Improved microRNA cloning and analysis of RNA splicing

Yi Long

Imperial College London

Department of Metabolism, Digestion and Reproduction

Thesis submitted to Imperial College London for the degree of Doctor of Philosophy

Declaration of Authorship

I, Yi Long, declare that this thesis titled, 'Improved microRNA cloning and analysis of RNA splicing' and the work presented in it are my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. For Chapter 5 I contributed strongly to Tables 5.2, 5.3 and 5.5 and made minor contributions to Tables 5.6 and 5.7. I did not contribute to Fig 5.3 (Chapter 5, Discussion).

Copyright declaration

Acknowledgements

I would like to thank my supervisor Dr Nick Dibb for all of his advice, support, and a few coffee meetings over the years, for his guidance and support and the scientific discussions about the many ideas contributing to this work, without his advice this work could not have been done. His advice for my project is very mature and useful, which he obtained from his work experience. Thank you Dr Nick Dibb for your advice and support for my project.

I also want to thank Dr Cui Wei, who gave me very good advice on my life, and I also want to thank her previous PhD student zhang shuchen in her group, her advice is also very helpful. Thanks to Rupa, Chin and Rakhee for their advice in the lab and for always being helpful. I would also like to thank the fourth floor IRDB members, for all the help and advice.

Special thanks to Diana Alexieva for all of her support and understanding, especially in the past few years. And thank you also for your support and help with my experimental technology. I would also like to thank Dr Leandro Castellano and Hayley Kendall-Berry for their generous help over the years.

My special thanks go to my family for the continuous support and encouragement throughout this and previous endeavours. This work was supported by the Genesis Research Trust.

Abstract

MicroRNAs are single stranded RNA molecules of 19-25 nucleotides in length that are evolutionary conserved. MicroRNAs anneal through complementarity to messenger RNAs and can cause either inhibition of target messenger RNA translation or promotion of messenger RNA degradation. Mutations of miRNA genes can disrupt normal development and many other physiological processes.

IsomiRs are miRNA variants that are generated by imprecise Drosha or Dicer cleavage, many of these variants would be expected to target new messenger RNA and so may be of biological importance. We previously demonstrated that the expression of miRNAs and isomiRs can differ significantly between tissues as shown by cloning and sequence analysis. Here we used a bioinformatics approach to screen cell lines for markedly different ratios of miRNA:isomiR production, in order to establish model systems to investigate isomiR function.

We confirmed that the first step in the cloning of miRNAs by the commonly used Illumina protocol is biased but could be remedied by changing the condition of the ligation reaction and by the use of variable bases at the end of the 3' adaptor RA3. We found that similar modifications could not reduce bias in the second step of cloning miRNAs to the standard 5' RNA adapter RA5. We confirmed that the second step of cloning could be improved by using a combined RA3:RA5 adapter suggested to us by Somagenics. This required the successful identification of a suitable ligation enzyme and a reversible block for the 3' end of adapter RA3:RA5. Our preliminary results indicate how the Somagenics protocol could be further improved.

In a second bioinformatics project we show how a sequencing database of spliced mRNA called Snaptron can be used to predict the effect of splice site mutations upon cryptic splice site activation and exon skipping, using BRCA1 and BRCA2 splice site mutations for the initial analysis. We discuss that this empirical approach compares favourably to *in silico* methods and is also of value for the analysis of other types of splicing mutations.

Publication

Background splicing as a predictor of aberrant splicing in genetic disease

Alexieva D, **Long Y**, Sarkar R, Dhayan H, Bruet E, Winston Rm, Vorechovsky I, Castellano L, and Dibb N.J

RNA Biol. 2022;19(1):256-265. doi: 10.1080/15476286.2021.2024031. Epub 2021 Dec 31.

Alexieva, Long and Sarkar are joint first authors

Table of Contents

Abbreviation

Ago	Argonaute protein
APS	ammonium persulfate
ATP	adenosine triphosphate
AOs	antisense oligonucleotids
BRCA1	Breast Cancer gene 1
BRCA2	Breast Cancer gene 2
BSA	bovine serum albumin
bss	background splice site
css	cryptic splice sites
cDNA	complementary DNA
DGCR8	DiGeorge syndrome critical region 8
DMEM	Dulbecco's modified eagle medium
dNTP	deoxynucleotide triphosphate
dsRNA	double-stranded RNA
DNA	deoxyribonucleic acid
DTT	Dithiothreitol
DBASS	database of aberrant splice sites
DAVID	annotation, visualization and integrated discovery database
ESTs	expressed sequence tags
EDTA	ethylenediaminetetraacetic acid
EB	Ethidium Bromide
FBS/FCS	foetal bovine serum/foetal calf serum
GEO	gene expression omnibus
HL60	Human caucasian promyelocytic leukemia cell line
H929	Human Caucasian IgA-producing plasmacytoma
HCT116	Human colorectal carcinoma cell line
HGMD	human genome mutation database
Kb	kilobase
LOVD	Leiden Open Variation Database online
miRNAs	microRNAs
mRNA	messenger RNA
miRNP	microribonucleoprotein complex
MRE	microRNA recognition element
MCF-7	Human breast cancer cell line
nt	nucleotide
ncRNA	non-coding RNA

PAGE	polyacrylamide gel electrophoresis
P-bodies	processing bodies
PBS	phosphate-buffered saline
pre-miRNAs	precursor miRNAs
pri-miRNAs	primary miRNAs
PAZ	Piwi-Argonaute-Zwille
PCR	polymerase chain reaction
RPMI	Roswell Park Memorial Institute
RT	reverse transcription
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNA	ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
RS	Recursive sites
SDS	sodium dodecyl sulfate
siRNA	small interfering RNA
SSC buffer	saline-sodium citrate buffer
ssRNA	single-stranded RNA
T4 Rnl1/2	T4 RNA ligase 1/2
TAE buffer	Tris-acetate-EDTA buffer
TBE buffer	Tris-borate-EDTA buffer
TBS	Tris buffered saline
Tris	Trishydroxymethylaminomethane
TE buffer	Tris-EDTA buffer
TEMED	N, N, N’, N'-tetraethylmethylethane-1,2-diamine
THP-1	human leukemia monocytic cell line
UV	ultraviolet
U-2 OS	Human osteosarcoma cells
µl	microlitre(s)
3’ UTR	3’ untranslated region
VUS	variant of unknown significance

List of Figures

List of Tables

Chapter 1-Introduction

1.1 Alternative RNA splicing and microRNAs

Eukaryote genes are subject to two major regulatory events following their transcription, RNA splicing and miRNA regulation. In the nucleus the vast majority of primary RNA transcripts undergo splicing, which acts both to remove introns from primary RNA transcripts and to generate alternatively spliced RNAs (BreitbartAndreadis and Nadal-Ginard, 1987, Fu and Ares, 2014, Quinn and Chang, 2016). Those alternatively spliced RNA transcripts that encode proteins are translated in the cytoplasm to produce alternative protein isoforms (Yoshida et al., 2011, Darman et al., 2015, DeBoever et al., 2015, Ilagan et al., 2015, Zhang et al., 2015, Suzuki et al., 2019) (Chapter5). RNA splicing allows single genes to encode many RNA and protein isoforms and notably it has been estimated that splicing allows the *DSCAM* gene of *Drosophila*, which encodes an axon guidance receptor, to encode over 38,000 receptor isoforms (Schmucker et al., 2000). Approximately 25% of mutations that cause human disease do so because they cause aberrant splicing (Scotti and Swanson, 2016, Baralle and Buratti, 2017). Splicing mutations that cause disease usually impair the splice sites of individual genes but may sometimes impair the splicing machinery itself (Chapter 5).

1.2 MicroRNAs regulate gene silencing

MicroRNAs are encoded by a family of more than 1000 human genes and are single stranded RNA molecules of 18-25 nucleotides in length that function as post-transcriptional regulators of gene expression (Hammond, 2015, KozomaraBirgaoanu and Griffiths-Jones, 2019, Leitao and Enguita, 2022, Naeli et al., 2022). The expression of most mRNAs is regulated by microRNAs and therefore miRNAs regulate a wide range of activities including cell proliferation, development and differentiation (FilipowiczBhattacharyya and Sonenberg, 2008, Felekkis et al., 2010, KrolLoedige and

Filipowicz, 2010, Lu and Rothenberg, 2018). MicroRNAs are transcribed from genes in intronic or intergenic locations and most miRNA genes are conserved between species (Rodriguez et al., 2004, Ying and Lin, 2006, Kim and Kim, 2007, Felekkis et al., 2010, Wong and Rasko, 2021). Usually microRNAs negatively regulate gene expression by destabilizing target transcripts and inhibiting the translation of proteins (Shivdasani, 2006, Vanicek, 2014, Couzigou et al., 2017). As such miRNAs are an important component of the overall regulation of both RNA and protein turnover within cells (Gushchina et al., 2018, Gan et al., 2019, Paterson et al., 2019, Rolfs et al., 2021, RossLanger and Jovanovic, 2021, Alagar Boopathy et al., 2023).

1.3 Discovery of microRNA

The first microRNA, *lin-4* was discovered in the *nematode Caenorhabditis elegans* by the research groups of Ruvkun and Ambros (LeeFeinbaum and Ambros, 1993, WightmanHa and Ruvkun, 1993, Bartel, 2004, Carthew and Sontheimer, 2009, LadomeryMaddocks and Wilson, 2011). These groups identified a number of genes that are essential for *worm* development and surprisingly one of these genes called *lin-4* did not encode a protein but instead encoded a small single stranded RNA that was shown to anneal to target sites on the 3' UTR of a mRNA transcript encoded by the gene *lin-14*, which is another gene that is essential for *worm* development (WightmanHa and Ruvkun, 1993, Hristova et al., 2005, Greene et al., 2023). *Lin-4* is specific to *worms* (LeeFeinbaum and Ambros, 1993, WightmanHa and Ruvkun, 1993), however, a second miRNA called *let-7* was discovered in *C.elegans* that is conserved with flies and humans (Pasquinelli et al., 2000, Reinhart et al., 2000, Galagali and Kim, 2020). The finding that *let-7* was conserved stimulated the miRNA field and soon many other miRNA genes were discovered in *C. elegans*, *Drosophila melanogaster* and human genomes (Lagos-Quintana et al., 2001, Lau et al., 2001, Lee and Ambros, 2001, Galagali and Kim, 2020).

1.4 MicroRNA biogenesis and RNA interference

Lin-4 and *let-7* were shown to encode single stranded miRNAs of about 22 nucleotides but larger sized overlapping RNA transcripts of about 70 bases were also observed and the gene sequences of *lin-4* and *let-7* indicated that the larger 70 bases transcripts were likely to form a double stranded structure (LeeFeinbaum and Ambros, 1993, WightmanHa and Ruvkun, 1993, Pasquinelli et al., 2000, Lee et al., 2002, SchulmanEsquela-Kerscher and Slack, 2005, Galagali and Kim, 2020). This observation stimulated several groups to investigate an RNAse called Dicer, which was known to generate single stranded short interfering RNA (siRNA) from double stranded RNA for the purpose of RNA interference. Knock-down or mutation of Dicer was found to cause the accumulation of longer precursor RNA indicating that Dicer is required to generate both siRNA and miRNA (Grishok et al., 2001, Hutvágner et al., 2001, Ketting et al., 2001, Knight and Bass, 2001, SvobodovaKubikova and Svoboda, 2016, SongAlluin and Rossi, 2022) (Fig 1.1).

RNA interference is a term given to the observation that viral infection of plants or the injection, transfection or even digestion of double stranded RNA by various animals can result in the silencing of target mRNAs with a homologous sequence (Fire, 2007, Chen and Rechavi, 2022). It is generally thought that RNA interference evolved as a defense mechanism against viruses, which often produce double stranded viral RNA and was then adapted for the production of miRNAs (Fire, 2007, SweversLiu and Smagghe, 2018, FátyolFekete and Ludman, 2020). A main difference between miRNA and siRNA production is that siRNAs are typically generated from exogenous RNA, such as viruses, whereas miRNAs are generated from endogenous host genes (Ketting et al., 2001, Carthew and Sontheimer, 2009, Szelenberger et al., 2019, Traber and Yu, 2023). However, RNA interference may also suppress endogenous transposon activity (Tabara et al., 1999, ElbashirLendeckel and Tuschl, 2001, RussoHarrington and Steiniger, 2016) and miRNAs can be transferred between cells (Arroyo et al., 2011, Cortez et al., 2011, Vickers et al., 2011, Makarova et al., 2021). Another difference is that individual siRNAs show complete complementarity to a specific target RNA (because the siRNA is

made from the target) that is then degraded, whereas individual miRNAs often regulate the expression of multiple mRNAs through partial complementarity and largely inhibit translation; although they can also cause mRNA degradation, particularly in plants (Lam et al., 2015, Neumeier and Meister, 2020) (see below Figure 1.1).

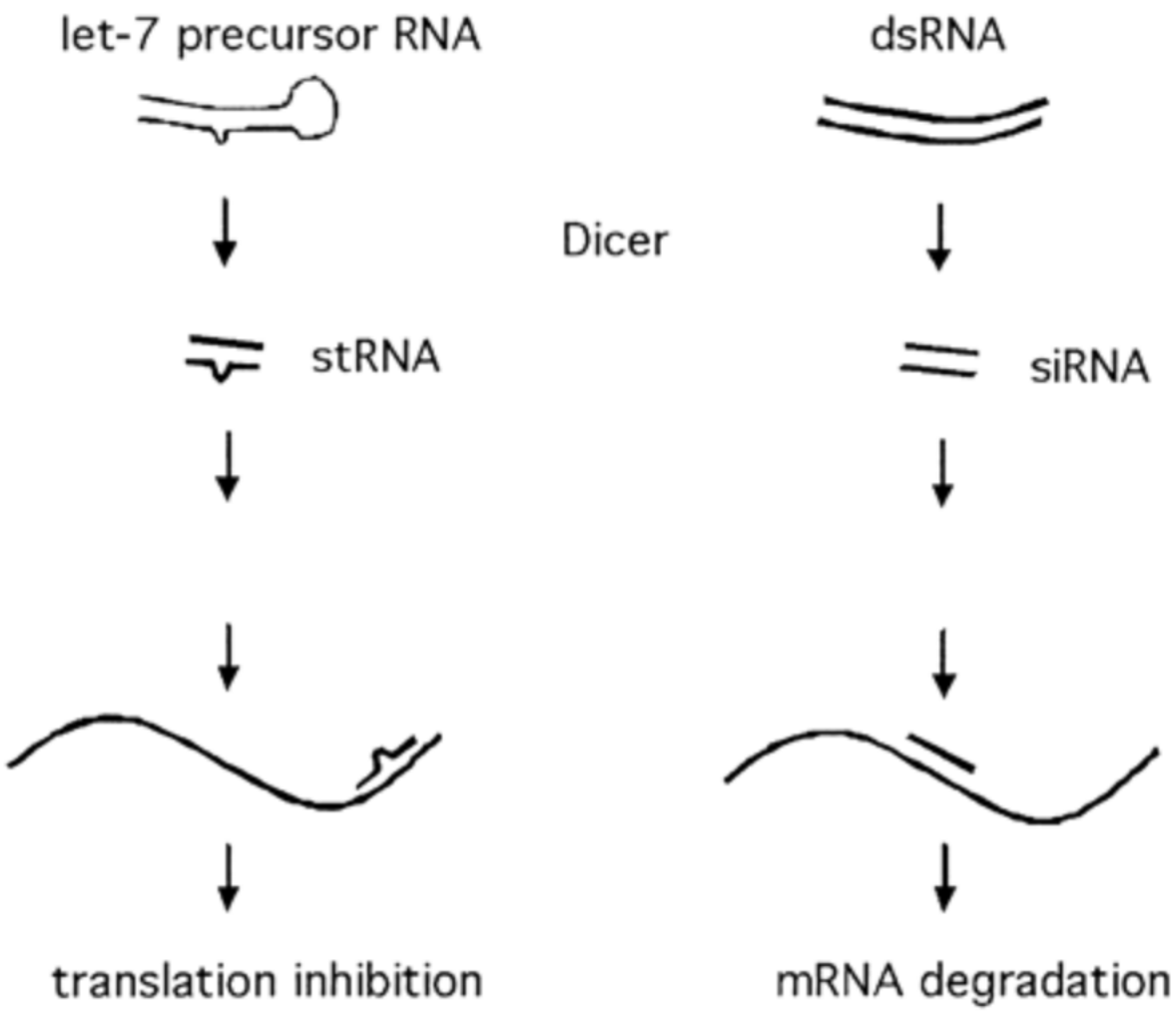


Figure 1.1 From Ketting et al 2001. The authors illustrate that Dicer has a role in both miRNA and siRNA production. Other groups made the same conclusion (Grishok et al., 2001, Hutvágner et al., 2001, Ketting et al., 2001, Knight and Bass, 2001, Hutvágner and Zamore, 2002). Ketting et al 2001 also show that *C.elegans* mutants for the gene *dcr-1* (which encodes Dicer) did not produce *let-7* miRNA but instead accumulated longer *let-7* precursor RNA. Copyright permission obtained from author and Cold Spring Harbor Laboratory Press.

The study of miRNA transcription in the nucleus (Lee et al., 2002, Ergin and Çetinkaya, 2022), using both a single miRNA gene (*miR-30a*) and a polycistronic cluster of 3 different miRNAs *(mir-23, mir-27, mir-24-2* containing three miRNA genes), led to the conclusions illustrated in Fig 1.2. The authors show that both monocistronic and polycistronic miRNA genes first produce a relatively large transcript (compared to pre-miRNA) that they termed a primary miRNA (pri-miRNA) (Lee et al., 2002, Vilimova and Pfeffer, 2023). By analysing nuclear and cytoplasmic fractions they showed that pre-miRNA was generated from a larger pri-miRNA precursor in the nucleus and that mature miRNA was made from pre-miRNA by Dicer in the cytoplasm (Lee et al., 2002, Michlewski and Cáceres, 2019, Vishlaghi and Lisse, 2020).

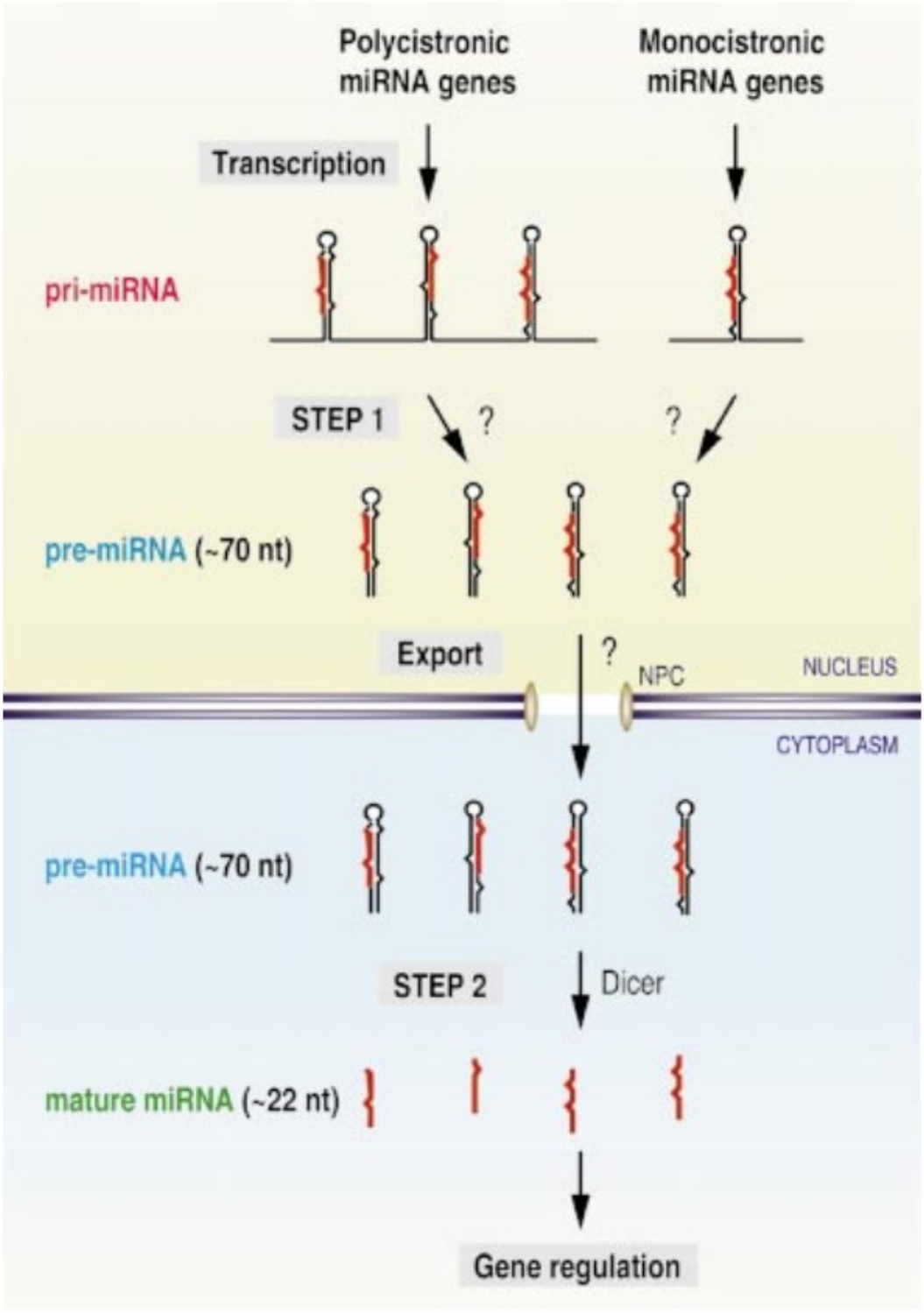

Polycistronic miRNA genes
Monocistronic miRNA genes
Transcription
pri-miRNA
STEP 1
?
?
pre-miRNA (~70 nt)
Export
?
NPC
NUCLEUS
CYTOPLASM
pre-miRNA (~70 nt)
STEP 2
Dicer
mature miRNA (~22 nt)
Gene regulation

Figure 1.2 Initial and historical characterization of the miRNA biogenesis pathway, from (Lee et al., 2002**).** The authors showed that both monocistronic (top right) and polycistronic miRNA genes (top left) first produce a transcript that they termed a primary miRNA. Polycistronic miRNA genes are clusters of miRNA genes (for example *mir-23*, *mir-27*, *mir-24-2*) that are transcribed from the same promoter and *mir-30a* is an example of a monocistronic miRNA gene. By using nuclear and cytoplasmic fractions they showed that pre-miRNA was generated in the nucleus and mature miRNA in the cytoplasm. The authors identified Dicer as the enzyme that converts pre-miRNA in the cytoplasm to mature miRNA. The enzyme responsible for generating pre-miRNA from primary miRNA was unknown at that time. RNA polymerase II is responsible for microRNA gene transcription (CaiHagedorn and Cullen, 2004, Lee et al., 2004). Copyright permission obtained from author and European Molecular Biology Organization (EMBO) Press.

1.4.1 Drosha identified

Subsequently the Kim group identified an RNase called Drosha as the nuclear enzyme that cleaves primary into pre-miRNA (Lee et al., 2003). They sequenced the end of *pre-miRNA-30a* and found it had a 2-nucleotide overhang at its 3' ends, which is characteristic of cleavage by an RNase of the type III family. From the human genome sequence it could be deduced that there were three genes that encoded type III RNases: L44, Dicer and Drosha (Lee et al., 2003). At that time Drosha was known to be an RNase III of unknown specific function that was first identified by sequencing in *Drosophila* (Filippov et al., 2000). Drosha seemed the most obvious candidate to be involved in microRNA processing and this was confirmed by showing that siRNA against Drosha caused the accumulation of primary miRNA and that immunoprecipitated Drosha could cleave primary miRNA *in vitro* (Lee et al., 2003, Pong and Gullerova, 2018, Matsuyama and Suzuki, 2019).

Through RNAi and biochemical studies, a protein called DGCR8 was found to be another important component of the pri-miRNA processing complex (the microprocessor) together with Drosha (Han et al., 2004, Guo and Wang, 2019) (Fig 1.3). Fig 1.3 illustrates that pre-miRNA transport into the cytoplasm is mediated by exportin-5 (Yi et al., 2003). Exportin-5 had previously been shown to be important for the export of non-coding RNAs (Lei and Silver, 2002, Zhang et al., 2021) and therefore a possible role for Exportin-5 in miRNA transport was tested and proven using siRNA against exportin-

5 (Yi et al., 2003, Melo and Esteller, 2014, Ohtsuka et al., 2015). Once the pre-miRNA enters the cytoplasm it is processed by the same cytoplasmic machinery that generates siRNA (Fig 1.1, 1.4).

1.4.2 Dicer and RISC identified

Dicer and RISC, which is an acronym for an enzyme called RNA induced silencing complex (Fig 1.3), were soon identified (Hammond et al., 2000, Bernstein et al., 2001) following the discovery of RNA interference (Fire et al., 1998, Hamilton and Baulcombe, 1999, Li and Zamore, 2019) and later the same enzymes were shown to process miRNAs (Fig 1.1, 1.4) (Caudy et al., 2002, IshizukaSiomi and Siomi, 2002, Mourelatos et al., 2002, Stavast and Erkeland, 2019). As previously discussed, RNA interference is a widespread process that is characterised by the production of short RNAs of about 22 bases in length in response to viral infection of plants or transfection of long double stranded RNA into animal cells (Hamilton and Baulcombe, 1999, Hammond et al., 2000, Parrish et al., 2000, Zamore et al., 2000, Ketting et al., 2001, Liu et al., 2023). The Hannon lab and others (Tuschl et al., 1999) managed to reproduce this effect with biochemical extracts and using this assay were able to partially purify and to identify the RNase responsible for generating the short dsRNA from longer dsRNA, which they called Dicer (Hammond et al., 2000, Bernstein et al., 2001, LeeKim and Kim, 2023).

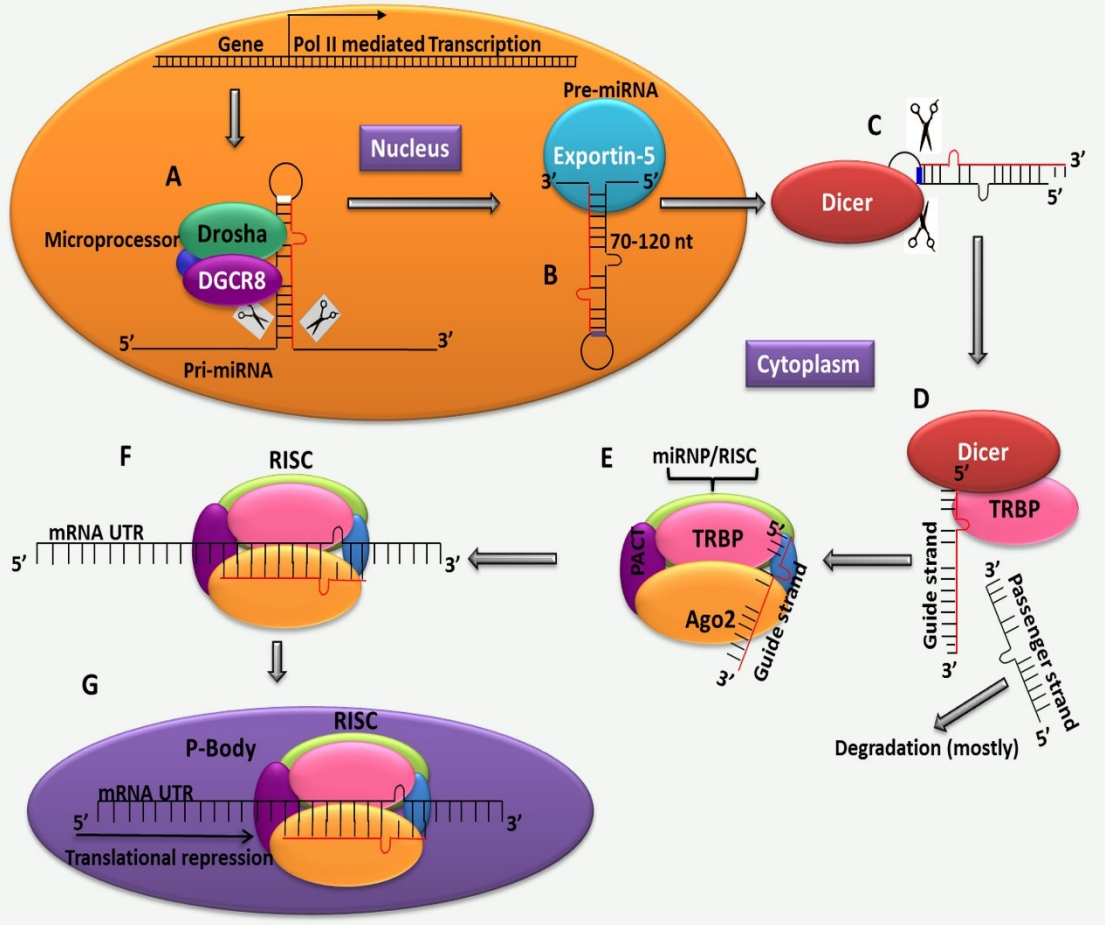


Figure 1.3 Adapted from (Bhaskaran and Mohan, 2014). A typical microRNA (miRNA) biogenesis pathway. (A) Primary miRNA is initially transcribed from miRNA genes by RNA polymerase II in the nucleus where it forms a hairpin that is processed by a large complex called microprocessor, which has an RNase type III enzyme called Drosha and the RNA binding protein DGCR8, to form a premature hairpin precursor with a length of about 70 nucleotides, called pre-miRNAs. (B) Pre-miRNA is exported into the cytoplasm via exportin 5. (C) In the cytoplasm the RNAse III enzyme named Dicer cleaves the loop region of the pre-miRNA to generate 18-23 double stranded mature miRNA molecules. (D) Dicer and a cellular protein called transactivation response RNA binding protein (TRBP) promotes Dicer-miRNA complex to RNA-induced silencing complex (RISC), which includes Argonaute 2 (Ago2) (E) The guide strand from the double stranded miRNA (C) is incorporated into the RNA-induced silencing complex (RISC) and the passenger strand is mostly degraded. (F) The guide strand guides RISC to the target mRNA where it anneals to sites in target mRNA. Extensive complementarity between the guide strand and the target mRNA activates mRNA degradation whereas partial complementarity inhibits translation. (G) It has been suggested but not proven that translation repression of mRNA by miRNAs may take place in P-bodies. Copyright permission obtained from author and SAGE Publishing.

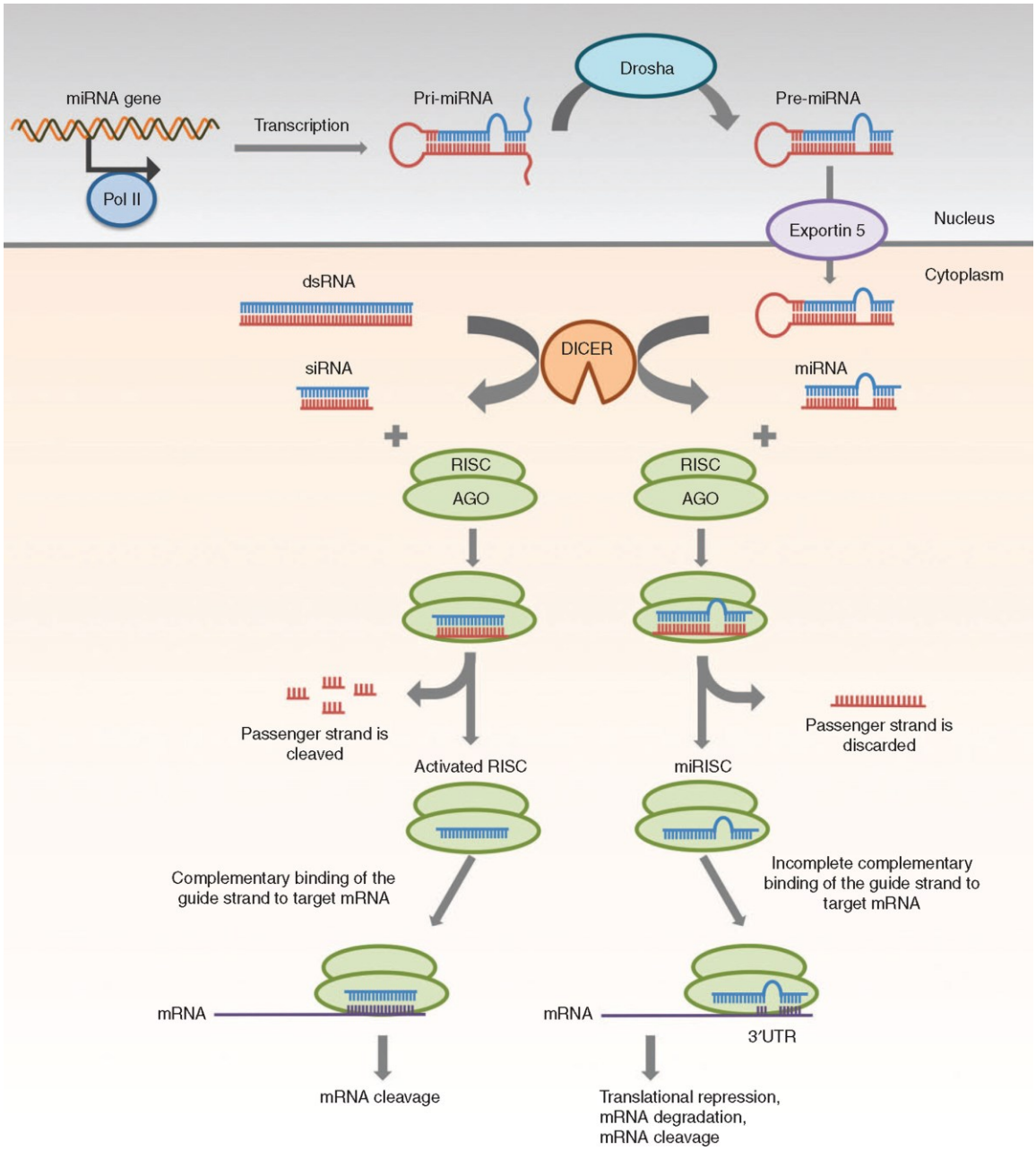


Figure 1.4 From (Lam et al., 2015) Double stranded RNA and pre-miRNA are processed in the cytoplasm by the same pathway.

Copyright permission obtained from author and Cell Press.

RISC (Fig 1.3, 1.4) was also identified by the Hannon lab and they showed that RISC degraded target mRNA that was homologous to whatever double stranded RNA was used to prime the biochemical extract (Hammond et al., 2000, Iwakawa and Tomari, 2022) (Fig 1.4, left pathway). Bernstein et al 2001 showed that Dicer and RISC could be separated by centrifugation and that the depletion of Dicer inhibits RISC from degrading specific mRNAs in response to transfection. This and other papers (Hamilton and Baulcombe, 1999, YangLu and Erickson, 2000) led to the model of RNA interference illustrated in Fig 1.4 which can be considered as three steps, where the double stranded RNA is diced, a single strand of the resulting short dsRNA is incorporated into RISC and used to scan for a matching target, which is degraded or sliced (Hammond et al., 2000, Bernstein et al., 2001, Jouravleva et al., 2022, Ranasinghe et al., 2022).

After Dicer processing, the miRNA duplex is unraveled by and loaded onto one of four human Argonaute proteins that bind to miRNA (see below) to form the RISC (Mourelatos et al., 2002, LiuFortin and Mourelatos, 2008, Kwak and Tomari, 2012, Nakanishi, 2022). This process is called miRNA loading, the other remaining strand (passenger strand) is largely complementary to the mature miRNA and if this strand is not loaded it will be degraded (LiuFortin and Mourelatos, 2008, LoiblArenz and Seitz, 2020, Ergin and Çetinkaya, 2022). Either strand of a miRNA duplex can be loaded on Ago proteins and the one which is loaded generally has the less stable 5' end (KhvorovaReynolds and Jayasena, 2003, Schwarz et al., 2003, Krol et al., 2004, Elkayam et al., 2012, Xiao and MacRae, 2022).

Fig 1.3 depicts a model of miRNA biogenesis that may of course change in the future. Kim et al (2016) demonstrated that knock-out of Drosha or Dicer in the HCT116 human colon carcinoma cell line had severe effects upon miRNA production but knock-out of exportin-5 had only modest effects, indicating that miRNA export can occur by other means. A small number of miRNAs, including miRNAs known as mirtrons, are processed independently of Drosha and an even smaller number of

miRNAs are made independently of Dicer (TreiberTreiber and Meister, 2019a, Salim et al., 2022).

1.4.3 siRNAs for mammalian cells

Elbashir et al., (2001) purified and then cloned and sequenced the short interfering RNA (siRNA) that are generated by the addition of longer dsRNA to *Drosophila* cell lysates. They found that the short siRNA consists predominantly of 21 or 22 base length RNAs that when annealed would be expected to generate double stranded RNA with short two base 3'overhangs. They showed that synthesised siRNA duplexes with two nucleotide 3' overhangs could mediate efficient target RNA cleavage in their *in vitro* system. Some of the short RNA duplexes they sequenced matched endogenous retrotransposons, which is consistent with the observation that some *C.elegans* mutants with deficient RNA interference show enhanced transposon mobilisation (Ketting et al., 1999, Tabara et al., 1999, RussoHarrington and Steiniger, 2016).

The above work led to the development of siRNA treatment for mammalian cells, which had previously proven resistant to dsRNA silencing because of the non-specific interferon response that is produced in response to long dsRNA, see (Tabara et al., 1999, Fire, 2007, Semple et al., 2022). However, synthetic 21 nucleotide siRNA duplexes could specifically silence target mRNAs in mammalian cell lines, so establishing the siRNA tool that is widely used today (Zamore et al., 2000, Caplen et al., 2001, ElbashirLendeckel and Tuschl, 2001, KobayashiTian and Ui-Tei, 2022).

1.5 Argonaute proteins

In parallel to the biochemical approaches described above, a genetic approach was developed by the groups of Fire and Mello to determine how double stranded RNA could cause the destruction of homologous target mRNA (Tabara et al., 1999, Hung and Slotkin, 2021). A genetic screen was devised

to select for mutants of *C.elegans* that were deficient for RNA interference and one of two genes (*rde-1*) that were initially identified encoded a protein of the Argonaute family (Tabara et al., 1999).

Argonaute proteins are now established as a key component of the RISC (Liu et al., 2004, KilikeviciusMeister and Corey, 2022) (Fig 1.3, 1.4, 1.5). *In vitro*, the minimal RISC needed for target cleavage is Ago2, a guide strand and Mg2+ (Liu et al., 2004, SchwarzTomari and Zamore, 2004) (Fig 1.5). Other proteins have been identified in the RISC (Fig 1.3E) and these are thought to assist Ago loading with miRNA (Kim and Kim, 2012, Nowak and Sarshad, 2021, Pérez-Cañamás et al., 2021). There is a family of 8 Arognaute proteins in humans (Sasaki et al., 2003) of which four Ago 1-4 bind to miRNA. Of these Ago2 is the only one of the four that has evident endonucelolytic slicer activity (Meister et al., 2004, Rand et al., 2004, Song et al., 2004, MüllerFazi and Ciaudo, 2019) and because of this Ago2 is the predominant mediator of RNA interference. Ago1, 3 or 4 are often found in the RISC with miRNA guides (Meister et al., 2004) and Ago2 slicer activity only operates when there is extensive complementarity between the Ago2:miRNA complex and its target mRNA (TreiberTreiber and Meister, 2019b) (Fig 1.5), which is seldom the case for miRNA targets.

1.5.1 Translation inhibition by miRNAs

Because Ago proteins 1, 3 and 4 do not have slicer activity and Ago2 rarely does when bound to miRNA it follows that the silencing effect of the Ago: miRNAs complex (Fig 1.3) in mammalian cells is likely to be achieved by translational inhibition rather than mRNA degradation. This is consistent with Fig 1.1, which depicts translational inhibition by miRNAs. The reason for this depiction is because the first reports of the *lin 4* miRNA showed that it acted to reduce the protein levels of *lin 14* rather than *lin 14* mRNA levels (LeeFeinbaum and Ambros, 1993, WightmanHa and Ruvkun, 1993, Greene et al., 2023). Nevertheless, target mRNA levels often reduce noticeably upon transfection with miRNAs (Lim et al., 2005) and it is generally thought that mRNA degradation does occur but only

after translation inhibition (Naeli et al., 2022). Such degradation cannot be directly caused by those Ago proteins without endonuclease activity but may occur by mechanisms that act upon most mRNAs (see below).

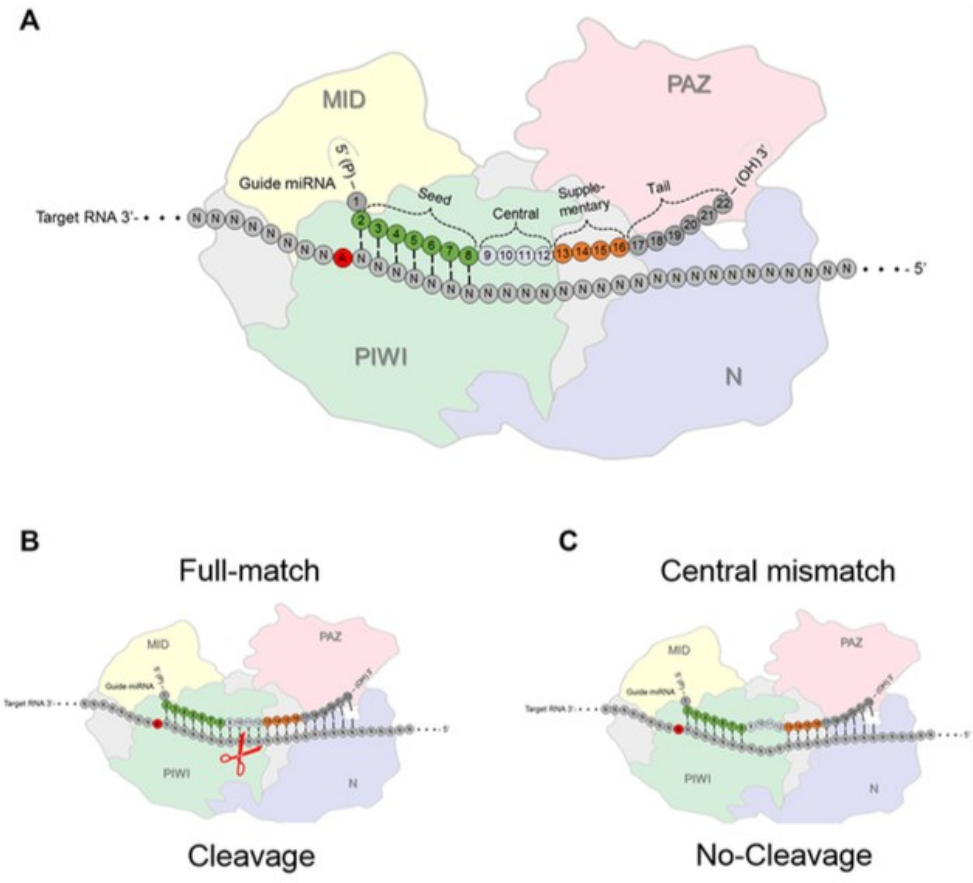


Figure 1.5 From (KilikeviciusMeister and Corey, 2022) A. Illustration of Ago bound to a miRNA and RNA target from known crystal structures. B, PIWI is responsible for cleavage of an RNA by Ago2 when perfect complementarity is present. C, illustrates that the miRNA sequence is mismatch for base pairing to the target RNA apart from the extreme 5' and 3' bases, which are clamped to Ago. A, C illustrate that mismatched bases between the miRNA and target RNA prevent cleavage. PAZ – PIWI-Argonaute-Zwille, N- N-terminal domain, PIWI – P-delement induced whimpy testes (this is an RNase) and MID – middle domain. Copyright permission obtained from author and Oxford University Press.

The mechanism of translation inhibition by miRNAs is an active area of research and there is increasing evidence that miRNAs can influence two mRNA-binding complexes, *eIF4F* and *Ccr4-Not*, which normally regulate the translation and turnover of most mRNAs (PillaiArtus and Filipowicz, 2004, Arthur and Djuranovic, 2018, Towler and Newbury, 2018, Wilczynska et al., 2019, MorrisCluet and Ricci, 2021, Naeli et al., 2022).

1.5.2 The function of the miRNA component of RISC

Pillai et al (2004) showed that tethering Ago2 alone to the 3'UTR of a luciferase mRNA caused inhibition of luciferase translation without affecting luciferase mRNA levels. The tethering was achieved by adding a short peptide to the amino terminus of recombinant Ago2 that could bind to a B box hairpin in the 3'UTR of the target luciferase mRNA. Just as importantly this result indicates that the primary function of miRNAs (which were absent in this experiment) is to guide Ago to target mRNAs.

Many mutagenesis studies have shown that mutations of a remarkably short region at the 5' end of a microRNA are best at preventing miRNAs from inhibiting the translation of a target mRNA in animal species and that this 5' region alone is sufficient to cause repression by some miRNAs (Doench and Sharp, 2004, Kloosterman et al., 2004, Brennecke et al., 2005, LaiTam and Rubin, 2005, Yan et al., 2019). In support of this finding some well established target sites of miRNAs only show strong complementarity to the 5' end of the miRNA (LaiTam and Rubin, 2005, Xiong et al., 2019). The critical 5' part of a miRNA is known as the seed region and is from nucleotides 2 to 8 (Fig 1.5). However, most of a miRNA sequence is conserved between species (Lim et al., 2003) indicating that the 3' part of miRNA must also be important for function. The 3' region contributes to the recognition of some

miRNA targets (LaiTam and Rubin, 2005, Grimson et al., 2007, Pu et al., 2019) (Fig 1.5C) and the passenger strand may also be important for some miRNAs (see below).

In plants, miRNAs often show full complementarity to target mRNAs and this is thought to be the reason why plant miRNAs cause mRNA degradation rather than translational inhibition (Bartel, 2004, Singh et al., 2023). In animals most miRNA targets contain at least a central mismatch with their guiding miRNA, which inhibit the activity of Ago2 endonuclease (Lim et al., 2005, Eichhorn et al., 2014, Becker et al., 2019) (Fig 1.5). The interchange between miRNA and siRNA function has been elegantly demonstrated by the finding that *let-7* miRNA and other miRNAs, which normally act to inhibit translation, can cause the degradation of mRNAs that have perfect complementarity (Hutvágner et al., 2001, Hutvágner and Zamore, 2002, ZengYi and Cullen, 2003, Meister et al., 2004). Conversely, siRNAs, which normally cause mRNA degradation, can repress the translation of mRNA targets that only have partial complementarity (DoenchPetersen and Sharp, 2003, ZengYi and Cullen, 2003, Doench and Sharp, 2004).

1.5.3 Co-operative binding

The limited complementarity between miRNAs and targets raises the question of how specificity is achieved. Wightman et al (1993) discussed that *lin-14* has seven target sites for *lin-4* perhaps because this facilitates co-operative binding and they suggested that synergistic binding of multiple *lin-4* miRNAs could generate a sharp down-regulation of lin-14 protein levels at a particular *lin-4* miRNA concentration during the *Larval 1* stage of *C.elegans* development. A number of experiments have since confirmed that miRNA repression becomes more efficient as the number of adjacent miRNA binding sites are increased (KilikeviciusMeister and Corey, 2022).

1.6 miRNA-mRNA target interaction prediction

Many groups have developed algorithms for predicting target sites of miRNAs, including miRanda, RNAhybrid, PicTar, TargetScan and miRtarget (Rodriguez et al., 2004, Ying and Lin, 2006, Kim and Kim, 2007). It is better to use a variety of algorithms for target prediction, because the different programs often predict different miRNA-binding sites. The computational algorithms are not completely precise, but they can be used to predict targets, which can be tested by experiment. There are also large database for researchers to analyse various possible interactions between miRNAs and mRNAs (Helwak et al., 2013, Grosswendt et al., 2014). High throughput experiments have shown that the target sites of miRNAs are not only located in the 3'-UTR, but can also be in 5' - UTR and coding regions (FormanLegesse-Miller and Coller, 2008, Zhou et al., 2009, Fabo and Khavari, 2023).

Because of the small size of the miRNA seed sequence (Fig 1.5), prediction algorithms for miRNA targets come up with a lot of false positives (Liu and Wang, 2019). Target sites can also be established by cross linking Ago that is bound to mRNA (Fig 1.5), digesting the mRNA and then immunoprecipitating Ago together with any protected target mRNA, which is then identified by RNA sequencing. A refinement of this process is to ligate the miRNA to the mRNA in order to identify both the target site and the targeting miRNA (Grosswendt et al., 2014, Fan et al., 2022).

A reservation about the CLIP (cross linking and immunoprecipitation) approach is that binding of miRNA to these mRNA sites might also identify sites that are not that important for repression of mRNA (Liu and Wang, 2019). The same authors point out that the more traditional method of identifying mRNAs whose expression decreases upon miRNA transfection (Lim et al., 2005) does identify mRNAs that are responsive to a miRNA (Liu and Wang, 2019). However, this traditional approach is subject to the criticism that such responses might be due to off-target or indirect effects

(KilikeviciusMeister and Corey, 2022). Target identification is evolving and for example (Liu and Wang, 2019) argue that it is valuable to cross-reference CLIP database and miRNA transfection database.

Despite the difficulties inherent to identifying MREs (microRNA response elements) and establishing their biological importance there have been many publications implicating the involvement of miRNAs and target mRNAs in disease, particularly cancer. Kilikevicius et al (2022) argue that most such studies fall short of providing conclusive insights. The authors have a number of reservations and in particular they list all recent publications (29 papers) that have reported cancer related functions of miRNAs in the HCT 116 colon carcinoma cell line. They point out that all 29 of the miRNAs that were studied are only expressed at very low levels by this cell line, arguably at levels that are not physiological (KilikeviciusMeister and Corey, 2022).

1.7 The function of microRNAs in biological processes and disease

Lin-4 and *let-7* microRNAs were identified as a result of the clear-cut effect of their mutation upon *C.elegans* development (LeeFeinbaum and Ambros, 1993, WightmanHa and Ruvkun, 1993, Pasquinelli et al., 2000, Britton et al., 2014). Similarly, mutations of at least four different plant miRNA genes have major effects upon normal plant development and flowering (Bartel, 2004). In humans, deletion of *mir-15* and *mir-16* contributes to the progression of chronic lymphocytic leukemia (Calin et al., 2002).

Miska et al (2007) reported that the majority of 95 different miRNA null mutants of *C.elegans* had no obvious phenotype. The same paper also questioned the functional relevance of miRNA overexpression studies because the implied involvement of two different overexpressed miRNAs in a physiological process was not supported by their subsequent knock-out. However, mice miRNA knock-outs have fared better with only 9 out of 28 having no obvious phenotype (ParkChoi and

McManus, 2010) (Table 1.1). Most of the mice knock-outs had cardiovascular or immune deficiencies but this is simply because the knock-outs were made by groups in those areas of research. Mice knock-outs of Dicer, Dgcr8, Drosha and Ago*2* are all embryonic lethals (ParkChoi and McManus, 2010, Dai et al., 2016). A resource of mice knock-out lines of the most highly expressed miRNAs has been initiated and are available for phenotype studies. Many microRNAs are very similar, therefore one of them deleted might not affect their phenotypes (Park et al., 2012). Very few of these mouse miRNA knock-outs were embryonic lethals (Park et al., 2012), which is consistent with the results for *C.elegans* (Miska et al., 2007).

Table 1.1 lists the antisense oligonucleotides that have been developed against the indicated miRNAs for the intended treatment of disease. The wide variety of disease in Table 1.1 reflects the widespread involvement of miRNAs in the underlying physiology. The miRNAs listed in Table 1.1 are considered to be pathologically overactive, which motivated the development of antisense oligonucleotides or anti-miRs. Some of the trials have been terminated, because of adverse side effects (Kim, 2023).

Table 1. The clinical trials of anti-miR miRNA inhibitors

			Clinical trial		
Drug name	miRNA	Diseases/disorders	Number	Phase	Status (Year)
Miravirsen (SPC3649)	miR-122	Hepatitis C	NCT00688012	I	Completed (2009)
			NCT00979927	I	Completed (2011)
			NCT01646489	I	Completed (2012)
			NCT01200420	II	Completed (2012)
				II	Unknown (2014)
			NCT01727934		
			NCT01872936	II	Unknown (2014)
pSil-miR200c/PMIS miR200a	miR-200a/c	Tooth Extraction Status Nos	NCT02579187	I	Withdrawn (2019)
RG-125 (AZD4076)	miR-103/107	Type 2 diabetes mellitus with non-alcoholic fatty liver disease	NCT02826525	I	Completed (2019)
		Non-alcoholic Steatohepatitis	NCT02612662	I	Active
MRG-110 (S95010)	miR-92	Wound healing	NCT03603431	I	Completed (2019)
		Cardiovascular Diseases	NCT03494712	I	Completed (2020)
CDR132L	miR-132	Heart Failure	NCT04045405	I	Completed (2020)
Cobomarsen (MRG-106)	miR-155	Lymphoma; Mycosis Fungoides; Leukemia	NCT02580552	I	Completed (2020)
		Cutaneous T-Cell Lymphoma/Mycosis Fungoides	NCT03837457	II	Terminated (2020)
			NCT03713320	II	Terminated (2020)
Lademirsen (RG-012)	miR-21	Alport Syndrome	NCT03373786	I	Completed (2019)
			NCT02855268	II	Recruiting
RGLS4326	miR-17	Polycystic Kidney Disease, Autosomal Dominant		I	Completed (2021)
			NCT04536688		
LNA-i-miR-221	miR-221	Multiple Myeloma, Refractory; Hepatocarcinoma; Advanced Solid Tumor	NCT04811898	I	Recruiting

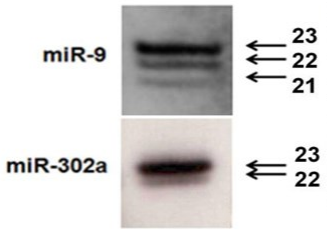
Notes. Information taken from https://clinicaltrials.gov

Table 1.1 From (KilikeviciusMeister and Corey, 2022) trials of antisense oligonucleotides (or equivalent) against the indicated miRNAs for disease treatment.

It is estimated that about 10% of human miRNAs are secreted from cells in exosomes (Arroyo et al., 2011, Cortez et al., 2011, Vickers et al., 2011) and it is certainly the case that miRNAs can be detected in blood, urine and other bodily fluids (Condrat et al., 2020, Wang and Chen, 2021). A large number of clinical studies have been published that show promise for diagnosing a wide range of diseases by assaying miRNA biomarkers from body fluids (Telonis et al., 2017, Condrat et al., 2020, Wang and Chen, 2021). There are concerns about the reproducibility of the assays that are used to detect miRNAs, which are largely based upon quantitative RT-PCR techniques (Condrat et al., 2020, Wang and Chen, 2021) and it remains to be seen if the use of miRNA biomarkers will become established clinical practice (Condrat et al., 2020).

1.8 miRNA variants: IsomiRs

Our lab previously sequenced three miRNA libraries from a human embryonic stem cell line H1, derived neural cells and foetal mesenchymal stem cells (Tan et al., 2014). A striking and consistent feature was that each miRNA gene also produced a substantial number of isomiRs that had 5' or 3' deletions or additions compared to the canonical miRNA that is listed in miRbase (KozomaraBirgaoanu and Griffiths-Jones, 2019, Bofill-De Ros et al., 2022). This result was confirmed by northern blotting (see Figure 1.6) and has also been reported by other labs (Baran-Gale et al., 2013, Bofill-De Ros et al., 2022).



MiRNA	Sequence	Length	Seq. No.
MiR-9	TCTTTGGTTATCTAGCTGTATGA	23	587
	CTTTGGTTATCTAGCTGTATGA	22	116
	TCTTTGGTTATCTAGCTGTATG	22	49
	TCTTTGGTTATCTAGCTGTATGAA	24	37
	CTTTGGTTATCTAGCTGTATGAA	23	15
	TCTTTGGTTATCTAGCTGTAT	21	13
	TCTTTGGTTATCTAGC	16	13
	TTTGGTTATCTAGCTGTATGA	21	10
MiR-302a	TAAGTGCTTCCATGTTTTGGTGA	23	54
	TAAGTGCTTCCATGTTTTGGTG	22	18
	AAGTGCTTCCATGTTTTGGTGA	22	16

Figure 1.6 From (Tan et al., 2014) IsomiRs detected by sequencing in embryonic stem cells can also be detected by northern blotting and so cannot be cloning or sequencing artifacts. The canonical miRNA is shaded in yellow.

1.8.1 miRNA and 5'isomiR target predicting

About half of the miRNAs that were sequenced (Tan et al., 2014) had 3' alterations and only 8% had 5' alterations. However, the 5' alterations would be expected to have a large effect upon mRNA targeting (Woods et al., 2020), as illustrated in Figure 1.7, which shows that a single base difference can create a large number of novel predicted targets for the isomiRs of *miR-9* or of *miR-302a*. 3' alterations have also been reported to have different activities compared with the corresponding canonical miRNA (Jones et al., 2009, Yamane et al., 2017, Yu et al., 2017).

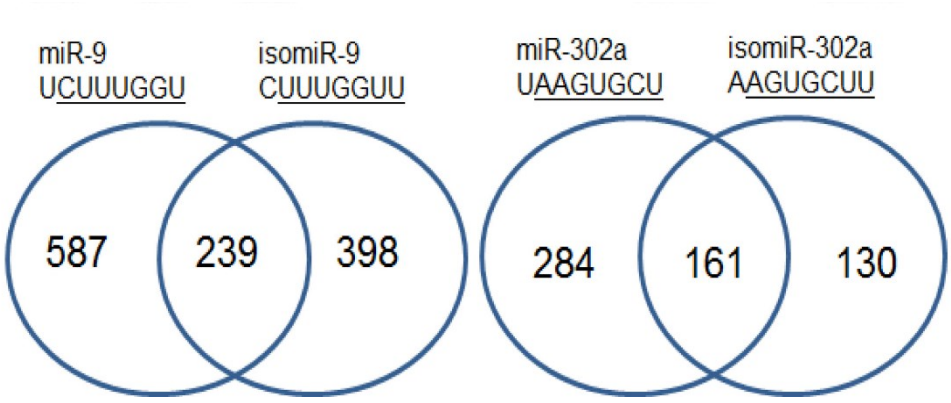


Figure 1.7 (Adapted from (Tan et al., 2014)). The Venn diagrams show that there is a surprisingly small proportion of shared predicted targets between *miR-9* and a common i*somiR-9* that has single base deletion at its 5' end. A similar analysis is also shown for *miR-302a*. The predictions were made by using TargetScanHuman (canonical) and TargetScan custom (isomiRs). Copyright permission obtained from author and Oxford University Press.

Tan et al (2014) confirmed that *isomiR-9* was able to target new mRNAs, such as DNMT3B *in vitro* and showed that it was possible to use the target site of DNMT3B to make a sponge that was specific for *isomiR-9* but not for the canonical *miR-9*.

1.8.2 MicroRNA and isomiR cleavage

IsomiRs are miRNA variants that are created largely by imprecise Drosha or Dicer cleavage (Landgraf et al., 2007, Morin et al., 2008a, Wu et al., 2009, Lee et al., 2010, Guo et al., 2011, NeilsenGoodall and Bracken, 2012, Guo et al., 2014) (Fig 1.8). IsomiRs can be longer or shorter by a small number of bases (usually one base) at the 5' or 3' ends compared to the most common miRNA (the canonical sequence) (Orbán, 2023).

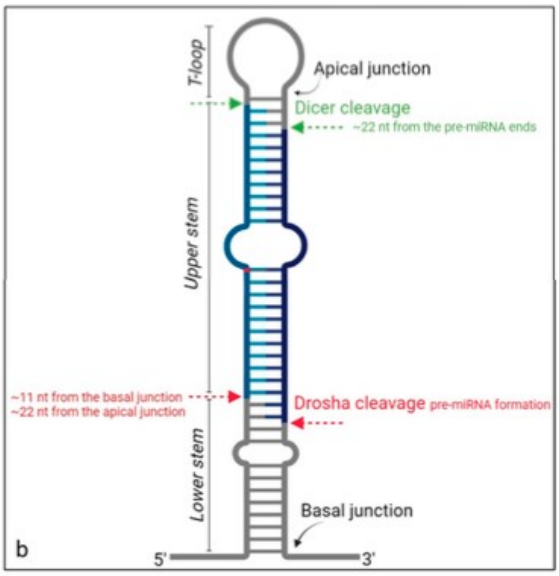


Figure 1.8 From (Zelli et al., 2021). Dicer and Drosha cleavage sites indicated on a pri-miRNA structure (see Figs 1.3, 1.4).

Copyright permission obtained from author and Multidisciplinary Digital Publications Institute (MDPI).

Kim et al., (2021) synthesized all 1881 pri-miRNAs (Fig 1.8) that were listed in miRBase version 21 and treated each pri-miRNA with purified Drosha:Dgcr8 (the microprocessor complex). They found that 8% of pri-miRNAs had more than one cleavage site for Drosha, indicating that some isomiRs can be produced by Drosha:Dgcr8 activity alone. They further examined primary *mir-142-5p* which was cleaved at one site by purified Drosha:Dgcr8 *in vitro* but cut in three different places *in vivo* (Wu et al., 2009). The authors confirmed this result *in vivo* by using their synthetic *pri-mir-142* to transfect HEK293T cells and confirming that three isomiRs were made. The authors concluded that additional auxiliary factors must be used *in vivo* for the generation of additional isomiRs by Drosha (Kim et al., 2021).

Kim et al., (2021) also identified a large number of pri-miRNAs that could not be cleaved by purified Drosha:Dgcr8 *in vitro*. After eliminating those miRNAs that might be processed by a non-canonical route or might require additional factors they suggest that 627 of the miRNAs listed in miRBase are false miRNAs because as well as not being cut by Drosha they are also known to be poorly expressed and/or unlikely to form a stable hairpin loop structure.

Following cleavage by Drosha and Dicer some miRNAs are further trimmed by exonucleases or tailed with adenines or uracils by nucleotidyl transferases (Tomasello et al., 2021, Zelli et al., 2021). Tailing with adenines or uracils is associated with more or less miRNA stability respectively (Tomasello et al., 2021).

Canonical miRNAs can be derived from either the 5p or 3p arm of a miRNA (Kuo et al., 2015). If derived from the 5p arm then the 5' end of the mature miRNA is cut and generated by Drosha (Fig 1.8). If the mature canonical miRNA is derived from the 3p arm then the 5' end of miRNA is cut by Dicer (Fig 1.8). The passenger strand refers to the miRNA arm that is not incorporated into RISC (Fig 1.3, 1.4). However, there are many examples where both of the arms of a miRNA are incorporated into RISC and some examples where the passenger strand is the predominant mature miRNA in certain tissues (Jagadeeswaran et al., 2010, Young et al., 2022). From our 3 miRNA libraries (see above) we found that an average of 77% of the miRNAs expressed only the guide strand, 17% expressed both guide and passenger strand and 6% expressed the passenger strand only (Tan GC PhD thesis https://spiral.imperial.ac.uk/handle/10044/1/39358).

