing has been discovered downstream of the SNP (chr2:101,493,692-101,493,860) it was decided to investigate the potential impact that the SNP might have on RNA editing.

Current understanding of post-transcriptional regulation of miRNA maturation demonstrates the presence of a regulatory network in which several factors may be responsible for regulation of specific mature miRNA (Thomson et al, 2006). One mechanism by which cells may regulate miRNA expression is by RNA editing (Luciano *et al*, 2004). As covered in

section 1.4.4 RNA editing is undertaken by deaminase enzymes, which modify nucleotides in an RNA sequence so that they differ from the genomic sequence. In the scope of miRNA biogenesis, a polymorphism or nucleotide substitution can lead to a conformational change resulting in impaired biogenesis by DICER processing (Fig. 1.13). This mechanism is implicated in the regulation of miR-151.

miR151 is known to undergo RNA editing at nucleotides 45 and 48 (numbered from 5' end). Editing at either one of the two sites is enough to impair processing by DICER even though no substantial conformational change occurs to the terminal hairpin loop.

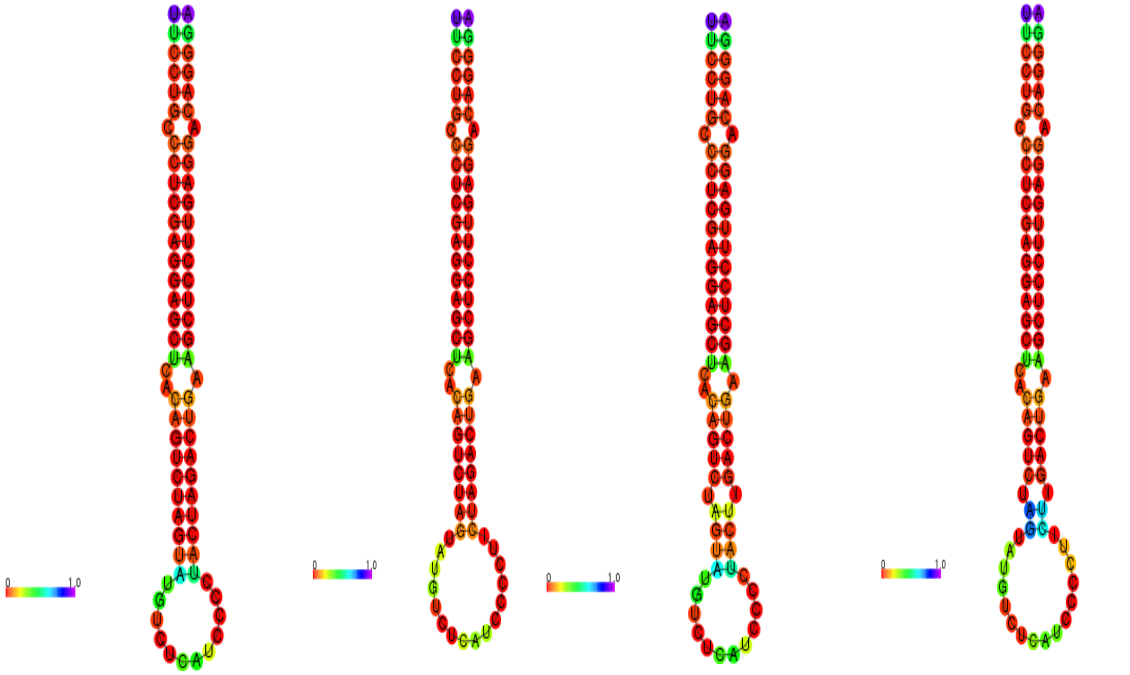


Figure 5.7: A change in energy within the terminal loop of both single edited hairpins is noticeable. It is only however when the hairpin is edited at both sites that processing by DICER is fully prevented. In the presence of both sites being edited there is a conformational change in the hairpin loop and a significant change in entropy around the loop. This figure was produced using Vienna RNAfold.

5.4.1 Rs1811399 is not an *in vivo* candidate for RNA editing.

Demonstrating that a sequence of RNA is edited *in vivo* is relatively simple. It involves sequencing the genomic DNA and a cDNA copy of a particular area and then comparing the two sequences. If there is any discrepancy then editing may have taken place. This will be performed for 6 human cell lines.

Table 5.3 below identifies the nucleotide present at the rs1811399 locus using both genomic DNA and sequence derived from transcribed RNA.

Name of Cell Line	**Genomic DNA sequence of rs1811399**	**cDNA sequence of rs1811399**
HeLa	**AC**	**AC**
SH-SY5Y	**AC**	**AC**
HI2162	**AA**	**AA**
HI2577	**AC**	**AC**
HEK293	**AC**	**AC**
HI2477	**AA**	**AA**

Table 5.3: Genomic DNA was isolated from the following cell lines: HeLa (Cervix), SH-SY5Y (neuroblastoma), HI2162, HI2477 and HI2577 (human lymphoblast) and HEK-293 (kidney). This was genotyped at the rs1811399 locus and the sequence compared against cDNA from all cell lines.

RNA editing had not happened at rs1811399 within the cells (Table 5.3). If there had been a discrepancy then it would raise interesting questions for RNA editing control of miRNA expression.

It may be possible that the miR-1273 hairpin is RNA edited and this was investigated in the next section.

5.4.2 Novel miR1273 hairpin is not an *in vivo* candidate for RNA editing.

Similar to the above, genomic DNA and cDNA for the nmiR-1273 was sequenced for discrepancies.

```
Sequence_nmiR-1273_gDNA      GGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGAGA 50
Sequence_nmiR-1273_cDNA      GGCATGAGAATCGCCTGAACCCGAGAGGCGGAGGTTGCAGTGAGCCGAGA 50
                             **************************************************

Sequence_nmiR-1273_gDNA      TCATGCTACTGCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTC-- 97
Sequence_nmiR-1273_cDNA      TCATGCTACTGCGCTCCAGCCTGGGTGACAGAGCGAGACTCCATCTCAA 99
                             *********************************************
```

Figure 5.8: Alignment of gDNA and cDNA for the nmiR-1273 locus. No RNA editing can be said to have occurred as there is no sequence discrepancy.

Both Table.5.3 and Fig. 5.8 demonstrate that in the cell no RNA editing takes place on these hairpins. There exists the possibility that editing of these hairpins is regulated and therefore an *in vitro* experiment was performed to see if the hairpin is a satisfactory substrate.

5.4.3 Rs1811399 hairpin is not an *in vitro* candidate for RNA editing.

RNA editing occurs on hairpin substrates (Tian *et al*, 2011). In order to see if the rs1811399 hairpins (A and C alleles) are adequate substrates RNA probes will be produced using T7 polymerase. These RNA molecules were then incubated with protein extract and reverse transcribed. The precursor was then amplified using PCR and sequenced.

cDNA YourSeq

```
TCTGGAGGTC AGGGCATGGT GATCCAGCGG CTGCCTGACA GTCACTGCCC  50
AGAGCTTCCC TTACCATAA
```

Genomic chr2 :

```
agcaaagttg gagggtttca ttgcctcgta gcttaatttt tatttcaaat  101478940
caagggctgg tattaaccat agcttggcag tgcagaaggc tgtggtcagg  101478990
TCTGGAGGTC AGGGCATGGT GATCCAGCGG CTGCCTGACA GTCACTGCCC  101479040
AGAGCTTCCC TTACCATAAc cttcctcagt agactagaaa aggttttcag  101479090
gtttgcccag gttatccaca cgtaccatgg catagtttct ccagcaagtc  101479140
agcacactgc cgttccttt
```

Side by Side Alignment

```
000000001 tctggaggtcagggcatggtgatccagcggctgcctgacagtcactgccc 000000050
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101478991 tctggaggtcagggcatggtgatccagcggctgcctgacagtcactgccc 101479040

000000051 agagcttcccttaccataa 000000069
>>>>>>>>> |||||||||||||||||| >>>>>>>>>
101479041 agagcttcccttaccataa 101479059
```

Figure 5.9: A RNA molecule of the rs1811399 hairpin was produced with the C variant. After incubation with cell extract as previously described, RNA was reverse transcribed and sequenced. Comparison with reference genome in bottom panel highlights rs1811399 variant locus and no change with regards to any potential RNA editing.

cDNA YourSeq

```
TCTGGAGGTC AGGGCATGGT GATaCAGCGG CTGCCTGACA GTCACTGCCC  50
AGAGCTTCCC TTACCATAA
```

Genomic chr2 :

```
agcaaagttg gagggtttca ttgcctcgta gcttaatttt tatttcaaat  101478940
caagggctgg tattaaccat agcttggcag tgcagaaggc tgtggtcagg  101478990
TCTGGAGGTC AGGGCATGGT GATcCAGCGG CTGCCTGACA GTCACTGCCC  101479040
AGAGCTTCCC TTACCATAAc cttcctcagt agactagaaa aggttttcag  101479090
gtttgcccag gttatccaca cgtaccatgg catagtttct ccagcaagtc  101479140
agcacactgc cgttccttt
```

Side by Side Alignment

```
000000001 tctggaggtcagggcatggtgatacagcggctgcctgacagtcactgccc 000000050
>>>>>>>>> ||||||||||||||||||||||| |||||||||||||||||||||||| >>>>>>>>>
101478991 tctggaggtcagggcatggtgatccagcggctgcctgacagtcactgccc 101479040

000000051 agagcttcccttaccataa 000000069
>>>>>>>>> |||||||||||||||||| >>>>>>>>>
101479041 agagcttcccttaccataa 101479059
```

Figure 5.10: A RNA molecule of the rs1811399 hairpin was produced with the A variant. After incubation with cell extract as previously described RNA was reverse transcribed and sequenced. Comparison with reference genome in bottom panel highlights rs1811399 variant locus and no change with regards to any potential RNA editing.

As Figures 5.9 and 5.10 demonstrate, none of the hairpins is an adequate substrate for RNA editing. It also demonstrates that the RNA editing locus within intron 1 of *NPAS2* does not extend this far upstream.

5.4.4 Novel miR1273 hairpin is not an *in vitro* candidate for RNA editing.

In order to see if the nmiR-1273 is an adequate substrate, RNA probes will be produced using T7 polymerase. This RNA molecule will then be incubated with protein extract and reverse transcribed. The precursor will then be amplified using PCR and sequenced.

Figure 5.11: Sequence gained after *in vitro* RNA editing experiment. Bottom panel demonstrates alignment between sample (top row) and reference genotype (bottom row).

Figure 5.11 demonstrates that nmiR-1273 is not a candidate substrate for RNA editing.

6. Discussion

6.1 Summary of main findings

In summary, this PhD has established the possibility that rs1811399 SNP may influence the rate of maturation of a previously unknown miRNA gene (Fig. 5.5) thus leading to aberrant regulation of downstream genes, thus potentially contributing to a disease phenotype.

This work has further identified novel miRNA genes within intron 1 of *NPAS2* including one which shares sequence homology with an established miRNA family (miR-1273). We have also identified both miRNA species are expressed independently of the host gene and have tentatively identified the promoter region responsible for driving its expression.

6.2 Rs1811399 has been linked with the autism phenotype.

Nicholas *et al* (2007) conducted a screen of circadian clock genes and demonstrated that rs1811399 is associated with autism at a confidence level of p=0.018. As the frequency of CC homozygous individual across the globe is on average around 9% (Fig. 5.1), it is not immediately apparent that it would have a deleterious impact on health. Given the higher frequency of the A allele in autistic children (Nicholas *et al*., 2007), CC homozygous individuals may actually benefit from the ancestral allele which reduces the maturation of the rs1811399 mircoRNA. One needs to consider however that phenotypes such as autism are inherently complex and any impact a single SNP could have in regards to causing it needs to be vigorously analysed, especially one which happens to be located within an intron.

One factor which has become apparent since the advent of large scale sequencing is the linkages between SNPs, for example rs2954041 and rs3924999 located in the *NRG1* gene (Yang *et al*, 2003). This should be borne in mind, for it is possible that another SNP that it is

in linkage with rs1811399 that might give rise to a phenotype. An attempt at disproving this linkage was made by Nicholas *et al* (2007) via the repeated sequencing of the whole *NPAS2* gene for each positive sample. However, since the release of this paper, a mass of sequencing data has been made available increasing the number of SNPs known to be in linkage with rs1811399. Rs1811399 is now known to be in linkage with at least 20 variant SNPs of which 5 (rs12472319, rs12472321, rs13011414, rs4081946 and rs6750976) are associated with regulatory regions. It is currently unknown if any of these SNP impart a functional deficit, however, *in silico* analysis indicates any impact would be minimal to non-existent (Table 5.2). None of the linked SNPs were picked up upon in the initial genome screen carried out by Nicholas *et al* (2007) and none currently have an attributed phenotype.

It is possible that the rs1811399 SNP may contribute to increased risk of developing autism in a certain subset of population due to its location within a miRNA precursor. What was not clear from the initial study, as it was beyond its scope, was the other analysed mutations that were present across the genome with which the SNP might interact. The SNP itself cannot be fully responsible for causing a phenotype as the variant CC homozygous genotype can be found in up to 21% of certain populations (1000 Genomes Mexican ancestry from Los Angeles), thus if this SNP were responsible for causing a certain phenotype of autism one would expect 21% of the Mexican population in Los Angeles to have a form of autism. As this is not the case and because the A allele is more frequently found in autistic children (Nicholas et al., 2007), CC homozygous individuals may be protected by having a reduced maturation rate of the rs11399 microRNA.` This statement is supported by twin studies quoted above (Muhle, Trentacoste & Rapin, 2004), which demonstrate monozygotic twins have a concordance rate of only 60%. This implies a conjunction between genetic mutation and environmental factors in the aetiology of autism.

The remainder of this thesis will attempt to elucidate the method by which rs1811399 could contribute to a phenotype.

6.3 Rs1811399 A>C does not influence regulatory region.

There are several established mechanisms by which a SNP could induce aberrant protein expression which could potential lead to a phenotype.

As established in the introduction, RNA editing plays a pivotal role in tissue specific expression of several genes. It was not unreasonable to assess a role for rs1811399 in such a pathway. This hypothesis was given further potential impetus by the presence of a confirmed RNA editing site at 10 separate locations within the *NPAS2* gene, all of which are in intron 1 and are located within an area associated with the transcription of exon 2 (Kim *et al*, 2004). To ascertain the answer, cDNA and gDNA were sequenced at the locus for several tissue types and none demonstrated any variance at the SNP locus (Table 5.3). This was further confirmed with an *in vitro* editing assay which, suggested that the RNA hairpin encoded by the locus might not be a suitable substrate for RNA editing (Fig.5.10). This finding suggests that whilst *NPAS2* might be regulated by RNA editing there does not appear to be a role for rs1811399 within this mechanism.

Neither does it seem that the SNP has a role in interfering with transcription factor binding. Were the SNP to influence binding (either in a positive or negative fashion) it would potential lead to an explanation of the association between rs1811399 and the autism phenotype as described by Nicholas *et al* (2007). Albers *et al* (2012) demonstrated that within the *RBM8A* gene, a SNP can dramatically influence the expression of the mature gene within platelets resulting in thrombocytopenia with absent radius (TAR syndrome). As we were not in possession of a homozygous CC cell line, contrary to efforts to purchase one through AGRE, we resorted to an *in vitro* assay to measure binding affinity of the locus to transcription

factors. An *in silico* study carried out in preparation for this experiment did not reveal any active transcription factor binding regions in the region of rs1811399 so anticipation that this was the mechanism by which the SNP can contribute to a phenotype was low. Indeed experimental evidence demonstrated that the SNP does not seem to influence protein binding (Fig4.23).

The only regulatory method that went untested was CpG island methylation. This was attempted in a previous project by B.Nicholas (unpublished) and demonstrated that the region was not the focus of a methylation event.

Taken together these findings exclude the possibility of rs1811399; as a target for RNA editing, is in a transcription factor binding locus and is in a CpG island. These would suggest that the region around the SNP is not directly involved with expression of downstream exons and as such a different mechanism must be established as to why a phenotype is associated with the SNP.

6.4 A putative miRNA cluster in the first intron of *NPAS2*.

NPAS2 is a well characterised transcription factor which regulates the expression of genes in a circadian manner via its affinity for binding to E-box sequences. A hitherto unknown mechanism by which NPAS2 might regulate gene expression has been investigated in this thesis: intron 1 of *NPAS2* contains a putative miRNA cluster.

NPAS2 is an ideal candidate for hosting miRNA genes in that it matches the generalised description of an archetypal miRNA host gene: it contains a large intron which acts as a repository for the miRNA (Zhou & Lin, 2006), and is a large gene of 176.7kb. Golan *et al*, 2010 noted that the average host gene is 177kb, Hinske *et al* (2009) elucidates further by stipulating that the majority of intronic miRNA genes are located within 5' introns.

The first miRNA precursor within *NPAS2* was tentatively identified by Nicholas *et al* (2008) during an investigation into the link between rs1811399 and the autism phenotype. *In silico* studies revealed that the SNP was located within the midst of a hairpin loop secondary structure, which was predicted to be suitable for DROSHA processing. The same author (unpublished) was also able to identify a homologue of the miR-1273 family upstream of the rs1811399 locus. Experimental evidence of the existence of these two miRNA, miR-1273h and miR-6275 according to naming convention, has been suggested within this body of work.

Notably a third miRNA has been described in this thesis which originates from the same precursor as the new miR-6725. This miRNA, again following precedent, will be known as miR-6725-3p whilst the original will be known as miR-6725-5p.

The sequences for each miRNA are (seed highlighted):

miR1273h: AGGCAUGAGAAUCGCCUGAACC

miR6725-5p: CAGGUCUGGAGGUCAGGGCAUG

miR6725-3p: CAGUCACUGCCCAGAGCUUCCC

Further *in silico* studies have identified several other potential miRNA precursors. The precursors have been experimentally validated, however the mature forms proved intransigent. This would suggest either a form of post-transcriptional repression of certain miRNA genes within the cluster or that the larger hairpin sequences that can be cloned are elements of the larger cluster transcript and are not further processed. The fact that these structures form hairpin structures but do not encode for mature forms would stipulate that either these *were* miRNA genes which have lost their functions or might someday *become* functional miRNA genes through a gain of function mutation. This phenomenon may even be occurring within the hairpin located at co-ordinates >chr2: 101,479,347-101,479,450. Within

this region exists two SNPs, one of which (rs137918055 G>A) causes the hairpin structure to attain a standard hairpin configuration. The incidence of this SNP within populations, as scanned by the 1000genomes project, is low (A allele only accounts for 0.5% of all alleles within the Chinese population with the A|G genotype accounting for 1% and G|G 99%) which implies either a negative selection pressure upon the locus or that it is a *de novo* mutation which arose within the Chinese population.

A search of EST libraries and other cDNA datasets identifies EST BF761854 with the region coding for our novel miR1273 variant. Interestingly a second, downstream (chr2: 101,477,169-101,477,511) EST T59368 also contains regions of strong homology with the miR1273 family. This second EST is found on the junction of two repeating elements: an Alu repeat and a L1MC4 repeat. These junctions between two repeating elements are an established "breeding ground" for miRNA and are prime drivers in miRNA evolution (Borchert *et al*, 2011). A consortium led by CSHL has performed a deep sequencing of the small RNA transcriptome and have identified at chr2:101478987-101479018 a region of small RNA that was expressed within a breast carcinoma cell line. This region is contiguous with the 5' arm mature miRNA of the rs1811399 locus.

Thus this could give rise to a miRNA cluster of at least 4 mature miRNA, two of which share the same precursor and two others of distinct but related precursors.

6.5 Evolution of a primate specific miRNA cluster within the *NPAS2* gene.

The reason for why a cluster should have integrated or evolved within the *NPAS2* gene is difficult to answer. Our initial assumption was that either the cluster was expressed within a circadian fashion with its host gene or that it required the same spatial expression pattern as exhibited by *NPAS2*. These two hypotheses were based on the co-expression of miRNA genes and their host genes (Baskerville & Bartel, 2005). However, experimental evidence

disclosed that two of the precursor molecules and the three detectable mature miRNA genes did not share a similar expression pattern (see Figure 4.16).

The fact that our information demonstrates the arising of regions of high homology within intron 1 of *NPAS2* within higher primates indicate that the region is relatively young, having arisen after the split of old world monkeys from non-human apes at around 25 million years ago. The mechanisms by which such regions arise can include genome duplication, retro-transposon activity or chromosomal translocation to name but three.

Upon closer examination of the sequence and gene structure of the chromosomes, it becomes apparent that *NPAS2*, within primates, is within the same genetic context i.e. regions of synteny with human chromosome 2. The close evolutionary history of the primate chromosomes is suggested by the fact that the homology search for the mouse homologous regions returns no result. An intriguing fact to note is that the human chromosome 2 is the fusion of two primate chromosomes: 2A and 2B (Ventura *et al*, 2012). This is further evident by regions of homology within the chromosome (approximately chr2: 114360201-114361000) to known telomere regions.

