

Generation of a murine ES cell system
deficient in microRNA processing for the
identification of microRNA targets

Matthew P. Davis

The Wellcome Trust Sanger Institute
University of Cambridge
Hughes Hall College

This dissertation is submitted for the degree of Doctor of Philosophy
31st March 2009

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this thesis has been submitted for any other qualification.

This thesis is within the 300-page limit laid down by the Biology Degree Committee.

Matthew P Davis

Abstract

MicroRNAs (miRNAs) are 21-22nt RNA molecules that regulate mRNAs, generally by triggering their degradation or blocking their translation. This effect is mediated via direct binding of the miRNA to its mRNA target at sites of partial complementarity. The number of miRNAs annotated in miRBase has grown rapidly in the last decade. There are now 695 human miRNAs and 488 mouse miRNAs. The significance of miRNA mediated post-transcriptional regulation has led to rapid advances in our understanding of miRNA expression, biogenesis and functional mechanism. However, with miRNAs predicted to regulate up to 60% of the human genome, there is a necessity for the development of methods to identify miRNA target sets on a large scale. It is increasingly evident that miRNAs can be functional components of large regulatory networks. The complexity of these associations is compounded by the ability of multiple miRNAs to regulate the same target mRNA simultaneously. It is also understood that miRNAs with a high degree of sequence similarity at their 5' end may be functionally redundant; this makes the analysis of target associations more challenging.

To address these problems I have developed a system in mouse embryonic stem (ES) cells to simply and rapidly derive gene lists enriched for miRNA targets. DGCR8 is a doublestranded RNA binding protein essential for the first cleavage of miRNA primary transcripts in the canonical miRNA processing pathway and is required for the maturation of these miRNAs. I have disrupted miRNA processing by the targeted insertion of a gene trap cassette into the second allele of *Dgcr8* in cell lines that already bear a gene trap within their first allele. This led to a broad reduction of miRNA processing in these cells and a depletion of mature miRNAs. As a

consequence of the disruption of this locus I was able to identify a number of miRNAs that appear to be processed in DGCR8 independent manner.

I proceeded to transfect these cells with ES-cell-expressed miRNA mimics. I used microarrays to identify transcripts that are down regulated as a consequence of the miRNA reintroduction. By comparing transcripts that had been up regulated upon the depletion of *Dgcr8* to this set I was able to create miRNA target lists for mmu-miR-25 and mmu-miR-291a-3p. These lists should be enriched for functionally relevant, co-expressed targets, moderated for miRNA mimic over expression and to a large extent devoid of interference from target saturation and combinatorial regulation. The system should also not be susceptible to problems associated with functional redundancy. In total I identified 25 target candidates for miR-291a-3p and 40 candidates for miR-25. Amongst these genes are a number of oncogenes and tumour suppressor genes in addition to genes involved in cell cycle regulation and extra-cellular signal transduction.

In conclusion it appears that miRNAs play a fundamental role in the regulation of the ES cell transcriptome and as such are deserving of considerable future research. It is my belief that the method presented in this thesis could contribute significantly to this effort by providing substantial and experimentally derived miRNA candidate target lists upon which to base future hypotheses.

Acknowledgements

I would like to begin by thanking my supervisor Dr. Ian Dunham for his advice, guidance, understanding and patience throughout the last four and a half years. Ian has given time and support generously especially when it has been needed the most. I am extremely grateful for his commitment to helping me to the end of my PhD.

In addition I would like to thank Dr. Anton Enright and Dr. Bill Skarnes for their support and help both in and out of the lab and in particular for adopting me when my team left the Institute. This PhD would have been impossible without their experience and knowledge. My thanks also extends to Dr. Eric Miska for his advice as a member of my Thesis Committee and for offering his assistance with the use of Luminex 100 technology.

I also need to acknowledge the members of Team 62; Andy, Cat, Caroline, Charmain, Christoph, Dave, Gayle, Ian, James, Jamil, Jo, John, Lotte, Owen, and Sarah. Their scientific know how helped me throughout my PhD and there were times when the offer of a beer was really just what I needed, even if that beer was at the top of a very muddy field. I know that to single out anyone in particular would be unbelievably complicated so thank you all...

I was fortunate during the course of my PhD that I was able to meet a large number of kind and generous individuals. In particular, I would like to express my gratitude to Dr. Peri Tate and all the members of Teams 107 and 87 for not only allowing me to use their stem cell facilities but also for all of the counsel and tuition that that entailed.

The members of the Enright laboratory at the EBI deserve a special mention for teaching me bioinformatics. I particularly wish to thank Dr. Cei Abreu-Goodger for patiently answering all of my R queries and for practical discussions, guidance and mapping of sequence data. In addition I would like to thank Dr. Stijn van Dongen and Dr. Harpreet Saini for both advice and contributions to my array and sequencing analysis.

My help I received from the laboratories of Dr. Duncan Odom (CRI) and Dr. Eric Miska (Gurdon Institute), both in Cambridge, also thoroughly deserve my thanks. Especially Dr. Claudia Kutter for teaching me to clone small RNA libraries and perform Northern Blots and Dr. Cherie Blenkiron for training me to work with the Luminex 100 system.

I need to acknowledge Dr. Peter Ellis and Dr. Cordelia Langford for hybridising my arrays and the Sanger Institute sequencing groups for the Solexa sequencing of my libraries and for various other small scale sequencing jobs. I would also like to thank Bee Ling Ng for FACS sorting my cell samples and advice on their analysis.

Thank you to Dr. Christina Hedberg-Delouka for helping my PhD to run as smoothly as possible.

Several additional teams have adopted me in the last 2 years, including the teams of Dr. Inês Barroso and Dr. Matthew Hurles. Their lab and office space has been pretty useful and the personal support has been handy too. Thanks. In the same vein I am very grateful to Dr. Alison Coffey for an occasional calming influence and Maureen, Kath, Christine, Dominic and Ian for pink cake, coffee, lunch and for reading bits and pieces I have sent to them. To all of my friends, it has been a while since I have seen them but for their support I owe them all much.

Very special thanks to my family. Without their love and support this PhD would have been so much harder. I am unbelievably grateful. Finally, thank you to Diana. Her patience and understanding has been amazing and her love a great comfort. It would not have been possible without her.

Contents

Declaration	I
Abstract.....	II
Acknowledgements	IV
Figures.....	XII
Tables	XV
Abbreviations.....	XVI

Chapter 1: Introduction.....	1
1.1 miRNAs	1
1.1.1 MicroRNA processing pathway.....	5
1.1.1.1 Pri-miRNAs	6
1.1.1.1.1 Transcription of pri-miRNAs.....	6
1.1.1.1.2 Intronic miRNAs.....	7
1.1.1.2 Cleavage of primary miRNA transcript by the microprocessor protein complex.....	8
1.1.1.2.1 The microprocessor	8
1.1.1.2.2 Recognition and mechanism of pri-miRNA processing.....	10
1.1.1.2.3 Processing of intronic miRNAs	11
1.1.1.2.4 miRNAs within the 3'UTRs of protein coding genes.....	12
1.1.1.2.5 Co-transcriptional miRNA processing.....	12
1.1.1.3 The fate of the precursor miRNA.....	14
1.1.1.3.1 Pre-miRNA processing.....	14
1.1.1.3.2 Dicer associated proteins.....	15
1.1.1.4 Regulation of the miRNA processing pathway.....	16
1.1.1.5 miRNPs, the effector complexes of miRNA mediated regulation.....	17
1.1.1.6 Mirtrons and other exceptions to the canonical rules.....	19
1.1.2 Mechanism of miRNA function.....	21
1.1.2.1 Deadenylation: Destabilisation of targets through the removal of the poly-A tail	22
1.1.2.2 Translational inhibition.....	23
1.1.2.3 A role for the P-body.....	26
1.1.2.4 Reconciling the different mechanisms for miRNA mediated post-transcriptional regulation.....	26
1.1.2.5 Other miRNA mediated regulatory mechanisms.....	27
1.1.3 The rules of miRNA target recognition and target prediction algorithms	29
1.1.4 Endogenous siRNAs.....	37
1.2 Embryonic stem cells	38
1.2.1 Transcriptional networks for maintaining the stem cell state	38
1.2.2 ES cell cycle.....	41
1.2.3 miRNAs and mouse ES Cells.....	42

1.2.3.1 The role of miRNAs in stem cells – Perturbing the processing pathway	42
1.2.3.2 miRNA expression in stem cells.....	44
1.2.3.3 The role of miRNAs in stem cells.....	47
1.2.3.4 The role of miRNAs in stem cells – Lessons from cancer.....	48
1.3 Aims of this project.....	50
Chapter 2: Materials and Methods.....	53
2.1 Centrifuges.....	53
2.2 General techniques.....	53
2.2.1 DNA phenol chloroform extraction	53
2.2.2 DNA ethanol precipitation.....	54
2.2.3 Small RNA isopropanol precipitation.....	54
2.2.4 PCR	54
2.2.4.1 Primers.....	55
2.2.4.2 PCR programmes	56
2.2.5 Colony PCR	56
2.2.6 Gel electrophoresis.....	57
2.2.7 Bacterial culture	57
2.2.7.1 Antibiotic concentration	57
2.2.7.2 Bacterial transformation	58
2.2.8 Sequencing.....	58
2.2.9 Restriction digest	59
2.3 Generating a targeting trap vector.....	59
2.3.1 Long PCR	59
2.3.2 Preparing the pR3R4AsiSI plasmid.....	60
2.3.3 Cloning the <i>Dgcr8</i> fragment.....	61
2.3.4 Preparing the targeted trap vector	61
2.4 ES cell culture.....	62
2.4.1 Splitting cells	63
2.4.1.1 Method 1.....	63
2.4.1.2 Method 2.....	63
2.4.2 Growth conditions for specific cell lines.....	64
2.4.3 Freezing ES cells.....	65
2.4.4 Defrosting cells	65
2.4.5 Subcloning	66
2.4.6 Colony picking.....	66
2.4.7 Gene targeting/electroporation.....	67
2.5 Judging ES cell characteristics.....	68
2.5.1 Xgal staining.....	68
2.5.2 Immuno-staining	68
2.5.2.1 Preparing the slides	68
2.5.2.2 Immuno-staining the slides.....	69
2.5.2.3 Imaging immuno-stained slides	70
2.5.3 Optimised embryoid body (EB) formation protocol	70
2.5.4 Cell cycle assay	71
2.5.4.1 Cell growth.....	71
2.5.4.2 Fluorescence activated cell sorting (FACS).....	72
2.5.4.3 FACS analysis.....	72

2.6 Protein purification and Western blots.....	72
2.6.1 Protein purification.....	72
2.6.2 Western blot	73
2.7 RNA purification.....	74
2.7.1 Trizol	74
2.7.2 DNase treatment.....	75
2.7.3 SV purification.....	75
2.7.4 Harvesting HeLa S3 RNA	76
2.8 RT-PCR	76
2.9 mRNA expression profiling.....	77
2.9.1 Northern blot	77
2.9.1.1 Northern blot preparation	77
2.9.1.2 Northern probes.....	78
2.9.1.3 Hybridisation.....	79
2.9.2 Illumina expression arrays.....	80
2.9.2.1 Expression array preparation	80
2.9.2.2 Expression arrays	80
2.9.3 Computational analysis of expression arrays.....	81
2.9.4 Relationship plots.....	81
2.9.5 Identification of expression changes between cell lines.....	82
2.9.6 KEGG and GO analysis.....	83
2.9.7 Wilcoxon Rank Test of the expression changes of the targets of transcription factors (TFs) when compared to the general expression changes in DGCR8 depleted cell lines	84
2.9.8 Annotation of probes associated with potential miRNA targets	84
2.10 miRNA expression profiling.....	86
2.10.1 miRNA Northern blots	86
2.10.1.1 Preparing the blots.....	86
2.10.1.2 Probe labeling.....	87
2.10.1.3 miRNA Northern blot probes.....	87
2.10.1.4 Hybridisation of miRNA Northern blots.....	88
2.10.2 Luminex 100 analysis of mRNA expression.....	88
2.10.2.1 Running the Luminex 100 system.....	88
2.10.2.2 Computational analysis of the Luminex data.....	90
2.10.3 miRNA expression profiling using Illumina/Solexa high-throughput sequencing.....	90
2.10.3.1 Preparing Illumina/Solexa samples.....	90
2.10.3.2 Computational analysis of the high-throughput sequencing data.....	92
2.11 Optimised transfection protocols	93
2.11.1 LacZ siRNA transfection.....	94
2.11.1.1 siRNAs.....	94
2.11.1.2 Optimised RNAi protocol.....	94
2.11.1.3 LacZ siRNA transfection and slides.....	95
2.11.1.3.1 Slide preparation.....	95
2.11.1.3.2 Imaging slides siRNA transfected slides...	95
2.11.2 miRNA mimic transfection.....	95
2.11.2.1 miRNA mimics	95

2.11.2.2 miRNA transfection protocol.....	96
2.12 Optimisation of transfection conditions	96
2.12.1 siGLO siRNA transfection.....	96
2.12.1 KIF11 siRNA transfection.....	97
2.13 Solutions	98
2.13.1 General laboratory solutions.....	98
2.13.2 Xgal staining solutions	100
2.13.3 Western blot solutions.....	101
2.13.4 miRNA Northern blot solutions.....	102
2.13.5 Cell culture solutions.....	103
Chapter 3: Disrupting the <i>Dgcr8</i> locus.....	106
3.1 Aim.....	106
3.2 Introduction	106
3.3 Disrupting the <i>Dgcr8</i> locus.....	109
3.3.1 Experimental design.....	109
3.3.2 Selecting <i>Dgcr8</i> trapped cell lines	114
3.3.3 Confirmation of <i>Dgcr8</i> gene trapped cell lines.....	115
3.3.4 Subcloning <i>Dgcr8</i> cell lines.....	118
3.3.5 Generating a targeting vector.....	118
3.3.6 Identification of cells with a successfully targeted <i>Dgcr8</i> locus	124
3.3.7 Determining the functional significance of trap insertions through the use of miRNA Northern blots.....	130
3.3.8 Confirmation of the position of inserted targeted traps by gene-trap specific staining.....	134
3.4 Characterisation of <i>Dgcr8</i> expression and a broader spectrum of miRNAs	136
3.4.1 Detection of wild type and truncated <i>Dgcr8</i> transcripts.....	136
3.4.2 The expression of ES cell miRNAs.....	138
3.5 Investigation of ES cell properties in <i>Dgcr8</i> ^{gt1/tm1} and <i>Dgcr8</i> ^{gt2/tm1} cell lines	141
3.5.1 Expression of core ES cell transcription factors.....	141
3.5.1.1 Oct4 expression measured by Western blot.....	141
3.5.1.2 Comparing Oct4 and SOX2 expression by the immuno-staining of cell cultures.....	142
3.5.2 Morphological phenotype of <i>Dgcr8</i> ^{gt1/tm1} cells	145
3.5.3 Investigation of <i>Dgcr8</i> ^{gt1/tm1} and <i>Dgcr8</i> ^{tm1, gt1/+} cell pluripotency	146
3.5.4 Flow sort for cell cycle.....	151
3.6 Discussion.....	153
Chapter 4: The miRNA expression profile of ES cells depleted of functional <i>Dgcr8</i>	156
4.1 Aim.....	156
4.2 Introduction.....	156
4.3 Results	159
4.3.1 Use of the Luminex platform to profile mature miRNA expression.....	159

4.3.2 miRNA expression profiling of cell lines with Illumina/Solexa high throughput sequencing	166
4.4 Discussion.....	186
Chapter 5: The derivation of ES cell miRNA candidate target lists in a background depleted of endogenous miRNA expression.....	190
5.1 Aim.....	190
5.2 Introduction.....	190
5.3 Results	194
5.3.1 A comparison of growth conditions and their effect on cell phenotype	194
5.3.2 Expression profiles of <i>Dgcr8^{gt1,gt1/+}</i> , <i>Dgcr8^{gt1,gt2/+}</i> , <i>Dgcr8^{gt1/tm1}</i> , <i>Dgcr8^{gt2/tm1}</i> and <i>Dgcr8^{+/+}</i> cells.....	202
5.3.2.1 Comparison of the expression profiles of <i>Dgcr8^{gt1,gt1/+}</i> , <i>Dgcr8^{gt1,gt2/+}</i> , <i>Dgcr8^{gt1/tm1}</i> , <i>Dgcr8^{gt2/tm1}</i> and <i>Dgcr8^{+/+}</i> cells	202
5.3.2.2 Functional analysis of genes up regulated upon DGCR8 depletion	208
5.3.2.3 Identifying DGCR8 dependent alterations to the targets of the ES cell core transcriptional network.....	211
5.3.3 Reintroducing miRNA mimics to <i>Dgcr8^{gt1/tm1}</i> cells.....	216
5.3.3.1 Optimisation of the conditions for miRNA reintroduction	216
5.3.3.2 Demonstration of Oct4 expression in transfected cells	219
5.3.3.3 Selecting miRNA mimics to reintroduce into <i>Dgcr8^{gt1/tm1}</i> cells.....	222
5.3.3.4 Transfection time course to identify the optimal time point for cell lysis.....	223
5.3.3.5 miR-25 and miR-291a-3p potential target lists.....	228
5.3.3.6 Comparison of genes in the miR-291a-3p candidate target list to a previously published experimentally predicted miR-290 cluster target list.....	238
5.3.3.7 Candidate targets of particular interest.....	241
5.4 Discussion.....	245
Chapter 6: Discussion	249
6.1 Disruption of the <i>Dgcr8</i> locus to deplete mature miRNAs in mouse ES cells.....	249
6.2 The small RNA profile of <i>Dgcr8^{gt/tm1}</i> and <i>Dgcr8^{gt1,gt/+}</i> cells	250
6.2.1 DGCR8 independent miRNA processing.....	251
6.3 Generating miRNA candidate target lists.....	252
6.3.1 The influence of miRNAs on the ES cell transcriptome.....	252
6.3.2 The re-addition of miRNAs to the DGCR8 deficient ES cells and miRNA candidate target lists.....	254
6.4 Future work.....	256
6.4.1 Expanding the system for the generation of candidate target lists	256

6.4.2 Novel miRNAs and siRNAs expressed in mouse ES cells ..	258
6.4.3 The roles of DGCR8 independent miRNAs	259
6.4.4 Prospects for improved pri-miRNA annotation	259
6.4.5 Extending the investigation of the cellular phenotype upon mature miRNA depletion	260
6.4.6 An investigation into the role of miRNAs in ES cell adhesion	261
6.4.7 Use of a conditional mutation in an alternative system	262
6.5 Conclusion	263

References	264
-------------------------	------------

Appendices	CD
-------------------------	-----------

Appendix A: Maximum read depths for each of the miRNAs both pre and post normalisation in addition to log fold change and adjusted P-values for a comparison between normalised read depths in the heterozygotes and homozygous mutants

Appendix B: Additional GO term and KEGG pathway analyses

Appendix C: Candidate target lists for mmu-miR-25 and mmu-miR-291a-3p

Figures

Figure 1.1: A simplified outline of the miRNA processing pathway, depicting the two cleavage steps required to liberate the mature mRNA.....	5
Figure 1.2: Canonical microRNA processing pathway in vertebrate, with the introduction of mirtronic miRNAs at the pre-miRNA stage	13
Figure 1.3: The mechanisms of miRNA mediated regulation.....	28
Figure 1.4: The miRNA seed region and the definitions of n -mer target sites as given by http://www.targetscan.org	37
Figure 3.1 A: The structure of the ORF of the <i>Dgcr8</i> gene and the positions of inserted traps relative to protein domains.....	111
Figure 3.1 B: The structure of the gene trap cassettes used to disrupt the <i>Dgcr8</i> locus	111
Figure 3.2: Nested RT-PCR to confirm the gene trap position within two BayGenomics <i>Dgcr8</i> ^{gt/+} cell lines.....	116
Figure 3.3: Gene traps and vectors used to disrupt the <i>Dgcr8</i> locus	122
Figure 3.4: Principles of targeting vector construction.....	123
Figure 3.5 A: A schematic of the expected outcomes of the targeted vector insertion and the primers designed to distinguish these events.....	125
Figure 3.5 B: RT-PCR was used to distinguish the genotypes of ES cell lines resulting from the electroporation of the gene trap targeting vector	126
Figure 3.6: Northern blot to judge miRNA expression within a range of cell lines.....	131
Figure 3.7: Xgal staining of cell lines to determine β -geo (β -galactosidase) activity.....	135
Figure 3.8: Northern blot to demonstrate the loss of wild type <i>Dgcr8</i> transcript in <i>Dgcr8</i> ^{gt1/tm1} and <i>Dgcr8</i> ^{gt2/tm1} cell lines and the effect of the traps on transcript length.....	137
Figure 3.9: miRNA Northern blots demonstrating the loss of ES cell miRNA expression in <i>Dgcr8</i> ^{gt1/tm} and <i>Dgcr8</i> ^{gt2/tm1} cells	140
Figure 3.10: Western blot to demonstrate comparable Oct4 expression between cell lines.....	142
Figure 3.11: Immuno-staining of <i>Dgcr8</i> ^{tm1,gt1/+} and <i>Dgcr8</i> ^{gt1/tm1} cells to demonstrate consistent expression of both Oct4 and <i>Sox2</i> within the cultures.....	144
Figure 3.12: An example of the morphological phenotype seen among <i>Dgcr8</i> ^{gt1/tm1} cells when compared to <i>Dgcr8</i> ^{tm1,gt1/+} controls	146
Figure 3.13: Images of EBs cultured on low attachment plates after 8 days of culture in the absence of LIF.....	150
Figure 3.14: Representative images of cell morphologies seen amongst EBs spreading on gelatinised tissue culture plates.....	150
Figure 3.15: ES cells depleted in the expression of <i>Dgcr8</i> and mature miRNAs accumulate in the G1 phase of the cell cycle.....	152
Figure 4.1: A box plot of the raw fluorescence values for each miRNA specific bead derived from the Luminex miRNA profiling system across multiple samples.....	161
Figure 4.2: Box plots of bead/miRNA-associated MFIs following normalization.....	163

Figure 4.3: A comparison of the expression of miRNAs in <i>Dgcr8</i> ^{tm1,gt1/+} and <i>Dgcr8</i> ^{tm1,gt2/+} cells compared to <i>Dgcr8</i> ^{gt1/tm1} and <i>Dgcr8</i> ^{gt2/tm1} cells.....	165
Figure 4.4: The proportion of ncRNA matching reads in each library that are accounted for by each ncRNA type.....	169
Figure 4.5 A: A Comparison of the small RNA libraries derived from the replicate cell lines	171
Figure 4.5 B: A wide selection of ncRNAs are not affected by the depletion of functional <i>Dgcr8</i> and are thus used to normalise the samples ..	172
Figure 4.6: Normalised total maximum sequence depths for each ncRNA species in each cell line.....	174
Figure 4.7: The normalised maximum sequence depth for the library of ncRNAs was compared between samples.....	176
Figure 4.8: A demonstration of the significance of the change in expression of miRBase annotated miRNAs between <i>Dgcr8</i> ^{tm1,gt1/+} and <i>Dgcr8</i> ^{gt1/tm1} genotyped cells.....	178
Figure 4.9: The maximum read depth and log fold changes of miRNAs that are potentially processed in a DGCR8 independent manner in pairwise comparisons between each <i>Dgcr8gt/tm1</i> cell line and their respective heterozygous control	181
Figure 4.10: The maximum depth of reads which match the miR-320 hairpin sequence	184
Figure 4.11: A Northern blot for miR-320 expression in <i>Dgcr8</i> ^{+/+} , <i>Dgcr8</i> ^{tm1,gt1/+} , <i>Dgcr8</i> ^{tm1,gt2/+} , <i>Dgcr8</i> ^{gt1/tm1} and <i>Dgcr8</i> ^{gt2/tm1} cell lines.....	185
Figure 4.12: Maximum read depth for all miRNAs for which the expression pattern has been demonstrated by Northern blot in this thesis	188
Figure 5.1 A: A rootless tree depicting the degree of correlation between all of the mRNA array profiles on the basis of the complete set of normalized probe intensities.....	197
Figure 5.1 B: A rootless tree depicting the degree of correlation between all of the mRNA array profiles on the basis of a refined set of probe intensities	198
Figure 5.2: Sylamer plots to identify miRNA seed sequence enrichment or depletion within up-regulated or down-regulated genes following the depletion of DGCR8, conducted separately using arrays derived from cells cultured under the older culture conditions and the more recent culture methods	201
Figure 5.3: A Venn diagram depicting the number of probes registering significant expression changes between cell types.....	203
Figure 5.4 A: A plot depicting the average expression of each probe in the heterozygote and homozygous mutant cell lines against the log fold change of each probe following the depletion of DGCR8 from the ES cells.....	206
Figure 5.4 B: Sylamer plots to identify miRNA seed sequence enrichment or depletion within up-regulated or down-regulated genes following the depletion of DGCR8 based on expression data derived from cells grown under both the older and the most recent culture conditions	207