The seed sequences of the 5p and 3p strand have very different predicted targets, as would be expected (Griffiths-Jones et al., 2011, Young et al., 2022). The observation of switching between 5p

and 3p expression in different tissues is referred to as arm switching and is suggested to be a fundamental mechanism in the evolution of miRNA function (de Wit et al., 2009, Griffiths-Jones et al., 2011). The physiological role of arm switching remains elusive although there has been progress in understanding the mechanism of arm switching (Kim et al., 2020).

IsomiRs are likely to be active *in vivo* because they co-immunoprecipitate with Ago proteins and are also active in luciferase and cleavage assays (Azuma-Mukai et al., 2008, Morin et al., 2008b, Lee et al., 2010, Cloonan et al., 2011), however, that does not necessarily mean that they are of any importance (NeilsenGoodall and Bracken, 2012). We showed that the 5' isomiRs in one species can be the canonical miRNA in a different species, which added to the limited evidence that isomiRs are of functional importance (NeilsenGoodall and Bracken, 2012, Tan et al., 2014).

Although a 5' change of a miRNA has a big impact upon target predictions (Fig 1.7), there are still predicted targets in common and Cloonan et al (2011) present good evidence that isomiRs act cooperatively with their canonical counterparts to target common biological pathways. They also cogently argue that isomiRs may help to reduce off-target effects.

1.9 Canonical to isomiR sequencing ratios in different tissues

We observed that the expression of five out of 295 of the most highly expressed miRNAs in miRGator showed convincing 5' isomiR switching between tissues, which is analogous to previous observations of arm switching (see above).

miR-215-5p	Canonical miRNA AUGACCU……	IsomiR UGACCU……	Samples	QC let-7a-1-5p	QC mir-23a-3p
Lung	1	0.21	15	0.99	0.98
Mammary glandular cells	1	0.09	4	0.99	0.98
Liver	1	103	6	0.99	0.98
Kidney	1	11.8	3	0.99	0.98

Table 1.2 (Adapted from (Tan et al., 2014)). This table shows that canonical *miR-215-5p* is a minority species in liver or kidney. The isomiR sequencing reads are shown in proportion to the canonical reads, which have been given the value of 1. The QC *let-7a-1-5p* and QC *mir-23a-3p* results are quality controls that show the ratio of these microRNAs that contain the 5' most base/total number of miRNAs sequenced (genuine isomiRs of *let-7a-1-5p* and *mir-23a-3p* are very rare). This indicates that incomplete sequencing of the 5' end of a microRNA is unlikely to be responsible for the 10 to 100 fold excess of a shorter 5' isomiR form of *miR-215-5p* in kidney and liver. Copyright permission obtained from author and Oxford University Press.

Table 1.2 shows the most extreme differences that we found but we also saw many fold changes in the ratio of canonical to isomiR expression for *miR-101-1-3p*, *miR-106a-5p*, *miR-140-3p* and *miR-500a-3p* between tissues (Tan et al., 2014). None of these isomiRs were made in significant amounts by the *in vitro* Drosha cleavage assay described above (Kim et al., 2021), indicating that the production of these particular isomiRs requires additional factors to Drosha and Dgcr8, similarly to *miR-142-5p* (see above).

We would like to know whether the observed differences between the ratios of canonical to isomiRs in different tissues are biologically relevant and whether these changes purposefully change mRNA targeting. In order to test this, we first need to identify cell lines that show good differences in canonical: isomiR ratios.

We also want to investigate a report that the cloning efficiency of miRNAs (and presumably isomiRs) can vary by many orders of magnitude according to the nature of the miRNA ends (Zhang et al., 2013). The authors also suggest a simple solution, which is to add two random bases to the ends of the adapters that are normally used for cloning and are ligated to each end of a miRNA (Zhang et al., 2013). However, the authors use an older and more demanding cloning method, (involving a gel purification step) compared to the commonly used Illumina kit method. The Illumina kit method allows all of the cloning steps to occur sequentially in the same eppendorf tube and so allows libraries to be made from very small starting amounts of total RNA.

1.10 Methods of miRNA cloning for sequencing

Figure 1.9 illustrates the two adapter method that is most commonly used for cloning miRNAs (BenesovaKubista and Valihrach, 2021). In the original procedure small RNA was purified from polyacrylamide gels, dephosphorylated (to prevent it from self-ligating) and ligated to a 3' adapter with T4 RNA ligase. The required ligation products were identified and purified from a polyacrylamide gel, phosphorylated and then ligated to a 5' adapter and the required and final ligation products were again identified and isolated from a polyacrylamide gel and then amplified by RT-PCR (ElbashirLendeckel and Tuschl, 2001) (Fig 1.9).

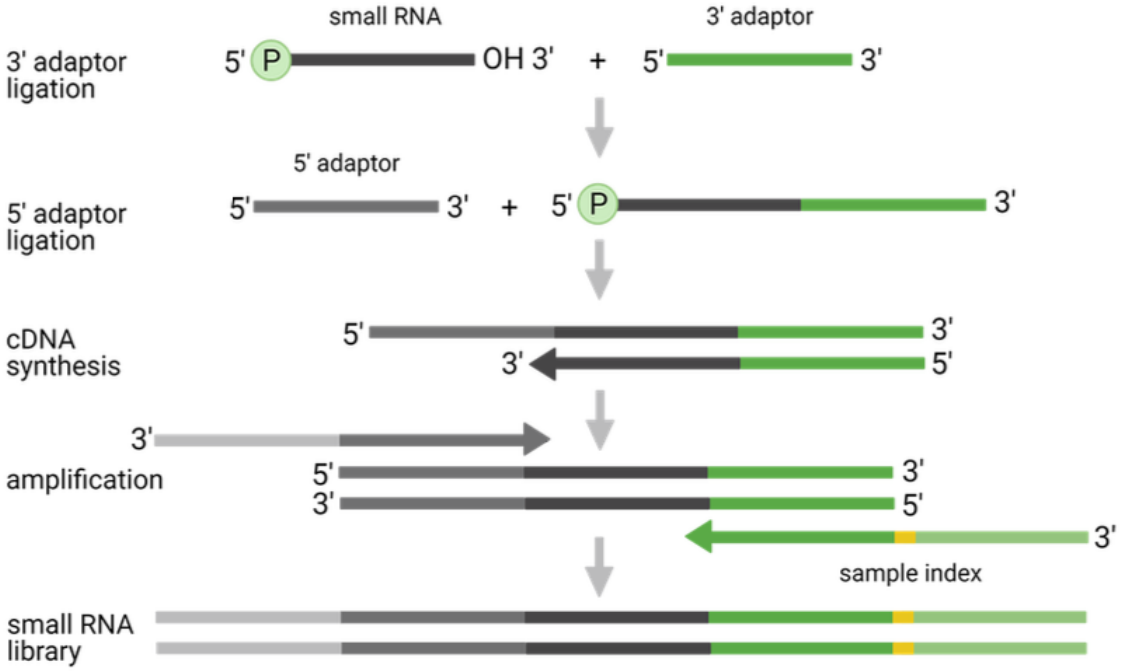


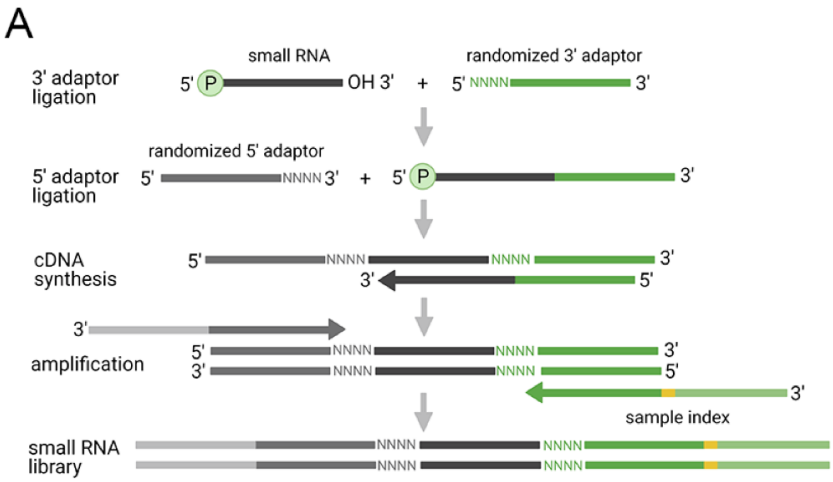
Figure 1.9 Two adapater method for cloning miRNA. From Benesova et al 2021 with permission.

Illumina modified this method so that the polyacrylamide gel purification steps were not required, which allowed libraries to be made from far smaller amounts of starting RNA (see Chapter 3). The current illumina protocol and kit was introduced in November 2010 (Baran-Gale et al., 2015). A number of other companies also sell kits that use the two adapter approach outlined above, although these may differ in minor details such as the method of removing unwanted adapter dimers (Baran-Gale et al., 2015).

However, considerable differences in sequencing reads were found between synthetic miRNA oligos following cloning by the two-adapter method and this was shown to be mainly due to marked differences in ligation efficiencies between different miRNAs to the cloning adapters, particularly to

the 5’ adapter (Hafner et al., 2011). Other groups also reported cloning bias by the two-adapter protocol (Linsen et al., 2009, Alon et al., 2011, Jayaprakash et al., 2011, Van Nieuwerburgh et al., 2011).

Zhang et al 2013 demonstrated that the ligation efficiency of those miRNAs that cloned poorly could be improved by changes to the ligation buffer and by adding variable bases to the 5’ end of the 3’ adapter and to the 3’ end of the 5’ adapter. This approach (Fig 1.10A) has been developed commercially (Perkin Elmer NetFlex).



B


miRNA
polyadenylation
AAAAAAAAAAAAAA
cDNA synthesis and
poly C tailing
CCCCCCCCC
AAAAAAAAAAAAAA
TTTTTTTTTTTTTTT
3' adapter
Template switching oligo
GGGGGGGGG
CCCCCCCCC
AAAAAAAAAAAAAA
TTTTTTTTTTTTTTT
amplicification
CCCCCCCCC
GGGGGGGGG
TTTTTTTTTTTTTTT
AAAAAAAAAAAAAA

C

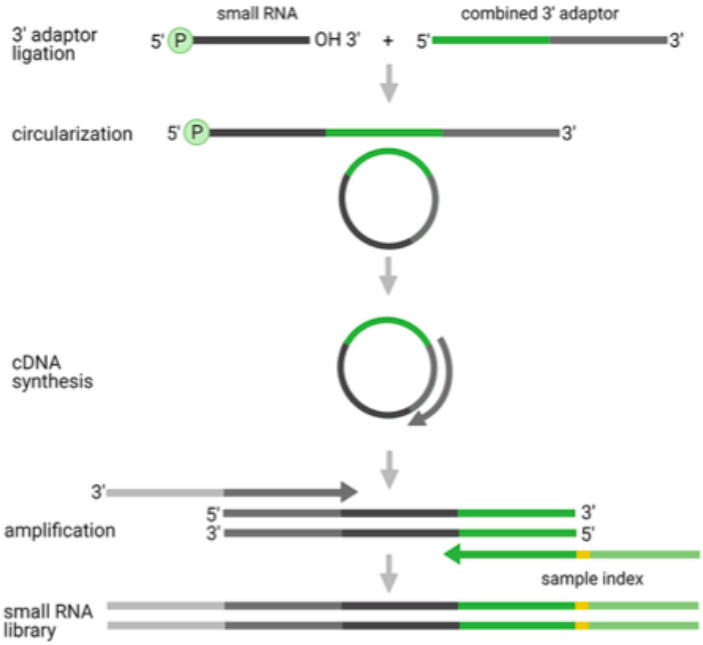

small RNA
combined 3' adaptor
3' adaptor ligation
5' P
OH 3'
+
5'
3'
circularization
5' P
3'
cDNA synthesis
3'
amplification
5'
3'
3'
5'
sample index
small RNA library

Figure 1.10 Methods to reduce miRNA cloning bias. A. Use of variable bases at the end of the 5' and 3' adapters. B. Polyadenylation and template switching. C. Circularisation. Diagrams from or based on Benesova et al 2021 with permission.

Fig 1.10B illustrates a method to clone miRNA that does not involve ligases and instead uses poly A polymerase to tail small RNAs with poly A which provides an annealing site for the 3' adapter. This method then takes advantage of the reverse transcriptase encoded by Moloney murine leukemia virus, which adds poly C to the 5' end of the newly synthesized cDNA and so provides an annealing site for the 5' adapter, a process referred to as template switching (Zhu et al., 2001).

Fig 1.10C illustrates a method developed by Somagenics that uses a combined adapter and so allows the second 5' adapter ligation step to occur by intramolecular ligation, which is more efficient than the usual intermolecular ligation (Fig 1.9).

A number of commercial kits are available to make miRNA libraries, and these have recently been tested by a number of independent groups (Barberán-Soler et al., 2018, Coenen-Stass et al., 2018, Dard-Dascot et al., 2018, Giraldez et al., 2018, Yeri et al., 2018, Godoy et al., 2019, Wong et al., 2019, Wright et al., 2019, Baldrich et al., 2020, Heinicke et al., 2020, Herbert et al., 2020, Androvic et al., 2021, BenesovaKubista and Valihrach, 2021). All groups have reported that randomized adapters (Fig 1.10A) give superior results to traditional adapters (Fig 1.9), see Benesova et al 2021. Fig 1.11 shows some of the results from a comprehensive multi-site study of commercially available miRNA cloning kits (Herbert et al., 2020).

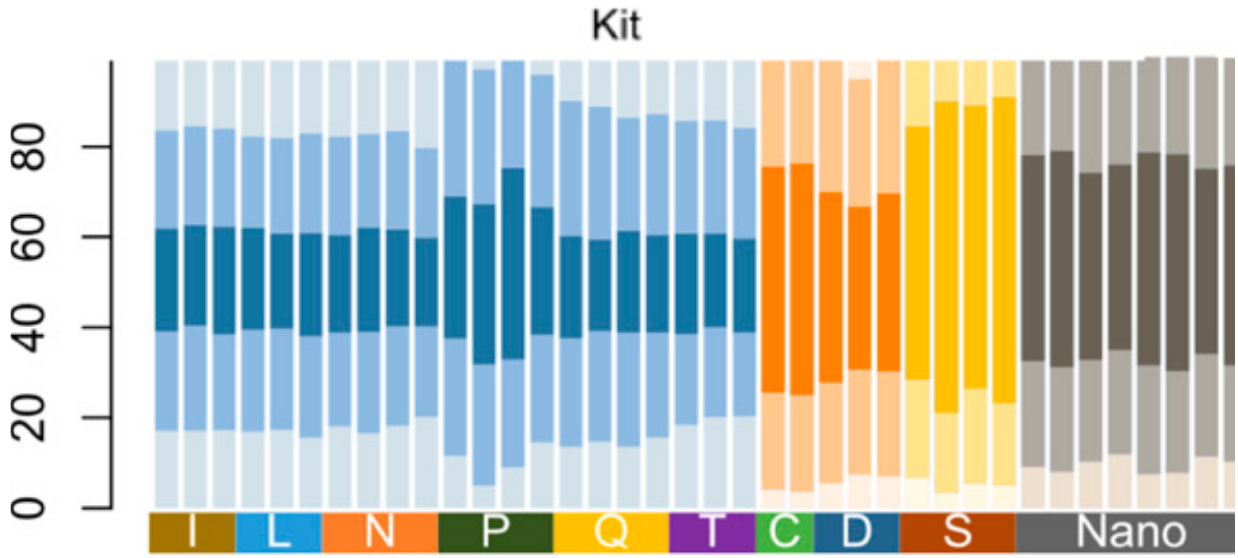


Figure 1.11 Multi-site study of commercially available miRNA cloning kits. A. Bar chart in which bars have been shaded in order to indicate the percentage of 960 synthetic miRNAs (miRXplore Universal Reference) that had sequencing reads within 2-fold of the median value (darkest shade) between 2 and 10-fold above or below the median value (medium shade) and greater than tenfold above or below the median value (lightest shade). **I**, Illumina TruSeq Small RNA Library Prep Kit; **L**, Lexogen Small RNA-Seq Library Prep Kit; **N**, NEBNext Small RNA Library Prep Set; **P**, PerkinElmer NextFlex Small RNA-Seq Kit v.3; **Q**, Qiagen QIAseq miRNA Library Kit; **T**, Trilink CleanTag Small RNA Library Prep Kit; **C**, Takara Bio (Clontech) SMARTer smRNA-Seq Kit; **D**, Diagenode CATS Small RNA-Seq Kit; **S**, Somagenics RealSeq-AC miRNA Library Kit; **Nano**, NanoString nCounter miRNA Expression Assay – this is a non-sequencing method that was used as a control by the authors. From Herbert et al 2020 with permission.

Five of the kits (I, L, N, Q and T) use the traditional two adapter method (Fig 1.9) and these have the smallest area of dark shading in the bars of Fig 1.11, which represent the proportion of the tested

960 synthetic miRNAs that had sequencing reads within two-fold of the median value. The lightest shades are most evident at the top of the bars for I, L, N, Q and T represent miRNAs that had greater than tenfold more reads than the median value. These miRNAs are particularly problematic as they constitute a very high percentage of the total miRNA sequence reads (Fig 1.12). From Figs 1.11 and 1.12 it can be seen that kits P (randomized adapters, Fig 1.10A), C and D (template switching, Fig 1.10B) and S (circularisation, Fig 1.10C), which were all developed to overcome cloning bias, have very few miRNAs that have sequencing reads that are tenfold above the median read value. However, as illustrated in Fig 1.11, 50% of miRNAs from the best performing kit (S) were still over or under-represented by 2 to 10-fold.

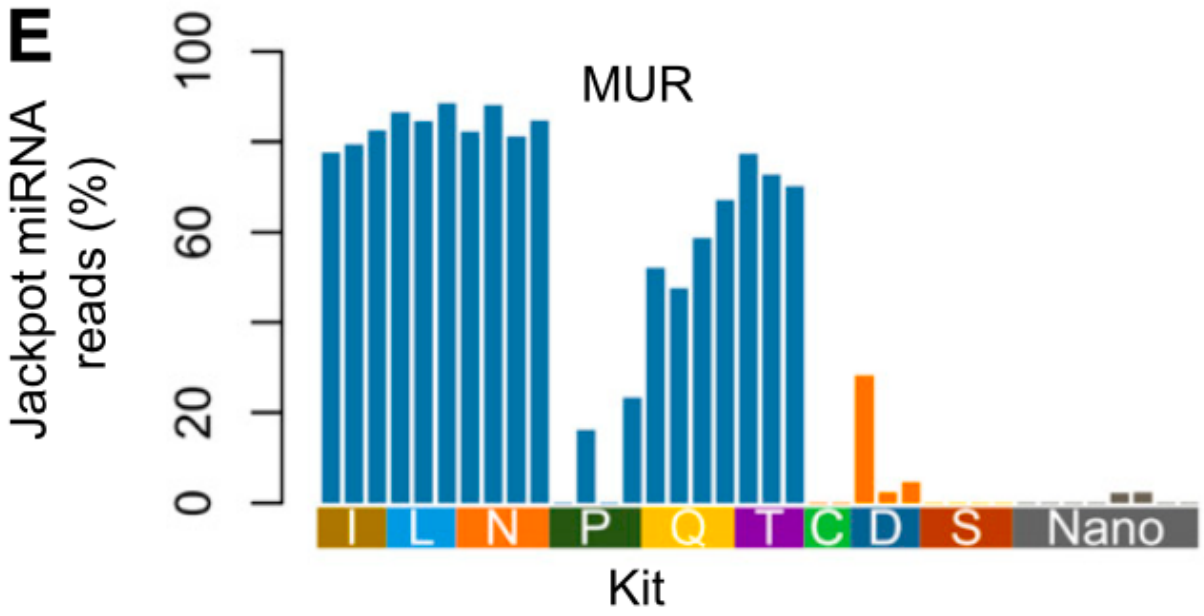


Figure 1.12 From Herbert et al 2020, with permission. The percentage of total reads attributable to a small number of 960 synthetic miRNAs from miRXplore that were easiest to clone (represented by the lightest shades at the top of bars in Fig 1.11).

1.11 Project Aims

1. To find cell lines with strong differences in isomiR expression. Such cell lines could then be used experimentally to identify specific mRNA targets of the isomiRs.
2. To investigate a reported bias problem with miRNA cloning and to incorporate a possible solution into the commonly used illumina method of miRNA cloning and sequencing.
3. To investigate whether the very poor intermolcular cloning of some miRNAs to the RA5 adapter could be improved by intramolecular cloning to a combined RA5RA3 adapter, as suggested by Somagenics.
4. To establish whether the more extensive RNA splicing data that is now available can be used to further improve the approach we developed to detect cryptic splice sites (Kapustin et al., 2011), with emphasis on BRCA1 and BRCA2. We will also test whether RNA splicing databases can be used to predict the effect of splice site mutations upon exon skipping.

Chapter 2-Materials and Methods

2.1 Materials

2.1.1 Cell culture

2.1.1.1 Cell Culture Reagents

Gibco™ Trypsin/ethylenediamine tetraacetic acid (EDTA) (0.05%), phenol red (Invitrogen, Gibco, catalogue number 25300120)

L-glutamine 200mM (100x) (Invitrogen, Gibco, Catalog number: 25030081)

DMEM (Dulbecco's Modified Eagle Medium) with high glucose and L-glutamine and phenol red (Invitrogen, Gibco, Catalog number: 11965118)

Roswell Park Memorial Institute (RPMI) 1640 Medium with L-glutamine, HEPES, High glucose and Phenol Red (Thermo fisher scientific, Gibco, Catalog number: A1049101)

McCoy's 5A (modified) Medium with L-glutamine, High glucose and Phenol Red (Thermo fisher scientific, Gibco, Catalog number: 16600082)

Heat Inactivated Foetal Bovine Serum (Thermo fisher scientific, Gibco, Catalog number: 16140071)

Penicillin/streptomycin (Life Technologies Limited, Gibco, Catalog number: 15140122)

Plasticware and cell culture plates/flasks (Life Sciences, Corning, T25 Product Number: 430639 and T75 Product Number: 431464U)

15 ml Centrifuge tube (Life Sciences, Corning, Product Number 430791)

Dimethyl sulfoxide (DMSO) (Sigma -Aldrich, CAS number: 67-68-5, EC number: 200-664-3)

2.1.1.2 Chemicals and Reagents and Kits

miRVana miRNA Isolation kit (Ambion, Catalog number: AM1560)

MTH adenylation kit (New England Biolabs, Catalog #: E2610S)

µMACS streptavidin kit (Miltenyi Biotec, Catalog number: 130-074-101)

Hybond N+ nylon membrane (Amersham, Thermo fisher scientific, Catalog number: 45-000-838)

SuperScript® IV (SSIV) Reverse Transcriptase (Thermo fisher scientific, Invitrogen, Catalog number: 18090010)

3M Sodium Acetate (Thermo fisher scientific, Invitrogen, Catalog number: AM9740)

10 mM dNTP mix (Thermo fisher scientific, Invitrogen, Catalog number: 18427013)

Agarose, molecular biology grade (Sigma -Aldrich, CAS number: 9012-36-6, EC number: 232-731-8)

50 x Denhardt's solution (Thermo fisher scientific, Invitrogen, Catalog number: 750018)

Ammonium persulfate (APS) (Sigma -Aldrich, CAS number: 7727-54-0, EC number: 231-786-5)

Ampicillin (Sigma -Aldrich, CAS number: 69-53-4)

Bovine serum albumin (BSA) (Sigma -Aldrich, CAS number: 9048-46-8, EC number: 232-936-2)

Bromophenol blue (Sigma -Aldrich, CAS number: 115-39-9, EC number: 204-086-2)

Chloroform (Sigma -Aldrich, CAS number: 67-66-3)

Nuclease-free water (Thermo fisher scientific, Invitrogen, Catalog number: AM9932)

Dithiothreitol (Thermo Scientific, Catalog number: R0861)

Ethanol (BDH, CAS number: 64-17-5)

Ethidium Bromide Solution (Sigma -Aldrich, CAS number: 1239-45-8)

Ethylenediamine tetraacetic acid (EDTA) (Sigma -Aldrich, CAS number: 60-00-4, EC number: 200-449-4)

SYBR™ Gold Nucleic Acid Gel Stain (Thermo fisher scientific, Invitrogen, Catalog number: S11494)

N, N, N', N'-Tetramethyl ethylenediamine (Temed) (Sigma -Aldrich, CAS number: 110-18-9, EC Index number: 203-744-6)

Phenol: Chloroform: isoamyl alcohol 25:24:1 (Sigma -Aldrich, CAS number: 136112-00-0, MDL number: MFCD00133763)

Restriction endonucleases and buffers (New England Biolabs)

RNaseOUT™ Ribonuclease inhibitor (Thermo fisher scientific, Invitrogen, Catalog number: 10777019)

Gel cassettes (Thermo fisher scientific, Invitrogen, Catalog number: NC2010)

2.1.1.3 Buffers and Solutions

10 x DNA loading buffer (Thermo fisher scientific, Invitrogen, Catalog number: 10816015)

or 0.2% (w/v) Bromophenol Blue, 40% (v/v) Glycerol, 100mM EDTA pH8.0

GlycoBlue (Thermo fisher scientific, Invitrogen, Catalog number: AM9516)

RNase ZAP (Thermo fisher scientific, Invitrogen, Catalog number: AM9780)

Super RX Blue Sensitive Film (Fujifilm, MXR, SKU: 100909)

Elution Buffer [10 mM Tris-HCl (pH 7.5), 1 mM EDTA]

Low Salt Buffer [0.15 M NaCl, 20 mM Tris-HCl (pH 7.5), 1 mM EDTA]

Sterile disposable scalpels (Swann-Morton, Code: 05XX)

Denhardt's Solution (50X) (Thermo fisher scientific, Invitrogen, Catalog number: 750018)

Isopropanol (Sigma -Aldrich, CAS number: 67-63-0, MDL number: MFCD00011674)

DNA Gel Loading Dye (6X) (Thermo Scientific™, Catalog number: R0611)

50 bases DNA ladder (New England Biolabs, Catalog number: N3236S)

Formamide (Sigma-Aldrich, CAS number: 75-12-7, product number: F9037)

2.1.1.4 PCR, Northern blotting and Cloning reagents

15% denaturing PAGE: 21 g urea (7M), 2.5 ml 10x TBE, 18.75 ml of 40% (w/v) 19:1 acrylamide:bis-acrylamide, adjust volume to 50 ml with water. Add 350 µl of 10% (w/v) ammonium persulphate (APS) and 17.5 µl of TEMED.

10% denaturing PAGE: as above 15% denaturing PAGE but with 12.5 ml of 40% (w/v) 19:1 acrylamide:bis-acrylamide, containing 7M urea and 0.5 x TBE. Add 350 µl of 10% (w/v) ammonium persulphate (APS) and 17.5 µl of TEMED.

12.5% non-denaturing PAGE: without urea, 15.66 ml of 40% (w/v) acrylamide: bisacrylamide (19:1), 5 ml of 10x TBE, and 29.34 ml of distilled water. Add 350 µl of 10% (w/v) ammonium persulphate (APS) and 17.5 µl of TEMED. Samples loaded with non-denaturing gel loading dye.

8% non-denaturing PAGE: without urea, 2 ml of 40% (w/v) acrylamide: bisacrylamide (19:1), 1 ml of 10x TBE, and 7 ml of distilled water. Add 100 µl of 10% (w/v) ammonium persulphate (APS) and 10 µl of TEMED. Samples loaded with non-denaturing gel loading dye.

Denaturing loading dye: 10 ml deionized formamide, 200 µl 0.5 M EDTA pH 8.0, 1 mg xylene cyanol FF, 1 mg bromophenol blue.

Non-denaturing loading dye: 0.02% w/v 1 M EDTA pH 8.0, 0.25% w/v xylene cyanol FF, 0.25% w/v bromophenol blue, 15% Ficoll in water.

1% Agarose gel: 1 g agarose, 10 ml 10x TAE, 90 ml water. Add 0.1 µg/ml of ethidium bromide.

2.1.1.5 Reagents and Buffers

10 x Phosphate buffer saline (PBS): 0.5 M NaH_2P0_4 (30 g/500 ml), 0.5 M Na_2HP0_4 (35.5 g/500 ml) pH 7.2 for stock solution. 200 ml of stock, add 8.76 g NaCl and make up to 1 liter with distilled water.

10 x TBE: 432 g Tris, 220 g Boric acid, 37.2 g EDTA, Water to 4 liters

20 x TAE: 0.8 M Tris, 0.4 M Sodium acetate, 20 mM EDTA, adjusted to pH 8 with acetic acid

20 x SSC: 175.3 g NaCl, 88.2 g Sodium citrate, water to 1 liter pH 7

TE Buffer: 10 mM Tris-HCL pH 7.4, 1 mM EDTA

DNA elution buffer: 0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% w/v SDS

2.1.2 List of oligos and primers

Chapter 3

For radiation P^{32} labelled miR-101-1-3p oligo

hsa-miR-101-1-3p GUACAGUACU… UACAGUACU…

Mature 5’-uacaguacugugauaacugaa-3’

Probe 5’-ttcagttatcacagtactgta-3’

RA3 5’PO4TGGAATTCTCGGGTGCCAAGG-dideoxyC3

NNRA3 5’PO4NNTGGAATTCTCGGGTGCCAAGG-dideoxyC3

RA5 5’GUUCAGAGUUCUACAGUCCGACGAUC-3’

RA5NN 5’GUUCAGAGUUCUACAGUCCGACGAUCNN-3’

RA5NNNN 5’GUUCAGAGUUCUACAGUCCGACGAUCNNNN-3’

RA5RA3 5’GTTCAGAGTTCTACAGTCCGACGATCTGGAATTCTCGGGTGCCAAGGC-3’

Red font – RNA oligos

miR-101-1-3p /5Phos/GUACAGUACUGUGAUAACUGAAG

miR-205 /5Phos/UCCUUCAUUCCACCGGAGUCUG

miR-214 /5Phos/ACAGCAGGCACAGACAGGCAGUC 23bases

RTP 5’GCCTTGGCACCCGAGAATTCCA-3’ 22bases

RP1 5’AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA-3’ 50 bases

RPI48 5’CAAGCAGAAGACGGCATACGAGATTGCCGAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA-3’ 63 bases

Blue font – bases that are complementary to RA5 and RA3 (see RT-PCR primers below).

STPNN 5’AUUCCANNCCACGUUCCCGUGG-3’

STP 5’GAAUUCCACCACGUUCCCGUGG-3’

5'dd-STP /5InvddT/GAAUUCCACCACGUUCCCGUGG-3’

5’ BiotinTEG-STP /5BiotinTEG/GAAUUCCACCACGUUCCCGUGG-3’

Chapter 4

5'biotin-TEG-HSA-mir-575 /5BiotinTEG/GCTCCTGTCCAACTGGCTC

5'biotin-TEG-HSA-mir-768-3p /5BiotinTEG/GTCAGCAGTTTGAGTGTCAGCATTGTGA

Biotin-TEGmiR-101-1-3p /5BiotinTEG/GTTATCACAGTACTGTAC

RP1 shorter 5'GTTCAGAGTTCTACAGTCCGAC-3'

RPI48 shorter 5'CCTTGGCACCCGAGAATTCC-3'

STPNN 5'AUUCCANNCCACGUUCCCGUGG-3'

NNRA3RA5NN-3'P (No6) - this is the vector used for making miRNA libraries

5'[Phos]NNTGGAATTCTCGGGTGCCAAGGGUUCAGAGUUCUACAGUCCGACGAUCNN[3'Phos] (51 bases)

Other RA3RA5 constructs

No 1 (NNRA3RA5--2OMe)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGUUCAGAGUUCUACAGUCCGACGAUrC-2OMe

No 2 (NNRA3RA5-Ome-rC)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGTTCAGAGTTCTACAGTCCGACGATrC-2OMe

No 3 (NNRA3RA5)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGTTCAGAGTTCTACAGTCCGACGATC

No 4 (NNRA3RA5)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGUUCAGAGUUCUACAGUCCGACGAUC

No 5 (NNRA3RA5NN)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGTTCAGAGTTCTACAGTCCGACGATCNN

alternative No5 (NNRA3RA5NN-3'P)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGTTCAGAGTTCTACAGTCCGACGATCNN/3Phos/

No 7 (NNRA3RA5-AmC3)

/5Phos/NNTGGAATTCTCGGGTGCCAAGGGTTCAGAGTTCTACAGTCCGACGATC-AmC3

NNRA3RA5 /5Phos/NNTGGAATTCTCGGGTGCCAAGGGUUCAGAGUUCUACAGUCCGACGAUmC

Further oligos

214RA3NN /5Phos/ACAGCAGGCACAGACAGGCAGUCTGGAATTCTCGGGTGCCAAGGNN

214-RA5NN /5Phos/ACAGCAGGCACAGACAGGCAGUCGTTCAGAGTTCTACAGTCCGACGATCNN

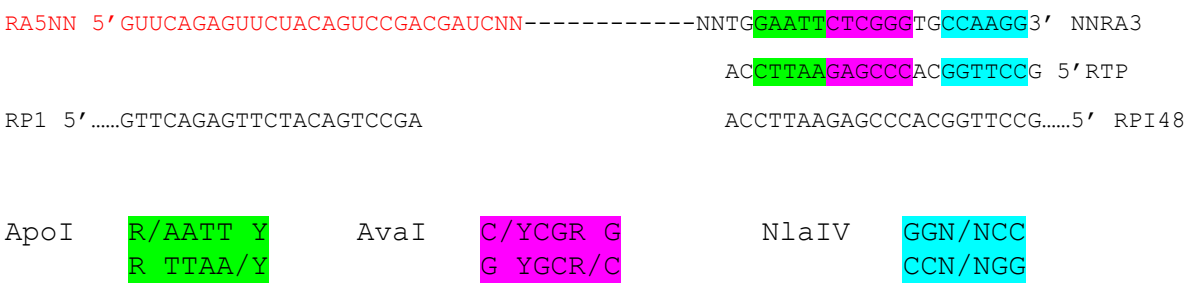
214RA3RA5 ACAGCAGGCACAGACAGGCAGUCTGGAATTCTCGGGTGCCAAGGGUUCAGAGUU

CUACAGUCCGACGAUC 70 bases

214RA3RA5 ACAGCAGGCACAGACAGGCAGTCTGGAATTCTCGGGTGCCAAGGGTTCAGAGTT

CTACAGTCCGACGATC

Detailed RT-PCR diagram of a microRNA (-----) cloned into NNRA3RA5NN (see Figs 3.2, 4.6). Restriction enzyme sites also marked (see Chapter 4).



For miR-214, which is 23 bases in length, the RT-PCR reaction shown above would be expected to generate a linear fragment of RP1 + RPI48 + miR-214 + 7 bases from the 3’ end of RA5NN + 2 bases from the 5’ end of NNRA3, which equals 145 bases.

The restriction enzyme sites were used to investigate and to deal with large sized bands unexpectedly generated by RT-PCR of circular ligations (see Chapter 4).

RT-PCR of circular NNRA3RA5NN-3’P (No6) would be expected to generate a linear fragment of RP1 + RPI48 + 7 bases from the 3’ end of RA5NN + 2 bases from the end of NNRA3 = 122 bases. This would be expected to be cut into double stranded fragments of 56 and 62 bases by ApoI (plus 4 bases single strand overhangs) and double stranded fragments 51 and 67 bases by AvaI (plus 4 bases single strand overhangs).

RT-PCR of circular 214RA3RA5 would be expected to generate a linear double stranded fragment of 141 bases (RP1 + RPI48 + mir-214 + 5 bases from the 3’ end of RA5).

This would be expected to generate double stranded fragments of 56 bases and 81 bases by ApoI and 51 bases and 86 bases by AvaI (plus 4 bases single stranded overhangs).

2.1.3 List of cell lines

List of cell lines

No	cell lines	Characteristics	Origins
			ATCC
U-2 OS	Human osteosarcoma cells	Adherent Cell line derived in 1964 from a moderately differentiated sarcoma of the tibia of a 15-year-old girl.	
THP-1	human leukemia monocytic cell line	Suspension Derived from the peripheral blood of a 1-year-old male with acute monocytic leukaemia	ATCC
MCF-7	Human breast cancer cell line	Adherent Derived from the pleural effusion of a 69-year-old Caucasian metastatic breast cancer (adenocarcinoma) in 1970	ATCC
HL60	Human caucasian promyelocytic leukemia	Suspension derived from peripheral blood leukocytes obtained by	ATCC

	cell line	leukopheresis of a 36-year-old Caucasian female with acute promyelocytic leukemia	
H929	Human caucasian IgA-producing plasmacytoma	Suspension from a malignant effusion in a 62-year-old Caucasian woman with myeloma	ATCC
HCT116	Human colorectal carcinoma cell line	Adherent Established from the primary colon carcinoma of an adult man	ATCC

2.1.4 List of Bioinformatics websites

TargetScan - http://www.targetscan.org/

PicTar - http://pictar.mdc-berlin.de/

UCSC Genome Browser - http://genome.cse.ucsc.edu/cgi-bin/hgBlat

NCBI Blast - http://blast.ncbi.nlm.nih.gov/Blast.cgi

Ensembl Genome Browser: http://www.ensembl.org/index.html

NCBI: http://www.ncbi.nlm.nih.gov/

Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/

Primer3: http://frodo.wi.mit.edu/

miRGator v3.0: http://mirgator.kobic.re.kr/

Blat (UCSC Genome Browser): https://genome.ucsc.edu/cgi-bin/hgBlat?command=start

miRBase: http://www.mirbase.org/

Nucleotide BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch

Primer-BLAST: http://www.ncbi.nlm.nih.gov/tools/primer-blast/

DAVID: http://david.abcc.ncifcrf.gov/

UCSC In-silico PCR: https://genome.ucsc.edu/cgi-bin/hgPcr

2.2 Methods

2.2.1 General cell culture

All culture dishes (Corning, Costar) and flasks (Corning) were suitable for sterile tissue culture and all serological plugged pipettes were purchased from Costar. Adherent cell lines were cultured in DMEM media (Dulbeco's Modified Medium (DMEM) (Invitrogen, Gibco) supplemented with 10% (v/v) heat inactivated Foetal Bovine Serum (FBS) (Life Technologies), 50 U/ml penicillin/streptomycin (Life Technologies Limited, Gibco) and 200 µM glutamine (Invitrogen, Gibco). McCoy's 5A (Modified) Medium (Life Technologies Limited, Gibco) supplemented with 10% (v/v) heat inactivated Foetal Bovine Serum (FBS) (Life Technologies), 50 U/ml penicillin/streptomycin (Life Technologies Limited, Gibco) and 200 µM glutamine (Invitrogen, Gibco) for adherent cells. All the cell culture experiments were carried out in a sterile condition Class II flow cabinet and all cells were grown at 37°C in 5% or 10% CO_2 (v/v).

Cells were passaged twice or three times per week when cells reached or neared confluency. Then all adherent cell lines were washed with PBS and incubated with 1 ml Trypsin-EDTA (0.25%) (Invitrogen, Gibco) for 2 to 5 minutes at 37°C in 5% CO_2 (v/v). Then, the cells were pipetted up and down to make a single cell suspension and re-suspended in 9 mls of warmed D10 media (DMEM + 10% FBS + glutamine + pen/strp) and centrifuged at 1000 rpm for 5 minutes. The cell pellet was resuspended in D10 and plated to the required density.

All suspension cell lines were grown in R10 ie RPMI 1640 Medium (Life Technologies Limited, Gibco) supplemented with 10% (v/v) heat inactivated FBS, 50 U/ml penicillin/streptomycin (Life Technologies Limited, Gibco) and 200 µM glutamine (Invitrogen, Gibco) in 5% CO_2 (v/v). Suspension cells were grown in 6 well plates and passaged weekly by putting 1/3 ml, 1/2 ml and 1 ml of cells into three plates containing 4 mls of R10 RPMI 1640 Medium (Life Technologies Limited, Gibco) supplemented with 10% (v/v) heat inactivated FBS, 50 U/ml penicillin/streptomycin (Life Technologies Limited, Gibco), 200 µM glutamine (Invitrogen, Gibco) and another three plates containing 4 mls R10 as control.

2.2.2 Freezing down cell lines

The freezing method of adherent cells and suspended cells was the same, except that detaching the adherent cells from the culture plates was required before the freezing procedure. Cells were grown in T75 flasks or six well plates until confluent. The cells were washed with 1x PBS and the adherent cells were treated with 1 ml of Trypsin-EDTA (0.25%) as described above and then the trypsin was inactivated by adding 9 mls of media. The cells were spun down at 1000 rpm for 5 minutes and the pellet was re-suspended in D10 media, R10 media or M10 media with 10% dimethyl sulfoxide (DMSO) and immediately aliquoted into 1 ml cryo-vials (0.7 ml per vial). The cells were frozen slowly in a cryo

freezing container containing isopropyl alcohol at -80°C at least for two days before being stored in liquid nitrogen.

2.2.3 Thawing the frozen cells

The frozen cells were taken from liquid nitrogen and were revived by thawing at 37°C in a water bath. The cells were then pipetted up and down and then added to 9 mls of warmed media in a falcon tube. The cells were centrifuged at 1000 rpm for 10 minutes. The media and DMSO were removed and a small amount of new media was added to the cell pellet which was gently pipetted into single cells and further diluted with fresh media, which was then transferred to the appropriate tissue culture dish and then incubated at 37°C in 5% CO_2 (v/v).

2.2.4 Total RNA extraction

The cells were lysed, and total RNA was extracted with TRIzol (Invitrogen) as instructed by the manufacturer. A T75 flask of sub-confluent to confluent cells could generate 20 µg - 100 µg of total RNA depending on the cell type. 5 mls of TRIzol reagent was used per 10 cm dish or T75 flask and incubated for 2-3 minutes at room temperature and then passed several times through a pipette to form a homogenous solution. The sample was transferred to eppendorf tubes and 0.2 ml of chloroform was added per 1 ml of TRIzol, shaken vigorously for 15 seconds and then centrifuged for 15 minutes in eppendorf tubes at 12,000 x g at 4°C. The aqueous phase was transferred to a new tube and 0.5 ml of isopropanol per 1 ml of TRIzol reagent was added and the tubes were inverted several times and left on ice for 10 minutes and then centrifuged at 12,000 x g for 10 minutes at 4°C. The RNA pellet was washed with 1 ml of 75% ethanol per 1 ml of TRIzol reagent and centrifuged for 5 minutes at 7500 x g at 4°C. The ethanol was removed and the pellet was resuspended in 20-50 µl

of nuclease-free water. The concentration of the RNA sample was measured by using a Nanodrop ND1000 spectrophotometer.

2.2.5 Small RNA extraction

About 100 µg of total RNA samples were enriched for small RNAs using the miRVana miRNA Isolation kit (Ambion) following the manufacturer's instructions. Small RNAs were precipitated with 1/3 volume of 100% ethanol (e.g. add 100 µl 100% ethanol to 300 µl aqueous phase), and then 2/3 volume of 100% ethanol (e.g. add 266 µl 100% ethanol to 400 µl of filtrate is recovered) miRVana miRNA Isolation kit (Ambion). The RNAs were centrifuged at 10,000 rpm in a microcentrifuge at 4°C. The RNAs were washed with miRNA wash solution 1 and wash solution 2/3 and after washing the small RNAs were recovered from 100 µl of nuclease-free water (Thermo fisher scientific).

2.2.6 RNA quality checking

Agarose powder was weighed out and 0.5 g was mixed with 50 mls 1x TAE buffer to make a 1% of agarose gel and the agarose was dissolved in a conical flask for 1-3 minutes in a microwave but not overboiled, because the buffer will be evaporated, and the concentration will not be accurate. Gels were cast in appropriate trays with suitable combs to produce wells and left for half hour to cool. Ethidium bromide solution (10 mg/ml) was added to the gel mix to make a final concentration of 0.5 µg/ml. Set gels were then placed in electrophoresis tanks and 1x TAE buffer was added to cover the gel and electrodes.

500 ng of total RNA of each sample was mixed with 6 x DNA loading dye (2 µl to each tube) and 10 µl of formamide added to a concentration of at least 60% and loaded into wells along with a suitable

DNA ladder. Before loading on the gel, the RNA sample with loading buffer was heated at 70°C for 5 minutes and then placed on ice.

The RNA samples were run on the gel at a constant 70 voltage of 4 V/cm for 1 hour and the agarose gel was stained with Ethidium Bromide (Sigma) and gel was viewed and imaged by UV (ultraviolet) light. For good quality total RNA the 28S ribosomal band was twice the intensity of the 18S RNA band.

2.2.7 Northern blot

2.2.7.1 Denaturing PAGE and RNA transfer

A 15% polyacrylamide denaturing PAGE gel was prepared by dissolving 21 g Urea (7M Urea) in 2.5 ml 10x TBE and 18.75 ml of 40% (w/v) 19:1 acrylamide: bis-acrylamide and adjust to 50 ml with water. After the urea dissolved, 500 µl of 10% (w/v) ammonium persulphate (APS) and 20 µl of TEMED were added and the gel was poured between glass plates and inserted a comb. Then 50-100 ng of RNA oligo was denatured at 70°C for 5 minutes, incubated on ice for 1-2 minutes and the RNA samples were loaded as equal amount and equal volume. The RNA samples were separated on the 15% polyacrylamide denaturing gel (7M Urea) in 0.5 x TBE buffer at 150-250V. The gel was stained with SYBR® Gold stain and viewed by UV light. The gel was washed by 0.5 x TBE to remove excess urea. Six pieces of Whatman filter paper, and 1 piece of Hybond N+ nylon membrane (Amersham) were cut the same size as the gel size and soaked in 0.5 x TBE. This is a semi-dry transfer apparatus was cleaned by RNaseZAP™ (Sigma-Aldrich®). The transfer 'sandwich' on the semi-dry transfer apparatus in order as following: three pieces of filter paper were placed on the semi-dry blot apparatus, followed by the membrane and then the gel was put on top of the membrane after that the remaining three pieces of filter paper were put on top of the gel. A long pipette was used to roll over the top of

the gel and filter papers 'sandwich' to squeeze out the bubbles and any excess liquid was removed. The semi-dry apparatus was run at 3.3 mA per cm^2 for 35 minutes (~5 V) and it is better not to exceed 20 V. The membrane was washed in 0.5 x TBE for 5 minutes, and then the RNA was fixed to the membrane by UV radiation in a UVP-CL-1000 UItraviolet cross-linked at 1200 microjoules, twice.

2.2.7.2 Preparation of oligonucleotide probe with [γ-^{32}P] ATP

Oligonucleotide probes were labelled with high specific activity [γ-^{32}P] ATP (6000Ci/mmol 10mCi/ml, Perkin Elmer®). The oligonucleotide was diluted 1:30 in nuclease-free water with 2.5 µl of 10X T4 polynucleotide kinase buffer (New England Biolabs) with 25 pmol of the oligonucleotide probe and 2 µl of [γ-^{32}P] ATP and made up to 25 µl with nuclease-free water, the last was adding 1 µl of T4 polynucleotide kinase enzyme (New England Biolabs). The sample mix was incubated at 37°C for 1 hour.

To ethanol precipitate the probe and remove any excess [γ-^{32}P] ATP, 55 µl of Tris-EDTA (pH 8.0) buffer, 1 µl of Glycoblue™ (ThermoScientific™) 15 mg/ml and followed by adding 20 µl of ammonium acetate (10M) were spun down, and finally were mixed with 250 µl of 100% Ethanol (kept at -20°C) and left on ice for 1-2 hours. The sample was centrifuged at maximum speed for 20 minutes at 4°C. The supernatant was transferred to another new tube, and the pellet was air dried before resuspending in 100 µl TE (Tris-EDTA) buffer. The radioactivity of the precipitate and the removed supernatant were checked with a Geiger counter, which used to consider a reading of 2:1 ratio of pellet: supernatant as a successful labelling. The resuspended pellet was stored at -80°C.

2.2.7.3 Hybridisation with [γ-^{32}P] ATP

The membrane was placed in a 60 ml container with the RNA side facing up and was treated with 2 x SSC + 0.1% SDS with gentle agitation for 10 minutes at room temperature. 15 mls of hybridisation buffer was consisting of 4.5 ml of 20 x SSC, 1.5 ml of Denhardt's solution (Thermo fisher scientific), 0.375 ml of 20% SDS, and added water up to 15 mls. The membrane was prehybridised with 4 mls of hybridisation buffer to each tube and then preheated for 30 minutes with constant rotation at 42°C. The pre-hybridisation buffer was removed and replaced with 1.5-2 mls of fresh hybridisation buffer plus 100 µl of [γ-^{32}P] ATP labelled oligonucleotide probe. The probe was hybridised to the membrane for at least 12 hours overnight at 42°C in the same conditions. The membrane hybridised with [γ-^{32}P] ATP labelled probe was washed 2-4 times with 2 x SSC + 0.1% SDS at room temperature, 10 minutes for each (1/3rd volume of the bottle). The membrane was wrapped in saran wrap and placed inside a cassette with an x-ray film on top in an intensifying screen at -80°C for at least 5 days or up to 10 days. The film was developed in a dark room using OPTIMAX film processor.

2.2.8 General Ligation and Cloning Protocol

2.2.8.1 Oligonucleotide Adenylation

Prior to ligation, the RA3 DNA oligonucleotide was adenylated at the 5' end using the MTH adenylation kit (New England Biolabs). 50-100 ng of RA3 was added to 2 µl of 10X 5' DNA adenylation reaction buffer, 2 µl of 1 mM ATP, 2 µl of Mth RNA ligase 100 pmol and made up to 18 µl with nuclease-free water and then 2 µl of Mth RNA ligase (100 pmol) was added. The sample was incubated at 65°C for 1 hour and the enzyme inactivated by incubation at 85°C for 10 minutes.

2.2.8.2 Ligation

2.2.8.2.1 3' Ligation reaction

This is oligonucleotide adenylation ligation. The adenylated oligo 3' adapter and microRNA were ligated with 3' ligation T4RNL2 kit (New England Biolabs). 1 µl of 50 ng 3' adenylated linker and 1 µl of P^{32} labelled miR-101-1-3p oligo or 50 ng of unlabelled oligo were added with 2 µl of 10X RNA Reaction buffer, 4 µl of 50% PEG8000, 1 µl of RNA ligase enzyme 2 truncated (Δ) and made up to 20 µl with nuclease-free water. The ligation was incubated at room temperature for 1 hour. The reaction was inactivated at 60°C for 20 minutes. RNA ligase 2 truncated only ligates oligos that are pre-adenylated.

2.2.8.2.2 5' Ligation reaction

The ligated 3' adenylated linker and microRNA then were added to 1 µl of 50-100 ng amounts of 5' adapter with 3 µl of 10X RNA Reaction buffer, 4 µl of 50% PEG8000, and 1 µl of RNA ligase enzyme 1 or 2, and 2 µl of ATP and then made up to 30 µl with nuclease-free water. The reaction was carried out at 25°C or 37°C temperature for 1-2 hours. The reaction was inactivated at 60°C for 15 minutes. The thermostable ligation reaction and a thermostable RNA ligase was used following the manufacture's instruction. After 5' ligation with oligos, the sample was added the equal volume of 2X loading buffer and briefly heated to 95°C and quickly cooled down on ice for 1-2 minutes. The sample was ready to load on 15% acrylamide gels containing 7 M urea and then stain the gel with SYBR® Gold stain (Invitrogen) and viewed by UV light.

2.2.8.2.3 Circular ligation using the NNRA3RA5NN-3'P microRNA cloning vector (see Fig 4.12)
Copy of Fig 4.12

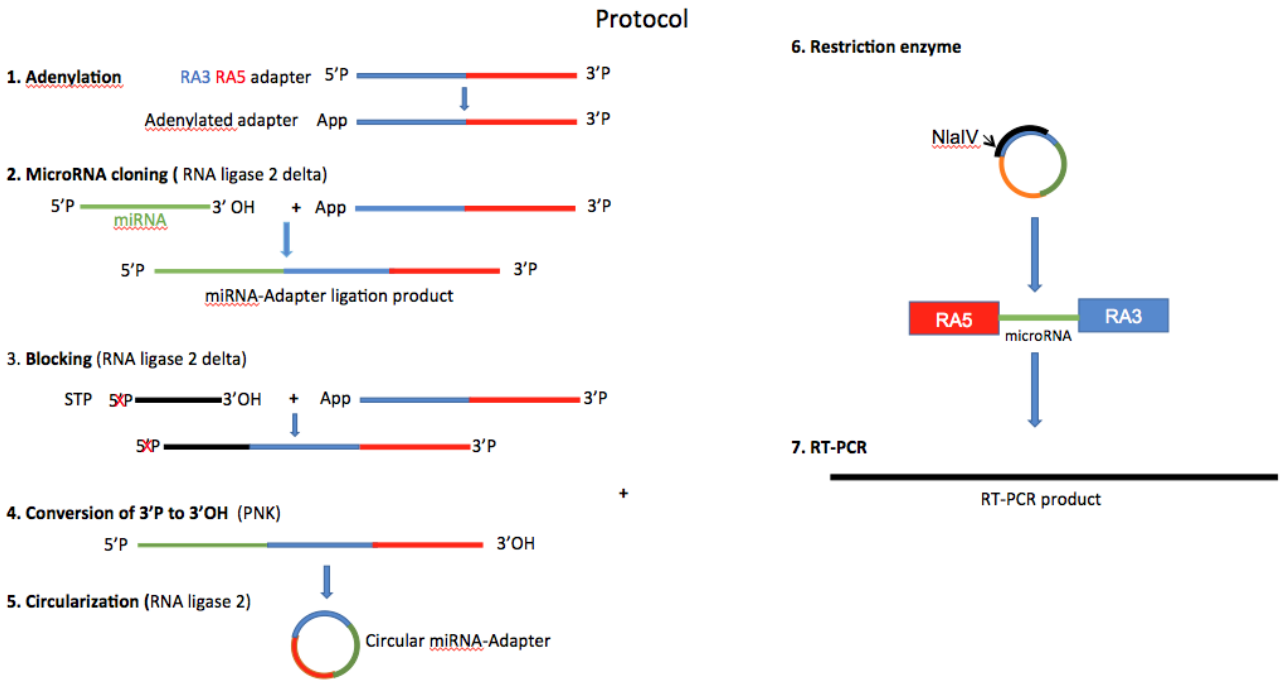


Figure 2.1 Illustration of the full protocol that was developed to clone miRNAs.

1. Adenylation

The NNRA3RA5NN-3'P oligonucleotide was adenylated at the 5' end using the MTH adenylation kit (New England Biolabs). 50-100 ng of NNRA3RA5NN-3'P was added to 2 µl of 10X 5' DNA adenylation reaction buffer, 2 µl of 1 mM ATP, 2 µl of Mth RNA ligase 100 pmol and made up to 18 µl with nuclease-free water and then 2 µl of Mth RNA ligase (100 pmol) was added. The sample was incubated at 65°C for 1 hour and the enzyme inactivated by incubation at 85°C for 10 minutes.

2. MicroRNA cloning

1 µl of 100 ng Ad-NNRA3RA5NN-3'P adapter was added with 2 µl of 1-2 µg total RNA or 1 µg small RNA (2.2.2) in nuclease-free water made up to 6 µl in a new 200 µl PCR tube and keep on ice. The mix was in the tube pipetted up and down and centrifuged briefly, after that the mix was placed on

the preheated thermal cycler and incubated at 70°C for 2 minutes. The tube was removed from the thermal cycler and placed on ice. Preheated the thermal cycler to 25°C. The following volumes were mixed in a new 200 µl PCR tube and kept on ice: 2 µl of 10X RNA Reaction buffer, 4 µl of 50% PEG8000, 1 µl of RNA ligase enzyme 2 truncated (Δ) and made up to 19 µl with nuclease-free water and then 1 µl of RNA ligase enzyme 2 truncated (Δ) was added. The ligation was incubated at 25°C for 1 hour. The reaction was inactivated at 60°C for 20 minutes.

Cloning synthetic miRNAs. 100 ng of adenylated NNRA3RA5NN-3'P adapter was ligated to 50 ng of synthetic microRNA oligos as described above for total RNA.

3. Blocking (STP ligation)

1 µl of STP (100 ng) was added and incubated at 25°C for 30 minutes.

4. Conversion of 3'P to 3'OH

Removal of 3'phosphate by PNK does not require ATP but its presence from step 2 is probably helpful because it should prevent unwanted 5'P removal. For 3' phosphate removal the following were added: 2.5 µl of 10X T4 PNK Reaction Buffer (New England Biolabs) and 1 µl (10 units) of T4 Polynucleotide Kinase (New England Biolabs) and the reaction made up to a total volume of 25 µl with nuclease-free water. The reaction was incubated at 37°C for 30 minutes and the enzyme was then heat inactivated at 65°C for 20 minutes.

5. Circularisaion

The thermal cycler was preheated to 70°C. The Ad-NNRA3RA5NN-3'P plus microRNA tube was placed on the preheated thermal cycler and incubated at 70°C for 2 minutes. The mix in the tube was removed from the thermal cycler and placed on ice. The thermal cycler was preheated at 25°C. 2 µl of 10 mM ATP was added to the tube of Ad-NNRA3RA5NN-3'P and pipetted up and down a few times. 1 µl of T4 RNA Ligase 1 or 2 was added to the Ad-NNRA3RA5NN-3'P/ATP mixture. The mixture was pipetted up and down a few times and placed on the preheated thermal cycler at 25°C for 1 hour. The tube was removed from the thermal cycler and kept on ice. This ligation was ready for the Reverse Transcribe RT-PCR following experiment.

6. Restriction enzyme treatment (see Chapter 4)

11 µl of ligated RNA sample, 1 µl of dNTPs and 1 µl of 100-250 ng of gene specific primer RTP was added and denatured at 65°C for 5 minutes and put on ice for 1-2 minutes. After that 2 µl of 10X rCutSmart™ Buffer, 1 µl of 2 U/ µl of NlaIV enzyme was added with nuclease-free water made up to a total volume of 20 µl. The samples were incubated at 37°C for 1 hour.

7. RT-PCR (see 2.5.1 to 2.5.2)

4 µl of 5x SSIV reverse transcriptase buffer, 1 µl of DTT, 0.5 µl of RNaseOUT and 1 µl of Superscript SSIV was added. The reaction for reverse transcriptase was incubated at 55°C for 10 minutes and inactived enzyme at 80°C for 10 minutes.

1 µl of 10 µM of RP1 and RPI48 primers mix was added, 1 µl of cDNA, 1.25 µl of DMSO and then made up to a total volume of 25 µl of 12.5 µl of 2X Dream taq mastermix – 25-35 for PCR cycles.

2.2.8.2.4 CircLigase™ I and CircLigase™ II Ligation

Circligase ssDNA ligase is a thermostable ATP-dependent ligase, however, CircLigase II ssDNA Ligase is a thermostable ligase that catalyzes intramolecular ligation. The linear ssDNAs were more than 25 nucleotides were circularized by CircLigase I or CircLigase II ssDNA Ligase, which was under standard reaction condictions. 50 ng of single-stranded DNA was added 2 µl of CircLigase 10X Reaction Buffer, 1 µl of 50 mM MnCl2, 1 µl of 1 mM ATP, 1 µl of CircLigase ssDNA Ligase (100 U) or 2 µl of CircLigase II 10X Reaction Buffer, 1 µl of CircLigase II ssDNA Ligase (100 U), 1 µl of 50 mM $MnCl_2$, 4 µl of 5 M Betaine and made up to 20 µl with nuclease-free water for CircLigase I and CircLigase II Ligation, separately. The sample was incubated at 60°C for 1 hour and inactivated the enzyme at 80°C for 10 minutes.

2.2.8.2.5 Exonuclease I and T7 Exonuclease

Exonuclease I was DNA specific exonuclease and catalysed to remove the nucleotides from linear single-stranded DNA from 3’ to 5’ direction. This enzyme with 1X Exonuclease I reaction buffer catalysed the release of 10 nmol of nucleotide in a total volume of 50 µl, which was incubated at 37°C for 30 minutes. This enzyme removes the linear single-stranded DNA or RNA.

T7 Exonuclease is also known as RNase T and is a double-stranded DNA specific exonuclease. The T7 Exonuclease starts at the 5' termini of linear or nicked double-stranded DNA and removes nucleotides from linear single-stranded or double-stranded DNA or RNA in a 5’ to 3’ direction. This enzyme with 1X NEBuffer 4 reaction buffer required to release 1 nmol of single dT nucleotides in a total volume of 50 µl, which was incubated at 25°C for 30 minutes. This enzyme is a single-stranded (ssDNA or RNA) specific exonuclease, which leaves behind double-stranded DNA in the sample.

2.2.9 General PCR and cloning

2.2.9.1 cDNA conversion

SuperScript® IV (SSIV) reverse transcriptase (Thermo fisher scientific) was to make complementary DNA using the manufacturer's protocol. After ligation of total RNA or microRNA oligos into Ad-NNRA3RA5NN-3'P, 11 µl of the ligation sample was mixed with 1 µl of 10 mM dNTPs, and 50 –250 ng of the oligo RTP. The tube was heated at 65°C for 5 minutes and left on ice for at least 1-2 minutes. 1 µl of 100 mM DTT, 4 µl of 5x SSIV buffer, 0.5 µl of RNaseOUT™ (40 U/µl) (Life Technologies™) and finally 1 µl of SuperScript® IV reverse transcriptase enzyme (200 U/µl) (Thermo fisher scientific) were added to the reaction. The sample was incubated at 55°C for 10 minutes and inactivated enzyme at 80°C for another 10 minutes. The sample was stored at -20°C.

2.2.9.2 Polymerase chain reaction (PCR)

Polymerase chain reaction was set up by using DreamTaq Master Mix (Life Technologies Limited). 12.5 µl of 2X DreamTaq Master Mix was used which consists of DreamTaq DNA polymerase, DreamTaq buffer, $MgCl_2$ and dNTPs. To this 1 µl of cDNA, 1 µl of a primer mix of 10 µM each of forward and reverse gene-specific primers, 1.25 µl of DMSO were added and 9.25 µl of nuclease-free water made up to a total volume of 25 µl for each sample. Thermo Hybaid PCR Express Thermal Cycler was performed for RT-PCR reaction. The PCR reaction was set at 94°C for 2 minutes, followed by 15 to 35 cycles of 94°C for 30 seconds, 50-60°C for 30 seconds and 70-72°C for 30 seconds and terminated at 70-72°C for 10 minutes.