Whilst the events on a karyotypic level might have been large scale with the regards to the rearrangement, the sequence within the locus of the *NPAS2* gene appear to have been constant. Across human, gorilla and chimp the *NPAS2* locus consists of 20 coding exons and 21 introns, each of approximately similar length. Downstream of the *NPAS2* gene in each case is a region on the reverse strand encoding for *TBC1D8*. Within humans and gorilla genomes *RPL31*, a ribonucleoprotein, is also annotated. The corresponding sequence however is also present within chimp but no gene has been annotated to the location within the Ensembl or UCSC databases.

Within intron 1 of *NPAS2*, the proposed miRNA cluster, however much has seemingly changed from when gorilla branched from the *Pan-Homo* genus. The novel miR-1273 locus is completely conserved across human and chimp, within gorilla however exists a SNP (chr2a: 98295758 A>G) which does not seem to interfere with folding of the pre-cursor miRNA. It does however influence the terminal loop, rendering it smaller than in the human gene. Interpreting this in light of Tsutsumi *et al* (2011) which states that terminal loops should contain at least 7 unbound nucleotides to maximise cleavage by Dicer efficiency it would appear that expression of gorilla novel miR1273 would not be as efficient as human or chimp novel miR1273.

Chimpanzees are in, according to Ensembl, possession of the mutant rs1811399 C allele; however I found a chimp cell line to be heterozygous (Section 2.1.1.j). This is reflected by the fact that in the EB-176J cell line, the mature form of the rs1811399 miRNA is identifiable. Gorilla also contain the rs1811399C allele, however they possess a second mutation within the hairpin (chr2a: 98299086) a G is substituted for an A. This has the intriguing property of, in the case of homozygous rs1811399C, reinstating the correct hairpin structure. Neither of these point mutations impacts upon the sequences of the mature miRNA.

It would seem therefore that the cluster had arisen in a functional form by the time the hominine genus had arisen as it is present in both chimp and gorilla which separated 10 million years ago. Since this split, gorillas have accrued a SNP in both novel miR1273 and the rs1811399 miRNA gene with the former potentially reducing expression level and the later restoring the canonical hairpin. This would imply that the cluster is part of the evolution of the hominids and play a role in regulatory pathways essential for higher primates.

6.6 Expression of the *NPAS2* intron 1 miRNA cluster is not dependent on host gene expression.

In the beginning it was assumed that miRNA genes which were located within the introns of protein coding genes were located so to make use of the protein coding gene's promoter region. Whilst this is indeed the case in certain examples it is not true for every miRNA.

Located as they are within the intron of a circadian clock gene one would assume that this would be the case to make use of the rhythmicity of the host gene's expression. The author undertook to ascertain whether a link was present by synchronising the circadian cycle of cells growing in culture and analysing the expression profile of both *NPAS2* and the miRNA genes. Whilst quantitative PCR was inconclusive, probably due to the short nature of the target transcript with regards to the miRNA, it was apparent utilising standard PCR to gain a qualitative answer as to the link between expression of host and miRNA.

Referring back to Figure 4.16 we have demonstrated that there is no requirement for expression of host gene mRNA for there to be expression of the miRNA cluster. This requirement was assessed utilising primers specific for several isoforms of NPAS2 but no overall link was detected. To cite an example, in asynchronous cells very little *NPAS2* expression was detectable whilst levels of miRNA precursor were consistent (Fig. 4.16). The next hypothesis to test was whether the precursor was consistently converted into mature form of miRNA or whether some form of post-transcriptional regulation was being inflicted. This does not seem to be the case as the mature form was identified from all stages of the circadian cycle and under various genotoxic conditions. A phenomenon of note; when *NPAS2* was induced via serum shock or genotoxic stress the level of expression of the precursor and mature forms did not alter. This was further proof of the independent nature of transcription of this cluster.

Consulting Monteys *et al* (2010) it was possible to begin an *in silico* examinations of upstream regions from the miRNA cluster to identify the promoter region. Utilising the UCSC genome browser it was possible to stipulate the conditions required for a putative transcription start region and by looking for histone methylation sites, DNase activity region, PolII binding sites and any other conserved transcription factor binding sites it was possible to identify a candidate region. Within this region, there was no CpG island detected by the ENCODE project. Monteys *et al* (2010) stipulate that this is the case in ~72% of all independently transcribed miRNA. Within the same region however existed a recognised PolII binding site (chr2:101474860-101474890 using hg19) and three transcription factor binding sites as described by the ENCODE project (these factors being GR at chr2:101473293-101473562, c-Fos at chr2:101474417-101474760 and GATA-2 at chr2:101474423-101474802 with all three coordinates using the hg19 system). Once identified it was possible to clone the region downstream of this promoter from a cDNA pool whereas the same was not possible for the area upstream of the predicted region. An assessment was undertaken of the putative transcription site by cloning it into a *E.coli* vector with the genomic sequence of miR-122, a miRNA with a strictly hepatic expression pattern, which allowed for expression of the miR within HeLa cells. Thus the promoter is clearly capable of driving expression of miRNA *in vivo*.

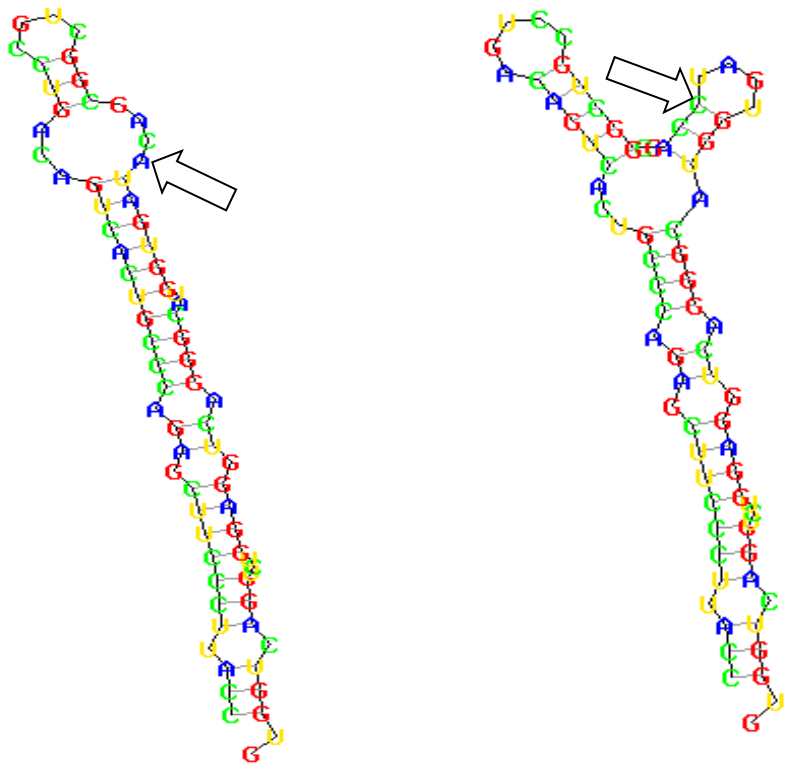
6.7 Rs1811399 A>C Influences maturation of novel miRNA.

In silico predictions have implied that a point mutation at rs1811399; C replacing A should distort the predicted secondary structure of a novel miRNA species. Interestingly whilst the initial *in silico* experimentation would seem to imply the mutation imbues an all or nothing type loss of function on the miRNA gene, the experimental work undertaken by the author (Figure 5.5) implies that even in the presence of the deleterious SNP some limited maturation

occurs. This finding is in keeping with other published works namely Harnprasopwat *et al* (2010) and Duan, Pak and Jin (2007) who discovered a SNP within pre-miR126 and pre-miR125, respectively, that influenced maturation rates of the miRNA. rs12976445 in pre-miR-125 for example reduces the expression of the mature form by 85%(Lehman *et al*, 2013).

The proposed model by which this mutation influences miRNA biogenesis is as follows:

A



A allele C allele

>rs1811399A

TGGTCAGGTCTGGAGGTCAGGGCATGGTGAT**A**CAGCGGCTGCCTGACAGTCACT

GCCCAGAGCTTCCCTTACC

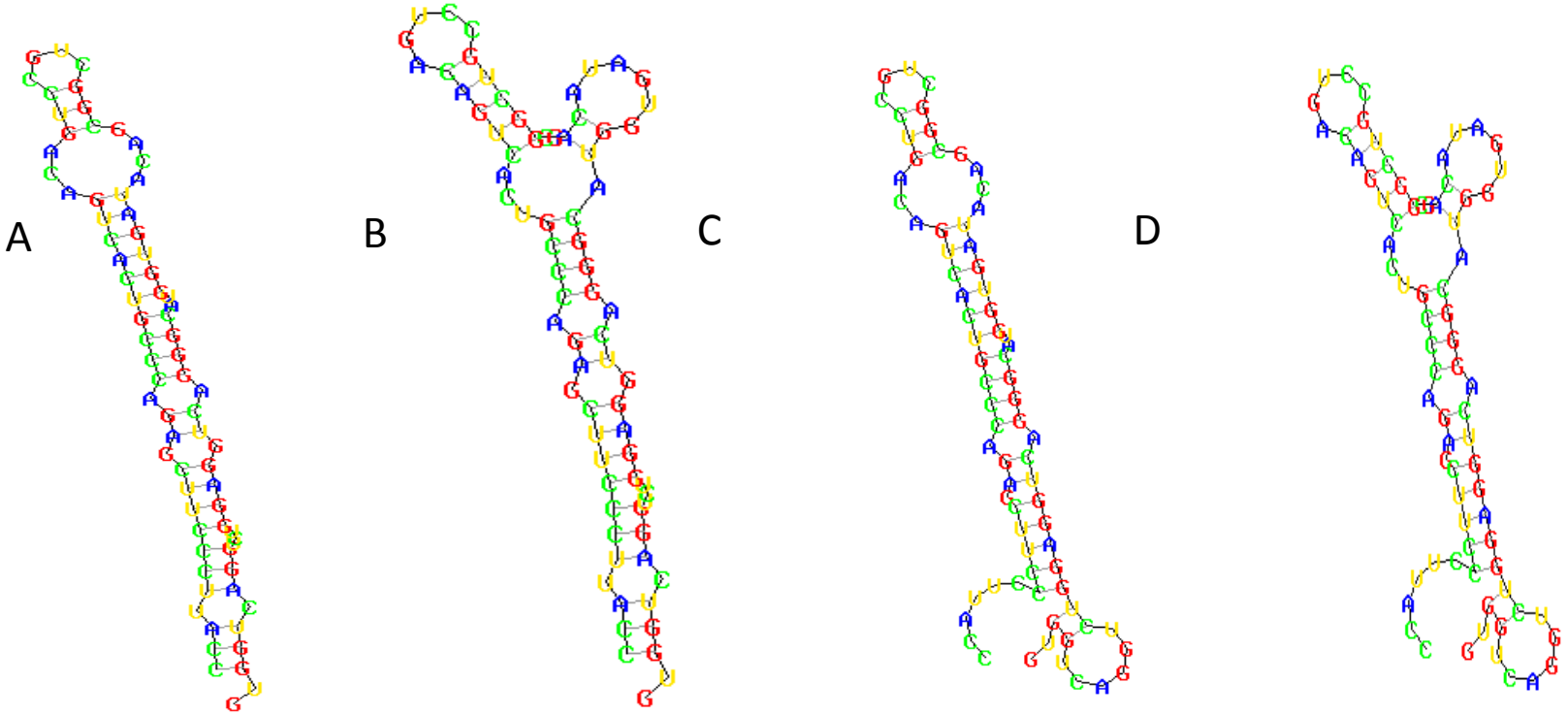
>rs1811399C

TGGTCAGGTCTGGAGGTCAGGGCATGGTGAT**C**CAGCGGCTGCCTGACAGTCACT

GCCCAGAGCTTCCCTTACC

Sequence of folding events within rs1811399A precursor

B

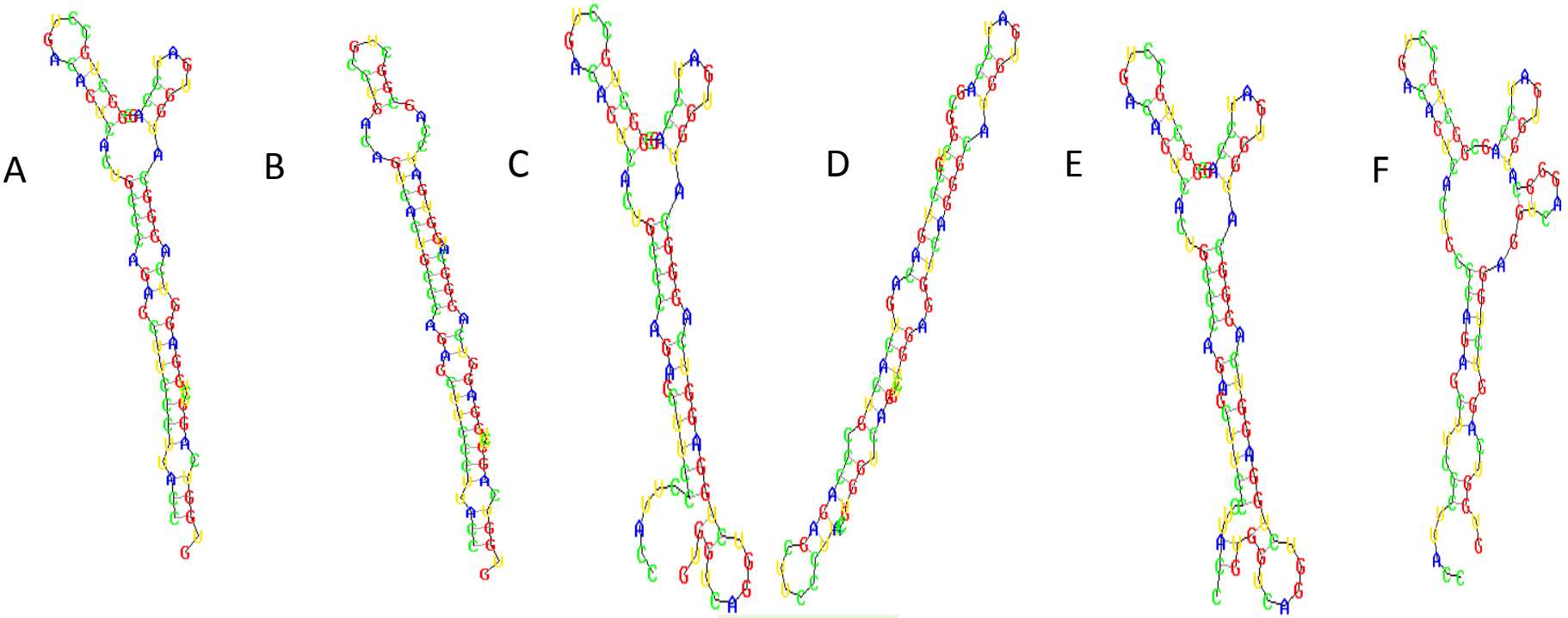


Probability of precursor existing in each state:

A	0.9002435
B	0.0996705
C	0.0000781
D	0.0000079

C

Sequence of folding events within rs1811399C precursor



Probability of precursor existing in each state:

A	0.8217119
B	0.1781900
C	0.0000655
D	0.0000267
E	0.0000038
F	0.0000022

Figure 6.1: Three panels depicting influence of rs1811399 SNP on RNA secondary structure. (A) Demonstrates the average calculated structure of the rs1811399 locus' hairpin and the influence the SNP can have on this average structure. (B) and (C) each describe the total number of calculated structures for the locus with each SNP, The algorithm utilised by the RNAShapes (Steffen *et al*, 2006) also allocates a probability that any hairpin can be found within a specific conformational configuration at any point in time.

Figure 6.1 A demonstrates the *average* predicted structure for the given nucleotide sequences. To arrive at this structure the total *Gibbs free energy* of the nucleotide sequence is deduced, this free energy being a physical measure of the amount of "useable" energy available within the sequence which could be used to "do work". When the algorithm has arrived at a number for the free energy given to the nucleotide sequence it then attempts to map a structure which amounts to the *minimum* possible free energy in order to satisfy the first and second laws of thermodynamics, briefly: the energy of the universe is always constant whilst the entropy (measure of order) is always increasing. The entropy of a system can be described as a measure of how close any system is to equilibrium (in thermodynamics equilibrium is a state at which no further change can occur) thus the lower the value of free energy within a given system is reflective of how likely the structure is to form. Using this knowledge we can see that the free energy for the A hairpin is -28.59 kcal/mol whilst that of the C hairpin is -27.60 kcal/mol. When compared with the free energy of the human let7a miRNA precursor of -35.15 kcal/mol or miR-24 at -26.79 kcal/mol and finally miR-29a at -25.16 kcal/mol we can note with satisfaction that both our hairpins match the free energy spectrum as required for miRNA precursors. We can therefore conclude that whilst the substitution of cytosine for adenine does influence the total free energy of the structure it does not remove it from the range at which miRNA precursors can spontaneously fold.

Steffen *et al* (2006) have devised a computer model which calculates all possible structural conformations for a given sequence and based upon their relative free energies can calculate the probability that you can find each sequence within a given secondary structure. Thus we can see with the A allele hairpin the greatest probability (~90%) is that of a hairpin structure (Fig.6.1B). The remaining 10% can be described as varied forked hairpin structures. With that of the C allele the picture is reversed. Circa 82% of all the structures can be described as forked hairpins with only ~17% being a plain hairpin (Fig.6.1C). Of interest is also the fact that the C allele hairpin has a greater number of alternate folding forms than that of the A allele. These facts are of significance when we consider the next step of the model.

As established in the introduction, the secondary structure of a miRNA precursor is essential for its processing by both DROSHA and DICER. As the mutation is within the terminal loop region it is hypothesised that rs1811399C influences DICER binding and recognition which would explain the incomplete processing of the rs1811399 precursor in the chicken DT40 cells (Fig.5.5).

Much work has been undertaken to understand the efficacy of the DICER enzyme. Tsutsumi *et al* (2011) revealed that the terminal loop region of a miRNA precursor is essential for recognition by DICER and that this region is recognised and selectively bound via DICER's helicase domain. Further binding is achieved via the PAZ domain which selects the 5' end of the pre-miRNA for processing before it is cut via the RNAse III domain. Tsutsumi *et al* (2011) stipulate that whilst the shape of the terminal loop is not strictly essential for correct recognition by DICER, the enzyme does have a structural requirement of having to have ~7nt length of ssRNA within the terminal loop. Thus we may conclude that as the rs1811399 C mutation limits the amount of ssRNA within the terminal loop region of the hairpin that DICER activity *in vivo* is reduced, thus limiting the amount of final product that can be generated.

The impact of a reduced amount of mature form will obviously cause aberrant regulation of target genes. As we have two mature miRNA encoded from the single hairpin that is affected by rs1811399 we would do well to consider Table 2, 3 and 4 which identify these miRNA as having important roles in neural development, cellular signalling and even the miRNA processing pathway. Whilst each of these pathways is under much regulation and the influence of one deranged miRNA precursor is slight, it becomes apparent that as the networks become more complex additional errors can accumulate. The end stage of this accumulation would be a disease phenotype. Such multifactorial causes are common in neurodevelopmental disorders (Fanous *et al* 2012).

6.8 Conclusion

As a body of work this PhD has conclusively identified three putative novel miRNA within intron 1 of the *NPAS2* gene. The project has also identified several hairpin structures, which may have been ancient miRNA or are miRNA within the process of evolving. The method of expressing this miRNA cluster has also been tested. Results suggest that miRNA within the region might be controlled by an intronic promoter region which allows for expression of the miRNA without requirement for host gene expression.

This PhD has also examined the role rs1811399C within the autism phenotype. The hypothesis stated that the C allele would prevent expression of the mature form. Evidence suggests that the C allele might alter the rate at which mature form is produced. This would have downstream implications with gene regulation, an established phenomenon within autism.

Appendix 1

Complete target list of *NPAS2* intron 1 miRNA cluster.

The author's own efforts were correlated with the Target Scan web-software (Lewis, Burge and Bartel 2005). Below are the tables containing all the targets for the miRNA along with the number of conserved sites within each target's 3' UTR.