Figure 5.4 C: The miRNA seed sequences highlighted as enriched or depleted by Sylamer analysis in gene lists ordered by LFC or t-statistic following the disruption of both alleles of the <i>Dgcr8</i> locus	208
Figure 5.5: Plots to assess the change in the expression of the transcriptional targets of various core transcription factors relative to the other genes following the depletion of DGCR8.....	212
Figure 5.6: Xgal staining results of LacZ siRNA transfected <i>Dgcr8^{gt1/tm1}</i> cells in the 6 well format to be used for miRNA mimic reintroduction experiments.....	219
Figure 5.7: Immuno-staining of transfected <i>Dgcr8^{gt1/tm1}</i> cells with LacZ and Oct4 specific antibodies to demonstrate siRNA transfection does not influence Oct4 expression	221
Figure 5.8: Sylamer plots querying gene lists ordered according to LFC and t-statistic following the transfection of DGCR8 depleted cells (<i>Dgcr8^{gt1/tm1}</i>) with miRNA mimics and lysed according to a time series	227

Tables

Table 1.1: The miRNA naming convention and definitions of relevant terms used throughout this thesis.....	4
Table 2.1: Primers used in during the course of this study.....	55
Table 2.2: Media quantities used to culture ES cells.....	63
Table 2.3: Culture plates and media quantity used for defrosting ES cells.....	66
Table 2.4: The sequences of DNA oligos used to make a size ladder for small RNA gel electrophoresis.....	87
Table 2.5: miRNA Northern blot probe sequences.....	87
Table 3.1: A description of the relative positions of important features within the structure of the <i>Dgcr8</i> transcript.....	112
Table 3.2: A summary of the genotype nomenclature used throughout this thesis.....	113
Table 3.3: A summary of the number of each kind of insertion event resulting from the targeted gene trap electroporation experiment.....	127
Table 3.4: Summary of the phenotypes seen amongst EB colonies after 12 days of culture in the absence of LIF.....	149
Table 4.1: The maximum mappable read depth of each of the small RNA libraries and the number of reads that subsequently map to the Ensembl and miRBase derived non0coding RNAs.....	167
Table 4.2: miRNAs that maintain relatively constant expression between the <i>Dgcr8</i> ^{gt¹,gt⁺} and <i>Dgcr8</i> ^{gt¹/tm¹} genotyped cell lines.....	179
Table 5.1: GO term and KEGG pathway analysis of genes up regulated upon the depletion of DGCR8.....	210
Table 5.2: Probes associated with Oct4, <i>Sox2</i> , <i>Nanog</i> , <i>Klf4</i> and c-Myc that exhibit a significant expression change between <i>Dgcr8</i> ^{gt¹,gt⁺} and <i>Dgcr8</i> ^{gt¹/tm¹} cell lines.....	213
Table 5.3: Tables summarizing the genes that are the suspected targets of mmu-miR-25 and mmu-miR-291a-3p as determined by transfection analysis.....	231

Abbreviations

ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BLAST	Basic Logical Alignment Search Tool
BMP	Bone morphogenic protein
bp	Base pairs
BSA	Bovine serum albumin
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
ChIP-Seq	ChIP-Sequencing
CRI	Cambridge Research Institute
DAS	Distributed annotation system
DDW	Double distilled water
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic acid
DPBS	Dulbecco's PBS
dsRNA	Double stranded RNA
EB	Embryoid body
EDTA	Ethylene-diamine-tetra-acetic acid
EGFP	Enhanced green fluorescent protein
ES cell	Embryonic stem cell
EST	Expressed sequence tag
FACS	Fluorescence activated cell sorting
FBS	Foetal bovine serum
GFP	Green fluorescent protein
GO	Gene ontology
GSK3	Glycogen synthase Kinase 3
IGTC	International Gene Trap Consortium
Indels	Insertion deletions
iPS cell	Induced pluripotent stem cell
IPTG	Isopropyl β -D-1-thiogalactopyranoside
IQR	Inter-quartile range
IRES	Internal ribosome entry site

kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LB	Luria-Bertani
LFC	Log fold change
LIF	Leukemia inhibitory factor
LNA	Locked Nucleic Acid
MAPK	Mitogen-activated protein kinase
MEFs	Mouse embryonic fibroblasts
mRNA	Messenger RNA
MFI	Median fluorescence intensity
miRNP	microRNA-associated ribonucleoprotein complex
miRNAs	microRNAs
ncRNAs	non-coding RNAs
nt	Nucleotides
ORF	Open reading frame
P-body	Processing body
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PI	Propidium Iodide
PITA	Probability of interaction by target accessibility
pol	RNA polymerase
polyA	polyadenylation
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
pSILAC	Pulsed stable isotope labeling with amino acids in cell culture
qRT-PCR	Quantitative reverse transcriptase polymerase chain reaction
RA	Retinoic acid
RACE	Rapid amplification of complementary DNA ends
RNA	Ribonucleic Acid
RNAi	RNA interference
RNP	Ribonucleoprotein complex

RISC	RNA-induced silencing complex
rRNA	Ribosomal RNA
RT-PCR	Reverse transcriptase polymerase chain reaction
SAPE	Streptavidin R-phycoerythrin
SDS	Sodium dodecyl sulphate
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
SSC	Saline sodium citrate
ssRNA	Single stranded RNA
TBE	Tris-Borate EDTA
TF	Transcription Factor
TK promoter	Thymidine kinase promoter
T_m	Melting temperature
tRNA	Transfer RNA
UTR	Untranslated region

Chapter 1: Introduction

The full potential of non-coding RNAs (ncRNAs) is only recently beginning to be realised (Mattick and Makunin, 2006). A prime example of this surge of interest in novel RNA functions is that of microRNAs (miRNAs). Not only has the number of annotated miRNAs increased rapidly, but it has become increasingly apparent that miRNA mediated post-transcriptional regulation is widespread and linked to many key biological processes, including cancer, development and embryonic stem cell self renewal and pluripotency (Gangaraju and Lin, 2009; Medina and Slack, 2008; Zhao and Srivastava, 2007). The identification of novel miRNAs has progressed so rapidly since the advent of this relatively new area of cellular research that there is very limited functional annotation of these miRNAs, bearing in mind the potential complexity of the regulatory networks within which each may participate. As a consequence, the aim of my research was to develop a system that would help to address this deficit in functional annotation through the derivation of a large number of experimentally supported candidate target messenger RNAs (mRNAs) for miRNAs expressed in mouse embryonic stem (ES) cells.

1.1 miRNAs

Metazoan miRNAs are ~21-22 nucleotide (nt) small RNA molecules which, as a general rule, guide a ribonucleoprotein complex (miRNP) to target mRNA molecules by partial complementarity between themselves and the mRNA molecule. The vast majority of mRNAs targeted by a miRNA are either degraded or translationally inhibited. This report will concentrate solely on the miRNAs of metazoans, as plant miRNAs have been hypothesized to

have evolved independently and therefore obey a related, but broadly non-applicable set of rules (Axtell and Bowman, 2008; Mallory and Bouche, 2008).

The related process of RNA interference (RNAi) is triggered by double stranded RNA (dsRNA) molecules, which are processed into small interfering RNAs (siRNAs); RNA molecules that are of approximately the same length as the miRNA. As a rule these are believed to guide ribonucleoprotein (RNP) complexes to mRNA targets, which they match with perfect complementarity. This leads to the cleavage of the target molecules at the point at which they bind. This mechanism of target regulation is not the same as that used by miRNAs in all but atypical circumstances.

Since the discovery of the first miRNA in 1993, the number of known miRNAs has rapidly increased, with the vast majority being identified since the turn of the century (Lee et al., 1993). Currently there are 695 human miRNAs and 488 mouse miRNAs registered in miRBase (Release 12) (Table 1.1A) (Griffiths-Jones et al., 2008). Originally believed to be exceptional, it is now clear that miRNA mediated post-transcriptional regulation has a major influence on both the RNA and protein expression profile of cells. Tens of miRNA species are expressed within every tissue (Landgraf et al., 2007) and each is expected to have hundreds of targets (Friedman et al., 2009). The network of miRNA-mediated control is complex, with elements of both combinatorial regulation by multiple miRNAs targeting the same molecule, and functional redundancy between different miRNA species. In addition feed forward and feedback loops have been recognized involving both miRNAs and proteins (Marson et al., 2008; Petrocca et al., 2008a).

In table 1.1B on the following page, I have included a list of definitions of miRNA related terms used through out this thesis for reference.

A

mmu-miR-92a-1

Name features	Purpose of feature
mmu-	Reference to the organism ('mmu-' = Mouse, 'has-' = human etc.)
92	Novel miRNAs are given a specific number
a	A letter appended to the miRNA number denotes closely related mature sequences, (eg. miR-92a and miR-92b)
-1	Identical mature sequences expressed from distinct loci are numbered sequentially
*	If ~22nt RNA molecules are derived from opposite strands of a precursor hairpin in a cloning study and one form clearly predominates the less dominant form is appended with a '*'
-3p/5p	If ~22nt RNA molecules are derived from opposite strands of a precursor hairpin in a cloning study and neither form clearly predominates the forms are appended with a '-3p' or '-5p' depending upon which strand of the miRNA hairpin the miRNA originates

B

Term	Definition
miRNA	~21-22nt RNA molecules responsible for guiding a miRNP to partially complementary mRNA molecules leading to post-translational regulation of targets.
pri-miRNA	Primary transcripts from within which miRNAs are initially transcribed.
pre-miRNA	Folded RNA hairpins containing the miRNA molecule. Released from the primary transcript by Drosha cleavage, exported from the nucleus and further processed by Dicer to release the miRNA.
microprocessor	A protein complex containing Drosha and DGCR8 proteins responsible for releasing pre-miRNAs from pri-miRNA transcripts.
miRNP/RISC	Protein complex containing miRNAs (or siRNAs) responsible for orchestrating miRNA guided post-transcriptional gene regulation.
mirtron	A small subset of miRNAs transcribed within short intronic structures. The ends of the pre-miRNA are determined by splice events and hence these are processed in a microprocessor independent manner.

Table 1.1: The miRNA naming convention and definitions of relevant terms used throughout this thesis.

A) This is an example of the miRNA naming convention used by miRBase. The names can be divided into 4 or 5 parts, each of which contains information concerning the miRNA's origins or limited information concerning miRNA:miRNA relationships. B) Definitions of terms used throughout.

1.1.1 MicroRNA processing pathway

Mature miRNAs are released from larger RNA molecules by two rounds of RNA cleavage. They are transcribed within primary miRNA (pri-miRNA) molecules, which fold to incorporate the miRNA within the stem of a hairpin structure. Initially the hairpin is released from this longer transcript by the RNase III family protein Drosha. This released hairpin is termed the precursor miRNA (pre-miRNA). Subsequently the loop is removed from the hairpin by the RNase III family protein Dicer and the miRNA is liberated (Fig.1.1). Next I shall discuss the biogenesis of miRNAs in more detail.

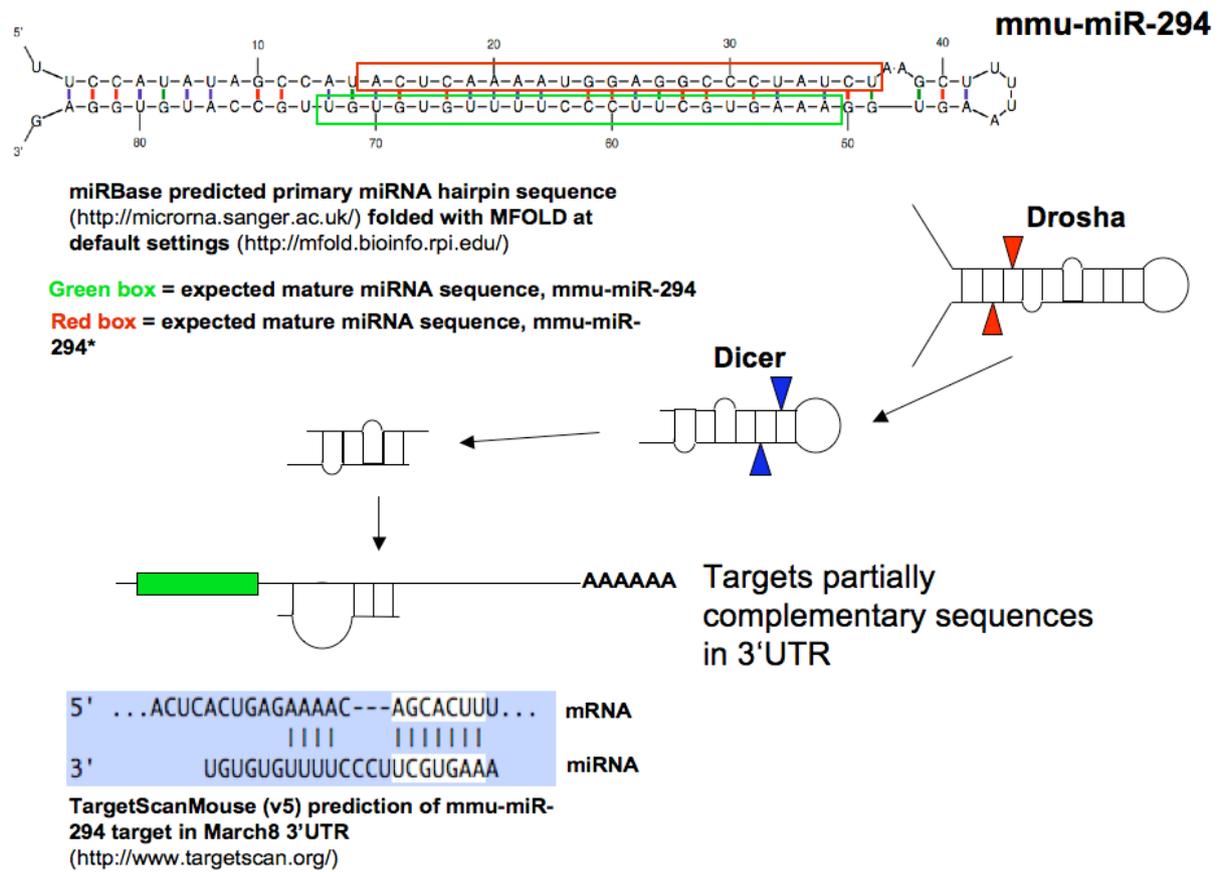


Fig.1.1: A simplified outline of the miRNA processing pathway, depicting the two cleavage steps required to liberate the mature miRNA. Sequence structures depict a representation of the local RNA sequence and structure into which the mature mmu-miR-294 miRNAs are embedded (http://microrna.sanger.ac.uk) (**top**) and a predicted target site interaction between miRNA and target sequence (http://www.targetscan.org) (**bottom**).

1.1.1.1 Pri-miRNAs

Pri-miRNAs can take on several forms. miRNAs can reside within introns or exons of independently transcribed ncRNA molecules or are transcribed along with the mRNAs of protein coding genes, embedded in their introns or untranslated regions (UTRs) (Fig.1.2) (Kim and Kim, 2007; Rodriguez et al., 2004). Multiple miRNAs can reside in clusters, transcribed together as a single unit, to be subsequently cleaved out and processed separately (Houbaviy et al., 2005).

1.1.1.1.1 Transcription of pri-miRNAs

In addition to miRNAs found within the introns of protein coding genes, independently transcribed non-coding pri-miRNA molecules are most commonly transcribed by polymerase II (pol II), although pol III regulated transcription cannot be excluded with upstream Alu repeat sequences seemingly capable of supporting miRNA expression from the human C19MC locus (Borchert et al., 2006). A range of these pri-miRNAs have been demonstrated to possess a 7-methyl guanosine cap and to be polyadenylated (Cai et al., 2004; Houbaviy et al., 2005; Lee et al., 2004a). In addition, α -amanitin, a pol II inhibitor, appears to reduce the levels of 7 pri-miRNAs within HeLa cells and pol II has been demonstrated to bind directly to the miR-23a-cluster promoter region (Lee et al., 2004a). Furthermore, the promoter region of hsa-miR-21, also identified in HeLa cells, is capable of supporting the transcription of functional mRNA transcripts which further suggests these promoter elements are capable of recruiting Pol II (Cai et al., 2004). The mmu-miR-290 cluster has been annotated in considerable detail. Houbaviy et al identified a pol II promoter region upstream of the miRNA cluster. This promoter region contained a TATA box conserved in *H. sapiens*, *M. musculus*, *B. Taurus* and *C. familiaris*, within 35 base pairs (bp) of the transcriptional start

site of the cluster (Houbaviy et al., 2005). More recently, through a comparison of regions found upstream of intergenic miRNAs in *C. elegans*, *H. sapiens*, *A. thaliana* and *O. sativa* to a collection of pol III and pol II promoters, Zhou *et al.* demonstrated that the vast majority of miRNAs appear to possess pol II promoters, with 100% of regions upstream (2000bp) of *C. elegans* pre-miRNAs and 96.3% of regions upstream of *H. sapiens* pre-miRNAs predicted to contain possible or definitive pol II promoters. The remaining regions were found to contain either possible pol III promoters or random sequence (Zhou et al., 2007).

1.1.1.1.2 Intronic miRNAs

Direct evidence that miRNAs may be expressed along with host mRNAs was demonstrated in a comparison of 90 human miRNAs to the NCBI expressed sequence tag (EST) database. Chimeric ESTs were found to contain both mRNA sequence and miRNA precursor sequence (Smalheiser, 2003). A recent and more comprehensive study (Kim and Kim, 2007) verified work performed earlier by Rodriguez et al. (Rodriguez et al., 2004), demonstrating that ~80% of the miRNAs that map to ESTs map to introns. Considered alongside all miRNAs (including those with no EST information or those which do not map to known genes), this intronic population accounts for ~50% of the miRNAs investigated in these two independent studies. Of the definitively intronic miRNAs from the Rodriguez *et al.* study, ~3/4 are within the introns of protein coding genes.

Where the intronic miRNAs appear on the same strand as the host gene, it is expected that the miRNAs will be transcribed along with the host transcript and subsequently processed. Microarrays have been employed to demonstrate correlation between the expression of miRNAs and their parent transcripts, again adding weight to the co-expression hypothesis

(Baskerville and Bartel, 2005). These results were replicated by deriving expression data for host transcripts by reverse transcriptase polymerase chain reaction (RT-PCR) and comparing these results to previously generated miRNA expression data (Rodriguez et al., 2004). Both intronic miRNAs tested in this way exhibited the same expression profile as their host transcript.

1.1.1.2 Cleavage of the primary miRNA transcript by the microprocessor protein complex

The initial miRNA processing step (release of pre-miRNAs from the pri-miRNA transcript) takes place within the nucleus. The characteristic 2 nt 3' overhangs and 5' monophosphate groups of mature miRNA duplexes prompted the identification of Drosha (RNASEN) as the enzyme that performs the initial restriction of the miRNA maturation process, as these features are also the byproduct of an RNase III cleavage reaction (Lee et al., 2003).

1.1.1.2.1 The microprocessor

Gregory *et al.* conducted an analysis of two Drosha containing protein complexes of differing sizes in HEK-293 cells (An apparent third complex containing only a smaller isoform of Drosha was not pursued further by this study) (Gregory et al., 2004). The authors found Drosha to be associating with multiple and varying proteins, including DEAD-box helicases, heterogeneous nuclear ribonucleoproteins and Ewing's sarcoma proteins. However, of the proteins tested, Drosha appeared to associate with DGCR8 (a dsRNA binding protein) alone in the smaller ~600kDa complex. This was also the complex which accounted for by far the largest proportion of pri-miRNA processing activity *in vitro*, processing pri-miRNAs into

pre-miRNAs in a site-specific manner. This protein complex constituting of Drosha and DGCR8 has been termed the microprocessor. The microprocessor-pri-miRNA processing activity could be replicated *in vitro* by combining recombinant Drosha and DGCR8, although alone, neither could process pri-miRNAs effectively and Drosha alone exhibited non-specific RNase activity. Examining the role of the microprocessor *in vivo*, siRNAs targeted to either Drosha or *DGCR8* were found to block miRNA processing at the pri-miRNA step. However it should be noted that *in vivo* depletion of three of the other components of the largest Drosha containing complex also had a small effect on mature miRNA levels although as this was much less significant than the effect seen with the smaller complex, the authors conclude it is “therefore more likely that the large Drosha-containing complex has a function in other RNA processing pathways” (Gregory et al., 2004).

A second study was published simultaneously, identifying the homologue of *DGCR8* (Pasha) as the partner of Drosha in *Drosophila* (*pasha*) and *C. elegans* (*pash-1*) (Denli et al., 2004). Again these two proteins co-immunoprecipitated in *Drosophila* S2 cells. Although PASHA co-precipitates with pri-miRNAs, the Drosha:PASHA interaction was unaffected by RNase treatment of the immunoprecipitates, implying a direct interaction between the two proteins. RNAi experiments targeting Pasha in both *Drosophila* and *C. elegans* lead to an accumulation of pri-miRNA and a depletion of mature miRNA as expected.

Fukuda *et al.* were only able to purify the larger of the two Drosha containing complexes, described above, from mouse cells, although they found DGCR8 within this complex (Fukuda et al., 2007). Drosha is thought to be involved in ribosomal RNA (rRNA) processing in addition to miRNA processing (Wu et al., 2000) and this complex was able to process both

miRNAs and rRNAs. This study implicated further proteins (DEAD-box RNA helicase subunits) in the processing of a subset of pri-miRNA sequences in addition to the minimal microprocessor. DGCR8 is not required for rRNA processing, however, as demonstrated in a *Dgcr8* mouse ES cell knock out experiment, in which no effect was seen on the levels of rRNAs caused by the removal of the functional protein (Wang et al., 2007).

Within this newly discovered microprocessor complex it appears that multiple copies of Drosha and DGCR8 interact. However, the enzymatic processing centre of Drosha is formed by intramolecular dimerisation of the two RNase sites within each Drosha molecule (Han et al., 2004).

This first excision defines one end of the mature miRNA sequence. This initial processing step proceeds co-transcriptionally; Drosha associating with the nascent strand in a DGCR8 dependent manner (Morlando et al., 2008).