All PCR products were run on an 8% or 12.5% non-denaturing acrylamide gel (PAGE) to check the product sizes. The 8% gel was mixed with 2 ml of 40% (w/v) acrylamide: bisacrylamide (19:1) (Sigma-Aldrich®), 1 ml of 10x TBE, and 7 ml of distilled water made up to a total volume of 10 ml in a 12 ml

falcon tube. After that 100 µl of 10% (w/v) ammonium persulfate (APS) and 10 µl of TEMED were added to the gel solution. The 12.5% non-denaturing gel was mixed with 15.66 ml of 40% (w/v) acrylamide: bisacrylamide (19:1) (Sigma-Aldrich®), 5 ml of 10x TBE, and 29.34 ml of distilled water. 350 µl of 10% (w/v) ammonium persulphate (APS) and 17.5 µl of TEMED were added to the gel solution. The gel solution in the tube was quickly poured into 1 mm gel cassettes (ThermoScientific™) and 10-12 well combs were positioned inside of cassette.

Each PCR sample was mixed with 2 µl of 6 x DNA loading dye containing bromophenol blue and xylene cyanol and these samples plus 3.5 µl of a 50 bases DNA ladder (New England Biolabs®) were run on a TBE non-denaturing acrylamide gel at 100 volts about 1 hour until the blue dye runs on the bottom of the gel and then added 1X SYBR® Gold staining solution incubation for 10-40 minutes at room temperature, which finally visualised for bands on a UV light box.

2.2.9.3 PCR product elution and DNA extraction from gel and sequencing

The GenElute DNA Gel Extraction kit was used for extracting DNA from agarose gels. DNA bands of the desired molecular weight were cut out from the gel and cut into small pieces with a sterile blade while visualizing under blue-light. The gel slice was weighed and 3 volumes of gel solubilization solution (0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% SDS) was added (for example: 100 mg gel slice, 300 µl of solution was added) to the gel pieces in a 1.5 ml microcentriguge tube and incubated with shaking (700 rpm) at 16°C overnight or for more than 2 hours at 37°C. The gel solution was filtered in a 0.42 µm spin-x centrifuge filter tubes (Costar® Spin-X®) and spun down for 2 minutes at 9500 x rpm at 4°C. The filtrate was removed and the remaining small pieces of gel on top of the filter were washed with fresh 100 µl of DNA elution buffer in the same conditions.

To precipitate the DNA, 45 µl of 3 M sodium acetate (1 in 10 of the total volume (350 + 50 µl), was added and then 1 µl of Glycoblue™ coprecipitant (ThermoScientific™, 15 mg/ml of glycogen) to and finally 1.3 mls (2.5-3 times of the total volume) of ice cold ethanol were added to each tube and the mixed tube was turned over several times and left on ice at least for 2 hours on ice or at -20°C overnight. The ethanol precipitate was spun down at 13,000 rpm for 30 minutes and then washed by 1 ml of 75% ethanol for 10 minutes. The DNA pellet was resuspended in 9 µl of nuclease-free water and 1 µl of 10 µM of the forward or reverse primer was added for sanger sequencing.

Sanger sequencing was located at the MRC CSC Genomics Core Laboratory at the Hammersmith Campus of Imperial College London or by GenewizUK. The obtained sequence was analysed by using the Nucleotide Blast ® online tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) in order to confirm the DNA position of the PCR sequences.

2.2.9.4 Enzymes selection

Enzymes selection and restriction endonuclease digestion restriction enzymes were selected by www.restrictionmapper.org website. ApoI, AvaI, EcoRI and NlaIV were used. Digestion mix was prepared in a 1.5 ml DNase free eppendorf tube. These restriction enzyme digests were performed by restriction enzymes (New NEB Biolabs) with the 1X recommended buffers and at the recommended reaction temperatures. All the enzymes were used with 2 µl of 10X restriction enzyme buffer 1, 2, 3 or 4 depending on which enzyme was chosen, 1 µl of restriction enzyme (New England Biolabs), 1 µl of 1-2 µg DNA sample or 100 ng of oligo product and nuclease-free water made up to a total volume of 20 µl and incubated at a recommended temperature for 1-2 hours and the reactions were inactivated at 65°C for 10 minutes and then put on ice for 5 minutes. The digested product was analysed by agarose gel electrophoresis.

2.2.10 Pull down

2.2.10.1 Incubation with streptavidin magnetic beads (NEB)

The µMACS streptavidin kit (Miltenyi Biotec) was used for pulldown experiment, and the protocol for isolating specific microRNAs provided by the manufacturer as follows. Complementary oligonucleotides to miR-101-1-3p, mir-575 and mir-768 were designed with Biotin-TEG on their 5' ends as shown in Section 2.8. These oligonucleotides were added in TEN buffer consisting of 10 mM Tris/HCl pH 8.0, 1 mM EDTA, 100 mM NaCl made up to 70 µl volume for each sample which was heated at 75°C denature for 5 mintutes immediately prior to annealing to 1 µl of 5 fmol/µl per RNA oligonucleotide of the miRNA reference library miRXplore.

The annealing temperatures and conditions were based on previous empirical results by Diana Alexieva (PhD Thesis https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.745277) and were normally around 37-45°C. The calculated temperature (for example miR-101-1-3p) was 43°C with adding 1 µl 0.5 µg of the biotinylated capture DNA (for example 5' end Biotin-TEGmiR-101-1-3p see Section 2.8). The final volume should not exceed 900 µl. (500 pmol of sing-strand 20 basepair biotin oligo nucleotide per mg; 36 µl of beads per oligo).

After the annealing step 36 µl of NEB magnetic streptavidin beads were added, vortexed and incubated for 10 minutes on ice or 4°C. For rare transcripts, a longer incubation time may be necessary. The beads were isolated by applying a magnet to the side of the tubes for 30 seconds and removing the supernatant. The beads were then washed by adding 100 µl of wash/binding buffer (0.5 M NaCl, 20 mM Tris-HCl (pH 7.5), 1 mM EDTA), which was vortexed to suspend and then applied to magnet on the side of the tubes for 30 seconds to 1 minute and then discard supernatant. Repeat wash twice. The total RNA samples were added to previously prepared magnet beads and vortexed to suspend the particles then incubate at room temperature for 10 minutes with every 3 minutes

flapped by hand. After that the tube with beads applied magnet 30 seconds to 1 minute and supernatant was removed. The 100 µl of wash/binding buffer was added vortex to suspend beads and then the supernatant was removed. Repeat washing twice with fresh wash buffer. 100 µl of the cold low salt buffer [0.15 M NaCl, 20 mM Tris-HCl (pH 7.5), 1 mM EDTA] was added to each bead, vortexed to suspend, which were applied magnet 30 seconds to 1 minute and then supernatant was removed.

The miRNA reference library miRXplore or the total RNAs that annealed to the biotinylated DNA sequences were eluted by adding 25 µl of elution buffer (10 mM Tris-HCl (pH 7.5), 1mM EDTA) that was preheated to 70°C. The microcentrifuge tube was vortexed to suspend beads and then incubated at room temperature for 2 minutes, after that applied magnet 30 seconds to 1 minute to transfer supernatant to a clean RNase-free microcentrifuge tube.

The eluted products were precipitated with NaCl/Isopropanol. The elution was precipitated by adding 350 µl of nuclease-free water, 40 µl of 1/10 5M NaCl and 1 µl of Glycogen. The mix was vortexed and 400 µl of Isopropanol was added and incubated for 15-20 minutes at 4°C, after that spun down at maximum speed for 30 minutes at 4°C. The pellet was washed three times with 70% ethanol at maximum speed for 5 minutes each time and diluted with 20 µl of nuclease-free water. The samples would be ready for loading directly onto the gel or for RT-PCR into sequencing.

2.2.10.2 Denaturing gel

10% denaturing polyacrylamide gel (40% (w/v) bisacrylamide (19:1)) containing 7M urea and 0.5 x TBE was prepared as previously. Each sample was loaded with the same volume and mixed with 6 x DNA loading dye that was run at 250 V until the bromophenol dye reached the bottom of the gel.

The gel was added 1X SYBR® Gold staining solution incubation for 10-40 minutes at room temperature, which finally visualised for bands on a UV light box.

2.3 Materials and methods for chapter 5

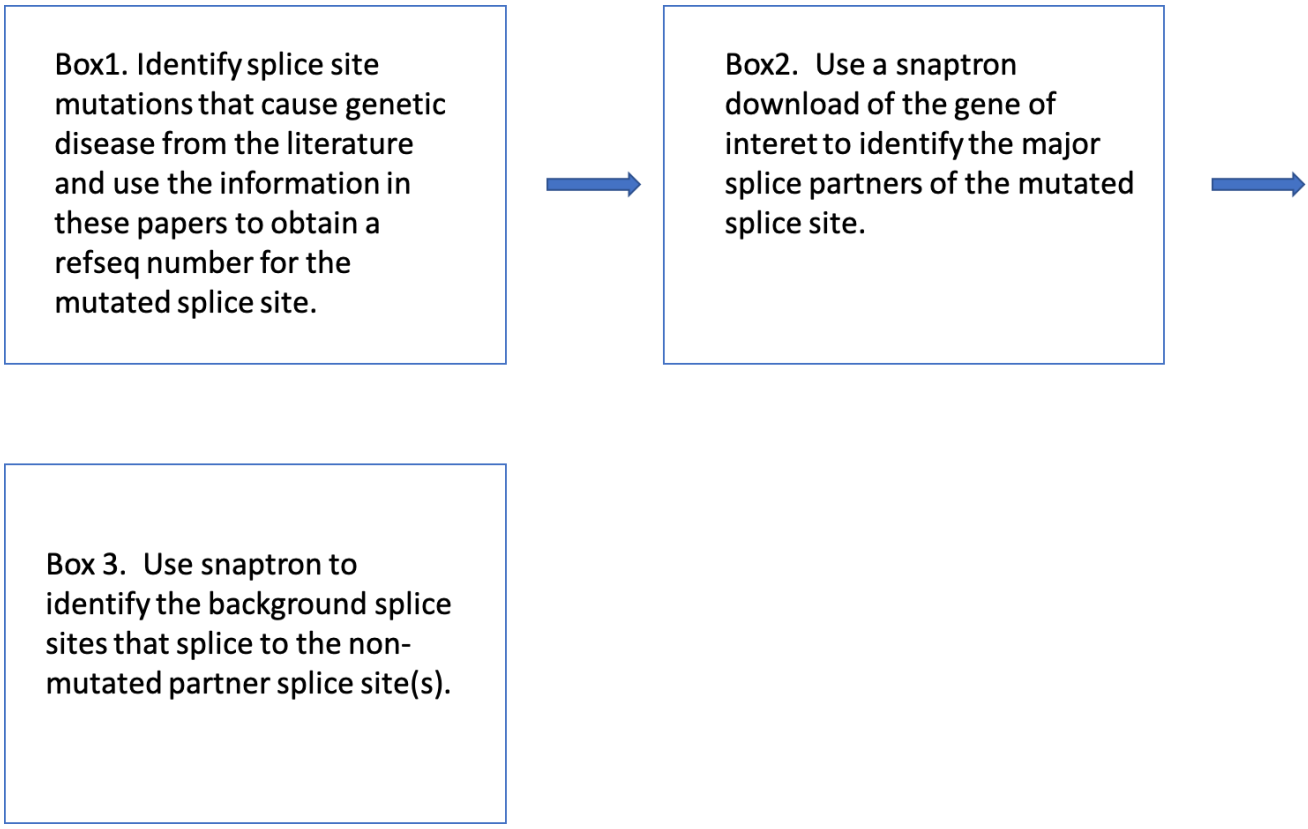


Figure 2.2 Flow diagram to identify background splice sites that are likely to be activated by splice site mutations. Box 1. The ref seq number can be obtained by a number of approaches. If sufficient sequencing data is given in the paper then use BLAT. The mutated splice site can also be identified from information about the size of the affected exon. Exon sizes of the gene of interest can be obtained from LOVD or from a snaptron download (Alexieva et al. 2022). Boxes 2 and 3: as illustrated in fig 5.1 and Tables 5. Note – you may have to change the fig 5.1 and Table 5.1 numbers if additional figures are added to the Introduction of Chapter 5.

Experimental reports of mutations that cause aberrant splicing of BRCA1 and BRCA2 were obtained from the database of aberrant splice sites (DBASS) the human genome mutation database (HGMD), the Leiden Open Variation Database online (LOVD) and by searching Pubmed (Buratti et al., 2011, Fokkema et al., 2011, Stenson et al., 2020). We used the BLAT tool (Kent, 2002) from UCSC website http://genome.ucsc.edu/ (Kent et al., 2002) to obtain genome reference numbers for relevant splice sites.

We then compared the above experimental database of aberrant splicing to the Snaptron database of spliced RNA sequences (Wilks et al., 2018). We downloaded Snaptron data for individual genes in a manner that allowed us to identify background splicing events that might be activated by splicing mutations (Fig 5.1). BRCA1 splicing data was downloaded from Snaptron by using the link (http://snaptron.cs.jhu.edu/srav1/snaptron?regions=BRCA1). RNA splicing data for any other gene can be obtained by changing BRCA1 to the required gene name ie (http://snaptron.cs.jhu.edu/srav1/snaptron?regions=BRCA2). To access the other spliced RNA database of Snaptron srav1 can be changed to srav2, gtex or tcga (see below).

The downloaded splicing reads for BRCA1 can be pasted into excel although we prefer LibreOffice Calc. Spreadsheet commands can then be used to order the splicing data as required. Choosing the BRCA1 splicing events with the highest reads identifies the exons and major alternative splice sites of BRCA1. This is useful in order to identify the exon skipping events highlighted in yellow in Table 5.1. To make Tables 5.1A and B we chose all splicing events involving the 3'ss 41219713 (Table 5.1A) or the 5'ss 41209068 (Table 5.1B).

Snaptron has four different RNA sequencing database that can be analysed. SRAv1 (hg19) and SRAv2 (hg38) are from the sequencing read archive at NCBI and contain 41 and 83M splice junctions

identified by sequencing, respectively. There are also two smaller database TCGA (hg38) and GTEx (hg38) with 37 and 29M junctions (Wilks et al., 2018). SRA – sequence read archive, GTEx – Genotype-Tissue Expression is the gene expression, TCGA – cancer genome atlas of DNA and RNA sequence data of relevance to cancer.

Statistical analysis. Probability values for Table 5.5 were obtained by binomial distribution analysis using the information that the average number of bss for each of the 199 different genes listed in DBASS5 and analysed in Table 5.5 is 5.8 and similarly the average number of bss for each of the 99 genes from DBASS3 is 6.56 (D et al., 2022). The chance of a css matching a top bss by chance is therefore 1/5.9 for DBASS5 and 1/6.56 for DBASS3. Use of a binomial probability calculator (https://www.anesi.com/binomial.htm) shows that the observation that 150 out of 237 5'css match the top bss is $p = 1 \times 10^{-56}$ by chance and similarly $p = 3.2 \times 10^{-23}$ for the observed match of 62 out of 110 3'css to the top bss.

The evident difference in results between rows 1 and 5 (columns B, C) for the DBASS5 data in Table 5.6 was tested for significance by a Pearson Chi square test and a Fisher's exact t-test for the equivalent DBASS data in rows 3 and 7 of Table 5.6.

Chapter 3-IsomiR analysis and microRNA cloning bias

3.1 Introduction

Our group previously identified five human miRNAs that are expressed at different ratios of canonical to 5' isomiR forms in different tissues. The most extreme observation was that 5' isomiRs of miR-215-5p were expressed at 10- and 100-fold levels in kidney and liver compared to the canonical miR-215-5p (Tan et al., 2014), Table 1. 2. This raises the possibility that the observed differences between the ratios of canonical to isomiRs in different tissues are biologically relevant and that these changes might purposefully change mRNA targeting. As a first step towards testing this hypothesis, we wanted to use database to identify cell lines that show good differences in canonical: isomiR ratios and to confirm that these differences occur in our same cell lines. Such cell lines could then be used experimentally to identify specific mRNA targets of the isomiRs.

There is also an important technical issue that we want to address. Zhang et al (2013) report that the efficiency of cloning miRNAs using the Illumina kit method can vary by many orders of magnitude according to the nature of the miRNA end. This obviously affects conclusions about the relative expression of isomiRs based on sequencing data alone. The authors suggest a simple solution, which is to include PEG8000 in the ligation mix and in addition to add two random bases to the ends of the 5' and 3' adapters that are normally used for cloning and are ligated to each end of a miRNA. However, the authors use an older and more demanding cloning method (involving a gel purification step) compared to the widely used Illumina kit method. A big advantage of the Illumina kit method is that it allows all of the cloning steps to occur sequentially in the same eppendorf tube, which in turn allows libraries to be made from very small starting amounts of total RNA.

Aims

1. To find cell lines with strong differences in isomiR expression.

2. To investigate a reported bias problem with miRNA cloning and to incorporate a possible solution into the commonly used illumina method of miRNA cloning and sequencing.

3.2 Results

3.2.1 Cell lines show isomiR switching

Our group previously identified five human miRNAs that are expressed at different ratios of canonical to 5' isomiR forms in a tissue specific manner (Tan et al., 2014). I screened miRGator for human cell lines that express these miRNAs. The most convincing canonical to isomiR differences were obtained for hsa-miR-101-1-3p (Table 3.1).

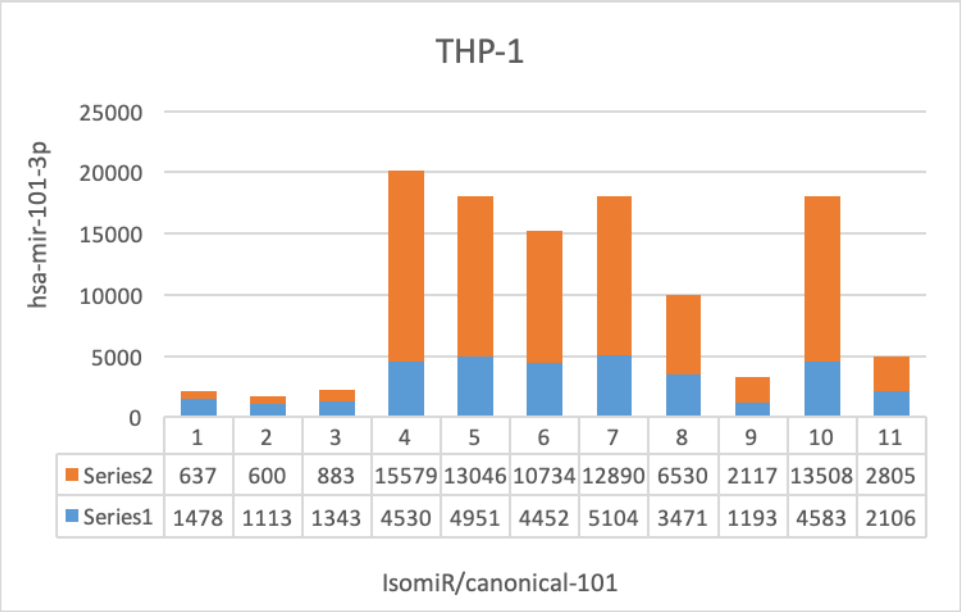
miRGator	**Sample ID**	**Canonical miRNA**	**IsomiR**			**hsa-let-7a-1-5p**
hsa-miR-101-1-3p		GUACAGUACU…	UACAGUACU…	Ratio	Samples	QC
Cell lines		normalised read counts	normalised read counts			
U2OS	s000574	1198	2493	2.08	1	0.99
SW480	s000578	785	2881	3.67	1	0.99
HL60	s000467	23014	5854	0.25	1	0.99
H929	s000428	2155	944	0.44	1	0.99
THP-1	s000447	1478	637	0.43	1	0.99
THP-1	s000448	1113	600	0.54	1	0.99
THP-1	s000449	1343	883	0.66	1	0.99
THP-1	s000638	4530	15579	3.44	1	0.96

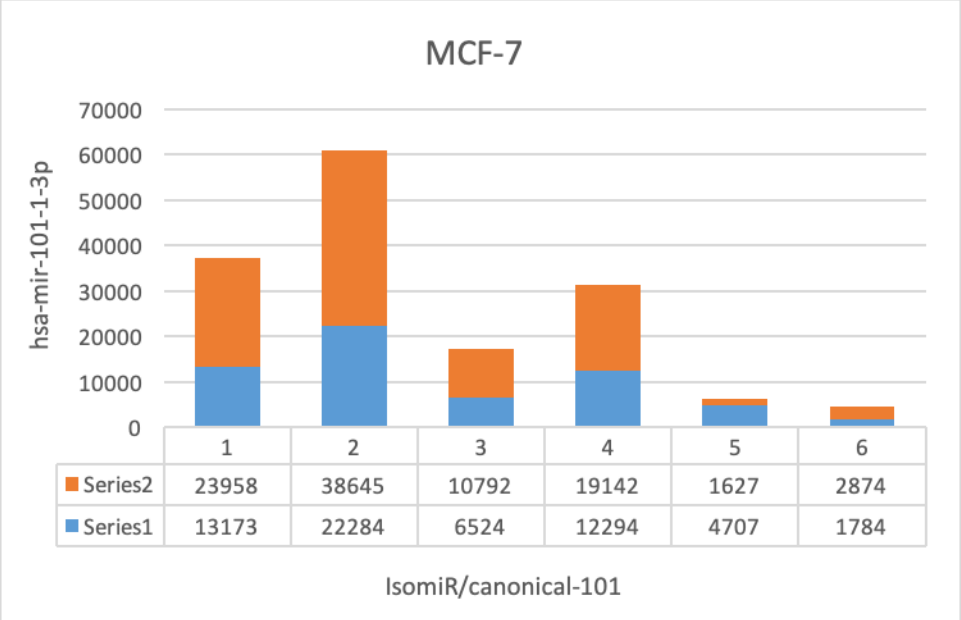
THP-1	s000639	4951	13046	2.64	1	0.96
THP-1	s000640	4452	10734	2.41	1	0.96
THP-1	s000641	5104	12890	2.53	1	0.96
THP-1	s000642	3471	6530	1.88	1	0.98
THP-1	s000643	1193	2117	1.77	1	0.98
THP-1	s000644	4583	13508	2.95	1	0.98
THP-1	s000645	2106	2805	1.33	1	0.98
Total		34324	79329	2.31	Total 11 samples	
MCF-7	s000049	13173	23958	1.82	1	0.96
MCF-7	s000050	22284	38645	1.73	1	0.95
MCF-7	s000051	6524	10792	1.65	1	0.96
MCF-7	s000052	12294	19142	1.56	1	0.95
MCF-7	s000572	4707	1627	0.35	1	0.99
MCF-7	s000580	1784	2874	1.61	1	0.99
Total		60766	97038	1.6	Total 6 samples	
WI-38	s000600	8669	2953	0.34	1	0.99
WI-38	s000601	511	196	0.38	1	0.99
WI-38	s000602	182	55	0.3	1	0.99
Total		9362	3204	0.34	Total 3 samples	
IMR90	s000120	5922	1997	0.34	1	0.99
IMR90	s000121	4164	1300	0.31	1	0.99
Total		10086	3297	0.33	Total 2 samples	

Table 3.1 Sequencing reads for the canonical miR-101-1-3p and a 5' isomiR from different cell lines in miRGator. Columns 1 and 2 list the cell lines and sample ID as listed in miRGator. Columns 3 and 4 show the normalized sequencing reads from miRGator for the canonical miR-101-1-3p and the most common indicated 5'isomiR. Column 5 shows the ratio obtained by dividing the reads in column 4 by the reads in column 3. Column 6 shows the number of samples for the sequencing reads and gives an average for repeat samples. Column 7 is a quality control for each sample that was made by dividing the reads for the full or complete sequence of hsa-let-7a-1 (this is normally invariant and appears to have no isomiRs) by the total reads for this miRNA (total reads will also include incomplete sequence reads of hsa-7a-1 due to poor sequencing).

Table 3.1 shows that as expected, cell lines WI-38, IMR90, HL60 and H929 expressed more of the canonical miR-101-1-3p than the isomiR. However, for cell lines THP-1, U2OS, SW480 and MCF-7 the opposite is observed. For THP-1 eight of eleven samples had more reads for the isomiR and for MCF-7 five out of six samples had more reads for the isomiR. The let-7a column is a quality control, values near to 1 show that the majority of let-7a reads are complete, which indicates that the reads for the 101 isomiR are not an artefact of incomplete sequencing.

Figure 3.1 is a graphical depiction of the canonical and isomiR sequencing reads shown in Table 3.1 for the cell lines THP-1, MCF-7 and WI-38.


THP-1
hsa-mir-101-3p
25000
20000
15000
10000
5000
0
1 2 3 4 5 6 7 8 9 10 11
Series2 637 600 883 15579 13046 10734 12890 6530 2117 13508 2805
Series1 1478 1113 1343 4530 4951 4452 5104 3471 1193 4583 2106
IsomiR/canonical-101


MCF-7
hsa-mir-101-1-3p
70000
60000
50000
40000
30000
20000
10000
0
1 2 3 4 5 6
Series2 23958 38645 10792 19142 1627 2874
Series1 13173 22284 6524 12294 4707 1784
IsomiR/canonical-101

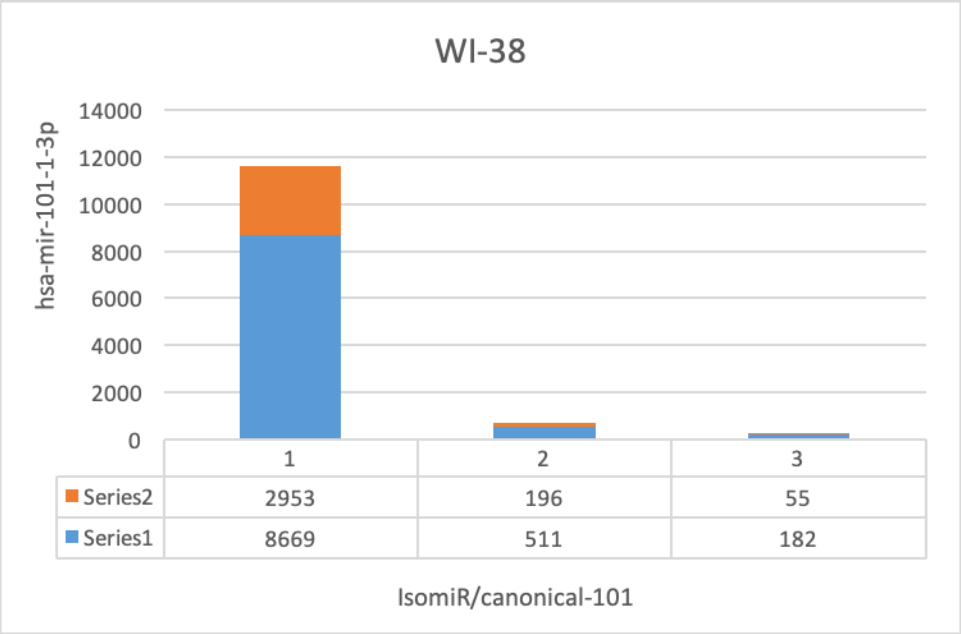


Figure 3.1 Graphical demonstration of the data of Table 3.1 for the cell lines THP-1, MCF-7 and WI-38 showing the proportion of canonical miRNA (blue) and 5'isomiR (orange) across the different samples.

We also identified a mouse cell line C2C12 that had 5 times more isomiR expression than the canonical mouse miR-101b; and another mouse cell line MIN6 that had 2 times less isomiR expression than the canonical miR-101b (Table 3.2). We also used let-7a-5p as the quality control, which had a value of 0.98, indicating that the sequences read data is reliable.

		Ratio of Isomir/Canonical				
		(UACAGUACU/GUACAGUACU)				
Cell line: MIN6	SRR933553	SRR933554	SRR933555	Cell line: C2C12	SRR835105	SRR835106
MMU-mir-101a	0.61	0.63	0.57	MMU-mir-101a	0.9	0.91
MMU-mir-101b	0.53	0.55	0.51	MMU-mir-101b	5.35	5.52
MMU-mir-101c	0	0	0	MMU-mir-101c	0	0
MMU-mir-140	2.24	1.87	1.77	MMU-mir-140	1.2	1.04
MMU-let-7a-5p (qc)	0.98	0.98	0.99	MMU-let-7a-5p (qc)	0.99	0.99

Table 3.2 Differences in the ratio of miR-101b canonical: isomiR in two mouse cell lines MIN6 (three samples) and C2C12 (two samples). The samples used to make the datasets are indicated by the SRR numbers. The mouse cell line data from the gene expression omnibus (GEO) was processed under the supervision of Dr Castellano using Linux system with mirdeep2 to make maps and then get sequencing reads.

We also used miRGator to analyse miR-140-3p and identified two potentially useful cell lines MCF7, which is a human breast cancer cell line and IMR90, which is a normal human lung fibroblast cell line. These two cell lines have different isomiR:miRNA ratios of 1.6 and 0.49 (Table 3.3).

miRGator	Canonical miRNA	IsomiR			hsa-let-7a-1-5p
hsa-miR-140-3p	UACCACAGGGU…	ACCACAGGG…	Ratio	Samples	UGAGGUAGUAGGUUGUAUAGUU
Cell lines	normalised read counts	normalised read counts			ratio
MCF7	4390	8394	1.91	1	0.96
MCF7	5656	11252	1.99	1	0.95
MCF7	2323	2497	1.07	1	0.96
MCF7	4217	5086	1.21	1	0.95
MCF7	3811	4534	1.19	1	0.99
MCF7	162	365	2.25	1	0.99
Total	**20559**	**32128**	1.6	Total 6 samples	
IMR90	30866	15701	0.51	1	0.99
IMR90	19975	9000	0.45	1	0.99
Total	**50841**	**24701**	0.49	Total 2 samples	

Table 3.3 Identifies two cell lines MCF7 and IMR90 that make the canonical miR-140-3p and the isomiRs indicated in different ratios. Column six is a quality control that was made by dividing the reads for the full or complete sequence of hsa-let-7a-1 by the total reads for this miRNA.

Table 3.4 compares the expression of canonical miR-140-3p and isomiR-140-3p in mouse tissues, the ratios do vary but not hugely between most tissues.

miRBase					
mmu-miR-140-3p	Canonical miRNA	IsomiR			mmu-let-7a-5p
Tissue	UACCACAGGGU…	ACCACAGGG…		Samples	UGAGGUAGUAGGU UGUAUAGUU
	normalised read counts	normalised read counts	Ratio		Ratio
Lung	28661	22188	0.77	1	0.99
Liver	6370	2864	0.45	1	0.99
Kidney	29031	25051	0.86	1	0.99
Pancreas	9384	11898	1.27	1	0.99
Skin	24051	31642	1.32	1	0.99
Skeletal muscle	30865	44679	1.45	1	0.99
Salivary glands	30311	44743	1.48	1	0.99

Table 3.4 The ratios of canonical miR-140-3p and isomiR are shown in different mouse tissues. Column one shows different mouse tissue types. Column two shows canonical miRNA (UACCACAGGGU) sequencing reads and column three shows IsomiR (ACCACAGGG) sequencing reads. The column four shows ratios by dividing the number of sequencing reads for the IsomiR by the number of canonical miRNA-140 reads. The column five shows sample number and column six shows let-7a-5p as the quality control, which had a value of 0.99, indicating that the sequences read data is reliable.

Tan et al (2014) reported that the liver produces 100-fold more of an isomiR of miR-215-5p than the canonical miR-215-5p and the kidney ten-fold more, this was the most dramatic example of tissue differences that they identified following a screen of 295 of the most expressed miRNAs in miRGator. I have confirmed this result for this tissue samples listed in miRGator and have also identified a

number of cell lines that produce canonical miR-215-5p, although the read numbers were rather low (Table 3.5). I have not found a cell line that produces a large proportion of isomiR-215-5p in repeated samples (Table 3.5). This might be because there are not many cell line samples listed in miRGator. In addition, there are relatively few reads for the mouse equivalent of isomiR-215-5p deposited in miRBase (data not shown).

miRGator	Sample ID	Canonical miRNA	IsomiR	canonical:isomiR ratio		hsa-let-7a-1-5p
hsa-miR-215-5p		AUGACCU…	UGACCU…			QC
cell lines		normalised read counts	normalised read counts		Samples	
SET2	s000005	42	4	10.5	1	0.98
MCF-7	s000049	15	12	1.25	1	0.96
MCF-7	s000050	28	8	3.5	1	0.95
MCF-7	s000051	9	6	1.5	1	0.96
MCF-7	s000052	11	12	0.916666667	1	0.95
MCF-7	s000572	17	20	0.85	1	0.99
MCF-7	s000580	2	8	0.25	1	0.99
IMR90	s000120	21	4	5.25	1	0.99
IMR90	s000121	23	0	>23	1	0.99
HeLa S2	s000122	5	609	0.008210181	1	0.99
HeLa S2	s000123	3	277	0.010830325	1	0.99
HeLa S2	s000124	1049	13	80.69230769	1	0.98
HeLa S2	s000125	326	6	54.33333333	1	0.97
HeLa S2	s000126	264	10	26.4	1	0.98
HeLa S2	s000127	368	5	73.6	1	0.97
HeLa S2	s000130	3	0	>3	1	0.99
HeLa S2	s000131	2	50	0.04	1	0.98
Naive39	s000408	8	2	4	1	0.99
GC136	s000409	10	1	10	1	0.99
Naive138	s000412	10	12	0.833333333	1	0.99
PC44	s000414	9	3	3	1	0.99
GCB385	s000416	22	6	3.666666667	1	0.99

GCB110	s000417	36	5	7.2	1	0.99
Ly3	s000418	42	41	1.024390244	1	0.99
BL115	s000421	9	6	1.5	1	0.99
BL134	s000422	11	17	0.647058824	1	0.99
MCL114	s000424	30	34	0.882352941	1	0.99
MCL112	s000425	12	53	0.226415094	1	0.99
U266	s000426	74	17	4.352941176	1	0.99
KMS12	s000427	67	4	16.75	1	0.99
L1236	s000429	11	7	1.571428571	1	0.99
L428	s000430	33	15	2.2	1	0.99
CLLU626	s000431	28	3	9.333333333	1	0.99
CLLM633	s000432	11	6	1.833333333	1	0.99
MALT413	s000433	22	6	3.666666667	1	0.99
ABC158	s000436	9	2	4.5	1	0.99
5-8F	s000445	85	2	42.5	1	0.99
5-8F	s000446	45	9	5	1	0.99
THP-1	s000447	10	0	#DIV/0!	1	0.99
THP-1	s000448	8	1	8	1	0.99
THP-1	s000449	13	0	>13	1	0.99
THP-1	s000638	85	0	>85	1	0.96
THP-1	s000639	85	0	>85	1	0.96
THP-1	s000640	63	3	21	1	0.96
THP-1	s000641	76	2	38	1	0.96
THP-1	s000642	62	3	20.66666667	1	0.98
THP-1	s000643	18	0	>18	1	0.98
THP-1	s000644	113	2	56.5	1	0.98
THP-1	s000645	35	2	17.5	1	0.98
K562	s000466	40	11	3.636363636	1	0.99
AG01522	s000571	21	1	21	1	0.92
A549	s000576	629	223	2.820627803	1	0.99
DLD2	s000579	23	6	3.833333333	1	0.99
MB-MDA231	s000581	9	13	0.692307692	1	0.99
HEK293	s000530	11	3	3.666666667	1	0.99
HEK293	s000531	2	0	>2	1	0.99
HEK293	s000532	14	1	14	1	0.99
HEK293	s000533	12	3	4	1	0.99
HEK293	s000534	28	13	2.153846154	1	0.98
HEK293	s000535	16	10	1.6	1	0.99

HEK293	s000536	154	21	7.333333333	1	0.98
HEK293	s000537	39	7	5.571428571	1	0.99
HEK293	s000538	78	1	78	1	0.99
HEK293	s000539	84	0	>84	1	0.99
HEK293	s000540	71	1	71	1	0.99
HEK293	s000541	15	0	>15	1	0.99
HEK293	s000542	102	1	102	1	0.99
HEK293	s000543	139	10	13.9	1	0.99
HeLa	s000573	3	1	3	1	0.99
HeLa	s000595	0	0	>0	1	0.99
HeLa	s000611	234	11	21.27272727	1	0.95
HeLa	s000612	1167	13	89.76923077	1	0.96
HeLa	s000613	1576	40	39.4	1	0.96
HeLa	s000631	2	0	>2	1	0.35
HeLa	s000632	1	0	>1	1	0.4
HeLa	s000633	2	0	>2	1	0.32
HeLa	s000634	1	0	>1	1	0.27
HeLa	s000656	8	1	8	1	0.96
WI-38	s000600	77	5	15.4	1	0.99
WI-38	s000601	1	0	>1	1	0.99
WI-38	s000602	2	0	>2	1	0.99

Table 3.5 Sequencing reads for the canonical *miR-215-5p* and a 5' isomiR from different cell lines in miRGator. Columns as described for Table 3.1.

3.2.2 Trying to improve the method for miRNA cloning

Zhang et al., (2013) reported that different miRNAs vary in their cloning efficiency by as much as 1000 fold using a standard Illumina kit. They identified mir-205 and miR-214 as poorly cloned miRNAs. We therefore looked at these and miR-101-1-3p. Inefficient cloning can occur at both or either of the RA3 or RA5 ligations (steps 2 and 4 of Fig 3.2).

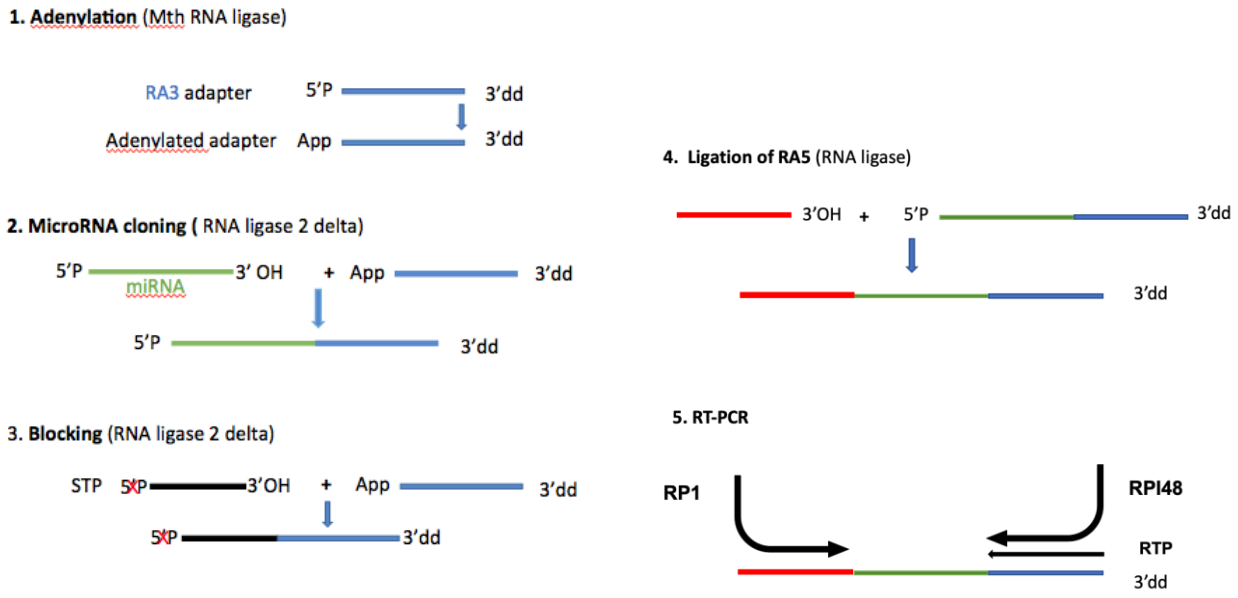


Figure 3.2 Illumina protocol for making a miRNA library. RA3 ligation is catalysed by a modified RNA ligase 2 (RNA ligase 2 delta), which requires a 5' adenylated substrate, in this case adenylated RA3. It is important to use the modified RNA ligase 2 delta as it can only make ligation products with adenylated RA3. RA3 cannot self-ligate as it is blocked at its 3' end with a dideoxy modification (dd).

Table 3.6 List of oligos used in these experiments.

RA3	5'PO4TGGAATTCTCGGGTGCCAAGG-dideoxyC3
NNRA3	5'PO4NNTGGAATTCTCGGGTGCCAAGG-dideoxyC3
RA5	5'GUUCAGAGUUCUACAGUCCGACGAUC-3'
RA5NN	5'GUUCAGAGUUCUACAGUCCGACGAUCNN-3'
RA5NNNN	5'GUUCAGAGUUCUACAGUCCGACGAUCNNNN-3'

miR-101-1-3p	5’PO4GUACAGUACUGUGAUAACUGAAG-3’
miR-205	5’PO4UCCUUCAUUCCACCGGAGUCUG-3’
miR-214	5’PO4ACAGCAGGCACAGACAGGCAGUC-3’
STP	5’GAAUUCCACCACGUUCCCGUGG-3’
RTP	5’GCCTTGGCACCCGAGAATTCCA-3’
RP1	5’AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA-3’
RPI48	5’CAAGCAGAAGACGGCATACGAGATTGCCGAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA-3’

RA3 was adenylated using the enzyme Mth RNA ligase (Biolabs) and successful adenylation was assayed by a slight change in mobility (Fig 3.3).

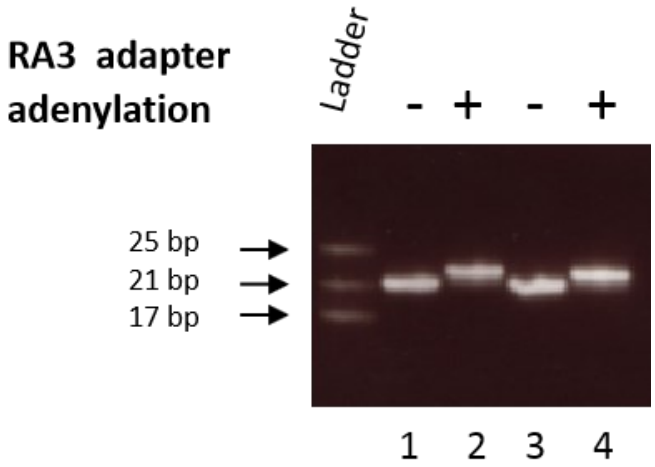


Figure 3.3 RA3 was adenylated using the enzyme Mth RNA ligase. The minus indicates without Mth RNA enzyme, and the plus indicates with Mth RNA enzyme. Samples were run on a 15% acrylamide gel.

We can assay RA3 ligation in vitro using P^{32} labelled substrates (Figure 3.4A. B) or by SybrGold staining (see remaining figures).

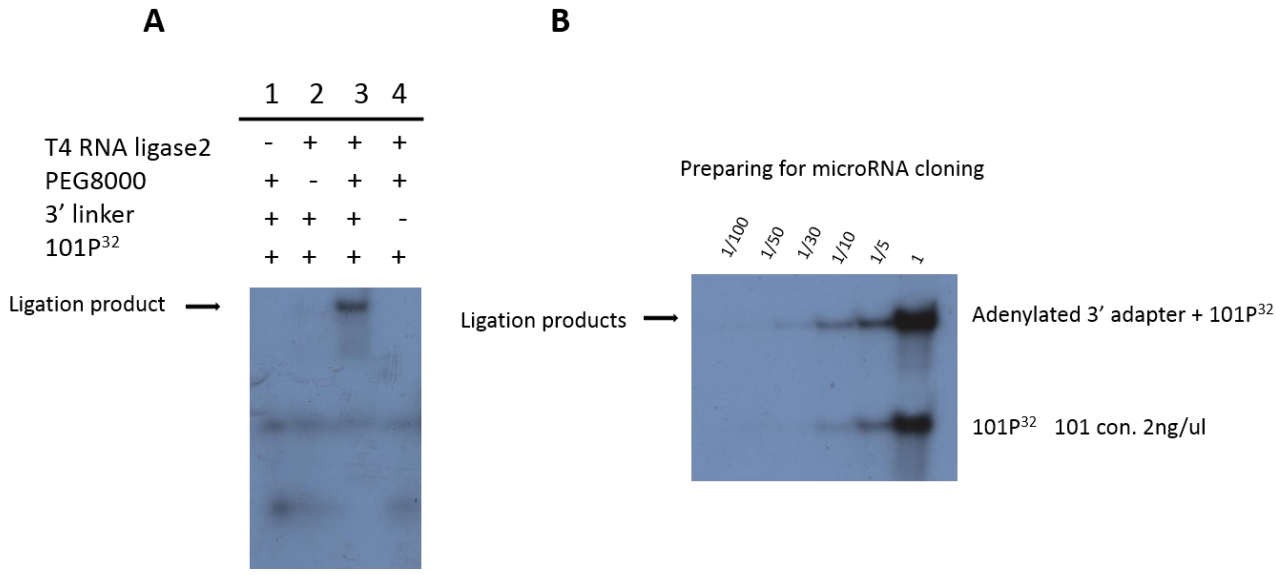


Figure 3.4 Adenylated RA3 adapter ligated to P^{32} labelled miR-101-1-3p oligo under the indicated conditions. A. 50 ng of adenylated RA3 adapter ligated to P^{32} labelled miR-101-1-3p oligo under the indicated conditions at room temperature for 90 minutes. The arrow indicates 3' adapter ligation product. B. 50 ng of adenylated RA3 adapter ligated to 2 ng of 101 oligo labelled with P^{32} (last lane) that was then diluted as indicated prior to ligation to 50 ng of RA3.

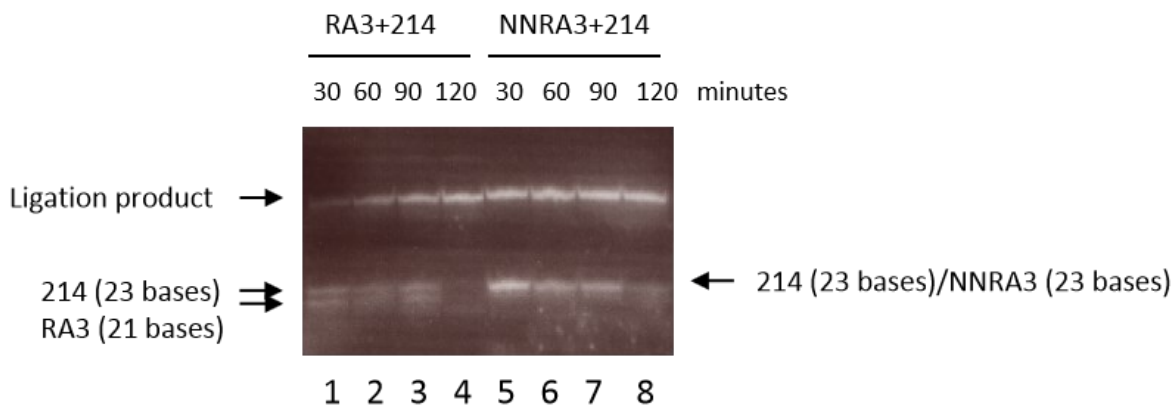
Figure 3.4A (compare lanes 2 and 3) illustrates the importance of including PEG8000 in the ligation buffer (this was included in all further experiments). Figure 3.4B is a titration and the last lane shows that there is a lot of unligated miR-101-1-3p (bottom band) even though there was a 25 fold excess

of the 3’ linker RA3 in the ligation incubation. Diluting the amount of miR-101-1-3p did not seem to markedly improve its ligation (lanes 1 to 5).

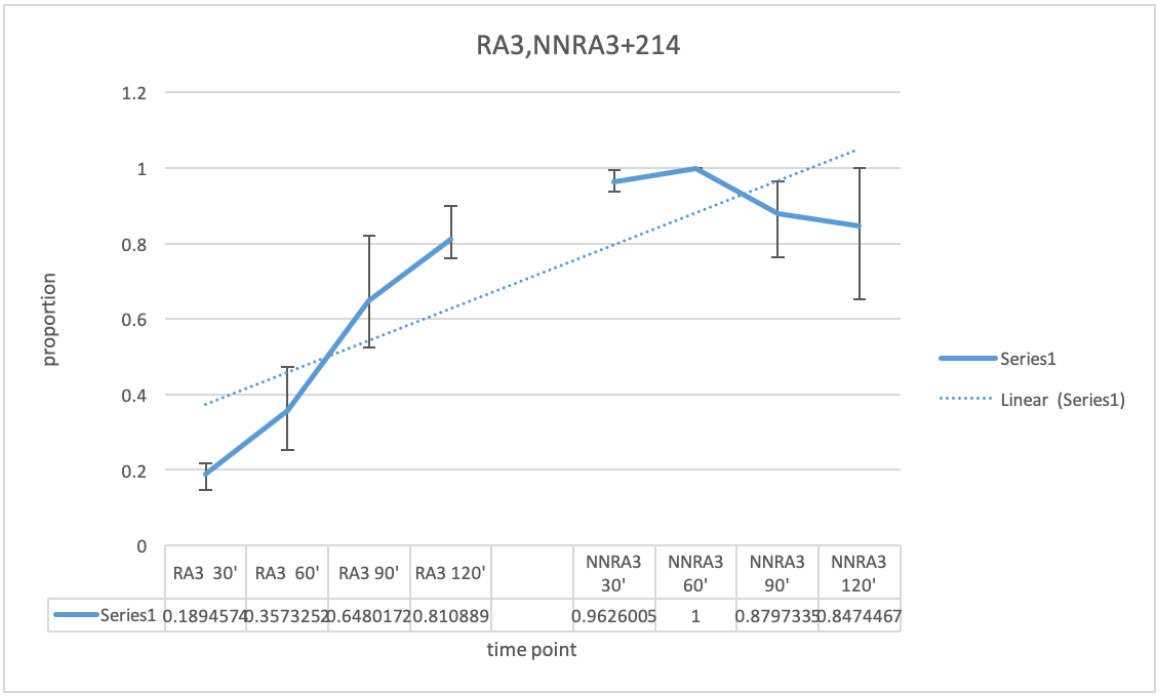
3.2.3 Optimization of conditions for the cloning of difficult microRNAs to RA3

A comparison of Figure 3.5 A, C and E confirms that miR-101-1-3p was ligated relatively inefficiently to the 3’ adapter RA3 compared to mir-205 and 214. The addition of two variable bases to the end of RA3 improved the cloning efficiency for mir-205 and 214, however, most of miR-101-1-3p was not ligated even to NNRA3 (Fig 3.5 C). The large error bars for Fig 3.5 D are reflective of the poor and variable ligation of miR-101-1-3p to RA3.

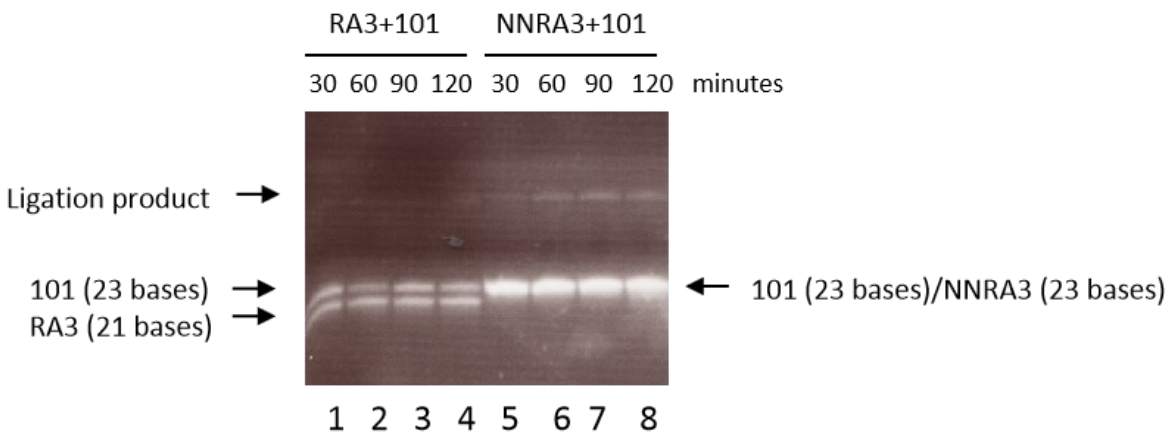
A



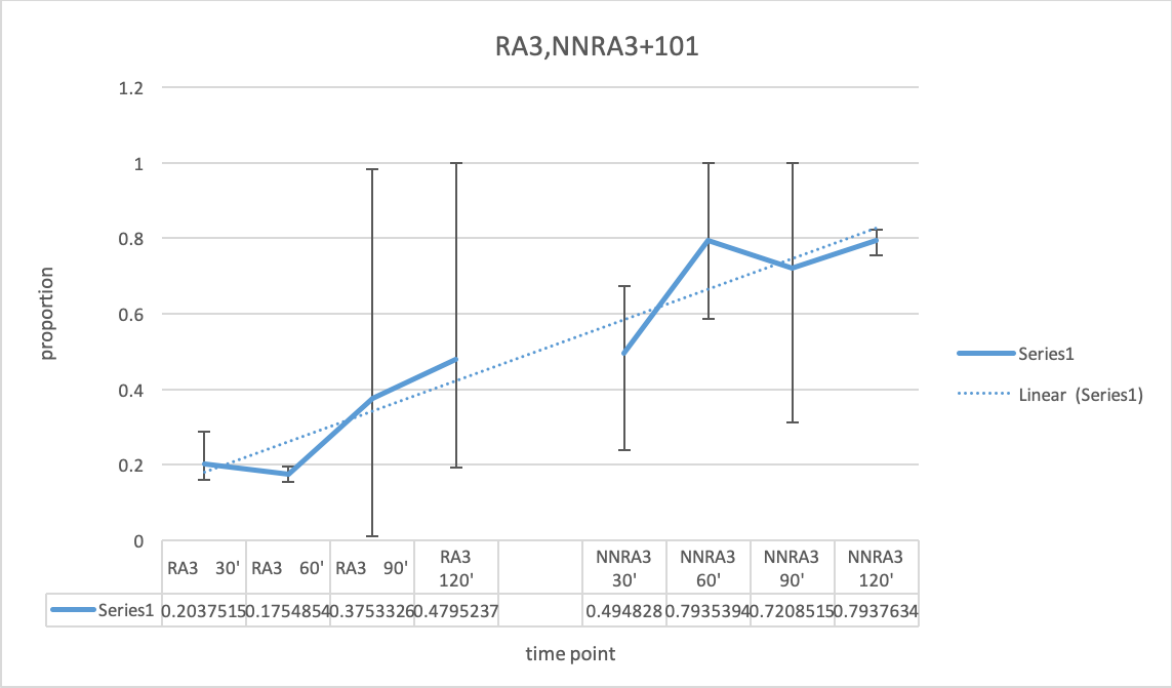
B


RA3,NNRA3+214
1.2
1
0.8
0.6
0.4
0.2
0
proportion
Series1
Linear (Series1)
RA3 30'
RA3 60'
RA3 90'
RA3 120'
NNRA3 30'
NNRA3 60'
NNRA3 90'
NNRA3 120'
Series1 0.1894574 0.3573252 0.6480172 0.810889 0.9626005 1 0.8797335 0.8474467
time point

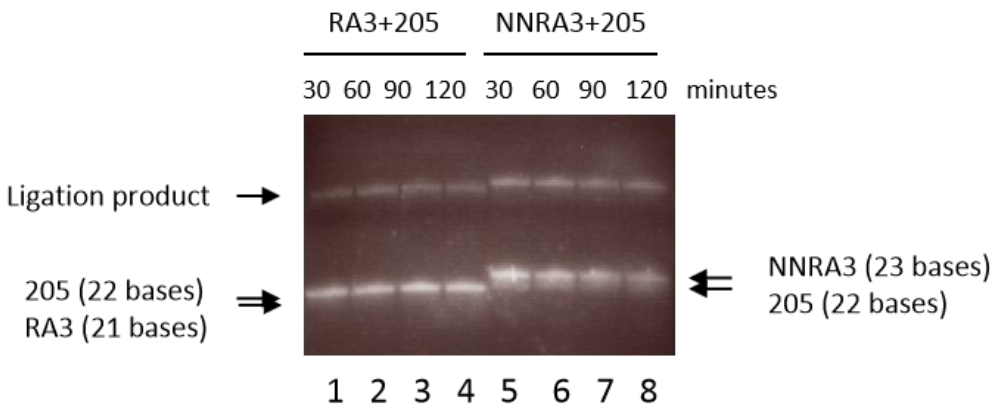
C


RA3+101
NNRA3+101
30 60 90 120 30 60 90 120 minutes
Ligation product
101 (23 bases)
RA3 (21 bases)
101 (23 bases)/NNRA3 (23 bases)
1 2 3 4 5 6 7 8

D


RA3,NNRA3+101
1.2
1
0.8
0.6
0.4
0.2
0
proportion
RA3 30'
RA3 60'
RA3 90'
RA3 120'
NNRA3 30'
NNRA3 60'
NNRA3 90'
NNRA3 120'
Series1
0.2037515
0.1754854
0.3753326
0.4795237
0.494828
0.7935394
0.7208515
0.7937634
time point
Series1
Linear (Series1)

E


RA3+205
NNRA3+205
30 60 90 120 30 60 90 120 minutes
Ligation product
205 (22 bases)
RA3 (21 bases)
NNRA3 (23 bases)
205 (22 bases)
1 2 3 4 5 6 7 8

F

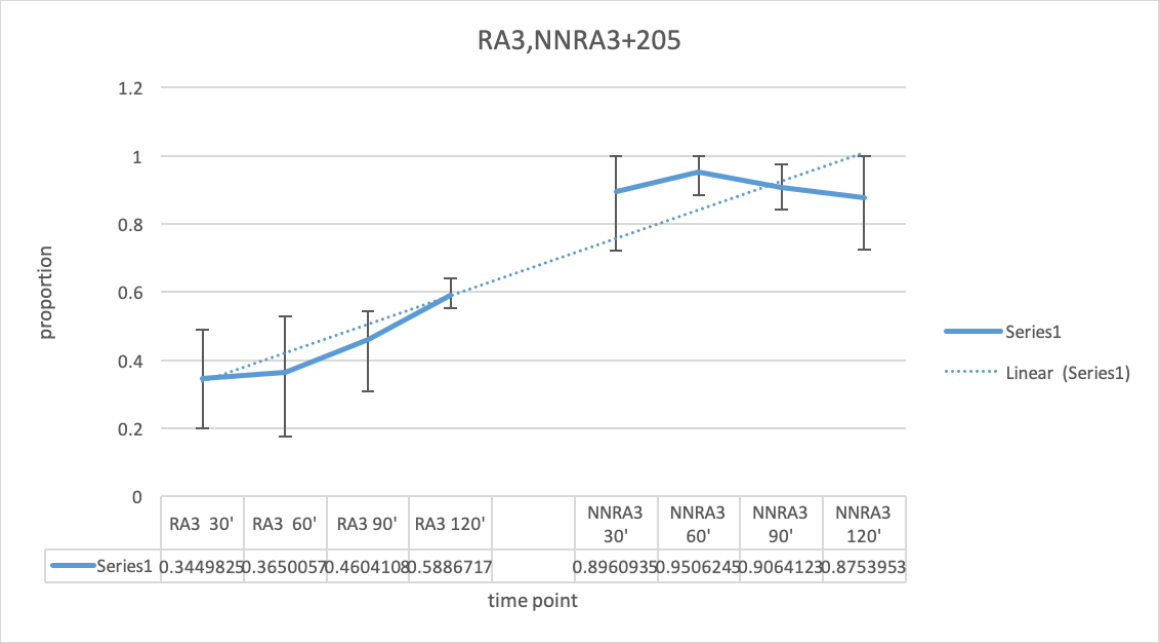


Figure 3.5 Ligation of RA3 or NNRA3 adapters to mir-214, miR-101-1-3p or mir-205 under the indicated conditions. 40 ng of RA3 and NNRA3 adapters were ligated to 40 ng of mir-214 (A, B), miR-101-1-3p (C, D) and mir-205 (E, F) for 30, 60, 90 and 120 minutes at room temperature. The ligation mixes were run on a 15% PAGE-urea gel and stained with SybrGold. Each experiment was repeated three times and averages were calculated by using image J with the most intense ligation product bands designated as 1. The error bars in B, D and F show the standard deviation of the data.

Figure 3.6 shows that ligation of miR-101-1-3p to NNRA3 but not RA3 could be improved by ligation for 4 hours at 16°C or alternatively at 4°C or room temperature overnight.

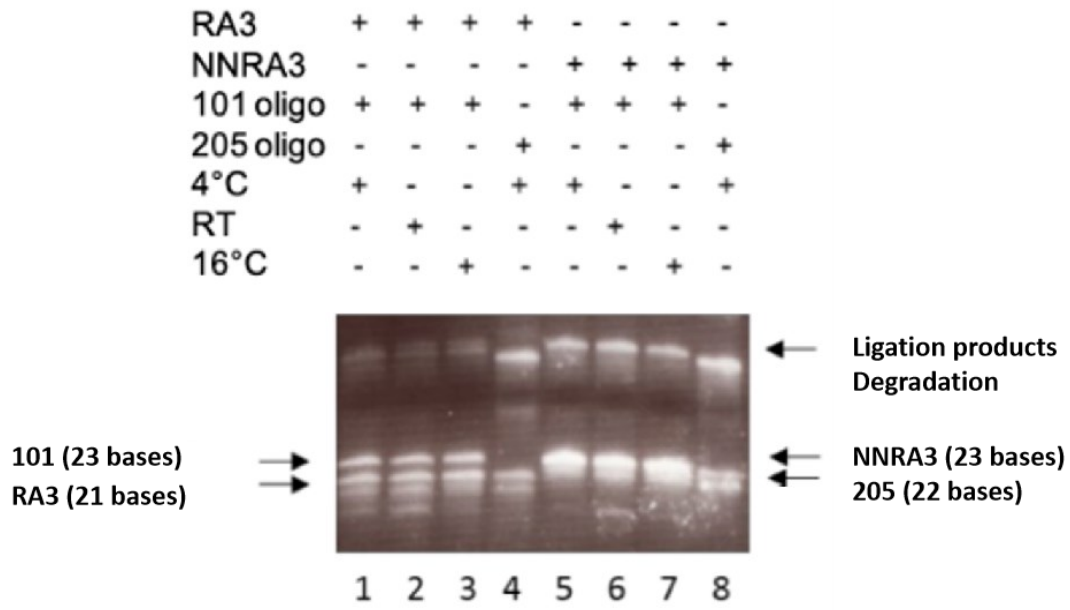


Figure 3.6 Ligation of miR-101-1-3p to NNRA3 but not RA3 could be improved by different temperature and time incubation. Ligation of 50 ng of miR-101-1-3p to 100 ng of NNRA3 but not RA3 could be improved by ligation for 4 hours at 16°C or alternatively at 4°C or room temperature overnight. 100 ng of RA3 and NNRA3 were ligated with 50 ng of 205 oligo at 4°C overnight, as indicated. Lanes 4 and 8 show ligations with mir-205 (lane 8), as a positive control.

3.2.4 Design of the STP oligo

The subsequent step in the Illumina protocol is to add a stop solution that prevents unused RA3 from ligating to RA5. The stop solution contains an oligo called STP that is ligated to RA3 by RNA ligase 2 delta and so mops up excess RA3. It isn't possible for the STP oligo to ligate to RA5 because the 5' end of STP is not phosphorylated (Table 3.6). STP ligation is an important step because it circumvents the need for gel purification in order to remove excess RA3 and was not included by Zhang et al., (2013). The STP oligo is designed to enhance its ligation to the 5' end of unused RA3 through complementarity (Fig 3.7).