Rs1811399 5' arm miRNA

		Conserved sites			
Target gene	**Gene name**	**total**	**8nt**	**7nt-m8**	**7nt-1A**
HIC2	hypermethylated in cancer 2	2	0	2	0
METAP1	methionyl aminopeptidase 1	2	0	2	0
MLL	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)	2	0	1	1
SENP5	SUMO1/sentrin specific peptidase 5	2	2	0	0
SOX11	SRY (sex determining region Y)-box 11	2	0	1	1
ZFHX4	zinc finger homeobox 4	2	0	1	1
ACTR1B	ARP1 actin-related protein 1 homolog B, centractin beta (yeast)	1	1	0	0
ADAMTSL3	ADAMTS-like 3	1	0	0	1
ADNP	activity-dependent neuroprotector homeobox	1	0	0	1
ARC	activity-regulated cytoskeleton-associated protein	1	0	1	0
ARID2	AT rich interactive domain 2 (ARID, RFX-like)	1	0	0	1
ARL6IP5	ADP-ribosylation-like factor 6 interacting protein 5	1	1	0	0
ATP13A3	ATPase type 13A3	1	1	0	0
ATP1A2	ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide	1	0	0	1

ATXN7L3	ataxin 7-like 3	1	0	1	0
B3GAT1	beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)	1	0	1	0
BBS4	Bardet-Biedl syndrome 4	1	0	1	0
BCL2	B-cell CLL/lymphoma 2	1	0	1	0
BCL2L2	BCL2-like 2	1	0	1	0
BMF	Bcl2 modifying factor	1	0	1	0
BTNL3	butyrophilin-like 3	1	0	1	0
C11orf47	chromosome 11 open reading frame 47	1	0	1	0
C11orf58	chromosome 11 open reading frame 58	1	0	1	0
C15orf27	chromosome 15 open reading frame 27	1	0	1	0
CALN1	calneuron 1	1	0	1	0
CASC3	cancer susceptibility candidate 3	1	0	1	0
CDCA7L	cell division cycle associated 7-like	1	0	1	0
COL4A6	collagen, type IV, alpha 6	1	1	0	0
CPEB4	cytoplasmic polyadenylation element binding protein 4	1	0	1	0
CREB5	cAMP responsive element binding protein 5	1	0	1	0
CRKL	v-crk sarcoma virus CT10 oncogene homolog (avian)-like	1	1	0	0
DAB1	disabled homolog 1 (Drosophila)	1	1	0	0
DDIT4	DNA-damage-inducible transcript 4	1	0	0	1
DNAJC6	DnaJ (Hsp40) homolog, subfamily C, member 6	1	0	1	0
EIF2C2	eukaryotic translation initiation factor 2C, 2	1	0	1	0
EIF4EBP2	eukaryotic translation initiation factor 4E binding protein 2	1	1	0	0
ENPEP	glutamyl aminopeptidase (aminopeptidase A)	1	1	0	0
FAM133A	family with sequence similarity 133, member A	1	0	1	0
FAT2	FAT tumor suppressor homolog 2 (Drosophila)	1	0	1	0

FBXW7	F-box and WD repeat domain containing 7	1	0	0	1
FGF7	fibroblast growth factor 7 (keratinocyte growth factor)	1	1	0	0
FLRT2	fibronectin leucine rich transmembrane protein 2	1	0	0	1
FOXD2	forkhead box D2	1	0	1	0
FOXN3	forkhead box N3	1	1	0	0
FOXO3	forkhead box O3	1	0	0	1
FRMD4B	FERM domain containing 4B	1	0	0	1
GAL3ST3	galactose-3-O-sulfotransferase 3	1	1	0	0
GAS2L1	growth arrest-specific 2 like 1	1	0	1	0
GATAD2A	GATA zinc finger domain containing 2A	1	0	0	1
GCC2	GRIP and coiled-coil domain containing 2	1	1	0	0
GDPD1	glycerophosphodiester phosphodiesterase domain containing 1	1	0	0	1
GLIS1	GLIS family zinc finger 1	1	0	1	0
GRIK2	glutamate receptor, ionotropic, kainate 2	1	0	1	0
GRIN2D	glutamate receptor, ionotropic, N-methyl D-aspartate 2D	1	0	1	0
GRLF1	glucocorticoid receptor DNA binding factor 1	1	0	1	0
H3F3B	H3 histone, family 3B (H3.3B)	1	0	1	0
HMG2L1	high-mobility group protein 2-like 1	1	0	1	0
HS2ST1	heparan sulfate 2-O-sulfotransferase 1	1	1	0	0
IDS	iduronate 2-sulfatase (Hunter syndrome)	1	0	1	0
IGSF9	immunoglobulin superfamily, member 9	1	1	0	0
INSIG1	insulin induced gene 1	1	1	0	0
IPO9	importin 9	1	1	0	0
ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	1	0	1	0
JARID1C	jumonji, AT rich interactive domain 1C	1	0	1	0

KCMF1	potassium channel modulatory factor 1	1	0	0	1
KLF3	Kruppel-like factor 3 (basic)	1	1	0	0
KPNA6	karyopherin alpha 6 (importin alpha 7)	1	0	1	0
LPHN1	latrophilin 1	1	0	1	0
LRRC59	leucine rich repeat containing 59	1	1	0	0
LRRC7	leucine rich repeat containing 7	1	1	0	0
MECP2	methyl CpG binding protein 2 (Rett syndrome)	1	0	0	1
MEST	mesoderm specific transcript homolog (mouse)	1	0	0	1
MMP16	matrix metallopeptidase 16 (membrane-inserted)	1	0	1	0
MRFAP1	Mof4 family associated protein 1	1	1	0	0
NCAM1	neural cell adhesion molecule 1	1	0	1	0
NFASC	neurofascin homolog (chicken)	1	1	0	0
NFYB	nuclear transcription factor Y, beta	1	0	1	0
NHLH1	nescient helix loop helix 1	1	0	1	0
NRIP3	nuclear receptor interacting protein 3	1	0	1	0
ONECUT2	one cut homeobox 2	1	0	1	0
PCP4L1	Purkinje cell protein 4 like 1	1	1	0	0
PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)	1	0	1	0
PHC3	polyhomeotic homolog 3 (Drosophila)	1	1	0	0
PHF15	PHD finger protein 15	1	0	0	1
PLXNA4	plexin A4	1	0	1	0
POU2F2	POU class 2 homeobox 2	1	0	1	0
PRELID2	PRELI domain containing 2	1	1	0	0
PRKAG1	protein kinase, AMP-activated, gamma 1 non-catalytic subunit	1	0	1	0

PSMF1	proteasome (prosome, macropain) inhibitor subunit 1 (PI31)	1	0	1	0
PTAFR	platelet-activating factor receptor	1	0	1	0
PTPN11	protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)	1	0	0	1
PURB	purine-rich element binding protein B	1	0	0	1
PVRL2	poliovirus receptor-related 2 (herpesvirus entry mediator B)	1	0	1	0
RAB1A	RAB1A, member RAS oncogene family	1	0	0	1
RC3H1	ring finger and CCCH-type zinc finger domains 1	1	0	1	0
RNASEN	ribonuclease type III, nuclear	1	0	0	1
RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2	1	0	1	0
S1PR2	sphingosine-1-phosphate receptor 2	1	0	1	0
SERTAD2	SERTA domain containing 2	1	0	1	0
SLC22A17	solute carrier family 22, member 17	1	0	1	0
SNAP29	synaptosomal-associated protein, 29kDa	1	1	0	0
SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	1	0	1	0
SNX1	sorting nexin 1	1	0	0	1
SP1	Sp1 transcription factor	1	0	1	0
SRC	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)	1	0	1	0
SRPK2	SFRS protein kinase 2	1	0	1	0
SSR1	signal sequence receptor, alpha (translocon-associated protein alpha)	1	0	0	1
SUPT7L	suppressor of Ty 7 (S. cerevisiae)-like	1	1	0	0
TCF4	transcription factor 4	1	0	0	1
TEF	thyrotrophic embryonic factor	1	0	1	0
TGOLN2	trans-golgi network protein 2	1	1	0	0

THAP6	THAP domain containing 6	1	0	0	1
TNKS	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase	1	0	0	1
TRPS1	trichorhinophalangeal syndrome I	1	0	0	1
UBE2D3	ubiquitin-conjugating enzyme E2D 3 (UBC4/5 homolog, yeast)	1	0	1	0
ZCCHC14	zinc finger, CCHC domain containing 14	1	0	1	0
ZFAND5	zinc finger, AN1-type domain 5	1	0	1	0
ZNF2	zinc finger protein 2	1	0	0	1
ZNF706	zinc finger protein 706	1	0	1	0
ZNF827	zinc finger protein 827	1	0	1	0
ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	1	0	0	1

Rs1811399 3' arm miRNA

		Conserved sites			
Target gene	**Gene name**	**total**	**8nt**	**7nt-m8**	**7nt-1A**
ARMC8	armadillo repeat containing 8	2	1	1	0
C6orf168	chromosome 6 open reading frame 168	2	1	0	1
MTMR9	myotubularin related protein 9	2	0	2	0
38412	membrane-associated ring finger (C3HC4) 5	1	0	1	0
A2BP1	ataxin 2-binding protein 1	1	1	0	0
ABHD2	abhydrolase domain containing 2	1	1	0	0
ACSL4	acyl-CoA synthetase long-chain family member 4	1	0	1	0
ADAM17	ADAM metallopeptidase domain 17 (tumor necrosis factor, alpha, converting enzyme)	1	0	1	0
ADAR	adenosine deaminase, RNA-specific	1	0	1	0
ADCY6	adenylate cyclase 6	1	0	1	0

AFF4	AF4/FMR2 family, member 4	1	1	0	0
AGPAT3	1-acylglycerol-3-phosphate O-acyltransferase 3	1	0	0	1
ALOX12	arachidonate 12-lipoxygenase	1	0	0	1
ANKRD28	ankyrin repeat domain 28	1	0	1	0
ANKRD49	ankyrin repeat domain 49	1	0	1	0
AP1G1	adaptor-related protein complex 1, gamma 1 subunit	1	1	0	0
BAZ2A	bromodomain adjacent to zinc finger domain, 2A	1	0	1	0
BRPF3	bromodomain and PHD finger containing, 3	1	0	1	0
BRWD1	bromodomain and WD repeat domain containing 1	1	0	0	1
C15orf57	chromosome 15 open reading frame 57	1	0	1	0
C18orf34	chromosome 18 open reading frame 34	1	0	1	0
C1orf128	chromosome 1 open reading frame 128	1	0	0	1
C6orf167	chromosome 6 open reading frame 167	1	0	1	0
C7orf60	chromosome 7 open reading frame 60	1	0	1	0
CALM1	calmodulin 1 (phosphorylase kinase, delta)	1	1	0	0
CALM3	calmodulin 3 (phosphorylase kinase, delta)	1	1	0	0
CCDC132	coiled-coil domain containing 132	1	0	0	1
CCNE2	cyclin E2	1	0	0	1
CDC2L6	cell division cycle 2-like 6 (CDK8-like)	1	1	0	0
CDC42SE1	CDC42 small effector 1	1	1	0	0
CECR2	cat eye syndrome chromosome region, candidate 2	1	1	0	0
CFL1	cofilin 1 (non-muscle)	1	0	1	0
CMIP	c-Maf-inducing protein	1	0	1	0
CNTN2	contactin 2 (axonal)	1	0	0	1
COL7A1	collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive)	1	0	1	0
CREBZF	CREB/ATF bZIP transcription factor	1	0	0	1
CSNK1A1	casein kinase 1, alpha 1	1	1	0	0

CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa	1	0	1	0
DAAM1	dishevelled associated activator of morphogenesis 1	1	1	0	0
DCX	doublecortex; lissencephaly, X-linked (doublecortin)	1	1	0	0
DDIT3	DNA-damage-inducible transcript 3	1	0	1	0
DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	1	0	1	0
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	1	0	1	0
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	1	0	1	0
DIP2B	DIP2 disco-interacting protein 2 homolog B (Drosophila)	1	1	0	0
DLL1	delta-like 1 (Drosophila)	1	0	0	1
DPYSL2	dihydropyrimidinase-like 2	1	0	1	0
EDA	ectodysplasin A	1	0	1	0
EGLN2	egl nine homolog 2 (C. elegans)	1	0	1	0
EGR1	early growth response 1	1	0	1	0
EID1	EP300 interacting inhibitor of differentiation 1	1	0	1	0
EIF5A2	eukaryotic translation initiation factor 5A2	1	0	1	0
EPB41	erythrocyte membrane protein band 4.1 (elliptocytosis 1, RH-linked)	1	0	1	0
EPC2	enhancer of polycomb homolog 2 (Drosophila)	1	0	1	0
EPHA10	EPH receptor A10	1	0	1	0
ESRRG	estrogen-related receptor gamma	1	0	1	0
ETV1	ets variant gene 1	1	0	1	0
FAM152A	family with sequence similarity 152, member A	1	0	1	0
FAM168B	family with sequence similarity 168, member B	1	0	1	0
FAM38B	family with sequence similarity 38, member B	1	0	1	0
FBXL11	F-box and leucine-rich repeat protein 11	1	0	1	0
FBXO41	F-box protein 41	1	1	0	0
FGFBP3	fibroblast growth factor binding protein 3	1	1	0	0
FLJ20309	hypothetical protein FLJ20309	1	0	1	0

FOXG1	forkhead box G1	1	0	1	0
FOXP2	forkhead box P2	1	0	0	1
FXR1	fragile X mental retardation, autosomal homolog 1	1	0	1	0
GABBR2	gamma-aminobutyric acid (GABA) B receptor, 2	1	0	1	0
GABPA	GA binding protein transcription factor, alpha subunit 60kDa	1	0	0	1
GABRE	gamma-aminobutyric acid (GABA) A receptor, epsilon	1	0	1	0
GALNT2	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 2 (GalNAc-T2)	1	0	1	0
GGNBP2	gametogenetin binding protein 2	1	0	1	0
GIGYF2	GRB10 interacting GYF protein 2	1	0	1	0
GMCL1	germ cell-less homolog 1 (Drosophila)	1	0	0	1
GMFB	glia maturation factor, beta	1	0	0	1
GNAO1	guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O	1	0	1	0
GPC4	glypican 4	1	0	1	0
GTF2I	general transcription factor II, i	1	0	1	0
H3F3B	H3 histone, family 3B (H3.3B)	1	0	1	0
HCN3	hyperpolarization activated cyclic nucleotide-gated potassium channel 3	1	0	1	0
HES1	hairy and enhancer of split 1, (Drosophila)	1	0	1	0
HES2	hairy and enhancer of split 2 (Drosophila)	1	0	1	0
HES5	hairy and enhancer of split 5 (Drosophila)	1	0	1	0
HMGCR	3-hydroxy-3-methylglutaryl-Coenzyme A reductase	1	0	1	0
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	1	0	0	1
HNRNPUL1	heterogeneous nuclear ribonucleoprotein U-like 1	1	1	0	0
HOXD9	homeobox D9	1	0	1	0
HRB	HIV-1 Rev binding protein	1	0	1	0
HSPC159	galectin-related protein	1	0	1	0

HTR3C	5-hydroxytryptamine (serotonin) receptor 3, family member C	1	1	0	0
IPO5	importin 5	1	0	1	0
ITPR2	inositol 1,4,5-triphosphate receptor, type 2	1	0	0	1
JHDM1D	jumonji C domain containing histone demethylase 1 homolog D (S. cerevisiae)	1	0	1	0
JPH1	junctophilin 1	1	0	0	1
KCNC1	potassium voltage-gated channel, Shaw-related subfamily, member 1	1	0	1	0
KIAA0152	KIAA0152	1	0	1	0
KIAA1045	KIAA1045	1	1	0	0
KIAA2022	KIAA2022	1	0	1	0
KIFAP3	kinesin-associated protein 3	1	1	0	0
KLHL21	kelch-like 21 (Drosophila)	1	0	1	0
LDLRAP1	low density lipoprotein receptor adaptor protein 1	1	0	1	0
LEP	leptin	1	0	1	0
LMOD3	leiomodin 3 (fetal)	1	0	1	0
LOC399947	similar to expressed sequence AI593442	1	1	0	0
LOC402665	hCG1651476	1	0	1	0
LOC440093	histone H3-like	1	0	1	0
MAN1C1	mannosidase, alpha, class 1C, member 1	1	0	1	0
MAP1B	microtubule-associated protein 1B	1	0	1	0
MAPK14	mitogen-activated protein kinase 14	1	0	1	0
MAPT	microtubule-associated protein tau	1	1	0	0
MARCKS	myristoylated alanine-rich protein kinase C substrate	1	0	0	1
MBNL1	muscleblind-like (Drosophila)	1	0	1	0
MCC	mutated in colorectal cancers	1	0	0	1
MED29	mediator complex subunit 29	1	0	1	0
MNX1	motor neuron and pancreas homeobox 1	1	0	1	0

MON2	MON2 homolog (S. cerevisiae)	1	1	0	0
MRPL42	mitochondrial ribosomal protein L42	1	0	1	0
MYST3	MYST histone acetyltransferase (monocytic leukemia) 3	1	0	1	0
MYT1L	myelin transcription factor 1-like	1	0	0	1
NAP1L1	nucleosome assembly protein 1-like 1	1	0	1	0
NAV2	neuron navigator 2	1	0	1	0
NCK1	NCK adaptor protein 1	1	0	1	0
NCOA4	nuclear receptor coactivator 4	1	0	1	0
NDST1	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1	1	0	1	0
NEUROD2	neurogenic differentiation 2	1	0	1	0
NOTUM	notum pectinacetylesterase homolog (Drosophila)	1	0	1	0
NR4A3	nuclear receptor subfamily 4, group A, member 3	1	0	1	0
ORC4L	origin recognition complex, subunit 4-like (yeast)	1	0	1	0
OSBPL2	oxysterol binding protein-like 2	1	0	1	0
PAPOLA	poly(A) polymerase alpha	1	0	1	0
PDCD7	programmed cell death 7	1	0	1	0
PDK4	pyruvate dehydrogenase kinase, isozyme 4	1	0	0	1
PGK1	phosphoglycerate kinase 1	1	1	0	0
PHLPPL	PH domain and leucine rich repeat protein phosphatase-like	1	0	1	0
PNPO	pyridoxamine 5'-phosphate oxidase	1	0	0	1
PPP2R1B	protein phosphatase 2 (formerly 2A), regulatory subunit A, beta isoform	1	0	1	0
PSMA2	proteasome (prosome, macropain) subunit, alpha type, 2	1	0	1	0
RBBP5	retinoblastoma binding protein 5	1	0	1	0
RBBP9	retinoblastoma binding protein 9	1	0	1	0
RBM33	RNA binding motif protein 33	1	0	1	0
RELN	reelin	1	0	1	0

RNF141	ring finger protein 141	1	0	1	0
RRM2	ribonucleotide reductase M2 polypeptide	1	1	0	0
SCN1A	sodium channel, voltage-gated, type I, alpha subunit	1	0	0	1
SCN2A	sodium channel, voltage-gated, type II, alpha subunit	1	0	0	1
SDC4	syndecan 4	1	0	1	0
SEMA4F	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4F	1	1	0	0
SGSM2	small G protein signaling modulator 2	1	0	1	0
SLC12A5	solute carrier family 12, (potassium-chloride transporter) member 5	1	0	1	0
SLC22A3	solute carrier family 22 (extraneuronal monoamine transporter), member 3	1	0	1	0
SLC24A3	solute carrier family 24 (sodium/potassium/calcium exchanger), member 3	1	0	1	0
SLC27A4	solute carrier family 27 (fatty acid transporter), member 4	1	0	1	0
SMAD4	SMAD family member 4	1	0	1	0
SMAD5	SMAD family member 5	1	0	1	0
SNCA	synuclein, alpha (non A4 component of amyloid precursor)	1	1	0	0
SOCS5	suppressor of cytokine signaling 5	1	0	0	1
SP7	Sp7 transcription factor	1	0	1	0
SPEN	spen homolog, transcriptional regulator (Drosophila)	1	0	1	0
SPOPL	speckle-type POZ protein-like	1	0	1	0
STAG3L4	stromal antigen 3-like 4	1	1	0	0
STX1A	syntaxin 1A (brain)	1	0	1	0
STXBP4	syntaxin binding protein 4	1	1	0	0
STXBP5L	syntaxin binding protein 5-like	1	0	0	1
SUFU	suppressor of fused homolog (Drosophila)	1	0	1	0

TCF4	transcription factor 4	1	1	0	0
TMEM132B	transmembrane protein 132B	1	0	1	0
TMEM70	transmembrane protein 70	1	0	1	0
TRAPPC3	trafficking protein particle complex 3	1	0	1	0
TRIM9	tripartite motif-containing 9	1	0	1	0
TSPYL5	TSPY-like 5	1	0	1	0
UBR1	ubiquitin protein ligase E3 component n-recognin 1	1	0	1	0
UNQ1887	signal peptide peptidase 3	1	0	1	0
USP37	ubiquitin specific peptidase 37	1	0	1	0
VGF	VGF nerve growth factor inducible	1	1	0	0
WDR45L	WDR45-like	1	1	0	0
WDR48	WD repeat domain 48	1	1	0	0
WIBG	within bgcn homolog (Drosophila)	1	1	0	0
WTAP	Wilms tumor 1 associated protein	1	0	1	0
YOD1	YOD1 OTU deubiquinating enzyme 1 homolog (S. cerevisiae)	1	0	1	0
YPEL2	yippee-like 2 (Drosophila)	1	0	1	0
ZBTB34	zinc finger and BTB domain containing 34	1	0	1	0
ZCCHC4	zinc finger, CCHC domain containing 4	1	0	1	0
ZDHHC20	zinc finger, DHHC-type containing 20	1	0	1	0
ZMIZ1	zinc finger, MIZ-type containing 1	1	0	1	0
ZNF618	zinc finger protein 618	1	0	1	0
ZXDC	ZXD family zinc finger C	1	0	1	0