1.1.1.2.2 Recognition and mechanism of pri-miRNA processing

Han *et al.* (Han et al., 2006) calculated the average local RNA structure of miRNAs in humans and flies. This structure consisted of an approximately 33bp stem (approximately 3 helical turns) with a terminal loop (for an example of a primary miRNA structure see Fig.1.1). The stem structure was flanked by single-stranded RNA segments. From observations discerned by the use of labeled transcripts containing various artificial mutations and structural alterations, and an immunopurified microprocessor complex, Han *et al.* proposed a model whereby DGCR8 binds firmly to the single stranded to double stranded junction of the pri-miRNA structure. Drosha binds to this anchoring complex transiently with

its RNase domains positioned approximately 11bp from the base of the stem; the site of RNase cleavage. Zeng et al. reached broadly similar conclusions with regard to microprocessor function. Again, using a series of *in vitro* assays in addition to transfected plasmid expression constructs they concurred that the microprocessor most efficiently processed a stem flanked at each end by lengths of single stranded RNA (Zeng and Cullen, 2005). However, in contrast to the other study, they proposed a mechanism whereby Drosha or a protein complex recognized the pre-miRNA hairpin loop and then mediated RNase cleavage ~22bp from that junction (Zeng et al., 2005). Han *et al.* attempted to reconcile some of the differences presented by these two conflicting models for microprocessor action by observing that the presence of a large terminal loop required in the Zeng hypothesis could be a consequence of the requirement for the single stranded RNA (ssRNA) at both ends of the duplex for efficient cleavage. Han *et al.* were unable to replicate the critical experiments of Zeng *et al.* where alterations to the position of the terminal loop of the pre-miRNA structure altered the site of cleavage accordingly.

1.1.1.2.3 Processing of intronic miRNAs

Until recently it had been assumed that miRNAs were processed from introns post-splice. However, Kim *et al.* have observed spliced ESTs that begin a few base pairs from a proposed Drosha cleavage site and contained sequences from the miRNA containing intron along with adjacent spliced exons from the parent transcript. This EST, which may be derived from a Drosha cleavage product, would imply that miRNAs are processed from unspliced introns within otherwise spliced mRNA molecules (Kim and Kim, 2007). Further experiments in HeLa cells confirmed the presence of partially spliced parent mRNA transcripts within which the miRNA containing intron appears to be spliced after other intronic sequences, implying

that the microprocessor may interfere with the splice machinery. Artificial expression constructs containing miRNAs within the introns of protein coding genes were used to demonstrate that splicing is not required for intronic miRNA processing. These constructs were also used to demonstrate that the presence of a miRNA within the intron of a gene did not significantly affect the levels of fully spliced mRNA (Kim and Kim, 2007). These observations imply that mRNAs cleaved by Drosha are subsequently spliced. The authors propose the “exon-tethering” model of Dye et al. as a possible explanation for this (Dye et al., 2006).

1.1.1.2.4 miRNAs within the 3’UTRs of protein coding genes

Intriguingly Rodriguez *et al.* identified 2 miRNAs that map within the exons of the 3’UTRs of protein coding genes, in a study of mouse and human miRNAs (ENSMUSG00000018171 and ENSG00000163430) (Rodriguez et al., 2004). It is worth noting that experiments conducted in Hek-293T cells with luciferase reporter genes harboring a pre-miRNA in their 3’UTR detected a large proportion of the truncated, Drosha processed, luciferase transcript in the cytoplasm and a much smaller fall in luciferase activity than expected. This prompted a hypothesis that the processed transcripts were still able to function as mRNAs despite the truncation event (Cai et al., 2004).

1.1.1.2.5 Co-transcriptional miRNA processing

Morlando *et al.* performed Drosha chromatin immunoprecipitation (ChIP) in HeLa cells. They discovered an enrichment of Drosha at the sites of expressed intronic and intergenic miRNAs that was both RNase sensitive and DGCR8 dependent (Morlando et al., 2008). The authors proceeded to investigate the process of co-transcriptional miRNA processing in

detail. In addition to copious other experiments, the authors used Northern blots to demonstrate the successful maturation of miRNAs expressed downstream of a poly-A site from highly unstable transcripts, generated by 3' transcription from unterminated β -globin gene constructs. Furthermore, the authors demonstrated the recruitment of exonucleases to the sites of Drosha cleavage in HeLa cells. These nucleases seem to enhance splicing efficiency following Drosha cleavage as depletion of these proteins by RNAi led to a reduction of splicing efficiency. The effect of this depletion was itself reduced upon the depletion of Drosha.

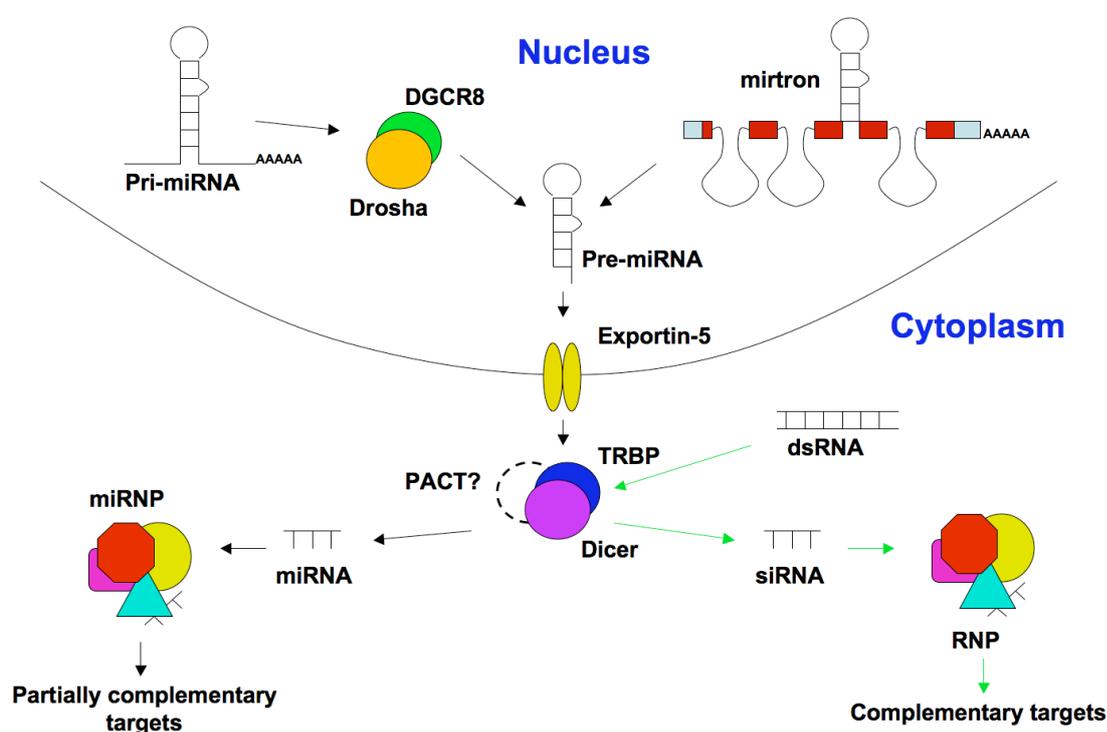


Fig.1.2: Canonical microRNA processing pathway in vertebrates, with the introduction of mirtronic miRNAs at the pre-miRNA stage. miRNAs are generally transcribed in longer pri-miRNA molecules, replete with secondary structure. Drosha (RNASEN) and DGCR8 operate in unison to liberate the pre-miRNA hairpin. This is exported from the nucleus to the cytoplasm by exportin 5 (XPO5). In the cytoplasm the pre-miRNA is further processed by Dicer (DICER1) with associated cofactors to release the mature miRNA (PACT (PRKRA) and TRBP (TARBP2)). One strand of the miRNA duplex is selected and incorporated into the miRNP, which it guides to target mRNA molecules to generally block translation or cause degradation via deadenylation. Mirtrons follow the same processing pathway for the most part, but are initially excised from the parent RNA molecule via a splice reaction (see section 1.1.1.6). The RNAi pathway overlaps with the miRNA processing

pathway at the Dicer cleavage stage. As research continues the distinction of these two pathways is becoming more murky and complicated by overlaps in components and function.

1.1.1.3 The fate of the precursor miRNA

Above I have discussed in detail the first cleavage event in the miRNA processing pathway required to release the pre-miRNA from the pri-miRNA transcript. This initial processing step takes place in the nucleus. The precursor hairpin (~70nt molecule) is then transported from the nucleus by exportin 5 (Yi et al., 2003; Zeng and Cullen, 2004) in cooperation with a Ran-GTP cofactor (Fig.1.2). Once in the cytoplasm the pre-miRNA is further processed by a second RNase III enzyme, Dicer, in order to liberate the mature miRNA from the larger RNA molecule.

1.1.1.3.1 Pre-miRNA processing

Dicer removes the loop from the end of the pre-miRNA, releasing the mature miRNA and once again leaving a 5' phosphate, 3' hydroxyl and a 2 nt 3' overhang, all of which are characteristic of miRNAs (Grishok et al., 2001; Ketting et al., 2001; Knight and Bass, 2001). It is worth noting that the miRNA processing pathways and the RNAi pathway converge at this point (Bernstein et al., 2001; Hutvagner et al., 2001) with Dicer responsible for the processing of dsRNAs in both pathways (Fig.1.2). These roles are performed by distinct Dicer orthologues in *Drosophila* (*Dcr-1* and *Dcr-2*) (Lee et al., 2004b), but in other animals from *C. elegans* (*dcr-1*) to *M. musculus* (*Dicer1*) and *H. sapiens* (*DICER1*) a single Dicer gene exists and this single protein performs both duties (Fig.1.2).

Like Drosha, Dicer has been shown to associate with dsRNA binding proteins (R2D2 and Loquacious in *Drosophila* and PACT (PRKRA) and TRBP (TARBP2) in human cell lines

and in the mouse (Chendrimada et al., 2005; Forstemann et al., 2005; Kok et al., 2007; Liu et al., 2006)) (Fig.1.2). However, unlike Drosha, which requires DGCR8 in order to cleave primary miRNAs *in vitro* (Han et al., 2004), Dicer is capable of cleaving both dsRNAs and pre-miRNAs *in vitro*, in the absence of a dsRNA binding partner (Chendrimada et al., 2005).

1.1.1.3.2 Dicer associated proteins

From this point on the understanding of the mechanism of miRNA function becomes more difficult to discern with considerable disagreement evident between a number of papers, some of which are at present very difficult to reconcile.

Within mammalian systems, Dicer has been found to interact *in vitro* and *in vivo* with PACT and TRBP in stable complexes. Chendrimada *et al.* found that TRBP bound Dicer and in turn allowed the association of Argonaute 2 (EIF2C2) (discussed later) to an siRNA associated complex (Chendrimada et al., 2005). They suggested that TRBP might therefore be involved in the initial stages of RNP/RISC (RNA-induced silencing complex) complex assembly (the miRNA and siRNA effector complexes). They also found that the depletion of TRBP in HEK-293 cells seemed to lead to the destabilisation of Dicer and a reduction of mature miRNA and siRNAs. Haase *et al.* did not demonstrate a fall in mature miRNA levels in cell lines depleted for TRBP (Haase et al., 2005), nor did they see a destabilisation of Dicer. They did however see that endogenous miRNAs had a reduced effect on reporter gene transcripts bearing perfectly complementary target sites in their UTRs following TRBP depletion, again implying a role for TRBP in RNP/RISC assembly.

Kok *et al.* found Dicer, PACT and TRBP to form trimeric complexes in HEK-293 cells and in mouse testicular tissue (Kok *et al.*, 2007). Through a series of processing experiments both in artificial systems and in human cell lines the authors found that Dicer cleavage of dsRNAs and short hairpin RNAs (shRNAs) was improved by the presence of both PACT and TRBP. However, in contrast to Haase *et al.* they also found that the Hek-293 cells depleted in TRBP remained susceptible to siRNA transfection but not shRNA transfection, whereas Haase *et al.* found that siRNA effects were abolished upon the depletion of TRBP in a Hek293T-REx cell line. Ultimately, Kok *et al.* suggested that TRBP and PACT function at the stage of siRNA production. These differences are difficult to reconcile and further experiments are necessary to clarify the situation.

1.1.1.4 Regulation of the miRNA processing pathway

It is clear that miRNAs co-expressed in clusters are not always present within the cell in equal quantities in their mature form. This would suggest that there are mechanisms whereby either the processing of mRNAs to their mature forms is regulated or that they are degraded at different rates. One such mechanism has been described for the regulation of the let-7 miRNA family by Lin28a (*LIN28*) and Lin28b (*LIN28B*) in vertebrates. Two systems have been proposed for lin-28's interaction with the miRNA. Newman *et al.* found that Lin28 binds to specific sequences in the loop of the let-7 hairpin, subsequently blocking the Drosha cleavage reaction (Newman *et al.*, 2008). Heo *et al.* found the block to be prior to Dicer cleavage of their target precursor. Again the authors found Lin-28 was able to bind the let-7 molecule but in this case the let-7 precursor appeared to acquire a uracil tail in a Lin-28 dependent manner and was more rapidly degraded than the standard precursor-miRNA (Heo *et al.*, 2008).

1.1.1.5 miRNPs, the effector complexes of miRNA mediated regulation

Both siRNAs and miRNAs are incorporated into RNPs which they then guide to target sequences in mRNAs. In the case of miRNAs in the vast majority of cases these target sequences are partially complementary to the miRNA sequence in metazoans. It could be expected that the strand of the miRNA selected to guide the miRNP is the strand with the least thermodynamically stable 5' end, with 5' instability being a property of functional siRNAs (Khvorova et al., 2003). However, with the advent of high throughput sequencing it is becoming more apparent that the so called “star” or “passenger” strand also plays an important part in post-transcriptional regulation of target sequences. This is supported by evolutionary data suggesting a biological role for both miRNA strands (Okamura et al., 2008).

At the core of the miRNP are the RISC-associated argonaute proteins (reviewed in Hutvagner *et al.* (Hutvagner and Simard, 2008)). These proteins are thought to be important effectors of miRNA function. Tethering of Argonaute-like proteins to mRNAs in the absence of miRNAs is sufficient to initiate post-translational control of the target sequence (Pillai et al., 2004), while Argonaute-like proteins have also been demonstrated to possess the “Slicer” activity required to cleave target sequences with perfect complementarity to an siRNA or miRNA (Liu et al., 2004; Meister et al., 2004).

Humans and mice possess four of these proteins AGO1-4 (EIF2C1, EIF2C2, EIF2C3, and EIF2C4). The PIWI module (Mid domain) of these proteins binds the 5' end of the miRNA, while the PAZ domain is thought to recognize the 2 nt 3' ssRNA overhang, produced by the

RNase III processing of the miRNAs and siRNAs. The AGO proteins are incorporated within larger protein structures. A recent study of human AGO1 and AGO2 identified a large number of proteins with numerous associated functions in what appeared to be 3 complexes of differing molecular weight (Hock et al., 2007). Interestingly, all four AGO proteins have been purified alongside a similar selection of proteins which implies a degree of functional redundancy between each of the AGO containing complexes (Landthaler et al., 2008).

Further studies have endeavored to identify the miRNAs associated with each AGO in humans. Argonaute-like proteins are known to have diversified in their function, with only human AGO2 exhibiting “slicer” activity required to cleave target sequences that are perfectly complementary to a miRNA or siRNA (Liu et al., 2004; Meister et al., 2004). Disruption of the *Eif2c2* gene in the mouse also leads to severe abnormalities and an embryonic lethal phenotype arguing against total functional redundancy amongst mammalian AGO proteins. It is therefore potentially surprising that evidence supports miRNAs binding to AGO proteins indiscriminately (Liu et al., 2004; Meister et al., 2004). It should be noted however that more recent experiments conducted through the immuno-precipitation of endogenous AGO proteins and the pyrosequencing and 454 sequencing of associated miRNAs, suggest that although overall the miRNA population from AGO1, AGO2 and AGO3 complexes are broadly similar, there are some differences which could direct the three complexes to subtly different target populations (Azuma-Mukai et al., 2008; Ender et al., 2008).

1.1.1.6 Mirtrons and other exceptions to the canonical rules

The advent of high throughput sequencing has allowed the small RNA complement of organisms and cell lines to be characterized beyond anything that was achievable through low throughput cloning methods, identifying rare miRNAs beyond the sequencing depth of more conventional methods. In addition, as these new methods do not require sequence complementarity for miRNA detection, they are able to easily profile these relatively rare small RNA species previously not profiled by microarray based detection strategies.

These techniques identified mirtrons as a new miRNA species that were not processed in the same way as the majority of miRNAs (see above). First discovered in *C. elegans* and *D. melanogaster* (Okamura et al., 2007; Ruby et al., 2007), these miRNAs are expressed within the short introns of other genes. However, rather than be sliced from the intron by Drosha, the entire intron is spliced by the splicing machinery and the spliced lariat is debranched to form a miRNA-precursor in a fashion that is microprocessor independent. Subsequently these pre-miRNAs seem to re-enter the canonical miRNA processing machinery at the stage of nuclear export, as defined by a series of RNAi experiments in S2 cells (Okamura et al., 2007; Ruby et al., 2007).

Berezikov *et al.* extended the search for mirtrons to mammals, proposing 19 mirtrons and 46 mirtron candidates based on high throughput sequencing data, restricting the search to short introns in mammalian genomes (Berezikov et al., 2007). Mechanistic evidence for the existence of mammalian mirtrons was provided by small RNA sequence libraries from Dicer and *Dgcr8* knockout mouse embryonic stem cell lines (Babiarz et al., 2008). Although by far the majority of miRNAs demonstrated a canonical requirement for both proteins, mirtrons

and other DGCR8 independent miRNA species were identified. Further sequence evidence was provided for miR-877 which had been predicted by Berezikov *et al.* as a mirtron, and the libraries also allowed the authors to demonstrate that these mirtrons are indeed DGCR8 independent as expected (Fig.1.2). miR-702 and miR-1981 were also identified as mirtrons. These mirtrons appear to be longer than those seen in invertebrates but still fold into pre-miRNA-like hairpins. miR-1982 also had a mirtron like structure and the same enzymatic dependencies. However, this intron folded into a structure reminiscent of dme-mir-1017 identified by Ruby *et al.* in *Drosophila*, with a single stranded RNA tail at one end of the hairpin proposed to be released by splicing (Ruby *et al.*, 2007). In order for this pre-miRNA to enter the canonical pathway, the authors hypothesise that this tail would be removed by an as yet unidentified nuclease.

In addition to mammalian mirtrons, a number of other hairpin, DGCR8-independent, Dicer-dependent miRNA precursors were identified. miR-320 produced the most abundant reads in this category. Unusually, it was not highly conserved beyond the hairpin pre-miRNA. This is in contrast to the majority of conserved miRNAs, as miRNAs must maintain the stem structure for a further helical turn (approximately) in order to provide the optimal substrates for the microprocessor. The authors also noted that the majority of the reads mapped to the 3' arm of this precursor, as would be expected if the 5' end did not possess the residual phosphate left by RNase III cleavage. Hence, fragments from this side of the hairpin would not be cloned successfully. miR-484 also fell into a category displaying similar features to miR-320. miR-1980 had a tailed hairpin-like structure reminiscent of miR-1982 (discussed above) although miR-1980 was not within an intron. Clearly there are a number of non-canonical pathways by which a small subset of miRNAs can be processed.

One of the most novel findings of this paper was the concept that miRNAs may also be processed from other non-coding RNAs under a certain set of circumstances. A stack of sequence reads resembling a distribution normally associated with a miRNA gene, mapped to an annotated transfer RNA (tRNA). This stack was again Dicer dependent but DGCR8 independent. The locus also seems to be capable of being processed as a tRNA and will fold into either a hairpin structure or a tRNA cloverleaf structure. As a tRNA, this locus is pol III transcribed.

As another example of DGCR8 independent processing of a non-coding RNA into a potentially functional miRNA, a human small nucleolar RNA (snoRNA) has recently been demonstrated to produce a functional miRNA (Ender et al., 2008). Although it seems to be a small proportion of this snoRNA that is converted into the miRNA, luciferase assays have been used to demonstrate that CDC2L6 may be an example of an endogenous target of this unconventional miRNA.

1.1.2 Mechanism of miRNA function

In the vast majority of cases, miRNAs guide the miRNP to transcripts with target sites partially complementary to the miRNA sequence, generally within a mRNA's 3'UTR. The miRNP is then responsible for the post-transcriptional regulation of the target transcript. The mechanisms by which metazoan miRNAs/miRNPs function are still poorly understood (For review see (Filipowicz et al., 2008)). This is in part due to a rapidly increasing plethora of exceptions to the general rules of miRNA function and partly due to conflicting hypotheses being proposed and experimentally supported by differing experimental evidence. Two broad

methods of post-transcriptional gene regulation are thought to predominate. The first is destabilization and degradation of the target transcripts, while the second is through translational inhibition. I will discuss each below together with a variety of exceptions.

Advances in the high throughput analysis of the cellular proteome has allowed an investigation of the relative contribution of these two mechanisms to the effect miRNAs have on cellular expression (Baek et al., 2008; Selbach et al., 2008). Using the new “pulsed stable isotope labeling with amino acids in cell culture” (pSILAC) method to differentially label protein samples in culture these papers show, through the over expression of miRNAs or disruption of endogenous miRNAs, that the majority of miRNA targets are either repressed at both the mRNA level and the translational level or that mRNA destabilisation seems to account for the changes. However, some targets were seen to be almost totally regulated at the translational level with no apparent change to mRNA levels.

1.1.2.1 Deadenylation: Destabilisation of targets through the removal of the poly-A tail

Of these two widely accepted mechanisms of miRNA action, target degradation is perhaps the most clearly understood. While investigating the role of miR-430 in the early development of the zebrafish, Giraldez *et al.* demonstrated that this miRNA triggered the deadenylation of its targets (Giraldez et al., 2006). Wu *et al.* noted the same phenomenon in a mammalian system (Wu et al., 2006). This deadenylation would be expected to precede miRNA triggered mRNA degradation. Indeed, it has been recently shown that the depletion of the deadenylation complex in *Drosophila* S2 cells leads to an enrichment of miRNA targets in these cells (Eulalio et al., 2009). The authors of this paper went on to predict that

“60% of transcripts up-regulated in AGO-1 depleted cells are normally degraded through deadenylation”.

Further to this they demonstrated that the depletion of decapping activators (*Ge-1* and *me31B*) also inhibited miRNA mediated degradation of targets, although there remains evidence that targets were still deadenylated despite their stability. This replicated earlier work by Eulalio *et al.* in 2007 (Eulalio *et al.*, 2007) and implies that the deadenylation of miRNA targets is followed by the removal of their 5' cap. It is noted that the degradation of a miRNA target appears to be independent of miRNA induced inhibition of translation. Even with a background of a total block in the initiation of reporter mRNA translation, the reporter construct cotransfected with a miRNA for which it bears targets is degraded more effectively than a reporter transfected with no targeting miRNA (Eulalio *et al.*, 2009; Giraldez *et al.*, 2006; Wu *et al.*, 2006).

The widespread destabilisation of miRNA targets has been utilised by the recent Sylamer program, which can identify enrichments of predicted miRNA targets within gene lists ordered depending upon mRNA expression changes following miRNA addition and disruption experiments (van Dongen *et al.*, 2008). These clear enrichments are seen either amongst the up regulated or down regulated gene sets depending upon whether the miRNAs are being added to or removed from the system.