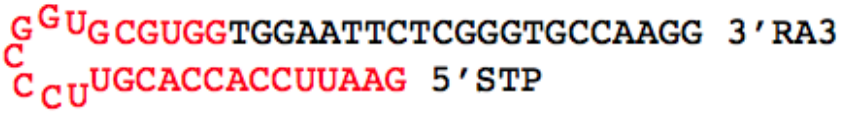
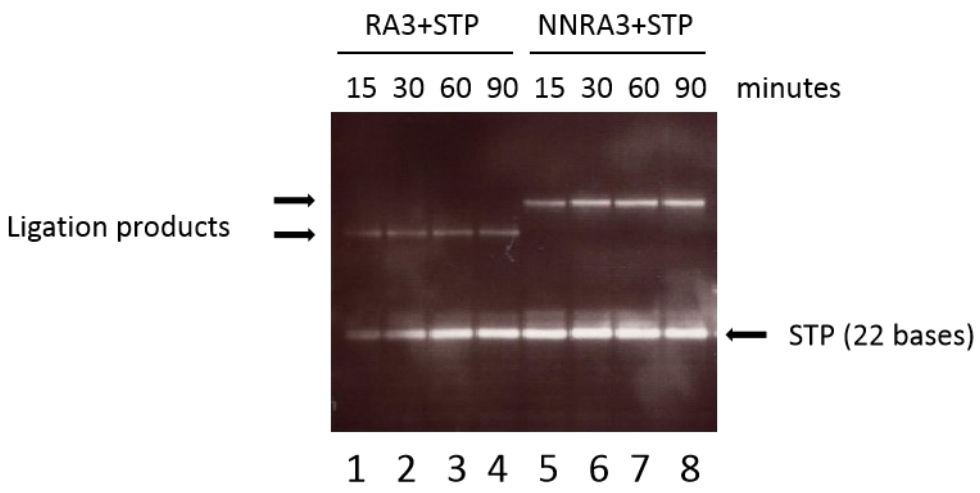


Figure 3.7 Complementarity pairing of STP designed to enhance its ligation to RA3. Black – RA3 sequence, red – STP oligo sequence from Illumina.

Because we want to use NNRA3 (NNTGGAATTC……) rather than RA3 (TGGAATTC……) we therefore designed a number of STP variants that could in theory accommodate the random bases at the 5' end of NNRA3. However, Figure 3.8 indicates that we could not detect any decrease in the efficiency of ligation of NNRA3 to the STP oligo provided by Illumina, in fact it ligated more efficiently. Similarly, we could not detect any increase in the efficiency of ligation of NNRA3 to modified STP oligos that are complementary to the random ends of NNRA3 (data not shown). We therefore conclude that at least part of the self-complementary features of STP responsible for the stem loop is perhaps not an essential feature. Consequently, the important STP step in the Illumina protocol can be easily adapted to NNRA3.

A


RA3+STP
NNRA3+STP
15 30 60 90 15 30 60 90 minutes
Ligation products
STP (22 bases)
1 2 3 4 5 6 7 8

B

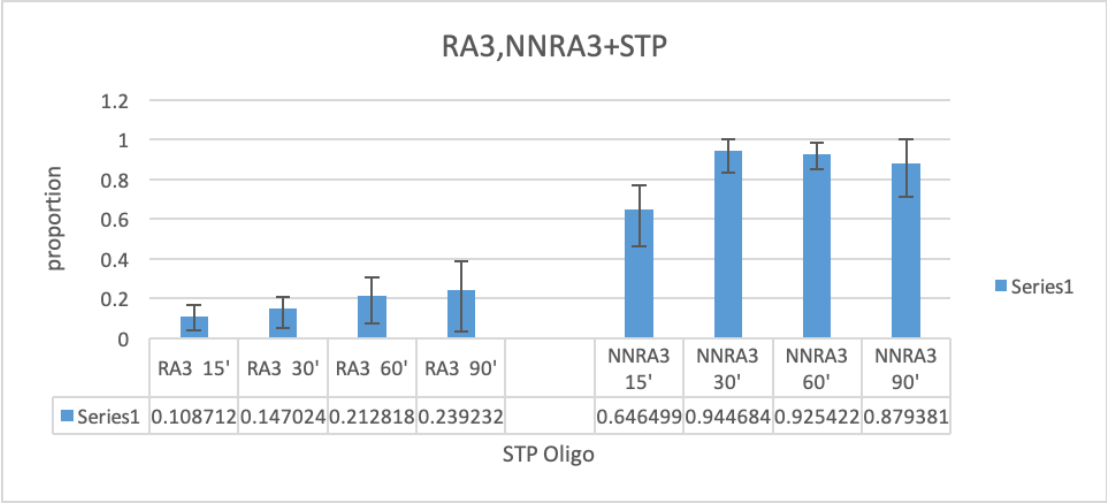
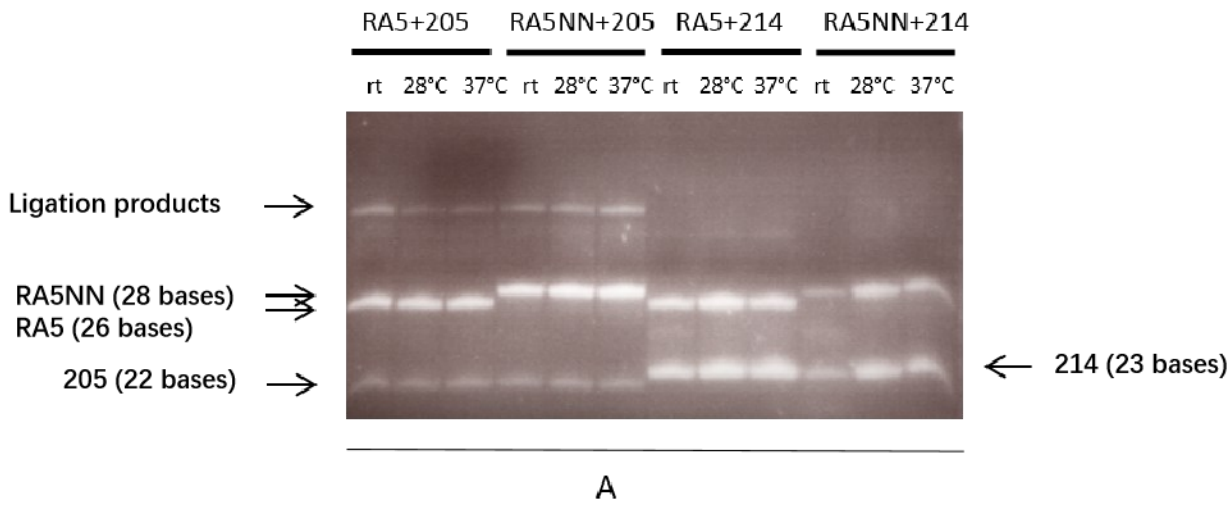

RA3,NNRA3+STP
proportion
1.2
1
0.8
0.6
0.4
0.2
0
RA3 15'
RA3 30'
RA3 60'
RA3 90'
NNRA3 15'
NNRA3 30'
NNRA3 60'
NNRA3 90'
Series1
0.108712
0.147024
0.212818
0.239232
0.646499
0.944684
0.925422
0.879381
STP Oligo
Series1

Figure 3.8 Ligation of STP to RA3 or NNRA3. A. The 3' adapters RA3 (50 ng) and NNRA3 (50 ng) were ligated to stop oligo (250 ng) for the indicated periods of time. B. Each experiment was repeated three times and averages were calculated by using image J with the most intense ligation product bands designated as 1. The error bars in B show the standard deviation of the data.

3.2.5 RA5 ligation to miRNAs

The next step is to ligate RA5 to the 5' ends of miRNAs previously ligated to RA3 or to NNRA3. However, we first studied the ligation of just RA5 to different miRNAs in vitro. Figure 3.9A shows that mir-214 did not ligate to RA5 or RA5NN at any of the tested temperatures. miR-205 did ligate but not efficiently over 2 hours incubation as there was unligated miR-205 at the bottom of the gel.



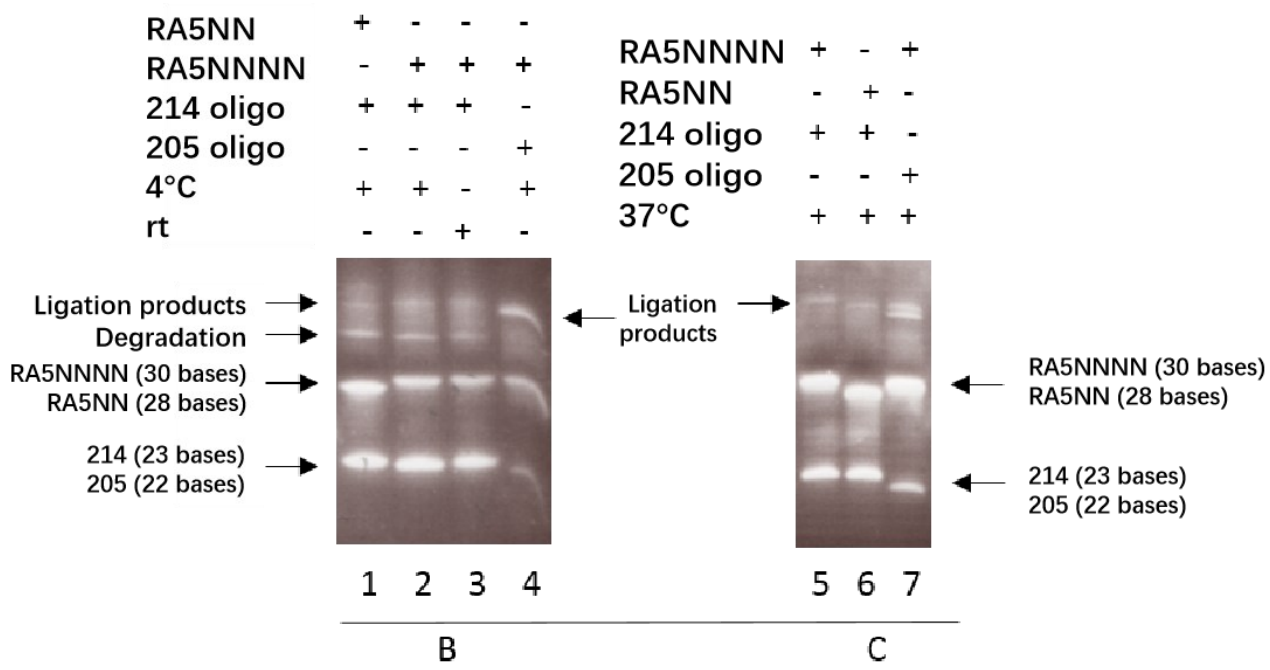
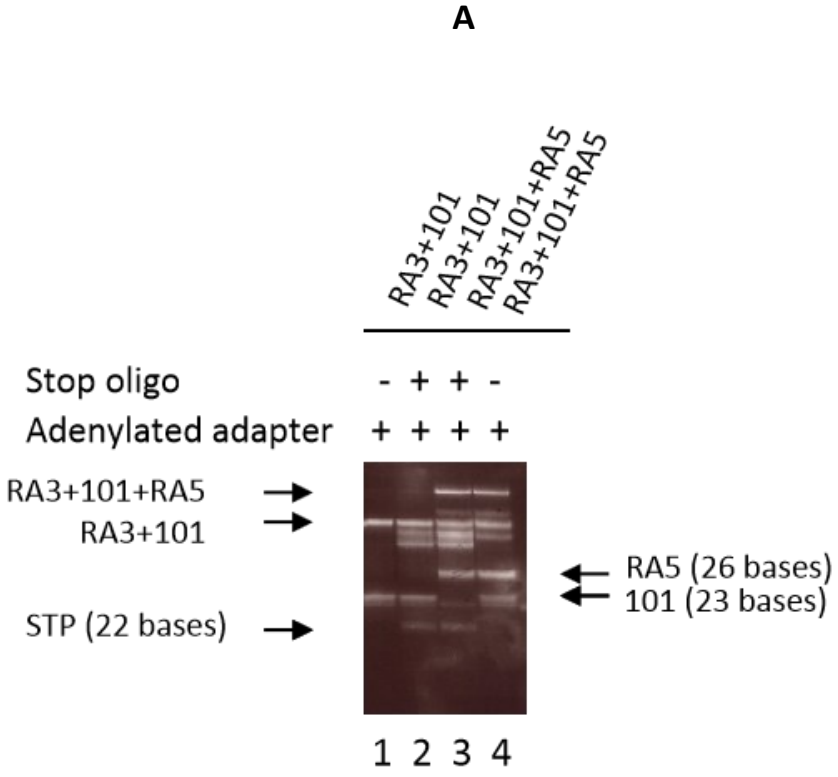


Figure 3.9 Ligation of RA5NN or RA5NNNN to mir-214 or mir-205 under the indicated conditions. A. Ligation of 80 ng of RA5 or RA5NN to 80 ng of miR-205 or 214 at the indicated temperatures for 2 hours. B & C. 100 ng of RA5NN or RA5NNNN was ligated with 100 ng of mir-214 or 205 at the indicated temperatures. Ligations at 4°C or room temperature (rt) were overnight, whereas ligations at 37°C were for 4 hours and in the presence of 20% DMSO.

We were unable to improve ligation of miR-214 substantially by varying temperature, length of incubation, addition of DMSO or used of an RA5 adapter with 4 variable bases at its 3' end (R Yi, personal communication). Although there was an improvement in the ligation of miR-205 (Fig 3.9B, C).

As expected, Figure 3.10 shows that the sequential ligation of miR-214 to NNRA3 and then to RA5NN was less efficient compared to miR-205 or 101.



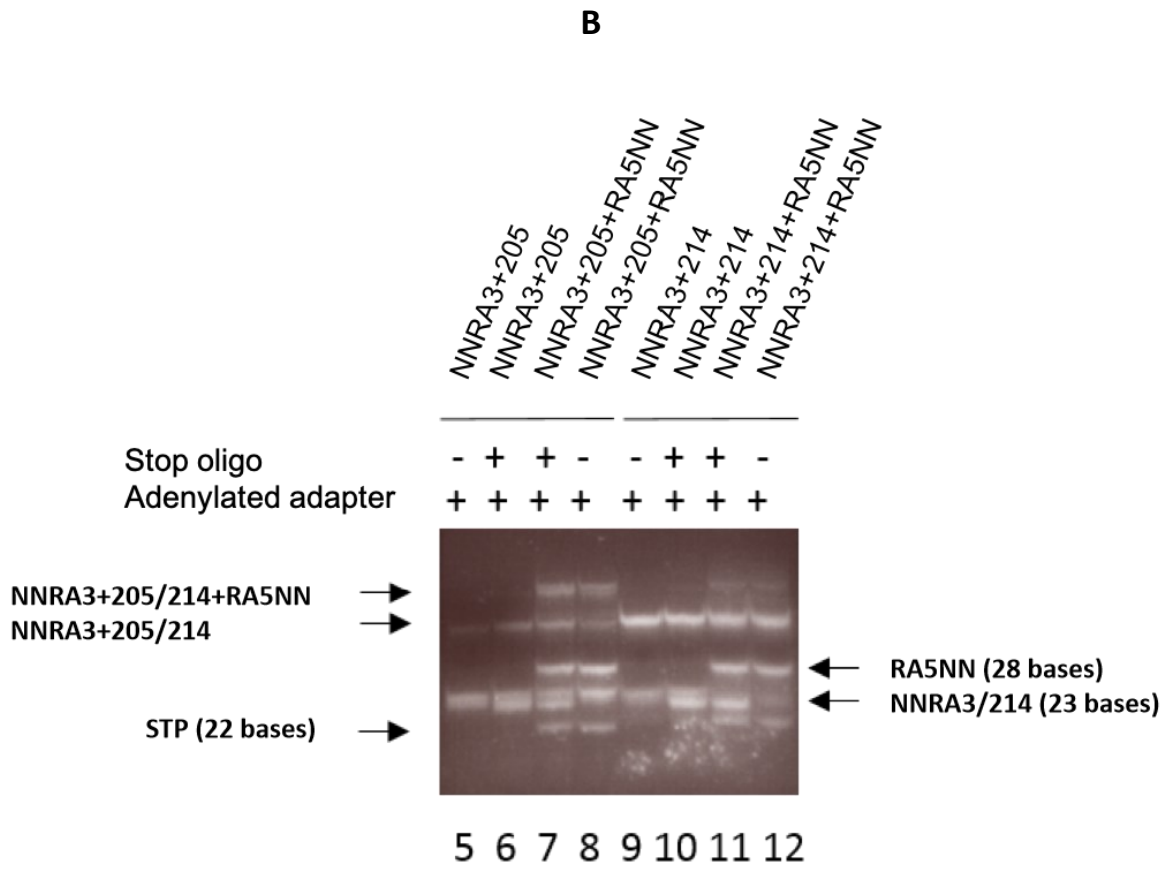


Figure 3.10 Sequential ligation of miR-101-1-3p, mir-205 or mir-214 to 3' and 5' adapters with or without the STP oligo. A. miR-101-1-3p ligated to RA3 (lanes 1 to 4) and then RA5 (lanes 3,4). B. mir-205 ligated to NNRA3 (lanes 5 to 8) and then RA5NN (lanes 7,8). B (lanes 9 to 12) mir-214 ligated to NNRA3 but was not ligated to RA5NN (lanes 11,12). Ligations used 50 ng of each miRNA oligo, 100 ng of RA3 or NNRA3, STP oligo and 50 ng of RA5 or RA5NN. The NNRA3 ligations were at 4°C overnight and the RA5NN ligations were at 37°C for 4 hours. As indicated for half of the incubations STP oligo was added after the 3' adapter ligation for 1 hour at 25°C.

3.3 Discussion

We identified cell lines that differed in their relative expression levels of miR-101-1-3p and miR-140-3p isomiRs and where the isomiRs showed greater expression than the canonical miRNA in at least one cell line (Tables 3.1 to 3.4). However, none of the cell lines showed isomiR expression that was ten to one hundred fold greater than canonical expression, as previously reported for a 5' isomiR of hsa-miR-215-5p in kidney and liver tissue (Tan et al., 2014).

In principle the cell lines described above could be used to test the effect of inhibition or overexpression of canonical: isomiR pairs of miR-101-1-3p or 140 on mRNA targeting. Tan et al (2014) constructed sponge vectors with repeated binding sites that were identified as specific target sites of either the canonical miRNA or an isomiR. Sponge vectors will compete with endogenous targets to bind to miRNA or isomiRs and an expected effect of this is to increase steady-state levels of candidate mRNA targets, which could be detected by RNA sequencing. Overexpression of microRNAs and isomiRs can be achieved by transfection into cell lines and possible effects upon mRNA levels could be determined by mRNA sequencing and further investigated by western blotting.

The most interesting candidate targets identified by the above methods could be further investigated. This will depend on the messenger RNA targets we identify and whether the targets generate or raise hypotheses about their function with respect to the specific biology of the cell line. For example, understanding the *in vivo* target of isomiR-215-5p might explain why the molecule is significantly increased by 100 times in the liver (see Table 1.2). We can use bioinformatics resources such as DAVID (annotation, visualization and integrated discovery database) (http://david.abcc.ncifcrf.gov/), which provides annotation tools to help understand the biological significance behind large lists of genes.

Other groups have confirmed that the relative expression of isomiRs varies between tissue types (Tomasello et al., 2021, Panzade et al., 2022, Smith and Hutvagner, 2022). The most common isomiRs have 3' changes but there are many reports of 5' isomiRs that are at least equally expressed compared to the canonical miRNA in certain tissues (Panzade et al., 2022, Smith and Hutvagner, 2022). The 5' isomiRs have been shown to have additional mRNA targets compared to the canonical miRNA (Tan et al., 2014), as might be expected (Lewis et al., 2003).

Tomasello et al (2021) have recently reviewed the literature that supports a functional role for isomiRs in repressing new and different targets. There are relatively few such papers but as discussed later (Chapter 6) it is not easy to prove a novel functional role for an isomiR. So far, there is evidence that the additional targeting of some 5' isomiRs contributes to cancer progression (Zelli et al., 2021, Li et al., 2022). An isomiR of miR-411 has been found to be upregulated in patients with ischemia and in fibroblasts under ischemic conditions and becomes five times more abundant than canonical miR-411. The isomiR but not the canonical mir-411 has an inhibitory effect upon cell migration indicating that these miRNAs may have distinct roles in angiogenesis (van der Kwast et al., 2020).

In agreement with Zhang et al., (2013) we found that ligation of the 3' adapter to some miRNAs was inefficient but could be greatly improved by adding variable bases to the end of the adapter and by extending the ligation time and including PEG 8000 (Figures 3.4, 3.4). We were also able to introduce the important STP step used by Illumina into the improved miRNA cloning technique developed by (Zhang et al., 2013), see Fig 3.8. We were unable to greatly improve the efficiency of ligation of certain miRNAs to the 5' adapter, despite adding variable bases to its 3' end, as suggested by (Zhang et al., 2013) and R Yi (personal communication), see Figure 3.9. Zhang et al., (2013) used a 50 fold excess of RA5NN to miRNA in their cloning experiments, whereas we used more stringent conditions of relatively equal amounts of each.

There is a commercially available miRNA cloning kit (Perkin Elmer Netflex) that uses adapters with random bases at the ends of both adapters for the first and second miRNA cloning steps (Fig 1.10A). However, although this method improves upon the kits that do not use random bases it is only a partial improvement, as discussed previously (Chapter 1, section 1.10). This indicates that one or both of the cloning steps is still quite biased despite the use of random bases. Our results indicate that the second cloning step is not necessarily improved by the use of random bases. In support of this, Hafner et al 2011 found that the second step of cloning to the 5' adapter (Fig 1.9) was the most inefficient. Also, as discussed by Benesova et al (2021), the circularization protocol developed by Somagenics addresses the inefficiency of the second cloning step but does not address cloning bias due to the first step, indicating that Somagenics regard the second cloning step as the most problematic. The template switching protocols that are available as commercial kits (Fig 1.10B) avoid using ligases, however, although the template switching protocols show less cloning bias than the Illumina protocol they still have their own cloning bias problem (see Introduction, 1.10).

Chapter 4-Intramolecular cloning of miRNAs

4.1 Introduction

In Chapter 3 we confirmed that ligation of the 3' adapter RA3 to some miRNAs could be greatly improved by using the ligation modifications described by Zhang et al (2013). However, we were unable to greatly improve the efficiency of ligation of certain miRNAs to the 5' adapter RA5, despite testing all of the improvements reported by Zhang et al., (2013) and further useful suggestions by the senior author R Yi (personal communication, and see Figs 3.9, 3.10B). A company called Somagenics was invited by us to give a webinar and they reported that they also could not efficiently clone certain miRNAs to the RA5 adapter of Illumina and were marketing a likely solution (Fig 4.1).

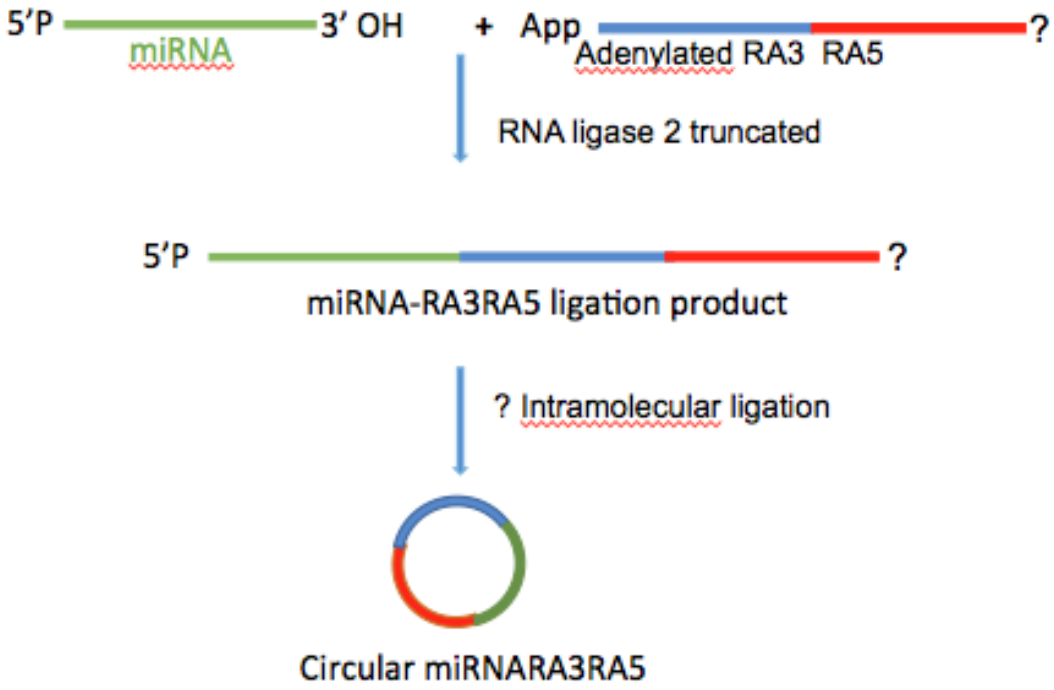


Figure 4.1 Intramolecular ligation of RA5 to miRNAs. Outline of an intramolecular ligation method suggested by Somagenics to improve the ligation of RA5 to the 5' ends of miRNAs. Ligation of the miRNA to the 5' end of RA3 is the same as the Illumina method (Fig 3.2). Subsequently RA5 is ligated intramolecularly to the miRNA to form a circle. RA5 and the miRNA are made of RNA and RA3 of DNA. The question marks indicate steps that were initially unknown to us (see text).

Somagenics first ligate a combined RA3:RA5 adapter to miRNA using truncated RNA ligase 2, as used by Illumina. For the second step they then use a ligase to circularise the miRNA previously ligated to RA3RA5 (Fig 4.1) because they found that the intramolecular circularisation step is far more efficient than the intermolecular RA5 ligation step that is still used by most companies such as Illumina.

At that time the kits sold by Somagenics did not work, although the defect was subsequently traced to a defective enzyme and rectified (Dr Leandro Castellano, personal communication). Furthermore the company would not tell us any of the details of their protocol including the nature of the enzyme used for the circularization, although this was subsequently published as normal RNA ligase 1 (Barberán-Soler et al., 2018). For these reasons and because we wanted to establish a protocol that was not dependent upon a kit, in part to be able to modify it (see below), we undertook the following steps.

Aim

To investigate whether the very poor intermolecular cloning of some miRNAs to the RA5 adapter could be improved by intramolecular cloning to a combined RA5RA3 adapter, as suggested by Somagenics.

4.2 Results

4.2.1 Trying to improve circular ligation

Previously we were unable to ligate mir-214RA3 to RA5 in an intermolecular ligation (Fig 3.10B). Fig 4.2 tests three enzymes for intramolecular ligation of a linear 214RA3RA5 molecule (RNADNARNA). The ligation efficiency is in order RNA ligase 2, RNA ligase 1, Circligase II and Circligase I, as judged by the relative amounts of linear and circular 214RA3RA5 in lanes 2 to 4. RA5 alone was not circularised by these three enyzmes (lanes 5 to 8) because there was no 5' phosphate group on the RA5 that was used. Similar results are reported in Figs S1 to S9 of Appendix 2. The circular form of 214RA3RA5 run more slowly than the linear form (compare lane 1 with lane 2 of Fig 4.2) and was also more resistant to exonuclease treatment (see Appendix 2 – exonuclease resistance of circular DNARNA).

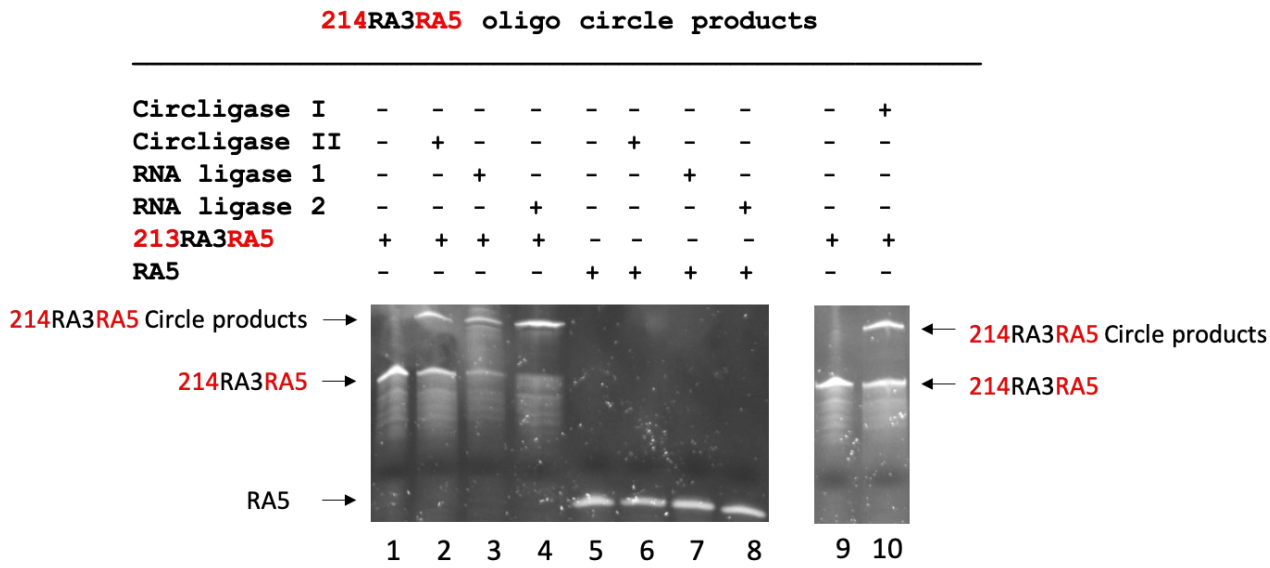


Figure 4.2 Circularisation of 214RA3RA5. All lanes included ATP. Lane 1, 100 ng of 214RA3RA5 untreated. Lanes, 2,3 and 4 100 ng of 214RA3RA5 treated with 100 U/µl Circligase II at 60°C for 1 hour (lane 2), 10 U/µl RNA liagse 1 at 25°C for 2 hours (lane 3), 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 4). Lane 5, 100 ng of RA5 no enzyme. Lanes 6 to 8, 100 ng of RA5 treated with 100 U/µl Circligase II at 60°C for 1 hour (lane 6), 10 U/µl RNA liagse 1 at 25°C for 2 hours (lane 7), 10 U/µl RNA ligase 2 at

37°C for 1 hour (lane 8). Lanes 9 and 10 are from Fig S2 and show 100 ng of 214RA3RA5 untreated (lane 9) or treated with 100 U/µl Circligase I at 60°C for 1 hour.

4.2.2 Identification of enzymes for circular ligation

Table 4.1 summarises the results for Fig 4.2 and further experiments that are presented in Appendix 2 (Figs S1 to S9). Table 4.1 Row 2 reports that 214RA3RA5 was ligated by intramolecular ligation with normal RNA ligase 2 and less efficiently by circligases I and II that act principally upon DNA. Row 4 summarises that circligases I and II could be used to ligate 214RA5NN provided that RA5NN is DNA based. However, the circligases are considerably more expensive than RNA ligase 2 and not quite so efficient at circular ligating 214RA3RA5 (row 2 Table 4.1).

	Circligase I		**Circligase II**		**RNA ligase 2**	
	+ ATP	- ATP	+ ATP	- ATP	+ ATP	- ATP
214-RA3-RA5 (DNA-DNA-DNA)	✓	✓	✓	✓	X	X
214-RA3-RA5 (RNA-DNA-RNA)	(✓)	?	(✓)	(✓)	✓	X
NNRA3RA5 (DNA-RNA)	✓	✓	✓	✓	✓	X
214-RA5NN (RNA-DNA)	✓	(✓)	✓	✓	X	X

Table 4.1 Intramolecular ligation results for the indicated linear molecules by circligases and RNA ligase 2. Red indicates oligos made from RNA bases and black DNA bases. Ticks and crosses show

whether the indicated oligonucleotides (column 1) could be efficiently circularized. RNA ligase 1 gave similar results to RNA ligase 2 (see below) but was not investigated so extensively. The question mark indicates variable results for Circligase I (see Discussion). The tick represents successful ligation.

Having established that the intramolecular ligation of mir214 to NNRA3RA5NN was efficient we next addressed the question of how to reversibly block the 3' end of NNRA3RA5NN (see Fig 4.1). In the Illumina protocol the 3' end of RA3 has a dideoxy group in order to prevent multiple ligations of RA3 by truncated RNA ligase 2 and also to prevent the adenylation enzyme from causing circularization of RA3 (Figure 3.2). The illumina RA3RA5 vector that is used by Somagenics for intramolecular ligation (see above) and by us also requires a block at the 3' end of RA5 to prevent unwanted ligations during the intital miRNA cloning. However, the normal dideoxy modification would prevent the subsequent intramolecular ligation step.

4.2.3 Reversible 3' blocking groups

We investigated the possible use of reversible terminators, which are used for second and next generation sequencing, but these modifications were only available for DNA bases (Professor Steven Benner, Harvard University, personal communication) rather than the 3' RNA base required by RNA ligase 2.

T4 polynucleotide kinase is commonly used to phosphorylate the 5' end of oligonucleotides but it can also dephosphorylate the 3'P groups that are formed during DNA strand breakage (HennerGrunberg and Haseltine, 1983). We therefore investigated the use of 3'P as a reversible modification in our miRNA cloning protocol. This is the method that Somagenics also use (Barberán-Soler et al., 2018) although we did not know that at the time.

Figure 4.3 shows a time course for PNK treatment of a synthetic oligo NNRA3RA5NN-3'P. After PNK treatment the oligo was incubated with RNA ligase 1, which acts similarly to RNA ligase 2 (Fig 4.1). The results show that PNK treatment for 30' was sufficient for RNA ligase 1 to convert NNRA3RA5NN-3'P into a circle (compare lane 4 with lanes 1 to 3 of Fig 4.3).

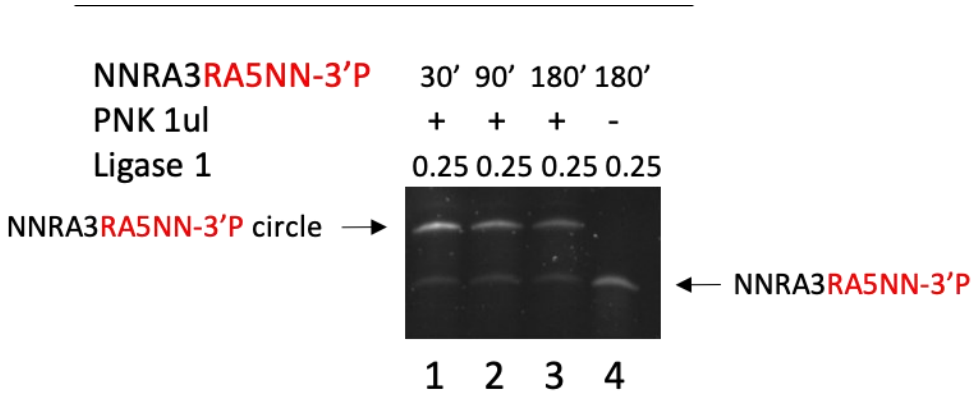


Figure 4.3 3'P removal. Lane 1, 50 ng of NNRA3RA5NN-3'P was treated with 10 U/µl PNK enzyme at 37°C for 30, 90 or 180 minutes and then 0.25 µl of 10 U/µl RNA ligase 1 enzyme at 25°C for 2 hours. Lane 4 is a control lane with no PNK treatment.

A comparison of Fig 4.3 with lanes 3 to 6 of Fig 4.4 shows that NNRA3RA5NN-3'P was not always fully converted to a circle form following PNK treatment to remove the 3'P and the addition of RNA ligase 2 to cause circularisation. In retrospect further PNK titration experiments might be helpful. Lanes 5 and 6 of Fig 4.4 indicate that adenylation may further inhibit circularisation by RNA ligase 2. However,

this was not considered to be an issue because the function of the adenyl group is to allow truncated RNA ligase 2 to ligate a miRNA to RA3 (Fig 4.1), which it can do very efficiently (Fig 3.2). Overall, there was little difference between the use of 0.5 and 1 µl of normal RNA ligase 2 in Fig 4.4.

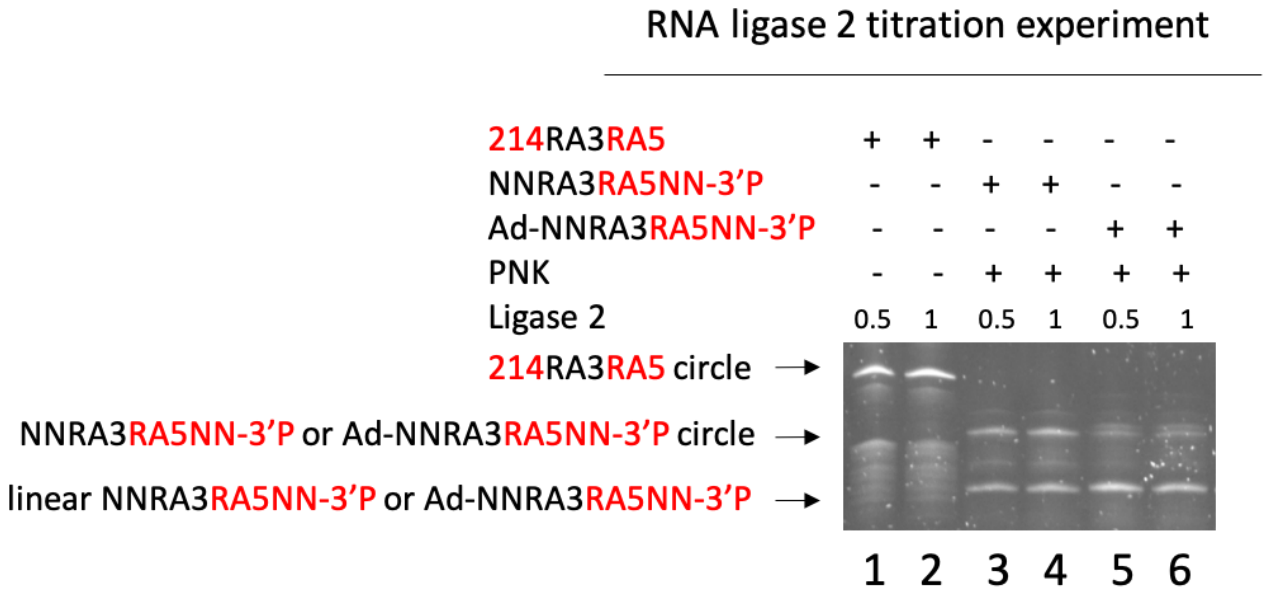


Figure 4.4 3'P removal. Lanes 1, 2: 100 ng of 214RA3RA5 was treated with 0.5 µl or 1.0 µl of 10 U/µl RNA Ligase 2 enzyme and incubated at 37°C for 30 minutes without PNK enzyme. Lanes 3,4: 100 ng of oligo NNRA3RA5NN-3'P was treated with 1 µl 10 U/µl PNK enzyme at 37°C for 30 minutes and adding 0.5 µl or 1.0 µl of 10 U/µl RNA Ligase 2 enzyme at 37°C for 30 minutes. Lane 5, 100 ng of Ad-NNRA3RA5NN-3'P was used 10 U/µl PNK enzyme at 37°C for 30 minutes and 0.5 µl of 10 U/µl RNA Ligase 2 enzyme incubated at 37°C for 30 minutes. Lane 6, 100 ng of Ad-NNRA3RA5NN-3'P was used 10 U/µl PNK enzyme at 37°C for 30 minutes and 1 µl of 10 U/µl RNA Ligase 2 enzyme incubated at 37°C for 30 minutes.

4.2.4 Cloning of miRNAs into NNRA3RA5NN-3'P

We next tested whether mir-214 or miR-101-1-3p could be ligated to NNRA3RA5NN-3'P and then circularised efficiently (Fig 4.5). Fig 4.5 lanes 9 and 10 show where 70 bases linear 214RA3RA5 and circularised 214RA3RA5 run. From this it can be deduced that the ligation of 214 to NNRA3RA5NN-3'P produced circle products in lanes 5 and 6 of Fig 4.5 and similarly for the ligation of mir101 to NNRA3RA5NN-3'P (lane 7). There was no appreciable difference for ligation of mir-214 and NNRA3RA5NN-3'P by RNA ligases 2 or 1 (lanes 5 and 6). Lanes 2 to 4 show partial circularisation of the vector NNRA3RA5NN-3'P (similar to Fig 4.4) and lane 8 shows that the ligation of the STP oligo to Ad-NNRA3RA5NN-3'P by truncated RNA ligase 2 generates an upper band that would be

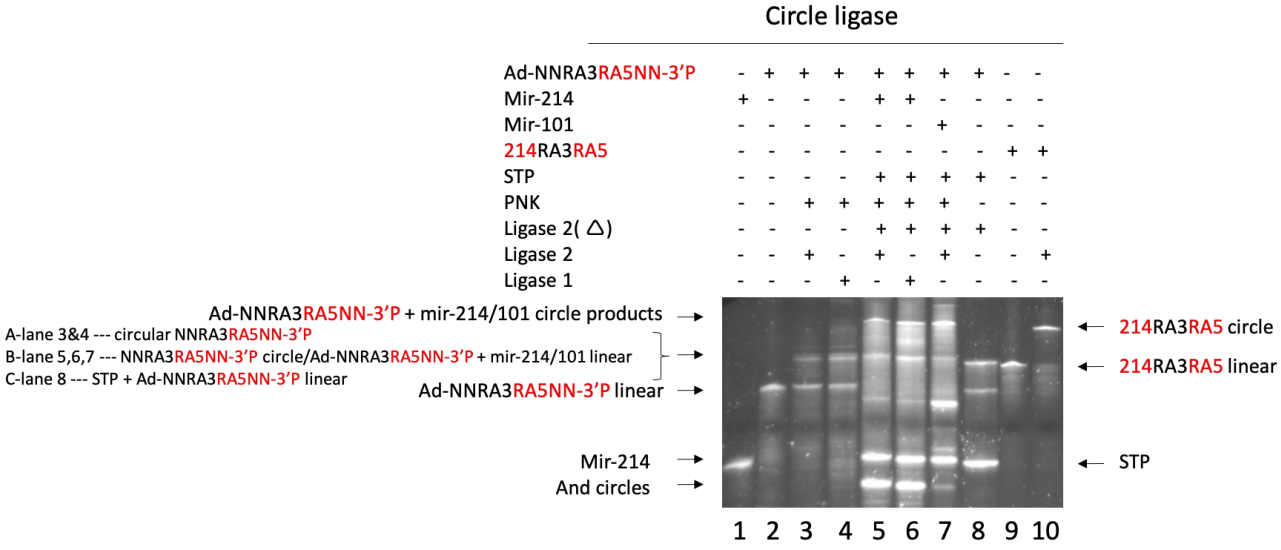


Figure 4.5 Cloning miR-101-1-3p and miR-214. Lane 1, 100 ng of mir-214 linear (23 bases) without adding any enzymes; Lane 2, 100 ng of Ad-NNRA3RA5NN-3'P (51 bases) without adding any enzymes. Lane 3, 100 ng of Ad-NNRA3RA5NN-3'P was used 10 U/µl PNK enzyme at 37°C for 30 minutes and 10 U/µl RNA Ligase 2 enzyme incubated at 37°C for 30 minutes. Lane 4, 100 ng of Ad-NNRA3RA5NN-3'P was used 10 U/µl PNK enzyme at 37°C for 30 minutes and 10 U/µl RNA Ligase 1 enzyme at 25°C for 2 hours. Lane 5, 300 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 300 ng of mir-214 by adding 200 ng of STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzyme incubated at 37°C for 30 minutes. Lane 6, 300 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 300 ng of mir-214 by adding 200 ng of STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 1 enzyme at 25°C for 2 hours. Lane 7, 300 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 300 ng of miR-101-1-3p by adding 200 ng of STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzyme incubated at 37°C for 30 minutes. Lane 8, 100 ng of pre-adenylated NNRA3RA5NN-3'P was ligated with 200 ng of STP by using 200 U/µl RNA ligase 2 Δ as arrows indicate. Lane 9, 100 ng of 214RA3RA5 linear (70b) without adding any enzymes; lane 10, 100 ng of 214RA3RA5 was circularised by using 10 U/µl RNA ligase 2 enzyme incubated at 37°C for 30 minutes.

expected to be 73 bases in length and indeed runs similarly to linear 214RA3RA5 (70 bases) in lane 9 and also runs similarly to circular NNRA3RA5NN in lanes 3 and 4. Consequently, the bands of this size in lanes 5 to 7 are difficult to interpret. These results are repeated in figures S10 to S12 in Appendix 2.

4.2.5 Reverse transcription of circle ligations

We next tested whether circular ligation products could be amplified by RT-PCR (Figs 4.6, 4.7).

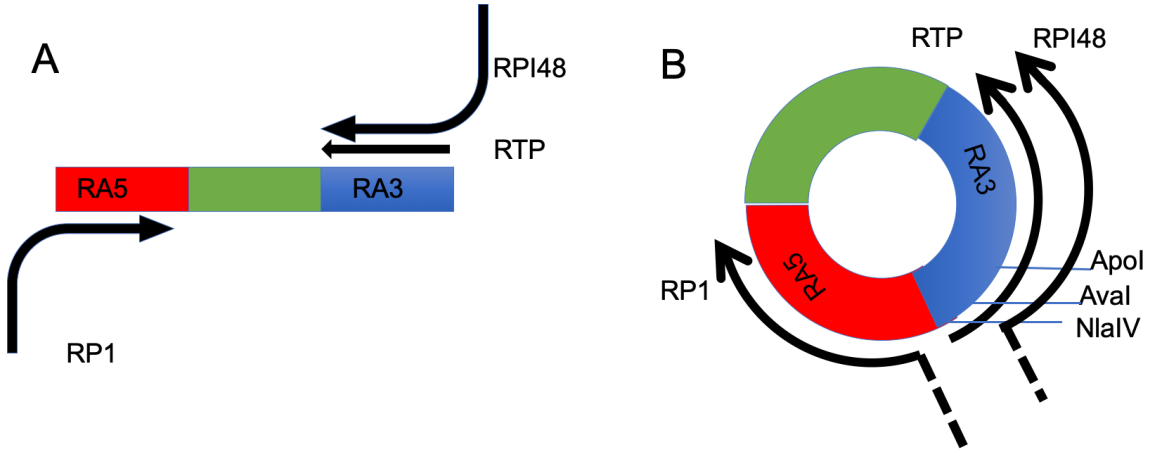


Figure 4.6 Illustration of RT-PCR for linears and circles. A. Illustration of the primers used for the standard Illumina method of miRNA amplification using the initial reverse transcription primer RTP, then primers RPI48 and RP1 for amplification. B. The same primers were used for RT-PCR of the circular version of A. Unique restriction sites ApoI, AvaI and NlaIV are shown. A PCR product from either A or B would have an expected size of RP1 and RPI48 (113 bases) plus the size of the miRNA and small regions of NNRA3 and RA5NN (9 bases), (see Materials and methods, RT-PCR primers).

Fig 4.6 shows that the same primers that are normally used by Illumina to prepare DNA libraries of miRNAs (Fig 4.6A) might also be used for circular forms of the same molecules (Fig 4.6B).

Fig 4.7 shows a lot of unexpected high molecular weight bands in all of the lanes involving PCR of circular molecules. In lane 7 the oligo 214RA3RA5 was previously circularized and then subjected to RT-PCR and it can be seen that multiple bands were produced rather than the expected single linear band of 141 bases. Similarly in lanes 3 and 6 multiple bands are seen rather than the expected single linear band of 122 bases (Materials and methods, RT-PCR primers). Lanes 1 and 4 are control lanes without PCR and were used to estimate the amount of input as 20 ng. Lanes 2 and 5 of Fig 4.7 are similar to lanes 1 and 4 except for an additional PCR step that has produced numerous higher molecular weight bands.

We initially thought that the large amount of higher molecular weight material running in all of the PCR lanes of Fig 4.7 might indicate a problem with RT-PCR rolling circle production, which has been previously reported to occur when trying to amplify circular DNA and produces multiple tandem copies of the starting material (Mohsen and Kool, 2016). We therefore tested whether the use of AvaI or ApoI enzymes (Fig 4.8) might resolve the possible RT-PCR rolling circle problem. Fig 4.8 repeats the observation of high molecular weight bands (lanes 2 and 3) following RT-PCR of NNRA3RA5NN-3'P or 214RA3RA5. These RT-PCR products were cut with either ApoI or AvaI and run on lanes 4 to 7. Several bands can be seen. Linear RT-PCR band of NNRA3RA5NN-3'P should be cut into fragments of 51 and 67 bases by AvaI and fragments of 56 and 62 bases by ApoI (Materials and methods, RT-PCR primers). A 214RA3RA5 PCR fragment would generate bands of 51 and 86 bases by AvaI and 56 and 81 bases by ApoI. This does arguably match the results of lanes 4 to 7, particularly as these lanes will also contain single stranded RP1 and RPI48 primers that are likely to run at the bottom of the gel. The gel would have benefitted from lanes showing PCR amplification of linear.

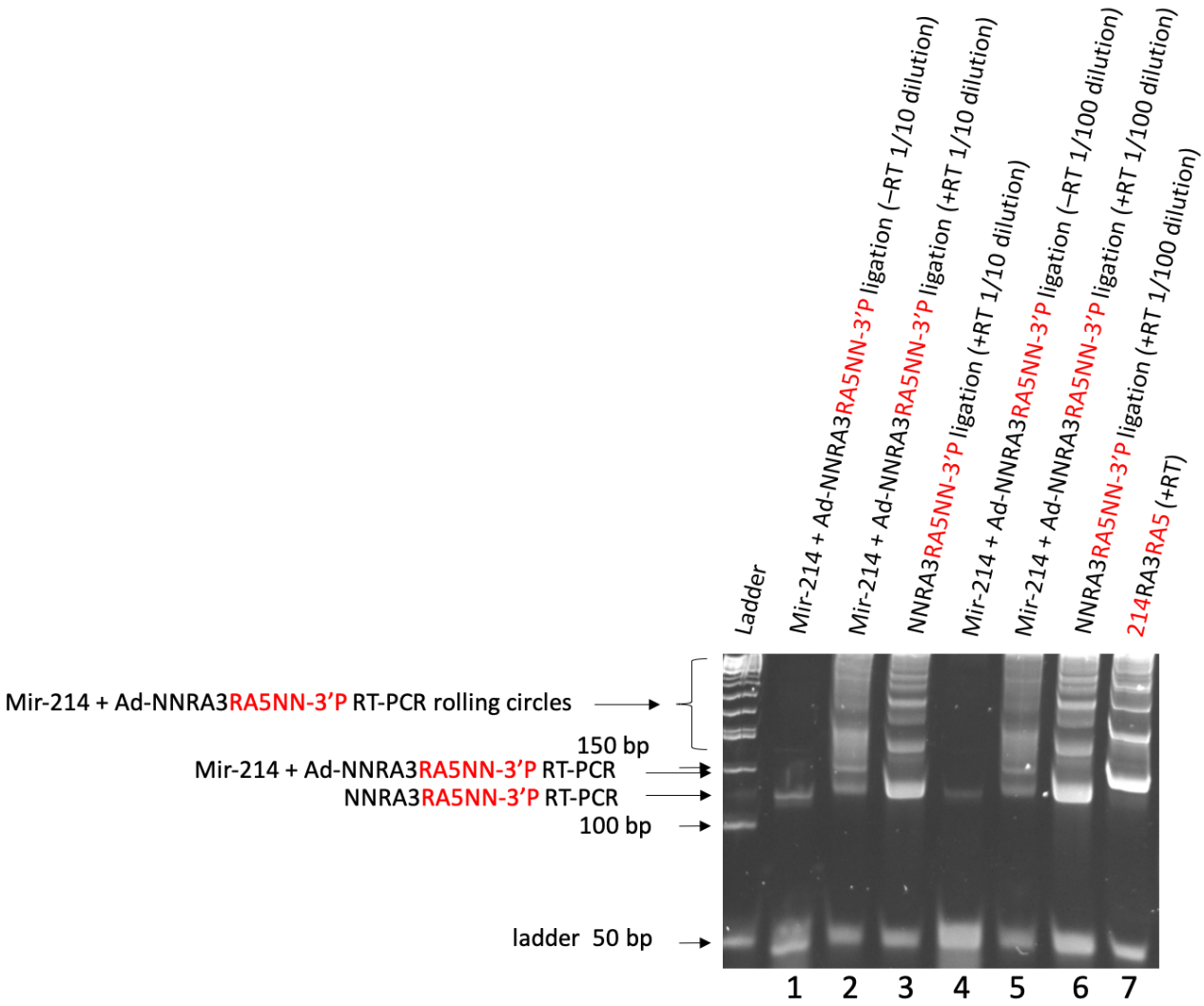


Figure 4.7 High molecular weight bands. Lanes 1, 2: 50 ng each of Mir-214 + Ad-NNRA3RA5NN-3'P was ligated without (lane 1) or with reverse transcriptase in RT-PCR amplification (lane 2). Lanes 3 and 6: 50 ng of NNRA3RA5NN-3'P circle with RT-PCR amplification. Lanes 4 and 5: 50 ng each of mir-

214 + Ad-NNRA3RA5NN-3'P circle without (lane 4) and with reverse transcriptase in RT-PCR (lane 5). Lane 7, 50 ng 214RA3RA5 circle with RT-PCR (All PCR reactions used 25 cycles).

NNRA3RA5NN-3'P and of linear 214RA3RA5 and their subsequent cutting with restriction enzymes. Overall, Fig 4.8 does not support a rolling circle problem because the cutting of any tandem repeats produced by rolling circle replication with either AvaI or ApoI should generate fragments of 122 bases for lanes 4 and 6 and 141 bases for lanes 5 and 7 of Fig 4.8 (see Materials and methods), which is not strongly evident in lanes 4 to 7 of Fig 4.8. This indicates that a different artefact is responsible for the high molecular weight bands seen in Figs 4.7 and 4.8.

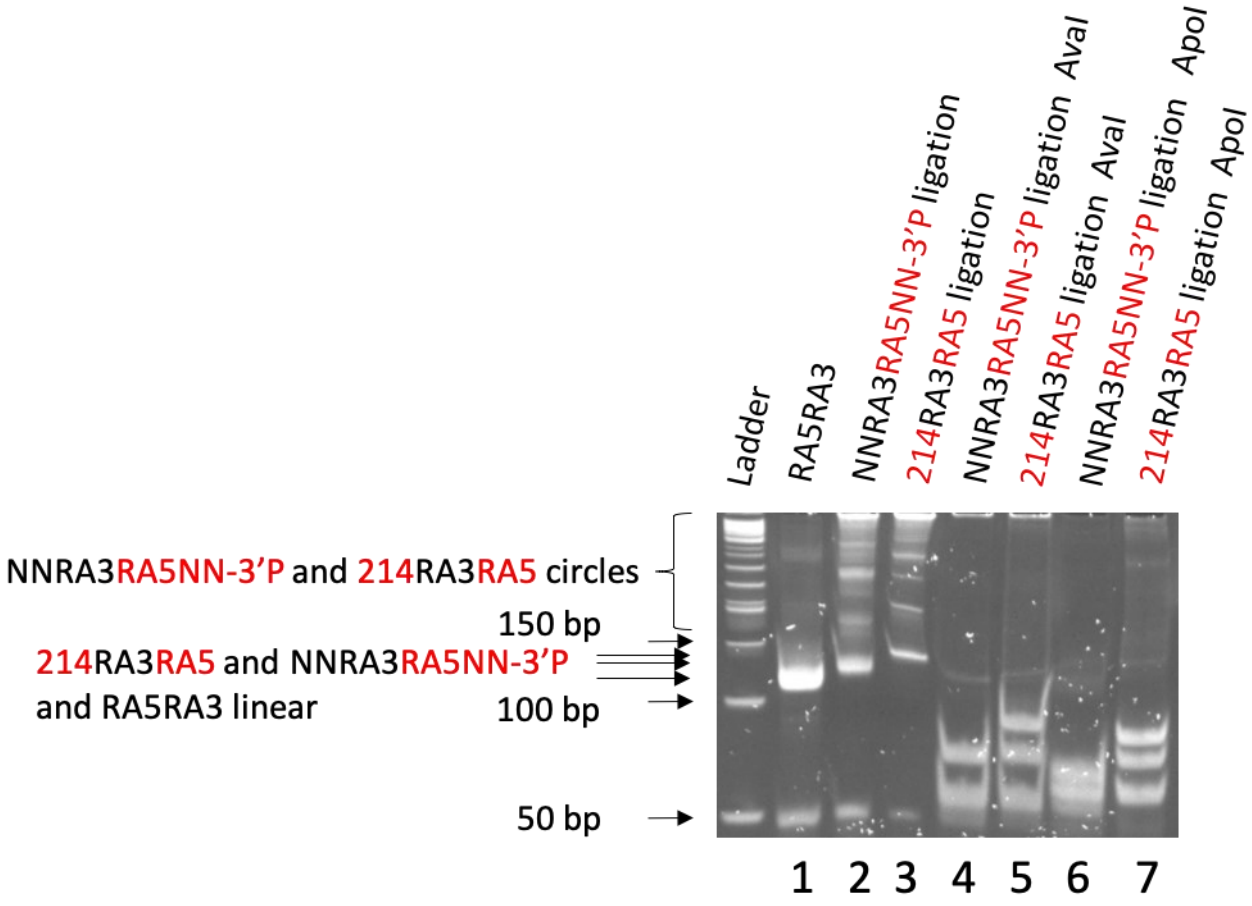


Figure 4.8 Analysis of high molecular weight bands. Lane 1 PCR (25 cycles) of the linear DNA oligo RA5RA3, and RT-PCR of circularized NNRA3RA5NN-3'P (lane 2) and circularized 214RA3RA5 (lane 3). Lanes 4 to 7 show two repeats of lanes 2 and 3 that were then treated with 10 U AvaI restriction enzyme (lanes 4 and 5) for 1 hour at 37°C or 20 U of ApoI (lanes 6, 7) for 1 hour at 37°C. NNRA3RA5NN-3'P was treated with 10 units of T4 Polynucleotide Kinase (PNK) that was incubated at 37°C for 30 minutes.

4.2.6 Method for stopping high molecular weight products from RT-PCR of circles

We decided to test if we could convert the circular products produced by intramolecular ligation (Fig 4.1) into a linear product by first cutting with the restriction enzyme NlaIV prior to the RT-PCR step. We chose NlaIV because it cuts close to the junction of RA3 and RA5 and so is less likely to affect the subsequent annealing of primers required for PCR (Fig 4.6B). Fig 4.9 shows a preliminary experiment where we test whether the annealing of the oligo RTP to RA3 (Fig 4.6) could efficiently generate double stranded DNA suitable for cutting by NlaIV. RTP is normally used to generate cDNA from RA3RA5 prior to PCR (Fig 4.6).

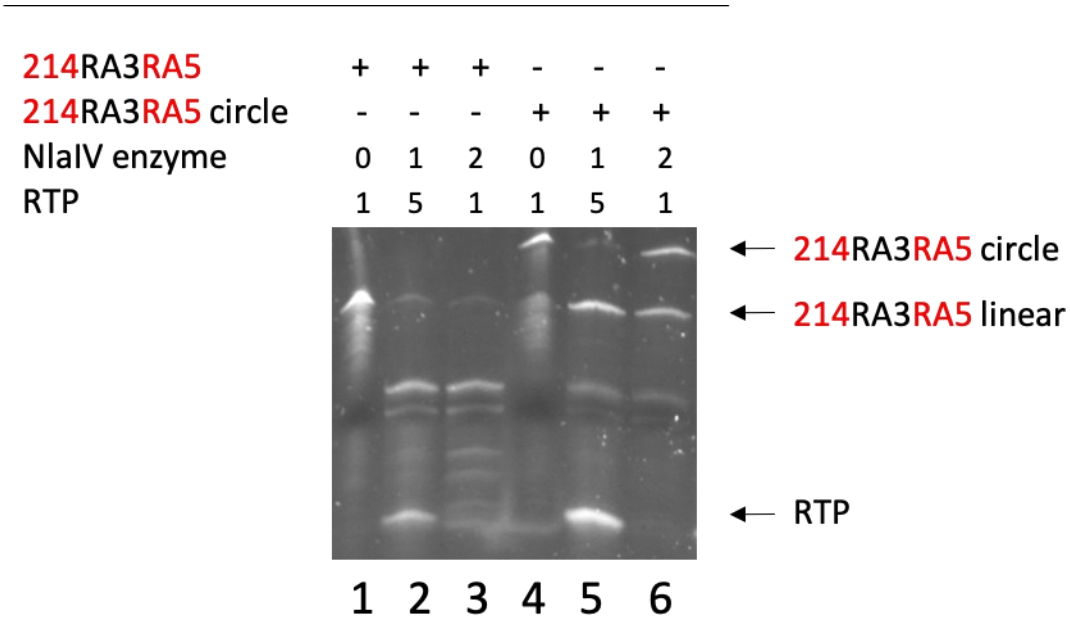


Figure 4.9 Method to linearise a single strand circle. Lane 1, 100 ng of the linear oligo 214RA3RA5. Lanes 2 and 3, 100 ng of linear 214RA3RA5 was annealed to 5 µl (lane 2) or 1 µl of RTP (50 ng/µl), heated to 65°C for 5 minutes and then cooled at 4°C for 2 minutes and then treated with 1 µl or 2 µl NlaIV (2 U/µl) at 37°C for 1 hour. Lane 4, 100 ng of previously made 214RA3RA5 circle. Lanes 5 and 6, 100 ng of 214RA3RA5 circle was annealed to 5 µl (lane 5) or 1 µl (lane 6) RTP (50 ng/µl) as described above and then incubated with 2 U/µl of NlaIV enzyme at 37°C for 1 hour.

Fig 4.9 lanes 4 to 6 show that circular 214RA3RA5 was efficiently converted into a linear form dependent upon sufficient RTP oligo and NlaIV, as expected (Fig 4.6). Fig 4.9 lanes 1 to 3 show the results for a similar treatment of linear 214RA3RA5. Successful NlaIV cutting would be expected to remove RA5 from 214RA3RA5 (Fig 4.6B), which is consistent with the smaller size of the major band in lanes 2 and 3.

Fig 4.10 tests NlaIV treatment following an intramolecular ligation reaction. Lane 4 is a positive control to show that circular 214RA3RA5 (previously made by intramolecular ligation) generates a lot of high molecular weight bands after RT-PCR. Lanes 2 and 3 compare the effect of NlaIV treatment prior to RT-PCR of a previous intramolecular ligation of mir214 to Ad-NNRA3RA5NN-3'P. In lane 3 a single band of similar size to the expected size of the linear RT-PCR product of 214RA3RA5 (145 bases) is marked with an asterisk.