Novel miR-1273-1 miRNA

Target gene	**Gene name**	**Conserved sites**

		total	**8nt**	**7nt-m8**	**7nt-1A**
PTGER3	prostaglandin E receptor 3 (subtype EP3)	2	2	0	0
USP47	ubiquitin specific peptidase 47	2	1	0	1
ANKRD45	ankyrin repeat domain 45	1	1	0	0
ANPEP	alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150)	1	1	0	0
ARHGAP17	Rho GTPase activating protein 17	1	1	0	0
ARID2	AT rich interactive domain 2 (ARID, RFX-like)	1	1	0	0
ATP7A	ATPase, Cu++ transporting, alpha polypeptide (Menkes syndrome)	1	1	0	0
ATXN7L3	ataxin 7-like 3	1	1	0	0
BBS9	Bardet-Biedl syndrome 9	1	1	0	0
BMI1	BMI1 polycomb ring finger oncogene	1	1	0	0
C2orf71	chromosome 2 open reading frame 71	1	1	0	0
C6orf89	chromosome 6 open reading frame 89	1	1	0	0
CCBE1	collagen and calcium binding EGF domains 1	1	1	0	0
CCDC117	coiled-coil domain containing 117	1	1	0	0
CHP	calcium binding protein P22	1	1	0	0
CLDN12	claudin 12	1	1	0	0

CROP	cisplatin resistance-associated overexpressed protein	1	1	0	0
DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	1	1	0	0
EEA1	early endosome antigen 1	1	1	0	0
EFNA4	ephrin-A4	1	1	0	0
ELAVL3	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 3 (Hu antigen C)	1	1	0	0
FAM126B	family with sequence similarity 126, member B	1	1	0	0
GABRB2	gamma-aminobutyric acid (GABA) A receptor, beta 2	1	1	0	0
GCS1	glucosidase I	1	1	0	0
IPO9	importin 9	1	1	0	0
KCMF1	potassium channel modulatory factor 1	1	1	0	0
LDHD	lactate dehydrogenase D	1	1	0	0
MYO1D	myosin ID	1	1	0	0
NEURL	neuralized homolog (Drosophila)	1	1	0	0
ONECUT2	one cut homeobox 2	1	1	0	0
OPRM1	opioid receptor, mu 1	1	1	0	0
PAQR5	progestin and adipoQ receptor family member V	1	1	0	0
PCNX	pecanex homolog (Drosophila)	1	1	0	0
PEG10	paternally expressed 10	1	1	0	0
PRICKLE2	prickle homolog 2 (Drosophila)	1	1	0	0
PRKCA	protein kinase C, alpha	1	1	0	0

PSKH1	protein serine kinase H1	1	1	0	0
PTCH1	patched homolog 1 (Drosophila)	1	1	0	0
RNF130	ring finger protein 130	1	1	0	0
RNF214	ring finger protein 214	1	1	0	0
SELI	selenoprotein I	1	1	0	0
SH3GLB1	SH3-domain GRB2-like endophilin B1	1	1	0	0
SOX4	SRY (sex determining region Y)-box 4	1	1	0	0
SSX2IP	synovial sarcoma, X breakpoint 2 interacting protein	1	1	0	0
STC2	stanniocalcin 2	1	1	0	0
SUV420H1	suppressor of variegation 4-20 homolog 1 (Drosophila)	1	1	0	0
TIAM1	T-cell lymphoma invasion and metastasis 1	1	1	0	0
TMEM198	transmembrane protein 198	1	1	0	0
TMEM87A	transmembrane protein 87A	1	1	0	0
TRPM3	transient receptor potential cation channel, subfamily M, member 3	1	1	0	0
UBE3A	ubiquitin protein ligase E3A (human papilloma virus E6-associated protein, Angelman syndrome)	1	1	0	0
VAMP5	vesicle-associated membrane protein 5 (myobrevin)	1	1	0	0
VEZF1	vascular endothelial zinc finger 1	1	1	0	0
WAPAL	wings apart-like homolog (Drosophila)	1	1	0	0
WIZ	widely interspaced zinc finger motifs	1	1	0	0

YTHDF3	YTH domain family, member 3	1	1	0	0
ZBTB9	zinc finger and BTB domain containing 9	1	1	0	0
ZNF664	zinc finger protein 664	1	1	0	0
ZNF831	zinc finger protein 831	1	1	0	0
ATRNL1	attractin-like 1	2	0	0	2
ARHGAP24	Rho GTPase activating protein 24	1	0	1	0
ARHGEF2	rho/rac guanine nucleotide exchange factor (GEF) 2	1	0	1	0
C1orf149	chromosome 1 open reading frame 149	1	0	1	0
CACNB1	calcium channel, voltage-dependent, beta 1 subunit	1	0	1	0
CCDC6	coiled-coil domain containing 6	1	0	1	0
CDR2L	cerebellar degeneration-related protein 2-like	1	0	1	0
CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)	1	0	1	0
CLPB	ClpB caseinolytic peptidase B homolog (E. coli)	1	0	1	0
CNOT6	CCR4-NOT transcription complex, subunit 6	1	0	1	0
CUGBP2	CUG triplet repeat, RNA binding protein 2	1	0	1	0
EIF4G2	eukaryotic translation initiation factor 4 gamma, 2	1	0	1	0
FAM122A	family with sequence similarity 122A	1	0	1	0
FAM135B	family with sequence similarity 135, member B	1	0	1	0
FAM175B	family with sequence similarity 175, member B	1	0	1	0
FBRS	fibrosin	1	0	1	0

FBXO34	F-box protein 34	1	0	1	0
FOXN2	forkhead box N2	1	0	1	0
GRIK2	glutamate receptor, ionotropic, kainate 2	1	0	1	0
KCNA6	potassium voltage-gated channel, shaker-related subfamily, member 6	1	0	1	0
KIAA0515	KIAA0515	1	0	1	0
KIAA1522	KIAA1522	1	0	1	0
MTAP	methylthioadenosine phosphorylase	1	0	1	0
MYCBP	c-myc binding protein	1	0	1	0
NAV2	neuron navigator 2	1	0	1	0
OSBPL6	oxysterol binding protein-like 6	1	0	1	0
PEX5	peroxisomal biogenesis factor 5	1	0	1	0
PTPRB	protein tyrosine phosphatase, receptor type, B	1	0	1	0
RP13-102H20.1	hypothetical protein FLJ30058	1	0	1	0
SNX18	sorting nexin 18	1	0	1	0
SP1	Sp1 transcription factor	1	0	1	0
THRB	thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)	1	0	1	0
TMEM185A	transmembrane protein 185A	1	0	1	0
VPRBP	Vpr (HIV-1) binding protein	1	0	1	0
XPO1	exportin 1 (CRM1 homolog, yeast)	1	0	1	0

YOD1	YOD1 OTU deubiquinating enzyme 1 homolog (S. cerevisiae)	1	0	1	0
ZADH2	zinc binding alcohol dehydrogenase domain containing 2	1	0	1	0
ZMAT3	zinc finger, matrin type 3	1	0	1	0
ZNF322A	zinc finger protein 322A	1	0	1	0
ZNF346	zinc finger protein 346	1	0	1	0
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	1	0	0	1
AP1S3	adaptor-related protein complex 1, sigma 3 subunit	1	0	0	1
CCDC4	coiled-coil domain containing 4	1	0	0	1
CCDC92	coiled-coil domain containing 92	1	0	0	1
CD164	CD164 molecule, sialomucin	1	0	0	1
CEBPG	CCAAT/enhancer binding protein (C/EBP), gamma	1	0	0	1
CPNE5	copine V	1	0	0	1
DCX	doublecortex; lissencephaly, X-linked (doublecortin)	1	0	0	1
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	1	0	0	1
DGKI	diacylglycerol kinase, iota	1	0	0	1
KCNU1	potassium channel, subfamily U, member 1	1	0	0	1
KCTD15	potassium channel tetramerisation domain containing 15	1	0	0	1
KIAA1324L	KIAA1324-like	1	0	0	1
LARP1	La ribonucleoprotein domain family, member 1	1	0	0	1

LIMCH1	LIM and calponin homology domains 1	1	0	0	1
LIN28	lin-28 homolog (C. elegans)	1	0	0	1
LMO4	LIM domain only 4	1	0	0	1
MARCKS	myristoylated alanine-rich protein kinase C substrate	1	0	0	1
MBOAT2	membrane bound O-acyltransferase domain containing 2	1	0	0	1
MYH10	myosin, heavy chain 10, non-muscle	1	0	0	1
NFYA	nuclear transcription factor Y, alpha	1	0	0	1
NUP50	nucleoporin 50kDa	1	0	0	1
PAX6	paired box 6	1	0	0	1
PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)	1	0	0	1
PDE7B	phosphodiesterase 7B	1	0	0	1
PFN2	profilin 2	1	0	0	1
PHF20L1	PHD finger protein 20-like 1	1	0	0	1
PHF21A	PHD finger protein 21A	1	0	0	1
PLXNA2	plexin A2	1	0	0	1
PRKCI	protein kinase C, iota	1	0	0	1
RALGPS1	Ral GEF with PH domain and SH3 binding motif 1	1	0	0	1
RPS6KA3	ribosomal protein S6 kinase, 90kDa, polypeptide 3	1	0	0	1
SAMD10	sterile alpha motif domain containing 10	1	0	0	1

SCYL2	SCY1-like 2 (S. cerevisiae)	1	0	0	1
SEMA6B	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6B	1	0	0	1
SMPD3	sphingomyelin phosphodiesterase 3, neutral membrane (neutral sphingomyelinase II)	1	0	0	1
SOX11	SRY (sex determining region Y)-box 11	1	0	0	1
SRGAP3	SLIT-ROBO Rho GTPase activating protein 3	1	0	0	1
TIMP3	TIMP metallopeptidase inhibitor 3 (Sorsby fundus dystrophy, pseudoinflammatory)	1	0	0	1
TMEM215	transmembrane protein 215	1	0	0	1
TMEM86A	transmembrane protein 86A	1	0	0	1
TTBK1	tau tubulin kinase 1	1	0	0	1
TTC9	tetratricopeptide repeat domain 9	1	0	0	1
UBE2G1	ubiquitin-conjugating enzyme E2G 1 (UBC7 homolog, yeast)	1	0	0	1
ZDHHC14	zinc finger, DHHC-type containing 14	1	0	0	1
ZDHHC5	zinc finger, DHHC-type containing 5	1	0	0	1
ZNF189	zinc finger protein 189	1	0	0	1
ZNF207	zinc finger protein 207	1	0	0	1
ZNF238	zinc finger protein 238	1	0	0	1

Novel miR-1273-2 miRNA.

	Gene name	**Conserved sites**

Target gene		total	8nt	7nt-m8	7nt-1A
ARFIP2	ADP-ribosylation factor interacting protein 2 (arfaptin 2)	2	1	1	0
ABHD4	abhydrolase domain containing 4	2	0	2	0
CAMKV	CaM kinase-like vesicle-associated	1	1	0	0
CBX1	chromobox homolog 1 (HP1 beta homolog Drosophila)	1	1	0	0
CC2D1B	coiled-coil and C2 domain containing 1B	1	1	0	0
CNTFR	ciliary neurotrophic factor receptor	1	1	0	0
ENAH	enabled homolog (Drosophila)	1	1	0	0
FMR1	fragile X mental retardation 1	1	1	0	0
FOXI2	forkhead box I2	1	1	0	0
GJA4	gap junction protein, alpha 4, 37kDa	1	1	0	0
GLT25D2	glycosyltransferase 25 domain containing 2	1	1	0	0
GTPBP1	GTP binding protein 1	1	1	0	0
INTS6	integrator complex subunit 6	1	1	0	0
KIAA0247	KIAA0247	1	1	0	0
LAG3	lymphocyte-activation gene 3	1	1	0	0
LIMS3	LIM and senescent cell antigen-like domains 3	1	1	0	0
LPIN3	lipin 3	1	1	0	0
LRCH1	leucine-rich repeats and calponin homology (CH) domain containing 1	1	1	0	0
LRFN2	leucine rich repeat and fibronectin type III domain containing 2	1	1	0	0
LRRC8A	leucine rich repeat containing 8 family, member A	1	1	0	0
MAT2A	methionine adenosyltransferase II, alpha	1	1	0	0
MS4A15	membrane-spanning 4-domains, subfamily A, member 15	1	1	0	0
OPN4	opsin 4	1	1	0	0
OPN5	opsin 5	1	1	0	0
PDE7B	phosphodiesterase 7B	1	1	0	0

PEG10	paternally expressed 10	1	1	0	0
RBPMS2	RNA binding protein with multiple splicing 2	1	1	0	0
RP1-21O18.1	kazrin	1	1	0	0
RPS6KL1	ribosomal protein S6 kinase-like 1	1	1	0	0
SCMH1	sex comb on midleg homolog 1 (Drosophila)	1	1	0	0
SIM2	single-minded homolog 2 (Drosophila)	1	1	0	0
SLC39A13	solute carrier family 39 (zinc transporter), member 13	1	1	0	0
STK40	serine/threonine kinase 40	1	1	0	0
USP47	ubiquitin specific peptidase 47	1	1	0	0
WFDC10A	WAP four-disulfide core domain 10A	1	1	0	0
AAK1	AP2 associated kinase 1	1	0	1	0
ACVR2B	activin A receptor, type IIB	1	0	1	0
ACVRL1	activin A receptor type II-like 1	1	0	1	0
APBB3	amyloid beta (A4) precursor protein-binding, family B, member 3	1	0	1	0
BAZ2A	bromodomain adjacent to zinc finger domain, 2A	1	0	1	0
BCORL1	BCL6 co-repressor-like 1	1	0	1	0
C14orf83	chromosome 14 open reading frame 83	1	0	1	0
C1orf190	chromosome 1 open reading frame 190	1	0	1	0
C9orf25	chromosome 9 open reading frame 25	1	0	1	0
CNP	2',3'-cyclic nucleotide 3' phosphodiesterase	1	0	1	0
CREBZF	CREB/ATF bZIP transcription factor	1	0	1	0
CSDC2	cold shock domain containing C2, RNA binding	1	0	1	0
DCLRE1B	DNA cross-link repair 1B (PSO2 homolog, S. cerevisiae)	1	0	1	0
DMWD	dystrophia myotonica, WD repeat containing	1	0	1	0
DNAJC5G	DnaJ (Hsp40) homolog, subfamily C, member 5 gamma	1	0	1	0
EIF2C1	eukaryotic translation initiation factor 2C, 1	1	0	1	0
GRHL2	grainyhead-like 2 (Drosophila)	1	0	1	0

HCRT	hypocretin (orexin) neuropeptide precursor	1	0	1	0
HMGB1	high-mobility group box 1	1	0	1	0
JARID1A	jumonji, AT rich interactive domain 1A	1	0	1	0
KCND1	potassium voltage-gated channel, Shal-related subfamily, member 1	1	0	1	0
KIAA1045	KIAA1045	1	0	1	0
KIAA1539	KIAA1539	1	0	1	0
KIF21B	kinesin family member 21B	1	0	1	0
LHFPL4	lipoma HMGIC fusion partner-like 4	1	0	1	0
MARK4	MAP/microtubule affinity-regulating kinase 4	1	0	1	0
MLL2	myeloid/lymphoid or mixed-lineage leukemia 2	1	0	1	0
PPP1R9B	protein phosphatase 1, regulatory (inhibitor) subunit 9B	1	0	1	0
SEMA4G	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4G	1	0	1	0
SEMA6C	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6C	1	0	1	0
SETX	senataxin	1	0	1	0
SIPA1L3	signal-induced proliferation-associated 1 like 3	1	0	1	0
SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1	1	0	1	0
SMC1A	structural maintenance of chromosomes 1A	1	0	1	0
SOCS3	suppressor of cytokine signaling 3	1	0	1	0
SOX10	SRY (sex determining region Y)-box 10	1	0	1	0
ST8SIA3	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 3	1	0	1	0
TMEM32	transmembrane protein 32	1	0	1	0
TUBA1A	tubulin, alpha 1a	1	0	1	0
VWA3A	von Willebrand factor A domain containing 3A	1	0	1	0
WDR42A	WD repeat domain 42A	1	0	1	0
AKAP11	A kinase (PRKA) anchor protein 11	1	0	0	1

ATXN7L1	ataxin 7-like 1	1	0	0	1
BMPR2	bone morphogenetic protein receptor, type II (serine/threonine kinase)	1	0	0	1
BTBD9	BTB (POZ) domain containing 9	1	0	0	1
C10orf104	chromosome 10 open reading frame 104	1	0	0	1
C16orf5	chromosome 16 open reading frame 5	1	0	0	1
C18orf62	chromosome 18 open reading frame 62	1	0	0	1
CDC42BPB	CDC42 binding protein kinase beta (DMPK-like)	1	0	0	1
CDX2	caudal type homeobox 2	1	0	0	1
CHIC1	cysteine-rich hydrophobic domain 1	1	0	0	1
CNNM1	cyclin M1	1	0	0	1
CUL5	cullin 5	1	0	0	1
DBX2	developing brain homeobox 2	1	0	0	1
DCUN1D3	DCN1, defective in cullin neddylation 1, domain containing 3 (S. cerevisiae)	1	0	0	1
ELAVL2	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)	1	0	0	1
FYCO1	FYVE and coiled-coil domain containing 1	1	0	0	1
GABRB3	gamma-aminobutyric acid (GABA) A receptor, beta 3	1	0	0	1
GSDML	gasdermin-like	1	0	0	1
HOXB3	homeobox B3	1	0	0	1
HOXC8	homeobox C8	1	0	0	1
KLF12	Kruppel-like factor 12	1	0	0	1
LARP6	La ribonucleoprotein domain family, member 6	1	0	0	1
LIN7A	lin-7 homolog A (C. elegans)	1	0	0	1
LOC116236	hypothetical protein LOC116236	1	0	0	1
OPRL1	opiate receptor-like 1	1	0	0	1
OTUB1	OTU domain, ubiquitin aldehyde binding 1	1	0	0	1
OTUB2	OTU domain, ubiquitin aldehyde binding 2	1	0	0	1

PNMAL2	PNMA-like 2	1	0	0	1
POLDIP2	polymerase (DNA-directed), delta interacting protein 2	1	0	0	1
PPP1R15B	protein phosphatase 1, regulatory (inhibitor) subunit 15B	1	0	0	1
PRCD	progressive rod-cone degeneration	1	0	0	1
PRM1	protamine 1	1	0	0	1
RAP2B	RAP2B, member of RAS oncogene family	1	0	0	1
RIC3	resistance to inhibitors of cholinesterase 3 homolog (C. elegans)	1	0	0	1
RUSC1	RUN and SH3 domain containing 1	1	0	0	1
SNIP	SNAP25-interacting protein	1	0	0	1
SOX4	SRY (sex determining region Y)-box 4	1	0	0	1
STX17	syntaxin 17	1	0	0	1
SUPT6H	suppressor of Ty 6 homolog (S. cerevisiae)	1	0	0	1
TARDBP	TAR DNA binding protein	1	0	0	1
TFAP2A	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	1	0	0	1
TMIGD1	transmembrane and immunoglobulin domain containing 1	1	0	0	1
TPRG1	tumor protein p63 regulated 1	1	0	0	1
TSPAN14	tetraspanin 14	1	0	0	1
UNC84B	unc-84 homolog B (C. elegans)	1	0	0	1
USP25	ubiquitin specific peptidase 25	1	0	0	1
VASP	vasodilator-stimulated phosphoprotein	1	0	0	1
ZBTB44	zinc finger and BTB domain containing 44	1	0	0	1
ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)	1	0	0	1
ZNF687	zinc finger protein 687	1	0	0	1