1.1.2.2 Translational inhibition

From very early in the study of miRNAs, translational inhibition has been recognised as a method by which miRNAs inhibit the expression of their target mRNAs (Wightman *et al.*,

1993). However, it remains far from clear how miRNAs achieve this. In particular, it is not obvious whether miRNAs inhibit translation at the initiation step or at a post initiation stage. Polysomal fractionation has been used to demonstrate blocks at both stages of translation. Inhibition of reporter mRNAs by let-7 has been demonstrated to cause a shift of target mRNA to the top of the sucrose gradient implying a block in the association of the target with ribosomes at the initiation of translation (Pillai et al., 2005). In contrast other studies have suggested that the inhibition of target mRNAs leads to no change in the polysomal association seen on these gradients, suggesting that the translational inhibition occurs post-initiation (Olsen and Ambros, 1999; Petersen et al., 2006).

A number of mechanisms have been suggested for the miRNA dependent regulation of translation at the stage of initiation. Wakiyama *et al.* noted that the polyadenylation of transcripts appears necessary for translational inhibition, implying that deadenylation may play a role in the inhibition (Wakiyama et al., 2007). It is known that the interaction between the poly-A tails of mRNAs and their caps enhance translation. However, contradictory results from subsequent studies refute this. Pillai *et al.* found that a poly-A tail is not required for translational inhibition (Pillai et al., 2005) and this was again corroborated by Wu *et al.* (Wu et al., 2006).

It also appears that the m(7)G-cap of the mRNA plays an important role in translational suppression. Pillai *et al.* found that mRNAs required a m(7)G-cap for repression with neither an internal ribosome entry site (IRES) nor tethered initiation factors acting as adequate substitutes (Pillai et al., 2005). Kiriakidou *et al.* noted that the AGO proteins possess a domain that resembles EIF4E, capable of binding the m(7)G-cap. They went on to

demonstrate that the disruption of this domain led to a loss of the block of initiation, implying that AGO proteins may disrupt the initiation complex (Kiriakidou et al., 2007). This finding was later extended to demonstrate that the block disrupts the recruitment of the 80S ribosome to the targeted transcripts (Mathonnet et al., 2007).

An alternative mechanism has been presented by Chendrimada *et al.*, who demonstrated an association between TRBP and EIF6. EIF6 associates with the 60S ribosomal associated factor and in doing so disrupts the assembly of the translationally competent 80S ribosome (Chendrimada et al., 2007).

miRNAs have also been shown to co-sediment with the polysome fraction on a sucrose gradient. By subsequently blocking mRNA translation with exogenous agents, it has been demonstrated that at least under specific circumstances miRNAs are associated with actively translated mRNAs (Maroney et al., 2006). In contrast to other publications, Petersen *et al.* found IRES dependent translation to remain susceptible to regulation by miRNAs/bulged-siRNAs. As mentioned above they also noted no change in polysomal occupancy on inhibition. Therefore, judging inhibition to occur post initiation, they then proceeded to inhibit translation initiation and noted more rapid dissociation of target mRNAs from polysomes than control mRNAs. They propose that miRNA-triggered premature release of target peptides from ribosomes is a cause of miRNA mediated translational inhibition (Petersen et al., 2006).

1.1.2.3 A role for the P-body

Processing bodies (P-bodies) are discrete cytoplasmic foci; a site of mRNA sequestration and degradation in the cytoplasm, (Reviewed (Eulalio et al., 2007a). It is becoming increasingly apparent that P-bodies are intricately involved in miRNA function. GW182 (a P-body associated protein) binds AGO proteins. Behm-Ansmant *et al.* revealed that GW182 is required for both translational inhibition and miRNA target degradation. They also noted that the CCR4:NOT1 deadenylation complex and DCP1 and DCP2 from the decapping complex, all of which are associated with the P-body, are required for target degradation (Behm-Ansmant et al., 2006). Given these associations and the localization of mRNAs to P-bodies in a miRNA dependent manner (Liu et al., 2005), it is surprising that the disruption of P-body integrity does not have a pronounced effect on miRNA dependent regulation of reporter constructs (Eulalio et al., 2007b).

1.1.2.4 Reconciling the different mechanisms for miRNA mediated post-transcriptional regulation

Reconciling these proposed differences in the miRNA functional mechanism is not easy. It is of note that the experiments were conducted in systems varying from *C. elegans* to human cell lines, and that for some of the findings cell extracts were used in addition to examining the system in a cellular environment and whole organisms. It is intriguing to think that the role of miRNAs and their mode of action may vary depending upon system and circumstance. Indeed Kong *et al.* has demonstrated that in HeLa cells, target mRNA constructs expressed from a simian virus 40 (SV40) promoter appear to be regulated at the initiation stage of translation, while mRNAs expressed from a thymidine kinase (TK)

promoter are repressed at a post-initiation stage. Surprisingly the promoter of the target gene appears to determine the form of translational inhibition (Kong et al., 2008).

1.1.2.5 Other miRNA mediated regulatory mechanisms

In addition, to the more widespread mechanisms of post-transcriptional gene silencing by miRNAs, there are what appear to be less common miRNA mechanisms. MiR-196 has been demonstrated to trigger the cleavage at a highly complementary site (with a single G:U wobble), in the 3'UTR of *Hoxb8* transcripts by the same mechanism used for siRNA directed mRNA cleavage (Yekta et al., 2004, Mansfield et al., 2004). In addition in mammalian systems, a cluster of maternally expressed miRNAs at the imprinted *antiPeg11* locus regulate a paternally expressed antisense transcript transcribed from a gene on the opposite strand (*Rtl1/Peg11*) (Davis et al., 2005). Again, these miRNAs appear to trigger the cleavage of their complementary mRNA target sequences. However, this method of targeting is massively outweighed by the more canonical targeting of partially complementary sites by miRNAs.

Under specific circumstances miRNAs can also cause the post-transcriptional up regulation of target transcripts (Vasudevan and Steitz, 2007). During cell cycle arrest three different miRNAs flipped their mode of regulation from suppression to activation.

In addition miRNAs have very recently been shown to activate (Place et al., 2008) and repress gene transcription in mammalian cells (Kim et al., 2008a). In both cases the miRNA appears to target each gene upstream of its start site. MiR-320 is encoded directly upstream

of the *POLR3D* gene and seems to direct Ago1 to the promoter and cause transcriptional silencing. MiR-320 expression also appears to correlate with H3K27me3 and EZH2 (a methyl transferase) at the promoter. In the case of transcriptional activation the miRNA does not match the target site with perfect complementarity, but miR-373 appears to cause the activation of E-cadherin (*CDH1*) and *CSDC2* by binding the promoter region, although the authors report no defined mechanism by which this occurs.

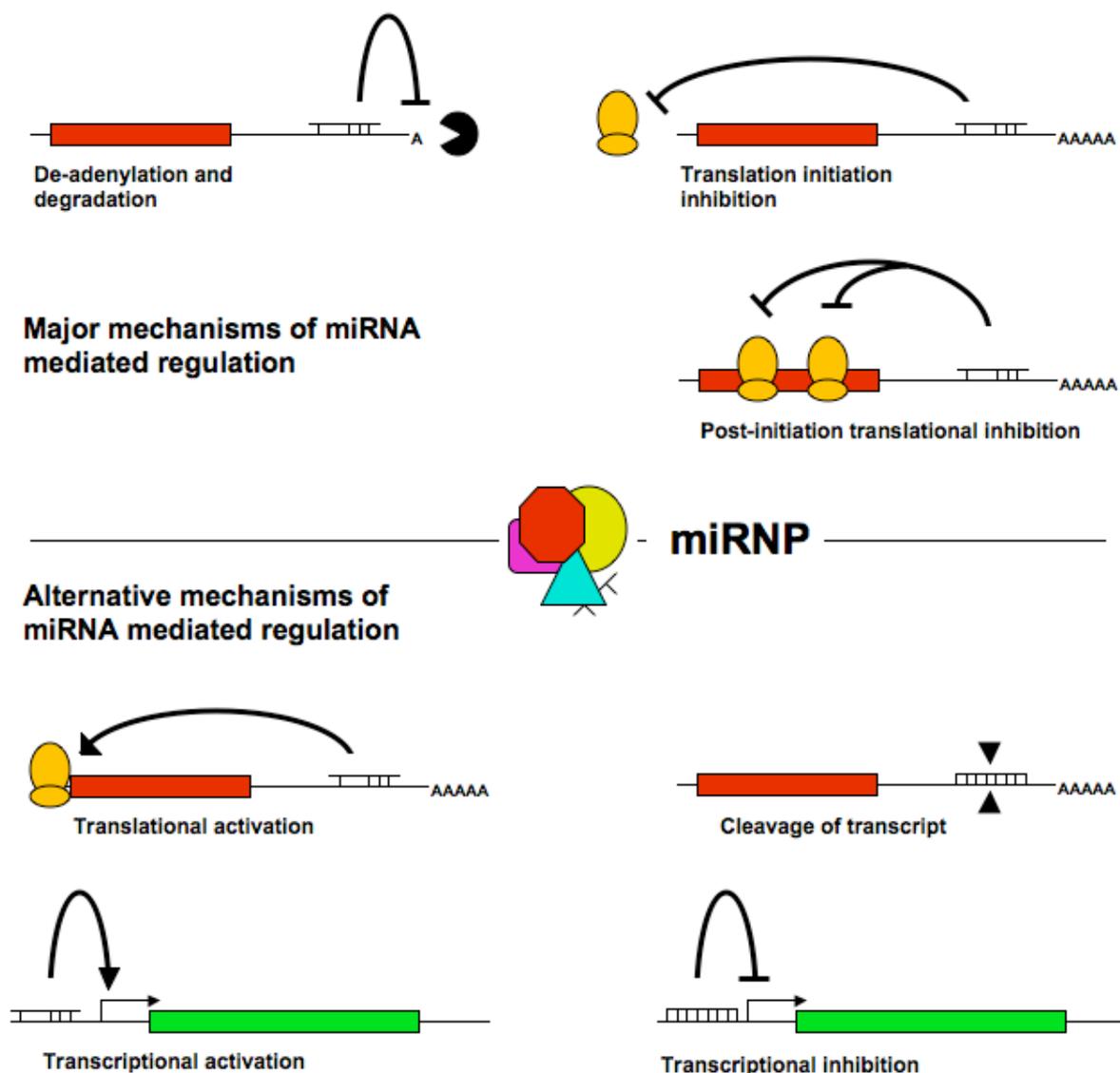


Fig.1.3: The mechanisms of miRNA mediated regulation. Top: Most common and widely researched miRNA/miRNP functions. Bottom: Alternative miRNA functions.

1.1.3 The rules of miRNA target recognition and target prediction algorithms

An increase in relevant data is leading to a better understanding of the rules which govern miRNA:mRNA target interactions and to rapid improvements in miRNA target prediction. In this section I intend to describe the progress in this field both before the inception of my studies and over the course of the last few years.

It was clear from some of the earliest miRNA targets identified experimentally that miRNAs bind targets of partial complementarity, with a preference for binding at the 5' end of the miRNA and within the 3'UTRs of mRNAs. MiRanda, one of the earliest miRNA target prediction algorithms, used filters based on these observations to select potential targets. Potential targets are identified by looking for complementarity across the length of the miRNA allowing G:U wobbles and a degree of mismatch. Preference is given to targets bound more completely at the 5' end. This initial matching process is followed by a series of filters which calculate both the thermal stability of the predicted targets and their evolutionary conservation across species. Multiple sites within the same UTR were summed to provide a list of the most confident predictions (Enright et al., 2003; John et al., 2004). A second method, TargetScan, used a signal/noise calculation based on the number of predicted targets for a true miRNA sequence divided by the number of targets predicted for shuffled sequence to define a set of rules for optimal target prediction (Lewis et al., 2003). This identified perfect complementarity to the “seed” sequences of each miRNAs (the region from bases 2-8 counted from the 5' end of the miRNA), (Fig.1.4) as one of the best predictors of miRNA targets. Better signal to noise ratios were achieved by a requirement for target site conservation. In addition to these criteria the program considered 3' complementarity, target

site free energy and the number of targets for each miRNA within a UTR as key criteria for target prediction. One perceived disadvantage of this method would be its propensity to miss targets with imperfect seed sequences, such as one of the let-7 target sites in the 3'UTR of the *lin-41* gene in *C. elegans* (Vella et al., 2004). At this early stage experimentally validated targets were rare so reporter assays were used to confirm target predictions. In brief, a segment from a 3'UTR containing a suspected target was cloned downstream of a luciferase open reading frame (ORF). An identical construct was also designed with point mutations within the target sites to disrupt miRNA binding. These target constructs were transfected into cells expressing the miRNAs of interest and the effect of the unmodified target site was monitored in relation to the mutated version to determine whether the relevant site is susceptible to miRNA targeting and induced post-transcriptional regulation (Lewis et al., 2003). Failing this, target prediction algorithms were assessed against the limited data available (John et al., 2004).

Studies using the reporter construct principle explained above and transfected into HeLa cells along with siRNA/miRNA duplexes, or in the presence of endogenously expressed miRNAs provided a further investigation of the properties of effective mRNA targets (Doench and Sharp, 2004). These experiments confirmed the importance of the 5' pairing of the miRNA to its target in leading to the down-regulation of the luciferase reporter, whereas pairing at the 3' end was deemed less important, although it was regarded as a modulating factor. G:U wobbles within the pairing appeared detrimental to miRNA control and miRNAs appeared to repress the reporter gene in a concentration dependent manner. Constructs with the target sites for multiple miRNAs/bulged-siRNAs in their 3' UTR also clearly demonstrated that miRNAs were able to control mRNAs in a combinatorial manner with multiple miRNA

regulating the same mRNA simultaneously by targeting different sites in the UTR. The fold repression of a construct with 2 sets of 2 target sites within its 3' UTR, each set partially complementary to a different siRNA, increased from approximately 3 fold to 8 fold if both siRNAs were introduced together as opposed to one at a time.

Brennecke *et al.* extended this work, expressing green fluorescent protein (GFP) with target sites in its 3'UTR in an imaginal disc of *Drosophila*. miRNAs expressed within a region of the same disc were used to assess the effectiveness of each target site (Brennecke et al., 2005). The authors found that perfect complementarity within the region from 5' bases 2-8 of the miRNA was sufficient to confer repression of a target without complementarity within the 3' end of the miRNA. Complementarity between bases 1-8 of the miRNA and the target provided even greater repression of the target mRNA. They found no correlation with the pairing energy of the 5' bases. Targets with limited 5' pairing to the miRNA (miRNA bases 2-5) were functional but required substantial pairing at the 3' end of the miRNA. Again G:U bases within the seed appeared detrimental but potentially tolerable for more limited function. Ultimately the authors defined 3 types of target; 1) Canonical sites with substantial pairing at both the 5' and 3' ends of the miRNA; 2) Seed sites with 5' pairing but little 3' pairing; 3) 3' compensatory sites with at least 4 bases paired within the seed and strong pairing at the 3' end of the miRNA. The seed sequences of true miRNAs were found to be more conserved within 3' UTRs than random sequences, while regions adjacent to these sites were rarely conserved. In addition the authors estimated that there are probably only 1-20 3' compensatory sites per miRNA.

Transfection of miRNA duplexes for miR-1 and miR-124 into HeLa cells caused a reorganization of the cellular expression profile, as measured by microarray, to more closely resemble the profiles of those tissues within which these miRNAs are normally expressed (Lim et al., 2005). The transfection of miRNAs and the use of microarray technology allowed the experimental investigation of miRNA-target interactions on a large scale. Once again down-regulated transcripts demonstrated an enrichment for sequences complementary to bases 2-7 of the transfected miRNA, again revealing the significant contribution of this region of the miRNA to target selection.

The release of the chicken genome led to an update of TargetScan; TargetScanS (Lewis et al., 2005). Using the same signal to noise ratio as before to judge effectiveness, by extending required conservation to include 5 genomes, including the chicken genome, the algorithm was stripped back to predict targets based solely on the conservation of seed sequences (miRNA bases 2-7) (Fig.1.4). The signal to noise ratios were improved by requiring a Watson-Crick match between base 8 of the miRNA and the target (7mer-m8 seed) or by requiring the target base opposite the first base of the miRNA to be an A (7mer-t1A). Requiring both conditions to be met improved the signal to noise ratio still further (8mer). Imperfect seeds increased the associated noise. The authors also found a relatively faint signal for miRNA targets existing within the ORF of genes.

At a similar time a new algorithm was released; PicTar. This attempted to account for the synergistic and combinatorial effects of multiple targets within the same 3' UTR mentioned in Doench *et al.* (Doench and Sharp, 2004) to predict genes most likely to be under miRNA control. The program identifies miRNA base 1-7 and 2-8, 7nt complementary sites in 3'

UTRs and then calculates miRNA:target 3' complementarity. It subsequently filters sites for free energy of the association and by requiring “anchors” in multiple UTR alignments and uses this to make the initial target calls. The method subsequently provides a PicTar score for multiple targets within the same UTR.

At the time of the inception of my PhD, Giraldez *et al.* derived zebrafish embryos lacking both maternal and zygotic Dicer (*dicer1*) function (Giraldez et al., 2005). Expression changes were judged by array and initially genes up-regulated upon miRNA removal were derived from these embryos and compared to transcripts whose expression was altered upon the re-addition of miR-430 by microinjection (Giraldez et al., 2006). The intersection of genes whose expression was up-regulated in the Dicer mutant when compared to the two alternative conditions were searched for an enrichment of miR-430 seed sequences in their 3' UTRs. Of the 328 genes in the intersected region with annotated 3' UTRs, there was a significant enrichment for the miR-430 seed sequence.

At this point a comparison was made between the targets predicted by different methods (Sethupathy et al., 2006). Using sets of experimentally verified targets, the authors of this paper tested each algorithm for its ability to identify targets within this set. Interestingly miRanda, TargetScanS and PicTar algorithms could only identify approximately 45-50% of the targets and roughly 2/3 of conserved targets, with miRanda making roughly 7000-8000 more predictions than PicTar or TargetScanS. Also notable was that although PicTar and TargetScanS appeared to share a large proportion of their predictions, when predictions for these three methods were overlapped, the intersection covered only 40% of conserved targets.

The programmes were clearly making substantially different predictions, suggesting that a number of rules for target prediction remained to be found.

Further use of Luciferase reporter assays demonstrated that nonconserved 7 or 8nt miRNA “seed” matches could affect the translation of reporter genes in the same way as TargetScanS predicted conserved targets (Farh et al., 2005). This suggested that a whole potentially important class of miRNAs was being missed by demanding conservation of targets as a factor for their prediction.

Subsequently, miRanda has been updated for use in miRBase (miRBase Targets) (Griffiths-Jones et al., 2008). Within miRBase Targets the miRanda algorithm is used to calculate scores for a particular miRNA’s predicted target sites based on complementarity, weighted for increased significance for complementarity at the 5’ end of the miRNA. These scores are then incorporated into a *P*-value calculation along with the number of additional sites for a specific miRNA in the relevant 3’UTR. Conservation is also considered within the *P*-value calculation. Targets are no longer judged according to their thermostability or the requirement to be conserved, although generally targets with a higher degree of conservation will be attributed with a more significant *P*-value.

Currently, high throughput target identification is beginning to have an impact on elucidating miRNA targeting rules, filling the void left by the slow pace of miRNA target identification. Grimson *et al.* analysed data from 11 miRNA over expression experiments in HeLa cells (Grimson et al., 2007). Following transfection the RNA of these cells was purified and arrays were used to assess mRNA degradation on a large scale. They confirmed a multiplicative

effect of multiple targets within the same 3'UTR although intervals of 8-40nt produced an even stronger repressive effect. Pairing energy was seen to be a bad indicator of the efficacy of 3' pairing. Instead 4 base contiguous pairing starting at miRNA bases 13 to 16 had the greatest effect on down regulation. Again the authors concluded that canonical sites described by Brennecke *et al.* are rare (Brennecke et al., 2005). Functional sites were generally found in regions rich in AU bases. Interestingly, although 8mer sites were seen to have no detectable effect on expression when found in 5' UTRs, they appeared to have a mild effect on expression when found in the ORF of a gene. This dampened effect seen with 8mers within an ORF extended 15nt into the 3'UTR. In addition targets near the ends of the 3'UTR were seen to have a greater effect on expression than those in the middle. These additional findings were used to calculate a "context score" which was applied to both conserved and nonconserved 7mer and 8mer seed sites in TargetScanS. This score was seen to be a good predictor of the efficacy of miRNA target sites when used in concert with luciferase target site confirmation assays described above.

TargetScanS has recently been updated again, using a new method to normalize the speed of evolution between different sets of UTRs and therefore removing the requirement to separate conserved and non-conserved sites into independent groups (Friedman et al., 2009). This new method allows a much more sensitive analysis of sites and increases the proportion of predicted target genes for a combined set of all human miRNAs to approximately 60% of human protein coding genes at a "conservation cutoff of 1.0". They note that the order of signal to background ratios for the conservation of seed sequences reflects the target site efficacy, placing target "seed" sequences in an effective order of 8mer > 7mer-m8 > 7mer-tA1 > 6mer > offset 6mer (complementary to bases 3-8 of the miRNA sequence) (Fig.1.4). At

the same cutoff as above, of the non-complementary seed sequences only 8mer seeds with a bulge appeared to be under effective selection suggesting non-seed targets are rare. Compensatory sites also appeared to be rare with an estimated 4.5 such sites conserved for each miRNA family. An estimated 4.9% of all preferentially conserved target sites were supplemented by conserved 3' pairing.

It is worth noting that the significance of the 5' complementarity between miRNAs and their targets implies that miRNA families, which share a high degree of sequence similarity, are likely to target the same mRNA target sites and consequently to be redundant in function to some extent, if co-expressed.

In addition to these most widely used algorithms many further methods exist for target prediction. The “Probability of interaction by target accessibility” algorithm (PITA) is one of the most original (Kertesz et al., 2007). Instead of using conservation to select functional targets, PITA uses a filter based on the thermodynamic stability of secondary structures at the target site to filter targets into those likely to be effective and those that are not. Furthermore, they found a tendency for miRNA seed sequences to be positioned in thermodynamically accessible regions of the UTR.

Recently, functional miRNA targets sites situated within the coding regions of several genes have been identified in the mouse (Tay et al., 2008). Multiple targets have been found in the ORF of Oct4 (*Pou5f1*), *Sox2* and *Nanog* genes. In some cases these target sites cross exon boundaries, are not conserved or have incomplete base pairing within the seed region. All of

these factors combined suggest that there are still likely to be many functional target sites not identified by prediction methods for a variety of reasons.