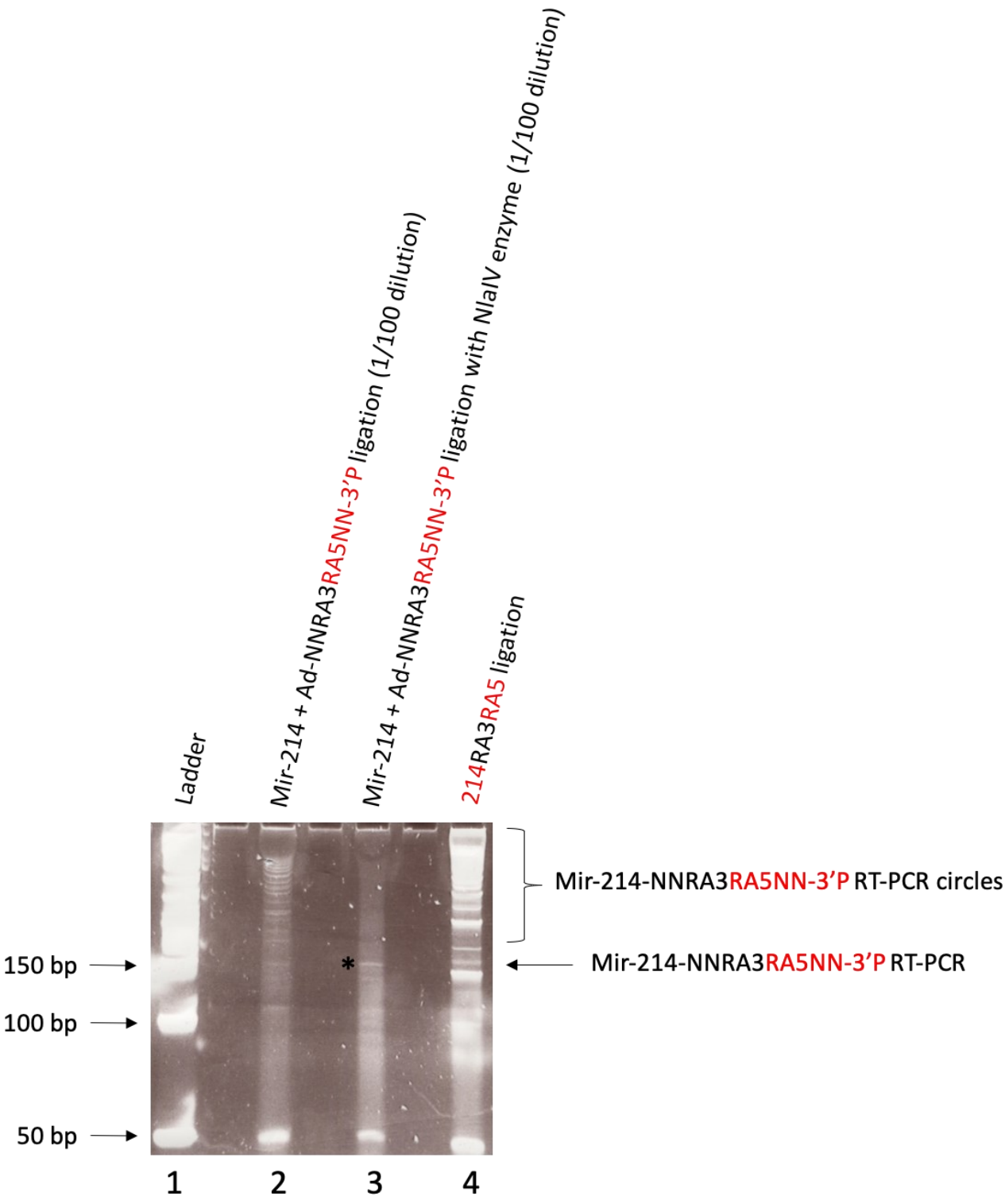

Ladder
Mir-214 + Ad-NNRA3RA5NN-3'P ligation (1/100 dilution)
Mir-214 + Ad-NNRA3RA5NN-3'P ligation with NlaIV enzyme (1/100 dilution)
214RA3RA5 ligation
Mir-214-NNRA3RA5NN-3'P RT-PCR circles
Mir-214-NNRA3RA5NN-3'P RT-PCR
150 bp
100 bp
50 bp
*
1
2
3
4

Figure 4.10 Cloning test of NlaIV. Lane 1, DNA ladder. Lane 2, mir-214 + Ad-NNRA3RA5NN-3'P circle ligation with the 1 in 100 dilution and RT-PCR. Lane 3, mir-214 + Ad-NNRA3RA5NN-3'P circle ligation and then was treated with 2 U/µl NlaIV restriction enzyme at 37°C for 1 hour, which was made 1 in 100 dilution and then RT-PCR. Lane 4, 214RA3RA5 circle was used to do RT-PCR as a positive control (RT-PCR 15 cycles). The circle ligation method is illustrated in Fig 4.12 and detailed in Chapter 2.

4.2.7 The STP oligo

Fig 4.11 tests the importance of the STP oligo, which is ligated to any unused adapter that is left over during the cloning protocol (fig 4.12). The STP oligo is an important step in the illumina protocol that prevents unused RA5 from ligating to RA5 (see 3.2.4). The STP oligo works by complementary pairing to the 5' end of unused RA3 and this step obviates the need for a gel purification step (see 3.2.4). Surprisingly we found that STP still ligated efficiently to NNRA3 despite the expected disruption to complementary pairing by the NN random bases (see Fig 3.8 and FigS15).

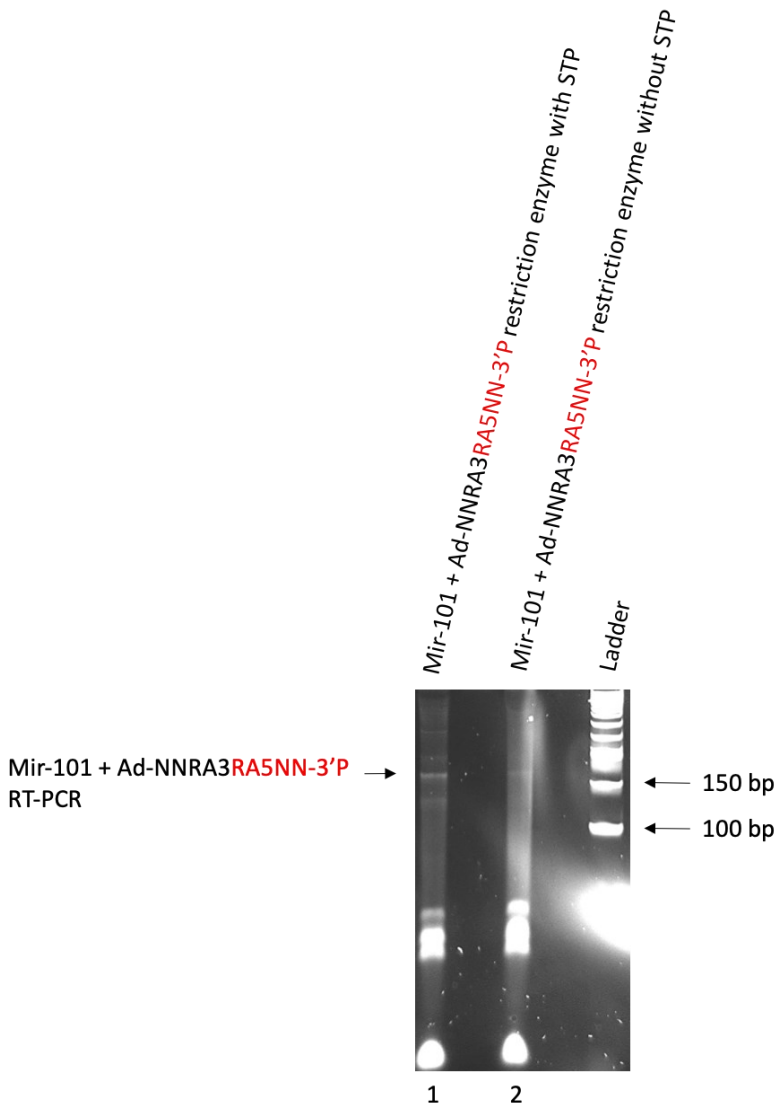


Figure 4.11 Cloning test of the STP oligo. Lanes 1 and 2, miR-101-1-3p + Ad-NNRA3RA5NN-3'P were ligated as outlined in Fig 4.12 with (lane 1) or without STP oligo (lane 2) and then treated with 2 U/µl NlaIV restriction enzyme treatment at 37°C for 1 hour. After this the reaction was diluted between 1 in 10 and 100 for RT-PCR (15 cycles).

A comparison of lanes 1 and 2 of Fig 4.11 shows that the desired ligation product (arrowed) was more intense when the STP oligo was included. This indicates but does not prove the likely value of including the STP oligo in the full protocol (Fig 4.12) but further repeats are required.

4.2.8 MicroRNA libraries

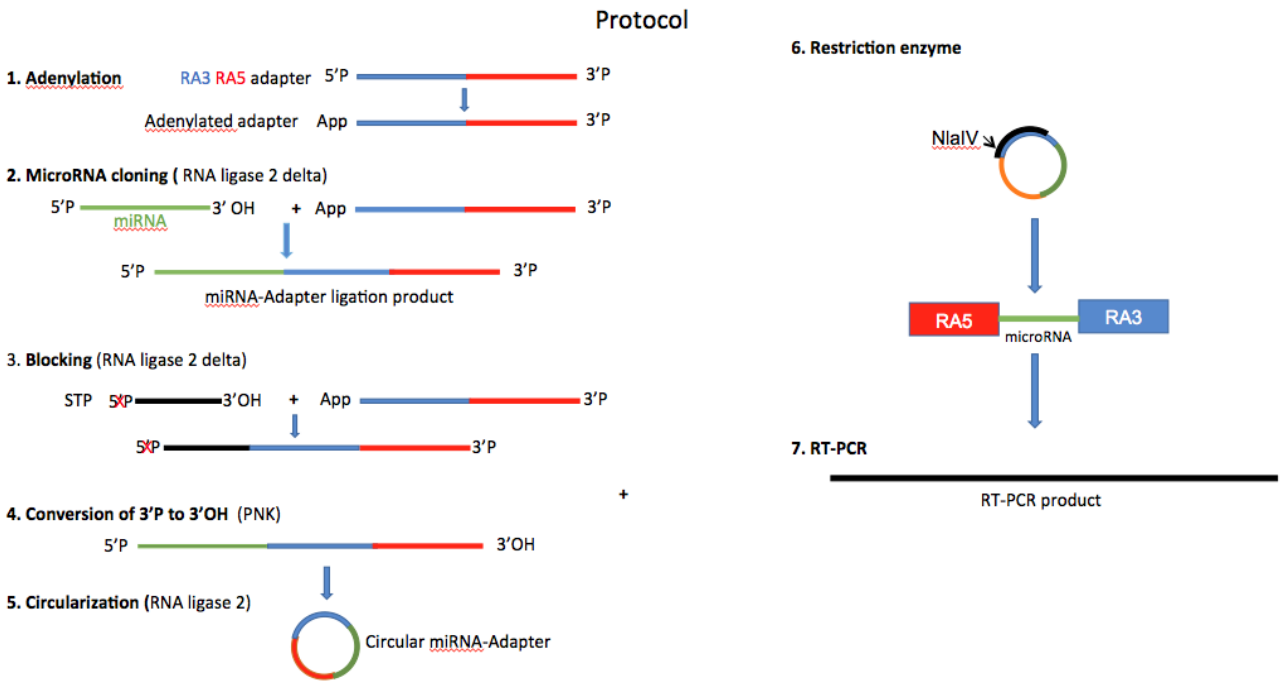


Figure 4.12 Illustration of the full protocol that was developed to clone miRNAs. For details see Materials and methods and text.

Fig 4.13 uses the protocol illustrated in Fig 4.12 to clone a miRNA reference library miRXplore (Miltenyi Biotec) that includes 999 different microRNAs with a size range from 18 to 29 bases (lanes 2 and 4). The major bands in lanes 2 and 4 are similar in size to the linear RA5RA3 and NNRA3RA5NN-3'P controls in lanes 1 and 3. There is a faint but discernable band in lanes 2 and 4 running just above the 150 bases band in the DNA ladder (third band up of the unmarked lane), which although faint is in the expected size range for the cloned miRNA reference library. Lanes 5 and 6 are controls without NlaIV treatment and also repeat the higher molecular weight band problem reported above (Fig 4.7). A comparison of lanes 3 with 5 and 6 with 7 shows only partial resolution of this problem by NlaIV treatment.

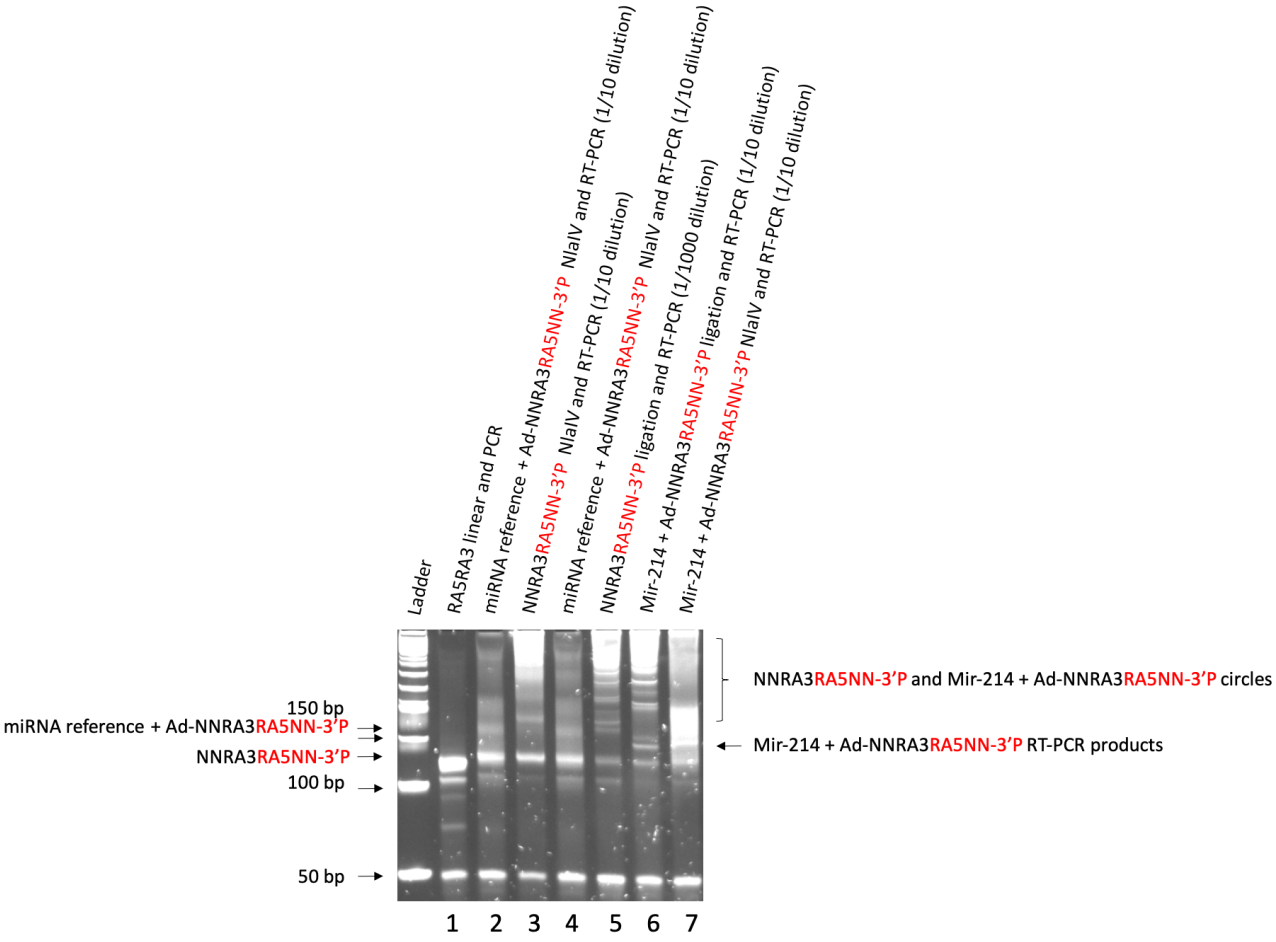


Figure 4.13 Cloning a miRNA reference library. All samples were amplified by RT-PCR (25 cycles) using the primers illustrated in Fig 4.6. Lane 1, RA5RA3 linear oligo control. Lanes 2 and 4, 500 ng of miRNA reference library cloned into 100 ng of Ad-NNRA3RA5NN-3'P, cut with 2 U/µl NlaIV for 1h at 37°C and amplified by RT-PCR as illustrated in Fig 4.12. Lanes 3 and 5, NNRA3RA5NN-3'P circle ligation and then treated with (lane 3) or without (lane 5) NlaIV treatment and RT-PCR. Lanes 6 and 7, ligation of mir-214 to Ad-NNRA3RA5NN-3'P and treatment with (lane 7) or without NlaIV (lane 6) and RT-PCR.

Fig 4.14 shows repeats for the cloning of mir101 (lane 1) with STP and the miRNA reference library cloning (lane 2).

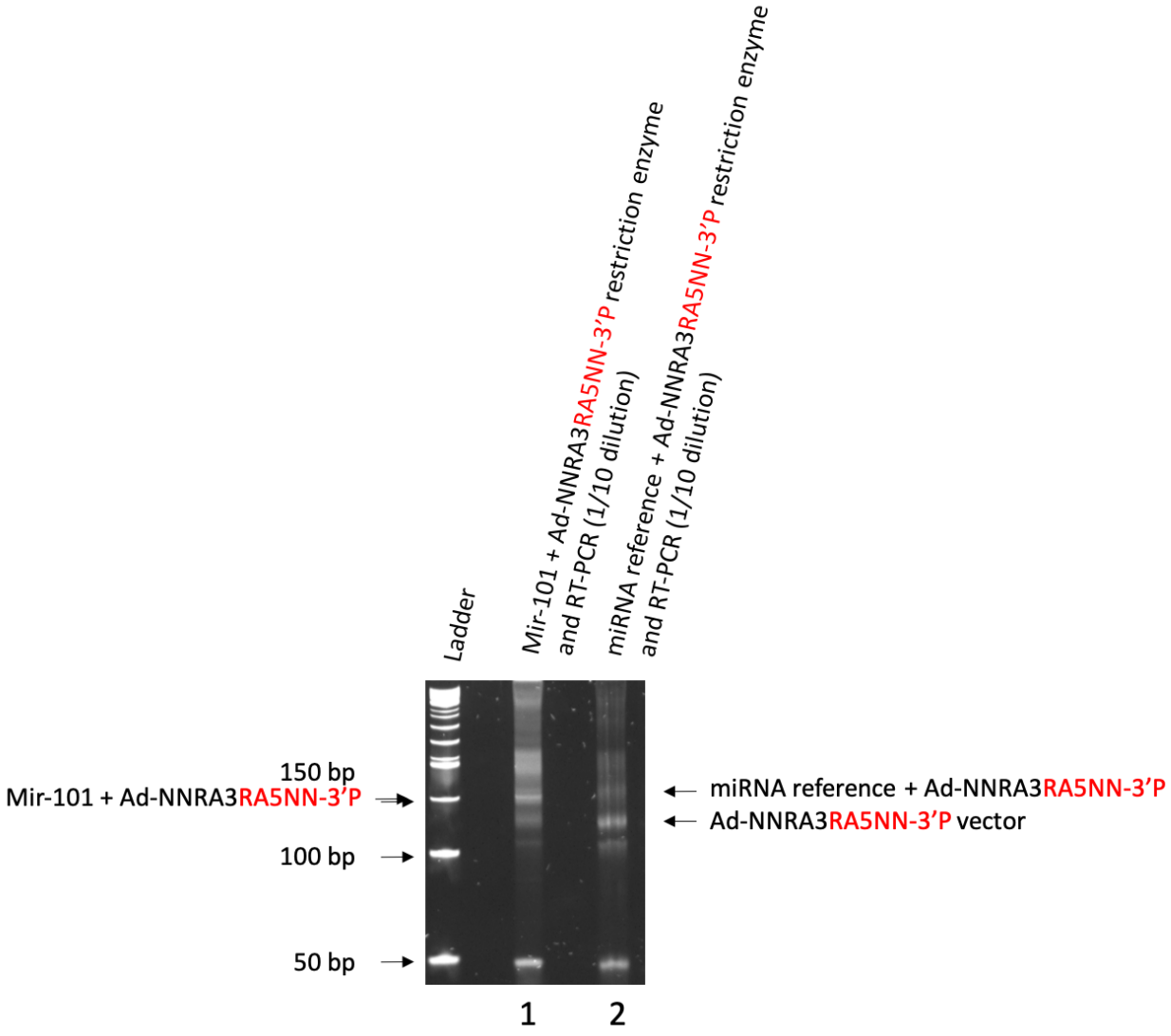


Figure 4.14 Comparison of miR-101-1-3p and reference library cloning. Lane 1, 100 ng of oligo miR-101-1-3p ligated to 100 ng of Ad-NNRA3RA5NN-3'P as outlined Fig 4.12, 25 cycles for RT-PCR. Lane 2, 5 fmol/µl per RNA oligonucleotide of the miRXplore Reference ligated to 100 ng of Ad-NNRA3RA5NN-3'P.

Fig 4.15 shows a repeat cloning of the microRNA reference library into Ad-NNRA3RA5NN-3'P (Fig 4.12) up until the generation of cDNA at the start of step 6 of Fig 4.12, following which the cDNA mix was titrated prior to PCR. The results show that 1µl of cDNA was better than 1/10, 1/100, and 1/1000 dilutions.

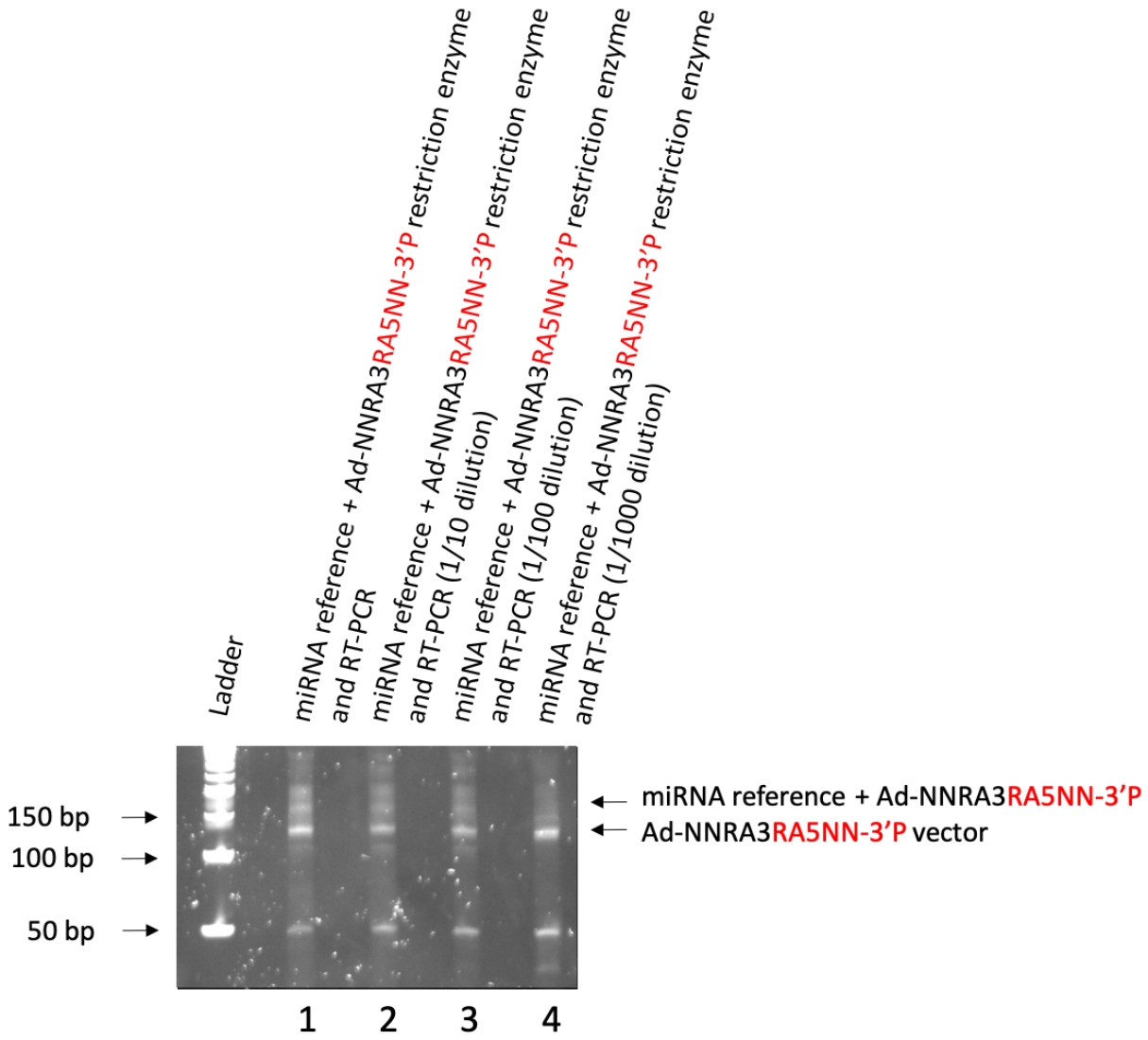


Figure 4.15 RT-PCR titration following cloning. Lanes 1 to 4, 1 µl and ten fold dilutions of miRNA reference library generated as described in Fig 4.12 (25 cycles RT-PCR). 100 ng of Ad-NNRA3RA5NN-3'P and 1 µl of miRXplore (5 fmol/µl) were ligated.

We next tested whether the faint wide band that ran above the Ad-NNRA3RA5NN-3'P PCR product in Fig 4.15 consisted of a mix of reference miRNAs cloned into Ad-NNRA3RA5NN-3'P. We devised a pulldown protocol for one of the shortest miRNAs (mir-575, 19 bases) and one of the longest (mir-768, 28 bases) in the reference library and for miR-101-1-3p (23 bases), which was also in the library. After the RT-PCR step of Fig 4.12 each miRNA was separately pulled down by using a complementary oligo attached to biotin, as outlined in Fig 4.16 and then further amplified by RT-PCR but using shorter primers in order to see any differences in size more easily. The pulldown method is based upon previous work by Diana Alexieva (PhD Thesis https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.745277).

4.2.9 A miRNA pulldown protocol

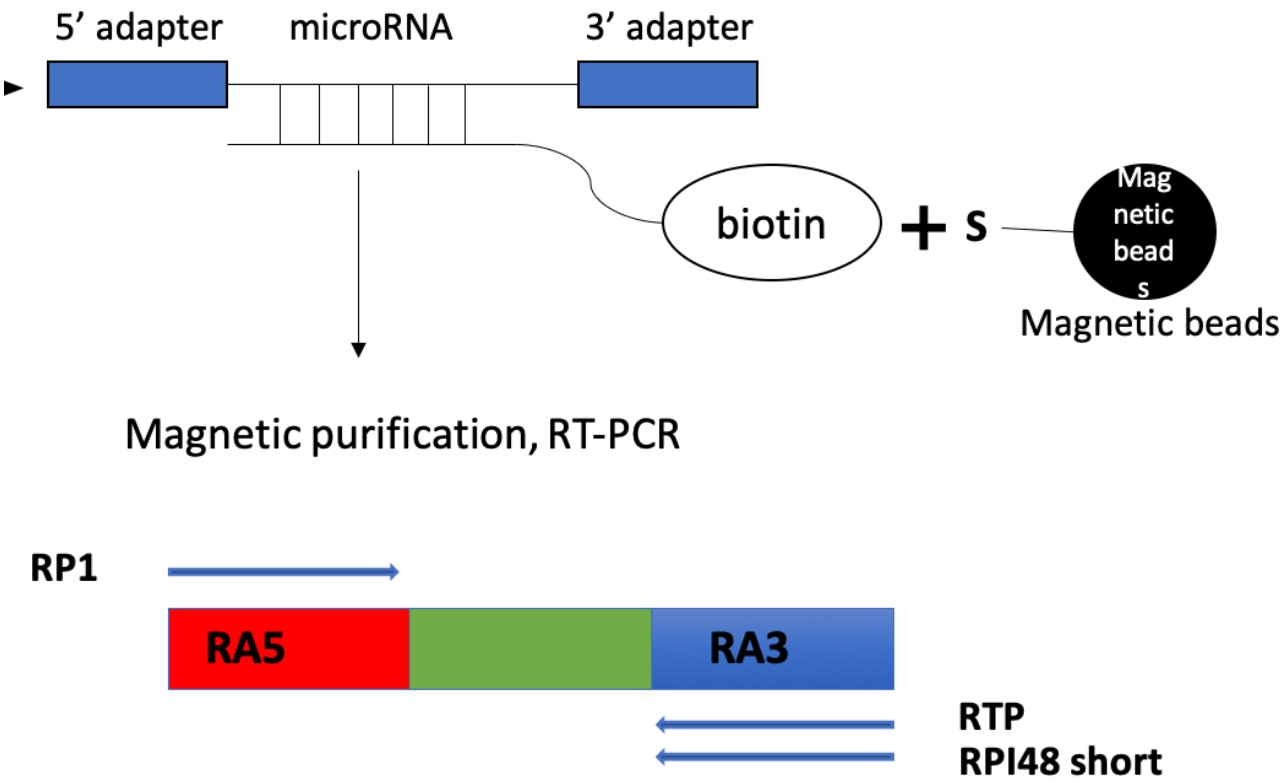


Figure 4.16 Outline of a pulldown protocol. Following cloning of the miRNA reference library in RA3RA5 individual miRNA was pulled down by using complementary oligos attached to biotin. The pulldown miRNAs were then amplified by RT-PCR using shorter primers that of RP1: 22 bases and RPI48: 20 bases that annealed to RA5 and RA3. Oligonucleotides that were used for this experiment are listed in Chapter 2 (2.6 pull down and 2.8 list of oligos and primers).

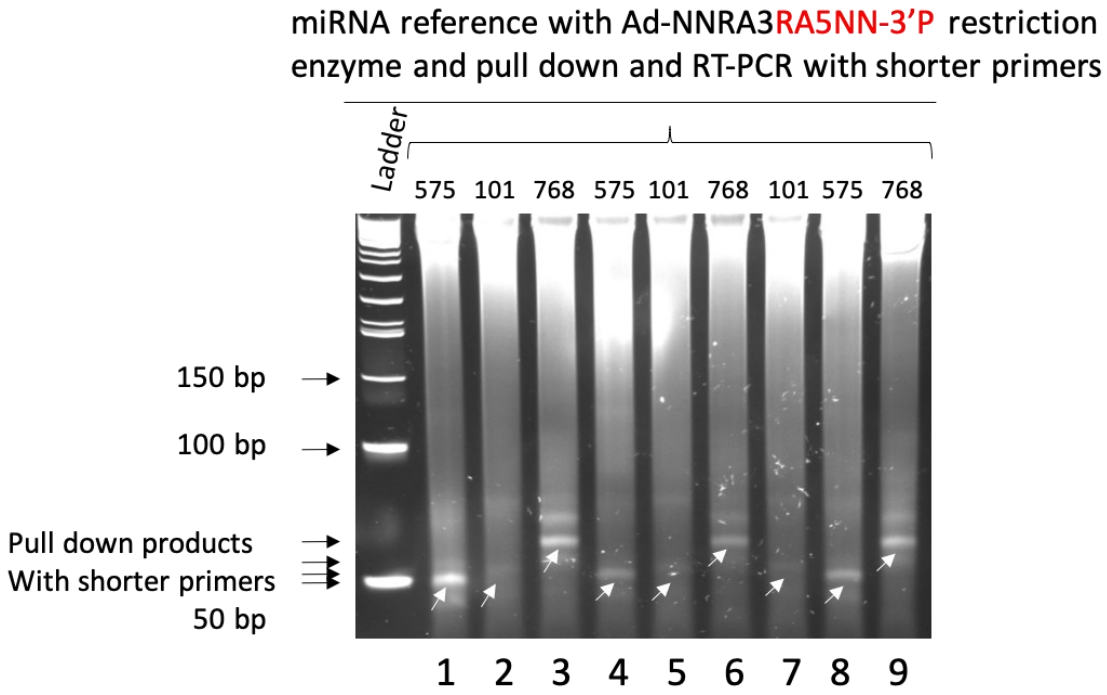


Figure 4.17 Pulldown of the shortest and longest miRNAs in the reference library. Lanes 1 to 3, Pulldowns of mir-575 (lane 1), miR-101-1-3p (lane 2) and mir-768 (lane 3) from a miRNA reference ligated into Ad-NNRA3RA5NN-3'P as described in Fig 4.12. Each pulldown was further amplified by RT-PCR (25 cycles) using primers illustrated in Fig 4.16. Lanes 4 to 9 are two duplicates from the same experiment. The bottom mol wt markers are 50 and 100 bases. All samples were run on an 8% non-denaturing PAGE gel.

Fig 4.17 lanes 1 to 3 show three bands (marked by arrows) that subject to confirmation match the expected size of the PCR fragment for the three different miRNAs (mir-575: 70 bases; mir101: 74 bases; miR768: 79 bases). Lanes 4 to 9 are repeats. The faintest band is seen for miR-101-1-3p (lanes 2,5,7) possibly because the complementary oligo used for the pulldown was relatively short.

4.2.10 Cloning total RNA from cell lines

Fig 4.18 shows the analysis of cloning total RNA from the cell line HCT 116 (see materials and methods) into the vector Ad-NNRA3RA5NN-3'P using the protocol depicted in Fig 4.12. Lane 5 is a control showing the cloning of a single miRNA into Ad-NNRA3RA5NN-3'P as previously analysed (Figs 4.13, 4.14). Comparison of lanes 1 and 5 indicates the likely position of Ad-NNRA3RA5NN-3'P only and the immediate bands above Ad-NNRA3RA5NN-3'P are presumptive total RNA cloned into Ad-NNRA3RA5NN-3'P (lane 1) and miR-101-1-3p cloned into Ad-NNRA3RA5NN-3'P (lane 5). The shorter primers that were used previously (Fig 4.17) were not successful in this experiment (Fig 4.18 lanes 3,4).

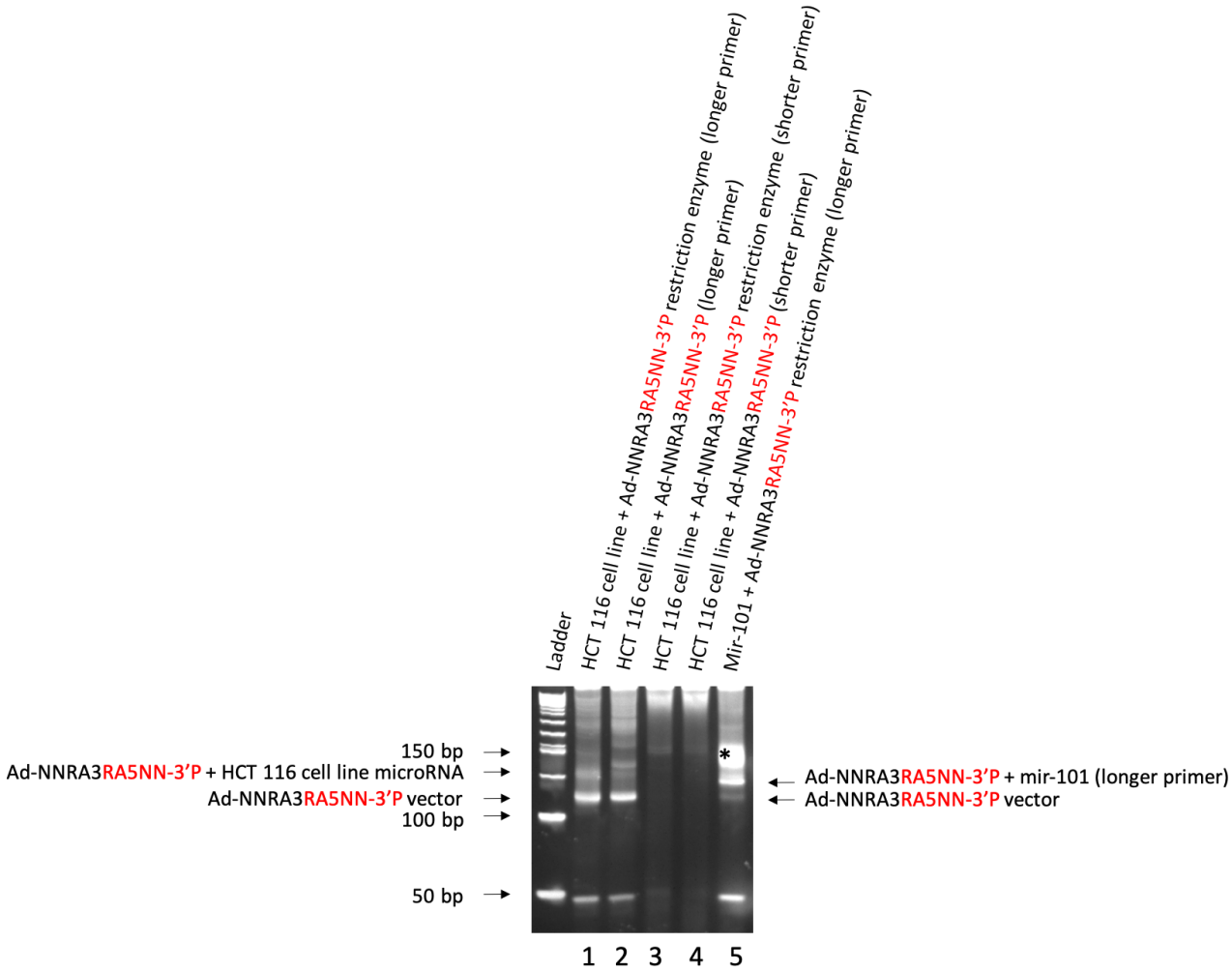


Figure 4.18 miRNA cloning from HCT116 cells. All the experiments with RT-PCR. Total RNA (see materials and methods) was isolated from the cell line HCT 116. Lane 1,2, HCT 116 total RNA cloned into Ad-NNRA3RA5NN-3'P with or without NlaIV treatment and RT-PCR with longer primers; lane 3,4, repeat of lanes 1 and 2 but using shorter RT-PCR primers. Lane 5, miR-101-1-3p cloned into Ad-NNRA3RA5NN-3'P with NlaIV treatment and RT-PCR with long primers * Unknown artefact.

Fig 4.19 shows the results for total RNA cloning from other cell lines HL60, H929, U2OS cell lines and miR-101-1-3p as a control. The PCR product of the vector only is labelled and is evident in lanes 1,2,4,6 and 8 and perhaps faintly in lane 10. This product was generated using the longer primers

illustrated in Fig 4.6 and is of the expected size. The band immediately above the vector is of the correct size to represent microRNAs from these cell lines cloned into the vector in all of these lanes. Repeat experiments with shorter primers for the RT-PCR step (lanes 3,5,7,9,11) were not successful.

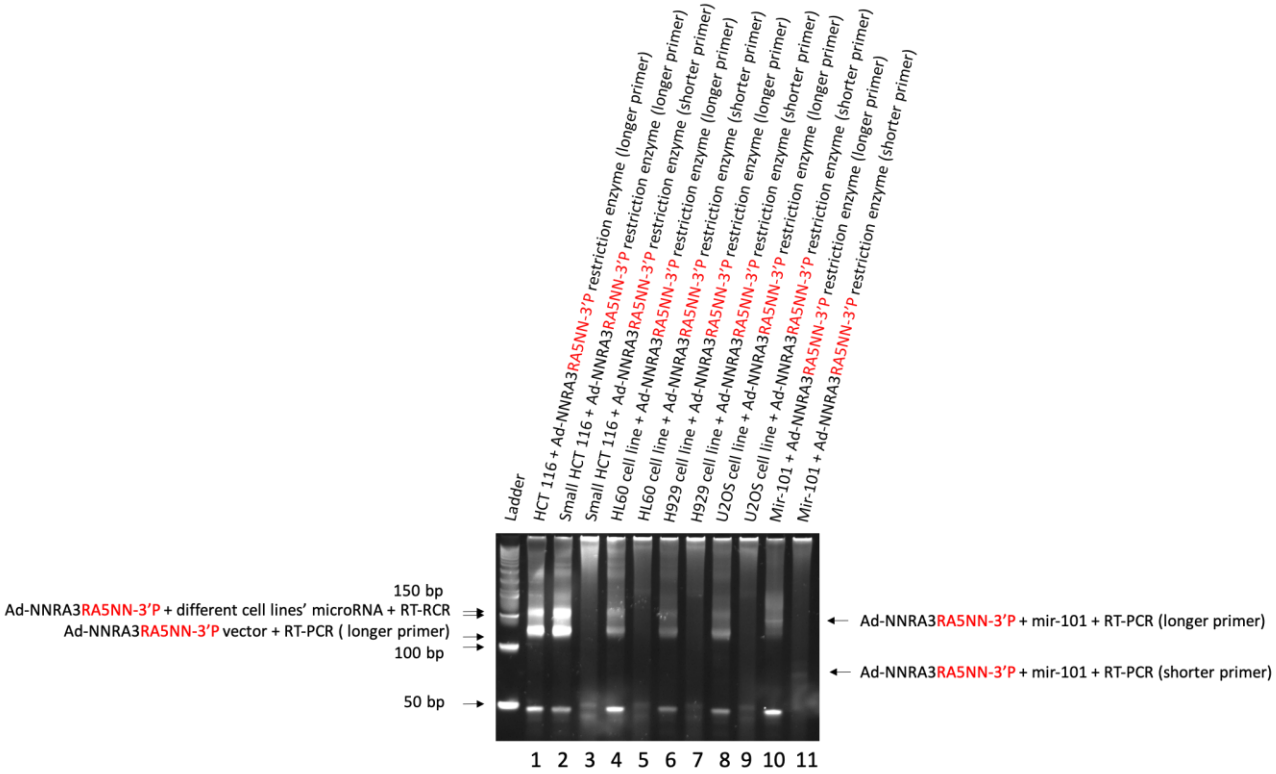


Figure 4.19 MicroRNA cloning from other cell lines. Lane 1, total RNA extract from HCT 116 cell line. Lanes 2, 3, small RNA extract from HCT 116 cell line. Lanes 4 to 9, total RNA from HL60, H929, and U2OS cell lines. Lanes 10, 11, miR-101-1-3p control. All RNA samples and miR-101-1-3p were cloned into Ad-NNRA3RA5NN-3'P as depicted in Fig 4.12 and analysed by RT-PCR for 25 cycles using longer or shorter primers. In all reactions 2 µg of total RNA was ligated to 150 ng of vector. All samples were run on a 12.5% non-denaturing PAGE gel.

Fig 4.20 shows a similar analysis of microRNA libraries made from cell lines including THP-1 and HL60.

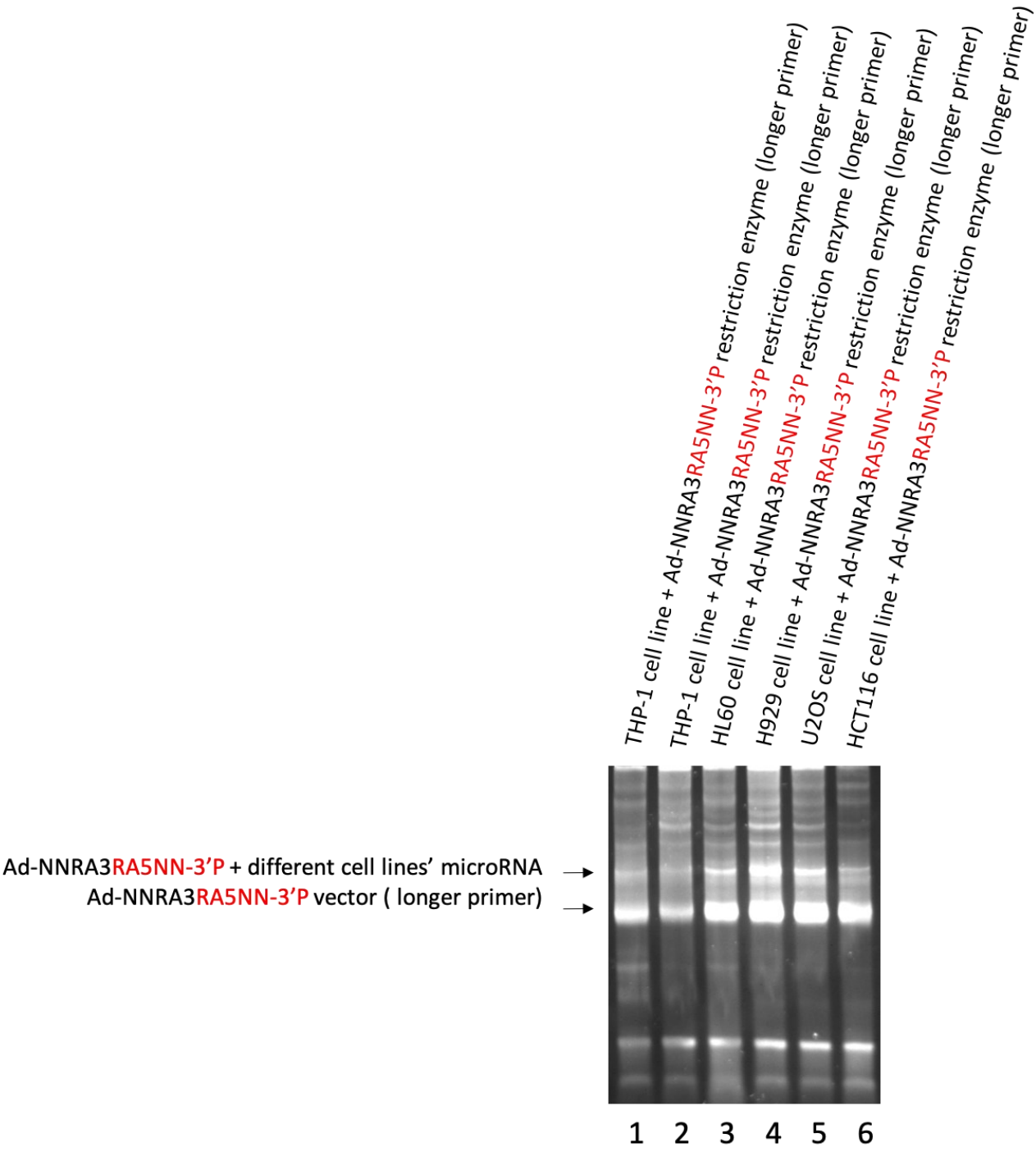


Figure 4.20 MicroRNA cloning from further cell lines. All the experiments with RT-PCR. Lane 1, THP-1 cell line total RNA (2 μg) ligated with Ad-NNRA3RA5NN-3'P vector (150 ng) with adding stop oligo (150 ng) and used NlaIV restriction enzyme treatment with longer primers by RT-PCR (20 cycles). Lane

2,3,4,5,6, the same repeat experiment with Ad-NNRA3RA5NN-3'P vector (50 ng) and THP-1, HL60, H929, U2OS, HCT116 cell lines' small RNA (1 µg).

4.2.11 Pulldowns of miR-575 and miR-101-1-3p from the indicated cell lines

The first three lanes of Fig 4.21 show the analysis of the microRNA reference library (see Fig 4.15) cloned into Ad-NNRA3RA5NN-3'P. In this case the library was amplified by RT-PCR using short primers (Fig 4.16), which would be expected to generate linear bands of 47 bases for RA3RA5-3'P and 20 to 30 bases more for microRNA clones. Subject to confirmation, the indicated bands in lanes 1 to 3 are likely products of Ad-NNRA3RA5NN-3'P (51 bases vector) and vector plus miRNA. Lanes 4 to 7 of Fig 4.21 show the results of a pulldown of miR575 and miR-101-1-3p from total RNA libraries made from the indicated cell lines. The libraries and pull downs were amplified by RT-PCR with both the shorter (Fig 4.16) and longer RP1 and RPI48 primers.

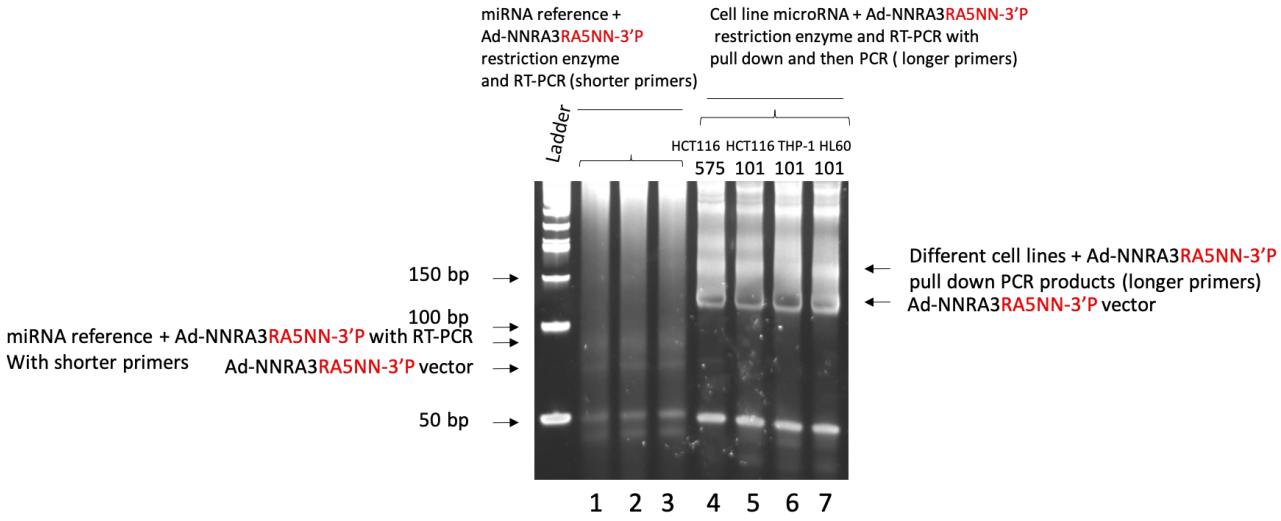


Figure 4.21 MicroRNA pulldown experiments. Lanes 1,2,3, are three repeat ligations of the test microRNA library into Ad-NNRA3RA5NN-3'P (Fig 4.12, 4.13) with shorter primers for RT-PCR (20 cycles). Lanes 4 to 7 show the results for pulldowns of miR-575 or mir101 made from total RNA (lanes

4,5) from the cell lines HCT 116 or total RNA (lanes 6,7) from THP-1 and HL60. Longer primers used for RT-PCR for 15 cycles. All samples were run on a 12.5% non-denaturing PAGE gel.

4.3 Discussion

In agreement with Somagenics (Barberán-Soler et al., 2018), we found that the 5' end of miR214 could be ligated efficiently to the 3' end of RA5 by intramolecular ligation (Fig 4.1), whereas intermolecular cloning of RA5 to mir214 was noticeably inefficient compared to other miRNAs (Fig 3.10). The protocol that we developed to clone miRNAs is illustrated in Fig 4.12. The circularisation step (step 5 of Fig 4.12) was suggested to us by Somagenics before they published or were prepared to reveal details of their protocol. Similarly and independently to Somagenics we decided upon RNA ligase 2 as the enzyme to catalyse circularisation, although there are alternatives (Table 4.1). We also came to the same conclusion regarding the 3' reversible modification step (Fig 4.12 step 4). The Somagenics protocol works and is available in kit form. Although we anticipated this, our reason for developing this protocol ourselves was to facilitate the introduction of additional steps such as a pulldown method for specific miRNAs (Fig 4.16) and to be able to introduce improvements. The Somagenics protocol has been tested against a range of other protocols that are available in kit form and was still found to show biased cloning although less so than the other tested methods (Wright et al., 2019, Herbert et al., 2020). There are perhaps some modifications to the Somagenics technique that if introduced might further decrease cloning bias (see General Discussion).

We were not able to prove that the method we developed works. Samples were sent for sequencing and were analysed by Dr Leandro Castellano, who has RNA sequencing expertise, but the only sequencing reads were of vector only. The samples were sequenced towards the end of the project and there was not the opportunity to repeat these experiments. The results shown in Figs 4.17 and 4.19 are perhaps the most convincing of successful miRNA cloning. However, a problem is that the samples analysed in Fig 4.17 would not have been suitable for sequencing as the shorter primers that

were used would not have annealed to the sequencing primers that are used. Making a good miRNA library is difficult and we may have had better results given more time and careful choice of samples.

There are some steps in our protocol that could be improved. The PNK step (Fig 4.12 step 4) produced variable results. Fig 4.3 indicates nearly complete removal of the 3'P group by PNK whereas only partial removal of the 3'P group is indicated by Figs 4.4 and 4.5. A careful titration of PNK concentration, amount of substrate and treatment time may help with this step. Although Somagenics have a diagram to illustrate their use of a 3'P block, the removal of this group is not mentioned at all in the text or methods (Barberán-Soler et al., 2018). The reversible terminators that are used for second generation sequencing would seem a good option for a reversible 3' block for NNRA3RA5NN-3'P but are only available for DNA (Professor Steven Benner, Harvard University, personal communication). We have some preliminary results to indicate that the O-Methyl attached to the 2' position of the 3' base of NNRA3RA5-2OMe inhibits RNA ligase 2 but not circligase II (Fig S9). However, we have reservations about using circligase for the circularisation step (see below).

We also had problems using circular constructs for PCR reactions, with the production of unexpected higher molecular weight products (Fig 4.6, 4.8). This wasn't a problem we encountered with PCR of linear DNA. This was a nuisance because it impeded the interpretation of our cloning protocol and gels. Although a solution was found (Figs 4.12 step 5 and 6), it would be better not to have this problem. The use of a different PCR enzyme, primers or reaction conditions could be investigated. Somagenics use the same method of RT-PCR to us, as first used by Illumina (Fig 4.6).

Circligase II is considered to be an ATP dependent enzyme but works without ATP addition. The reason for this is that ATP is required to adenylate the enzyme and this adenyl group is then transferred to the 5'PO4 end of the nucleic acid substrate (WoodSabatini and Hajduk,

2004). Circligase II is largely adenylated upon purchase and is therefore active without added ATP, which is consistent with the results of Table 4.1. This means that circligase II must be used in stoichiometric amounts in order to catalyse ligation. Circligase I is considered to be ATP dependent and is provided with an ATP solution in addition to the reaction buffer. Therefore, the results for Circligase I in table 4.1 were surprising. In our hands the enzyme worked without ATP addition and occasionally worked better without ATP (Table 4.1, Appendix 2). Possibly this was because the circligase I or ATP solutions were faulty. This was not something we pursued because circligase I is expensive and RNA ligase 2 was better at circularisation (Table 4.1).

Chapter 5-Background splicing and genetic disease

Background

During the course of my PhD a database of spliced mRNA was published (Wilks et al., 2018) that offered an ideal resource for improving a bioinformatics method previously devised by our lab for the detection of cryptic splice sites (Kapustin et al., 2011). This is a project to which I was well suited because of my previous bioinformatics work (see Chapter 3).

5.1 Introduction

Splicing mutations cause 15 to 25% of human genetic disease (Scotti and Swanson, 2016, Baralle and Buratti, 2017). Most of these mutations disrupt intron splice sites (Fig 5.1) or generate de novo splice sites. De novo splice sites are new splice sites that are created directly by the mutation,

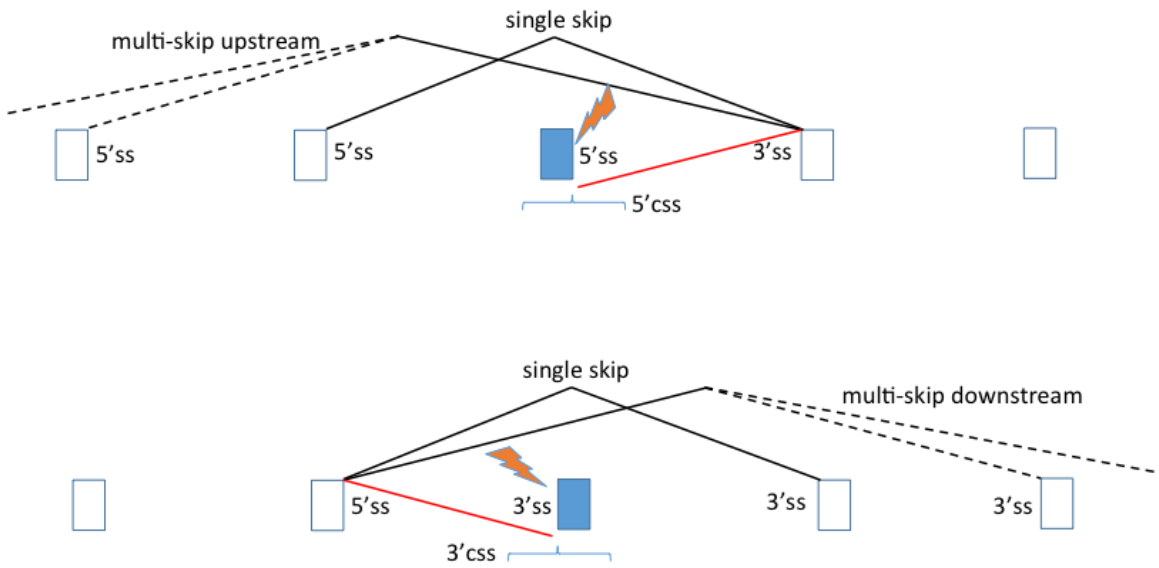


Figure 5.1 Aberrant mRNA splicing events that are usually activated by mutations of the 5' or 3'ss of introns. A. The upper panel illustrates that mutation of the 5'ss of an intron typically causes the non-mutated partner 3'ss to splice instead with a 5'css or to splice with upstream intron 5'ss, which causes single or multiple exons skipping. The lower panel of Fig 5.1B shows 3'ss mutations typically cause the partner 5'ss to splice instead with cryptic 3'ss or with downstream intronic 3'ss, which results in single or multiple exon skipping. The brackets indicate that the large majority of cryptic splice site (css) are located within 1000 bases of the mutated 5' or 3' intron splice site.

which means that the mutation is part of the new splice site (Buratti et al., 2011). Some de novo splice site mutations also lead to the activation of pseudoexons (Fig 5.2A, B).

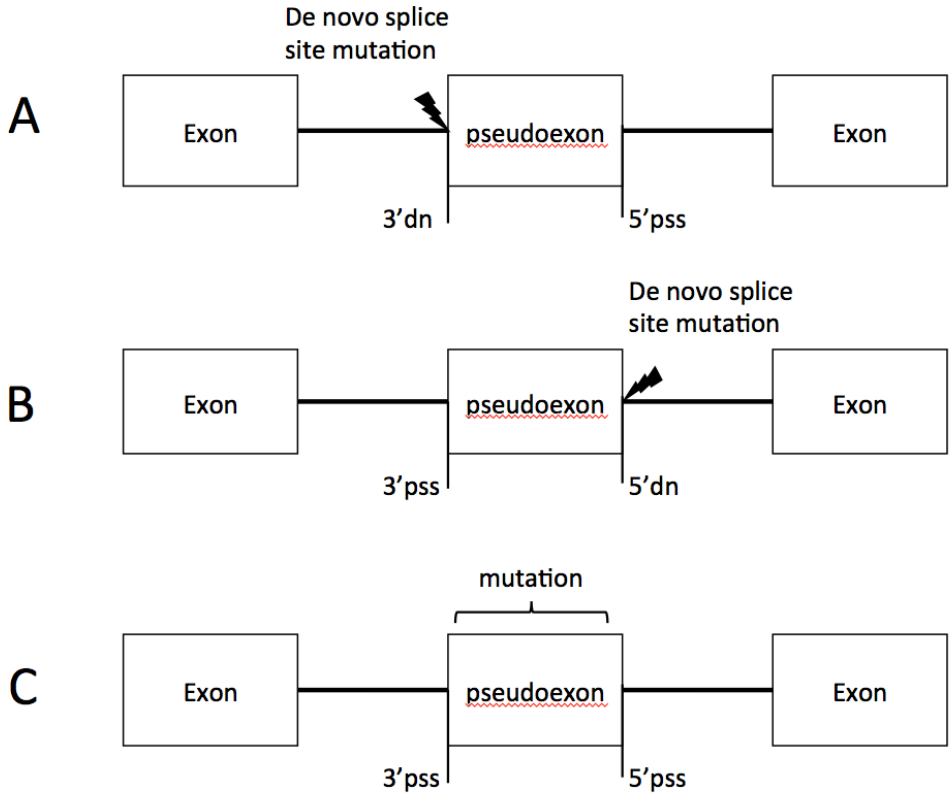


Figure 5.2 Illustrates three mutation types that can generate pseudoexons (Alexieva et al., 2022). Figure 5.2A shows that a 3' de novo splice site mutation may also activate a downstream 5' pseudo splice site (pss) to create a pseudoexon on the other site. Similarly, 2B shows that a 5' de novo splice site mutation can activate an upstream 3' pss. 2C illustrates that mutations other than de novo splice site mutations can also create pseudoexons, usually these mutations disrupt or create splicing auxiliary sequences that lie within the pseudoexon.

Fig 5.2C illustrates the effect of another class of splice site mutations that generate pseudoexons through the disruption of splicing auxiliary motifs. These motifs are present in both introns and exons and are binding sites for splicing proteins that may enhance or suppress the recognition of splice sites (Culler et al., 2010).

Recently mutations of the splicing machinery itself have been discovered to cause leukemia and other cancers (Yoshida et al., 2011, Darman et al., 2015, DeBoever et al., 2015, Ilagan et al., 2015, Zhang et al., 2015, Suzuki et al., 2019).

A number of *in silico* programs have been developed to analyse patient DNA sequences for mutations that might cause aberrant splicing (JianBoerwinkle and Liu, 2014, Moles-Fernandez et al., 2018, OhnoTakeda and Masuda, 2018, Alvarez et al., 2021, DawesJoshi and Cooper, 2022). This is a valuable goal because it is often very difficult to analyse patient RNA samples for possible splicing defects (JianBoerwinkle and Liu, 2014, Baralle and Buratti, 2017, Moles-Fernandez et al., 2018, Alvarez et al., 2021). It is important to know if a patient has a splicing mutation in order to make a correct diagnosis and subsequently to consider splicing therapy. There is a particular need to identify mutations outside splice sites that might disrupt splicing, understandably such mutations are the most difficult to identify with *in silico* methods. The problem faced by *in silico* models is that splicing auxiliary motifs do not have strong consensus sequences and consequently there are many confounding false positives in any DNA sequence (Culler et al., 2010). Furthermore, *in silico* methods seem better suited to predicting whether a variant of unknown significance (vus) is likely to disrupt splicing rather than predicting the exact effect of the mutation (JianBoerwinkle and Liu, 2014, Baralle and Buratti, 2017, Moles-Fernandez et al., 2018, DawesJoshi and Cooper, 2022).