Appendix 2 List of SNPs in linkage with rs1811399

Variation Name	Chromosome	Position on Chromosome (bp)	Variant Alleles	Clinical significance	Phenotype name	Associated variant risk allele
rs73967685	2	101476598	T/A	None	N/A	N/A
rs187015324	2	101476699	A/G	None	N/A	N/A
rs189862993	2	101476794	C/T	None	N/A	N/A
rs181495383	2	101476795	G/A	None	N/A	N/A
rs12619609	2	101476824	G/A	None	N/A	N/A
rs34080732	2	101476835	-/T	None	N/A	N/A
rs59825425	2	101476874	C/T	None	N/A	N/A
rs1118509	2	101476892	T/C	None	N/A	N/A
rs114516729	2	101476903	T/G	None	N/A	N/A
rs139141368	2	101476918	C/A	None	N/A	N/A
rs72627422	2	101476971	G/T	None	N/A	N/A
rs185632116	2	101477071	A/G	None	N/A	N/A
rs2082816	2	101477085	C/T	None	N/A	N/A
rs76199684	2	101477159	G/A	None	N/A	N/A
rs189998798	2	101477225	G/A	None	N/A	N/A
rs74793064	2	101477237	C/T	None	N/A	N/A
rs75802064	2	101477238	G/A	None	N/A	N/A
rs3042733	2	101477499	-/TT	None	N/A	N/A
rs145456219	2	101477517	C/T	None	N/A	N/A
rs147680618	2	101477598	G/A	None	N/A	N/A

rs59313264	2	101477679	T/C	None	N/A	N/A
rs142354778	2	101477745	G/T	None	N/A	N/A
rs182279589	2	101477856	G/C	None	N/A	N/A
rs144587893	2	101477904	G/A	None	N/A	N/A
rs75803056	2	101477912	C/T	None	N/A	N/A
rs145701823	2	101477961	G/A	None	N/A	N/A
rs149014911	2	101477975	A/G	None	N/A	N/A
rs185120643	2	101478034	C/A	None	N/A	N/A
rs60275557	2	101478230	G/A	None	N/A	N/A
rs117623721	2	101478360	C/T	None	N/A	N/A
rs34505682	2	101478415	-/C	None	N/A	N/A
rs143817583	2	101478453	A/G	None	N/A	N/A
rs192325412	2	101478518	A/C	None	N/A	N/A
rs76376883	2	101478530	G/A	None	N/A	N/A
rs183699561	2	101478598	C/T	None	N/A	N/A
rs79741991	2	101478644	C/T	None	N/A	N/A
rs1435511	2	101478658	A/G	None	N/A	N/A
rs78536120	2	101478682	C/G	None	N/A	N/A
rs188744318	2	101478736	T/C	None	N/A	N/A
rs192859484	2	101478744	G/A	None	N/A	N/A
rs57185786	2	101478749	A/G	None	N/A	N/A
rs148151442	2	101478803	T/C	None	N/A	N/A
rs183407056	2	101478820	T/G	None	N/A	N/A
rs74366569	2	101478827	T/G	None	N/A	N/A

rs1811399	2	101479014	C/A	None	N/A	N/A
rs60386505	2	101479014	A/C	None	N/A	N/A
rs186974771	2	101479056	A/G	None	N/A	N/A
rs141328566	2	101479064	C/T	None	N/A	N/A
rs145350615	2	101479089	A/G	None	N/A	N/A
rs74734518	2	101479401	G/A	None	N/A	N/A
rs137918055	2	101479424	G/A	None	N/A	N/A
rs142441654	2	101479517	G/C	None	N/A	N/A
rs191470343	2	101479525	A/G	None	N/A	N/A
rs183636491	2	101479577	G/C	None	N/A	N/A
rs10196787	2	101479799	T/C	None	N/A	N/A
rs187996669	2	101480136	C/T	None	N/A	N/A
rs148884504	2	101480330	TC/-	None	N/A	N/A
rs193271178	2	101480400	C/T	None	N/A	N/A
rs983287	2	101480401	G/A	None	N/A	N/A
rs201415124	2	101480505	T/G	None	N/A	N/A
rs150516718	2	101480517	A/T	None	N/A	N/A
rs138379884	2	101480603	A/G	None	N/A	N/A
rs185454932	2	101480621	C/T	None	N/A	N/A
rs190362640	2	101480677	G/A	None	N/A	N/A
rs181686912	2	101480730	G/A	None	N/A	N/A
rs116473377	2	101480820	A/C	None	N/A	N/A
rs2043534	2	101480885	C/T	None	N/A	N/A

TMP_ESP_2_101480886	2	101480886	G/A	None	N/A	N/A
TMP_ESP_2_101480922	2	101480922	T/C	None	N/A	N/A
rs149334280	2	101480953	A/T	None	N/A	N/A
rs13034472	2	101481119	G/C	None	N/A	N/A
rs184329763	2	101481274	A/G	None	N/A	N/A
rs76510652	2	101481333	A/G	None	N/A	N/A
rs13017728	2	101481348	T/G	None	N/A	N/A
rs188682724	2	101481355	A/C	None	N/A	N/A
rs181111795	2	101481395	C/T	None	N/A	N/A
rs74388834	2	101481397	T/C	None	N/A	N/A
rs185371416	2	101481448	G/A	None	N/A	N/A

Appendix 3

Quantitative PCR.

Attempts were made to quantify the levels of transcript in an attempt to accurately determine the level of inhibition presented by rs1811399C. Unfortunately this proved not to be possible. Methodology used was from Ro & Yan (2010) and reads as follows:

RNA isolation

a. Harvest cells.

b. Centrifuge at ~5,000 rpm for 2 min at room temperature.

c. Remove PBS

d. Add 600 µl of Lysis/Binding buffer on ice.

e. Vortex for 40 sec to mix.

f. Add 1/10 volume of miRNA Homogenate Additive on ice and mix well by vortexing.

g. Add an equal volume (660 µl) of Acid-Phenol: Chloroform

h. Mix thoroughly by inverting the tubes several times.

i. Centrifuge at 10,000 rpm for 5 min at RT.

j. Recover the aqueous phase.

k. Add 1/3 volume of 100% ethanol at RT to the recovered aqueous phase.

l. Mix thoroughly.

m. Transfer 700 µl of the mixture into a Filter Cartridge within a collection tube.

n. Centrifuge at 10,000 rpm for 15 sec at RT.

o. Collect the filtrate (the flow-through).

p. Add 2/3 volume of 100% ethanol at RT to the flow-through.

q. Mix thoroughly.

r. Transfer up to 700 µl of the mixture into a new Filter Cartridge.

s. Centrifuge at 10,000 rpm for 15 sec at RT.

t. Discard the flow-through, and repeat until all of the filtrate mixture is through the

filter.

u. Apply 700 µl of miRNA Wash Solution 1 to the filter.

v. Centrifuge at 10,000 rpm for 15 sec at RT.

w. Discard the flow-through.

x. Apply 500 µl of miRNA Wash Solution 2/3 to the filter.

y. Centrifuge at 10,000 rpm for 15 sec at RT.

z. Discard the flow-through and repeat #27.

aa. Centrifuge at 12,000 rpm for 1 min at RT.

ab. Apply 100 µl of pre-heated (95 °C) Elution Solution to the centre of the filter, and close the cap.

ac. Leave the filter tube for 1 min at RT.

ad. Centrifuge at 12,000 rpm for 1 min at RT.

ae. Store it at −20 °C until used.

Polyadenylation

a. Set up a reaction mixture with a total volume of 50 µl in a 0.5 ml tube containing

100 ng-2 µg of small RNAs, 10 µl of 5× E-PAP Buffer, 5 µl of 25 mM MnCl2, 5 µl

of 10 mM ATP, 1 µl (2 U) of E. coli Poly(A) Polymerase I and RNase-free water

(up to 50 µl).

b. Mix well and spin the tube briefly.

c. Incubate for 1.5 hr at 37 °C.

Reverse Transcription

a.Mix 2 µg of tailed RNAs, 1 µl (1 µg) of miRTQ primers, and RNase-free water (up to 21

µl) in a PCR tube.

b. Incubate for 10 min at 65 °C and for 5 min at 4 °C.

c. Add 1 µl of 10 mM dNTP mix, 1 µl of RNaseOUT, 4 µl of 10× RT buffer, 4 µl of

0.1 M DTT, 8 µl of 25 mM MgCl2, and 1 µl of Optimax Reverse Transcriptase to the mixture.

d. Mix well and spin the tube briefly.

e. Incubate for 60 min at 50 °C and for 5 min at 85 °C for inactivation of the reaction.

f. Add 1 µl of RNase H to the mixture.

g. Incubate for 20 min at 37 °C.

h. Add 60 µl of nuclease-free water.

qPCR

a. Set a reaction with a total volume of 25 µl in a PCR tube containing 1 µ1 of small

RNA cDNAs, 1 µl (5 pm) of a miRNA-specific primer, 1 µl (5uM) of RTQ-UNIr, 12.5 µl of BioRad SYBR qPCR Master Mix and 9.5 µl of nuclease-free water.

b. Mix well and spin the tube briefly.

c. Start PCR or qPCR with the conditions: [at 95°C for 10 min, then 40 cycles (at

95°C for 15 sec, and at 60°C for 1 min)].

4. Run 2 µl of the PCR or qPCR products along with a 100 bp DNA ladder on a 2%

agarose gel.

Analysis

Whilst the method allowed me to clone and sequence the mature miRNA forms from a cDNA pool; it proved difficult to consistently accrue accurate qPCR results. Even using the Let7a primers suggested in the paper I was unable to gain a meaningful result. Upon running an aliquot of each reaction on a 3% gel I was often confronted with the following:

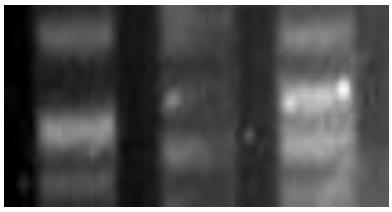
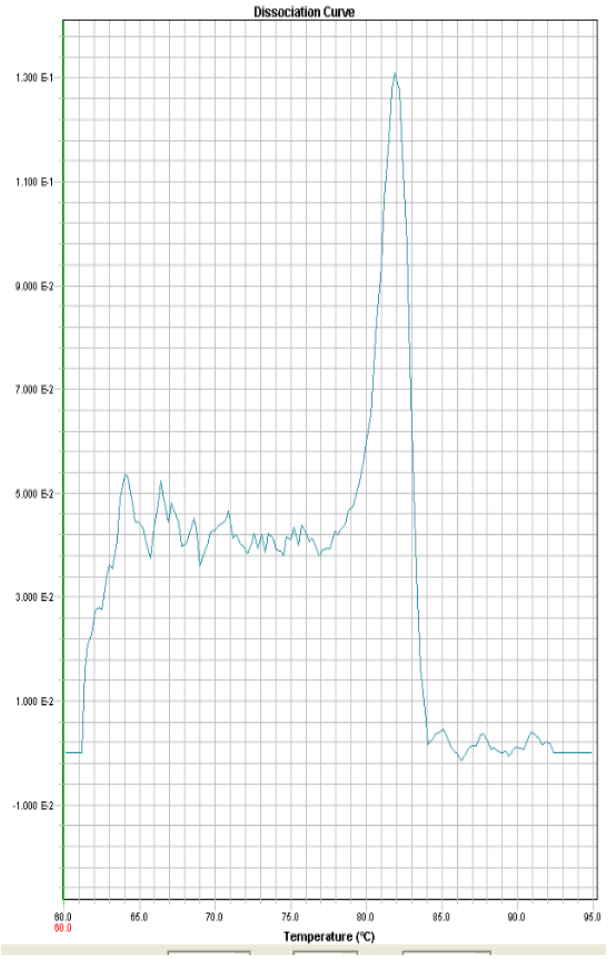
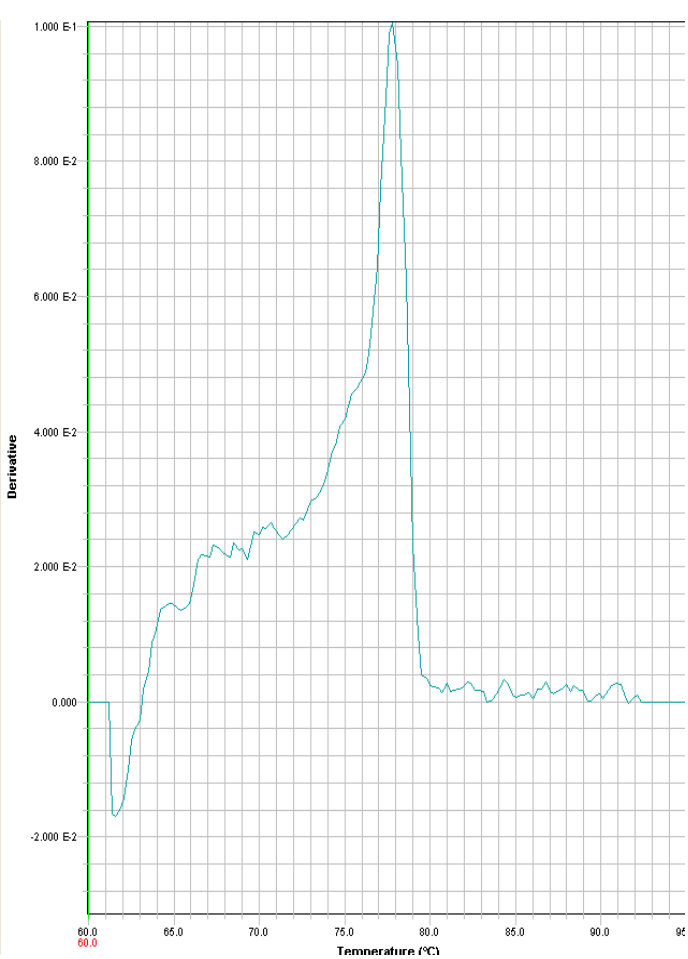


Figure App.1: Smearing and multiple bands visible on qPCR analysis of small RNA pool. Usually a symptom of the following: annealing temperature too low, insufficient salt concentration within solution, unspecific primers or in the case of reverse transcription too many secondary structures within the RNA.

Each of the hypotheses above were tested as far as reasonably possible. Annealing temperatures were increased as was salt. The only one impervious to change was the sequence of the primer, what with Let7a only being 20nt long. As such primers for other miRNA were ordered but each exhibited a similar phenomenon. The primers were tested using melt curve analysis and found to be favourable.

Let7a rs1811399





Each step prior to the qPCR was then assayed to identify any contributing errors.

RNA extraction was repeated carefully to avoid contamination or degradation. Some DNA contamination was present within the initial preparation but the utilisation of the specific miR-TQ primer would negate any influence the genomic DNA could have. Regardless, all RNA preps which were routinely treated with DNase were also now precipitated in lithium chloride. Upon re-suspension the purity was measured using spectrophotometry and any sample scoring below 1.8 was discarded.

The whole process was again repeated culminating in the qPCR much to the same result; even with different parameters and different primers.

There are many potential factors for the continued failure of this experiment:

- Poly-adenylation's random distribution of adenine on the small RNA result in various tail lengths which the qPCR detects.

- RNA secondary structures inhibit efficient RT. As the original method calls for a RT reaction at 60 degrees C and our in lab enzyme at the time could only work between 37-42 might account for the failure of the qPCR.

Appendix 4

A bioinformatics survey was conducted using the UCSC browser's Table Browser to correlate all known miRNA which have SNPs located within their precursors.

Chrom.	Start Co-ords	End Co-ords	miRNA name	SNP name	Alleles	Allele Freqs	SNP2	Alleles	Allele Freq	SNP 3	Alleles	Allele Freq
1	167967897	167967964	hsa-mir-1255b-2	rs79639536	A,G,	0.982633,0.017367,						
1	19223564	19223642	hsa-mir-1290	rs75705742	A,G,	0.016045,0.983955,						
1	41220026	41220118	hsa-mir-30e	rs112439044	C,T,	0.989680,0.010320,						
1	67094122	67094200	hsa-mir-3117	rs12402181	A,G,	0.268304,0.731696,						
1	143163749	143163822	hsa-mir-3118-2	rs75563322	A,G,	0.453881,0.546119,						
1	170120518	170120603	hsa-mir-3119-1	rs58602811	G,T,	0.014155,0.985845,						
1	170120518	170120603	hsa-mir-3119-2	rs58602811	G,T,	0.014155,0.985845,						
1	241295571	241295646	hsa-mir-3123	rs78240175	A,G,	0.030356,0.969644,						
1	249120575	249120642	hsa-mir-3124	rs12081872	C,T,	0.989159,0.010841,	rs115160731	A,C,	0.012142,0.987858,			
1	150524404	150524490	hsa-mir-4257	rs74743733	A,G,	0.015178,0.984822,						
1	209796788	209796855	hsa-mir-4260	rs7544976	G,T,	0.986301,0.013699,						
1	151518271	151518367	hsa-mir-554	rs79661940	G,T,	0.023851,0.976149,						
1	168344761	168344859	hsa-mir-557	rs78825966	C,T,	0.933651,0.066349,						
1	117637264	117637350	hsa-mir-942	rs115372145	C,T,	0.984389,0.015611,						
2	114340535	114340673	hsa-mir-1302-3	rs7589328	C,T,	0.662249,0.337751,						
2	208133998	208134148	hsa-mir-1302-4	rs116237969	A,T,	0.971813,0.028187,						

2	241395417	241395506	hsa-mir-149	rs71428439	A,G,	0.855059,0.144941,	rs2292832	C,T,	0.591793,0.408207,			
2	562160804	562161194	hsa-mir-216a	rs41291179	A,T,	0.934703,0.065297,						
2	207647957	207648032	hsa-mir-3130-1	rs2241347	A,G,	0.403285,0.596715,	rs2241347	A,G,	0.403285,0.596715,	rs115772313	A,G,	0.019948,0.980052,
2	207647957	207648032	hsa-mir-3130-2	rs115772313	A,G,	0.019948,0.980052,	rs2241347	A,G,	0.403285,0.596715,	rs2241347	A,G,	0.403285,0.596715,
2	220771222	220771286	hsa-mir-4268	rs4674470	C,T,	0.204893,0.795107,						
2	476048013	476049009	hsa-mir-559	rs114803590	C,T,	0.012576,0.987424,	rs58450758	C,T,	0.799726,0.200274,			
2	176032360	176032437	hsa-mir-933	rs79402775	A,G,	0.042466,0.957534,						
3	155377445	155378015	hsa-mir-4270	rs79922312	A,G,	0.988291,0.011709,						
3	124451285	124451363	hsa-mir-544b	rs10934682	G,T,	0.174825,0.825175,						
3	449033739	449034073	hsa-mir-564	rs114636202	A,G,	0.982220,0.017780,	rs2292181	C,G,	0.061279,0.938721,			
3	170824452	170824548	hsa-mir-569	rs73037390	A,C,	0.986770,0.013230,						
3	195426271	195426368	hsa-mir-570	rs9860655	C,T,	0.055302,0.944698,						
3	197401366	197401447	hsa-mir-922	rs114814977	A,C,	0.017780,0.982220,						
3	189547710	189547798	hsa-mir-944	rs75715827	C,T,	0.042009,0.957991,						
4	102251458	102251571	hsa-mir-1255a	rs28664200	C,T,	0.350046,0.649954,						
4	364279887	364280050	hsa-mir-1255b-1	rs6841938	A,G,	0.092609,0.907391,						
4	7461754	7461845	hsa-mir-4274	rs12512664	A,G,	0.676203,0.323797,						

4	110409853	110409951	hsa-mir-576	rs77639117	A,T,	0.950903,0.049097,						
4	115577914	115578010	hsa-mir-577	rs34115976	C,G,	0.881897,0.118103,						
4	19881101	19882004	hsa-mir-943	rs1077020	C,T,	0.277325,0.722675,						
5	179225277	179225346	hsa-mir-1229	rs2291418	C,T,	0.969732,0.030268,						
5	132763287	132763398	hsa-mir-1289-2	rs115216563	A,C,	0.971379,0.028621,						
5	153726665	153726807	hsa-mir-1294	rs13186787	A,G,	0.988104,0.011896,						
5	154065335	154065421	hsa-mir-1303	rs77055126	C,T,	0.032152,0.967848,						
5	159912358	159912457	hsa-mir-146a	rs2910164	C,G,	0.402693,0.597307,						
5	92956401	92956494	hsa-mir-2277	rs111409602	A,G,	0.024669,0.975331,						
5	153975571	153975632	hsa-mir-3141	rs936581	A,G,	0.098339,0.901661,						
5	17088993	17089083	hsa-mir-4277	rs115200817	A,G,	0.017346,0.982654,	rs12523324	A,G,	0.647832,0.352168,			
5	54466359	54466450	hsa-mir-449a	rs10061133	A,G,	0.872719,0.127281,						
5	54468089	54468181	hsa-mir-449c	rs112705432	C,T,	0.988128,0.011872,	rs35770269	A,T,	0.746582,0.253418,			
5	36147993	36148090	hsa-mir-580	rs115089112	C,T,	0.012142,0.987858,						
5	168690604	168690698	hsa-mir-585	rs62376935	C,T,	0.894975,0.105025,	rs62376934	A,G,	0.194166,0.805834,			
6	120336324	120336403	hsa-mir-3144	rs68035463	A,C,	0.220246,0.779754,	rs67106263	A,G,	0.220146,0.779854,			
6	18572014	18572111	hsa-mir-548a-1	rs12197631	G,T,	0.108676,0.891324,						