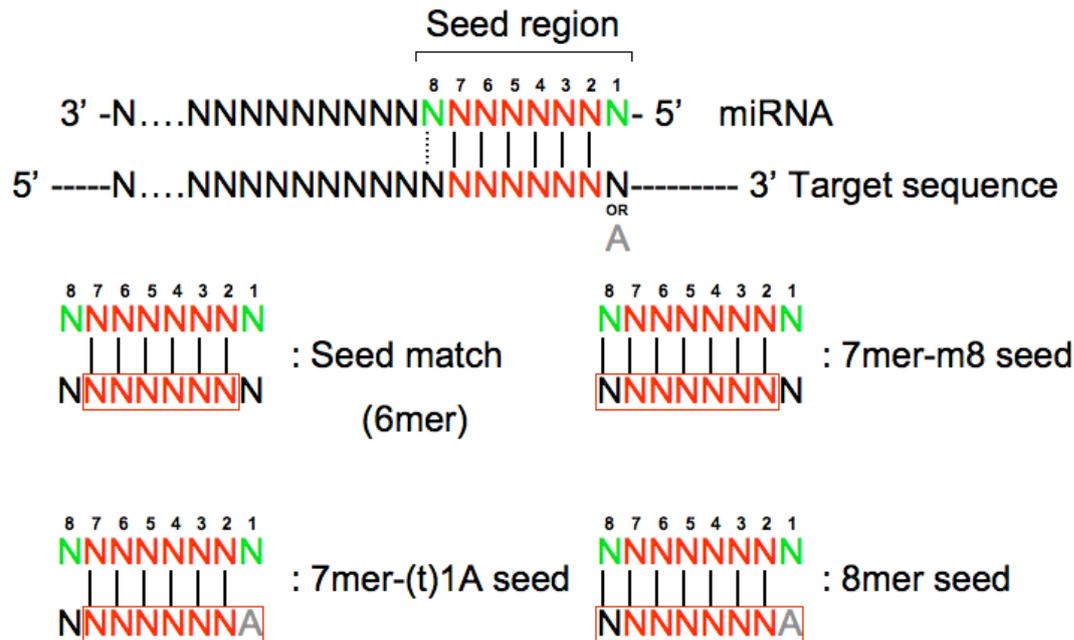


Fig.1.4: The miRNA seed region and the definitions of $-mer$ target sites as given by <http://www.targetscan.org>. The red box outlines the target bases within the mRNA referred to in each case. Black lines indicate complementary bases. Red bases indicate the core 6mer. Numbered bases are counted from the 5' end of the miRNA.

1.1.4 Endogenous siRNAs

Also of interest as a closely related small RNA family is the recent discovery of endogenous siRNAs in mammals. Until very recently it had been assumed that endogenous siRNAs were restricted to those species that possess an RNA-dependent-RNA-polymerase, an enzyme not found in mammals, capable of generating long dsRNA substrates for Dicer, which would in turn generate siRNAs perfectly complementary to their targets. As a result it was assumed that RNAi in mammals was only targeted by dsRNA supplied from external sources. This theory was compounded by the presence of an interferon response in vertebrates that would be triggered by long dsRNAs present in the cells. However, the work of Babiarz *et al.*

(Babiarz et al., 2008) alongside the work of other teams (Tam et al., 2008; Watanabe et al., 2008) have now identified endogenously derived siRNAs from repeat sequences and/or pseudogenes in mouse ES cells and oocytes. These siRNAs appear to be functional. In the case of Babiarz *et al.* they are demonstrated to be derived in a Dgcr8-independent, Dicer-dependent manner. Tam *et al.* speculate that these siRNAs will operate by a conventional RNAi mechanism, leading to the degradation of their targets, while Watanabe *et al.* demonstrate that Dicer (DICER1) and AGO2 (EIF2C2) (slicer competent AGO) both appear to be required for the endogenous siRNA driven regulation of transposons and pseudogenes, through examining the expression levels of predicted targets in knockout oocytes. It is also worth noting that both oocytes and embryonic stem cells are devoid of the dsRNA dependent interferon response.

1.2 Embryonic stem cells

Mouse ES cells are cells derived from the inner cell mass of the blastocyst (Kaufman et al., 1983; Martin, 1981). They are pluripotent, (able to differentiate into all somatic lineages and, if introduced into pre-implantation embryos, colonise all foetal lineages including the germ cells in addition to some extra-embryonic tissues) and they are capable of self-renewal (Beddington and Robertson, 1989; Bradley et al., 1984).

1.2.1 Transcriptional networks for maintaining the stem cell state

At the centre of the transcriptional network responsible for maintaining the pluripotency of stem cells reside a set of transcription factors, that include Oct4 (Nichols et al., 1998; Niwa et al., 2000), *Sox2* (Avilion et al., 2003; Kopp et al., 2008) and *Nanog* (Mitsui et al., 2003). The perturbation of any of the factors given above has been demonstrated to have a profound

effect on the stability of the undifferentiated state. Expression of these genes has subsequently been used as a marker for embryonic stem cells.

The molecular factors underlying pluripotency have been further elucidated through the generation of induced pluripotent stem (iPS) cells by the introduction of artificially expressed gene combinations into differentiated somatic cell types. Initially the factors required for this reprogramming were Oct4, *Sox2*, *Klf4* and c-Myc (*Myc*), introduced to the cells using retroviruses (Takahashi and Yamanaka, 2006). The exact role that each of these genes plays has yet to be fully understood, as have the intricacies of the process. Subsequent experiments have reduced the number of factors required to derive iPS cells. Nakagawa *et al.* were able to derive iPS cells without c-Myc, a known oncogene (Nakagawa *et al.*, 2008) while Huangfu *et al.* derived human iPS cells using Oct4 and *Sox2* and the histone deacetylase inhibitor, valproic acid (Huangfu *et al.*, 2008). Given the role of *Nanog* in the maintenance of pluripotency in embryonic stem cells and its ability to improve the transfer of pluripotency to cells in fusion experiments (Silva *et al.*, 2006), it is perhaps surprising that *Nanog* was not one of the required factors. As techniques and methods have been improved, iPS cells have been generated that are broadly comparable to ES cells and that are capable of producing germline competent chimeras to contribute to further generations (Okita *et al.*, 2007).

In addition to the endogenous transcriptional network required to maintain ES cells in an undifferentiated state, exogenous signals are also necessary (reviewed in (Okita and Yamanaka, 2006)). Traditionally mouse ES cells have been maintained in culture in the presence of leukaemia inhibitory factor (LIF) and serum or bone morphogenetic protein (BMP) (Ying *et al.*, 2003). The BMP (or serum) induction of inhibitor-of-differentiation

genes and LIF induction of STAT3 signaling pathways ensures that the cells are unable to differentiate. Recently it has been demonstrated, however, that these two pathways are not necessarily required to maintain stem cell identity, but may instead dampen the effects of exogenous, pro-differentiation stimuli (Ying et al., 2008). These culture conditions can be substituted by the disruption of the mitogen-activated protein kinase (MAPK) pathway and the glycogen synthase kinase 3 (GSK3) pathway.

Attempts are being made to integrate both the transcriptional networks for self-renewal and pluripotency and the effectors of important cell signaling pathways into a unified system through ChIP and affinity purification experiments. Wang *et al.* used biotinylated proteins to purify the interacting partners from the pluripotency network, identifying interactions between a large number of proteins with known involvement in differentiation or the integrity of the inner cell mass (Wang et al., 2006). They later followed this work with a ChIP-Chip survey of the promoter occupancy of the factors identified by Takahashi *et al.* (Takahashi and Yamanaka, 2006) and a selection of previously identified interacting partners; 9 proteins in total (*Nanog*, *Oct4*, *Sox2*, *Klf4*, *Myc*, *Dax1 (Nr0b1)*, *Nac1 (Nacc1)*, *Zfp281*, *Rex1 (Zfp42)*) (Kim et al., 2008b). This work demonstrated the apparent complexity of the coordinated regulation of the network's targets. Approximately 800 gene promoters were bound by 4 or more of the proteins tested while approximately 50% were bound by only a single factor. MYC and Rex1 appeared to bind a different set of targets from the other transcription factors (TFs) tested. 96% of MYC promoters had a H3K4me3 signature that implies the chromatin region is open and active. This complements data that suggested that MYC targets have a generally greater expression in ES cells than genes with promoters bound by the other transcription factors. The targets of the other genes included proteins both expressed and

depleted in ES cells. Interestingly, the gene sets with a greater number of this subset of transcription factors associated with their promoters seem to have greater expression in ES cells.

Additional, independent, ChIP studies have been conducted repeating a number of these experiments using slightly different protocols. Chen *et al.* performed a series of ChIP-sequencing (ChIP-Seq) assessments for a similar set of genes (*Nanog*, *Oct4*, *Sox2*, *Klf4*, *c-Myc* (*Myc*), *n-Myc* (*Mycn*), *Esrrb*, *E2f1*, *Zfx*, *Smad1*, *Stat3*, *Tcfcp2l1*, *Ctcf*) (Chen et al., 2008). The authors identified sites bound by combinations of these TFs and once again they found groups of factors whose binding correlated into clusters, with *c-Myc* tending to bind with different factors to NANOG, SOX2 and Oct4. They noticed p300 (EP300) histone acetylase, known to associate with enhancer elements, tended to bind sites with 3-6 other TFs from the NANOG, SOX2, Oct4 group. They proceeded to find that 60% of genes that are up regulated in ES cells, when compared to differentiated cells tend to be associated with target sites enriched for NANOG, Oct4, SOX2, SMAD1, STAT3, *c-Myc* and *n-Myc* binding.

1.2.2 ES cell cycle

Mouse embryonic stem cells have a drastically shortened G1 phase of their cell cycle, compared to most somatic cells, which allow them to proliferate rapidly. This is the result of constitutively active cyclinE:CDK2, low cyclinD and CDK4 activity and permanently hyperphosphorylated Rb protein. Within somatic cells mitogen signaling induces the activity of CyclinD:CDK in early G1 phase. The increase in cyclinD:CDK activity ultimately leads to the hyperphosphorylation of Rb. Rb hyperphosphorylation allows CyclinE:CDK2 to become active and hence the cell cycle can proceed into S phase. Mouse ES cells essentially remove

the requirement for G1 mitogen control. Mitogens can also trigger the differentiation of ES cells, so this omission allows the cell cycle to continue without this risk of mitogen induced differentiation (reviewed in (Orford and Scadden, 2008)).

1.2.3 miRNAs and mouse ES Cells

1.2.3.1 The role of miRNAs in stem cells – Perturbing the processing pathway

In 2003 an attempt was made to breed a mouse with a homozygous, null Dicer gene, in order to further explore the role of miRNAs in development (Bernstein et al., 2003). However, homozygous mutants displayed an embryonic lethal phenotype. Development appears severely disrupted prior to embryonic day 7.5 (E7.5) with loss of embryonic Oct4 staining implying the loss of the stem cell population in early development. Subsequently the role of miRNAs in stem cells has become a focus for a number of laboratories.

Early experiments with hybrid DT40 chicken cells containing human chromosome 21 seemed to support the notion that Dicer (DICER1) dependent siRNAs derived from peri-centromeric transcripts played a role in centromeric structure, heterochromatin formation and cell division (Fukagawa et al., 2004). This work is complemented by later experiments in Dicer deficient mouse embryonic stem cells.

As expected, the removal of Dicer from the mouse ES cells seemed to lead to a loss of mature miRNAs and an accumulation of pre-miRNA transcripts (Kanellopoulou et al., 2005). However, it also seemed to lead to an increase in dsRNA species derived from centromeric

satellite repeats, which the authors suggest may be processed into short RNA species in a Dicer dependent manner. This correlates with changes in methylation and histone modification profiles. These findings remain controversial, however, in particular the concept of Dicer playing a major and direct role in chromatin modification via an RNA mediated process. Indeed, these findings were not replicated by a later study that found no changes in centromeric satellite associated DNA methylation or histone modification in Dicer deficient mouse ES cells (Murchison et al., 2005). However, the potential for a Dicer mediated role in the maintenance of heterochromatin, makes it more difficult to interpret the cellular functions of miRNAs in ES cells from the results of these knockout studies in isolation.

More recently, Wang *et al.* generated a *Dgcr8* conditional knockout mouse embryonic stem cell line (Wang et al., 2007). As DGCR8 plays no part in the generation of siRNAs from dsRNA substrates, DGCR8 was considered to be a more likely candidate for the generation of a miRNA specific phenotype. The *Dgcr8* null genotype also proved to be embryonic lethal and the *Dgcr8* mutants replicated a number of other phenotypes seen in Dicer mutant stem cells. All three mutant stem cell lines had a reduced rate of proliferation with both Murchison *et al.* and Wang *et al.* detecting an increase the number of cells in the G1 phase of the cell cycle in the absence of miRNAs (Murchison et al., 2005; Wang et al., 2007).

Furthermore, the loss of miRNAs had a profound effect on the differentiation potential of the stem cells. Dicer deficient cells were unable to form teratomas when injected subcutaneously into immuno-deficient mice (Kanellopoulou et al., 2005). Embryoid bodies formed from these cells grew for 8-10 days and then arrested and none of the markers of differentiation tested were expressed in these Dicer knockout embryoid bodies. In contrast, *Dgcr8* knockout

cells appeared to have a less profoundly compromised differentiation potential (Wang et al., 2007). They successfully formed teratomas following subcutaneous injection, although they appeared largely undifferentiated in structure. They also expressed a number of the markers of various cell lineages upon embryoid body induced differentiation and embryoid body growth did not arrest. The authors do note however that the mutant stem cells appear unable to silence pluripotency markers in the course of differentiation, with a larger proportion of mutant cells reverting to ES cell like growth following a period of induced differentiation than seen with control cells.

The differences evident between the phenotypes of these various knock out cell lines may in part be explained by molecular functions specific to either DGCR8 or Dicer. Pyrosequencing of the small RNA fraction of Dicer knockout cell lines and wild-type ES cells failed to identify the population of Dicer dependent centromeric heterochromatin associated siRNAs hypothesized by Kanellopoulou *et al* (Calabrese et al., 2007). However, as discussed in the sections “Mirtrons and other exceptions to the canonical rules” (see section 1.1.1.6) and “Endogenous siRNAs” (see section 1.1.4), a number of Dicer specific small RNAs have been identified by Illumina and 454 sequencing that may account for some of these phenotypic discrepancies (Babiarz et al., 2008).

1.2.3.2 miRNA expression in stem cells

Initially, ES cell miRNA expression profiles were constructed by cloning small RNAs and subsequently sequencing them. This process identified large numbers of previously unannotated miRNAs and revealed a number of miRNAs that seemed to be present specifically within mouse ES cells (Houbaviy et al., 2003). Perhaps the most notable, novel

miRNAs were found within a cluster on Chromosome 7 (the miR-290 cluster). A number of the miRNAs that have been subsequently ascribed to this cluster share a common seed sequence (7mer-1A; GCACTTA; mmu-miR-291a-3p, -291b-3p, -292-3p, -294, -295). By sharing the same seed it is expected that there can be a considerable overlap between the targets of these miRNAs. In 2004 a similar study was conducted in human ES cells (Suh et al., 2004). This study identified an apparently ES cell specific cluster (has-miR-371, -372, -373), that is orthologous to the mmu-miR-290 cluster. The human ES cells also expressed an orthologous miR-302 cluster that had been sequenced in the mouse ES cells. miR-302 shares a seed sequence with the mmu-miR-290 cluster.

As the number of annotated miRNA sequences was expanded, microarrays were designed to profile cells for the expression of known miRNAs (Laurent et al., 2008), as well as quantitative RT-PCR (qRT-PCR) assays that could profile miRNAs from a single ES cell (Tang et al., 2006). The latter could prove an important strategy for the dissection of self renewal as ES cells form notoriously heterogeneous culture populations, containing small populations of spontaneously differentiated cells and cells expressing marker proteins at different levels (reviewed in (Silva and Smith, 2008)).

Ultimately, high throughput sequencing techniques have been employed to ascertain miRNA expression with a degree of sensitivity and specificity, unobtainable with the aforementioned techniques. The depth at which these new technologies allow the miRNA expression pool to be sampled allows miRNAs expressed at a low level to be detected and annotated. Pyrosequencing of the miRNA population from mouse ES cells attributes the majority (70-76%) of miRNA expression to 6 loci in the genome, some of which are home to clusters of

miRNAs (Calabrese et al., 2007). These include the mmu-miR-290 cluster (the most highly expressed cluster) and a cluster containing mmu-miR-467a and its paralogues. Once again mmu-miR-467a shares the same 7mer-1A seed as mmu-miR-291a-3p.

By contrast, human ES cells seem to express has-miR-302a and its paralogues at a far greater level than the has-miR-371 cluster, as measured by Solexa sequencing (Morin et al., 2008). Although this study identified many isomers of the canonical (miRBase annotated) mature miRNA sequences, with a variety of 5' and 3' extensions that may alter the seed sequence, this predominance of miR-302 over miR-371 to miR-373 is replicated by further studies, measured by pyrosequencing (Bar et al., 2008). It is intriguing that the predominant miRNAs expressed in human and mouse ES cells are different, while maintaining a common miRNA seed sequence, perhaps underlying a degree of redundancy in function between the two miRNA families. It is also worth bearing in mind that there are known phenotypic differences between mouse ES cells and human ES cells and any comparisons made between the two systems should be made with due caution (Discussed in (Tesar et al., 2007)).

A ChIP-Seq study has investigated the transcriptional control of miRNAs in mouse ES cells and the association of Oct4, SOX2, NANOG and TCF3 (a further TF) at miRNA promoters and correlated this with miRNA expression (Marson et al., 2008). It seems that as with protein coding genes, these transcription factors control miRNAs that are both activated and repressed in ES cells. In this way a putative and simple series of networks have been constructed, demonstrating the roles of miRNA in both coherent and incoherent feed-forward control of ES cell protein expression, fine tuning protein expression and poisoning the cells for differentiation.

1.2.3.3 The role of miRNAs in stem cells

Even if the recent prediction that 60% of human genes are likely to be targeted by miRNAs is considered to be an over-estimate (Friedman et al., 2009), when considering the ever-increasing number of annotated miRNAs the number of functional targets annotated for miRNAs remains tiny.

In mouse ES cells, miRNAs modulate the activity of DNA methyltransferases. Independent studies reported that miRNAs from the miR-290 cluster (mmu-miR-291-3p, -292-3p, -294 and -295) post-transcriptionally regulate the expression of *Rbl2* which in turn down-regulates *Dnmt3a* and *Dnmt3b* expression (Benetti et al., 2008; Sinkkonen et al., 2008). Dicer deficient mouse ES cells have decreased levels of these methyltransferases (in addition to *Dnmt1*) and exhibit global hypomethylation, substantial telomere length changes and increases in telomeric recombination (Benetti et al., 2008). The introduction of the miRNAs listed above into these cells led to decreases in *Rbl2* levels and increases in *Dnmt3a* and *Dnmt3b*. This regulation of *Dnmt3a* and *Dnmt3b* expression, via *Rbl2*, by the miR-290 cluster also has an important role in the silencing of the Oct4 promoter upon ES cell differentiation (Sinkkonen et al., 2008). Dicer deficient mouse ES cells are unable to efficiently silence the Oct4 promoter by methylation, upon differentiation. Interestingly, the transfection of the miR-290 cluster upon differentiation rescued this phenotype. In order to identify targets in this study, the authors transfected the Dicer knock out cells, deficient in mature miRNAs, with miRNA mimics and then assessed the resultant changes in mRNA levels by array. This system allowed the authors to identify miRNA targets in a system with no interference from functionally redundant miRNA families.

As implied by the results of the miRNA processing pathway disruption experiments described above, miRNAs also contribute to the regulation of the cell cycle. Wan *et al.* performed a screen of 266 mouse miRNAs; transfecting them into *Dgcr8* knock out mouse ES cells and examining the cell cycle for changes (Wang *et al.*, 2008). Once again they identified mmu-miR-291-3p, -292-3p, -294, -295 and miRNAs sharing their seed sequence (including miR-302a and homologues) as increasing the proliferative rate of the mutant ES cells. Further investigation of the mechanism by which these miRNAs achieved this revealed p21 (*Cdkn1a*), a cyclinE-CDK2 inhibitor as a target of mmu-miR-291-3p, -292-3p, -294, -295 and miR-302d.

A further intriguing result published by Lin *et al.* concerned the transfection of cancerous cell lines with a retrovirus expressing the miR-302 cluster (Lin *et al.*, 2008a). This was sufficient to convert these cell lines into an apparently pluripotent state, combined with the expression of various human ES cell markers, including Oct4. This implies that miRNAs may have a significant role to play at the very centre of the regulatory network which controls pluripotency.

1.2.3.4 The role of miRNAs in stem cells – Lessons from cancer

A large number of the miRNAs that are highly expressed in mouse and human ES cells have also been demonstrated to play a role in the pathogenesis of various forms of cancer. Intensive studies conducted in cancer and the identification of multiple targets of these miRNAs provides a resource when considering miRNA roles in stem cells.

Hsa-miR-373 (an orthologue of the mmu-miR-290 cluster of miRNAs) has been demonstrated to contribute to the migratory potential of cancer cell lines in part through the target gene *CD44* (Huang et al., 2008). In addition, a screen for miRNAs that can confer a degree of resistance to oncogenic-induced senescence identified both has-miR-372 and has-miR-373 as candidates (Voorhoeve et al., 2006). These miRNAs target *LATS2*, which is known to influence G1 to S phase transition in the cell cycle.

The miR-17-92 cluster was renamed oncomiR-1 after it was found to possess oncogenic potential (He et al., 2005). The cluster has been demonstrated to play a vital role in development, with its deletion in mice causing death at birth (Ventura et al., 2008). This cluster belongs to a set of three highly conserved clusters present in the mouse genome (including mmu-miR-106b-25 and mmu-miR-106a-363). Deletion of either of the other two clusters had no visible phenotype. However, when both mmu-miR-17-92 and mmu-miR-106b-25 were deleted, the embryos died before E15. This more severe phenotype suggests that the mmu-miR-17-92 cluster and the mmu-miR-106b-25 are to some extent redundant in their functions. This functional relationship is discussed in a review by Petrocca *et al.* (Petrocca et al., 2008b). These two clusters cooperate in the regulation of the TGF β signaling pathway. MiR-106b, -93, -17 and -20a down regulate p21, while miR-25 and miR-92a-1 target Bim (a pro-apoptotic gene (*Bcl2l1*)). Both clusters are under *Myc* and *E2f1* regulation and both clusters (miR-106b, -93, -17 and -20a) create a feedback loop to down-regulate *E2f1*.

A further ES cell expressed miRNA, miR-21, down regulates a number of tumour suppressor genes and is over expressed in a wide variety of cancers. It has recently been found to target

the tumour suppressor gene *PDCD4* (Lu et al., 2008), in addition to previously described targets *PTEN* and *TPMI*. Contrary to this observation, identifying a miRNA acting as an oncogene, Liu *et al.* identified *CCND1*, *CCND3*, *CCNE1* and *CDK6* as miR-16 targets (Liu et al., 2008a). The authors used a HepG2 cell line depleted for Droscha (*RNASEN*) to investigate the extent of miRNA post-transcriptional control and to search expression profiles for genes mis-regulated in cancers that could be miRNA targets. As would be expected miR-16 over expression caused A549 cells (human lung carcinoma) to accumulate in G1 phase.

1.3 Aims of this project

While there have been constant and rapid advances in the field of miRNA target identification, there remains a large discrepancy between the number of known miRNA targets and the number of annotated miRNAs in miRBase. As a consequence there is a need for simple and effective methods by which to generate large numbers of experimentally supported miRNA-target interactions. Such data could then be used to both optimize miRNA target prediction algorithms and to directly annotate miRNAs with associated functional information, upon which to build hypotheses for further experimentation.

One novel approach developed as this project was being conceived was adopted by Giraldez *et al.* in zebrafish maternal-zygotic Dicer (*MZDicer*) mutant embryos (Giraldez et al., 2006). In order to eliminate the maternally contributed *dicer* activity from the early zebrafish embryo, Giraldez *et al.* conducted a germ line replacement experiment, whereby they depleted the germ cells from a wild type fish and reconstitute the germ line with *dicer1*^{-/-} cells. Hence the offspring of a cross between these fish would no longer exhibit any endogenous Dicer activity. These *MZDicer* embryos allowed for the identification of a large

number of putative miRNA targets through the ectopic expression of a miRNA in conditions where endogenous targets are expressed. Target-miRNA associations were judged by monitoring the down-regulation of mRNA by expression array, following the injection of miRNAs into the mutant embryos. By drawing an intersection between those genes down-regulated when miRNAs are added and mRNAs up-regulated upon the removal of Dicer, a list of gene transcripts predicted to be enriched for specific miRNA targets was derived. The prediction of these putative targets would not be restricted by repression of other miRNAs working in concert with the miRNA of interest or by miRNAs sharing seeds with the introduced miRNA, which may mask its targets through functional redundancy and potential saturation of target sites. In this way the predicted target set would be expected to be “cleaner” than sets derived from over expressed miRNAs in unrelated systems (Lim et al., 2005), or target sets derived from studies in which a single miRNA has been disrupted, regardless of the aforementioned redundancy.