Our lab established that css can be identified by a bioinformatics approach that is based on our finding that css are already active in normal genes, albeit at very low levels (Kapustin et al., 2011). This was established by comparing the position of rarely used splice sites with known css that are activated in human disease. However, this approach was limited to a minority of genes for which there was sufficient data from expressed sequence tags (ESTs). Since that time a large amount of RNA-sequencing data has been deposited, which should in theory strongly increase the power of css prediction.

The Snaptron database is a very large database of spliced RNA-seq reads made from over 70,000 human samples (Wilks et al., 2018). This database was made by first identifying RNA transcripts with deleted regions through alignment to the genome sequence. Not all deletions are generated by splicing, therefore a filter was introduced whereby the 5' end of the deletion must match a GT or GC and the 3'end of the deletion must match an AG dinucleotide. These are the most common dinucleotides by far at the ends of human introns.

As expected, the Snaptron database has the highest reads for RNA transcripts that have deletions due to intron removal or to alternative splicing. However, the Snaptron database has far more splicing events with very low reads (which we call background splicing) than splicing events with the high reads necessary for intron removal or functional alternative splicing. The Snaptron database shows that there are three main types of background splicing: splicing between 5' and 3' background splice sites (bss) located throughout exons and introns; low level exon skipping between normal intron splice sites and low level splicing between bss and intron ss. The last two categories of background splicing are illustrated in Fig 5.1.

BRCA1 and BRCA2

BRCA1 and BRCA2 are DNA repair genes (Patel et al., 1998, Moynahan et al., 1999) that are expressed in most normal tissues but were originally identified as breast cancer susceptibility genes by genetic linkage studies of familial breast and ovarian cancer (Black, 1994). Heterozygous mutations of just one of the two alleles for BRCA1 are strongly predisposing to cancer (Konishi et al., 2011) and similarly for BRCA2 (Warren et al., 2003). Consequently BRCA1 and BRCA2 have been extensively screened and tested for possible splicing mutations to an exacting standard (Whiley et al., 2014) and so provide extensive and reliable experimental data for comparison with the snaptron database (see below).

Aim

To establish whether the more extensive RNA splicing data that is now available can be used to further improve the approach we developed to detect cryptic splice sites (Kapustin et al., 2011), with emphasis on BRCA1 and BRCA2. We will also test whether RNA splicing databases can be used to predict the effect of splice site mutations upon exon skipping.

5.2 Results

Experimental reports of mutations that cause aberrant splicing of BRCA1 and BRCA2 were obtained from the database of aberrant splice sites (DBASS) the human genome mutation database (HGMD), the Leiden Open Variation Database online (LOVD) and by searching Pubmed (Buratti et al., 2011, Fokkema et al., 2011, Stenson and Ciorba, 2020) and see Materials and methods.

5.2.1 Mutated 5'ss or 3'ss cause cryptic splice site activation and/or exon skipping

The top panel of Fig 5.1 shows that mutation of a 5'ss of an intron usually activates a nearby 5'css in place of the mutated 5'ss or causes exon skipping - where the 5'ss of an upstream exon splice with the 3'ss partner of the mutated 5'ss. There is a similar effect caused by mutations of the 3'ss of an intron (Fig 5.1 bottom panel). It is difficult to predict the effect of splice site mutations by current *in silico* methods (JianBoerwinkle and Liu, 2014, Baralle and Buratti, 2017, Moles-Fernandez et al., 2018, Alvarez et al., 2021, DawesJoshi and Cooper, 2022).

Table 5.1A illustrates a possible method to predict the precise effect of intron splice site mutations by using large database of spliced RNA.

	Chr	5'ss	3'ss	intron size (bp)	Distance of 5' bss from 5'ss 41222944	Reads
1	chr17	41277293	41219713	57581	-54349	1
2	chr17	41234420	41219713	14708	-11476	1
3	chr17	41228504	41219713	8792	-5560	5
4	chr17	41223037	41219713	3325	-93	2
5	chr17	41222944	41219713	3232	0	148299
6	chr17	41222938	41219713	3226	6	4
7	chr17	41222879	41219713	3167	65	243
8	chr17	41222875	41219713	3163	69	78
9	chr17	41222820	41219713	3108	124	1
10	chr17	41221068	41219713	1356	1876	2
11	chr17	41221056	41219713	1344	1888	5
12	chr17	41219800	41219713	88	3144	1

	Chr	5'ss	3'ss	intron size (bp)	Distance of 3'bss From 3'ss 41203135	Reads
1	chr17	41209068	41208322	747	-5187	1
2	chr17	41209068	41208085	984	-4950	1
3	chr17	41209068	41206426	2643	-3291	21
4	chr17	41209068	41205984	3085	-2849	9
5	chr17	41209068	41203468	5601	-333	7
6	chr17	41209068	41203321	5748	-186	2
7	chr17	41209068	41203186	5883	-51	24
8	chr17	41209068	41203135	5934	0	153347
9	chr17	41209068	41203127	5942	8	206
10	chr17	41209068	41202208	6861	927	9
11	chr17	41209068	41201212	7857	1923	1298
12	chr17	41209068	41201189	7880	1946	1
13	chr17	41209068	41199721	9348	3414	26
14	chr17	41209068	41197820	11249	5315	13
15	chr17	41209068	41197749	11320	5386	1
16	chr17	41209068	41180702	28367	22433	1
17	chr17	41209068	41170817	38252	32318	1
18	chr17	41209068	40851396	357673	351739	1

Table 5.1 Snaptron splicing data for BRCA1. (I contributed to this Table). A. The Snaptron splicing data shows all splicing events involving the 3'ss 41219713 (hg 19) of intron 16 of the wild type BRCA1 on chromosome 17. As expected, the splice reads to its 5'ss partner 41222944 are highest (148299, blue shading). Low level background splicing to the 5'ss of upstream exons are shaded yellow and row 4 shows background splicing (2 reads) to an exonic 5'ss 93 bases upstream from the normal 5'ss. Rows 6 to 12 show background splicing between 3'ss 41219713 and 5'ss at the indicated positions within the intron. B. All of the splicing events involving the 5'ss (41209068) of intron 20 of BRCA1. Blue and yellow shading as above, rows 1 to 7 show background splicing within the intron and remaining rows are splicing events between the 5'ss 41209068 downstream of its normal 3'ss partner (41203135).

In this example we examine a mutation of the 5'ss 41222944 (shaded in red) of intron 16 of BRCA1 which was first reported to activate a 5' css at position +69 (Scholl et al., 1999) and subsequently at position +65 by other groups (Wappenschmidt et al., 2012, Colombo et al., 2013, Baert et al., 2017). Table 5.1A lists all of the splicing events involving the normal 3'ss splice site partner (3'ss 41219713) of the mutated 5'ss, as illustrated in Fig 5.1 top panel. This data is from the Snaptron spliced RNA database and is for wild type BRCA1. The row shaded in blue shows that as expected there are a large number of reads (148299) for splicing between 3'ss 41219713 and its normal 5'ss partner 41222944. The yellow shading (rows 1 to 3) shows that there are low numbers of background reads (5, 1 and 1) for splicing between 3'ss 41219713 and the 5'ss of upstream introns. Row 4 shows that there are 2 reads for splicing between 3'ss 41219713 and a background 5'ss splice site (bss) at 41223037, 93 bases upstream from the normal intron 5'ss. There are also reads for splicing between 3'ss 41219713 and 7 different bss downstream of the normal 5'ss, ie within the intron (rows 6 to 12). It can be seen that the bss with the most reads of 243 and 78 (shaded grey) exactly match the css at +65 and +69 that known to be activated by mutation of the 5'ss 41222944 (see above). The relatively high reads for the +65 and +69 bss perhaps indicates why both css were identified (Scholl et al., 1999) and (Wappenschmidt et al., 2012, Colombo et al., 2013, Baert et al., 2017).

Similarly, Table 5.1B compares the known effect of mutation of the 3'ss 41203135 (hg19) of intron 20 of BRCA1 with the background splicing events involving its normal 5'ss partner 41209068. The mutation of 3'ss 41203135 is known to activate single exon skipping between the 5'ss 41209068 and the 3'ss 41201212 in the downstream intron plus weaker activation of a 3'css 41203127 located at +8 bases (Wappenschmidt et al., 2012, Colombo et al., 2013). These two aberrant splicing events also have the most background splicing reads of 1298 and 206 in Table 5.1B. Overall, Table 5.1 shows that potential aberrant splicing events (such as those illustrated in Figure 5.1) already occur at very low background levels in normal BRCA1 and that the aberrant splicing events that are activated by splice site mutations may use the most active bss.

5.2.2 The effects of all of the splice site mutations of BRCA1

Table 5.2 lists the effects of all of the splice site mutations of BRCA1 that we could find (columns 1 and 2, materials and methods) and compares these to background splicing data from Snaptron (columns 3 to 8). The key background splicing data from Table 5.1A and B is listed in rows 13 and 32 and from this it can be seen how the numbers in columns 3 to 8 were obtained. We identified 17 css (Table 5.2, column 2) that are activated by either 5' or 3' ss mutations of BRCA1 splice sites in patients and 15 of these css exactly match bss obtained from the Snaptron RNA splicing data for normal BRAC1 (column 3), the two exceptions in column 3 are shaded and discussed in Table S1 of Appendix 3. Twelve of the 15 bss that match css had the highest reads of any bss within 1000 bases of the mutated splice site (column 4).

Table 5.2 BRCA1

	1	2	3	4	5	6	7	8
	Mutated splice site 5' ss	Experimental summary	css match in Snaptron	Snaptron css rank	Snaptron css reads	Snaptron top bss reads	Snaptron single skip reads	Snaptron double skip reads
1	41276033 (exon 2)	skip				11	898	0
2	41267742 (exon 3)	skip				24	2622	104
3	41258472 (exon 5)	Css(-22) & skip	Yes	1(4)	7375	7375	3629	178
4	41256884 (exon 6)	Css (-9)	Yes	1(5)	327	327	5	6
5	41256138 (exon 7)	Css (-62)	No	0(0)	0	0	3	2
6	41251791 (exon 8)	skip				0	83	0
7	41249260 (exon 9)	skip				31	2193	3
8	41243451 (exon 11)	skip (and alt ss enhancement)				2	346	44

9	41242960 (exon 12)	skip				115	1	18
10	41234420 (exon 13)	skip				0	164	0
11	41228504 (exon 14)	single and double skip (weak)				1	971	58
12	41226347 (exon 15)	Css (-11), single and double skip	Yes	1(1)	494	494	476	356
13	41222944 (exon 16)	Css (65,69)	Yes, yes	1(5), 2(5)	243, 78	243, 78	0	5
14	41219624 (exon 17)	Css (153, weak) & skip	Yes	1(3)	36	36	1030	1
15	41215890 (exon 18)	skip				179	140	17
16	41215349 (exon 19)	skip				23	1	56
17	41209068 (exon 20)	Css (87) & skip	Yes	1(1)	19	19	52	0
18	41203079 (exon 21)	skip				4	1298	0
19	41201137 (exon 22)	Css (156, weak) & skip	Yes	1(1)	6	6	2732	26
20	41199659 (exon 23)	Css (5, weak) & skip	Yes	1(2)	247	247	300	132

3’ss

21	41267797 (exon 3)	Css (7)	Yes	1(3)	5	5	2622	178
22	41258551 (exon 5)	skip				18	3629	6
23	41256974 (exon 6)	Css (-59)	Yes	1(4)	15	15	5	2
24	41256279 (exon 7)	Css (-10)	No	0(1)	0	81	3	0
25	41251898 (exon 8)	Css (-69)	Yes	1(1)	4	4	83	3
26	41247940 (exon 10)	skip			339	339	840	44
27	41246878 (exon 11)	skip & alt skip			3	3	206,364	18
28	41219713 (exon 17)	skip			12	12	1030	17
29	41215969 (exon 18)	skip			11	11	140	56
30	41215391 (exon 19)	skip			1	1	1	0

31	41209153 (exon 20)	Css (13, weak) & strong skip	Yes	2(2)	1	3	52	0
32	41203135 (exon 21)	Css (8, weak) & skip	Yes	1(5)	206	206	1298	26
33	41201212 (exon 22)	skip			1147	1147	2732	132
34	41199721 (exon 23)	skip			26	26	300	0
35	41197820 (exon 24)	Css (11)	Yes	2(4)	5	26	0	0

Table 5.2 Comparision of the effect of BRCA1 splice site mutations with background splicing sites from Snaptron. (I contributed to this Table). Column 1 List of previously analysed BRCA1 5'splice sites and 3'splice site mutations. Column 2 Experimental of css activation and/or exon skipping caused by the splice site mutations. Column 3 Match between css and bss. Column 4 Relative reads of the bss that matches the css compared to other bss, where 1(4) means there are four bss within 1000 bases of the mutated ss and the bss that matches the css has the most reads of these four. Column 5 Reads for the bss that matches the css. Column 6 Gives the reads of the top bss, most of these matches the css. Columns 7,8. Reads for single and double exon skips. The shaded boxes indicate discrepancies between the experimental and Snaptron data that are discussed further in Table S1 of 2. Table S1 also gives the references for the experimental data in column 2 of Table 5.2.

19 of the 35 splice site mutations of BRCA1 in Table 5.2 activate exon skipping rather than css and eight of the splice site mutations do both (Table 5.2, column 2). We noticed that mutations that cause skipping but not css activation tend to have higher background reads for exon skipping than for any bss within 1000 bases of the splice site mutation (compare columns 6 and 7).

The background splicing reads that do not agree with the experimental results are shaded in columns 6 & 7. There are four examples (rows 9, 15, 24 and 35), where the highest background reads are for candidate css that nevertheless were not activated as css, for rows 15 and 24 these experimental results were repeated by different groups (Table S1, Appendix 3). Rows 9, 15, 24 and 35 are examples of likely false positives (see Discussion).

The shading in rows 16, 20 and 25 also highlight differences between predictions by background splicing reads and the experimental results. For row 25, the primers that were used to detect the css at -69 would not have detected the possible double exon skip indicated by the background splicing reads. Similarly, there are reasons to believe that further experimentation might resolve differences with the background splicing reads in rows 16 and 20 (Table S1, Appendix 3).

The results for Table 5.2 row 5 have been established by repeated experiment and so the background splicing reads falsely predicts single exon skipping over the activation of a css at -62. The background splicing reads for row 31 are largely in agreement with the major experimental result of exon skipping as a result of the indicated splice site mutation. There is a minor disagreement in that background splicing favours the activation of a weak css at position -474 rather than +11 (Table S1, Appendix 3).

5.2.3 A similar analysis of BRCA2

Table 5.3 shows a similar analysis of BRCA2. We found 18 reports of css activation splice site mutations of BRCA2 (column 3) and 14 of these matched bss. Of these 14, eight had the highest reads of all bss within 1000 bases of the splice site mutation (column 4). There are 24 reports of mutations that cause mainly single exon skipping (rows 2-5, 8,9,12-17, 19-27, 29, 33, 35, 37) and Snaptron shows that 22 of the 25 have the highest background reads for a single exon skip than for any other background splicing event.

Table 5.3

BRCA2	1	2	3	4	5	6	7	8

	Mutated splice site 5’ ss	Experimental summary	Snaptron match to css	Snaptron css rank	Snaptron css reads	Snaptron top bss reads	Snaptron single skip reads	Snaptron double skip reads
1	32889805 (exon 1)	Css (-99)	Yes	7(10)	5	763	0	0
2	32890665 (exon 2)	skip				133	37	0
3	32893463 (exon 3)	skip				2	1858	8
4	32899322 (exon 4)	skip				59	131	101
5	32900288 (exon 5)	skip				7	424	35
6	32900420 (exon 6)	skip and double skip				0	4	62
7	32900751 (exon 7)	skip and quadruple skip				1	56	22(double), 144(triple) 194 (quadruple)
8	32903630 (exon 8)	skip				12	19	1
9	32915334 (exon 11)	skip				1	195	1
10	32921034 (exon 13)	skip and double skip				0	14	64
11	32929426 (exon 14)	Css (5)	No	0(1)	0	21	9	0
12	32930747 (exon 15)	skip				86	55	0
13	32932067 (exon 16)	Css (-100 weak) and skip	Yes	1(2)	2	2	6	5
14	32936831 (exon 17)	skip				1	15	0
15	32937671 (exon 18)	skip and css (-991 weak)	Yes	3(3)	1	149	431	390
16	32944695 (exon 19)	skip				1	742	24
17	32945238 (exon 20)	skip				2	61	6

18	32950929 (exon 21)	Css(46)	Yes	1(2)	141	141	2	0
19	32954051 (exon 23)	skip				1	34	0
20	32954283 (exon 24)	skip				1	0	0 (double), 2(triple)
21	32969071 (exon 25)	skip				1	3	0
Mutated splice site 3' ss								
22	32890558 (exon 2)	skip and double skip (minor)				3	37	8
23	32893213 (exon 3)	skip				9	1858	101
24	32900237 (exon 5)	skip				4	424	62
25	32900378 (exon 6)	skip				2	5	23
26	32900635 (exon 7)	skip				1	56	1
27	32903579 (exon 8)	skip				41	19	4
28	32905055 (exon 9)	Css (73) and skip (minor)	Yes	1(3)	29	29	4	220
29	32928997 (exon 14)	Css (246 weak) and skip	Yes	1(2)	1	1	9	0
30	32930564 (exon 15)	Css (13)	Yes	5(9)	7	37	55	5
31	32931878 (exon 16)	Css (44)	Yes	2(5)	46	189	6	0
32	32936659 (exon 17)	Css (20,69), skip	No, Yes	0(2), 1(2)	0, 23	23	15	390
33	32937315 (exon 18)	skip				4	431	24
34	32944538 (exon 19)	Css (14)	Yes	2(3)	11	18	742	6
35	32945092 (exon 20)	Css(12 minor) and skip	No	0(2)	0	1	61	0

36	32950806 (exon 21)	Css (-43) and insertion	No	0(0)	0	0	2	1
37	32953453 (exon 22)	css (484, weak) and skip	Yes	1(6)	146	146	719	0
38	32953886 (exon 23)	Css (51) and skip (weak)	Yes	1(7)	1223	1223	34	0
39	32954143 (exon 24)	Css (7)	Yes	2(3)	11	47	0	0
40	32968825 (exon 25)	Css(27) and skip	Yes	1(1)	2	2	3	0

Table 5.3 Comparision of the effect of BRCA2 splice site mutations with background splicing sites from Snaptron. (I contributed to this Table). Columns and shading as in Table 5.2.

5.2.4 Analysis of cryptic splice sites (css) that cause a wide range of medical syndromes

In order to expand this analysis we next compared the css that are listed in the DBASS database of aberrant splice sites (Buratti et al., 2011) to Snaptron background splicing data.

Aberrant ss	DBASS5	DBASS3
css	**459**	**182**
Unusual css	**13**	**31**
De novo ss created	**34**	**123**
De novo ss enhanced	**95**	**11**
Pseudoexon	**71**	**14**
Pseudoexon unusual	**14**	**6**

Table 5.4 Summary of the numbers and different types of aberrant splicing events listed in DBASS. (Courtesy of Dr N Dibb). Css are activated by splice site mutations and usually lie within 1000 bases of the mutated ss. DBASS5 lists 5'css that are activated by 5'ss mutations and DBASS3 lists 3'css that are activated by 3'ss mutations, see Figure 5.1. Unusual css means that the css is activated by a less common splice site mutation that lies outside the core consensus sequence of the splice site. De novo splice sites are new splice sites that are created directly by the mutation, so that the mutation

is part of the new splice site. We have subdivided these into mutations that create the GT or GC of a new 5'ss (or the AG of a new 3'ss) and into mutations that enhance already existing but dormant GT or GC 5' de novo splice sites or AG 3' de novo splice sites. De novo splice site mutations can also lead to the activation of pseudoexons (Fig 5.2A, B) but unusual pseudoexons are not created by de novo ss mutations but by mutations that usually affect auxiliary splicing motifs (Fig 5.2C).

Table 5.4 shows that DBASS lists 459 5'ss mutations that activate 5' css and these are known to cause 199 different medical syndromes (Alexieva et al., 2022). We systematically chose and analysed the first listed 5'ss mutation out of the 459 listed in DBASS5 (Table 5.4) that cause each of the 199 medical syndromes similarly we chose the first 3'ss mutations responsible for the 99 medical syndromes listed in DBASS3 (Alexieva et al., 2022). Because DBASS entries have been made chronologically a range of intron ss mutations at different gene positions were covered. Table 5.5 shows that 201 out of 237 experimentally identified 5'css (some of the 199 mutations generated more than one css) matched bss and that 150 of these matched bss with the most reads ($p = 1x\ 10^{-56}$, see Materials and methods). Similar results were found for 3'css listed in DBASS3, where 62 out of 110 css matched bss with the most reads ($p = 3.2 \times 10^{-23}$). The reason why approximately 15% of 5'css and 12% of 3'css did not match bss (Table 5.5) was usually because there were no bss for comparison. Where bss data was available, we found that css did not match a bss in only 2-3 % of cases, listed under poor match in Table 5.5. The match between BRCA css and bss is also summarised in Table 5.5. DBASSw refers to a category of splice site mutations in DBASS5 or DBASS3

Table 5.5

5'css	No. of css analysed	No. that match snaptron bss	Top match	Poor match	Source
DBASS5	237	201 (85%)	150 (75%)	9	(Alexieva et al., 2022)
DBASS5w	14	11	10	2	(Alexieva et al., 2022)
BRCA1	10	9	8	1	Table 5.2

BRCA2	5	4	2	1	Table 5.3
3'css					
DBASS3	110	97 (88%)	62 (64%)	2	(Alexieva et al., 2022)
DBASS3w	39	38	31	0	(Alexieva et al., 2022)
BRCA1	7	6	4	1	Table 5.2
BRCA2	13	10	6	2	Table 5.3

Table 5.5 General match between bss and css.

DBASS5 summarises the match of between 237 css that are activated by 199 mutations of the 5'ss of different genes responsible each medical syndrome listed in DBASS5. DBASS3 summarised a similar anlaysis of 110 css activated by 99 different 3'ss mutations. DBASS5w and DBASS3w refer to splice site mutations in DBASS that lie just outside the core 5' or 3' splice site consensus. BRCA1 and BRCA2 show summaries of Tables 5.2 and 5.3.

that lie outside the core region of the splice site (see Table 5.4) but still activate css. These unusual and presumably weaker ss mutations activate css that match bss with particularly high reads (Alexieva et al., 2022). Overall, this analysis shows that DBASS 5' and 3' css match bss in 85 and 88% of cases and that usually the css matches the bss with the most reads.

5.2.5 Cryptic splice sites versus exon skipping

Table 5.6 is from Alexieva et al (2022), to which I contributed, and compares background splicing information for the DBASS database of css and an exon skipping database (Divina et al., 2009, Buratti et al., 2011).

Table 5.6

	A Experimental results	B Snaptron data skip>css reads	C css>skip reads	D Total css reads	E Total skip reads
	DBASS5				
1	70 css only	11	59	105757	6884
2	36 css + skip	23	13	10112	143955
	DBASS3				
3	18 css only	2	16	26791	2659
4	22 css + skip	11	11	15839	31527
	5' skip database				
5	79 skip only	71	8	5978	217587
6	3 skip + css	1	2	9395	2852
	3' skip database				
7	65 skip only	54	11	17346	349439
8	4 skip + css	2	2	1939	3185

Table 5.6 Cryptic splice site (css) activation versus exon skipping. (I contributed to this Table of original data). The experimental results listed in column A are from the DBASS css database and an exon skip database (Divina et al., 2009, Buratti et al., 2011) and they show the numbers of reports of css activation only, exon skipping only or both in response to 5' or 3'ss mutations. Columns B and C are from Snaptron and show how the samples divide with respect to the relative number of reads for single exon skipping versus the number of reads for the bss that matches the css. For examples that do not report a css or more rarely report a css that does not match a bss we used the read numbers of the top bss (bss with the most reads within 1000 bases of the mutated ss). Columns D and E are from Snaptron and show the total css and single exon skip read count and shaded examples are discussed (see text).

Table 5.6 row 1 shows that for the 70 reports of css activation only in response to a splice site mutation, the reads for the css are greater than the reads for single exon skipping in 59 out of 70 cases. By contrast Table 5.6 row 5 shows that for the 79 reports of exon skipping only in response to a splice site mutation the background reads for single exon skipping are greater than the reads for the top bss in 71 out of 79 cases. The probability of obtaining these contrasting ratios by chance is p

= 6 x 10^{-19}, see Materials and methods. The total number of reads listed in columns D and E are also strongly supportive of this finding. Similar results are seen for splice site mutations that activate 3'css or 3' exon skipping only (rows 3 and 7, p = 1.4 x 10^{-8}).

Overall Table 5.6 shows that when the background reads for single exon skipping are greater than the background reads for any candidate css, then exon skipping usually occurs in response to a splice site mutation and so confirms this initial indication from the BRCA1 and BRCA2 analysis (Tables 5.2, 5.3). The exceptions to this general finding are shaded in Table 5.6.

5.2.6 Multiple exon skipping

Alexieva et al (2022) also compiled a list of all experimental reports of multiple exon skipping from the DBASS and exon skip database, these reports were compared to background splicing data and are reproduced in Table 5.7.

Table 5.7

A	B	C	D	E	F	G	H
Splice site mutations that cause multiple exon skipping			Snaptron data				
			Single skip reads	Double skip reads	css	css match	css reads
	Gene	Experimental effect					
1	LAMP2A	Single and double exon skip (similar ratio).	64	80	no		
2	LAMP2B	weak css and strong single exon skip, no double skip	8	0	yes	×	0
3	LAMP2C	weak single exon skip and strong double exon skip	1	25	no		
4	p67-PHOX	css, single and double skip, relative ratios not given	26	1	yes	✓	3
5	PKLR	css, single (major event) and double skips	0	0	yes	×	0
6	ATP7A	css, single (major event) and double skips	340	84	yes	✓	24

7	COL5A1	css (x2, weakest), exon skip, double exon skip (major)	1	60	yes	✓✓	10,6
8	HPRT1	css (20%), exon skip (60%), double skip (20%)	26	914	yes	✓	410
9	ALDH3A2	Single and double exon skip (strongest)	748	4364	no		
10	ATM	Single skip (90%) and double skip (10%)	734	56	no		
11	CAPN3	Double exon skip reported (single skip unclear)	8	2	no		
12	ECHA	Single (major) and double exon skip (minor)	22	285	no		
13	NTRK1	Single (stronger) and double exon skip.	9	42	no		
14	SEDL	single and double exon skip (ratio not clear).	194	135	no		
15	WT1	Single and double exon skip (similar amounts).	0	6	no		
16	ALDH3A2	Single and double exon skip. Ratio not given	6111	1536	no		
17	ATM	Single and double exon skip. Ratio not given.	1052	683	no		
18	BTK	Triple exon skip and css only	1	11, 27 (triple)	yes	✓	66
19	KCNQ1	Double exon skip only.	0	295	no		
20	BRCA1	Single and double skip (weak)	971	58	no		
21	BRCA1	css, single (major events) and double skip (minor)	476	356	yes	✓	494
22	BRCA2	Single and double skip (major effect for 1 of 2 reports)	4	62	no		
23	BRCA2	Single and quadruple skip	56	22, 144 (triple), 194 (quad)	no		
24	BRCA2	Single and double skip (major effect)	14	64	no		
25	BRCA2	Single and double skip (minor)	37	8	no		
26	DMD	double skip only	7	83	no		
27	DMD	skip and double skip (ratio not given)	4	67	no		
28	DMD	skip and double skip (ratio not given)	19	0	no		
29	DMD	css, skip (strongest) and double skip (weakest)	92	1	yes	✓	12
30	DMD	skip, double skip, triple skip (ratio not clear)	0	21, 6 (triple)	no		
31	DMD	css, skip, double skip (ratio not clear)	13	1	yes	✓	1
32	SLC35A1	css (major) single skip, double skip (weakest)	2919	650	yes	✓	26028
33	FGA	multiple css reported but no single or double exon skipping	1	11	yes	✓×✓	1,0,3
34	COL5A1	Single and double exon skips not reported	10	27	yes	×	0
35	STK11	only css reported	0	2058	yes	✓✓	79,56
36	COL7A1	only a css reported	895	494	yes	✓	151
37	FBN1	Single exon skip only	9	55	no		
38	BRCA1	css at -62 reported but not single or double exon skipping	3	2	yes	×	0
39	BRCA2	css and a single exon skip reported but not a double skip	15	390	yes	×✓	0,23
40	DMD	single exon skip reported	0	8	no		

41	DMD	single exon skip reported but not a triple skip	3	0, 10(triple)	no		
42	DMD	css and single exon skip but not a double skip reported	11	21	yes	✓	1

Table 5.7 Multi-exon skipping events (Alexieva et al., 2022). (I contributed to this Table). Experimental reports of mutations that cause multi-exon skipping compared to background splicing predictions. Genes are listed in column B and the experimental results are listed in column C and also column F. Snaptron data is compared in columns D, E and G and H. For shading see text.

Rows 1 to 3 of Table 5.7 concern proteins LAMP2A, B and C which are generated by alternative splicing from a common 5'ss and three alternative 3'ss (Di Blasi et al., 2008). The authors report that mutation of the common 5'ss has different effects upon single or double exon skipping by each 3' alternative ss (column C). These differences in skipping correlate well with the relevant background splicing reads (columns D and E). Other matches in Table 5.7 include reports of double exon skips only (rows 19 and 26) or mainly double exon skipping (rows 3, 7, 9, 22 and 24) and how these correlate with higher background reads for double skips than for single exon skips in Snaptron. Similarly, the reports of css activation and triple exon skipping (row 18) are a good match to the background splicing reads, and single and quadruple exon skipping (row 23) are a good match to the background splicing reads.

There are ten examples (rows 33 to 42) in Table 5.7 where the experimental results do not match the multiple exons skip predictions from Snaptron. There are seven examples (rows 8, 12, 13, 15, 18, 28 and 30) where there is some but not exact agreement. In addition, 6 out of 23 css reported in Table 5.7 did not match a bss. For the css of row 5, Snaptron has no bss with which to compare and for row 2 the css has a non-consensus sequence, which is filtered from Snaptron (Wilks et al., 2018). This analysis shows that high background reads for multiple exon skips is a reasonable indication that these events will occur in response to splice site mutations.

5.3 Discussion

Our analysis of large database of css (Table 5.6) and exon skipping events (Table 5.7) confirm that the effect of splice site mutations upon both css activation or exon skipping (Figure 5.1A) can be largely predicted from background splicing information for normal genes, as first indicated by the analyses of BRCA1 and BRCA2 (Tables 5.2, 5.3). Table 5.5 shows that 85% of 5'ccs and 88% of 3'css match bss and that 75% of 5' css and 64% of 3'css match bss with the most reads. When exon skipping only is caused by a splice site mutation this correlates with higher background reads for skipping compared to candidate css reads in 125/143 (87%) of examples (Table 5.6). Table 5.7 shows that the experimental reports of multiple exons skipping caused by splicing mutations also correlate well with background splicing reads. Consequently, an initial consideration of background splicing gives a useful indication of the primer design required to investigate the likely effect of an intron splice site mutation and should help to interpret RT-PCR results.

Tables 5.3 to 5.6 were analysed using the original SRAv1 database of spliced RNA, which has 41 million spliced RNA reads in total (see Materials and methods). There is now a larger database SRAv2 that has 83 million reads and Table 5.8 indicates how this increases the number of matches between css and bss for BRCA1, BRCA2 and the DBASS database of css.

Table 5.8

A	B	C
	SRAv1	SRAv2
BRCA1	15/17	17/17
BRCA2	14/18	16/18
DBASS5'css	201/237 (85%)	219/237 (92%)
DBASS3'css	97/110 (88%)	101/110 (92%)

Table 5.8 Match between css of BRCA1, BRCA2 and the DBASS database of 5'css and 3'css with SRAv1 and SRAv2 background splice sites (Alexieva et al., 2022). (I contributed to this Table of original data).

As RNA splicing database increase in size this will presumably further reduce the number of css that do not match bss. However, there is likely to be a larger and consistent percentage of bss that lie within 1000 bases of a splice site yet are not activated as css by splice site mutations, ie false positive css. For example, some of the splice site mutations of BRCA1 and BRCA2 activate css that match a bss but do not match the top bss, examples of these are shaded in column 6 of Tables 5.2 and 5.3. Some of these bss might be identified as css in future, Table 5.1A provides such an example, however, most of the BRCA1 and BRCA2 mutations have been repeatedly analysed. The upper limit to the percentage of top bss that might be false positives can be estimated from Table 5.5, where 51 out of 201 (25%) of the css that match bss do not match bss with the top reads. The shaded numbers in Table 5.6 may also be due to false positive css. In these cases, skipping occurs rather than css activation despite the presence of a nearby bss with more reads in Snaptron than the skipping event.

Unlike the SRAv database, the GTEX database of spliced RNA transcripts is made from normal tissue and this smaller database (17m reads) was also analysed by Alexieva et al (2022) in order to confirm that background splicing is a property of normal genes.

Alexieva et al (2022) also compared the other types of splicing mutations with Snaptron splicing data. After intron splice site mutations, the next largest category is mutations that create de novo splice sites (Table 5.4). De novo splice sites may also create pseudoexons (Figure 5.2A, B). Alexieva et al (2022) used the Snaptron data to ask if the site of a de novo mutation was already weakly active prior to the activating de novo mutation. Because of the snaptron filtering system we could only test for

mutations that enhanced already existing 5’ GT or GC sites or exisiting 3’ AG sites. Fig 5.3 column 1 shows that the large majority of enhanced de novo mutations (irrespective of whether they activate pseudoexons) occur at sites that were already active as background splice sites. Columns 2, 3 and 4 of Fig 5.3 apply to both enhanced and created de novo mutations and show that the pseudoexon splices sites (pss) that are sometimes activated by de novo mutations to generate a pseudoexon (Figure 5.2A, B) largely match background splice sites. Figure 5.3 column 4 illustrates that the pss that are used to create a pseudoexon usually match bss that are nearest to the 5’ or 3’ de novo mutation ((D et al., 2022) Fig 2A,B).

Figure 5.3

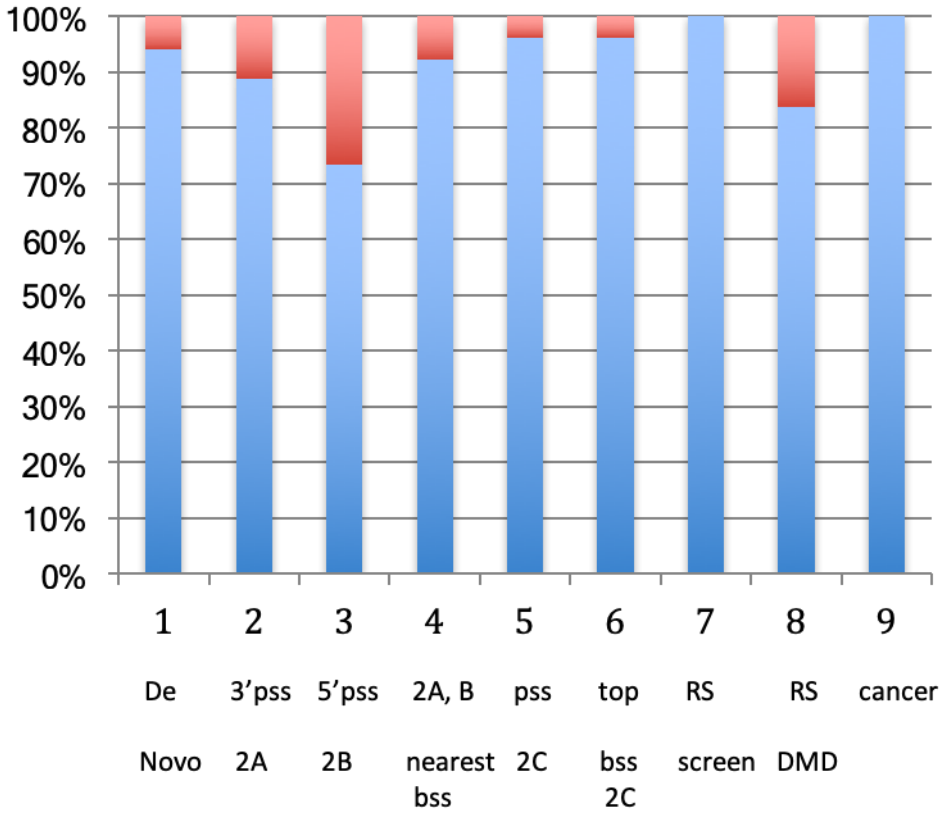


Figure 5.3 Match between background splice sites (bss) with de novo splice sites, pseudoexon ss (pss), recursive ss (RS) and aberrant ss in cancer (Alexieva et al., 2022). Columns **1**: 47/50 match between bss and 'enhanced' de novo ss. **2, 3**: 63/71 and 14/22 match between bss and the 3' or 5' pss of pseudoexons type I (Fig 5.2A, B). **4**: 71/77 bss that match the pseudo ss of type I pseudoexons are nearest to the causative de novo mutation. **5**: 50/52 match of bss to the 3' or 5' ss of pseudoexons type II (Fig 5.2C). **6**: 48/50 type II pss match intron bss with top 3 reads. **7**: 20/20 match between bss and 3' recursive splice sites identified in a genome screen. **8**: 124/148 match between bss and 3'RS and 5'RS of DMD. **9**: 72/72 match between bss and aberrant ss activated by mutations of the spliceosome.

Pseudoexons that are created by mutations other than de novo splice sites (Figure 5.2C) showed the strongest match between their 3' and 5' splice sites with bss. In these cases, the pseudosplice sites usually matched a 3' and 5' bss with a top three read compared to other bss within the same intron. Mutations that might impact splicing because they create or suppress splicing auxiliarly motifs are

among the most difficult to assess by in silico methods, largely because such motifs do not have strong consensus sequences. The analysis by Alexieva et al (2022) provides the additional information that such mutations normally lie within or are close to pseudoexons that already have relatively high background activity.

Large introns are removed in sections by a process called recursive splicing, which is a process that occurs in the nucleus in order to generate mature mRNA. Figure 5.3 colums 7 and 8 show that recursive splice sites strongly match bss and furthermore match bss with relatively high reads compared to other bss. The pss of type II pseudoexons (Figure 5.2C) also match bss with relatively high reads and Keegan et al have shown that such pseudoexons are likely to be generated from recursive splice sites. Finally, Fig 5.3 column 9 shows that the vast majority of cryptic splice sites that are activated in cancer as a result of mutations of the splicing machinery match bss.

Perhaps the most useful background splicing data for an experimentalist is the indication of multiple exon skipping (Table 5.7), which would be missed by the standard primers that are often used to look for single exon skiping or css activation. In addition, background splicing information may also inform splicing therapy. Alexieva et al (2022) demonstrate how background splicing data can identify those rare cryptic splice sites that can be usefully targetted by antisense oligonucleotides (AOs). The use of AOs to induce potentially beneficial exon skipping for the treatment of Duchene muscular dystrophy sometimes activates double exon skipping as an unwanted side effect (Aartsma-Rus et al., 2005, Wilton et al., 2007). Table 5.7 shows that background splicing can predict mutiple exon skipping events caused by mutations and background splicing information may also predict multiple exon skipping by AOs (Alexieva et al., 2022).

Chapter 6-General Discussion

It is now widely accepted that isomiRs are fully functional and that the small differences in sequence at the 5' end of an isomiR can have a surprisingly large impact upon targeting, at least *in vitro* (Table 1.2, Fig 1.7). What is still unclear is whether or not the differences in targeting between a canonical and a derived 5' isomiR are as important as the common targets (NeilsenGoodall and Bracken, 2012, Tomasello et al., 2021). As discussed in Chapter 3 there are relatively few papers that have been able to establish a functional role for specific targeting by isomiRs. Part of the reason for this concerns the availability of tools to modulate specific isomeric forms (NeilsenGoodall and Bracken, 2012), although the development of isomiR specific sponges should be helpful in that regard (Tan et al., 2014).

A remaining problem is one that is general to the miRNA field, namely the likely overlapping and redundant nature of miRNA genes. Although miRNAs were discovered through clear-cut effects upon development in *C.elegans* (see Introduction), a systematic study of miRNA knock-outs in the same organism showed that the majority of 95 miRNA null mutants had no obvious phenotype (Miska et al., 2007). The same paper also questioned the functional relevance of miRNA overexpression studies, where the implied involvement of an overexpressed miRNA in a physiological process was not supported by its subsequent knock-out. Kilikevicius et al (2022) have stressed the importance of studying miRNAs that are expressed at levels that are likely to be of physiological relevance. These are important reservations to bear in mind prior to investigating differences in function of a miRNA and a derived isomiR.

We found that the extreme differences in the ratio of canonical: isomiR levels for human miR-215-5p in liver and kidney compared to other tissues (Tan et al., 2014) (Introduction) was not evidently conserved in mice (Chapter 3) and although we identified human cell lines that showed reasonable differences in expression of some isomiRs (Tables 3.1 to 3.3). We decided to concentrate upon the miRNA cloning project. This was largely because we found marked differences in the cloning efficiencies of different miRNAs by the standard Illumina method (Chapter 3), which clearly needed to be addressed.

In agreement with Zhang et al (2013) we found that the poor cloning of miR-101-1-3p to the RA3 adaptor (Fig 3.2) could be easily improved by adding two variable bases to the 5' end of RA3 and by small changes to the ligation incubation (Fig 3.6). We could not however improve the cloning of miR-214 to the RA5 adapter by adding variable bases (Fig 3.2, 3.9, 3.10). A solution to this problem was provided by Somagenics who kindly told us of their use of a combined RA3RA5 adaptor, which they said was much more efficient (Fig 4.1). We quickly confirmed this to be the case by testing and showing efficient intramolecular ligation of oligonucleotides of the general sequence 214RA3RA5 by several single stranded RNA and DNA ligases (Figs 4.2, S1 to S10). As discussed in Chapter 4 we then had to find a reversible block for the 3' end of the combined adapter RA3RA5 and to overcome problems with the RT-PCR of circular molecules. Because of these delays we were not able to perfect our miRNA cloning protocol nor test its efficiency of cloning or amount of bias.

There is now a wide range of commercial kits available that use various ways of making miRNA libraries, all of which were developed to reduce cloning bias (BenesovaKubista and Valihrach, 2021). Strategies have also been devised for cloning those miRNAs that lack a 5'P or have a 2-O-Me group at the 3' ribose (BenesovaKubista and Valihrach, 2021). Bias in miRNA cloning protocols is assessed by cloning and then sequencing a miRNA reference library such as miRXplore (Fig 1.11, 4.14) and then comparing the sequencing results with the known amounts of miRNAs in the library. The protocol

from Somagenics has shown the least bias in two comprehensive tests of miRNA cloning kits from a range of companies (Barberán-Soler et al., 2018, Herbert et al., 2020), although one of the two tests is a publication from Somagenics. Another study did not notice any improvement with the Somagenics protocol although they had reservations about whether their test was fair (BenesovaKubista and Valihrach, 2021). What is clear is that a considerable amount of biased cloning occurs for all of the various miRNA cloning methods including the Somagenics protocol (BenesovaKubista and Valihrach, 2021) and see section 1.10 and Chapter 4 Discussion.

We note that the sequence of the combined RA3RA5 adapter published by Somagenics (Barberán-Soler et al., 2018) does not have variable bases at either the 3' end of RA5 nor the 5' end of RA3. We do not know if variable bases at the 3' end of RA5 might improve matters but we would certainly expect an improvement in cloning efficiency of some miRNAs by having variable bases at the 5' end of RA3, as shown by (Zhang et al., 2013) and confirmed by us in Chapter 3. Fig 3.7 illustrates that having variable bases at the 5' end of RA3 might be expected to hinder the ligation of the Illumina STP oligo, which is an important step that is used to block unused RA3 adapter from subsequently ligating to RA5 (Fig 3.2). Somagenics use a related blocking method that also relies upon base pairing of a splint oligo to RA3 (Barberán-Soler et al., 2018). In practice we found that variable bases at the 5' end of RA3 did not hinder the ligation of the Illumina STP oligo to NNRA3 (Fig 3.8) nor to the combined NNRA3RA5NN adapter that is used by Somagenics (Fig S15, Appendix 2). Consequently, this is a potential improvement that could be easily introduced.

The method we developed in Chapter 5 for analysing aberrant splicing events is quite distinctive to the *in silico* methods. We essentially use an empirical method that searches large spliced RNA database for background splice sites (bss) that normally are rarely used (Fig 5.1, Table 5.1). However, as shown here and previously (Kapustin et al., 2011) these bss can become strong splice sites when a normal splice site is disrupted by mutation (Table 5.1). We also show that bss are fundamental to

all known examples of aberrant splicing including cancer (Fig 5.3, Alexieva et al 2022). Our method is particularly useful for predicting the effect of a splice site mutation with regards to cryptic splice site activation, exon skipping or even multiple exon skipping (see Chapter 5 Discussion).

In silico methods seem better suited to predicting whether a variant of unknown significance (vus) is likely to disrupt splicing rather than predicting the exact effect of the mutation (JianBoerwinkle and Liu, 2014, Baralle and Buratti, 2017, Moles-Fernandez et al., 2018, DawesJoshi and Cooper, 2022). Our method can also contribute to assessing whether a vus might disrupt splicing. We show that de novo splice sites and pseudoexons splice sites usually originate from bss (Fig 5.3), which would seem useful information to cross-reference. The filtering system used by Snaptron (GT/GC and AG deletion end points) prevents us from analyzing all classes of de novo mutations, but the filtering system can in principle be changed. We also found that the very large majority of pseudoexons of the type illustrated in Fig 5.2C are semi-dormant in that their splice sites are highly active bss prior to their further activation by mutations (D et al., 2022). Consequently, any vus that lies within or close to such semi-dormant pseudoexons is more likely to be of significance.

Dawes et al (2022) independently published the same approach to predicting 5'css (Fig 6.1), which we discuss below.

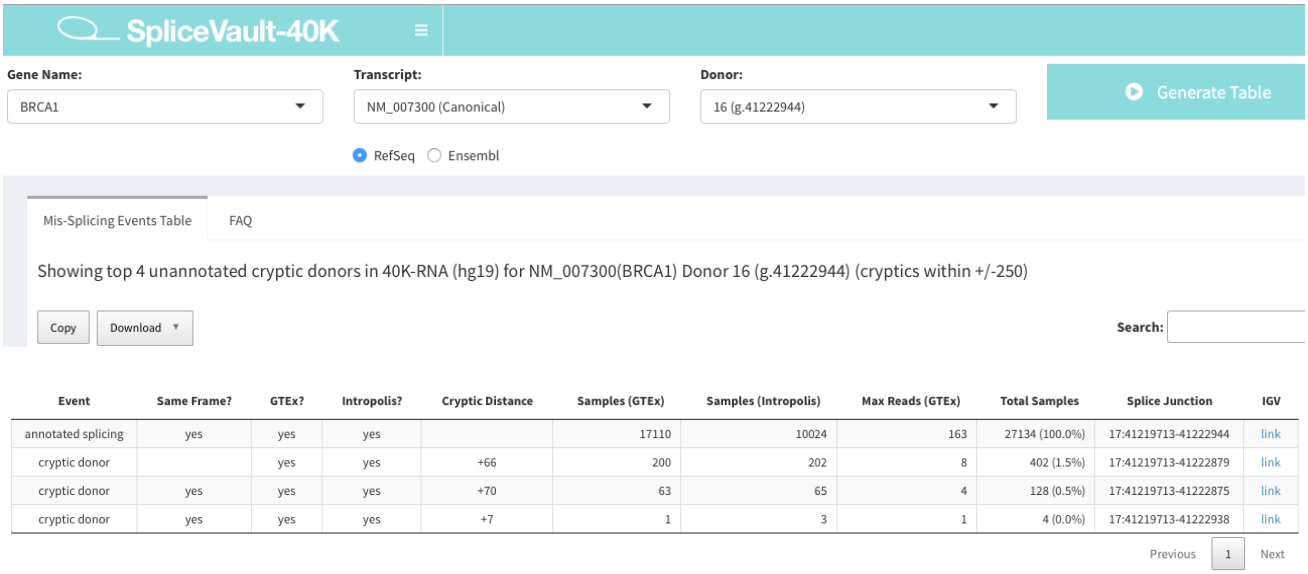


Figure 6.1 Illustration of BRCA1 cryptic splice site predictions by Splicevault-40k. https://kidsneuro.shinyapps.io/splicevault-40k

The top half of Fig 6.1 illustrates one of the strong points of the paper (DawesJoshi and Cooper, 2022) in that it provides a web resource to look for the top four 5'css within 250 bases of any annotated 5'ss on any gene. In Fig 6.1 we have chosen BRCA1 and the same donor site that we analyse in Table 5.1A as a useful comparison. It can be seen that there are 163 reads for annotated or normal splicing between the 5'ss 41222944 and the 3'ss 41219713 (Fig 6.1). This is much lower than the read number shown in Table 5.1A (148299) possibly because Fig 6.1 shows reads per sample rather than total reads. Nevertheless Fig 6.1 detects the use of two background splice sites at positions +65 and +69 that are known to be activated by mutation of the 5'ss 41222944 (Table 5.1A), these are listed as +66 and +70 in Fig 6.1. The authors discuss that their method could be extended to 3'css and to exon skipping as we have shown (Alexieva et al., 2022). In fact, the skipping information becomes available simply by not filtering it out (Table 5.1). Another strong point of the paper is that the authors formally demonstrate that their method (and therefore ours) compares favourably to cryptic splice site

predictions by a number of *in silico* methods including the deep learning tool SpliceAI (Jaganathan et al., 2019), which is considered to be the strongest of the *in silico* methods (DawesJoshi and Cooper, 2022).

We previously provided a web resource called cryptic splice finder (Kapustin et al., 2011), which was similar to Splice-Vault 40K but is no longer maintained. Table 5.1 illustrates that the information provided by CSF or SpliceVault can also be obtained simply by downloading the splicing data for an individual gene into a spread sheet and then ordering the information as shown (Table 5.1) but it would of course be useful to obtain the exact same data more quickly through a web resource. Other possible improvements to our method may result from the likely expansion of the GTEx RNA sequencing database, which we can also analyse (Alexieva et al., 2022). This database catalogues the RNA splicing data for different tissue types, which will allow tissue of most relevance to the patient to be analysed.

Future studies

IsomiR analysis

For future work we would use our same approach (Chapter 3) to analyse more recent miRNA databases as they are likely to be larger and of better quality. The goal would be identified tenfold or more changes in expression of an isomiR between human cell lines, which ideally would also be replicated in mice. This would be a good starting point for subsequent experimental projects that aim to test if the observed differences in expression of isomiRs between different cell types are of functional relevance. Any experimental analysis should be conducted to the high standards described by (Chu et al., 2020, KilikeviciusMeister and Corey, 2022).

miRNA cloning

We would want to optimise and then test our cloning protocol. Our protocol is similar to the one used by Somagenics, which is so far the most successful of the various methods that are available (Chapter 6, Discussion). Unlike Somagenics, we incorporated variable bases at the 5'end of the RA5 adapter, which is a likely advantage that we would like to test. The optimisations we require are discussed in Chapter 4 and concern the more efficient removal of the 3'P reversible blocker at the 3'end of the RA3 part of our cloning vector and establishing a less complicated way of amplifying circular constructs by PCR. Once we have optimized our protocol we would test it by cloning and sequencing the miRXplore miRNA library, which is a commonly used method used to test cloning bias (BenesovaKubista and Valihrach, 2021). We would also test if our method improved upon the Somagenics protocol.

Another advantage of having our own protocol, rather than a kit, is the ability to modify it in order for example to further develop the pulldown technique described in Figs 4.16 and 4.17. In principle this would allow single miRNAs to be analysed by cloning and sequencing, which should be cheaper and less time consuming than having to clone and sequence entire miRNA libraries. In addition, because we tested and developed each step of the cloning protocol for preparing miRNAs for sequencing, we can easily introduce and test further improvements.

Splicing and genetic disease

The main advance here would be to either establish our own database or to identify RNA sequencing data that is not pre-filtered as is the case for snaptron (Wilks et al., 2018). Filtering is used to remove RNA transcripts that have deletions that are not caused by splicing. However, the filter used by Snaptron also removes genuine RNA splicing that uses non-canonical splice sites. We would prefer to use a different filter that we described in Kapustin et al (2011) where the only restriction is that

one end of the mRNA deletion should match an established splice site. There is no restriction on the other end of the splice site, which means that this filtering method is well suited to detected aberrant splicing caused by splice site mutations (Fig 5.1). This method would, for example, allow us to test whether de novo splice site mutations that we describe as 'created' were already active prior to the mutation.

References

AARTSMA-RUS, A., DE WINTER, C. L., JANSON, A. A., KAMAN, W. E., VAN OMMEN, G. J., DEN DUNNEN, J. T. & VAN DEUTEKOM, J. C. 2005. Functional analysis of 114 exon-internal AONs for targeted DMD exon skipping: indication for steric hindrance of SR protein binding sites. *Oligonucleotides,* 15**,** 284-97.

ALAGAR BOOPATHY, L. R., BEADLE, E., GARCIA-BUENO RICO, A. & VERA, M. 2023. Proteostasis regulation through ribosome quality control and no-go-decay. *Wiley Interdiscip Rev RNA***,** e1809.

ALON, S., VIGNEAULT, F., EMINAGA, S., CHRISTODOULOU, D. C., SEIDMAN, J. G., CHURCH, G. M. & EISENBERG, E. 2011. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res,* 21**,** 1506-11.

ALVAREZ, M. E. V., CHIVERS, M., BOROVSKA, I., MONGER, S., GIANNOULATOU, E., KRALOVICOVA, J. & VORECHOVSKY, I. 2021. Transposon clusters as substrates for aberrant splice-site activation. *RNA Biol,* 18**,** 354-367.

ANDROVIC, P., BENESOVA, S., ROHLOVA, E., KUBISTA, M. & VALIHRACH, L. 2021. Small RNA-sequencing for Analysis of Circulating miRNAs: Benchmark Study. *bioRxiv***,** 2021.03.27.437345.

ARROYO, J. D., CHEVILLET, J. R., KROH, E. M., RUF, I. K., PRITCHARD, C. C., GIBSON, D. F., MITCHELL, P. S., BENNETT, C. F., POGOSOVA-AGADJANYAN, E. L., STIREWALT, D. L., TAIT, J. F. & TEWARI, M. 2011. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc Natl Acad Sci U S A,* 108**,** 5003-8.

ARTHUR, L. L. & DJURANOVIC, S. 2018. PolyA tracks, polybasic peptides, poly-translational hurdles. *Wiley Interdiscip Rev RNA,* 9**,** e1486.

AZUMA-MUKAI, A., OGURI, H., MITUYAMA, T., QIAN, Z. R., ASAI, K., SIOMI, H. & SIOMI, M. C. 2008. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proc Natl Acad Sci U S A,* 105**,** 7964-9.

BAERT, A., DEPUYDT, J., VAN MAERKEN, T., POPPE, B., MALFAIT, F., VAN DAMME, T., DE NOBELE, S., PERLETTI, G., DE LEENEER, K., CLAES, K. B. & VRAL, A. 2017. Analysis of chromosomal radiosensitivity of healthy BRCA2 mutation carriers and non-carriers in BRCA families with the G2 micronucleus assay. *Oncol Rep,* 37**,** 1379-1386.

BALDRICH, P., TAMIM, S., MATHIONI, S. & MEYERS, B. 2020. Ligation bias is a major contributor to nonstoichiometric abundances of secondary siRNAs and impacts analyses of microRNAs. *bioRxiv***,** 2020.09.14.296616.

BARALLE, D. & BURATTI, E. 2017. RNA splicing in human disease and in the clinic. *Clin Sci (Lond),* 131**,** 355-368.

BARAN-GALE, J., FANNIN, E. E., KURTZ, C. L. & SETHUPATHY, P. 2013. Beta cell 5'-shifted isomiRs are candidate regulatory hubs in type 2 diabetes. *PLoS One,* 8**,** e73240.

BARAN-GALE, J., KURTZ, C. L., ERDOS, M. R., SISON, C., YOUNG, A., FANNIN, E. E., CHINES, P. S. & SETHUPATHY, P. 2015. Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. *Front Genet,* 6**,** 352.

BARBERÁN-SOLER, S., VO, J. M., HOGANS, R. E., DALLAS, A., JOHNSTON, B. H. & KAZAKOV, S. A. 2018. Decreasing miRNA sequencing bias using a single adapter and circularization approach. *Genome Biol,* 19**,** 105.

BARTEL, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-97.

BECKER, W. R., OBER-REYNOLDS, B., JOURAVLEVA, K., JOLLY, S. M., ZAMORE, P. D. & GREENLEAF, W. J. 2019. High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2. *Mol Cell,* 75**,** 741-755.e11.

BENESOVA, S., KUBISTA, M. & VALIHRACH, L. 2021. Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. *Diagnostics (Basel),* 11.

BERNSTEIN, E., CAUDY, A. A., HAMMOND, S. M. & HANNON, G. J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature,* 409**,** 363-6.

BHASKARAN, M. & MOHAN, M. 2014. MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. *Vet Pathol,* 51**,** 759-74.

BLACK, D. 1994. Familial breast cancer. BRCA1 down, BRCA2 to go. *Curr Biol,* 4**,** 1023-4.

BOFILL-DE ROS, X., HONG, Z., BIRKENFELD, B., ALAMO-ORTIZ, S., YANG, A., DAI, L. & GU, S. 2022. Flexible pri-miRNA structures enable tunable production of 5' isomiRs. *RNA Biol,* 19**,** 279-289.

BREITBART, R. E., ANDREADIS, A. & NADAL-GINARD, B. 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu Rev Biochem,* 56**,** 467-95.

BRENNECKE, J., STARK, A., RUSSELL, R. B. & COHEN, S. M. 2005. Principles of microRNA-target recognition. *PLoS Biol,* 3**,** e85.

BRITTON, C., WINTER, A. D., GILLAN, V. & DEVANEY, E. 2014. microRNAs of parasitic helminths - Identification, characterization and potential as drug targets. *Int J Parasitol Drugs Drug Resist,* 4**,** 85-94.

BURATTI, E., CHIVERS, M., HWANG, G. & VORECHOVSKY, I. 2011. DBASS3 and DBASS5: databases of aberrant 3'- and 5'-splice sites. *Nucleic Acids Res,* 39**,** D86-91.

CAI, X., HAGEDORN, C. H. & CULLEN, B. R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA,* 10**,** 1957-66.

CALIN, G. A., DUMITRU, C. D., SHIMIZU, M., BICHI, R., ZUPO, S., NOCH, E., ALDLER, H., RATTAN, S., KEATING, M., RAI, K., RASSENTI, L., KIPPS, T., NEGRINI, M., BULLRICH, F. & CROCE, C. M. 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A,* 99**,** 15524-9.

CAPLEN, N. J., PARRISH, S., IMANI, F., FIRE, A. & MORGAN, R. A. 2001. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc Natl Acad Sci U S A,* 98**,** 9742-7.

CARTHEW, R. W. & SONTHEIMER, E. J. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell,* 136**,** 642-55.

CAUDY, A. A., MYERS, M., HANNON, G. J. & HAMMOND, S. M. 2002. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev,* 16**,** 2491-6.

CHEN, X. & RECHAVI, O. 2022. Plant and animal small RNA communications between cells and organisms. *Nat Rev Mol Cell Biol,* 23**,** 185-203.

CHU, Y., KILIKEVICIUS, A., LIU, J., JOHNSON, K. C., YOKOTA, S. & COREY, D. R. 2020. Argonaute binding within 3'-untranslated regions poorly predicts gene repression. *Nucleic Acids Res,* 48**,** 7439-7453.

CLOONAN, N., WANI, S., XU, Q., GU, J., LEA, K., HEATER, S., BARBACIORU, C., STEPTOE, A. L., MARTIN, H. C., NOURBAKHSH, E., KRISHNAN, K., GARDINER, B., WANG, X., NONES, K., STEEN, J. A., MATIGIAN, N. A., WOOD, D. L., KASSAHN, K. S., WADDELL, N., SHEPHERD, J., LEE, C., ICHIKAWA, J., MCKERNAN, K., BRAMLETT, K., KUERSTEN, S. & GRIMMOND, S. M. 2011. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol,* 12**,** R126.

COENEN-STASS, A. M. L., MAGEN, I., BROOKS, T., BEN-DOV, I. Z., GREENSMITH, L., HORNSTEIN, E. & FRATTA, P. 2018. Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. *RNA Biol,* 15**,** 1133-1145.

COLOMBO, M., DE VECCHI, G., CALECA, L., FOGLIA, C., RIPAMONTI, C. B., FICARAZZI, F., BARILE, M., VARESCO, L., PEISSEL, B., MANOUKIAN, S. & RADICE, P. 2013. Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS One,* 8**,** e57173.

CONDRAT, C. E., THOMPSON, D. C., BARBU, M. G., BUGNAR, O. L., BOBOC, A., CRETOIU, D., SUCIU, N., CRETOIU, S. M. & VOINEA, S. C. 2020. miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells,* 9.

CORTEZ, M. A., BUESO-RAMOS, C., FERDIN, J., LOPEZ-BERESTEIN, G., SOOD, A. K. & CALIN, G. A. 2011. MicroRNAs in body fluids--the mix of hormones and biomarkers. *Nat Rev Clin Oncol,* 8**,** 467-77.

COUZIGOU, J. M., LAURESSERGUES, D., ANDRÉ, O., GUTJAHR, C., GUILLOTIN, B., BÉCARD, G. & COMBIER, J. P. 2017. Positive Gene Regulation by a Natural Protective miRNA Enables Arbuscular Mycorrhizal Symbiosis. *Cell Host Microbe,* 21**,** 106-112.

CULLER, S. J., HOFF, K. G., VOELKER, R. B., BERGLUND, J. A. & SMOLKE, C. D. 2010. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res,* 38**,** 5152-65.

D, A., Y, L., R, S., H, D., E, B., RM, W., I, V., L, C. & N, J. D. 2022. Background splicing as a predictor of aberrant splicing in genetic disease. *RNA Biol,* 19**,** 256-265.

DAI, L., CHEN, K., YOUNGREN, B., KULINA, J., YANG, A., GUO, Z., LI, J., YU, P. & GU, S. 2016. Cytoplasmic Drosha activity generated by alternative splicing. *Nucleic Acids Res,* 44**,** 10454-10466.

DARD-DASCOT, C., NAQUIN, D., D'AUBENTON-CARAFA, Y., ALIX, K., THERMES, C. & VAN DIJK, E. 2018. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics,* 19**,** 118.