6	57254929	57255010	hsa-mir-548u	rs4994089	G,T,	0.843236,0.156764,	rs2397267	A,G,	0.500000,0.500000,	rs2894843	A,C,	0.545455,0.454545,
7	129410222	129410332	hsa-mir-182	rs76481776	C,T,	0.942726,0.057274,						
7	1062568	1062662	hsa-mir-339	rs72631831	A,G,	0.010989,0.989011,						
7	73125646	73125727	hsa-mir-4284	rs11973069	C,T,	0.970803,0.029197,						
7	73605527	73605624	hsa-mir-590	rs6971711	C,T,	0.951249,0.048751,						
7	127721912	127722012	hsa-mir-593	rs73721294	C,T,	0.979471,0.020529,						
7	158325409	158325505	hsa-mir-595	rs4909237	C,T,	0.820091,0.179909,	rs114151270	C,T,	0.988725,0.011275,			
8	129021143	129021202	hsa-mir-1206	rs2114358	C,T,	0.304762,0.695238,						
8	129162361	129162434	hsa-mir-1208	rs56863230	C,G,	0.020364,0.979636,	rs2648841	A,C,G,T,	0.000911,0.001821,0.808743,0.188525,			
8	10682882	10682953	hsa-mir-1322	rs59878596	C,T,	0.941606,0.058394,						
8	113655721	113655812	hsa-mir-2053	rs10505168	A,G,	0.621508,0.378492,						
8	26906369	26906480	hsa-mir-548h-4	rs73235381	C,T,	0.958368,0.041632,						
8	1765396	1765473	hsa-mir-596	rs116450906	C,T,	0.985256,0.014744,	rs61388742	C,T,	0.077099,0.922901,			
9	69002238	69002321	hsa-mir-1299	rs79965448	C,T,	0.437500,0.562500,	rs78546776	A,C,G,	0.250000,0.500000,0.250000,	rs76266132	C,T,	0.562500,0.437500,
9	97572243	97572339	hsa-mir-2278	rs356125	A,G,	0.039433,0.960567,						
9	18573303	18573377	hsa-mir-3152	rs13299349	A,G,	0.217890,0.782110,						

10	14478574	14478660	hsa-mir-1265	rs11259096	C,T,	0.122592,0.877408,						
10	105154009	105154158	hsa-mir-1307	rs7911488	A,G,	0.736986,0.263014,						
10	104196268	104196341	hsa-mir-146b	rs76149940	C,T,	0.987424,0.012576,						
10	135061014	135061124	hsa-mir-202	rs12355840	C,T,	0.281037,0.718963,						
10	115933863	115933938	hsa-mir-2110	rs17091403	C,T,	0.955184,0.044816,						
10	144251978	144252076	hsa-mir-4293	rs12780876	A,T,	0.238620,0.761380,	rs12220909	C,G,	0.042340,0.957660,			
10	131641562	131641638	hsa-mir-4297	rs114362263	C,T,	0.977884,0.022116,						
10	24564613	24564710	hsa-mir-603	rs11014002	C,T,	0.918765,0.081235,						
10	29833932	29834026	hsa-mir-604	rs2368393	C,T,	0.303636,0.696364,	rs2368392	C,T,	0.703574,0.296426,			
10	53059332	53059415	hsa-mir-605	rs2043556	A,G,	0.734424,0.265576,						
10	102734741	102734841	hsa-mir-608	rs4919510	C,G,	0.619683,0.380317,						
10	29891192	29891275	hsa-mir-938	rs12416605	C,T,	0.870434,0.129566,						
11	93466839	93466930	hsa-mir-1304	rs2155248	A,C,	0.890554,0.109446,	rs79759099	A,G,	0.975737,0.024263,	rs79462725	C,G,	0.024718,0.975282,
11	61582632	61582712	hsa-mir-1908	rs174561	C,T,	0.309132,0.690868,						
11	64658608	64658718	hsa-mir-192	rs11231898	A,G,	0.019645,0.980355,						
11	18409333	18409407	hsa-mir-3159	rs79957895	A,G,	0.021461,0.978539,						
11	48118333	48118410	hsa-mir-3161	rs73466882	A,T,	0.986758,0.013242,						

11	87909669	87909761	hsa-mir-3166	rs35854553	A,T,	0.111749,0.888251,						
11	126858353	126858438	hsa-mir-3167	rs670637	C,T,	0.188525,0.811475,	rs670637	C,T,	0.188525,0.811475,	rs634171	A,T,	0.811588,0.188412,
11	1880693	1880766	hsa-mir-4298	rs75966923	A,C,	0.036686,0.963314,						
11	94199660	94199746	hsa-mir-548l	rs11020790	C,T,	0.956927,0.043073,	rs13447640	C,T,	0.958363,0.041637,			
11	65211928	65212028	hsa-mir-612	rs550894	G,T,	0.829537,0.170463,	rs12803915	A,G,	0.168861,0.831139,			
12	120151438	120151529	hsa-mir-1178	rs7311975	C,T,	0.131135,0.868865,	rs74614893	A,G,	0.010399,0.989601,			
12	50627924	50627995	hsa-mir-1293	rs113699072	A,G,	0.036073,0.963927,						
12	113132838	113132981	hsa-mir-1302-1	rs74647838	A,G,	0.025571,0.974429,						
12	54385521	54385631	hsa-mir-196a-2	rs11614913	C,T,	0.624280,0.375720,						
12	26026952	26027012	hsa-mir-4302	rs11048315	A,G,T,	0.100546,0.898999,0.000455,						
12	95228173	95228289	hsa-mir-492	rs2289030	C,G,	0.231453,0.768547,						
12	65016288	65016385	hsa-mir-548c	rs17120527	A,G,	0.944242,0.055758,						
12	81329514	81329612	hsa-mir-618	rs2682818	A,C,	0.231779,0.768221,						
13	40238170	40238272	hsa-mir-4305	rs67976778	C,T,	0.552416,0.447584,						
13	92003567	92003645	hsa-mir-92a-1	rs9589207	A,G,	0.034025,0.965975,						
14	101510534	101510620	hsa-mir-1185-2	rs11844707	A,G,	0.073998,0.926002,						
14	101507699	101507782	hsa-mir-300	rs12894467	C,T,	0.429220,0.570780,						

14	101522555	101522637	hsa-mir-323b	rs56103835	C,T,	0.303832,0.696168,	rs75330474	C,T,	0.964963,0.035037,			
14	100774195	100774293	hsa-mir-345	rs72631832	C,T,	0.989030,0.010970,						
14	101489661	101489757	hsa-mir-411	rs111906529	C,T,	0.016438,0.983562,						
14	101531783	101531874	hsa-mir-412	rs61992671	A,G,	0.764813,0.235187,						
14	55344830	55344911	hsa-mir-4308	rs28477407	C,T,	0.747604,0.252396,						
14	103005980	103006063	hsa-mir-4309	rs12879262	C,G,	0.073315,0.926685,						
14	101533060	101533138	hsa-mir-656	rs58834075	C,T,	0.930201,0.069799,						
15	44085856	44085957	hsa-mir-1282	rs11269	G,T,	0.857473,0.142527,						
15	93447628	93447705	hsa-mir-3175	rs1439619	A,C,	0.603636,0.396364,	rs114901994	A,G,	0.970945,0.029055,			
15	42491767	42491864	hsa-mir-627	rs2620381	A,C,	0.915525,0.084475,						
15	70371710	70371807	hsa-mir-629	rs78212770	C,G,	0.939341,0.060659,						
15	75645951	75646026	hsa-mir-631	rs5745925	-,CT,	0.070588,0.929412,						
16	70064248	70064325	hsa-mir-1972-2	rs57629257	C,T,	0.813157,0.186843,						
16	593276	593366	hsa-mir-3176	rs8054514	G,T,	0.218063,0.781937,						
16	820182	820277	hsa-mir-662	rs74656628	A,T,	0.127235,0.872765,	rs9745376	A,G,	0.064278,0.935722,			
17	41522173	41522253	hsa-mir-2117	rs7207008	A,T,	0.426488,0.573512,						
17	925715	925799	hsa-mir-3183	rs72812091	A,G,	0.023830,0.976170,	rs2663345	C,T,	0.486212,0.513788,			

17	28444096	28444190	hsa-mir-423	rs6505162	A,C,	0.499782,0.500218,						
17	13446845	13446963	hsa-mir-548h-3	rs9913045	A,G,	0.412113,0.587887,						
17	61021575	61021673	hsa-mir-633	rs17759989	A,G,	0.968692,0.031308,						
17	66420591	66420689	hsa-mir-635	rs77279010	A,G,	0.970945,0.029055,						
18	35237097	35237178	hsa-mir-4318	rs11873400	A,G,	0.931076,0.068924,	rs117767519	G,T,	0.014731,0.985269,			
19	54191734	54191821	hsa-mir-1283-1	rs57111412	A,G,	0.766257,0.233743,						
19	54261485	54261572	hsa-mir-1283-2	rs71363366	C,G,	0.986770,0.013230,						
19	13947253	13947331	hsa-mir-27a	rs895819	A,C,G,T,	0.000168,0.340963,0.000168,0.658701,						
19	18392886	18392971	hsa-mir-3188	rs7247237	C,T,	0.720203,0.279797,	rs7247767	A,G,	0.720507,0.279493,			
19	47212549	47212602	hsa-mir-320e	rs10423365	A,G,	0.550363,0.449637,						
19	10341088	10341161	hsa-mir-4322	rs114399468	A,G,	0.017780,0.982220,						
19	54240098	54240188	hsa-mir-516b-1	rs78861479	A,G,	0.025563,0.974437,						
19	54238130	54238217	hsa-mir-518d	rs116220629	C,T,	0.983955,0.016045,	rs74704964	C,T,	0.980936,0.019064,			
19	54185412	54185499	hsa-mir-520f	rs75598818	A,G,	0.053954,0.946046,						
19	54245765	54245853	hsa-mir-520h	rs56013413	A,G,	0.080822,0.919178,						
19	14640354	14640452	hsa-mir-639	rs114121047	C,T,	0.023417,0.976583,						

20	60639857	60639941	hsa-mir-3195	rs1886009	C,T,	0.652651,0.347349,						
20	61870130	61870194	hsa-mir-3196	rs113297757	A,G,	0.023266,0.976734,	rs744591	A,C,	0.468934,0.531066,			
20	61918159	61918218	hsa-mir-4326	rs6062431	C,G,T,	0.285908,0.713638,0.000453,						
20	58883531	58883625	hsa-mir-646	rs6513496	C,T,	0.237857,0.762143,	rs6513497	G,T,	0.156534,0.843466,			
20	62573983	62574079	hsa-mir-647	rs73147065	A,G,	0.787699,0.212301,						
20	62550793	62550882	hsa-mir-941-1	rs2427556	A,G,	0.458920,0.541080,	rs6089780	A,G,	0.337125,0.662875,			
20	62551100	62551189	hsa-mir-941-2	rs113258585	C,G,	0.742922,0.257078,	rs34604519	C,G,	0.781478,0.218522,			
20	62551212	62551301	hsa-mir-941-3	rs12625445	C,G,	0.749773,0.250227,	rs35544770	A,G,	0.030538,0.969462,	rs12625454	C,G,	0.725455,0.274545,
21	15017095	15017171	hsa-mir-3118-5	rs116935353	A,T,	0.028336,0.971664,						
21	14778704	14778781	hsa-mir-3156-3	rs2747232	A,T,	0.021918,0.978082,						
22	26951177	26951289	hsa-mir-548j	rs4822739	C,G,	0.859072,0.140928,	rs12161068	C,T,	0.028285,0.971715,			
22	23165269	23165365	hsa-mir-650	rs5996397	C,G,	0.855383,0.144617,						
X	151562883	151562964	hsa-mir-105-2	rs72631816	A,T,	0.010989,0.989011,						
X	45606420	45606530	hsa-mir-222	rs72631825	A,G,	0.018519,0.981481,						
X	83480759	83480836	hsa-mir-548i-4	rs72632467	A,G,	0.090494,0.909506,						
X	80950052	80951022	hsa-mir-651	rs60180387	C,T,	0.978082,0.021918,						
X	145076301	145076378	hsa-mir-888	rs5965660	G,T,	0.103440,0.896560,						

X	145109311	145109390	hsa-mir-891a	rs5965990	A,G,	0.021511,0.978489,						
X	135633036	135633119	hsa-mir-934	rs73558572	A,G,	0.028311,0.971690,						

References

The International HapMap Consortium, Sep. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.
URL http://dx.doi.org/10.1038/nature09298

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., McVean, G. A., Oct. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
URL http://dx.doi.org/10.1038/nature09534

Abravaya, K., Phillips, B., Morimoto, R. I., Nov. 1991. Attenuation of the heat shock response in HeLa cells is mediated by the release of bound heat shock transcription factor and is modulated by changes in growth and in heat shock temperatures. *Genes & Development* **5**: 2117–2127.
URL http://dx.doi.org/10.1101/gad.5.11.2117

Adams, J. C., Seed, B., Lawler, J., Sep. 1998. Muskelin, a novel intracellular mediator of cell adhesive and cytoskeletal responses to thrombospondin-1. *The EMBO Journal* **17**: 4964–4974.
URL http://dx.doi.org/10.1093/emboj/17.17.4964

Akhtar, R. A., Reddy, A. B., Maywood, E. S., Clayton, J. D., King, V. M., Smith, A. G., Gant, T. W., Hastings, M. H., Kyriacou, C. P., Apr. 2002. Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus. *Current Biology* **12**: 540–550.
URL http://view.ncbi.nlm.nih.gov/pubmed/11937022

Albers, C. A., Paul, D. S., Schulze, H., Freson, K., Stephens, J. C., Smethurst, P. A., Jolley, J. D., Cvejic, A., Kostadima, M., Bertone, P., Breuning, M. H., Debili, N., Deloukas, P., Favier, R., Fiedler, J., Hobbs, C. M., Huang, N., Hurles, M. E., Kiddle, G., Krapels, I., Nurden, P., Ruivenkamp, C. A., Sambrook, J. G., Smith, K., Stemple, D. L., Strauss, G., Thys, C., van Geet, C., Newbury-Ecob, R., Ouwehand, W. H., Ghevaert, C., Apr. 2012. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature Genetics* **44**: 435-439
URL http://dx.doi.org/10.1038/ng.1083

Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H. H., Spatafora, J. W., Carrington, J. C., Dec. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in arabidopsis thaliana. *Nature Genetics* **36**: 1282–1290.
URL http://dx.doi.org/10.1038/ng1478

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T., Margalit, H., 2005. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Research* **33**: 2697–2706.
URL http://dx.doi.org/10.1093/nar/gki567

Alvarez-Saavedra, M., Antoun, G., Yanagiya, A., Oliva-Hernandez, R., Cornejo-Palma, D., Perez-Iratxeta, C., Sonenberg, N., Cheng, H.-Y. Y., Feb. 2011. miRNA-132 orchestrates chromatin remodeling and translational control of the circadian clock. *Human Molecular Genetics* **20**: 731–751.
URL http://dx.doi.org/10.1093/hmg/ddq519

Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., Jewell, D., May 2003. MicroRNAs and other tiny endogenous RNAs in c. elegans. *Current Biology* **13**: 807–818.
URL http://view.ncbi.nlm.nih.gov/pubmed/12747828

Balsalobre, A., Damiola, F., Schibler, U., Jun. 1998. A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell* **93**: 929–937.
URL http://view.ncbi.nlm.nih.gov/pubmed/9635423

Bandele, O. J., Wang, X., Campbell, M. R., Pittman, G. S., Bell, D. A., Jan. 2011. Human single-nucleotide polymorphisms alter p53 sequence-specific binding at gene regulatory elements. *Nucleic Acids Research* **39**: 178–189.
URL http://dx.doi.org/10.1093/nar/gkq764

Bar, I., Dec. 2000. The evolution of cortical development. an hypothesis based on the role of the reelin signaling pathway. *Trends in Neurosciences* **23**: 633–638.
URL http://dx.doi.org/10.1016/s0166-2236(00)01675-1

Barik, S., Sep. 2008. An intronic microRNA silences genes that are functionally antagonistic to its host gene. *Nucleic Acids Research* **36**: 5232–5241.
URL http://dx.doi.org/10.1093/nar/gkn513

Barr, I., Smith, A. T., Senturia, R., Chen, Y., Scheidemantle, B. D., Burstyn, J. N., Guo, F., May 2011. DiGeorge critical region 8 (DGCR8) is a double-cysteine-ligated heme protein. *The Journal of Biological Chemistry* **286**: 16716–16725.
URL http://dx.doi.org/10.1074/jbc.m110.180844

Barreiro, L. B., Laval, G., Quach, H., Patin, E., Quintana-Murci, L., Mar. 2008. Natural selection has driven population differentiation in modern humans. *Nature Genetics* **40**: 340–345.
URL http://dx.doi.org/10.1038/ng.78

Bartlett, D. W., Davis, M. E., Jan. 2006. Insights into the kinetics of siRNA-mediated gene silencing from live-cell and live-animal bioluminescent imaging. *Nucleic Acids Research* **34**: 322–333.
URL http://dx.doi.org/10.1093/nar/gkj439

Baskerville, S., Bartel, D. P., Mar. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**: 241–247.
URL http://dx.doi.org/10.1261/rna.7240905

Borchert, G. M., Holton, N. W., Williams, J. D., Hernan, W. L., Bishop, I. P., Dembosky, J. A., Elste, J. E., Gregoire, N. S., Kim, J.-A. A., Koehler, W. W., Lengerich, J. C., Medema, A. A., Nguyen, M. A., Ower, G. D., Rarick, M. A., Strong, B. N., Tardi, N. J., Tasker, N. M., Wozniak, D. J., Gatto, C., Larson, E. D., May 2011. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mobile Genetic Elements* **1**: 8–17.
URL http://dx.doi.org/10.4161/mge.1.1.15766

Botstein, D., White, R. L., Skolnick, M., Davis, R. W., May 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314–331.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1686077/

Brown, F. A., Webb, H. M., Oct. 1948. Temperature relations of an endogenous daily rhythmicity in the fiddler crab. *Physiological Zoology* **21**: 371–381.
URL http://view.ncbi.nlm.nih.gov/pubmed/18891156

Cai, Y., Wang, J., Ren, C., Ittmann, M., 2012. Frequent heterogeneous missense mutations of GGAP2 in prostate cancer: implications for tumor biology, clonality and mutation analysis. *PloS One* **7**: e32708
URL http://dx.doi.org/10.1371/journal.pone.0032708

Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J., Schubert, M., Jan. 2011. microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evolution & Development* **13**: 15–27.
URL http://dx.doi.org/10.1111/j.1525-142x.2010.00452.x

Carson, J. P., Zhang, N., Frampton, G. M., Gerry, N. P., Lenburg, M. E., Christman, M. F., Mar. 2004. Pharmacogenomic identification of targets for adjuvant therapy with the topoisomerase poison camptothecin. *Cancer Research* **64**: 2096–2104.
URL http://view.ncbi.nlm.nih.gov/pubmed/15026349

Carthew, R. W., Sontheimer, E. J., Feb. 2009. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**: 642–655.
URL http://dx.doi.org/10.1016/j.cell.2009.01.035

Cavadini, G., Petrzilka, S., Kohler, P., Jud, C., Tobler, I., Birchler, T., Fontana, A., Jul. 2007. TNF-alpha suppresses the expression of clock genes by interfering with e-box-mediated transcription. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 12843–12848.
URL http://dx.doi.org/10.1073/pnas.0701466104

Cerqueira, N. M., Fernandes, P. A., Ramos, M. J., 2007. Understanding ribonucleotide reductase inactivation by gemcitabine. *Chemistry* **13**: 8507–8515.
URL http://dx.doi.org/10.1002/chem.200700260

Chang, A. C. M., Reddel, R. R., Jun. 1998. Identification of a second stanniocalcin cDNA in mouse and human: Stanniocalcin 2. *Molecular and Cellular Endocrinology* **141**: 95–99.
URL http://dx.doi.org/10.1016/s0303-7207(98)00097-5

Chen, K., Rajewsky, N., Feb. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics* **8**: 93–103.
URL http://dx.doi.org/10.1038/nrg1990

Chen, Z., McKnight, S. L., Dec. 2007. A conserved DNA damage response pathway responsible for coupling the cell division cycle to the circadian and metabolic cycles. *Cell Cycle* **6**: 2906–2912.
URL http://view.ncbi.nlm.nih.gov/pubmed/18000400