We decided to attempt a similar approach in a cell-based system in mammals, providing an easily manipulated foundation for the generation of miRNA target lists in a mammalian context. The aims were as follows:

- 1) To develop cell lines with a depletion of endogenous miRNA expression through the disruption of the *Dgcr8* locus using ES cell gene traps and homologous recombination.
- 2) To assess the expression of miRNAs in these mutant cells and control cell lines and so identify miRNAs expressed endogenously in a wild type and heterozygous background and to search for DGCR8 independent miRNA expression.

- 3) To optimise transfection conditions for the reintroduction of miRNA mimics into these cells.
- 4) To reintroduce selected miRNAs expressed in control cell lines into the miRNA depleted cells and monitor mRNA expression via microarrays.
- 5) To determine gene lists enriched for miRNA targets by comparing genes up-regulated upon removal of miRNAs from the system to genes down-regulated upon miRNA reintroduction.

Chapter 2: Materials and Methods

All solutions used are listed at the end of the chapter.

All chemicals were obtained from Sigma unless otherwise stated.

2.1 Centrifuges

Listed below are the centrifuges used at various stages of this project:

- Eppendorf Centrifuge 5417R – used for all protocols unless otherwise stated
- Eppendorf Centrifuge 5702 – ES Cell culture (section 2.4), Optimised EB formation protocol (section 2.5.3) and Cell cycle assay: Cell Growth (section 2.5.4.1)
- Thermo Electron Heraeus Sepatech Megafuge 1.0 – Fluorescence activated cell sorting (section 2.5.4.2) and Harvesting HeLa S3 RNA (section 2.7.4)
- Beckman J6-MC – Trizol RNA purification of large volumes in 50ml Falcon tubes (section 2.7.1)
- Thermo Electron Heraeus Pico 21 – Preparing Illumina/Solexa samples (section 2.10.3.1) and Northern blot preparation (section 2.9.1.1)
- Eppendorf Plate Centrifuge 5403 – KIF11 siRNA transfection (section 2.12.2)

2.2 General techniques

2.2.1 Phenol chloroform extraction

0.5 vol. phenol and 0.5 vol. chloroform (24 chloroform : 1 isoamylalcohol) were added to the DNA. The tube was mixed by shaking and spun at maximum for 15 minutes to separate phases. The aqueous phase was removed to a fresh tube and an

equal volume of chloroform (24 chloroform : 1 isoamylalcohol) was added. The solution was mixed by shaking and the phases were separated by spinning for 5 minutes. The aqueous layer was removed to a separate tube and ethanol precipitated (see section (2.2.2)).

2.2.2 DNA ethanol precipitation

2 volumes of ice cold 100% ethanol and 1/10 volume of 3 M sodium acetate were added to the DNA in aqueous solution. The solution was mixed by inversion and placed at -20°C overnight. The DNA was then pelleted by spinning at maximum for 30 minutes at 4°C. The DNA was washed with >500 µl of 70%-75% ethanol and spun for 5 minutes at 4°C. Finally the pellet was air dried and resuspended in T_{0.1}E.

2.2.3 Small RNA isopropanol precipitation

An equal volume of isopropanol was added to the RNA sample and 3 µl of GlycoBlue (Ambion). The contents were mixed by inversion and placed on ice for 30 minutes. The spin and washes were conducted following the ethanol precipitation protocol (see section 2.2.2), but using RNase free solutions. The pellet was resuspended in RNase free water.

2.2.4 PCR

Standard polymerase chain reaction (PCR) was conducted using KOD Hotstart Polymerase (Novagen) in either 25 µl (2.5 µl KOD Buffer, 2.5 µl 2 mM dNTPs, 1 µl 25 mM MgSO₄, 8.3 µl DDW + Template, 8.2 µl 3x Sucrose/Cresol Solution, 1 µl 15 µM Primer 1, 1 µl 15 µM Primer 2, 0.5 µl KOD Polymerase) or 15 µl reactions (1.5 µl KOD Buffer, 1.5 µl 2 mM dNTPs, 0.6 µl 25 mM MgSO₄, 4.9 µl DDW + Template,

5 µl 3x Sucrose/Cresol Solution, 0.6 µl 15 µM Primer 1, 0.6 µl 15 µM Primer 2, 0.3 µl KOD Polymerase) using a Peltier Thermal Cycler. If stated, the Expand 20kb PLUS PCR System (Roche) was used to amplify long PCR products, according to the manufacturers instructions, with the MEXP62 PCR programme (see section 2.2.4.2).

2.2.4.1 Primers

Primers were designed using the Primer3 programme (<http://frodo.wi.mit.edu/>) unless otherwise stated.

Number	Sequence	Pre-designed?
1	TACGGATCTGGAAGTCAAG	
2	ATTCAGGCTGCGCAACTGTTGGG	
3	CTCAAGGTCCGCCCTGTTTA	
4	AGTATCGGCCTCAGGAAGATCG	
5	AGGCACTCATGGAGGATCTG	
6	GTCACCCATTCCATGGTTTC	
7	GAGCTGGATGGAGCTGTAGG	
8	CGCCACCTTCAAAGTTGTT	
9	CATTTGGGGACTCCTTGATG	
10	GTGACCTGGCCAGTAGACCA	
11	AATTGGCGCGCCCCTGGAGTAGGCATGTTGATTTAC	
12	AATTGGCGCGCCATGCTGAGACAAGACTGGAAACCAC	
13	GAAGGTCTCTGTGCTCCCAAG	
14	CAAGATCTGAATCTGGGTGGTG	
15	AGAGAAGTGTGGTGCAGGTG	
16	GGGTGAGATCAAGGCTTCC	
17	GCGGATAACAATTTACACAGGA	pUCR
18	TGTAACGACGGCCAGT	M13 Forward 1
19	GGAGAAAGGCGGACAGGTAT	
20	CCTTGAAGGACTCCAATAGGG	
21	GCTGCAGGAGTAAGGACAGG	
22	GTGGATGAAGAGGCCTTGAA	
23	TTCTTTGGTTTTCGGGACC	
24	GTTTTCGGGACCTGGGAC	
25	GCTGGGCTGTTGTCTCCATA	
26	CATCTTGGGTTTCTTCCGAGT	
27	CGACGACCCATTCAACTTCT	
28	TCGAGCACTGCATACTCCAC	
29	GTAACGACGGCCAGT	M13 Forward 2
30	GGAAACAGCTATGACCATG	M13 Reverse

Table 2.1: Primers used in during the course of this study

2.2.4.2 PCR programmes

MKOD6030

1. 94°C – 2 minutes
2. 94°C – 15 seconds
3. 60°C – 30 seconds
4. 72°C – 1 minute

Repeat steps 2-4 30x

5. 72°C – 5 minutes

MEXP62

1. 92°C – 2 minutes
2. 92°C – 10 seconds
3. 62°C – 30 seconds
4. 68°C – 10 minutes

Repeat steps 2-4 10x

5. 92°C – 10 seconds
6. 62°C – 30 seconds
7. 68°C – 10 minutes and 10 seconds + 10 seconds/cycle

Repeat steps 5-7 24x

8. 68°C – 7 minutes

2.2.5 Colony PCR

A bacterial colony or ice scrape was picked into 100 µl DDW. 1µl of this solution was then used as the PCR template. A 15 µl or 25 µl KOD Hot-start (Novagen) PCR

reaction was then conducted with the MKOD6030 PCR programme. Products were subsequently size separated by gel electrophoresis (see section 2.2.6).

2.2.6 Gel electrophoresis

Agarose gel electrophoresis was used to size separate nucleic acids. 0.6%–2% agarose gels were prepared with 1x TBE containing ethidium bromide (400 ng/ml). Samples were loaded in 1x sucrose / cresol solution. Generally samples were loaded alongside a 1kb size marker (Invitrogen #15615-024). Mini-gels (50 ml) and larger gels (250 ml) were run at a voltage and for a period that would allow suitable size separation of nucleic acids. Nucleic acids were visualised with UV light on a transilluminator (UVP) and digital images were captured with a UVP system.

2.2.7 Bacterial culture

MACH1 *E. coli* (Invitrogen) cells, made competent by the rubidium chloride method, were kindly provided by James Grinham. Bacteria were cultured in Luria-Bertani (LB) broth or on LB agar plates supplemented with appropriate antibiotics (see below) at 37°C. Liquid cultures were incubated in a shaking incubator (37°C, 300 rpm). Blue/white selective plates were made by supplementing 500 ml of LB agar with 40 ng Xgal dissolved in 800 µl Dimethyl formamide and 2.5 ml 0.1 M Isopropyl β-D-1-thiogalactopyranoside (IPTG). Bacteria were stored frozen at -70°C in LB supplemented with 7.5% glycerol.

2.2.7.1 Antibiotic concentrations

Ampicillin: Final concentration of 100 µg/ml

Chloramphenicol: Final concentration of 30 µg/ml

Kanamycin: Final concentration of 50 µg/ml

2.2.7.2 Bacterial transformation

0.5 µl of plasmid was added to a frozen Thermowell 96-well plate (Costar), rested in a Stratacooler cooling block (Stratagene), and allowed to cool. Cells were defrosted on ice and 10 µl of MACH1 cells (Invitrogen) was added to each well containing plasmid. The plate and block were incubated in a Stratacooler (Stratagene) for 20 minutes. The cells were subsequently heat shocked for 40 seconds at 42°C (Peltier thermal cycler) and returned to the ice box for a further 2 minutes. Finally the cells were removed from the ice box and 90 µl of LB was added to each well. The cells were placed in a 37°C incubator for 1.5 hours and then the total volume was plated to warmed agar plates.

2.2.8 Sequencing

PCR fragments purified with the Qiagen Qiaquick Gel Extraction kit were diluted 1/5 -1/10 in Double distilled water (DDW) and 1µl of this template was added to 1µl of the appropriate primer (5µM) and 5µl of DDW. For plasmid sequencing 200 ng of plasmids purified by Qiagen Qiaprep Spin Miniprep were added to ~20 ng of the appropriate primer and the volume of the mix was adjusted to 7µl with DDW. These reaction mixtures were passed to the Sanger Core Sequencing Facility for sequencing. Gap4 (Staden Sequence Analysis Package software) was used to compile the sequencing results and join matching sequences. Basic Logical Alignment Search Tool (BLASTN) was used to compare consensus sequence against the mouse genome in Ensembl (<http://www.ensembl.org/>) or the Gap4 database was used to compare the sequences against an expected sequence based on Ensembl annotation

(ENSMUST0000009321, ENSMUST00000115633 (*Dgcr8*), ENSMUST00000023292 (*Arsa*)) and plasmid sequences.

2.2.9 Restriction digest

Restriction enzymes were supplied by NEB. Buffers were diluted as appropriate for the enzyme in question and bovine serum albumin (BSA) was added as applicable. Reaction volume varied as noted in individual protocols. On occasion 1/3 vol. of Sucrose/Cresol Solution was added to each reaction. Unless otherwise stated, 3 μ l of restriction enzyme was added to each reaction, the reaction was incubated for 1 hour at 37°C and a further 3 μ l of enzyme was added followed by a further incubation. Reactions were heat inactivated at 65-70°C for 20 minutes and stored frozen. Digested plasmid was size separated by gel electrophoresis alongside non-digested plasmid to ensure the digest had reached completion (section 2.2.6). The digest was repeated if necessary.

2.3 Generating a targeting trap vector

2.3.1 Long PCR

129Sv strain derived bacterial artificial chromosomes (BACs) (Adams et al., 2005) within the *Dgcr8* region (judged by an Ensembl mapped distributed annotation system (DAS) trace (v33, m34)) were selected. Where the sequences were obtainable from the sequence repository (nfs/repository/p305/MSE-WG-58230), BLASTN was used to remap the end sequences against the Ensembl database to check their annotated location. BACs were obtained from the Sanger Clone Resource Centre. The BAC plasmid was purified using the Qiagen Quiaprep Spin Miniprep. ~800ng of BAC PCR templates were digested by *NotI* (NEB) in a 200 μ l reaction volume (see section

2.2.9). RestrictionMapper 3 (<http://www.restrictionmapper.org/>) was used in order to ensure that the fragment required for amplification was not digested by this restriction enzyme. Oligonucleotide primers for long PCR were designed with an in house programme (courtesy of David Beare); *longPCR*. Prior to primer design mouse repeat sequences from the region upon which the primers were based were repeat masked (RepeatMasker v.3.0.8). The designed primers were to amplify a region ~5-6 kb in length, upstream of the *Dgcr8* gene trap (see section 3.3.5 for a full description of primer design). An *AscI* site and a short, 4-base spacer were added to the 5' the end of each primer. The Expand 20 kb PLUS PCR System (Roche) was used for Long PCR with the MEXP62 PCR programme (see section 2.2.4). The PCR products were size separated by gel electrophoresis (see section 2.2.6) on a 0.7% agarose gel and purified with the Qiagen Qiaquick Gel Extraction kit. 2.3 µg of the fragment was then digested with the *AscI* restriction enzyme (NEB) in a 200 µl reaction (see section 2.2.9).

2.3.2 Preparing the pR3R4AsiSI plasmid

DB3.1 *E. coli* containing the pR3R4AsiSI plasmid (generously donated by the Skarnes laboratory) were streaked onto chloramphenicol and ampicillin selective LB agar plates and cultured overnight at 37°C. A starter culture was set up in LB with ampicillin (8 hours, 37°C, 300 rpm). This starter culture was then diluted 1/500 for overnight culture (37°C, 300 rpm), again in LB with ampicillin. The plasmid was subsequently prepared by Qiagen HiSpeed Plasmid Maxikit and ethanol precipitated to increase the concentration. Then 5 µg of the plasmid was digested with an *AscI* (NEB) in a 200 µl reaction (see section 2.2.9). The digest was separated by 0.6% agarose gel electrophoresis (see section 2.2.6) and purified with a Qiagen Qiaquick

Gel Extraction kit. ~1 µg of the restricted plasmid was treated with Antarctic Phosphatase (NEB) in a volume of 50.5 µl, according to the manufacturers instructions.

2.3.3 Cloning the *Dgcr8* fragment

The *Dgcr8* fragment was cloned into the prepared pR3R4AsiSI backbone using the Roche Rapid Ligation Kit according to the manufacturer's recommendations. The ligated plasmid was then transformed into MACH1 cells (Invitrogen) (see section 2.2.7.2). In order to judge which bacterial colonies contained the correctly inserted fragment, colony PCR (see section 2.2.5) was conducted with primers 13, 14, 15, 16, 17 and 18 (see section 2.2.4.1), paired as appropriate (See Fig.3.4). The plasmid was purified with a Qiagen Qiaprep spin miniprep kit. The plasmid was sequenced by the Sanger Institute Core Sequencing Facility with primers 14, 15, 17 and 18 (see section 2.2.4.1).

2.3.4 Preparing the targeted trap vector

The *Dgcr8* fragment was transferred from the pR3R4AsiSI plasmid to the pL3/L4_(+)_GT1T2hygroP2EGFP plasmid (Generously donated by the Skarnes laboratory) with Gateway L/R Clonase II (Invitrogen) according to the manufacturers protocol. The pL3/L4_(+)_GT1T2hygroP2EGFP plasmid was then transformed into MACH1 cells (Invitrogen) (see section 2.2.7.2). Colony PCR (see section 2.2.5) was used to identify colonies containing the correctly inserted fragment, with primers 14, 17, 15 and 18 (see section 2.2.4.1). Colonies were cultured in LB with kanamycin and frozen.

A starter culture was set up in LB with kanamycin (8 hours, 37°C, 300 rpm). The starter culture was subsequently diluted 1/500 in fresh LB with kanamycin and cultured overnight. The plasmid was purified by HiSpeed Plasmid Maxi Kit (Qiagen) according to the manufacturers protocol, using 1.5x the usual amount of P1, P2 and P3 solutions to increase prep yield. The plasmid was phenol chloroform purified (see section 2.2.1) and resuspended in 200 µl T_{0.1}E.

50 µg of the plasmid was digested with *HindIII* (NEB) in a 500 µl reaction (No Sucrose/Cresol solution, 15 µl of enzyme introduced over 3 hours at 37°C). 0.2 µl of each reaction was size separated on a 0.7% agarose gel by electrophoresis (see section 2.2.6) alongside non-digested plasmid to ensure the digest had proceeded to completion. Finally the restricted plasmid was ethanol precipitated (see section 2.2.2), but the precipitation itself preceded for 15 minutes on ice prior to centrifugation. In addition, following the initial 70% ethanol wash a further 500 µl of 70% ethanol was added to the plasmid. In a tissue culture hood the ethanol was removed and the plasmid was air dried for 25 minutes. The pellet was resuspended at room temperature, overnight, in DPBS (-CaCl₂ and MgCl₂) (Gibco).

2.4 ES cell culture

Cell lines were maintained in ES Cell culture medium at 37°C, 7% CO₂ in gelatinized tissue culture treated plates and flasks (48, 24, 12 and 6 well plates (Falcon), T25 and T75 flasks (Corning)). Plates were gelatinized by adding 0.1% gelatin to the well prior to plating and aspirating away the excess. Medium was supplemented with selective agents appropriate to the cellular genotype (See below for specific culture conditions (see section 2.4.2)). Medium was replaced every day, unless otherwise

stated in the text. Cell images were taken with an Olympus IX51 microscope and Olympus DPSOft software at 10x relief contrast. Spinning of cells was conducted in an Eppendorf centrifuge 5702.

Well Size	Media (ml)
48	0.6
24	1.5
12	3
6	6
T25	10
T75	30

Table 2.2: Media quantities used to culture ES cells

2.4.1 Splitting cells

2.4.1.1 Method 1

Cells were washed once with DPBS (-CaCl₂ and MgCl₂) (Gibco). 1x Trypsin was then added to the well. Cells were incubated for 2 minutes at 37°C. Trypsin was diluted 1:10 with ES cell culture medium. Cells were spun out of the medium (1,200 rpm, 3 minutes) to remove trypsin and the cell pellet was resuspended in fresh ES cell culture medium. The appropriate proportion of the cells was transferred to a fresh well or flask. The well/flask was topped up with an appropriate quantity of growth medium and the well/flask was mixed by pipetting.

2.4.1.2 Method 2

As Method 1, however, cells were incubated in trypsin for 3 minutes. In addition once the trypsin was diluted with fresh media the cells were transferred to fresh plates directly without spinning.

2.4.2 Growth conditions for specific cell lines

E14 mouse ES cells were used as a wild type control cell line ($Dgcr8^{+/+}$) as this is the parent cell line for BayGenomics gene trap cell lines used as the basis of this study. These cells were split every 2-3 days as they approached confluence. The normal split ratios were 1/8-1/10 and depended upon the condition and confluence of the cells.

Heterozygous gene trap cell lines from BayGenomics were generally maintained in media supplemented with 150 $\mu\text{g}/\text{ml}$ G418 (Geneticin – Gibco). These cells were split every 2-3 days as they approached confluence. The normal split ratios were 1/8-1/10. These culture conditions were maintained upon the removal of selection prior to RNA lysis. Prior to electroporation these cells were split by Method 2 (see section 2.4.1.2). Due to a labeling problem in transit from the supplier (the labels fell off when thawing the vials) it is not possible to distinguish the XH157 and XG058 cell lines and be sure of which is which without a specific PCR or cloning strategy. This is because the gene traps in these cell lines are within the same intron. Hence from this point the cells have been arbitrarily ascribed the names $Dgcr8^{gt1/+}$ and $Dgcr8^{gt2/+}$. Subsequently these cell lines have been demonstrated to behave very similarly (Chapter 4 and 5) and as the identity of the cell lines appears to have limited biological relevance it was decided that these experiments would prove time consuming and of little significance.

The $Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$ cell lines were maintained in media supplemented with 100 $\mu\text{g}/\text{ml}$ hygromycin B (Calbiochem). These cells were split every 2-3 days as they approached confluence. The normal split ratios were 1/3.5-1/5. Upon the removal of selection for experiments the split ratio varied between 1/3.5-1/6.

Dgcr8^{gt1/tm1} and *Dgcr8^{gt2/tm1}* cells lines were maintained in media supplemented with 100 µg/ml hygromycin B (Calbiochem). These cells were split every 2-3 days as they approached confluence. The normal split ratio was 1/3.5. Upon the removal of selection for experiments the split ratio varied between 1/3.5-1/6.

Dgcr8^{gt/tm1} and *Dgcr8^{tm1,gt1/+}* cells generally varied between passages 13 and 26 when used for experiments. The cells behaved consistently throughout.

2.4.3 Freezing ES cells

Cells were trypsinised as if to be split and the trypsination was stopped with a 1:10 dilution of ES cell culture medium (see section 2.4.1). The cells were spun out of this medium (1,200 rpm, 3 minutes) and resuspended in an appropriate volume of freeze medium (500 µl per vial; culture medium supplemented with 10% Dimethyl sulphoxide (DMSO) and filter sterilised). Cells were transferred to Nunc CryoTube vials and placed at -70°C. Subsequently the cells were moved into liquid nitrogen storage.

2.4.4 Defrosting cells

In the evening, cells were removed from liquid nitrogen and warmed in the hand until defrosted. Cells were immediately transferred drop by drop to 10 ml of non-selective culture medium. Cells were swirled to mix and spun at 1,200 rpm for 3 minutes to pellet. The size of the pellet was used to judge the appropriate size of well into which the cells should be plated (See table 2.3). The pellet was subsequently resuspended in the appropriate quantity of non-selective culture medium and transferred to a

gelatinized plate. Early next morning the medium was replaced with the selective medium that corresponds to the cells' genotype.

Pellet size (mm) <i>Dgcr8^{+/+}</i>	Pellet size (mm) <i>Dgcr8^{gt/tm1}</i>	Well Size	Media for defrost
Barely visible	1mm	48	1ml
1mm	2mm	24	2.4ml
2mm	3mm	12	4ml
3mm	NA	6	8ml

Table 2.3: Culture plates and media quantity used for defrosting ES cells

2.4.5 Subcloning

Cells were trypsinised for 3 minutes. Trypsination was stopped by the addition of media. Cells were spun out of media and resuspended in fresh non-selective media. Cells were mixed gently to ensure they reached a single cell suspension. 10^4 cells were plated to 100 mm tissue culture treated plates (Corning) with 10 mls of 150 $\mu\text{g/ml}$ G418 selective media (Geneticin – Gibco). Plates were fed every other day with 10-12 ml of fresh selective media. 10-11 days later 12 colonies of each cell line were picked (see section 2.4.6).