DARMAN, R. B., SEILER, M., AGRAWAL, A. A., LIM, K. H., PENG, S., AIRD, D., BAILEY, S. L., BHAVSAR, E. B., CHAN, B., COLLA, S., CORSON, L., FEALA, J., FEKKES, P., ICHIKAWA, K., KEANEY, G. F., LEE, L., KUMAR, P., KUNII, K., MACKENZIE, C., MATIJEVIC, M., MIZUI, Y., MYINT, K., PARK, E. S., PUYANG, X., SELVARAJ, A., THOMAS, M. P., TSAI, J., WANG, J. Y., WARMUTH, M., YANG, H., ZHU, P., GARCIA-MANERO, G., FURMAN, R. R., YU, L., SMITH, P. G. & BUONAMICI, S. 2015. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep,* 13**,** 1033-45.

DAWES, R., JOSHI, H. & COOPER, S. T. 2022. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. *Nat Commun,* 13**,** 1655.

DE WIT, E., LINSEN, S. E., CUPPEN, E. & BEREZIKOV, E. 2009. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res,* 19**,** 2064-74.

DEBOEVER, C., GHIA, E. M., SHEPARD, P. J., RASSENTI, L., BARRETT, C. L., JEPSEN, K., JAMIESON, C. H., CARSON, D., KIPPS, T. J. & FRAZER, K. A. 2015. Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. *PLoS Comput Biol,* 11**,** e1004105.

DI BLASI, C., JARRE, L., BLASEVICH, F., DASSI, P. & MORA, M. 2008. Danon disease: a novel LAMP2 mutation affecting the pre-mRNA splicing and causing aberrant transcripts and partial protein expression. *Neuromuscul Disord,* 18**,** 962-6.

DIVINA, P., KVITKOVICOVA, A., BURATTI, E. & VORECHOVSKY, I. 2009. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur J Hum Genet,* 17**,** 759-65.

DOENCH, J. G., PETERSEN, C. P. & SHARP, P. A. 2003. siRNAs can function as miRNAs. *Genes Dev,* 17**,** 438-42.

DOENCH, J. G. & SHARP, P. A. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev,* 18**,** 504-11.

EICHHORN, S. W., GUO, H., MCGEARY, S. E., RODRIGUEZ-MIAS, R. A., SHIN, C., BAEK, D., HSU, S. H., GHOSHAL, K., VILLEN, J. & BARTEL, D. P. 2014. mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell,* 56**,** 104-15.

ELBASHIR, S. M., LENDECKEL, W. & TUSCHL, T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev,* 15**,** 188-200.

ELKAYAM, E., KUHN, C. D., TOCILJ, A., HAASE, A. D., GREENE, E. M., HANNON, G. J. & JOSHUA-TOR, L. 2012. The structure of human argonaute-2 in complex with miR-20a. *Cell,* 150**,** 100-10.

ERGIN, K. & ÇETINKAYA, R. 2022. Regulation of MicroRNAs. *Methods Mol Biol,* 2257**,** 1-32.

FABO, T. & KHAVARI, P. 2023. Functional characterization of human genomic variation linked to polygenic diseases. *Trends Genet,* 39**,** 462-490.

FAN, X., ZOU, X., LIU, C., PENG, S., ZHANG, S., ZHOU, X., ZHU, J. & ZHU, W. 2022. Identify miRNA-mRNA regulation pairs to explore potential pathogenesis of lung adenocarcinoma. *Aging (Albany NY),* 14**,** 8357-8373.

FÁTYOL, K., FEKETE, K. A. & LUDMAN, M. 2020. Double-Stranded-RNA-Binding Protein 2 Participates in Antiviral Defense. *J Virol,* 94.

FELEKKIS, K., TOUVANA, E., STEFANOU, C. & DELTAS, C. 2010. microRNAs: a newly described class of encoded molecules that play a role in health and disease. *Hippokratia,* 14**,** 236-40.

FILIPOWICZ, W., BHATTACHARYYA, S. N. & SONENBERG, N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet,* 9**,** 102-14.

FILIPPOV, V., SOLOVYEV, V., FILIPPOVA, M. & GILL, S. S. 2000. A novel type of RNase III family proteins in eukaryotes. *Gene,* 245**,** 213-21.

FIRE, A., XU, S., MONTGOMERY, M. K., KOSTAS, S. A., DRIVER, S. E. & MELLO, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature,* 391**,** 806-11.

FIRE, A. Z. 2007. Gene silencing by double-stranded RNA (Nobel Lecture). *Angew Chem Int Ed Engl,* 46**,** 6966-84.

FOKKEMA, I. F., TASCHNER, P. E., SCHAAFSMA, G. C., CELLI, J., LAROS, J. F. & DEN DUNNEN, J. T. 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat,* 32**,** 557-63.

FORMAN, J. J., LEGESSE-MILLER, A. & COLLER, H. A. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A,* 105**,** 14879-84.

FU, X. D. & ARES, M., JR. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet,* 15**,** 689-701.

GALAGALI, H. & KIM, J. K. 2020. The multifaceted roles of microRNAs in differentiation. *Curr Opin Cell Biol,* 67**,** 118-140.

GAN, J., LEESTEMAKER, Y., SAPMAZ, A. & OVAA, H. 2019. Highlighting the Proteasome: Using Fluorescence to Visualize Proteasome Activity and Distribution. *Front Mol Biosci,* 6**,** 14.

GIRALDEZ, M. D., SPENGLER, R. M., ETHERIDGE, A., GODOY, P. M., BARCZAK, A. J., SRINIVASAN, S., DE HOFF, P. L., TANRIVERDI, K., COURTRIGHT, A., LU, S., KHOORY, J., RUBIO, R., BAXTER, D., DRIEDONKS, T. A. P., BUERMANS, H. P. J., NOLTE-'T HOEN, E. N. M., JIANG, H., WANG, K., GHIRAN, I., WANG, Y. E., VAN KEUREN-JENSEN, K., FREEDMAN, J. E., WOODRUFF, P. G., LAURENT, L. C., ERLE, D. J., GALAS, D. J. & TEWARI, M. 2018. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat Biotechnol,* 36**,** 746-757.

GODOY, P. M., BARCZAK, A. J., DEHOFF, P., SRINIVASAN, S., ETHERIDGE, A., GALAS, D., DAS, S., ERLE, D. J. & LAURENT, L. C. 2019. Comparison of Reproducibility, Accuracy, Sensitivity, and Specificity of miRNA Quantification Platforms. *Cell Rep,* 29**,** 4212-4222.e5.

GREENE, S., HUANG, J., HAMILTON, K., TONG, L., HOBERT, O. & SUN, H. 2023. The heterochronic LIN-14 protein is a BEN domain transcription factor. *Curr Biol,* 33**,** R217-r218.

GRIFFITHS-JONES, S., HUI, J. H., MARCO, A. & RONSHAUGEN, M. 2011. MicroRNA evolution by arm switching. *EMBO Rep,* 12**,** 172-7.

GRIMSON, A., FARH, K. K., JOHNSTON, W. K., GARRETT-ENGELE, P., LIM, L. P. & BARTEL, D. P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell,* 27**,** 91-105.

GRISHOK, A., PASQUINELLI, A. E., CONTE, D., LI, N., PARRISH, S., HA, I., BAILLIE, D. L., FIRE, A., RUVKUN, G. & MELLO, C. C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell,* 106**,** 23-34.

GROSSWENDT, S., FILIPCHYK, A., MANZANO, M., KLIRONOMOS, F., SCHILLING, M., HERZOG, M., GOTTWEIN, E. & RAJEWSKY, N. 2014. Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol Cell,* 54**,** 1042-1054.

GUO, L., YANG, Q., LU, J., LI, H., GE, Q., GU, W., BAI, Y. & LU, Z. 2011. A comprehensive survey of miRNA repertoire and 3' addition events in the placentas of patients with pre-eclampsia from high-throughput sequencing. *PLoS One,* 6**,** e21072.

GUO, L., ZHAO, Y., YANG, S., ZHANG, H. & CHEN, F. 2014. A genome-wide screen for non-template nucleotides and isomiR repertoires in miRNAs indicates dynamic and versatile microRNAome. *Mol Biol Rep,* 41**,** 6649-58.

GUO, W. T. & WANG, Y. 2019. Dgcr8 knockout approaches to understand microRNA functions in vitro and in vivo. *Cell Mol Life Sci,* 76**,** 1697-1711.

GUSHCHINA, L. V., KWIATKOWSKI, T. A., BHATTACHARYA, S. & WEISLEDER, N. L. 2018. Conserved structural and functional aspects of the tripartite motif gene family point towards therapeutic applications in multiple diseases. *Pharmacol Ther,* 185**,** 12-25.

HAFNER, M., RENWICK, N., BROWN, M., MIHAILOVIĆ, A., HOLOCH, D., LIN, C., PENA, J. T., NUSBAUM, J. D., MOROZOV, P., LUDWIG, J., OJO, T., LUO, S., SCHROTH, G. & TUSCHL, T. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *Rna,* 17**,** 1697-712.

HAMILTON, A. J. & BAULCOMBE, D. C. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science,* 286**,** 950-2.

HAMMOND, S. M. 2015. An overview of microRNAs. *Adv Drug Deliv Rev,* 87**,** 3-14.

HAMMOND, S. M., BERNSTEIN, E., BEACH, D. & HANNON, G. J. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature,* 404**,** 293-6.

HAN, J., LEE, Y., YEOM, K. H., KIM, Y. K., JIN, H. & KIM, V. N. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev,* 18**,** 3016-27.

HEINICKE, F., ZHONG, X., ZUCKNICK, M., BREIDENBACH, J., SUNDARAM, A. Y. M., S, T. F., LEITHAUG, M., DALLAND, M., FARMER, A., HENDERSON, J. M., HUSSONG, M. A., MOLL, P., NGUYEN, L., MCNULTY, A., SHAFFER, J. M., SHORE, S., YIP, H. K., VITKOVSKA, J., RAYNER, S., LIE, B. A. & GILFILLAN, G. D. 2020. Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol,* 17**,** 75-86.

HELWAK, A., KUDLA, G., DUDNAKOVA, T. & TOLLERVEY, D. 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell,* 153**,** 654-65.

HENNER, W. D., GRUNBERG, S. M. & HASELTINE, W. A. 1983. Enzyme action at 3' termini of ionizing radiation-induced DNA strand breaks. *J Biol Chem,* 258**,** 15198-205.

HERBERT, Z. T., THIMMAPURAM, J., XIE, S., KERSHNER, J. P., KOLLING, F. W., RINGELBERG, C. S., LECLERC, A., ALEKSEYEV, Y. O., FAN, J., PODNAR, J. W., STEVENSON, H. S., SOMMERVILLE, G., GUPTA, S., BERKELEY, M., KOEMAN, J., PERERA, A., SCOTT, A. R., GRENIER, J. K., MALIK, J., ASHTON, J. M., PIVARSKI, K. L., WANG, X., KUFFEL, G., MESA, T. E., SMITH, A. T., SHEN, J., TAKATA, Y., VOLKERT, T. L., LOVE, J. A., ZHANG, Y., WANG, J., XUEI, X., ADAMS, M. & LEVINE, S. S. 2020. Multisite Evaluation of Next-Generation Methods for Small RNA Quantification. *J Biomol Tech,* 31**,** 47-56.

HRISTOVA, M., BIRSE, D., HONG, Y. & AMBROS, V. 2005. The Caenorhabditis elegans heterochronic regulator LIN-14 is a novel transcription factor that controls the developmental timing of transcription from the insulin/insulin-like growth factor gene ins-33 by direct DNA binding. *Mol Cell Biol,* 25**,** 11059-72.

HUNG, Y. H. & SLOTKIN, R. K. 2021. The initiation of RNA interference (RNAi) in plants. *Curr Opin Plant Biol,* 61**,** 102014.

HUTVÁGNER, G., MCLACHLAN, J., PASQUINELLI, A. E., BÁLINT, E., TUSCHL, T. & ZAMORE, P. D. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science,* 293**,** 834-8.

HUTVÁGNER, G. & ZAMORE, P. D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science,* 297**,** 2056-60.

ILAGAN, J. O., RAMAKRISHNAN, A., HAYES, B., MURPHY, M. E., ZEBARI, A. S., BRADLEY, P. & BRADLEY, R. K. 2015. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res,* 25**,** 14-26.

ISHIZUKA, A., SIOMI, M. C. & SIOMI, H. 2002. A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes Dev,* 16**,** 2497-508.

IWAKAWA, H. O. & TOMARI, Y. 2022. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol Cell,* 82**,** 30-43.

JAGADEESWARAN, G., ZHENG, Y., SUMATHIPALA, N., JIANG, H., ARRESE, E. L., SOULAGES, J. L., ZHANG, W. & SUNKAR, R. 2010. Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development. *BMC Genomics,* 11**,** 52.

JAGANATHAN, K., KYRIAZOPOULOU PANAGIOTOPOULOU, S., MCRAE, J. F., DARBANDI, S. F., KNOWLES, D., LI, Y. I., KOSMICKI, J. A., ARBELAEZ, J., CUI, W., SCHWARTZ, G. B., CHOW, E. D., KANTERAKIS, E., GAO, H., KIA, A., BATZOGLOU, S., SANDERS, S. J. & FARH, K. K. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell,* 176**,** 535-548.e24.

JAYAPRAKASH, A. D., JABADO, O., BROWN, B. D. & SACHIDANANDAM, R. 2011. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res,* 39**,** e141.

JIAN, X., BOERWINKLE, E. & LIU, X. 2014. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med,* 16**,** 497-503.

JONES, M. R., QUINTON, L. J., BLAHNA, M. T., NEILSON, J. R., FU, S., IVANOV, A. R., WOLF, D. A. & MIZGERD, J. P. 2009. Zcchc11-dependent uridylation of microRNA directs cytokine expression. *Nat Cell Biol,* 11**,** 1157-63.

JOURAVLEVA, K., GOLOVENKO, D., DEMO, G., DUTCHER, R. C., HALL, T. M. T., ZAMORE, P. D. & KOROSTELEV, A. A. 2022. Structural basis of microRNA biogenesis by Dicer-1 and its partner protein Loqs-PB. *Mol Cell,* 82**,** 4049-4063 e6.

KAPUSTIN, Y., CHAN, E., SARKAR, R., WONG, F., VORECHOVSKY, I., WINSTON, R. M., TATUSOVA, T. & DIBB, N. J. 2011. Cryptic splice sites and split genes. *Nucleic Acids Res,* 39**,** 5837-44.

KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res,* 12**,** 656-64.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KETTING, R. F., FISCHER, S. E., BERNSTEIN, E., SIJEN, T., HANNON, G. J. & PLASTERK, R. H. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev,* 15**,** 2654-9.

KETTING, R. F., HAVERKAMP, T. H., VAN LUENEN, H. G. & PLASTERK, R. H. 1999. Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell,* 99**,** 133-41.

KHVOROVA, A., REYNOLDS, A. & JAYASENA, S. D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell,* 115**,** 209-16.

KILIKEVICIUS, A., MEISTER, G. & COREY, D. R. 2022. Reexamining assumptions about miRNA-guided gene silencing. *Nucleic Acids Res,* 50**,** 617-634.

KIM, H., KIM, J., YU, S., LEE, Y. Y., PARK, J., CHOI, R. J., YOON, S. J., KANG, S. G. & KIM, V. N. 2020. A Mechanism for microRNA Arm Switching Regulated by Uridylation. *Mol Cell,* 78**,** 1224-1236.e5.

KIM, K., BAEK, S. C., LEE, Y. Y., BASTIAANSSEN, C., KIM, J., KIM, H. & KIM, V. N. 2021. A quantitative map of human primary microRNA processing sites. *Mol Cell,* 81**,** 3422-3439.e11.

KIM, Y. 2023. Drug Discovery Perspectives of Antisense Oligonucleotides. *Biomol Ther (Seoul),* 31**,** 241-252.

KIM, Y. & KIM, V. N. 2012. MicroRNA factory: RISC assembly from precursor microRNAs. *Mol Cell,* 46**,** 384-6.

KIM, Y. K. & KIM, V. N. 2007. Processing of intronic microRNAs. *EMBO J,* 26**,** 775-83.

KLOOSTERMAN, W. P., WIENHOLDS, E., KETTING, R. F. & PLASTERK, R. H. 2004. Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res,* 32**,** 6284-91.

KNIGHT, S. W. & BASS, B. L. 2001. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science,* 293**,** 2269-71.

KOBAYASHI, Y., TIAN, S. & UI-TEI, K. 2022. The siRNA Off-Target Effect Is Determined by Base-Pairing Stabilities of Two Different Regions with Opposite Effects. *Genes (Basel),* 13.

KONISHI, H., MOHSENI, M., TAMAKI, A., GARAY, J. P., CROESSMANN, S., KARNAN, S., OTA, A., WONG, H. Y., KONISHI, Y., KARAKAS, B., TAHIR, K., ABUKHDEIR, A. M., GUSTIN, J. P., CIDADO, J., WANG, G. M., COSGROVE, D., COCHRAN, R., JELOVAC, D., HIGGINS, M. J., ARENA, S., HAWKINS, L., LAURING, J., GROSS, A. L., HEAPHY, C. M., HOSOKAWA, Y., GABRIELSON, E., MEEKER, A. K., VISVANATHAN, K., ARGANI, P., BACHMAN, K. E. & PARK, B. H. 2011. Mutation of a single allele of the cancer susceptibility gene BRCA1 leads to genomic instability in human breast epithelial cells. *Proc Natl Acad Sci U S A,* 108**,** 17773-8.

KOZOMARA, A., BIRGAOANU, M. & GRIFFITHS-JONES, S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Res,* 47**,** D155-D162.

KROL, J., LOEDIGE, I. & FILIPOWICZ, W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet,* 11**,** 597-610.

KROL, J., SOBCZAK, K., WILCZYNSKA, U., DRATH, M., JASINSKA, A., KACZYNSKA, D. & KRZYZOSIAK, W. J. 2004. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem,* 279**,** 42230-9.

KUO, W. T., SU, M. W., LEE, Y. L., CHEN, C. H., WU, C. W., FANG, W. L., HUANG, K. H. & LIN, W. C. 2015. Bioinformatic Interrogation of 5p-arm and 3p-arm Specific miRNA Expression Using TCGA Datasets. *J Clin Med,* 4**,** 1798-814.

KWAK, P. B. & TOMARI, Y. 2012. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat Struct Mol Biol,* 19**,** 145-51.

LADOMERY, M. R., MADDOCKS, D. G. & WILSON, I. D. 2011. MicroRNAs: their discovery, biogenesis, function and potential use as biomarkers in non-invasive prenatal diagnostics. *Int J Mol Epidemiol Genet,* 2**,** 253-60.

LAGOS-QUINTANA, M., RAUHUT, R., LENDECKEL, W. & TUSCHL, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science,* 294**,** 853-8.

LAI, E. C., TAM, B. & RUBIN, G. M. 2005. Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes Dev,* 19**,** 1067-80.

LAM, J. K., CHOW, M. Y., ZHANG, Y. & LEUNG, S. W. 2015. siRNA Versus miRNA as Therapeutics for Gene Silencing. *Mol Ther Nucleic Acids,* 4**,** e252.

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., SCHERMER, B., BISSELS, U., INMAN, J., PHAN, Q., CHIEN, M., WEIR, D. B., CHOKSI, R., DE VITA, G., FREZZETTI, D., TROMPETER, H. 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. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell,* 129**,** 1401-14.

LAU, N. C., LIM, L. P., WEINSTEIN, E. G. & BARTEL, D. P. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science,* 294**,** 858-62.

LEE, L. W., ZHANG, S., ETHERIDGE, A., MA, L., MARTIN, D., GALAS, D. & WANG, K. 2010. Complexity of the microRNA repertoire revealed by next-generation sequencing. *Rna,* 16**,** 2170-80.

LEE, R. C. & AMBROS, V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science,* 294**,** 862-4.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75**,** 843-54.

LEE, Y., AHN, C., HAN, J., CHOI, H., KIM, J., YIM, J., LEE, J., PROVOST, P., RÅDMARK, O., KIM, S. & KIM, V. N. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature,* 425**,** 415-9.

LEE, Y., JEON, K., LEE, J. T., KIM, S. & KIM, V. N. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *Embo j,* 21**,** 4663-70.

LEE, Y., KIM, M., HAN, J., YEOM, K. H., LEE, S., BAEK, S. H. & KIM, V. N. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J,* 23**,** 4051-60.

LEE, Y. Y., KIM, H. & KIM, V. N. 2023. Sequence determinant of small RNA production by DICER. *Nature,* 615**,** 323-330.

LEI, E. P. & SILVER, P. A. 2002. Protein and RNA export from the nucleus. *Dev Cell,* 2**,** 261-72.

LEITAO, A. L. & ENGUITA, F. J. 2022. A Structural View of miRNA Biogenesis and Function. *Noncoding RNA,* 8.

LEWIS, B. P., SHIH, I. H., JONES-RHOADES, M. W., BARTEL, D. P. & BURGE, C. B. 2003. Prediction of mammalian microRNA targets. *Cell,* 115**,** 787-98.

LI, C. & ZAMORE, P. D. 2019. RNA Interference and Small RNA Analysis. *Cold Spring Harb Protoc,* 2019.

LI, X., MICHELS, B. E., TOSUN, O. E., JUNG, J., KAPPES, J., IBING, S., NATARAJ, N. B., SAHAY, S., SCHNEIDER, M., WÖRNER, A., BECKI, C., ISHAQUE, N., FEUERBACH, L., HEßLING, B., HELM, D., WILL, R., YARDEN, Y., MÜLLER-DECKER, K., WIEMANN, S. & KÖRNER, C. 2022. 5'isomiR-183-5p|+2 elicits tumor suppressor activity in a negative feedback loop with E2F1. *J Exp Clin Cancer Res,* 41**,** 190.

LIM, L. P., LAU, N. C., GARRETT-ENGELE, P., GRIMSON, A., SCHELTER, J. M., CASTLE, J., BARTEL, D. P., LINSLEY, P. S. & JOHNSON, J. M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature,* 433**,** 769-73.

LIM, L. P., LAU, N. C., WEINSTEIN, E. G., ABDELHAKIM, A., YEKTA, S., RHOADES, M. W., BURGE, C. B. & BARTEL, D. P. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev,* 17**,** 991-1008.

LINSEN, S. E., DE WIT, E., JANSSENS, G., HEATER, S., CHAPMAN, L., PARKIN, R. K., FRITZ, B., WYMAN, S. K., DE BRUIJN, E., VOEST, E. E., KUERSTEN, S., TEWARI, M. & CUPPEN, E. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods,* 6**,** 474-6.

LIU, J., CARMELL, M. A., RIVAS, F. V., MARSDEN, C. G., THOMSON, J. M., SONG, J. J., HAMMOND, S. M., JOSHUA-TOR, L. & HANNON, G. J. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science,* 305**,** 1437-41.

LIU, S., HAN, Y., LI, W. X. & DING, S. W. 2023. Infection Defects of RNA and DNA Viruses Induced by Antiviral RNA Interference. *Microbiol Mol Biol Rev,* 87**,** e0003522.

LIU, W. & WANG, X. 2019. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol,* 20**,** 18.

LIU, X., FORTIN, K. & MOURELATOS, Z. 2008. MicroRNAs: biogenesis and molecular functions. *Brain Pathol,* 18**,** 113-21.

LOIBL, N., ARENZ, C. & SEITZ, O. 2020. Monitoring Dicer-Mediated miRNA-21 Maturation and Ago2 Loading by a Dual-Colour FIT PNA Probe Set. *Chembiochem,* 21**,** 2527-2532.

LU, T. X. & ROTHENBERG, M. E. 2018. MicroRNA. *J Allergy Clin Immunol,* 141**,** 1202-1207.

MAKAROVA, J., TURCHINOVICH, A., SHKURNIKOV, M. & TONEVITSKY, A. 2021. Extracellular miRNAs and Cell-Cell Communication: Problems and Prospects. *Trends Biochem Sci,* 46**,** 640-651.

MATSUYAMA, H. & SUZUKI, H. I. 2019. Systems and Synthetic microRNA Biology: From Biogenesis to Disease Pathogenesis. *Int J Mol Sci,* 21.

MEISTER, G., LANDTHALER, M., PATKANIOWSKA, A., DORSETT, Y., TENG, G. & TUSCHL, T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell,* 15**,** 185-97.

MELO, S. A. & ESTELLER, M. 2014. Disruption of microRNA nuclear transport in human cancer. *Semin Cancer Biol,* 27**,** 46-51.

MICHLEWSKI, G. & CÁCERES, J. F. 2019. Post-transcriptional control of miRNA biogenesis. *Rna,* 25**,** 1-16.

MISKA, E. A., ALVAREZ-SAAVEDRA, E., ABBOTT, A. L., LAU, N. C., HELLMAN, A. B., MCGONAGLE, S. M., BARTEL, D. P., AMBROS, V. R. & HORVITZ, H. R. 2007. Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS Genet,* 3**,** e215.

MOHSEN, M. G. & KOOL, E. T. 2016. The Discovery of Rolling Circle Amplification and Rolling Circle Transcription. *Acc Chem Res,* 49**,** 2540-2550.

MOLES-FERNANDEZ, A., DURAN-LOZANO, L., MONTALBAN, G., BONACHE, S., LOPEZ-PEROLIO, I., MENENDEZ, M., SANTAMARINA, M., BEHAR, R., BLANCO, A., CARRASCO, E., LOPEZ-FERNANDEZ, A., STJEPANOVIC, N., BALMANA, J., CAPELLA, G., PINEDA, M., VEGA, A., LAZARO, C., DE LA HOYA, M., DIEZ, O. & GUTIERREZ-ENRIQUEZ, S. 2018. Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? *Front Genet,* 9**,** 366.

MORIN, R. D., AKSAY, G., DOLGOSHEINA, E., EBHARDT, H. A., MAGRINI, V., MARDIS, E. R., SAHINALP, S. C. & UNRAU, P. J. 2008a. Comparative analysis of the small RNA transcriptomes of Pinus contorta and Oryza sativa. *Genome Res,* 18**,** 571-84.

MORIN, R. D., O'CONNOR, M. D., GRIFFITH, M., KUCHENBAUER, F., DELANEY, A., PRABHU, A. L., ZHAO, Y., MCDONALD, H., ZENG, T., HIRST, M., EAVES, C. J. & MARRA, M. A. 2008b. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res,* 18**,** 610-21.

MORRIS, C., CLUET, D. & RICCI, E. P. 2021. Ribosome dynamics and mRNA turnover, a complex relationship under constant cellular scrutiny. *Wiley Interdiscip Rev RNA,* 12**,** e1658.

MOURELATOS, Z., DOSTIE, J., PAUSHKIN, S., SHARMA, A., CHARROUX, B., ABEL, L., RAPPSILBER, J., MANN, M. & DREYFUSS, G. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev,* 16**,** 720-8.

MOYNAHAN, M. E., CHIU, J. W., KOLLER, B. H. & JASIN, M. 1999. Brca1 controls homology-directed DNA repair. *Mol Cell,* 4**,** 511-8.

MÜLLER, M., FAZI, F. & CIAUDO, C. 2019. Argonaute Proteins: From Structure to Function in Development and Pathological Cell Fate Determination. *Front Cell Dev Biol,* 7**,** 360.

NAELI, P., WINTER, T., HACKETT, A. P., ALBOUSHI, L. & JAFARNEJAD, S. M. 2022. The intricate balance between microRNA-induced mRNA decay and translational repression. *Febs j*.

NAKANISHI, K. 2022. Anatomy of four human Argonaute proteins. *Nucleic Acids Res,* 50**,** 6618-6638.

NEILSEN, C. T., GOODALL, G. J. & BRACKEN, C. P. 2012. IsomiRs--the overlooked repertoire in the dynamic microRNAome. *Trends Genet,* 28**,** 544-9.

NEUMEIER, J. & MEISTER, G. 2020. siRNA Specificity: RNAi Mechanisms and Strategies to Reduce Off-Target Effects. *Front Plant Sci,* 11**,** 526455.

NOWAK, I. & SARSHAD, A. A. 2021. Argonaute Proteins Take Center Stage in Cancers. *Cancers (Basel),* 13.

OHNO, K., TAKEDA, J. I. & MASUDA, A. 2018. Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdiscip Rev RNA,* 9.

OHTSUKA, M., LING, H., DOKI, Y., MORI, M. & CALIN, G. A. 2015. MicroRNA Processing and Human Cancer. *J Clin Med,* 4**,** 1651-67.

ORBÁN, T. I. 2023. One locus, several functional RNAs-emerging roles of the mechanisms responsible for the sequence variability of microRNAs. *Biol Futur,* 74**,** 17-28.

PANZADE, G., LI, L., HEBBAR, S., VEKSLER-LUBLINSKY, I. & ZINOVYEVA, A. 2022. Global profiling and annotation of templated isomiRs dynamics across Caenorhabditis elegans development. *RNA Biol,* 19**,** 928-942.

PARK, C. Y., CHOI, Y. S. & MCMANUS, M. T. 2010. Analysis of microRNA knockouts in mice. *Hum Mol Genet,* 19**,** R169-75.

PARK, C. Y., JEKER, L. T., CARVER-MOORE, K., OH, A., LIU, H. J., CAMERON, R., RICHARDS, H., LI, Z., ADLER, D., YOSHINAGA, Y., MARTINEZ, M., NEFADOV, M., ABBAS, A. K., WEISS, A., LANIER, L. L., DE JONG, P. J., BLUESTONE, J. A., SRIVASTAVA, D. & MCMANUS, M. T. 2012. A resource for the conditional ablation of microRNAs in the mouse. *Cell Rep,* 1**,** 385-91.

PARRISH, S., FLEENOR, J., XU, S., MELLO, C. & FIRE, A. 2000. Functional anatomy of a dsRNA trigger: differential requirement for the two trigger strands in RNA interference. *Mol Cell,* 6**,** 1077-87.

PASQUINELLI, A. E., REINHART, B. J., SLACK, F., MARTINDALE, M. Q., KURODA, M. I., MALLER, B., HAYWARD, D. C., BALL, E. E., DEGNAN, B., MÜLLER, P., SPRING, J., SRINIVASAN, A., FISHMAN, M., FINNERTY, J., CORBO, J., LEVINE, M., LEAHY, P., DAVIDSON, E. & RUVKUN, G. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature,* 408**,** 86-9.

PATEL, K. J., YU, V. P., LEE, H., CORCORAN, A., THISTLETHWAITE, F. C., EVANS, M. J., COLLEDGE, W. H., FRIEDMAN, L. S., PONDER, B. A. & VENKITARAMAN, A. R. 1998. Involvement of Brca2 in DNA repair. *Mol Cell,* 1**,** 347-57.

PATERSON, R. W., GABELLE, A., LUCEY, B. P., BARTHÉLEMY, N. R., LECKEY, C. A., HIRTZ, C., LEHMANN, S., SATO, C., PATTERSON, B. W., WEST, T., YARASHESKI, K., ROHRER, J. D., WILDBURGER, N. C., SCHOTT, J. M., KARCH, C. M., WRAY, S., MILLER, T. M., ELBERT, D. L., ZETTERBERG, H., FOX, N. C. & BATEMAN, R. J. 2019. SILK studies - capturing the turnover of proteins linked to neurodegenerative diseases. *Nat Rev Neurol,* 15**,** 419-427.

PÉREZ-CAÑAMÁS, M., HEVIA, E., KATSAROU, K. & HERNÁNDEZ, C. 2021. Genetic evidence for the involvement of Dicer-like 2 and 4 as well as Argonaute 2 in the Nicotiana benthamiana response against Pelargonium line pattern virus. *J Gen Virol,* 102.

PILLAI, R. S., ARTUS, C. G. & FILIPOWICZ, W. 2004. Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis. *Rna,* 10**,** 1518-25.

PONG, S. K. & GULLEROVA, M. 2018. Noncanonical functions of microRNA pathway enzymes - Drosha, DGCR8, Dicer and Ago proteins. *FEBS Lett,* 592**,** 2973-2986.

PU, M., CHEN, J., TAO, Z., MIAO, L., QI, X., WANG, Y. & REN, J. 2019. Regulatory network of miRNA on its target: coordination between transcriptional and post-transcriptional regulation of gene expression. *Cell Mol Life Sci,* 76**,** 441-451.

QUINN, J. J. & CHANG, H. Y. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet,* 17**,** 47-62.

RANASINGHE, P., ADDISON, M. L., DEAR, J. W. & WEBB, D. J. 2022. Small interfering RNA: Discovery, pharmacology and clinical development-An introductory review. *Br J Pharmacol*.

RAND, T. A., GINALSKI, K., GRISHIN, N. V. & WANG, X. 2004. Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. *Proc Natl Acad Sci U S A,* 101**,** 14385-9.

REINHART, B. J., SLACK, F. J., BASSON, M., PASQUINELLI, A. E., BETTINGER, J. C., ROUGVIE, A. E., HORVITZ, H. R. & RUVKUN, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature,* 403**,** 901-6.

RODRIGUEZ, A., GRIFFITHS-JONES, S., ASHURST, J. L. & BRADLEY, A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res,* 14**,** 1902-10.

ROLFS, Z., FREY, B. L., SHI, X., KAWAI, Y., SMITH, L. M. & WELHAM, N. V. 2021. An atlas of protein turnover rates in mouse tissues. *Nat Commun,* 12**,** 6778.

ROSS, A. B., LANGER, J. D. & JOVANOVIC, M. 2021. Proteome Turnover in the Spotlight: Approaches, Applications, and Perspectives. *Mol Cell Proteomics,* 20**,** 100016.

RUSSO, J., HARRINGTON, A. W. & STEINIGER, M. 2016. Antisense Transcription of Retrotransposons in Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. *Genetics,* 202**,** 107-21.

SALIM, U., KUMAR, A., KULSHRESHTHA, R. & VIVEKANANDAN, P. 2022. Biogenesis, characterization, and functions of mirtrons. *Wiley Interdiscip Rev RNA,* 13**,** e1680.

SASAKI, T., SHIOHAMA, A., MINOSHIMA, S. & SHIMIZU, N. 2003. Identification of eight members of the Argonaute family in the human genome. *Genomics,* 82**,** 323-30.

SCHMUCKER, D., CLEMENS, J. C., SHU, H., WORBY, C. A., XIAO, J., MUDA, M., DIXON, J. E. & ZIPURSKY, S. L. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell,* 101**,** 671-84.

SCHOLL, T., PYNE, M. T., RUSSO, D. & WARD, B. E. 1999. BRCA1 IVS16+6T-->C is a deleterious mutation that creates an aberrant transcript by activating a cryptic splice donor site. *Am J Med Genet,* 85**,** 113-6.

SCHULMAN, B. R., ESQUELA-KERSCHER, A. & SLACK, F. J. 2005. Reciprocal expression of lin-41 and the microRNAs let-7 and mir-125 during mouse embryogenesis. *Dev Dyn,* 234**,** 1046-54.

SCHWARZ, D. S., HUTVÁGNER, G., DU, T., XU, Z., ARONIN, N. & ZAMORE, P. D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell,* 115**,** 199-208.

SCHWARZ, D. S., TOMARI, Y. & ZAMORE, P. D. 2004. The RNA-induced silencing complex is a Mg2+-dependent endonuclease. *Curr Biol,* 14**,** 787-91.

SCOTTI, M. M. & SWANSON, M. S. 2016. RNA mis-splicing in disease. *Nat Rev Genet,* 17**,** 19-32.

SEMPLE, S. L., AU, S. K. W., JACOB, R. A., MOSSMAN, K. L. & DEWITTE-ORR, S. J. 2022. Discovery and Use of Long dsRNA Mediated RNA Interference to Stimulate Antiviral Protection in Interferon Competent Mammalian Cells. *Front Immunol,* 13**,** 859749.

SHIVDASANI, R. A. 2006. MicroRNAs: regulators of gene expression and cell differentiation. *Blood,* 108**,** 3646-53.

SINGH, A., JAIN, D., PANDEY, J., YADAV, M., BANSAL, K. C. & SINGH, I. K. 2023. Deciphering the role of miRNA in reprogramming plant responses to drought stress. *Crit Rev Biotechnol,* 43**,** 613-627.

SMITH, C. M. & HUTVAGNER, G. 2022. A comparative analysis of single cell small RNA sequencing data reveals heterogeneous isomiR expression and regulation. *Sci Rep,* 12**,** 2834.

SONG, J. J., SMITH, S. K., HANNON, G. J. & JOSHUA-TOR, L. 2004. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science,* 305**,** 1434-7.

SONG, M. S., ALLUIN, J. & ROSSI, J. J. 2022. The Effect of Dicer Knockout on RNA Interference Using Various Dicer Substrate Small Interfering RNA (DsiRNA) Structures. *Genes (Basel),* 13.

STAVAST, C. J. & ERKELAND, S. J. 2019. The Non-Canonical Aspects of MicroRNAs: Many Roads to Gene Regulation. *Cells,* 8.

STEFFENSEN, A. Y., DANDANELL, M., JØNSON, L., EJLERTSEN, B., GERDES, A. M., NIELSEN, F. C. & HANSEN, T. 2014. Functional characterization of BRCA1 gene variants by mini-gene splicing assay. *Eur J Hum Genet,* 22**,** 1362-8.

STENSON, P. D., MORT, M., BALL, E. V., CHAPMAN, M., EVANS, K., AZEVEDO, L., HAYDEN, M., HEYWOOD, S., MILLAR, D. S., PHILLIPS, A. D. & COOPER, D. N. 2020. The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. *Hum Genet*.

STENSON, W. F. & CIORBA, M. A. 2020. Nonmicrobial Activation of TLRs Controls Intestinal Growth, Wound Repair, and Radioprotection. *Front Immunol,* 11**,** 617510.

SUZUKI, H., KUMAR, S. A., SHUAI, S., DIAZ-NAVARRO, A., GUTIERREZ-FERNANDEZ, A., DE ANTONELLIS, P., CAVALLI, F. M. G., JURASCHKA, K., FAROOQ, H., SHIBAHARA, I., VLADOIU, M. C., ZHANG, J., ABEYSUNDARA, N., PRZELICKI, D., SKOWRON, P., GAUER, N., LUU, B., DANIELS, C., WU, X., FORGET, A., MOMIN, A., WANG, J., DONG, W., KIM, S. K., GRAJKOWSKA, W. A., JOUVET, A., FEVRE-MONTANGE, M., GARRE, M. L., NAGESWARA RAO, A. A., GIANNINI, C., KROS, J. M., FRENCH, P. J., JABADO, N., NG, H. K., POON, W. S., EBERHART, C. G., POLLACK, I. F., OLSON, J. M., WEISS, W. A., KUMABE, T., LOPEZ-AGUILAR, E., LACH, B., MASSIMINO, M., VAN MEIR, E. G., RUBIN, J. B., VIBHAKAR, R., CHAMBLESS, L. B., KIJIMA, N., KLEKNER, A., BOGNAR, L., CHAN, J. A., FARIA, C. C., RAGOUSSIS, J., PFISTER, S. M., GOLDENBERG, A., WECHSLER-REYA, R. J., BAILEY, S. D., GARZIA, L., MORRISSY, A. S., MARRA, M. A., HUANG, X., MALKIN, D., AYRAULT, O., RAMASWAMY, V., PUENTE, X. S., CALARCO, J. A., STEIN, L. & TAYLOR, M. D. 2019. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature,* 574**,** 707-711.

SVOBODOVA, E., KUBIKOVA, J. & SVOBODA, P. 2016. Production of small RNAs by mammalian Dicer. *Pflugers Arch,* 468**,** 1089-102.

SWEVERS, L., LIU, J. & SMAGGHE, G. 2018. Defense Mechanisms against Viral Infection in Drosophila: RNAi and Non-RNAi. *Viruses,* 10.

SZELENBERGER, R., KACPRZAK, M., SALUK-BIJAK, J., ZIELINSKA, M. & BIJAK, M. 2019. Plasma MicroRNA as a novel diagnostic. *Clin Chim Acta,* 499**,** 98-107.

TABARA, H., SARKISSIAN, M., KELLY, W. G., FLEENOR, J., GRISHOK, A., TIMMONS, L., FIRE, A. & MELLO, C. C. 1999. The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell,* 99**,** 123-32.

TAN, G. C., CHAN, E., MOLNAR, A., SARKAR, R., ALEXIEVA, D., ISA, I. M., ROBINSON, S., ZHANG, S., ELLIS, P., LANGFORD, C. F., GUILLOT, P. V., CHANDRASHEKRAN, A., FISK, N. M., CASTELLANO, L., MEISTER, G., WINSTON, R. M., CUI, W., BAULCOMBE, D. & DIBB, N. J. 2014. 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res,* 42**,** 9424-35.

TELONIS, A. G., MAGEE, R., LOHER, P., CHERVONEVA, I., LONDIN, E. & RIGOUTSOS, I. 2017. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res,* 45**,** 2973-2985.

TOMASELLO, L., DISTEFANO, R., NIGITA, G. & CROCE, C. M. 2021. The MicroRNA Family Gets Wider: The IsomiRs Classification and Role. *Front Cell Dev Biol,* 9**,** 668648.
TOWLER, B. P. & NEWBURY, S. F. 2018. Regulation of cytoplasmic RNA stability: Lessons from Drosophila. *Wiley Interdiscip Rev RNA,* 9**,** e1499.
TRABER, G. M. & YU, A. M. 2023. RNAi-Based Therapeutics and Novel RNA Bioengineering Technologies. *J Pharmacol Exp Ther,* 384**,** 133-154.
TREIBER, T., TREIBER, N. & MEISTER, G. 2019a. Publisher Correction: Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nat Rev Mol Cell Biol,* 20**,** 321.
TREIBER, T., TREIBER, N. & MEISTER, G. 2019b. Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nat Rev Mol Cell Biol,* 20**,** 5-20.
TUSCHL, T., ZAMORE, P. D., LEHMANN, R., BARTEL, D. P. & SHARP, P. A. 1999. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev,* 13**,** 3191-7.
VAN DER KWAST, R., WOUDENBERG, T., QUAX, P. H. A. & NOSSENT, A. Y. 2020. MicroRNA-411 and Its 5'-IsomiR Have Distinct Targets and Functions and Are Differentially Regulated in the Vasculature under Ischemia. *Mol Ther,* 28**,** 157-170.
VAN NIEUWERBURGH, F., SOETAERT, S., PODSHIVALOVA, K., AY-LIN WANG, E., SCHAFFER, L., DEFORCE, D., SALOMON, D. R., HEAD, S. R. & ORDOUKHANIAN, P. 2011. Quantitative bias in Illumina TruSeq and a novel post amplification barcoding strategy for multiplexed DNA and small RNA deep sequencing. *PLoS One,* 6**,** e26969.
VANICEK, J. 2014. Predicting the Genes Regulated by MicroRNAs via Binding Sites in the 3' Untranslated and Coding Regions. *Chimia (Aarau),* 68**,** 629-32.
VICKERS, K. C., PALMISANO, B. T., SHOUCRI, B. M., SHAMBUREK, R. D. & REMALEY, A. T. 2011. MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat Cell Biol,* 13**,** 423-33.
VILIMOVA, M. & PFEFFER, S. 2023. Post-transcriptional regulation of polycistronic microRNAs. *Wiley Interdiscip Rev RNA,* 14**,** e1749.
VISHLAGHI, N. & LISSE, T. S. 2020. Dicer- and Bulge Stem Cell-Dependent MicroRNAs During Induced Anagen Hair Follicle Development. *Front Cell Dev Biol,* 8**,** 338.
WANG, H. & CHEN, Y. H. 2021. microRNA Biomarkers in Clinical Study. *Biomolecules,* 11.
WAPPENSCHMIDT, B., BECKER, A. A., HAUKE, J., WEBER, U., ENGERT, S., KÖHLER, J., KAST, K., ARNOLD, N., RHIEM, K., HAHNEN, E., MEINDL, A. & SCHMUTZLER, R. K. 2012. Analysis of 30 putative BRCA1 splicing mutations in hereditary breast and ovarian cancer families identifies exonic splice site mutations that escape in silico prediction. *PLoS One,* 7**,** e50800.
WARREN, M., LORD, C. J., MASABANDA, J., GRIFFIN, D. & ASHWORTH, A. 2003. Phenotypic effects of heterozygosity for a BRCA2 mutation. *Hum Mol Genet,* 12**,** 2645-56.
WHILEY, P. J., DE LA HOYA, M., THOMASSEN, M., BECKER, A., BRANDÃO, R., PEDERSEN, I. S., MONTAGNA, M., MENÉNDEZ, M., QUILES, F., GUTIÉRREZ-ENRÍQUEZ, S., DE LEENEER, K., TENÉS, A., MONTALBAN, G., TSERPELIS, D., YOSHIMATSU, T., TIRAPO, C., RAPONI, M., CALDES, T., BLANCO, A., SANTAMARIÑA, M., GUIDUGLI, L., DE GARIBAY, G. R., WONG, M., TANCREDI, M., FACHAL, L., DING, Y. C., KRUSE, T., LATTIMORE, V., KWONG, A., CHAN, T. L., COLOMBO, M., DE VECCHI, G., CALIGO, M., BARALLE, D., LÁZARO, C., COUCH, F., RADICE, P., SOUTHEY, M. C., NEUHAUSEN, S., HOUDAYER, C., FACKENTHAL, J., HANSEN, T. V., VEGA, A.,

DIEZ, O., BLOK, R., CLAES, K., WAPPENSCHMIDT, B., WALKER, L., SPURDLE, A. B. & BROWN, M. A. 2014. Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem,* 60**,** 341-52.

WIGHTMAN, B., HA, I. & RUVKUN, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-62.

WILCZYNSKA, A., GILLEN, S. L., SCHMIDT, T., MEIJER, H. A., JUKES-JONES, R., LANGLAIS, C., KOPRA, K., LU, W. T., GODFREY, J. D., HAWLEY, B. R., HODGE, K., ZANIVAN, S., CAIN, K., LE QUESNE, J. & BUSHELL, M. 2019. eIF4A2 drives repression of translation at initiation by Ccr4-Not through purine-rich motifs in the 5'UTR. *Genome Biol,* 20**,** 262.

WILKS, C., GADDIPATI, P., NELLORE, A. & LANGMEAD, B. 2018. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics,* 34**,** 114-116.

WILTON, S. D., FALL, A. M., HARDING, P. L., MCCLOREY, G., COLEMAN, C. & FLETCHER, S. 2007. Antisense oligonucleotide-induced exon skipping across the human dystrophin gene transcript. *Mol Ther,* 15**,** 1288-96.

WONG, A. C. H. & RASKO, J. E. J. 2021. Splice and Dice: Intronic microRNAs, Splicing and Cancer. *Biomedicines,* 9.

WONG, R. K. Y., MACMAHON, M., WOODSIDE, J. V. & SIMPSON, D. A. 2019. A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. *BMC Genomics,* 20**,** 446.

WOOD, Z. A., SABATINI, R. S. & HAJDUK, S. L. 2004. RNA ligase; picking up the pieces. *Mol Cell,* 13**,** 455-6.

WOODS, S., CHARLTON, S., CHEUNG, K., HAO, Y., SOUL, J., REYNARD, L. N., CROWE, N., SWINGLER, T. E., SKELTON, A. J., PIRÓG, K. A., MILES, C. G., TSOMPANI, D., JACKSON, R. M., DALMAY, T., CLARK, I. M., BARTER, M. J. & YOUNG, D. A. 2020. microRNA-seq of cartilage reveals an overabundance of miR-140-3p which contains functional isomiRs. *Rna,* 26**,** 1575-1588.

WRIGHT, C., RAJPUROHIT, A., BURKE, E. E., WILLIAMS, C., COLLADO-TORRES, L., KIMOS, M., BRANDON, N. J., CROSS, A. J., JAFFE, A. E., WEINBERGER, D. R. & SHIN, J. H. 2019. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genomics,* 20**,** 513.

WU, H., YE, C., RAMIREZ, D. & MANJUNATH, N. 2009. Alternative processing of primary microRNA transcripts by Drosha generates 5' end variation of mature microRNA. *PLoS One,* 4**,** e7566.

XIAO, Y. & MACRAE, I. J. 2022. The molecular mechanism of microRNA duplex selectivity of Arabidopsis ARGONAUTE10. *Nucleic Acids Res,* 50**,** 10041-10052.

XIONG, P., SCHNEIDER, R. F., HULSEY, C. D., MEYER, A. & FRANCHINI, P. 2019. Conservation and novelty in the microRNA genomic landscape of hyperdiverse cichlid fishes. *Sci Rep,* 9**,** 13848.

YAMANE, D., SELITSKY, S. R., SHIMAKAMI, T., LI, Y., ZHOU, M., HONDA, M., SETHUPATHY, P. & LEMON, S. M. 2017. Differential hepatitis C virus RNA target site selection and host factor activities of naturally occurring miR-122 3′ variants. *Nucleic Acids Res,* 45**,** 4743-4755.

YAN, L., LIANG, M., HOU, X., ZHANG, Y., ZHANG, H., GUO, Z., JINYU, J., FENG, Z. & MEI, Z. 2019. The role of microRNA-16 in the pathogenesis of autoimmune diseases: A comprehensive review. *Biomed Pharmacother,* 112**,** 108583.

YANG, D., LU, H. & ERICKSON, J. W. 2000. Evidence that processed small dsRNAs may mediate sequence-specific mRNA degradation during RNAi in Drosophila embryos. *Curr Biol,* 10**,** 1191-200.

YERI, A., COURTRIGHT, A., DANIELSON, K., HUTCHINS, E., ALSOP, E., CARLSON, E., HSIEH, M., ZIEGLER, O., DAS, A., SHAH, R. V., ROZOWSKY, J., DAS, S. & VAN KEUREN-JENSEN, K. 2018. Evaluation of commercially available small RNASeq library preparation kits using low input RNA. *BMC Genomics,* 19**,** 331.

YI, R., QIN, Y., MACARA, I. G. & CULLEN, B. R. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev,* 17**,** 3011-6.

YING, S. Y. & LIN, S. L. 2006. Current perspectives in intronic micro RNAs (miRNAs). *J Biomed Sci,* 13**,** 5-15.

YOSHIDA, K., SANADA, M., SHIRAISHI, Y., NOWAK, D., NAGATA, Y., YAMAMOTO, R., SATO, Y., SATO-OTSUBO, A., KON, A., NAGASAKI, M., CHALKIDIS, G., SUZUKI, Y., SHIOSAKA, M., KAWAHATA, R., YAMAGUCHI, T., OTSU, M., OBARA, N., SAKATA-YANAGIMOTO, M., ISHIYAMA, K., MORI, H., NOLTE, F., HOFMANN, W. K., MIYAWAKI, S., SUGANO, S., HAFERLACH, C., KOEFFLER, H. P., SHIH, L. Y., HAFERLACH, T., CHIBA, S., NAKAUCHI, H., MIYANO, S. & OGAWA, S. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature,* 478**,** 64-9.

YOUNG, C., CAFFREY, M., JANTON, C. & KOBAYASHI, T. 2022. Reversing the miRNA -5p/-3p stoichiometry reveals physiological roles and targets of miR-140-3p miRNAs. *Rna,* 28**,** 854-864.

YU, F., PILLMAN, K. A., NEILSEN, C. T., TOUBIA, J., LAWRENCE, D. M., TSYKIN, A., GANTIER, M. P., CALLEN, D. F., GOODALL, G. J. & BRACKEN, C. P. 2017. Naturally existing isoforms of miR-222 have distinct functions. *Nucleic Acids Res,* 45**,** 11371-11385.

ZAMORE, P. D., TUSCHL, T., SHARP, P. A. & BARTEL, D. P. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell,* 101**,** 25-33.

ZELLI, V., COMPAGNONI, C., CAPELLI, R., CORRENTE, A., CORNICE, J., VECCHIOTTI, D., DI PADOVA, M., ZAZZERONI, F., ALESSE, E. & TESSITORE, A. 2021. Emerging Role of isomiRs in Cancer: State of the Art and Recent Advances. *Genes (Basel),* 12.

ZENG, Y., YI, R. & CULLEN, B. R. 2003. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc Natl Acad Sci U S A,* 100**,** 9779-84.

ZHANG, J., LIEU, Y. K., ALI, A. M., PENSON, A., REGGIO, K. S., RABADAN, R., RAZA, A., MUKHERJEE, S. & MANLEY, J. L. 2015. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proc Natl Acad Sci U S A,* 112**,** E4726-34.

ZHANG, X., LIU, F., YANG, F., MENG, Z. & ZENG, Y. 2021. Selectivity of Exportin 5 binding to human precursor microRNAs. *RNA Biol,* 18**,** 730-737.

ZHANG, Z., LEE, J. E., RIEMONDY, K., ANDERSON, E. M. & YI, R. 2013. High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol,* 14**,** R109.

ZHOU, X., DUAN, X., QIAN, J. & LI, F. 2009. Abundant conserved microRNA target sites in the 5'-untranslated region and coding sequence. *Genetica,* 137**,** 159-64.

ZHU, Y. Y., MACHLEDER, E. M., CHENCHIK, A., LI, R. & SIEBERT, P. D. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques,* 30**,** 892-7.

Appendix 1

Appendix 1

Enzymes for circularisation

Table 4.1 (Copied below) summarises the data that is presented in Fig 4.2 (Chapter 4) and Figs S1 to S8 (see below)

Copy of Table 4.1

	Circligase I		Circligase II		RNA ligase 2	
	+ ATP	- ATP	+ ATP	- ATP	+ ATP	- ATP
214-RA3-RA5 (DNA-DNA-DNA)	✓	✓	✓	✓	X	X
214-RA3-RA5 (RNA-DNA-RNA)	(✓)	?	(✓)	(✓)	✓	X
NNRA3RA5 (DNA-RNA)	✓	✓	✓	✓	✓	X
214-RA5NN (RNA-DNA)	✓	(✓)	✓	✓	X	X

Copy of Table 4.1

Fig S1 shows that RNA ligase 2 requires ATP in order to work (compare lanes 7 and 8) and also confirms that the expected inhibitory effect of readily available 3' blockers 2'-O-Methyl Ribose C (2'OMe-rC) and 3' Amino Modifier C3 (AmC3) upon circularisation (compare lanes 1 and 2 with control lane 3). Lane 6 shows that pre-adenylation can also cause circularisation.

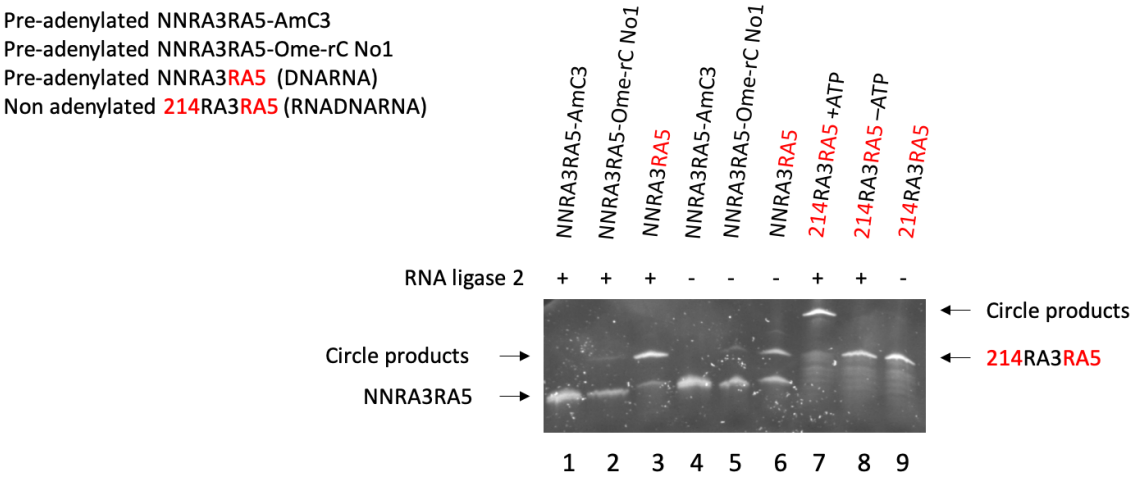


Figure S1 ATP-dependent RNA ligase 2 used to test with different substrate block end. Lane 1, 100 ng of the NNRA3RA5-AmC3 treated with 10 U/µl of RNA ligase 2 at 37°C for 1 hour. Lane 2, 100 ng of the NNRA3RA5-Ome-rC treated with 10 U/µl of RNA ligase 2 at 37°C for 1 hour. Lane 3, 100 ng of NNRA3RA5 (not blocked) treated with 10 U/µl of RNA ligase 2 at 37°C for 1 hour. Lanes 4 to 6 are a repeat of lanes 1 to 3 but without RNA ligase 2 treatment. Lane 7, 100

ng of 214RA3RA5 treated with 10 U/µl RNA ligase 2 at 37°C for 1 hour. Lane 8, 100 ng of 214RA3RA5 treated with 10 U/µl of RNA ligase 2 but without ATP. Lane 9, 100 ng of 214RA3RA5 untreated. All reactions except lane 8 contained 400 µM ATP. Red font indicates that the oligo is made of RNA.

Main findings of Figs S2 to S7.

Circligase II enzyme circularised the DNA molecule 214RA3RA5 irrespective of ATP, as expected (Fig S3 lane 2, Fig S5 lane 7, Chapter 4 Discussion). Also as expected the DNA oligo 214RA3RA5 was not circularised by RNA ligase 2 (Fig S2 lane 12, Fig S3 lane 3, Fig S4 lane 4). The oligo RNA: DNA: RNA molecule 214RA3RA5 was circularised by RNA ligase 2 but only in the presence of ATP (Fig S1 lanes 7,8). The ratio of linear to circular products in lanes 6 to 8 of Fig S2 shows that RNA ligases 1 and 2 could circularize 214RA3RA5 more efficiently than Circligase I. Circligase II could also weakly circularize 214RA3RA5 (Fig S3 lane 5, Fig S4 lane 8). The cloning vector oligo NNRA3RA5 was circularised by all three enzymes (Fig S1 lane 3 and Fig S5, lanes 1-4) and as expected circularization by RNA ligase 2 required ATP (Fig S5 lane 9, Fig S1 lane 3). Oligo 214RA5NN could only be circularised by the circligases (Fig S3 lanes 7,8) but not RNA ligase 2 (Fig S3 lanes 9). Unexpectedly, Circligase I could circularise the DNA oligo214RA3RA5 without ATP and also unexpectedly this circularization was on occasion inhibited by ATP (Figure S4, compare lanes 1 and 2). Similar results are seen for the circularization of the DNA oligo NNRA3RA5 (Fig S6). There was also evidence of oligo degradation by Circligase I on occasion (Fig S7, Fig S3 lane 4, Fig S4 lanes 6,7).

Lane 2 of Fig S2 shows that circularization of NNRA3RA5-2OMe by circligase I was not blocked by a 2OMe group compared to RNA ligases 1 and 2 (lanes 3 and 4).

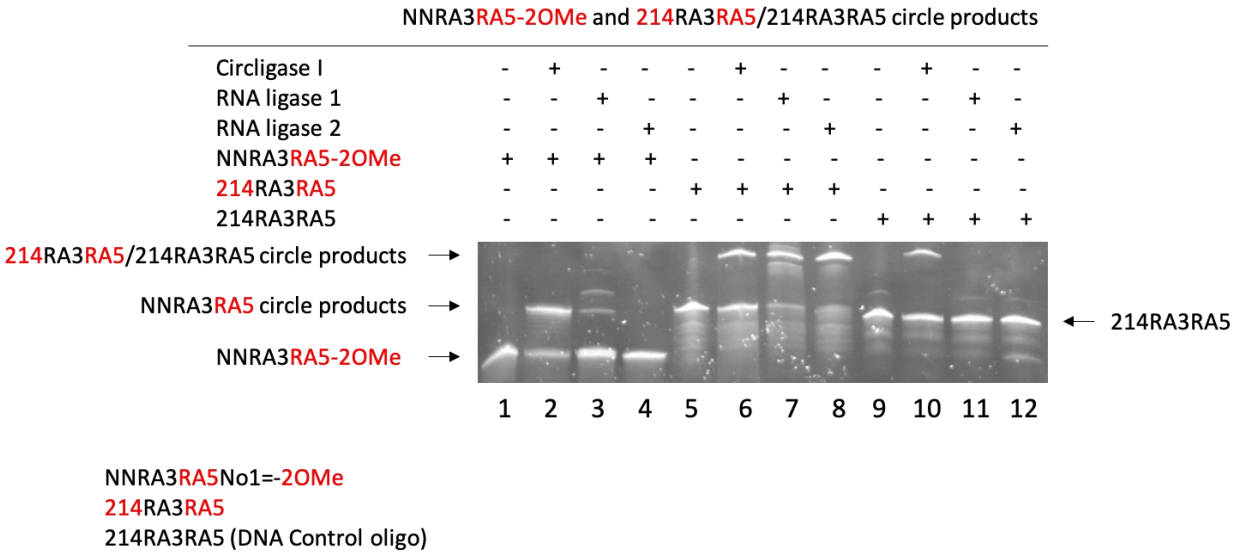


Figure S2 Circularization of NNRA3RA5-2OMe, 214RA3RA5 and 214RA3RA5. All lanes included ATP. Lanes 1 to 4, 100 ng of NNRA3RA5-2OMe: untreated (lane 1), treated with 100 U/µl Circligase I at 60°C for 1 hour (lane 2), treated with 10 U/µl RNA liagse 1 at 25°C for 2 hours (lane 3) and treated with 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 4). Lanes 5 to 8 are the same as lanes 1 to 4 except that used 100 ng of 214RA3RA5 instead of NNRA3RA5-2OMe. Similarly, lanes 9 to 12 are a repeat of lanes 1 to 4 except that 100 ng of 214RA3RA5 was used.

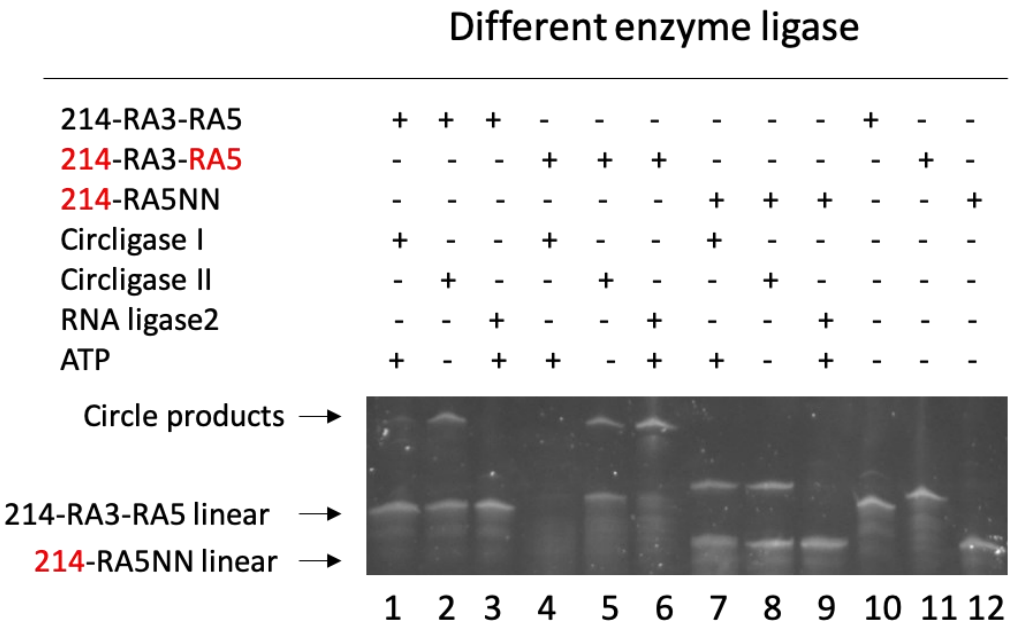


Figure S3 Three different ligase enzymes used to test 214RA3RA5, 214RA3RA5 and 214-RA5NN circular ligase. Lanes 1 to 3, 100 ng of 214RA3RA5 treated with 100 U/µl circligase I plus 1 mM ATP at 60°C for 1 hour (lane 1), 100 U/µl circligase II at 60°C for 1 hour (lane 2), 10 U/µl RNA ligase 2 at 37°C for 1 hour. Lanes 4 to 5, 100 ng of 214RA3RA5 treated with 100 U/µl Circligase I with 1 mM ATP at 60°C for 1 hour (lane 4), 100 U/µl Circligase II at 60°C for 1 hour (lane 5), 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 6). Lanes 7 to 9, 100 ng of 214-RA5NN treated with 100 U/µl circligase I with 1 mM ATP at 60°C for 1 hour (lane 7), 100 U/µl circligase II without 1 mM ATP at 60°C for 1 hour (lane 8), 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 9). Lanes 10 to 12, 100 ng of 214RA3RA5, 214RA3RA5 and 214-RA5NN only.