Cuperus, J. T., Fahlgren, N., Carrington, J. C., Feb. 2011. Evolution and functional diversification of MIRNA genes. *The Plant Cell* **23**: 431–442.
URL http://dx.doi.org/10.1105/tpc.110.082784

Dame, R. T., May 2005. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular Microbiology* **56**: 858–870.
URL http://dx.doi.org/10.1111/j.1365-2958.2005.04598.x

de Wit, E., Linsen, S. E., Cuppen, E., Berezikov, E., Nov. 2009. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Research* **19**: 2064–2074.
URL http://dx.doi.org/10.1101/gr.093781.109

Deltour, S., Pinte, S., Guérardel, C., Leprince, D., Sep. 2001. Characterization of HRG22, a human homologue of the putative tumor suppressor gene HIC1. *Biochemical and Biophysical Research Communications* **287**: 427–434.
URL http://dx.doi.org/10.1006/bbrc.2001.5624

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F. F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M. M., Gascuel, O., Jul. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**: 465–469.
URL http://dx.doi.org/10.1093/nar/gkn180

Dioum, E. M., Rutter, J., Tuckerman, J. R., Gonzalez, G., Gilles-Gonzalez, M.-A., McKnight, S. L., Dec. 2002. NPAS2: A Gas-Responsive transcription factor. *Science* **298**: 2385–2387.
URL http://dx.doi.org/10.1126/science.1078456

Dragunow, M., Faull, R., Sep. 1989. The use of c-fos as a metabolic marker in neuronal pathway tracing. *Journal of Neuroscience Methods* **29**: 261–265.
URL http://view.ncbi.nlm.nih.gov/pubmed/2507830

Duan, R., Pak, C., Jin, P., May 2007. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Human Molecular Genetics* **16**: 1124–1131.
URL http://dx.doi.org/10.1093/hmg/ddm062

Dubois, M.-F., Nguyen, V. T., Bonnet, F., Bensaude, O., Marshall, N. F., Dahmus, G. K., Dahmus, M. E., Mar. 1999. Heat shock of HeLa cells inactivates a nuclear protein phosphatase specific for dephosphorylation of the c-terminal domain of RNA polymerase II. *Nucleic Acids Research* **27**: 1338–1344.
URL http://dx.doi.org/10.1093/nar/27.5.1338

Dunham, C. M., Dioum, E. M., Tuckerman, J. R., Gonzalez, G., Scott, W. G., Gilles-Gonzalez, M.-A. A., Jul. 2003. A distal arginine in oxygen-sensing heme-PAS domains is essential to ligand binding, signal transduction, and structure. *Biochemistry* **42**: 7701–7708.
URL http://dx.doi.org/10.1021/bi0343370

Efferth, T., Fu, Y.-J. J., Zu, Y.-G. G., Schwarz, G., Badireenath, V. S., Wink, M., 2007. Molecular target-guided tumor therapy with natural products derived from traditional chinese medicine. *Current Medicinal Chemistry* **14**: 2024–2032.
URL http://view.ncbi.nlm.nih.gov/pubmed/17691944

Engels, B. M., Hutvagner, G., Oct. 2006. Principles and effects of microRNA-mediated post-transcriptional gene regulation. *Oncogene* **25**: 6163–6169.
URL http://dx.doi.org/10.1038/sj.onc.1209909

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B. E., May 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.
URL http://dx.doi.org/10.1038/nature09906

Fanous, A. H., Zhou, B., Aggen, S. H., Bergen, S. E., Amdur, R. L., Duan, J., Sanders, A. R., Shi, J., Mowry, B. J., Olincy, A., Amin, F., Cloninger, R. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Freedman, R., Dudbridge, F., Holmans, P. A., Ripke, S., Gejman, P. V., Kendler, K. S., Levinson, D. F., Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, Dec. 2012. Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *The American Journal of Psychiatry* **169**: 1309–1317.
URL http://dx.doi.org/10.1176/appi.ajp.2012.12020218

Fernandez-Valverde, S. L., Taft, R. J., Mattick, J. S., Oct. 2010. Dynamic isomiR regulation in drosophila development. *RNA* **16**: 1881–1888.
URL http://dx.doi.org/10.1261/rna.2379610

Filipowicz, W., Jul. 2005. RNAi: the nuts and bolts of the RISC machine. *Cell* **122**: 17–20.
URL http://dx.doi.org/10.1016/j.cell.2005.06.023

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., Mello, C. C., Feb. 1998. Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature* **391**: 806–811.
URL http://dx.doi.org/10.1038/35888

Franken, P., Dijk, D.-J. J., May 2009. Circadian clock genes and sleep homeostasis. *The European Journal of Neuroscience* **29**: 1820–1829.
URL http://dx.doi.org/10.1111/j.1460-9568.2009.06723.x

Franken, P., Dudley, C. A., Estill, S. J. J., Barakat, M., Thomason, R., O'Hara, B. F., McKnight, S. L., May 2006. NPAS2 as a transcriptional regulator of non-rapid eye movement sleep: genotype and sex interactions. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 7118–7123.
URL http://dx.doi.org/10.1073/pnas.0602006103

Garcia, J. A., Zhang, D., Estill, S. J., Michnoff, C., Rutter, J., Reick, M., Scott, K., Diaz-Arrastia, R., McKnight, S. L., Jun. 2000. Impaired cued and contextual memory in NPAS2-deficient mice. *Science* **288**: 2226–2230.
URL http://dx.doi.org/10.1126/science.288.5474.2226

Gatfield, D., Le Martelot, G., Vejnar, C. E., Gerlach, D., Schaad, O., Fleury-Olela, F., Ruskeepää, A.-L. L., Oresic, M., Esau, C. C., Zdobnov, E. M., Schibler, U., Jun. 2009. Integration of microRNA miR-122 in hepatic circadian gene expression. *Genes & Development* **23**: 1313–1326.
URL http://dx.doi.org/10.1101/gad.1781009

Gatfield, D., Schibler, U., May 2007. Proteasomes keep the circadian clock ticking. *Science* **316**: 1135–1136.
URL http://dx.doi.org/10.1126/science.1144165

Gillesgonzalez, M., Gonzalez, G., Jan. 2005. Heme-based sensors: defining characteristics, recent developments, and regulatory hypotheses. *Journal of Inorganic Biochemistry* **99**: 1–22.
URL http://dx.doi.org/10.1016/j.jinorgbio.2004.11.006

Golan, D., Levy, C., Friedman, B., Shomron, N., Apr. 2010. Biased hosting of intronic microRNA genes. *Bioinformatics* **26**: 992–995.
URL http://dx.doi.org/10.1093/bioinformatics/btq077

Gong, J., Tong, Y., Zhang, H.-M. M., Wang, K., Hu, T., Shan, G., Sun, J., Guo, A.-Y. Y., Jan. 2012. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. *Human Mutation* **33**: 254–263.
URL http://dx.doi.org/10.1002/humu.21641

Gong, X., Bacchelli, E., Blasi, F., Toma, C., Betancur, C., Chaste, P., Delorme, R., Durand, C. M., Fauchereau, F., Botros, H. G. G., Leboyer, M., Mouren-Simeoni, M.-C. C., Nygren, G., Anckarsäter, H., Rastam, M., Gillberg, I. C., Gillberg, C., Moreno-De-Luca, D., Carone, S., Nummela, I., Rossi, M., Battaglia, A., International Molecular Genetic Study of Autism Consortium (IMGSAC), Jarvela, I., Maestrini, E., Bourgeron, T., Sep. 2008. Analysis of x chromosome inactivation in autism spectrum disorders. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* **147B**: 830–835.
URL http://dx.doi.org/10.1002/ajmg.b.30688

Gregory, R., Chendrimada, T., Shiekhattar, R., 2006. MicroRNA biogenesis: Isolation and characterization of the microprocessor complex. In: Ying, S.-Y. (Ed.), *MicroRNA Protocols. Vol. 342 of Methods in Molecular Biology™*. Humana Press, pp. 33–47.
URL http://dx.doi.org/10.1385/1-59745-123-1\%253a33

Gregory, R. I., Shiekhattar, R., May 2005. MicroRNA biogenesis and cancer. *Cancer Research* **65**: 3509–3512.
URL http://dx.doi.org/10.1158/0008-5472.can-05-0298

Griffiths-Jones, S., Hui, J. H., Marco, A., Ronshaugen, M., Feb. 2011. MicroRNA evolution by arm switching. *EMBO Reports* **12**: 172–177.
URL http://dx.doi.org/10.1038/embor.2010.191

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., Bartel, D. P., Oct. 2008. Early origins and evolution of microRNAs and piwi-interacting RNAs in animals. *Nature* **455**: 1193–1197.
URL http://dx.doi.org/10.1038/nature07415

Grupe, A., Li, Y., Rowland, C., Nowotny, P., Hinrichs, A. L., Smemo, S., Kauwe, J. S., Maxwell, T. J., Cherny, S., Doil, L., Tacey, K., van Luchene, R., Myers, A., Wavrant-De Vrièze, F., Kaleem, M., Hollingworth, P., Jehu, L., Foy, C., Archer, N., Hamilton, G., Holmans, P., Morris, C. M., Catanese, J., Sninsky, J., White, T. J., Powell, J., Hardy, J., O'Donovan, M., Lovestone, S., Jones, L., Morris, J. C., Thal, L., Owen, M., Williams, J., Goate, A., Jan. 2006. A scan of chromosome 10 identifies a novel locus showing strong association with late-onset alzheimer disease. *American Journal of Human Genetics* **78**: 78–88.
URL http://dx.doi.org/10.1086/498851

Gu, M., Lima, C. D., Feb. 2005. Processing the message: structural insights into capping and decapping mRNA. *Current Opinion in Structural Biology* **15**: 99–106.
URL http://dx.doi.org/10.1016/j.sbi.2005.01.009

Guhaniyogi, J., Brewer, G., Mar. 2001. Regulation of mRNA stability in mammalian cells. *Gene* **265**: 11–23.
URL http://dx.doi.org/10.1016/s0378-1119(01)00350-x

Guil, S., Cáceres, J. F., Jul. 2007. The multifunctional RNA-binding protein hnRNP a1 is required for processing of miR-18a. *Nature Structural & Molecular Biology* **14**: 591–596.
URL http://dx.doi.org/10.1038/nsmb1250

Guo, H., Ingolia, N. T., Weissman, J. S., Bartel, D. P., Aug. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–840.
URL http://dx.doi.org/10.1038/nature09267

Hadjiagapiou, C., Giannoni, F., Funahashi, T., Skarosi, S. F., Davidson, N. O., May 1994. Molecular cloning of a human small intestinal apolipoprotein b mRNA editing protein. *Nucleic Acids Research* **22**: 1874–1879.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308087/

Hagarman, J. A., Motley, M. P., Kristjansdottir, K., Soloway, P. D., Jan. 2013. Coordinate regulation of DNA methylation and H3K27me3 in mouse embryonic stem cells. *PLoS ONE* **8**: e53880+.
URL http://dx.doi.org/10.1371/journal.pone.0053880

Han, J., Lee, Y., Yeom, K.-H. H., Nam, J.-W. W., Heo, I., Rhee, J.-K. K., Sohn, S. Y. Y., Cho, Y., Zhang, B.-T. T., Kim, V. N., Jun. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**: 887–901.
URL http://dx.doi.org/10.1016/j.cell.2006.03.043

Harnprasopwat, R., Ha, D., Toyoshima, T., Lodish, H., Tojo, A., Kotani, A., Aug. 2010. Alteration of processing induced by a single nucleotide polymorphism in pri-miR-126. *Biochemical and Biophysical Research Communications* **399**: 117–122.
URL http://dx.doi.org/10.1016/j.bbrc.2010.07.009

He, C., Li, Z., Chen, P., Huang, H., Hurst, L. D., Chen, J., Jan. 2012. Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA–host gene coevolution. *Nucleic Acids Research* **40**: 4002–4012.
URL http://dx.doi.org/10.1093/nar/gkr1312

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I., Stadler, P., of Bioinformatics Computer Labs 2004, T. S., 2005, Feb. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**: 25.
URL http://dx.doi.org/10.1186/1471-2164-7-25

Hinske, L. C. C., Galante, P. A., Kuo, W. P., Ohno-Machado, L., Oct. 2010. A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics* **11**: 533.
URL http://dx.doi.org/10.1186/1471-2164-11-533

Hoffman, A. E., Zheng, T., Ba, Y., Zhu, Y., Sep. 2008. The circadian gene NPAS2, a putative tumor suppressor, is involved in DNA damage response. *Molecular Cancer Research* **6:** 1461–1468.
URL http://dx.doi.org/10.1158/1541-7786.mcr-07-2094

Hong, C. I., Zámborszky, J., Csikász-Nagy, A., May 2009. Minimum criteria for DNA damage-induced phase advances in circadian rhythms. *PLoS Computational Biology* **5**: e1000384+.
URL http://dx.doi.org/10.1371/journal.pcbi.1000384

Hopfer, S. M., Sunderman, F. W., Dec. 1988. Hypothermia and deranged circadian rhythm of core body temperature in nickel chloride-treated rats. *Research Communications in Chemical Pathology and Pharmacology* **62**: 495–505.
URL http://view.ncbi.nlm.nih.gov/pubmed/3222527

Hu, V. W., Sarachana, T., Kim, K. S. S., Nguyen, A., Kulkarni, S., Steinberg, M. E., Luu, T., Lai, Y., Lee, N. H., Apr. 2009. Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Research* **2**: 78–97.
URL http://dx.doi.org/10.1002/aur.73

Huang, H.-S., Matevossian, A., Whittle, C., Kim, S. Y., Schumacher, A., Baker, S. P., Akbarian, S., Oct. 2007. Prefrontal dysfunction in schizophrenia involves Mixed-Lineage leukemia 1-Regulated histone methylation at GABAergic gene promoters. *The Journal of Neuroscience* **27**: 11254–11262.
URL http://dx.doi.org/10.1523/jneurosci.3272-07.2007

Hutvagner, G., Oct. 2006. MicroRNAs and cancer: issue summary. *Oncogene* **25**: 6154–6155.
URL http://dx.doi.org/10.1038/sj.onc.1209917

Hwang, H.-W., Wentzel, E. A., Mendell, J. T., Apr. 2009. Cell–cell contact globally activates microRNA biogenesis. *Proceedings of the National Academy of Sciences* **106**: 7016–7021.
URL http://dx.doi.org/10.1073/pnas.0811523106

Jetten, A. M., Kurebayashi, S., Ueda, E., 2001. The ROR nuclear orphan receptor subfamily: critical regulators of multiple biological processes. *Progress in Nucleic Acid Research and Molecular Biology* **69**: 205–247.
URL http://view.ncbi.nlm.nih.gov/pubmed/11550795

Johnson, C. P., Myers, S. M., , the Council on Children With Disabilities, Nov. 2007. Identification and evaluation of children with autism spectrum disorders. *Pediatrics* **120**: 1183–1215.
URL http://dx.doi.org/10.1542/peds.2007-2361

Kanner, L., 1968. Autistic disturbances of affective contact. *Acta Paedopsychiatrica* **35**: 100–136.
URL http://view.ncbi.nlm.nih.gov/pubmed/4880460

Kawamata, T., Tomari, Y., Jul. 2010. Making RISC. *Trends in Biochemical Sciences* **35**: 368–376.
URL http://dx.doi.org/10.1016/j.tibs.2010.03.009

Khvorova, A., Reynolds, A., Jayasena, S. D., Oct. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**: 209–216.
URL http://view.ncbi.nlm.nih.gov/pubmed/14567918

Kim, D. D. Y., Kim, T. T. Y., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., Gabriel, A., Sep. 2004. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Research* **14**: 1719–1725.
URL http://dx.doi.org/10.1101/gr.2855504

Knight, J. C., 2009. Human genetic diversity : functional consequences for health and disease. Oxford University Press.
URL http://www.worldcat.org/isbn/9780199227709

Köhler, A., Hurt, E., Oct. 2007. Exporting RNA from the nucleus to the cytoplasm. *Nature reviews Molecular Cell Biology* **8**: 761–773.
URL http://dx.doi.org/10.1038/nrm2255

Konopka, R. J., Benzer, S., Sep. 1971. Clock mutants of drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **68**: 2112–2116.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC389363/

Kozak, M., Nov. 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**: 13–37.
URL http://dx.doi.org/10.1016/j.gene.2005.06.037

Kuhn, R., Cahn, C. H., Sep. 2004. Eugen bleuler's concepts of psychopathology. *History of Psychiatry* **15**: 361–366.
URL http://dx.doi.org/10.1177/0957154x04044603

Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., Monaco, A. P., Oct. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**: 519–523.
URL http://dx.doi.org/10.1038/35097076

Lamia, K. A., Storch, K.-F., Weitz, C. J., Sep. 2008. Physiological significance of a peripheral tissue circadian clock. *Proceedings of the National Academy of Sciences.* **105**: 15172–15177
URL http://dx.doi.org/10.1073/pnas.0806717105

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.-U. U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.-I. I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M., Tuschl, T., Jun. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**: 1401–1414.
URL http://dx.doi.org/10.1016/j.cell.2007.04.040

Lavery, D. J., Schibler, U., Oct. 1993. Circadian transcription of the cholesterol 7 alpha hydroxylase gene may involve the liver-enriched bZIP protein DBP. *Genes & Development* **7**: 1871–1884.
URL http://view.ncbi.nlm.nih.gov/pubmed/8405996

Lee, B.-K., Bhinge, A. A., Battenhouse, A., McDaniell, R. M., Liu, Z., Song, L., Ni, Y., Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E., Iyer, V. R., Jan. 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Research* **22**: 9–24.
URL http://dx.doi.org/10.1101/gr.127597.111

Lee, J. A., Carvalho, C. M. B., Lupski, J. R., Dec. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**: 1235–1247.
URL http://dx.doi.org/10.1016/j.cell.2007.11.037

Lee, Y., Jeon, K., Lee, J.-T. T., Kim, S., Kim, V. N., Sep. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal* **21**: 4663–4670.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC126204/

Lehnert, S., Kapitonov, V., Thilakarathne, P. J., Schuit, F. C., 2011. Modeling the asymmetric evolution of a mouse and rat-specific microRNA gene cluster intron 10 of the sfmbt2 gene. *BMC Genomics* **12:** 257
URL http://dx.doi.org/10.1186/1471-2164-12-257

Leung, J. W.-C. W., Ghosal, G., Wang, W., Shen, X., Wang, J., Li, L., Chen, J., Mar. 2013. Alpha thalassemia/mental retardation syndrome x-linked gene product ATRX is required for proper replication restart and cellular resistance to replication stress. *The Journal of Biological Chemistry* **288**: 6342–6350.
URL http://dx.doi.org/10.1074/jbc.m112.411603

Lewis, B. P., Burge, C. B., Bartel, D. P., Jan. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**: 15–20.
URL http://dx.doi.org/10.1016/j.cell.2004.12.035

Li, J., Pauley, A. M., Myers, R. L., Shuang, R., Brashler, J. R., Yan, R., Buhl, A. E., Ruble, C., Gurney, M. E., Sep. 2002. SEL-10 interacts with presenilin 1, facilitates its ubiquitination, and alters a-beta peptide production. *Journal of Neurochemistry* **82**: 1540–1548.
URL http://dx.doi.org/10.1046/j.1471-4159.2002.01105.x

Li, S., Mead, E., Liang, S., Tu, Z., Dec. 2009. Direct sequencing and expression analysis of a large number of miRNAs in aedes aegypti and a multi-species survey of novel mosquito miRNAs. *BMC Genomics* **10**: 581-596.
URL http://dx.doi.org/10.1186/1471-2164-10-581

Lin, S.-L. L., Miller, J. D., Ying, S.-Y. Y., 2006. Intronic microRNA (miRNA). *Journal of Biomedicine & Biotechnology* **2006**: 26818
URL http://dx.doi.org/10.1155/jbb/2006/26818

Lo-Castro, A., Galasso, C., Cerminara, C., El-Malhany, N., Benedetti, S., Nardone, A. M., Curatolo, P., Jun. 2009. Association of syndromic mental retardation and autism with 22q11.2 duplication. *Neuropediatrics* **40**: 137–140.
URL http://dx.doi.org/10.1055/s-0029-1237724

Luciano, D. J., Mirsky, H., Vendetti, N. J., Maas, S., Aug. 2004. RNA editing of a miRNA precursor. *RNA* **10**: 1174–1177.
URL http://dx.doi.org/10.1261/rna.7350304

Magnone, M. C. C., Jacobmeier, B., Bertolucci, C., Foà, A., Albrecht, U., Feb. 2005. Circadian expression of the clock gene per2 is altered in the ruin lizard (podarcis sicula) when temperature changes. *Molecular Brain Research* **133**: 281–285.
URL http://dx.doi.org/10.1016/j.molbrainres.2004.10.014