2.4.6 Colony picking

Culture medium was aspirated from the 100 mm (Corning) plate and replaced with 10 ml DPBS (-CaCl₂ and MgCl₂) (Gibco). Colonies were transferred with a pipette to a gelatinized 48 well plate in 50 μl DPBS (-CaCl₂ and MgCl₂). 50 μl 2x Trypsin was added to each well and incubated at 37°C for 5 minutes. 1ml of 150 $\mu\text{g/ml}$ G418 selective medium (Geneticin – Gibco) was added to the wells and they were mixed gently to disperse the cells. Subsequently cells were maintained under 150 $\mu\text{g/ml}$

G418 selective medium. Ultimately, 2 vials of each cell line were frozen (see section 2.4.3) and RNA was purified from each cell line by the Promega SV Total RNA Isolation System (see section 2.7.3) for RT-PCR to check the cell line genotype.

2.4.7 Gene targeting/electroporation

Cells were seeded from T25 to T75 and cultured for 2 days, until approaching confluence. The T75 was trypsinised for 4 minutes and subsequently the cells were collected in 10mls of media. Cells were spun down for 3 minutes at 1,200 rpm. Cells were resuspended in DPBS (-CaCl₂ and MgCl₂) (Gibco), counted and spun once more. Cells (normally $2-5 \times 10^7$) were resuspended in 700 μ l of room temperature PBS. Cells were transferred to a microcentrifuge tube containing the DNA to be transfected (~ 50 μ g assuming 100% recovery from plasmid digest), pipetted up and down once to mix and transferred to a .4mm electroporation cuvette (BioRad). Cells were immediately electroporated with a Gene Pulser II (800 V, 3.0 μ F with an approximate time constant of 0.04 ms). Cells were left at room temperature for 20 minutes to recover, diluted in non-selective medium and plated onto 3x 100 mm gelatinized plates (Corning) (5×10^6 , 2.5×10^6 and 1×10^6 cells). The following day, the media was replaced with 10 ml of media containing 120 μ g/ml hygromycin B. The media was replaced daily for 2 weeks. After 9 days selection was dropped to 100 μ g/ml hygromycin B to aid cell growth. From 11 to 15 days after plating, colonies of a suitable size were picked to 48 well plates (see section 2.4.6 substituting G418 selection for hygromycin selection). Subsequently cells were maintained in 100 μ g/ml hygromycin B. RNA was purified from an over confluent 24-well-plate well of each cell line by SV RNA Purification (see section 2.7.3) for genotyping by RT-PCR. Ultimately 70%-90% confluent 48 well plates containing cultured colonies were

frozen by replacing the growth media with freeze media and storing them at -70°C until genotyping was complete. To defrost, cells were removed from the -70°C freezer and as soon as the freeze media melted it was replaced with $>1\text{ml}$ of non-selective media. The next day this was replaced with selective, $100\ \mu\text{g/ml}$ hygromycin B growth media.

2.5 Judging ES cell characteristics

2.5.1 Xgal staining

Cells were washed once with PBS and then Fix Buffer was added. Plates were incubated for 30 minutes at room temperature. Each well was then washed twice with wash buffer. Stain Buffer was then added to the cells and they were incubated overnight at 37°C in a sealed Tupperware box. Subsequently Stain Buffer was removed from the cells, the cells were washed once with wash buffer and each well was stored under Fix Buffer. Cells were imaged at a magnification of 10x with relief contrast.

2.5.2 Immuno-staining

2.5.2.1 Preparing the slides

Prior to plating to slides, all cells were maintained for 2 days in non-selective media. On the second day cells were plated to gelatinized chamber slides (Nunc 8 well slides, VWR 62407-335). 2.1×10^4 *Dgcr8^{tm1.gt1/+}* cells and 4.2×10^4 *Dgcr8^{gt1/tm1}* cells were plated in $300\ \mu\text{l}$ of non-selective media per well. Approximately 24 hours later the media was removed and replaced with $300\ \mu\text{l}$ of 4% paraformaldehyde solution. The slides were then incubated for 20 minutes at room temperature. The

paraformaldehyde solution was then replaced with 400 μ l of DPBS (-CaCl₂ and MgCl₂) (Gibco) and the slides were incubated at room temperature for a further 5 minutes. Finally the DPBS was replaced with a further 400 μ l of DPBS and the slides were wrapped in parafilm and stored at 4°C.

2.5.2.2 Immuno-staining the slides

Each well was washed 3x with PBS. After the final wash, 100 μ l of 5 mM NH₄Cl in PBS was added to each well and incubated for 10 minutes at room temperature. This solution was replaced with 0.2% Triton in PBS and incubated for 10 minutes at room temperature. The wells were washed a further 3x with PBS. Following the third wash the cells of each well were blocked with 100 μ l of 5% donkey serum in PBS and incubated for 20 minutes at room temperature. Next the primary antibodies were added to the cells, diluted in 100 μ l 5% donkey serum (Anti-Oct4 (Goat) Santa Cruz – sc8628 at 1/50 OR Anti-SOX2 (Rabbit) Chemicon International – AB5603 at 1/500). Slides were incubated for 1 hour at room temperature in a tip box with a damp cloth to maintain humidity. After an hour each well was washed 4x with PBS and blocked for a further 20 minutes with 5% donkey serum. Next the secondary antibodies were diluted in 5% donkey serum and 100 μ l of the mixture was added to the appropriate wells for 1hr at room temperature (Alexa Fluor 594 Donkey anti-goat – Invitrogen – 1/500 OR Alexa Fluor 488 Donkey anti-rabbit – Invitrogen – 1/500). Subsequently cells were washed a further 4x with PBS. Finally the chambers were removed and the slides were dipped in RO water to rinse. A drop of Vectorshield with DAPI (Vector Laboratories) was added to each well. A cover slip was added and sealed to the slide with nail varnish. In parallel to the addition of primary and secondary antibodies to

the same wells, secondary antibodies were also applied to further wells with no primary antibody to allow an assessment of non-specific staining.

2.5.2.3 Imaging immuno-stained slides

A Leica TCS SP5 confocal microscope was used to take images of the slides at 20x magnification with oil. Oct4 images were taken with the following settings: Sequential setting 1 - Laser Line UV (405): 64%, PMT1 - ~412-489nm, Sequential setting 2 – Laser line Visible (594): 40%, PMT3 - ~608-690nm. SOX2 images were taken with the following settings: Sequential setting 1 - Laser Line UV (405): 100%, PMT1 - ~412-489nm, Sequential setting 2 – Laser line Visible (488): 75%, PMT2 - ~492-566nm. Laser power, Gain and Offset levels for each PMT were set to obtain the optimal image with minimum background. All settings were maintained for the images of both cell lines taken with relation to the same primary antibodies. A Z-series was taken of cells in each well (10 slices through 488 or 594 fluorescence visible in test wells and 5 slices through any 488 or 594 fluorescence visible in control wells). Final images were derived by a “maximum” projection of the Z-series and these were subsequently overlaid.

2.5.3 Optimised embryoid body (EB) formation protocol

Dgcr8^{tm1,gt1/+} and *Dgcr8^{gt1/tm1}* cells were cultured for 4 days in non-selective media prior to EB formation. A T25 flask of cells was trypsinised for 3 minutes. This reaction was stopped with the addition of media without LIF (-LIF). The cells were spun down (1,200 rpm for 3 minutes) and resuspended in 10 ml -LIF media. Cells were counted and diluted to 10^4 cells per ml for *Dgcr8^{+/+}* and *Dgcr8^{tm1,gt1/+}* lines and 4×10^4 cells for *Dgcr8^{gt1/tm1}* cells. Cells were subsequently plated to a non-tissue culture

treated, round-bottomed, 96 well plate; 100µl per well. The following day a further 100 µl of -LIF media was added. On day 3, EBs were transferred to 24 well low attachment plates (Costar – 3473) and fed with 1.5 ml -LIF media. From day 4 onwards EBs were fed every other day. On days 5-8 half of the EBs were cultured with retinoic acid (RA) (1 µM). On day 8 the EBs were transferred to gelatinised tissue culture treated 12 well plates (Falcon). On day 12 the media was removed from the EBs and 0.8 ml 4% Paraformaldehyde Fixing Solution was added to the wells. EBs were fixed for 20 minutes at room temperature and then washed twice with PBS before being stored at 4°C in PBS.

2.5.4 Cell cycle assay

2.5.4.1 Cell growth

Dgcr8^{tm1.gt1/+} and *Dgcr8^{gt1/tm1}* cells were cultured for 2 days in non-selective media prior to plating for cell cycle analysis. Cells were plated to gelatinised T75 flasks (Cell numbers: 186 x 10⁴ *Dgcr8^{+/+}* cells, 246 x 10⁴ *Dgcr8^{tm1.gt1/+}* cells and 304 x 10⁴ or 380 x 10⁴ *Dgcr8^{gt1/tm1}* cells) in non-selective media. After 2 days the cells were 60-75% confluent. Cells were washed with DPBS (-CaCl₂ and MgCl₂) (Gibco) and trypsinised for 4 minutes with 3ml 1x Trypsin. The reaction was stopped with 17 ml non-selective media and the cell suspension was spun for 3 minutes at 1,200 rpm. Cells were washed in 20 ml DPBS (-CaCl₂ and MgCl₂) (Gibco) and respun. The pellet was subsequently resuspended in 1 ml DPBS (-CaCl₂ and MgCl₂) (Gibco) and 9ml of ice cold 70% ethanol was added dropwise while vortexing. Cells were stored in this ethanol at 4°C.

2.5.4.2 Fluorescence activated cell sorting (FACS)

Cell spinning for staining was conducted with a Thermo Electron Corp Heraeus Sepatech Megafuge 1.0. 5 mls of cell suspension was spun for 5 minutes at 1,200 rpm to pellet cells. Cells were washed once in 10 ml and once in 5 ml of PBS, spinning between washes to remove supernatant. Cells were then counted in the second wash and spun out of PBS. Cells were resuspended in 1 ml of propidium iodide (PI) solution per 2×10^6 cells. Cells were stained overnight and filtered to remove clumps (30 μ m mesh filter, Partec).

Stained samples were analysed with a Beckman Coulter (Cytomics FC-500) flow cytometer with a 20mW 488nm air-cooled argon laser. A 620nm band pass filter was used to collect the fluorescence emitted by the PI stained cells. 20,000 events were collected per sample.

2.5.4.3 FACS analysis

The cell cycle profile of each cell line was determined with *FlowJo* v7.2.5. Samples were gated to remove the signal from cell doublets and the Watson algorithm was used at default settings to determine the cellular profile.

2.6 Protein purification and Western blots

2.6.1 Protein purification

Trapped cells were maintained for 2 days in non-selective media. Subsequently all cells were plated to gelatinised tissue culture treated 100mm plates (Corning). 136×10^4 *Dgcr δ ^{+/+}* cells, 180×10^4 *Dgcr δ ^{tm1.gt/+}* cells and 223×10^4 *Dgcr δ ^{gt/tm1}* were seeded per plate. Cells were maintained for a further 2 days in non-selective media and then

lysed at 70-85% confluence. Media was aspirated from cells and they were washed with DPBS (-CaCl₂ and MgCl₂) (Gibco) (2x20 ml). 360 µl of Protein Lysis Buffer was added to the cells and the plates were placed on ice for 5 minutes. The base of the well was scraped with a pipette tip and the lysate was transferred to an Eppendorf tube. The lysate was passed repeatedly (16x) through a 21G hypodermic needle and then spun at 14,000 rpm for 10 minutes at 4°C. The lysate was then removed from the cellular debris and placed in a fresh tube. Lysate was snap frozen with dry ice and ethanol and stored at -70°C. Samples were diluted 1/50 and the protein was quantified with Bradfords Reagent, according to the manufacturers protocol.

2.6.2 Western blot

50ug of each protein sample was separated on 4-12% Bis-Tris Gels (Invitrogen) using an XCell SureLock Mini-Cell (Invitrogen). Protein samples were prepared as described by the manufacturer with the inclusion of 1 µl of NuPAGE Reducing Agent. Gels were run in MOPS SDS Running Buffer (Invitrogen) in parallel with 3 µl of MagicMark XP protein standard (Invitrogen). Proteins were transferred to Hybond-ECL filter (Amersham) in a Mini-Trans-Blot Cell (BioRad) in Western Transfer Buffer (10% Methanol, 1x NuPAGE Transfer Buffer (Invitrogen)) at 100 V for 1.25 hours.

Filters were blocked for 1 hr in Western Blocking Solution with gentle rocking. Subsequently the blocking solution was replaced with fresh solution and the primary antibodies were added. The anti-Oct4 (1/1000 Santa Cruz – sc8628) and the anti- α -tubulin antibody (1/5000 Abcam – ab7291) were added together. The filters were rocked gently for 1 hour. Each membrane was rinsed 3 times for 10 minutes in

Western Washing Solution with gentle rocking. Initially the peroxidase conjugated secondary antibody appropriate for the Oct4 primary antibody (1/7500 Sigma-Aldrich A4174) was added in Western Blocking Solution and the filters were rocked for 2 hours. The Blocking Solution was subsequently removed and the filters were rinsed 3x in Western Washing Solution for 10 minutes with rocking. 5mls of Western Lightning Enhanced Luminol Reagent and 5 ml Western Lightning Oxidising Reagent (Perkin Elmer) were added to the membranes and they were incubated for 1 minute in the dark with rocking. Excess fluid was blotted from the membranes with 3MM Whatman paper and the membranes were wrapped in Saran wrap and exposed to film for a variety of durations to optimize the image obtained. The filters were then washed 3 times with Western Washing Solution for 10 minutes with rocking and then Western Blocking Solution was added with a 1/7500 dilution of peroxidase conjugated secondary antibody (Sigma-Aldrich A6782, appropriate for binding the anti- α -tubulin primary antibody). Membranes were incubated for a further hour with this solution, rinsed 3 times and developed as described above.

2.7 RNA purification

2.7.1 Trizol

Cells were lysed with Trizol reagent (Invitrogen). Cells were washed twice with DPBS (-CaCl₂ and MgCl₂) (Gibco) and Trizol was added (7.5 ml per T75, 1 ml per 6 well). The plate was rocked gently. After the Trizol cleared it was homogenized by pipetting and transferred to a Falcon tube. Lysate was stored at -70°C. RNA purification essentially followed manufacturers recommendation. However, large volumes of Trizol chloroform mix (T75) were centrifuged at 4200 rpm (Beckman J6-MC) for 1 hour at 4°C in a 50 ml Falcon tube to separate the aqueous phase. The

aqueous phase was subsequently divided between 5 Eppendorfs tubes for ethanol precipitation of the RNA. Smaller volumes of Trizol:chloroform (6 well) were phase separated with Eppendorf phaselock gel tubes (Heavy Gel) spun for 15 minutes at $< 12,000g$, $4^{\circ}C$. RNA pellets were stored under 75% ethanol at $-20^{\circ}C$. Pellets were resuspended in RNase free water for applications. Once resuspended RNA was stored at $-70^{\circ}C$. RNA was generally quantified by Nanodrop (Although early quantification was performed by Eppendorf Bio photometer) and an aliquot was size separated by agarose gel electrophoresis (see section 2.2.6) to ensure its integrity before use.

2.7.2 DNase treatment

When mentioned in the text Trizol purified RNA was TURBO DNase treated according to the manufacturers protocol (Ambion). DNase treated RNA was cleaned with a Qiagen RNeasy MiniElute Cleanup Kit according to the manufacturers protocol.

2.7.3 SV purification

RNA was purified using the SV Total RNA isolation System (Promega) from a single, confluent 24-well-microtitre-plate well of ES cells. Cells were lysed with $300 \mu l$ SV RNA Lysis buffer and incubated at $70^{\circ}C$ for 3 minutes, followed by centrifugation at $12,000-14,000 g$ at room temperature for 10 minutes. $200 \mu l$ of 96% ethanol was added to the cleared lysate. The lysate was passed through a SV Total RNA Isolation System spin column. From this point on the manufacturers recommendations were followed. RNA was quantified by Eppendorf Bio photometer.

2.7.4 Harvesting HeLa S3 RNA

HeLa S3 cells were cultured in suspension in Hams F12 media supplemented with 10% Foetal Bovine Serum and 1x Penicillin-Streptomycin-Glutamine (Invitrogen). 3x T75 flasks containing a total of approximately 5.6×10^7 cells in 180 ml media were lysed for RNA with Trizol. Briefly, cells were separated into 6 50 ml falcons and spun at 1,200 rpm for 5 minutes to pellet. Each pellet was resuspended in 5 ml PBS, combined to a single 50 ml Falcon and spun again to pellet. The pellet was resuspended in 30 ml PBS to wash, spun and lysed with 7.5 ml Trizol. Lysate was stored at -70°C . Lysate was processed as explained in section 2.7.1.

2.8 RT-PCR

Roughly 300 ng of SV purified RNA (or 100ng of DNase treated Trizol purified RNA if stated in the text) was reverse transcribed with Superscript II (Invitrogen) according to the manufacturers instructions. Reverse transcription was primed by either random hexamer (Roche) or oligo dT primers (Invitrogen).

Nested PCR was conducted to amplify fragments from the complementary DNA (cDNA). This consisted of 2 rounds of 15 μl or 25 μl KOD Hot-Start PCR reactions (see section 2.2.4). The template for the first round was 1-2 μl of the previously described RT reaction primed by the external primer set. The second round template was 1-2 μl of a 1/100 to 1/500 dilution of the completed first round reaction and was primed by an internal primer set. The MKOD6030 PCR programme was used for each round (see section 2.2.4.2). 3 μl of the product of the second reaction was size separated by agarose gel electrophoresis (see section 2.2.6).

For sequencing the products of the complete second round reaction were size separated by agarose gel electrophoresis and the bands of the correct size were excised from the gel and purified using the Qiagen Qiaquick Gel Extraction Kit according to the manufacturers instructions. 4µl of the purified fragments were subjected to gel electrophoresis to check the specificity of the band purification (see section 2.2.6). Purified bands were sequenced by the Sanger Core Sequencing Facility (see section 2.2.8).

2.9 mRNA expression profiling

2.9.1 Northern blot

2.9.1.1 Northern blot preparation

Dgcr8^{tm1,gt1/+}, *Dgcr8^{tm1,gt2/+}*, *Dgcr8^{gt1/tm1}*, *Dgcr8^{gt2/tm1}*, *Dgcr8^{gt1/+}* and *Dgcr8^{gt2/+}* cell lines were cultured for 2 days in non-selective media and then all cells were plated to T75 flasks (Cell numbers: 186 x 10⁴ *Dgcr8^{+/+}*, *Dgcr8^{gt1/+}* and *Dgcr8^{gt2/+}* cells, 246 x 10⁴ *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* cells and 304 x 10⁴ *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cells). RNA was purified by Trizol (see section 2.7.1) and 100 µg of RNA was cleaned with an RNeasy Mini Kit according to the manufacturers protocol. The RLT and RPE washes were collected as these contain the small RNA fraction, suitable for Solexa sequencing. RNA was eluted in 50 µl of water. The eluate was subsequently reapplied to the same column and re-eluted to maximise RNA yield. The RNA in the eluate was quantitated with an Agilent Technologies 2100 Bioanalyser using a Eukaryotic Total RNA Nano Chip. Half of the eluate was then processed with the PolyAtract mRNA Isolation System III (Promega) to remove non-polyadenylated RNA, including ribosomal RNA (rRNA), from the samples. The concentration of the polyadenylation (polyA) selected RNA was quantified again by Bioanalyser, as

above, to ensure the removal of rRNA was successful. All subsequent centrifugation steps were conducted with a Thermo Electron Corp Heraeus Pico 21 centrifuge at maximum unless otherwise stated. The mRNA was ethanol precipitated (625 μ l 100% ethanol, 25 μ l 3 M Sodium Acetate, 3 μ l GlycoBlue (Ambion); overnight at -20°C; Spin 60 minutes, 4°C, 700 μ l 70% ethanol wash; resuspend in 10 μ l RNase free water). The original, pre-polyA selection Bioanalyser results were used to estimate approximately equal quantities to load on the Northern and the samples were made up to 10 μ l with RNase free water. The RNA was mixed with 2x RNA loading buffer (17% formaldehyde, 50% formamide, 1x MOPS, 5% glycerol, 0.05% bromophenol blue, 0.05% xylene cyanol), heated to 70°C for 10 minutes and cooled on ice. A gel was prepared (1% Agarose, 1x MOPS, 2% Formaldehyde, 1/50,000 SybrGreen (Invitrogen)) and pre-run for 30 minutes at 70 V, 4°C in 1x MOPS Buffer. The samples were loaded alongside a 0.5-10 kb ladder (Invitrogen). Samples were run for ~5 hours at 70-200 V, 4°C. Nucleic acids were visualised with UV light on a transilluminator and digital images were captured with a UVP system. The gel was rinsed in 10x SSC. The gel was transferred over night by capillary blot to a Hybond XL membrane (GE Healthcare) with 10x SSC and UV cross-linked (UVP).

2.9.1.2 Northern probes

Primers were designed using ENSMUST00000115633 and ENSMUST00000009321 as the template sequences. *Dgcr8*^{+/+} Trizol-purified-RNA derived cDNA was used as the template for probe amplification. Probes were amplified in 25 μ l KOD Hot-Start PCR (Novagen) reactions (see section 2.2.4) using primers 25 and 26 (3') or 27 and 28 (5') (see section 2.2.4.1), following the MKOD6030 programme. Amplified probes were purified by gel electrophoresis (see section 2.2.6) followed by Qiagen Qiaquick

Gel Extraction. The fragments were A-tailed (1 μ l NEB Buffer, 2 μ l dATP (Amersham), 1 μ l AmpliTaq (Perkin Elmer), ~250ng DNA (3-6 μ l), to 10 μ l with DDW. 70°C for 30 minutes), ligated into the pGEM-T-Easy vector (Promega) with the Roche Rapid Ligation Kit and transformed into MACH1 cells (Invitrogen) (see section 2.2.7.2). The successful transformant colonies were Blue/white selected on Xgal containing ampicillin selective plates (Cultured overnight at 37°C). Colony PCR was conducted with flanking plasmid primers 29 and 30 to check the fragment insertion (see section 2.2.6). Plasmids containing the correct inserts were prepared with a Qiaprep Spin Miniprep Kit (Qiagen). These plasmids were sequenced with primers 25, 26, 27, 28, 29 and 30 as applicable (see section 2.2.8). The probes were finally amplified by KOD Hot-Start PCR in 25 μ l reactions, with a 1/10,000 dilution of the correctly identified, purified plasmids as template and probe specific primers given above. The total PCR reaction was subjected to gel electrophoresis and the probes were purified by Qiaquick Gel Extraction (Qiagen).

2.9.1.3 Hybridisation

Probe labeling and hybridization were performed by Dr. Claudia Kutter. Probes were radiolabeled with [$a^{32}P$]-dCTPs using the Random Labeling kit (Invitrogen). The probe was subsequently purified with a G50 spin column. Pre-hybridisation and hybridisations were performed using 15 ml PerfectHybTMPlus buffer (Sigma) at 55°C in a rolling incubator. The 5' probe was added to the hybridization buffer and hybridized overnight at 55°C. Filters were washed once with Wash Buffer 1 (2x SSC, 0.1% SDS) for 5 minutes at room temperature. The filters were washed a further 2x in Wash Buffer 2 (0.1x SSC, 0.1% SDS) at 55°C for 10 minutes and 30 minutes respectively. The filters were wrapped and exposed to a phosphoimager screen (GE

Healthcare) for 2.5 hours. The filter was subsequently stripped twice with boiling 0.1% SDS for 20 minutes until only background radioactivity was detected. The 3' probe was hybridized as before and exposed to a phosphoimager overnight.