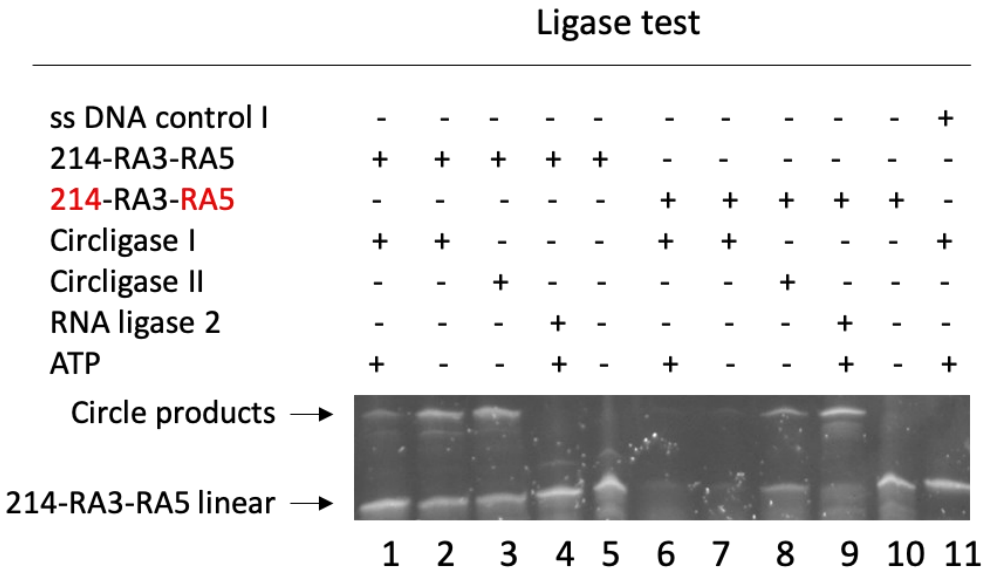


Figure S4 The 214RA3RA5 and 214RA3RA5 were used to test the circligase I, circligase II and RNA ligase 2 circular ligation efficiency. Lanes 1 to 5, 100 ng of 214RA3RA5 treated with 100 U/µl circligase I with or without 1 mM ATP at 60°C for 1 hour (lanes 1, 2), treated 100 U/µl circligase II at 60°C for 1 hour (lane 3), treated with 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 4), untreated control (lane 5). Lanes 6 to 10, 100 ng of 214RA3RA5 treated with 100 U/µl circligase I with and without 1 mM ATP at 60°C for 1 hour (lanes 6,7), treated with 100 U/µl circligase II at 60°C for 1 hour (lane 8), treated with 10 U/µl RNA ligase 2 at 37°C for 1 hour (lane 9), untreated (lane 10). Lane 11, ss DNA control treated with 100 U/µl circligase I and 1 mM ATP.

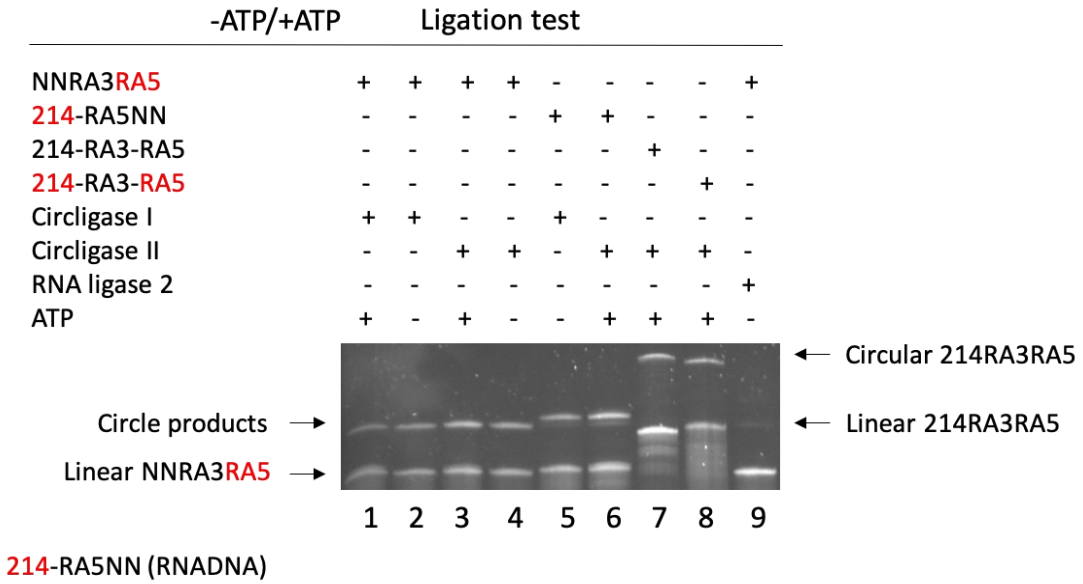


Figure S5 NNRA3RA5, 214RA5NN, 214RA3RA5 and 214RA3RA5 with three different ligase enzyme efficiency tests with or without ATP. Lanes 1 to 4, 100 ng of NNRA3RA5 was treated with 100 U/µl circligase I (lanes 1, 2) or 100 U/µl circligase II (lanes 3, 4) with or without 1 mM ATP at 60 ˚C for 1 hour. Lanes 5,6 100 ng of 214RA5NN was treated with 100 U/µl circligase I without 1 mM ATP buffer (lane 5) and circligase II with 1 mM ATP buffer (lane 6) at 60°C for 1 hour. 100 ng of 214RA3RA5 and 214RA3RA5 (lanes 7,8) were treated with 100 U/µl circligase II with 1 mM ATP buffer at 60°C for 1 hour. Lane 9, 100 ng of NNRA3RA5 treated with 10 U/µl RNA ligase 2 without 400 µM ATP in the buffer at 37°C for 1 hour.

Fig S6 shows that the smaller DNA oligo NNRA3RA5 could be circularized by Circligase I in the presence or absence of ATP.

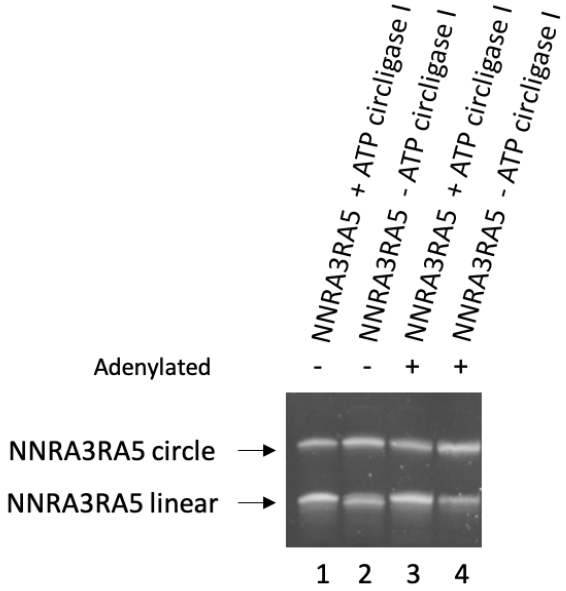


Figure S6 NNRA3RA5 circular ligation by circligase I with or without ATP. 100 ng of NNRA3RA5 treated with 100 U/µl circligase I with 1 mM ATP (lane 1), treated with 100 U/µl circligase I without ATP (lane 2). 100 ng of pre-adenylated NNRA3RA5 treated with 100 U/µl circligase I with 1 mM ATP (lane 3), 100 ng of pre-adenylated NNRA3RA5 treated with 100 U/µl circligase I without ATP (lane 4). All reactions were at 60°C for 1 hour.

Fig S7 lanes 6 to 10 show that circligase II efficiency was not greatly affected by different bases at the 3' end of RA5. A comparison of lanes 1 to 5 with 6 to 10 indicates that circligase I might degrade oligos in the absence of ATP.

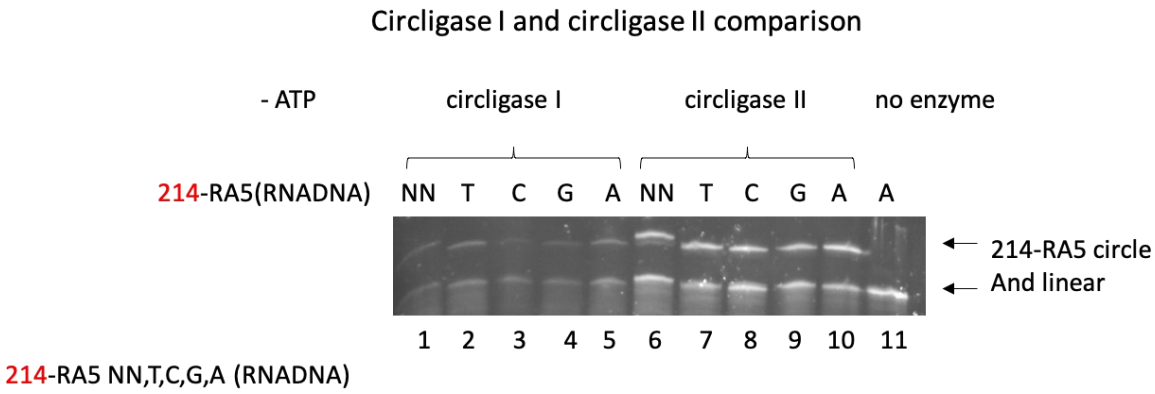


Figure S7 214-RA5 used to test the ligation efficiency by using circligase I and II without ATP. Lanes 1-5, 100 ng of 214-RA5 with the indicated different bases at the 3' end were treated with 100 U/µl circligase I without 1 mM ATP at 60°C for 1 hour. Lanes 6-10 are a repeat but with 100 U/µl circligase II without 1 mM ATP at 60°C for 1 hour. Lane 11 is a control without enzyme treatment.

Fig S8 shows that circularization of the cloning vector 214RA5NN by either circligase I or circligase II was largely complete within 5 minutes of treatment. In this experiment there was no evident loss of oligo in the lanes treated with circligase I. Circligase I is normally supplied with ATP (see Discussion). Consequently, the buffer without ATP was made by us. Further experiments would be required to clarify this.

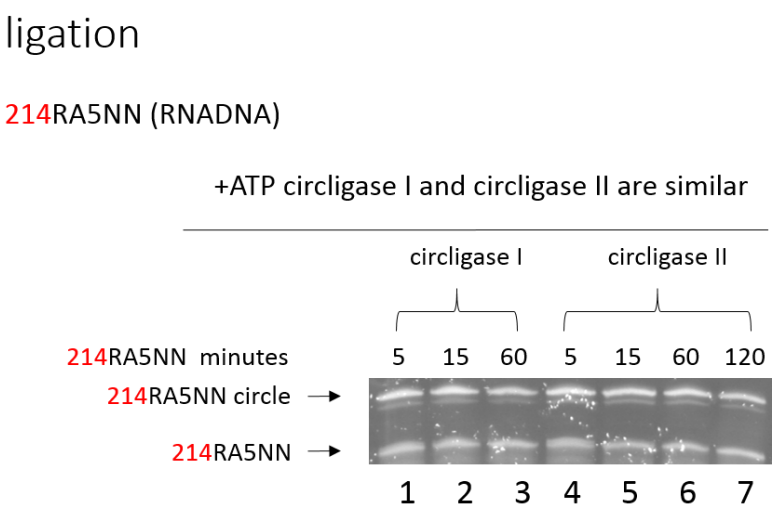


Figure S8 The time course for 214RA5NN with ATP by using circligase I and II. Lanes 1, 2 and 3, 100 ng of 214RA5NN was treated with 100 U/µl circligase I and 1 mM ATP buffer for 5, 15 or 60 minutes. Lanes 4, 5, 6 and 7, 100 ng of 214RA5NN was treated with 100 U/µl circligase II with 1 mM ATP buffer for 5, 15, 60 and 120 minutes.

2OMe block

Figure S9 shows that a 2OMe group on the 3' base of RA3RA5 blocks RNA ligase 2 but not circligase II. This together with the results of Fig S2 (lanes 1 to 4) raises the possibility of using NNRA3RA5-2OMe or NNRA3RA5-2OMe as a miRNA cloning vector where the microRNA ligation is catalysed by RNA ligase 2 delta and circularisation is by a circligase. This would require further experimental confirmation.

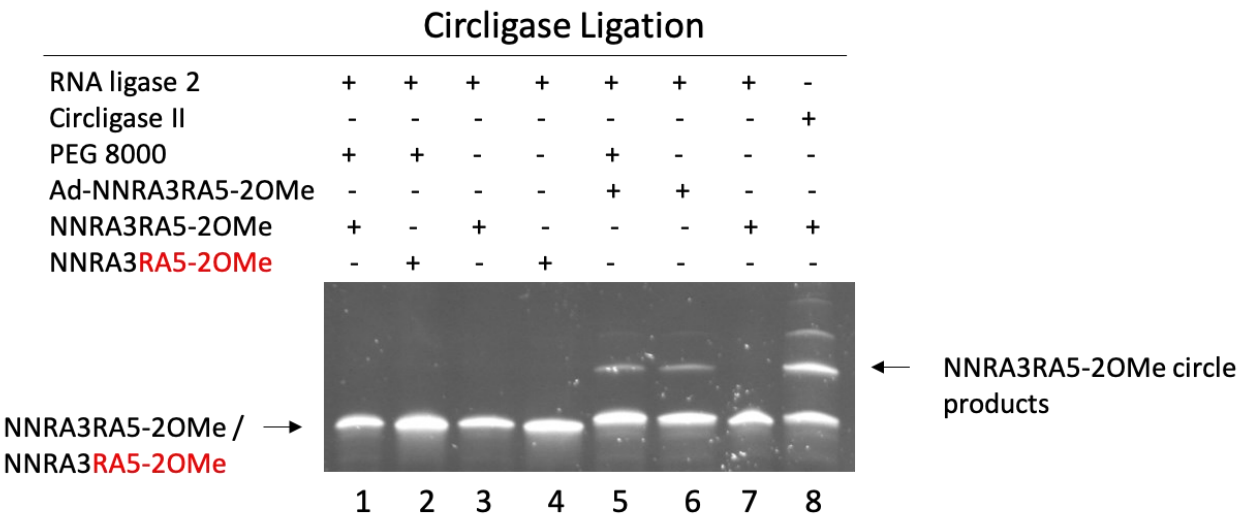


Figure S9 The circligase II and RNA ligase 2 used to test circle ligase efficiency substrate with -2OMe block end. Lane 1,2 100 ng of NNRA3RA5-2OMe and 100 ng of NNRA3RA5-2OMe were treated with 1 µl of 10 U/µl RNA ligase 2 with 10% (w/v) PEG 8000 at 37°C for 1 hour. Lanes 3,4 100 ng of NNRA3RA5-2OMe and 100 ng of NNRA3RA5-2OMe were treated with 1 µl of 10 U/µl RNA ligase 2 without 10% (w/v) PEG 8000 at 37°C for 1 hour. Lanes 5 and 6, 100 ng of pre-adenylated NNRA3RA5-2OMe was treated with 1 µl of 10 U/µl RNA ligase 2 with (lane 5) or

without (lane 6) 10% (w/v) PEG 8000 at 37°C for 1 hour. Lane 7, 100 ng of NNRA3RA5-2OMe was treated with 1 µl of 10 U/µl RNA ligase 2 at 37°C for 1 hour. Lane 8, 100 ng of NNRA3RA5-2OMe was treated with 1 µl of 100 U/µl Circligase II at 60°C for 1 hour.

MicroRNA cloning

Figures S10 to S12 are similar experiments to Fig 4.5 (Chapter 4). Fig S10 (lanes 3 to 6) show that the RNA oligo mir-214 can be cloned into Ad-NNRA3RA5NN-3'P (using the protocol outlined in Fig 4.12). Similarly, mir-214 can be weakly cloned into the DNA oligo Ad-NNRA3RA5NN-3'P by circligase II (Fig S10 lane 8) but not RNA ligase 1 (Fig S10 lane 7).

Fig S11 lanes 3 and 4 repeat the observation (Fig 4.5 lanes 5 to 7) that linear Ad-NNRA3RA5NN-3'P can be ligated with miR-101-1-3p or mir-214 to form likely circular products, as indicated. A comparison of lanes 3 and 5 show that the likely circular forms of NNRA3RA5NN-3'P and of mir-214 + Ad-NNRA3RA5NN-3'P are more resistant to T7 exonuclease treatment, as would be expected for circular products.

Fig S12 lanes 1 to 3 repeat the observation that mir-214 cloned into Ad-NNRA3RA5NN-3'P by RNA ligase 2 delta is more efficiently circularised with RNA ligase 2 rather than circligase II.

Lanes 5 to 8 repeat the circligase is best used for circularising mir-214 cloned into the DNA molecule Ad-NNRA3RA5NN.

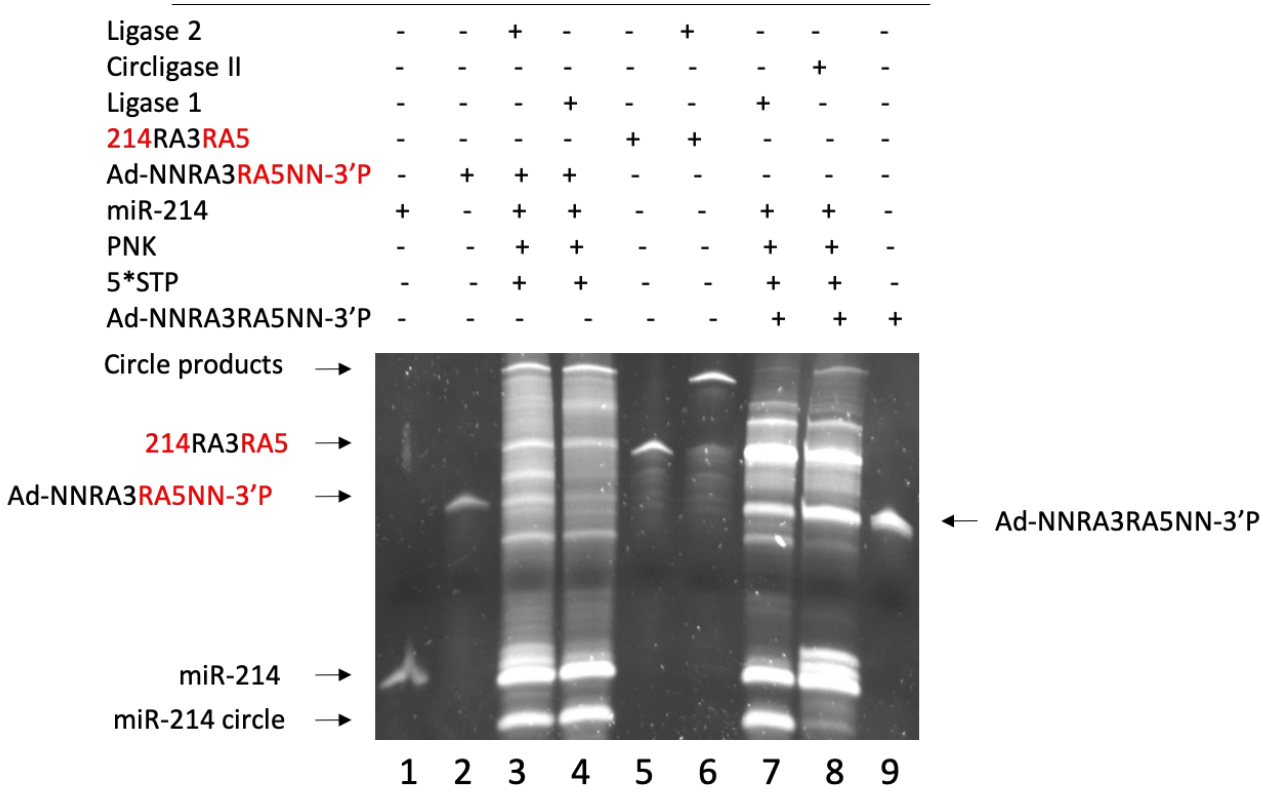


Figure S10 Circligase II, ligases 1 and 2 were used to test mir-214 ligated with Ad-NNRA3RA5NN-3'P for circularised efficiency. Lane 1, 60 ng of mir-214 linear without adding enzyme; lane 2, 60 ng of Ad-NNRA3RA5NN-3'P linear without adding enzyme. Lane 3, 600 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzyme; Lane 4, 600 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase

1. Lane 5, 60 ng of 214RA3RA5 linear without adding enzyme; lane 6, 100 ng of 214RA3RA5 was circularised by using 10 U/µl RNA ligase 2. Lane 7, 600 ng of pre-adenylated NNRA3RA5NN-3'P was ligated with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 1 enzyme, Lane 8, 600 ng of pre-adenylated NNRA3RA5NN-3'P was ligated with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then using 100 U/µl circligase II. Lane 9, 100 ng of Ad-NNRA3RA5NN-3'P linear without adding enzyme. 5*STP is the same as the STP oligo used in Chapter 3 except that it has a 5'blocking group.

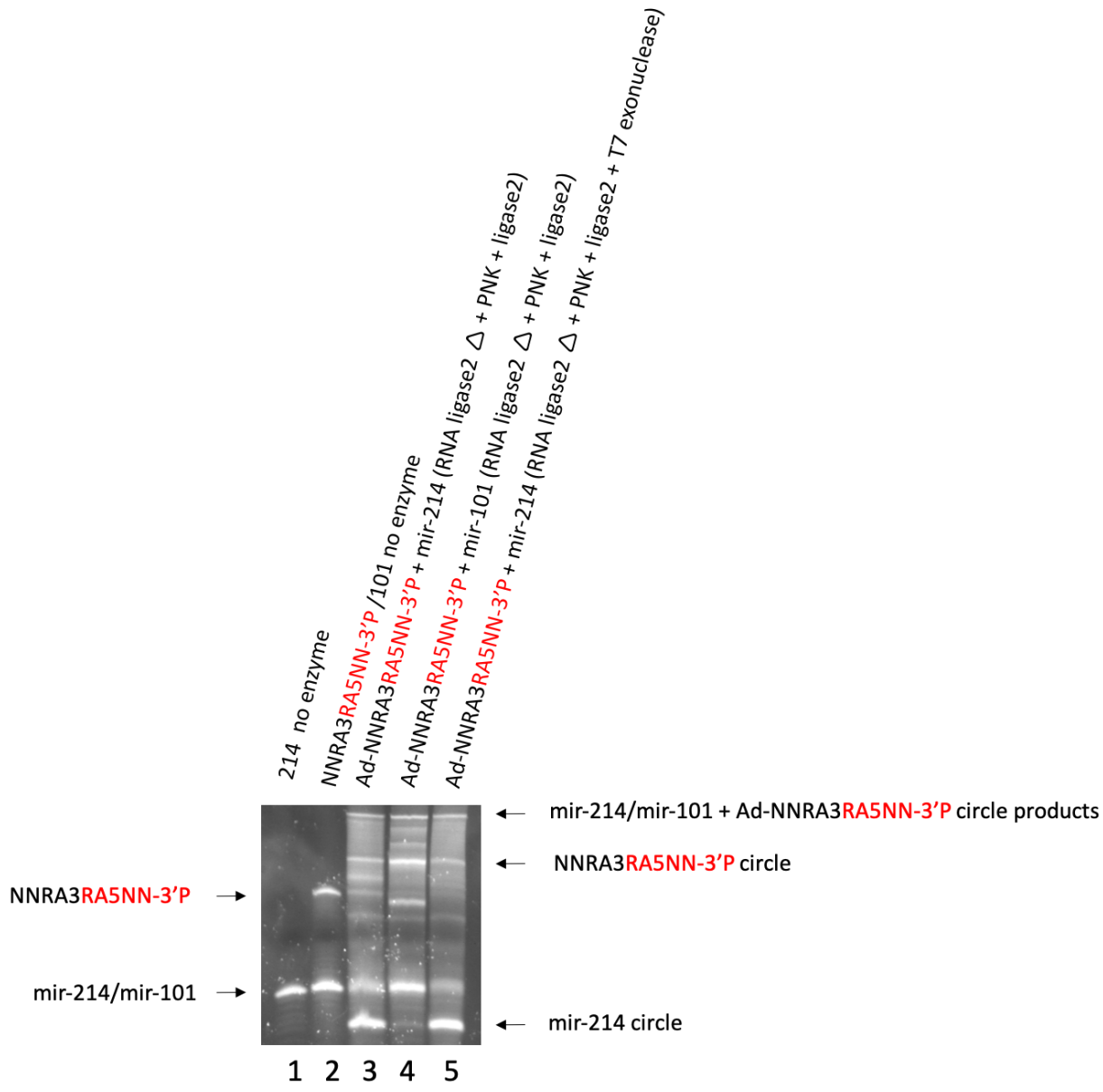


Figure S11 Mir-214/101 was circularized Ad-NNRA3RA5NN-3'P and used T7 exonuclease to clean the background. Lane 1, 100 ng of mir-214 no treatment; lane 2, 100 ng of miR-101-1-3p and 100 ng of NNRA3RA5NN-3'P no treatment. Lane 3, 100 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 100 ng of mir-214 by using 200 U of 1 µl RNA ligase 2

Δ at 25°C for 1 hour + 20 units of PNK at 37°C for 30 minutes and then 1 µl of 10 U/µl RNA ligase 2 enzyme at 37°C for 30 minutes; Lane 4, 100 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 100 ng of miR-101-1-3p by using 200 U/µl RNA ligase 2 Δ at 25°C for 1 hour + 20 units of PNK at 37°C for 30 minutes and then 10 U/µl RNA ligase 2 enzyme at 37°C for 30 minutes as upper arrow indicates; lane 5, 100 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 100 ng of mir-214 by using 200 U/µl RNA ligase 2 Δ at 25°C for 1 hour + 20 units of PNK at 37°C for 30 minutes and then 10 U/µl RNA ligase 2 enzyme at 37°C for 30 minutes, which was then treated with 10 U/µl T7 exonuclease was incubated at 25°C for 30 minutes, finally made the clean faint NNA3RA5NN-3'P circle and mir-214 circle as arrows indicate.

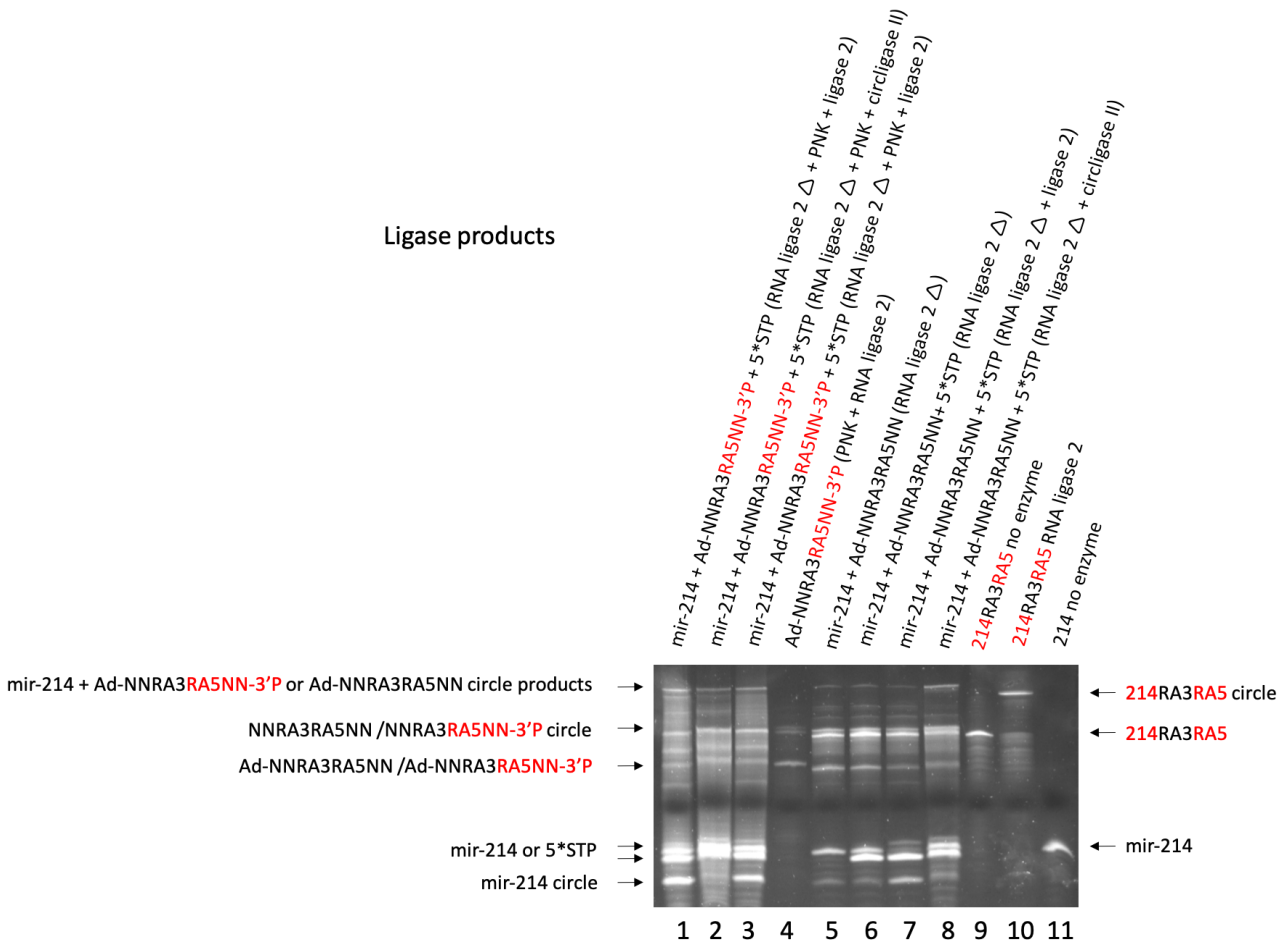


Figure S12 Mir-214 was circularised with Ad-NNRA3RA5NN-3'P or Ad-NNRA3RA5NN-3'P by RNA ligase 2 or circligase II. Lane 1, 600 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzymes, Lane 2, 600 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 600 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then using 100 U/µl circligase II. Lane 1, enzyme was inactivated only after adding 5*STP but lane 2 the enzyme was inactivated after each step. Lane 3, 400 ng of pre-adenylated NNRA3RA5NN-3'P was circularised with 400 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ + 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzymes, and lane 3 the experiment was inactivated after each step. Lane 4, 400 ng of pre-adenylated NNRA3RA5NN-3'P was used by adding 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzymes, but the Ad-NNRA3RA5NN-3'P was not made a circle. Lane 5, 400 ng of Ad-NNRA3RA5NN was ligated with 400 ng of mir-214 by using 200 U/µl RNA ligase 2 Δ. Lane 6, 400 ng of Ad-NNRA3RA5NN was ligated with 400 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ. Lane 7, 400 ng of Ad-NNRA3RA5NN was ligated with 400 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ and then 10 U/µl RNA ligase 2 enzymes. Lane 8, 400 ng of Ad-NNRA3RA5NN was circularised with 400 ng of mir-214 by adding 200 ng of 5*STP with 200 U/µl RNA ligase 2 Δ and then 100 U/µl circligase II as the upper band indicates. Lane 5,6,7, experiments did not make any circle except Ad-NNRA3RA5NN self-circles, and lane 7 was inactivated after adding 5*STP, but lane 8 the experiment was inactivated after each step. Lane 7, mir-214 ligated Ad-NNRA3RA5NN did not convert into circle with ligase enzyme. Lane 9, 100 ng of 214RA3RA5 linear without adding enzyme. Lane 10, 100 ng of 214RA3RA5 was circularised by using 10 U/µl RNA ligase 2 as the upper band indicates. Lane 11, 100 ng of mir-214 linear without adding enzyme.

Exonuclease resistance of circular DNA.

A comparison of lanes 3 and 5 of Fig S11 indicates that circular products are more resistant to exonuclease treatment, as expected. A comparison of lanes 3 and 4 of Fig S13 shows that exonuclease I degrades linear NNRA3RA5NN-3'P, whereas a comparison of lanes 5 to 7 indicates that the circular form of NNRA3RA5NN-3'P is resistant to exonuclease I and to exonuclease T (T7 Exonuclease). Fig S14 is a preliminary experiment which shows that the 5' to 3' exonuclease T cannot degrade the STP oligo (because it lacks a 5'P) nor 5*STP (this is the same as STP except that it has a 5' block), whereas the 3' to 5' exonuclease I can degrade both STP and 5*STP as expected.

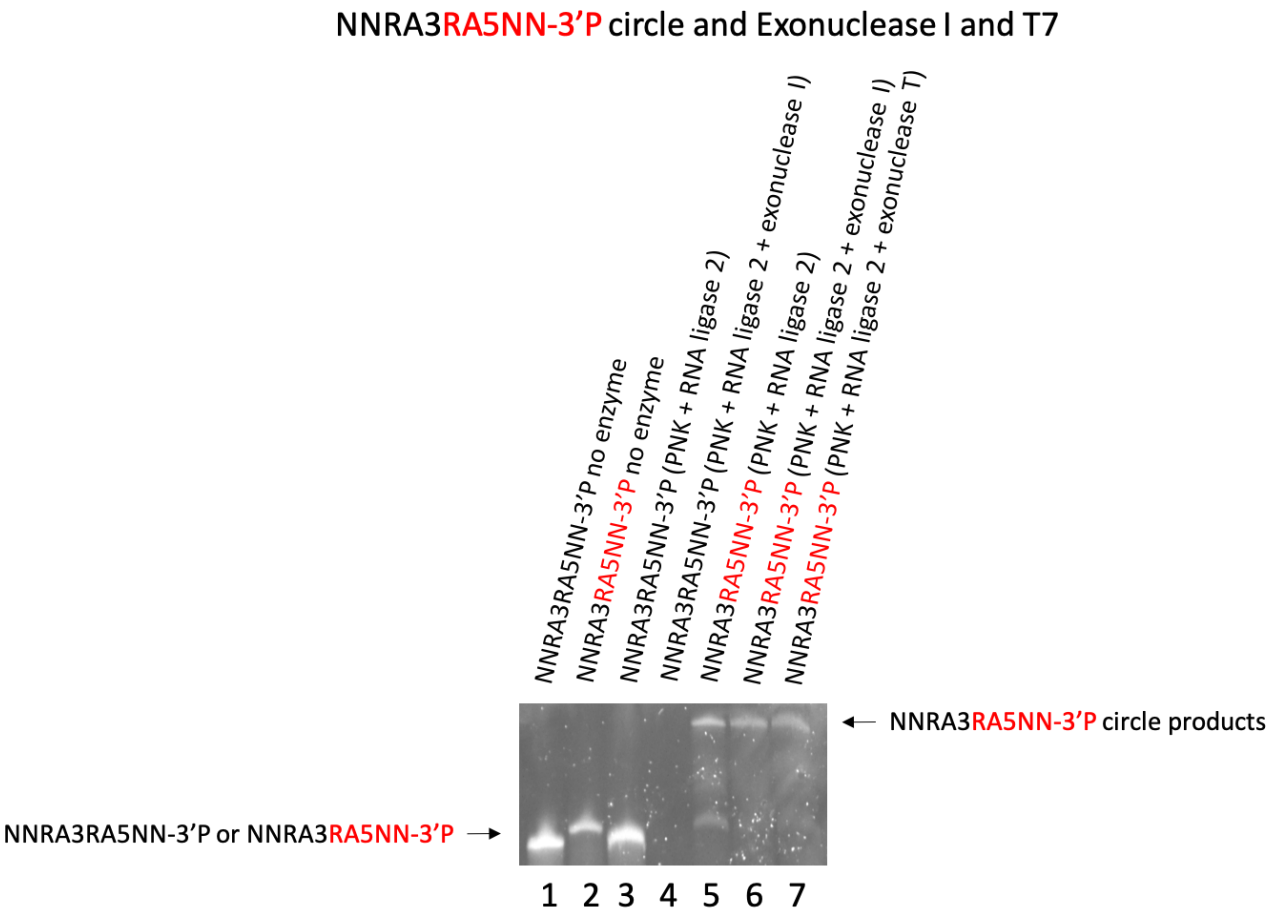


Figure S13 NNRA3RA5NN-3'P was used to test the exonuclease I and T7 Exonuclease clean the background efficiency. Lane 1, 100 ng of NNRA3RA5NN-3'P was a linear without adding any enzymes; lane 2, 100 ng of NNRA3RA5NN-3'P was a linear without adding any enzymes. Lane 3, 100 ng of NNRA3RA5NN-3'P was treated with 10 U/µl PNK and 10 U/µl RNA Ligase 2 enzyme, and the lane 4, 100 ng of NNRA3RA5NN-3'P was added one more step than lane 3 treated with 20 U of exonuclease I incubated at 37°C for 45 minutes to make a clean background. The lane 5, 100 ng of NNRA3RA5NN-3'P was treated with 10 U/µl PNK and then 10 U/µl RNA ligase 2 enzyme, which made it circle; the lane 6 and lane 7, 100 ng of NNRA3RA5NN-3'P was added one more step than lane 5, in lane 6 NNRA3RA5NN-3'P was treated with 20 U of exonuclease I incubated at 37°C for 45 minutes, and in lane 7

NNRA3RA5NN-3'P was treated with 10 U/µl T7 Exonuclease at 25°C for 30 minutes, finally made clean faint NNRA3RA5NN-3'P circle products.

Mir-214 and Ad-NNRA3RA5NN/Ad-NNRA3RA5NN-3'P made circle products with circligase II or RNA ligase 2 enzyme and also used 214RA3RA5 as a control (Figure S12). We want to know which exonuclease I or T 7 exonuclease will degrade the STP or 5*STP oligos, and then I set up the experiment (Figure below).

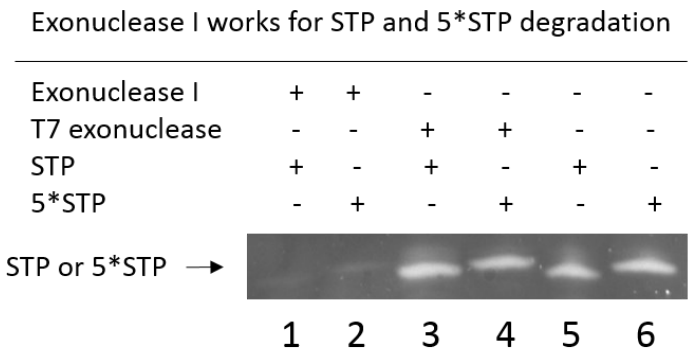


Figure S14 The Exonuclease I and T7 exonuclease were used to test clean up the STP and 5*STP. Lane 1,2, 100 ng of STP and 100 ng of 5*STP were degraded by adding 20 U Exonuclease I which make a clean background, but lane 3,4, 10 U/µl T7 exonuclease Incubate at 25°C for 30 minutes did not degradate 100 ng of STP or 100 ng of 5*STP which is the same as no enzymes. Lane 5,6, 100 ng of STP and 100 ng of 5*STP linear were without any degradation enzymes.

5*STP oligo

Fig S15 is related to Fig3.8 of chapter 3 and confirms that STP and a close derivative 5*STP are ligated equally well to Ad-NNRA3RA5NN-3'P. This is a prerequisite for step 3 of our miRNA cloning protocol illustrated in Fig 4.12 of Chapter 4. We do not know the nature of the extra bands in lanes 7 to 9, this was not pursued.

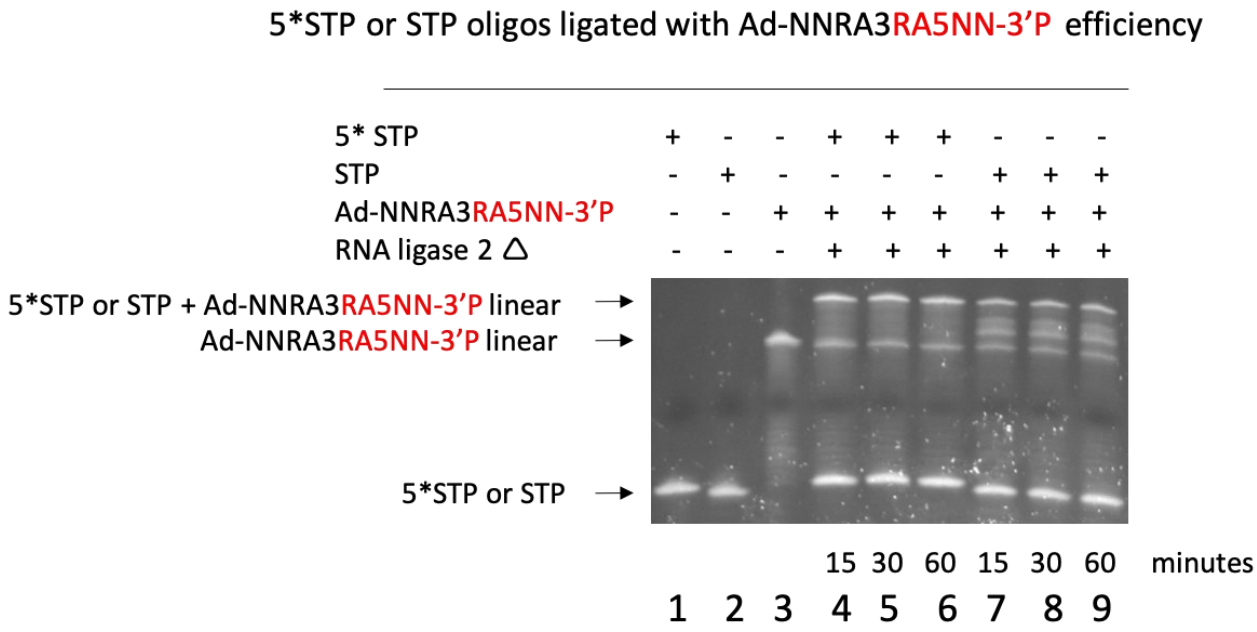


Figure S15 The time course used to test 5*STP or STP ligated with Ad-NNRA3RA5NN-3'P efficiency. Lane 1, 100 ng of 5*STP linear without adding enzyme; lane 2, 100 ng of STP linear without adding enzyme. Lane 3, 100 ng of Ad-NNRA3RA5NN-3'P linear without adding enzyme. Lane 4,5,6, 100 ng of 5*STP was ligated with 100 ng of Ad-NNRA3RA5NN-3'P by using 200 U/µl

RNA ligase 2 Δ for time course experiment at 15 minutes, 30 minutes, 60 minutes. Lane 7,8,9, 100 ng of STP was ligated with 100 ng of Ad-NNRA3RA5NN-3'P by using 200 U/µl RNA ligase 2 Δ for time course experiment at 15 minutes, 30 minutes, 60 minutes. 5*STP is the same as STP (5'GAAUUCCACCACGUUCCCGUGG) except that it has a 5'block (5' Inverted Dideoxy-T /5InvddT/ STP).

Appendix 2

Table S1 A discussion of the shaded examples from Table 5.2 and a list of references for column 2 of Table 5.2.

Table 5.2 Row 5

Chen et al 2006 PMID: 16619214; Thomassen et al., (2012) PMID: 21769658 and Colombo et al., (2013) PMID: 23451180 report that a mutation of the 5'ss 41256138 causes the activation of a 5'css 41256200(hg19) at -62, this deletes 62 bases from the end of exon 7.

	Various 5'ss	3'ss	Intron size	strand	distance of 5'bss from normal 5'ss	reads	Notes
chr17	41277293	41251898	25396	-	-21155	3	
chr17	41277287	41251898	25390	-	-21149	3	
chr17	41277198	41251898	25301	-	-21060	1	Chen et al 2006 PMID: 16
chr17	41267742	41251898	15845	-	-11604	1	
chr17	41258472	41251898	6575	-	-2334	2	Double skip
chr17	41256884	41251898	4987	-	-746	3	Single skip
chr17	41256138	41251898	4241	-	0	97874	CCTTCCTTGgtaaaacca
chr17	41253959	41251898	2062	-	2179	5	

The above information was used to make row 5 of Table 5.2 and shows all splicing events involving the partner 3'ss 41251898 of the mutated normal 5'ss 41256138. This 5'ss also splices to an alternative 3'ss at 41251895, similar results to above (data now shown). There are no background reads for the css at -62 in Snaptron SRAv1 although there is a single read for this event in the larger SRAv2 databse.

The main contradiction between experiment and background splicing in both SRAv1 and SRAv2 is that there are more background reads for single (3 reads) and double exon skips (2 reads) than for the experimentally identified css at -62. The RT-PCR primers listed in Table S2 of Thomassen et al (2012) anneal to exon 5 and 11 so they would have seen the single and double skips of exon 7 and of exons 6 & 7 predicted by Snaptron if they had occurred. Similarly Colombo et al 2013 used primers located in exons 5 and 8. There is a report of both exon 7 skipping and activation of css-62 by this 5'ss mutation (Steffensen et al., 2014), using minigene analysis. Overall, the background splicing data underestimates the use of css -62 following mutation of the nearby 5'ss.

Table 5.2 Row 9, bss (41242251) with 115 reads not detected by experiment.

chr17	41277287	41234593	42695-	-34327	1		
chr17	41251791	41234593	17199-	-8831	17		
chr17	41249260	41234593	14668-	-6300	26		
chr17	41247862	41234593	13270-	-4902	18	Thomassen et al (2012)	eport that mutation of th
chr17	41246760	41234593	12168-	-3800	2		
chr17	41243451	41234593	8859-	-491	1		
chr17	41242960	41234593	8368-	0	151941	ACCACTCAGgtaaaaagc	tttttgaagCAGAGGGAT
chr17	41242251	41234593	7659-	709	115		
chr17	41241766	41234593	7174-	1194	1		
chr17	41239979	41234593	5387-	2981	1		
chr17	41239287	41234593	4695-	3673	1		

Thomassen et al., (2012) detected a skip as a result of a mutation of the 5'ss 41242960 and would likely have detected the css at 41242251 if it had been used at a reasonable frequency.

Consequently, the background ss at +709 is probably a false positive, with regards to css potential. The intron is quite large (8368 bases) and so is likely to contain recursive splice sites (see text).

Table 5.2 Row 15

chr17	41234420	41215391	19030	-	-18530	5
chr17	41228504	41215391	13114	-	-12614	1
chr17	41222944	41215391	7554	-	-7054	17
chr17	41219624	41215391	4234	-	-3734	140
chr17	41219471	41215391	4081	-	-3581	1
chr17	41215906	41215391	516	-	-16	179
chr17	41215899	41215391	509	-	-9	1
chr17	41215890	41215391	500	-	0	180038
chr17	41215699	41215391	309	-	191	14

PMID: 23239986 Wappenschmidt et al (2012) report skipping (41219624 to 41215391) following mutation of the 5'ss 41215890+1G to C. Baert et al., 2018 PMID: 29280214 also report similar results. No groups have reported the activation of a css at -16 (179 reads), indicating that it is a false positive.

Table 5.2 Row 16, bss (41215906) and double exon skip not detected by expt.

chr17	41277293	41209153	68141	-	-61944	2		
chr17	41277287	41209153	68135	-	-61938	2		
chr17	41277198	41209153	68046	-	-61849	1		
chr17	41249260	41209153	40108	-	-33911	1		
chr17	41242960	41209153	33808	-	-27611	1		
chr17	41234420	41209153	25268	-	-19071	2		
chr17	41228504	41209153	19352	-	-13155	3		
chr17	41222944	41209153	13792	-	-7595	34		
chr17	41219624	41209153	10472	-	-4275	56		
chr17	41215906	41209153	6754	-	-557	23		
chr17	41215890	41209153	6738	-	-541	1		
chr17	41215349	41209153	6197	-	0	182438	CTGAATGAGgtaagtact	ttctttcagCATGATTTT
chr17	41210349	41209153	1197	-	5000	54	PMID: 23239986 Wapp	
chr17	41210307	41209153	1155	-	5042	2		
chr17	41210265	41209153	1113	-	5084	3		
chr17	41210240	41209153	1088	-	5109	593		
chr17	41210233	41209153	1081	-	5116	3		
chr17	41209911	41209153	759	-	5438	1		
chr17	41209402	41209153	250	-	5947	1		
chr17	41209264	41209153	112	-	6085	38		

Wappenschmidt et al (2012) report that ivs19+2T>G causes exon skipping (Snaptron has 1 read for this, see above) but did not detect bss 41215906, which has 23 reads. Use of this predicted css would also cause exon skipping plus the additional deletion of the last 16 bases of the upstream exon. The primer used to detect the skip would also have detected the predicted css 41215906, if it had been used, so a possible false positive. Also, Snaptron has 56 reads for the double exon skip between 41219624 and 41209153. The oligo listed as being used for the RT-PCR would not have detected this event (if it occurred).

Table 5.2 Row 20

chr17	41251791	41197820	53972 -	-52132	1
chr17	41247862	41197820	50043 -	-48203	6
chr17	41234420	41197820	36601 -	-34761	1
chr17	41228504	41197820	30685 -	-28845	9
chr17	41222944	41197820	25125 -	-23285	1
chr17	41219624	41197820	21805 -	-19965	2
chr17	41215349	41197820	17530 -	-15690	2
chr17	41209068	41197820	11249 -	-9409	13
chr17	41203079	41197820	5260 -	-3420	132
chr17	41201137	41197820	3318 -	-1478	300
chr17	41199659	41197820	1840 -	0	137637
chr17	41199654	41197820	1835 -	5	247
chr17	41199034	41197820	1215 -	625	6
chr17	41198188	41197820	369 -	1471	7

Four groups report that mutations affecting the 5'ss 41199659 caused exon skipping but only Yang et al 2003 detected activation of a css at +5, although they had a stronger RT-PCR band for the exon skip. The snaptron reads of 247 for the +5 css and 300 for the exon skip predicts similar activation of both splicing events. Whiley et al 2011 studied a +5G>C mutation that would have destroyed the +5 css. Ladopolou et al 2002 studied the same G>A mutation subsequently analysed by Yang et al 2003 and Rouleau et al 2010 studied a mutation 14 bases upstream from the end of exon 23. Yang et al 2003 analysed a homozygous mutation and showed that the G>A mutation did not entirely prevent normal intron removal. They were not able to resolve normal splicing from use of the +5css by RT-PCR but did so by sequencing of isolated clones made from the RT-PCR band. Both Ladopolou et al 2002 and Rouleau et al 2010 had the added complication of studying a heterozygous mutation from patient material, which would have added to the difficulties (see above) of detecting activation of the +5 css.

Table 5.2 Row 21. Single exon skip not detected despite having far more reads (2622) than the detected css (5 reads).

chr17	41276033	41275715	319-	-7918	32		
chr17	41276033	41274898	1136-	-7101	3		
chr17	41276033	41273698	2336-	-5901	781		
chr17	41276033	41273694	2340-	-5897	4038		
chr17	41276033	41271949	4085-	-4152	11		
chr17	41276033	41271283	4751-	-3486	340		
chr17	41276033	41269753	6281-	-1956	1		
chr17	41276033	41268941	7093-	-1144	1		
chr17	41276033	41267861	8173-	-64	1		
chr17	41276033	41267836	8198-	-39	2		
chr17	41276033	41267797	8237-	0	112381	TCCCATCTGgtaagtcag	ccctgctagTCTGGAGTT
chr17	41276033	41267790	8244-	7	5	Wappenschmidt et al., (2012). Mutation of -1G to C	
chr17	41276033	41264814	11220-	2983	1	Vreeswijk et al 2009 IVS2 -6T>A (de novo ss).	
chr17	41276033	41264620	11414-	3177	2		
chr17	41276033	41262602	13432-	5195	5		
chr17	41276033	41262598	13436-	5199	21		
chr17	41276033	41262526	13508-	5271	1		
chr17	41276033	41258560	17474+	9237	2		
chr17	41276033	41258551	17483-	9246	2622		
chr17	41276033	41258520	17514-	9277	1		
chr17	41276033	41257977	18057-	9820	1		
chr17	41276033	41257033	19001-	10764	1		
chr17	41276033	41256974	19060-	10823	178		
chr17	41276033	41256279	19755-	11518	1		
chr17	41276033	41246878	29156-	20919	45		
chr17	41276033	41243050	32984-	24747	4		

Wappenschmidt et al (2012). Mutation is -1G to C, which changes the 3' css to GCTACTCTGGAG/TT. Perhaps in addition to inactivating the 3'ss this mutation also increases the strength of the css, which might be why it was detected over the intron skip?

Table 5.2 Row 24. Css -10 not predicted, plus a ss at -177 with 81 reads not seen.

chr17	41256884	41256456	429-	-177	81		
chr17	41256884	41256279	606-	0	111753	GTTTGGAGTgtaagtgtt	tttacagATGCAAACAG
chr17	41256884	41251898	4987-	4381	3	Gutiérrez-Enríquez et al.,	
chr17	41256884	41251895	4990-	4384	7		

Gutiérrez-Enríquez et al., (2017) and Chen et al 2006 report the activation of a 3'css ten bases downstream of 41256279 following its mutation (ivs6 -1G>T and ivs6-2delA). Primers used were also capable of detecting a possible skip between 41256884 and 41251898.

Css they see is quite close to the mutation, which may account for its strong showing? Additional note: this css matches a bss in the larger SRAv2 database with 3 reads (see below). The bss at -177 with 81 reads is a likely false positive.

chr17	43104867	43104439	429	-	-177	560
chr17	43104867	43104262	606	-	0	300364
chr17	43104867	43104252	616	-	10	3
chr17	43104867	43104233	635	-	29	1
chr17	43104867	43099881	4987	-	4381	15
chr17	43104867	43099878	4990	-	4384	12

Table 5.2 Row 25. Single exon skip not detected despite having far more reads (83) than the reported css (4).

chr17	41256138	41254083	2056	-	-2185	1
chr17	41256138	41254011	2128	-	-2113	1
chr17	41256138	41251967	4172	-	-69	4
chr17	41256138	41251898	4241	-	0	97874
chr17	41256138	41251895	4244	-	3	34012
chr17	41256138	41249607	6532	-	2291	2
chr17	41256138	41249307	6832	-	2591	83
chr17	41256138	41247940	8199	-	3958	3
chr17	41256138	41246878	9261	-	5020	65
chr17	41256138	41243050	13089	-	8848	1
chr17	41256138	41197740	58399	-	54158	1

PMID: 10323242 A ten base deletion (41251922 to 41251912) activates the 3'css 41251967. The primers that were used for RT-PCR would not have detected a possible exon skip between 41256138 and 41249307.

Table 5.2 Row 31, possible false positive bss with 3 reads at 41209360 not detected.

Wappenschmidt et al (2012) PMID: 23239986 report strong exon skipping (41215349 to 4120315) and weaker activation of the 3'css 41209140 due to a mutation of the 3'ss 41209153-1 G to T.

chr17	41215349	41215225	125	-	-6072	1
chr17	41215349	41214575	775	-	-5422	1
chr17	41215349	41213312	2038	-	-4159	1
chr17	41215349	41210387	4963	-	-1234	216
chr17	41215349	41210384	4966	-	-1231	417
chr17	41215349	41209360	5990	-	-207	3
chr17	41215349	41209153	6197	-	0	182438
chr17	41215349	41209150	6200	+	3	9
chr17	41215349	41209140	6210	-	13	1
chr17	41215349	41203135	12215	-	6018	52
chr17	41215349	41199721	15629	-	9432	2
chr17	41215349	41197820	17530	-	11333	2
chr17	41215349	41197650	17700	-	11503	1

The strongest effect was exon skipping which agrees with the Snaptron reads. Bss 41209360 has more reads than bss 4129140 (which was activated as a weak css) but is further away from the mutated intron ss. Possible that the mutation may also have strengthened bss 4129140. Note: bss 41209150 is on the + transcript and therefore not relevant.

Table 5.2 Row 35, possible false positive bss with 26 reads not detected. Baert et al., 2018 PMID: 29280214 report the activation of 3'css 41197809 following mutation of 3'ss 41197820 (-1G>A). The bss that matches the 3'css has 5 reads whereas another bss at -474 with 26 reads was not reported.

chr17	41199659	41199417	243	-	-1597	8
chr17	41199659	41199384	276	-	-1564	111
chr17	41199659	41198899	761	-	-1079	18
chr17	41199659	41198294	1366	-	-474	26
chr17	41199659	41198070	1590	-	-250	5
chr17	41199659	41197820	1840	-	0	137637
chr17	41199659	41197809	1851	-	11	5
chr17	41199659	41196991	2669	-	829	3
chr17	41199659	41196095	3565	-	1725	4
chr17	41199659	41194773	4887	-	3047	2
chr17	41199659	41188436	11224	+	9384	1
chr17	41199659	41188433	11227	-	9387	78

	Mutated splice site	References
	5' ss	
1	41276033 (exon 2)	Baert et al 2018 PMID: 29280214; Colombo et al 2013 PMID: 23451180
2	41267742 (exon 3)	Baert et al 2018 PMID: 29280214 Brose et al., 2004 PMID: 15345110
3	41258472 (exon 5)	Yang et al 2003 PMID: 12915465;Mene´ndez et al 2012 PMID: 21735045;Thery et al 2011 PMID: 21673748; Raponi et al 2011 PMID: 21309043, Sanz et al 2011 PMID: 20215541; Claes et al 2002
4	41256884 (exon 6)	Thomassen et al 2012 PMID: 21769658
5	41256138 (exon 7)	Thomassen et al 2012 PMID: 21769658; Chen et al 2006 PMID: 16619214; Colombo et al 2013 PMID: 23451180 PMID: 23451180
6	41251791 (exon 8)	Pyne et al 1999 PMID: 10479726 Colombo et al (2013)
7	41249260 (exon 9)	Whiley et al 2011 PMID: 21394826 PMID: 21394826
8	41243451 (exon 11)	Wappenschmidt et al 2012 PMID: 23239986
9	41242960 (exon 12)	Thomassen et al 2012 PMID: 21769658.
10	41234420 (exon 13)	Mene´ndez et al 2012 PMID: 21735045, Thomassen et al 2012 PMID: 21769658
11	41228504 (exon 14)	Mene´ndez et al 2012 PMID: 21735045; Santos et al 2014 PMID: 22684231
12	41226347 (exon 15)	Whiley et al 2011 PMID: 21394826 PMID: 21394826; Baert et al 2018 PMID: 29280214
13	41222944 (exon 16)	Wappenschmidt et al 2012 PMID: 23239986;Colombo et al 2013 PMID: 23451180, Baert et al 2018 PMID: 29280214; Scholl et al 1999 PMID: 10406662
14	41219624 (exon 17)	Brose et al 2004 PMID: 15345110 report the skip and weak css activation; Mene´ndez et al (2012) PMID: 21735045 report just Ahlborn et al 2015 PMID: 25724305 PMID: 25724305; Thomassen et al 2012 PMID: 21769658 reported only the 5'css.
15	41215890 (exon 18)	The skip; Wappenschmidt et al 2012 PMID: 23239986; Baert et al 2018 PMID: 29280214.
16	41215349 (exon 19)	Wappenschmidt et al 2012 PMID: 23239986.
17	41209068 (exon 20)	Elstrodt et al (2006) PMID: 16397213; Sanz et al.,(2010) PMID: 20215541; Tesoriero et al 2005 PMID: 16211554 PMID: 16211554; Colombo et al 2013 PMID: 23451180 – skip only.
18	41203079 (exon 21)	Ahlborn et al 2015 PMID: 25724305 Colombo et al 2013 PMID: 23451180
19	41201137 (exon 22)	Wappenschmidt et al 2012 PMID: 23239986,; Rouleau et al 2010 PMID: 20875879; Jarhelle et al 2017 PMID: 27495310 – detected only the skip. Eposito et al 2016 PMID: 28009814 detected only the css
20	41199659 (exon 23)	Yang et al 2003 PMID: 12915465; Whiley et al 2011 PMID: 21394826; Rouleau et al 2010 PMID: 20875879; Ladopoulou et al (2002) PMID: 12034536. Only Yang et al 2003 detected the css weakly.
	3'ss	
21	41267797 (exon 3)	Wappenschmidt et al 2012 PMID: 23239986
22	41258551 (exon 5)	Wappenschmidt et al 2012 PMID: 23239986; Tesoriero et al 2005 PMID: 16211554; Claes et al 2002 PMID 12037674
23	41256974 (exon 6)	Colombo et al 2013 PMID: 23451180; Jarhelle et al (2017) PMID: 27495310; Friedman et al (1994) PMID: 7894493.
24	41256279 (exon 7)	Gutiérrez-Enríquez et al (2017) PMID: 18712473.
25	41251898 (exon 8)	Li et al 1999 PMID: 10323242 (primers used would not have detected the skip)
26	41247940 (exon 10)	Tesoriero et al 2005 PMID: 16211554
27	41246878 (exon 11)	Keaton et al.,(2003) PMID: 14513821
28	41219713 (exon 17)	Colombo et al 2013 PMID: 23451180; Quiles et al (2016) PMID: 21990134.
29	41215969 (exon 18)	Wappenschmidt et al 2012 PMID: 23239986
30	41215391 (exon 19)	Wappenschmidt et al 2012 PMID: 23239986
31	41209153 (exon 20)	Wappenschmidt et al 2012 PMID: 23239986
32	41203135 (exon 21)	Wappenschmidt et al 2012 PMID: 23239986; Colombo et al 2013 PMID: 23451180; Thomassen et al 2012 PMID: 21769658 All groups report exon skipping, first two groups also report weak css activation.
33	41201212 (exon 22)	Wappenschmidt et al 2012 PMID: 23239986: Ahlborn et al 2015 PMID: 25724305
34	41199721 (exon 23)	Ahlborn et al 2015 PMID: 25724305
35	41197820 (exon 24)	Baert et al 2018 PMID: 29280214.