Maletic, V., Robinson, M., Oakes, T., Iyengar, S., Ball, S. G., Russell, J., Nov. 2007. Neurobiology of depression: an integrated view of key findings. *International Journal of Clinical Practice* **61**: 2030–2040.
URL http://dx.doi.org/10.1111/j.1742-1241.2007.01602.x

Mayr, F., Schütz, A., Döge, N., Heinemann, U., Aug. 2012. The lin28 cold-shock domain remodels pre-let-7 microRNA. *Nucleic Acids Research* **40**: 7492–7506.
URL http://dx.doi.org/10.1093/nar/gks355

McNamara, P., Seo, S.-b., Rudic, R. D., Sehgal, A., Chakravarti, D., FitzGerald, G. A., Jun. 2001. Regulation of CLOCK and MOP4 by nuclear hormone receptors in the vasculature. *Cell* **105**: 877–889.
URL http://dx.doi.org/10.1016/s0092-8674(01)00401-9

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P., Kaessmann, H., Jan. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Research* **23**: 34–45.
URL http://dx.doi.org/10.1101/gr.140269.112

Monteys, A. M. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y., Davidson, B. L., Mar. 2010. Structure and activity of putative intronic miRNA promoters. *RNA* **16**: 495–505.
URL http://dx.doi.org/10.1261/rna.1731910

Mori, M., Triboulet, R., Mohseni, M., Schlegelmilch, K., Shrestha, K., Camargo, F. D., Gregory, R. I., Feb. 2014. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell* **156**: 893–906.
URL http://view.ncbi.nlm.nih.gov/pubmed/24581491

Muhle, R., Trentacoste, S. V., Rapin, I., May 2004. The genetics of autism. *Pediatrics* **113**: e472–486.
URL http://dx.doi.org/10.1542/peds.113.5.e472

Mühlebach, M. D., Mateo, M., Sinn, P. L., Prüfer, S., Uhlig, K. M., Leonard, V. H., Navaratnarajah, C. K., Frenzke, M., Wong, X. X., Sawatsky, B., Ramachandran, S., McCray, P. B., Cichutek, K., von Messling, V., Lopez, M., Cattaneo, R., Dec. 2011. Adherens junction protein nectin-4 is the epithelial receptor for measles virus. *Nature* **480**: 530–533.
URL http://dx.doi.org/10.1038/nature10639

Musiyenko, A., Bitko, V., Barik, S., Mar. 2008. Ectopic expression of miR-126*, an intronic product of the vascular endothelial EGF-like 7 gene, regulates prostein translation and invasiveness of prostate cancer LNCaP cells. *Journal of Molecular Medicine* **86**: 313–322.
URL http://dx.doi.org/10.1007/s00109-007-0296-9

Napoli, C., Lemieux, C., Jorgensen, R., Apr. 1990. Introduction of a chimeric chalcone synthase gene into petunia results in reversible Co-Suppression of homologous genes in trans. *The Plant Cell* **2**: 279–289.
URL http://dx.doi.org/10.1105/tpc.2.4.279

Nava, C., Keren, B., Mignot, C., Rastetter, A., Chantot-Bastaraud, S., Faudet, A., Fonteneau, E., Amiet, C., Laurent, C., Jacquette, A., Whalen, S., Afenjar, A., Périsse, D., Doummar, D., Dorison, N., Leboyer, M., Siffroi, J.-P. P., Cohen, D., Brice, A., Héron, D., Depienne, C., May 2013. Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders. *European Journal of Human Genetics* **22**: 71–78
URL http://dx.doi.org/10.1038/ejhg.2013.88

Nicholas, B., Owen, M. J., Wimpory, D. C., Caspari, T. 2008 Autism-associated SNPs in the clock genes npas2, per1 and the homeobox gene en2 alter DNA sequences that show characteristics of microRNA genes. *Nature Precedings* **713**.
URL http://dx.doi.org/10.1038/npre.2008.2366.1

Nicholas, B., Rudrasingham, V., Nash, S., Kirov, G., Owen, M. J., Wimpory, D. C., Jan. 2007. Association of per1 and npas2 with autistic disorder: support for the clock genes&social timing hypothesis. *Molecular Psychiatry* **12**: 581–592.
URL http://dx.doi.org/10.1038/sj.mp.4001953

Ning, X.-H., Chi, E. Y., Buroker, N. E., Chen, S.-H., Xu, C.-S., Tien, Y.-T., Hyyti, O. M., Ge, M., Portman, M. A., Oct. 2007. Moderate hypothermia (30°C) maintains myocardial integrity and modifies response of cell survival proteins after reperfusion. A*merican Journal of Physiology - Heart and Circulatory Physiology* **293**: H2119–H2128.
URL http://dx.doi.org/10.1152/ajpheart.00123.2007

Noguchi, T., Wang, L. L., Welsh, D. K., Jun. 2013. Fibroblast PER2 circadian rhythmicity depends on cell density. *Journal of Biological Rhythms* **28**: 183–192.
URL http://dx.doi.org/10.1177/0748730413487494

Nozawa, M., Miura, S., Nei, M., Jan. 2010. Origins and evolution of MicroRNA genes in drosophila species. *Genome Biology and Evolution* **2**: 180–189.
URL http://dx.doi.org/10.1093/gbe/evq009

O'Geen, H., Echipare, L., Farnham, P. J., 2011. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in Molecular Biology* **791**: 265–286.
URL http://dx.doi.org/10.1007/978-1-61779-316-5_20

Ohno, S., 1970. Evolution by gene duplication., 1st Edition. Allen & Unwin; Springer-Verlag.
URL http://www.worldcat.org/isbn/0045750157

Oster, H., 2006. The genetic basis of circadian behavior. *Genes, Brain, and Behavior* **5**: 73–79.
URL http://dx.doi.org/10.1111/j.1601-183x.2006.00226.x

Oulas, A., Boutla, A., Gkirtzou, K., Reczko, M., Kalantidis, K., Poirazi, P., Jun. 2009. Prediction of novel microRNA genes in cancer-associated genomic regions–a combined computational and experimental approach. *Nucleic Acids Research* **37**: 3276–3287. URL http://dx.doi.org/10.1093/nar/gkp120

Ouyang, Y., Andersson, C. R., Kondo, T., Golden, S. S., Johnson, C. H., Jul. 1998. Resonating circadian clocks enhance fitness in cyanobacteria. *Proceedings of the National Academy of Sciences* **95**: 8660–8664. URL http://www.pnas.org/content/95/15/8660.abstract

Park, J.-E. E., Heo, I., Tian, Y., Simanshu, D. K., Chang, H., Jee, D., Patel, D. J., Kim, V. N., Jul. 2011. Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature* **475**: 201–205. URL http://dx.doi.org/10.1038/nature10198

Paul, C., Schöberl, F., Weinmeister, P., Micale, V., Wotjak, C. T., Hofmann, F., Kleppisch, T., Dec. 2008. Signaling through cGMP-dependent protein kinase i in the amygdala is critical for auditory-cued fear memory and long-term potentiation. *The Journal of Neuroscience* **28**: 14202–14212. URL http://dx.doi.org/10.1523/jneurosci.2216-08.2008

Pederson, T., Oct. 2010. Regulatory RNAs derived from transfer RNA? *RNA* **16**: 1865–1869. URL http://dx.doi.org/10.1261/rna.2266510

Piriyapongsa, J., Mariño Ramírez, L., Jordan, I. K., Jun. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**: 1323–1337. URL http://dx.doi.org/10.1534/genetics.107.072553

Pommier, Y., Oct. 2006. Topoisomerase i inhibitors: camptothecins and beyond. *Nature Reviews Cancer* **6**: 789–802. URL http://dx.doi.org/10.1038/nrc1977

Purschke, M., Laubach, H.-J. J., Anderson, R. R., Manstein, D., Jan. 2010. Thermal injury causes DNA damage and lethality in unheated surrounding cells: active thermal bystander effect. *The Journal of Investigative Dermatology* **130**: 86–92. URL http://dx.doi.org/10.1038/jid.2009.205

Rajagopalan, R., Vaucheret, H., Trejo, J., Bartel, D. P., Dec. 2006. A diverse and evolutionarily fluid set of microRNAs in arabidopsis thaliana. *Genes & Development* **20**: 3407–3425. URL http://dx.doi.org/10.1101/gad.1476406

Reilly, D. F., Westgate, E. J., FitzGerald, G. A., Aug. 2007. Peripheral circadian clocks in the vasculature. *Arteriosclerosis, Thrombosis, and Vascular Biology* **27**: 1694–1705. URL http://dx.doi.org/10.1161/atvbaha.107.144923

Rensing, L., Monnerjahn, C., Oct. 1996. Heat shock proteins and circadian rhythms. *Chronobiology International* **13**: 239–250. URL http://view.ncbi.nlm.nih.gov/pubmed/8889248

Ro, S., Song, R., Park, C., Zheng, H., Sanders, K. M., Yan, W., Dec. 2007. Cloning and expression profiling of small RNAs expressed in the mouse ovary. *RNA* **13**: 2366–2380.
URL http://dx.doi.org/10.1261/rna.754207

Ro, S., Yan, W., 2010. Detection and quantitative analysis of small RNAs by PCR. *Methods in Molecular Biology* **629**: 295–305.
URL http://dx.doi.org/10.1007/978-1-60761-657-3_19

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., Bradley, A., Oct. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Research* **14**: 1902–1910.
URL http://dx.doi.org/10.1101/gr.2722704

Romano, N., Macino, G., Nov. 1992. Quelling: transient inactivation of gene expression in neurospora crassa by transformation with homologous sequences. *Molecular Microbiology* **6**: 3343–3353.
URL http://dx.doi.org/10.1111/j.1365-2958.1992.tb02202.x

Rudic, R. D., McNamara, P., Reilly, D., Grosser, T., Curtis, A.-M. M., Price, T. S., Panda, S., Hogenesch, J. B., FitzGerald, G. A., Oct. 2005. Bioinformatic analysis of circadian gene oscillation in mouse aorta. *Circulation* **112**: 2716–2724.
URL http://dx.doi.org/10.1161/circulationaha.105.568626

Sailaja, K., Rao, V. R., Yadav, S., Reddy, R. R., Surekha, D., Rao, D. N., Raghunadharao, D., Vishnupriya, S., Jul. 2012. Intronic SNPs of TP53 gene in chronic myeloid leukemia: Impact on drug response. *Journal of Natural Science, Biology, and Medicine* **3**: 182–185.
URL http://dx.doi.org/10.4103/0976-9668.101910

Shende, V. R., Goldrick, M. M., Ramani, S., Earnest, D. J., Jul. 2011. Expression and rhythmic modulation of circulating MicroRNAs targeting the clock gene bmal1 in mice. *PLoS ONE* **6**: e22586+.
URL http://dx.doi.org/10.1371/journal.pone.0022586

Shenkar, R., Shen, M. H., Arnheim, N., Apr. 1991. DNase i-hypersensitive sites and transcription factor-binding motifs within the mouse e beta meiotic recombination hot spot. *Molecular and Cellular Biology* **11**: 1813–1819.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC359851/

Shibata, M., Kurokawa, D., Nakao, H., Ohmura, T., Aizawa, S., Oct. 2008. MicroRNA-9 modulates Cajal-Retzius cell differentiation by suppressing foxg1 expression in mouse medial pallium. *The Journal of Neuroscience* **28**: 10415–10421.
URL http://dx.doi.org/10.1523/jneurosci.3219-08.2008

Shomron, N., Levy, C., 2009. MicroRNA-biogenesis and Pre-mRNA splicing crosstalk. *Journal of Biomedicine & Biotechnology*. **2009**: Article ID 594678
URL http://dx.doi.org/10.1155/2009/594678

Siddiqui, N., Mangus, D. A., Chang, T.-C., Palermino, J.-M., Shyu, A.-B., Gehring, K., Aug. 2007. Poly(A) nuclease interacts with the c-terminal domain of polyadenylate-binding protein domain from Poly(A)-binding protein. *Journal of Biological Chemistry* **282**: 25067–25075.
URL http://dx.doi.org/10.1074/jbc.m701256200

Sidote, D., Majercak, J., Parikh, V., Edery, I., Apr. 1998. Differential effects of light and heat on the drosophila circadian clock proteins PER and TIM. *Molecular and Cellular Biology* **18**: 2004–2013.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC121430/

Smalheiser, N., Torvik, V., Jun. 2005. Mammalian microRNAs derived from genomic repeats. *Trends in Genetics* **21**: 322–326.
URL http://dx.doi.org/10.1016/j.tig.2005.04.008

Steffen, P., Voß, B., Rehmsmeier, M., Reeder, J., Giegerich, R., Feb. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**: 500–503.
URL http://dx.doi.org/10.1093/bioinformatics/btk010

Storch, K.-F. F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H., Weitz, C. J., May 2002. Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**: 78–83.
URL http://dx.doi.org/10.1038/nature744

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., Sommer, S. S., Rossi, J. J., Sep. 2009. SNPs in human miRNA genes affect biogenesis and function. *RNA* **15**: 1640–1651.
URL http://dx.doi.org/10.1261/rna.1560209

Tamaru, T., Hattori, M., Honda, K., Benjamin, I., Ozawa, T., Takamatsu, K., Sep. 2011. Synchronization of circadian per2 rhythms and HSF1-BMAL1:CLOCK interaction in mouse fibroblasts after Short-Term heat shock pulse. *PLoS ONE* **6**: e24521+.
URL http://dx.doi.org/10.1371/journal.pone.0024521

Tsutsumi, A., Kawamata, T., Izumi, N., Seitz, H. A., Tomari, Y., Oct. 2011. Recognition of the pre-miRNA structure by drosophila dicer-1. *Nature Structural Molecular Biology* **18**: 1153–1158.
URL http://dx.doi.org/10.1038/nsmb.2125

Tu, B. P., Kudlicki, A., Rowicka, M., McKnight, S. L., Nov. 2005. Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science* **310**: 1152–1158.
URL http://dx.doi.org/10.1126/science.1120499

Ünsal Kaçmaz, K., Mullen, T. E., Kaufmann, W. K., Sancar, A., Apr. 2005. Coupling of human circadian and cell cycles by the timeless protein. *Molecular and Cellular Biology* **25**: 3109–3116.
URL http://dx.doi.org/10.1128/mcb.25.8.3109-3116.2005

Van Maldergem, L., Hou, Q., Kalscheuer, V. M., Rio, M., Doco-Fenzy, M., Medeira, A., de Brouwer, A. P., Cabrol, C., Haas, S. A., Cacciagli, P., Moutton, S., Landais, E., Motte, J., Colleaux, L., Bonnet, C., Villard, L., Dupont, J., Man, H.-Y. Y., Aug. 2013. Loss of function of KIAA2022 causes mild to severe intellectual disability with an autism spectrum disorder and impairs neurite outgrowth. *Human Molecular Genetics* **22**: 3306–3314.
URL http://dx.doi.org/10.1093/hmg/ddt187

Varelas, X., Samavarchi-Tehrani, P., Narimatsu, M., Weiss, A., Cockburn, K., Larsen, B. G., Rossant, J., Wrana, J. L., Dec. 2010. The crumbs complex couples cell density sensing to hippo-dependent control of the TGF-β-SMAD pathway. *Developmental Cell* **19**: 831–844.
URL http://dx.doi.org/10.1016/j.devcel.2010.11.012

Vincent, J. B., Skaug, J., Scherer, S. W., Jan. 2000. The human homologue of flamingo, EGFL2, encodes a Brain-Expressed large Cadherin-Like protein with epidermal growth Factor-Like domains, and maps to chromosome 1p13.3-p21.1. *DNA Research* **7**: 233–235.
URL http://dx.doi.org/10.1093/dnares/7.3.233

Voellmy, R., Boellmann, F., 2007. Chaperone regulation of the heat shock protein response. *Advances in Experimental Medicine and Biology* **594**: 89–99.
URL http://dx.doi.org/10.1007/978-0-387-39975-1_9

Wakil, A. E., Francius, C., Wolff, A., Pleau-Varet, J., Nardelli, J., Jun. 2006. The GATA2 transcription factor negatively regulates the proliferation of neuronal progenitors. *Development* **133**: 2155–2165.
URL http://dx.doi.org/10.1242/dev.02377

Wang, Y., Zheng, Y., Luo, F., Fan, X., Chen, J., Zhang, C., Hui, R., Feb. 2009. KCTD10 interacts with proliferating cell nuclear antigen and its down-regulation could inhibit cell proliferation. *Journal of Cellular Biochemistry* **106:** 409–413.
URL http://dx.doi.org/10.1002/jcb.22026

Weickert, C. S. S., Miranda-Angulo, A. L., Wong, J., Perlman, W. R., Ward, S. E., Radhakrishna, V., Straub, R. E., Weinberger, D. R., Kleinman, J. E., Aug. 2008. Variants in the estrogen receptor alpha gene and its mRNA contribute to risk for schizophrenia. *Human Molecular Genetics* **17**: 2293–2309.
URL http://dx.doi.org/10.1093/hmg/ddn130

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., Peterson, K. J., 2009. The deep evolution of metazoan microRNAs. *Evolution & Development* **11**: 50–68.
URL http://dx.doi.org/10.1111/j.1525-142x.2008.00302.x

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., Sibly, R. M., Jun. 2003. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**:781–787.
URL http://www.genetics.org/cgi/content/abstract/164/2/781

Williams, S. K., Spence, H. J., Rodgers, R. R., Ozanne, B. W., Fitzgerald, U., Barnett, S. C., Sep. 2005. Role of mayven, a kelch-related protein in oligodendrocyte process formation. *Journal of Neuroscience Research* **81**: 622–631.
URL http://dx.doi.org/10.1002/jnr.20588

Wimpory, D., Nicholas, B., Nash, S., May 2002. Social timing, clock genes and autism: a new hypothesis. *Journal of Intellectual Disability Research* **46:** 352–358.
URL http://dx.doi.org/10.1046/j.1365-2788.2002.00423.x

Wu, L., Fan, J., Belasco, J. G., Mar. 2006. MicroRNAs direct rapid deadenylation of mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 4034–4039.
URL http://dx.doi.org/10.1073/pnas.0510928103

Xu, L., Qu, Z., Apr. 2012. Roles of protein ubiquitination and degradation kinetics in biological oscillations. *PLoS ONE* **7**: e34616+.
URL http://dx.doi.org/10.1371/journal.pone.0034616

Xue, Y., Wang, Q., Long, Q., Ng, B. L. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., Asan, MacArthur, D. G., Quail, M. A., Carter, N. P., Yang, H., Tyler-Smith, C., Sep. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* **19**: 1453–1457.
URL http://dx.doi.org/10.1016/j.cub.2009.07.032

Yang, M., May, W. S., Ito, T., Sep. 1999. JAZ requires the double-stranded RNA-binding zinc finger motifs for nuclear localization. *Journal of Biological Chemistry* **274**: 27399–27406.
URL http://dx.doi.org/10.1074/jbc.274.39.27399

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., Wang, J., Jul. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**: 75–78.
URL http://dx.doi.org/10.1126/science.1190371

Yuan, Z., Sun, X., Liu, H., Xie, J., Mar. 2011. MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. *PLoS ONE* **6:** e17666+.
URL http://dx.doi.org/10.1371/journal.pone.0017666

Zhang, L., Chia, J.-M., Kumari, S., Stein, J. C., Liu, Z., Narechania, A., Maher, C. A., Guill, K., McMullen, M. D., Ware, D., Nov. 2009. A Genome-Wide characterization of MicroRNA genes in maize. *PLoS Genetics* **5**: e1000716+.
URL http://dx.doi.org/10.1371/journal.pgen.1000716

Zhang, S., Lu, J., Zhao, X., Wu, W., Wang, H., Lu, J., Wu, Q., Chen, X., Fan, W., Chen, H., Wang, F., Hu, Z., Jin, L., Wei, Q., Shen, H., Huang, W., Lu, D., Jul. 2010. A variant in the CHEK2 promoter at a methylation site relieves transcriptional repression and confers reduced risk of lung cancer. *Carcinogenesis* **31**: 1251–1258.
URL http://dx.doi.org/10.1093/carcin/bgq089

Zhou, X., Ruan, J., Wang, G., Zhang, W., Mar. 2007. Characterization and identification of microRNA core promoters in four model species. *PLoS Computational Biology* **3**:e37+.
URL http://dx.doi.org/10.1371/journal.pcbi.0030037

Zhu, Y., Lu, Y., Zhang, Q., Liu, J.-J., Li, T.-J., Yang, J.-R., Zeng, C., Zhuang, S.-M., May 2012. MicroRNA-26a/b and their host genes cooperate to inhibit the G1/S transition by activating the pRb protein. *Nucleic Acids Research* **40:** 4615–4625.
URL http://dx.doi.org/10.1093/nar/gkr1278