2.9.2 Illumina expression arrays

2.9.2.1 Expression array preparation

For standard expression array profiling, unless otherwise stated, *Dgcr8^{tm1,gt/+}* and *Dgcr8^{tm1,gt/+}* cells were cultured for 2 days in non-selective media and then all cells were plated to T75 flasks (Cell numbers: 186 x 10⁴ *Dgcr8^{+/+}* cells, 246 x 10⁴ *Dgcr8^{tm1,gt/+}* cells and 304 x 10⁴ *Dgcr8^{gt1/tm1}* cells). Following 2 days of further culture sub-confluent cells were lysed with Trizol (see section 2.7.1). For expression profiling post miRNA mimic reintroduction cells were cultured in 6 well plates as described in the miRNA transfection section (see section 2.11.2.2).

10µg of Trizol purified RNA was cleaned up with an RNeasy MiniElute Cleanup Kit (Qiagen) according to the manufacturers protocol. The RNA was then quantified by Nanodrop and 500 ng of RNA was amplified and labeled with the Illumina Total Prep RNA Amplification Kit (Ambion) according to the manufacturers protocol either by me or by the Sanger Institute Microarray Facility.

2.9.2.2 Expression arrays

Microarrays were processed by the Sanger Institute Microarray Facility. Briefly, 1500 ng of biotinylated cRNA was hybridised to Illumina Mouse-6 v1.1 Expression BeadChips overnight at 58°C. These chips were washed, detected and scanned,

following the manufacturer's instructions. The scanner output was imported into BeadStudio software v.3.1.8 (Illumina).

2.9.3 Computational analysis of expression arrays

Array analysis was conducted with advice and help from Dr. Cei Abreu-Goodger. The general analysis of the expression array data was conducted in R 2.8.1/Bioconductor (<http://www.bioconductor.org/>) with additional packages (*affy*, *lumi*, *limma*, *ape*, *lumiMouseAll.db*, *GOstats*, *GO.db*, *R2HTML*, *annotate*, *KEGG.db* and *org.Mm.eg.db*). The function *addNuID2lumi* was used to ascribe nuID information to probe IDs, via probe sequences, from the Illumina Mouse-6 v1.1 annotation file *Mouse-6_v1_1_sequence.csv*. All arrays for both cell line expression analyses and miRNA transfection experiments were VST transformed (Lin et al., 2008b) and quantile normalised together.

2.9.4 Relationship plots

Sample relationships were determined using a distance metric based on Spearman rank-correlations; (1-cor). A distance tree was then computed using the Saitou and Nei (1987) method. Unrooted trees were plotted using the *ape* package. The subset of probes selected for correlation calculations was refined using the detection call from the Illumina Beadstudio output, restricting probe choice to only those called as “present” with a *P*-value < 0.05 in more than five of the samples. For further refinement, the normalised expression arrays corresponding to the cell line profiling arrays grown with culture method 2 (Section 5.3.1) were considered independently from the other array sets. A linear model was constructed including the arrays from all 5 cell lines (*Dgcr8^{tm1,gt1/+}*, *Dgcr8^{tm1,gt2/+}*, *Dgcr8^{gt1/tm1}*, *Dgcr8^{gt2/tm1}* and *Dgcr8^{+/+}*). The

limma package was used to determine the genes with significant change in expression when comparing the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells to the *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cell lines (P -value = 0.05, Log fold change = $\log_2(1.2)$). Subsequently “detected” and “significant” probes were used to recalculate the correlations and plot the relationships between all samples.

2.9.5 Identification of expression changes between cell lines

In order to compile gene lists for Sylamer analysis, four sets of comparisons were performed. First of all the normalised expression arrays from each growth method (see section 5.3.1) were considered separately using the *limma* package to generate probe lists ordered by log fold change and t-statistic as a result of comparisons between cells of differing genotypes. Similar gene lists were compiled using *limma* when all cellular expression arrays were considered together irrespective of the culture conditions both in the analysis and in the construction of the linear model upon which the analysis is based. In this case the method by which the cells were grown was provided to the linear model as an additional factor. The miRNA transfection experiment arrays were excluded from these analyses. The final comparisons were made between the post miRNA transfection expression arrays. The linear model for this comparison was constructed using the normalised array data from only the miRNA transfection experiments. Sylamer analyses (van Dongen et al., 2008) were performed by Dr. Cei Abreu-Goodger (seeds were defined as bases 7mer-1A and 7mer-m8. 740 mouse miRNA 7mers were compiled from miRBase version 12. P -values were Markov corrected based on words of length 4). In addition to ordered gene lists *limma* was also used to identify significant expression changes

between samples using the linear models constructed above, see text for the details of these comparisons.

2.9.6 KEGG and GO analysis

For gene ontology (GO) analysis probes with an inter-sample interquartile range (IQR) less than the median IQR for all probes were removed from the analysis. Next probes without Entrez IDs annotated in the *lumiMouseAll* library were also removed along with probes without a *lumiMouseAll* annotated GO term. Only a single probe, with the greatest IQR, was used for the analysis from sets of probes which target the same gene. Finally probe IDs were converted into Entrez IDs using the *lumiMouseAll* library. This gene set was used as the “gene universe” for the comparison. Lists of probes with significantly altered expression were filtered in the same way. Finally the *GOstats* package was used to calculate a *conditional* hypergeometric P-value for GO term enrichment (The *conditional* parameter calls on the algorithm to estimate whether a term is significantly enriched beyond the evidence contributed by a terms’ child terms within the GO tree). A P-value cut off of 0.01 was applied for the analysis. The categories were then filtered to leave only those containing 10 or more gene IDs. A similar process was applied for Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. However, probes were filtered for those with annotated KEGG terms rather than GO terms and the Category size was not limited. In addition the KEGG terms for the analysis were derived from the *org.Mm.eg.db* library.

2.9.7 Wilcoxon Rank Test of the expression changes of the targets of transcription factors (TFs) when compared to the general expression changes in DGCR8 depleted cell lines

The TF target predictions were downloaded from the supplementary material of Kim *et al.* and Chen *et al.* (Chen *et al.*, 2008; Kim *et al.*, 2008b). Targets proposed by Chen *et al.* with an “association score” greater than 0 were considered as potential TF targets. For each TF these targets were intersected with those of Kim *et al.* and targets predicted by both sets were used. The Illumina array probes were ordered according to log fold change (LFC) in a comparison between *Dgcr8^{tm1.gt/+}* and *Dgcr8^{gt/tm1}* cell lines. Probes were converted to gene symbols with the *lumiMouseAll* library. Probes with no symbol annotation were removed. For multiple probes annotated to the same gene, the mean LFC was used and the duplicate was removed. Genes present in the TF lists but not present on the array were removed. For each TF their targets’ associated LFC values were compared to the LFC values of the remainder of the genes on the array with a Wilcoxon Rank Sum test.

2.9.8 Annotation of probes associated with potential miRNA targets

The log fold change and *P*-value scores in the relevant microarray experiments were included (Sections 5.3.2.1 and 5.3.3.5). Gene symbol, gene name and Entrez ID were annotated using the *lumiMouseAll* annotation library. In order to build a more comprehensive annotation file, if no associated Entrez ID or gene name was available in this library an attempt was made to fill these gaps with an annotation file constructed by Dr. Cei Abreu-Goodger, directly from the Ensembl database (release 51). This file was also used as a source of Ensembl annotated transcripts mapped to each probe to be used for Sylamer seed counts. Vega transcripts

(<http://vega.sanger.ac.uk/index.html>) were selected in preference, if these were not present Ensembl transcripts (<http://www.ensembl.org/index.html>) were used and, if these were unavailable, EST (Expressed Sequence Tag) based transcripts were selected. If multiple transcripts were available within the same class, the transcript with the longest 3'UTR was used. The Sylamer programme was used to screen these UTRs for miRNA associated seed sequences (Dr. Stijn van Dongen) (van Dongen et al., 2008). Probe associated GO terms were identified with the *lumiMouseAll* database. These were filtered for GO terms from the “Biological process” ontology using the *annotate* package and these IDs were included in the probe annotation table. GO descriptions associated with these IDs were derived from the *GO.db* library. Probe associated KEGG pathway IDs were again derived from the *lumiMouseAll* library and the associated pathway names obtained from the *KEGG.db*. TargetScan version 4.2 (Grimson et al., 2007; Lewis et al., 2005) predictions were identified through the EntrezIDs associated with each probe. To do this mouse EntrezIDs were mapped to Human EntrezIDs using NCBI homologene annotation (<http://www.ncbi.nlm.nih.gov/homologene>). If multiple TargetScan predictions for the same gene were available (http://www.targetscan.org/vert_42/), the one with the best context score was recorded. miRBase v5 targets (<http://microrna.sanger.ac.uk/targets/v5/>) were also queried using the probe associated Entrez ID, where available. These were mapped to Ensembl transcripts using miRBase annotation, by Dr. Stijn van Dongen and subsequently used to identify transcript-associated targets and their “P-orthologous group” values. The smallest “P-orthologous group” value was included in the case of multiple predictions for the same gene. Finally, where multiple probes were associated with the same gene, the annotation was amalgamated to a single entry. In these cases the minimum *P*-values

and maximum log fold changes for associated array expression changes were included.

2.10 miRNA expression profiling

2.10.1 miRNA Northern blots

2.10.1.1 Preparing the blots

Unless otherwise stated cells were grown and RNA was purified as described in the section describing expression array preparation (see section 2.9.2.1). RNA samples were size separated by gel electrophoresis using 15% TBE-Urea gels (Invitrogen), in a XCell *SureLock* Mini-Cell, with Novex TBE-Urea Sample Buffer and Novex TBE Running Buffer (Invitrogen) according to the manufacturers protocol. 12.7 µg to 20 µg of total RNA was loaded per well. The same quantity of RNA was loaded in each well of a gel unless otherwise stated. RNA was run in parallel with a DNA oligo ladder (see Table 2.4). All buffers were prepared with HPLC grade water (BDH). Bottles and cylinders were rinsed with 0.1% SDS and 100 mM NaOH before use. Post-electrophoresis, the ladder was sliced from the gel and post-stained with ethidium bromide (Approximately 7 µl of ethidium bromide in 50 ml TBE) and visualised on a UV transilluminator and images were recorded (UVP).

The RNA was transferred to a Genescreen Plus nylon membrane (Perkin Elmer) using a BioRad Semi-Dry Electrophoretic Transfer cell. Briefly, the gel was equilibrated for 20 minutes in 0.5x TBE. 4 extra thick blot paper pads (BioRad) and the transfer membrane were soaked in 0.5x TBE and then stacked with the gel in the order of 2 pads, membrane, gel and then a further 2 pads. RNA was transferred at approximately

25 V for 45 minutes. While moist the membrane was UV crosslinked with a stratalinker (Stratagene) with 1000 μ J total energy and baked for 1 hour at 80°C.

Ladder	Oligo Sequence
97nt	TGCTGTTGACAGTGAGCGCTGCCTTGATGTTATTCCAGAG TAGTGAAGCCACAGATGTACTCTGGAATAACATCAAGGCA TTGCCTACTGCCTCGGA
65nt	GGCCAAGCTTGCCACCATGGACTACAAAGACGATGACGA CAAGGATATCTTAAGCGGCCGCGGCC
37nt	AATTGGCGCGCCGAAGTGTGGGAGATGTGCAGAGT
25nt	GTTAGAAGATTACCAAGATGCAGTG
20nt	ACATCTGCCAATCCATCTCA
18nt	GTTTTCGGGACCTGGGAC

Table 2.4: The sequences of DNA oligos used to make a size ladder for small RNA gel electrophoresis

2.10.1.2 Probe labelling

The probes used were DNA oligos complementary to the miRBase annotated miRNA targets (see Table 2.5). 2 μ l of 10 μ M oligo was added to 2 μ l 10x T4 PNK kinase buffer (NEB), 2.5 μ l of 32 P- γ -ATP (Amersham AA0018), 12 μ l DDW and 1 μ l T4 PNK Kinase (NEB). The probe was incubated for 1 hour at 37°C and heat inactivated for 10 minutes at 68°C. The probe was purified with a G-25 MicroSpin column (GE Healthcare) according to the manufacturer's protocol.

2.10.1.3 miRNA Northern blot probes

Probe	Sequence	Published Source?
miR-16	CGCCAATATTTACGTGCTGCTA	
miR-21	TCAACATCAGTCTGATAAGCTA	
miR-92a	CAGGCCGGGACAAGTGCAATA	
miR-130a	ATGCCCTTTTAACATTGCACTG	
miR-292-3p	ACACTCAAACCTGGCGGCACTT	Murchison <i>et al.</i> 2005
miR-293	ACACTACAACTCTGCGGCACT	Murchison <i>et al.</i> 2005
miR-320	TTCGCCCTCTCAACCCTGCTTTT	
Let-7a	AACTATACAACCTACTACCTCA	
U6snRNA	GCTTCACGAATTTGCGTGTATCCT	Murchison <i>et al.</i> 2005

Table 2.5: miRNA Northern blot probe sequences

2.10.1.4 Hybridisation of miRNA Northern blots

50ml Pre-hyb solution was preheated to 50°C. Pre-hyb solution was then applied to the miRNA filter in a sandwich box and incubated for 2+ hours at 50°C. Subsequently the Prehyb solution was replaced with 50 mls of fresh solution to which the prepared probe was added. The filter was incubated at 50°C overnight and then washed twice with 40 ml Primary Wash Solution for 10 minutes. The filters were subsequently washed a further two times with primary wash solution with 20 µl of 100 mM ATP, added for 30 minutes. Finally the filters were washed for 5 minutes in 80 ml Secondary Wash Solution. In general the wash volumes were doubled if multiple filters were washed together. All washes were conducted at 50°C with shaking. Finally the filters were sealed in saran wrap and exposed to film, with or without intensifiers, for varying periods to obtain the optimal image.

To strip filters, where necessary, boiling Strip Solution was poured on the filters and they were allowed to cool for 45 minutes to 1 hour with gentle rocking. This process was repeated if the first attempt did not remove sufficient probe. The filters were then rinsed in 2x SSC, wrapped in Saran wrap and exposed to film to judge the success of the stripping.

2.10.2 Luminex 100 analysis of mRNA expression

2.10.2.1 Running the Luminex 100 system

Cells were grown and RNA was purified as described in the section Expression Array Preparation (see section 2.9.2.1). RNA was quantitated by Nanodrop and the small RNA was labeled and bead coupled essentially as described in Blenkiron *et al.* (Blenkiron *et al.*, 2007) with a few minor alterations. Briefly, 10 µg of each sample

was spiked with 3x precontrol oligos at 3 fmoles/sample and the 18-26 mer fraction was purified by PAGE using the BioRad Criterion System. 3' and 5' adapter RNA-DNA hybrid oligos (3': pUUUaaccggaattccagt-idT, 5': acggaattcctcactAAA) were ligated to either end of the size selected RNA fragments with T4 RNA ligase, in between successive rounds of PAGE size selection to isolate the successfully ligated fractions. The cloned RNA species were reverse transcribed with SuperscriptII and the M37 primer (TACTGGAATTCGCGGTTA) and the 8 µl of the cDNA sample was subsequently used as the template for PCR amplification of the library with primers M37 and M33 (5' Biotin-CAACGGAATTCCTCACTAAA). The PCR products were precipitated and resuspended in 64 µl TE pH8 (Sigma). 3 biotinylated post controls were spiked into the samples (100 fmols of each). A mouse miRNA specific set of oligo probes corresponding to 319 miRNAs and assorted controls, each coupled to a colour-coded polystyrene bead were used to profile the miRNA profile of each cloned sample. These probe were divided into four separate bead sets to allow each probe to be coupled to a specific bead colour. The control probes were included in each set. Each sample was hybridized to all 4 bead sets. 33µl of the prepared bead pools were combined with 16 µl of each sample and samples were incubated overnight at 50°C. 1-2 bead blanks and negative TE pH8 (Sigma) controls were included with each sample. The beads were washed as described (Blenkiron et al., 2007) and a 1/100 dilution of SAPE (streptavidin R-phycoerythrin) was added to the beads. These were further incubated at 50°C for 10 minutes to allow the Biotin to bind the SAPE. Finally the median fluorescence intensity (MFI) of each coloured bead in each bead set was determined with a Luminex 100 machine and Starstation software (ACS, Sheffield, UK).

2.10.2.2 Computational analysis of the Luminex data

The analysis of the Luminex data was conducted in R 2.8.1/Bioconductor (<http://www.bioconductor.org/>). The MFI values for each sample were \log_2 transformed. The first replicate samples were considered independently from replicates 2 and 3. The probes for each sample were divided into separate bead sets. The MFIs of samples within each bead set were normalised by using a factor equal to the mean of the MFIs for the precontrol beads for the sample divided by the mean of the precontrol MFIs in all samples, for the relevant bead set. By normalising each bead set separately no assumption was made as to the quality of signal from each of the beadsets. However, following a comparison of the corrected precontrol means for each bead sets I was confident that all of the bead sets were behaving in a similar manner. Subsequently I normalised all samples across all beadsets using factors determined by dividing the average of the precontrol MFIs within each beadset by the average of the precontrol MFIs in all beadsets.

2.10.3 miRNA expression profiling using Illumina/Solexa high-throughput sequencing

2.10.3.1 Preparing Illumina/Solexa samples

The RLT and RPE washes, from the RNeasy Mini Kit (Qiagen) purification of RNA for Northern blot (see section 2.9.1), were subjected to isopropanol precipitation to collect the small RNA fraction which they contain (see section 2.2.3). An Agilent Technologies 2100 Bioanalyzer was used to assess the concentration and integrity of the RNA with a small RNA chip (Agilent). The Solexa libraries were prepared according to the Illumina “Preparing Samples for Analysis of Small RNA” protocol version 1 (2007) with a few exceptions. The initial purification stage was skipped and

the process begun with the addition of the 3' adaptor to the small RNA fraction. The RNA was denatured at 92°C for 2 minutes and snap cooled on ice. Subsequently a slightly altered reaction mixture was compiled (7.4 µl (240-560 ng) RNA, 1.1 µl SRA 3'Adapter, 3 µl DMSO (100%) (Qiagen), 1.5 µl RNA ligase Buffer, 1 µl RNase Out (Invitrogen), 1 µl RNA ligase). This reaction mixture was treated as described by Illumina and size separated on a Novex 15% TBE-urea gel, as described, with the SRA ladder and 10 bp DNA ladder (Invitrogen). The RNA between 35-65 bp was sliced from the gel and purified as described. The RNA was eluted from the gel slice overnight at 4°C prior to the ethanol precipitation (which included 3 µl of GlycoBlue (Ambion) to aid pellet recovery). Precipitated RNA was resuspended in 3.9 µl of RNase free water, heat denatured as above, and the 5'Adapter was ligated with a slightly altered reaction mix (3.9 µl RNA, 1.1 µl SRA 5'Adapter, 2 µl DMSO, 1 µl RNA ligase Buffer, 1 µl RNase OUT and 1 µl RNA ligase). The reaction mixture was treated as described in the protocol and size separated on a 10% TBE-urea gel with both ladders as before. A band of 60-100 bp was sliced from the gel and eluted as before. The eluted RNA was ethanol precipitated and resuspended in 4.5 µl of ultrapure water. The RNA was reverse transcribed and the cDNA was PCR amplified in 4 separate reactions for each template (1 µl RT reaction, 4 µl HF Buffer, 0.125 µl RT primer (CAAGCAGAAGACGGCATAACGA), 0.125 µl smRNA primer (AATGATACGGCGACCACCGACAGGTTTCAGAGTTCTACAGTCCGA), 0.2 µl Phusion Taq (NEB), 12.55 µl Water). The PCR programme was as described by Illumina. A 10% SequaGel PAGE gel was prepared to allow space for maximum RNA separation. The Sequagels were pre-run for 30 minutes at 150 V. 6x Loading Buffer was added to the PCR reactions. These were then heat denatured at 65°C for 15 minutes, snap chilled and the 4 PCR reactions for each sample were combined and

run in 2 wells of the gel alongside the 25 bp DNA ladder. The smRNA libraries were identified by size and sliced from the gel as described. The DNA was eluted from the gel slurry in 400 μ l of 0.3M NaCl overnight at room temperature and then ethanol precipitated as before. Finally the libraries were resuspended in 10 μ l Resuspension Buffer (Illumina). The concentration and integrity of the libraries were once again judged with an Agilent Technologies 2100 Bioanalyzer and a DNA 1000 Chip (Agilent). Libraries were stored at -80°C . The Illumina libraries were Solexa sequenced by the Sanger Institute Core Sequencing Facility (36-cycle Single-ended run, Illumina GA instrument).

2.10.3.2 Computational analysis of the high-throughput sequencing data

The initial quality control and data analysis of the high-throughput sequencing data was conducted by Dr. Cei Abreu-Goodger. Briefly, 3' adapter sequence was removed from the reads. Reads that comprised of a single nucleotide for $>75\%$ of their length or less than 16 bp long following the removal of adapter were stripped from the dataset. An "RNA" database was constructed including all of the mouse miRBase miRNA hairpins and all of the mouse RNA genes from Ensembl (excluding the miRNA genes). Ssaha2 was used to match all unique sequence reads to this database (Ning et al., 2001). If a read matched to different kinds of RNA within this dataset, it was removed. The parameters chosen for ssaha2 required that the aligned region must be 100% identical, have a length of at least 16 bases and comprise $>75\%$ of the actual read length. The aligned region had to start at least at position 2 of the read and with no insertions/deletions (indels) in the aligned region. Alignments on the reverse strand were ignored and only the best scoring alignment(s) for each read was considered.

The accumulated read depth at each base position within all of the RNA genes was calculated. Mature miRNAs were considered present if at least half of their length was covered by a read depth ≥ 1 , otherwise the miRNA was considered to possess a read depth of 0. For other miRNAs, the maximum read depth of the mature miRNA was used. For other RNA genes the maximum depth across the entire gene was considered. In addition to mapping the sequence reads generated as part of this study, the Illumina/Solexa data from the study by Babiarz *et al.* (GEO Database, GPL7195) (Babiarz *et al.*, 2008) were also remapped against the same RNA database for comparison.

I conducted further analysis of the data in R 2.8.1/Bioconductor (<http://www.bioconductor.org/>) using packages *lattice*, *ape*, *affy*, *limma* and *gplots*. In order to normalise the maximum read depths between samples, first RNAs with a maximum depth of 0 in all samples were removed from the dataset and for the remainder the maximum read depths were \log_2 transformed (following the addition of 1 to all values). All RNA species, other than the miRNAs, with a maximum read depth greater than 3 were used to normalise the samples. Based upon this subset of RNAs the maximum read depth values for all of the RNAs in all of the samples were Loess normalised (*affy* package). Subsequently, miRNAs with a normalised \log_2 read number of 0 in all of the *Dgcr8^{tm1,gt1/+}*, *Dgcr8^{tm1,gt2/+}*, *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell line samples were removed from the analysis. All the normalised samples were used to construct a linear model with the *limma* Bioconductor package and subsequently *limma* was used to compare the maximum read depth for each miRNA between the *Dgcr8^{tm1,gt/+}* and *Dgcr8^{tm1,gt/+}* cell lines. Broadly, miRNAs with an adjusted *P*-value < 0.05 and a fold change > 2 were considered to have significantly altered expression.

2.11 Optimised transfection protocols

2.11.1 LacZ siRNA transfection

2.11.1.1 siRNAs

Stealth RNAi LacZ Reporter Control siRNA – 20 μ M (Invitrogen #12935-147)

Stealth RNAi Negative Control Lo GC Duplex #2 – 20 μ M (Invitrogen #12935-110)

2.11.1.2 Optimised RNAi protocol

Dgcr8^{gt1/tm1} cells were cultured for 4 days in non-selective media. On the fourth day the cells were plated to gelatinised 6 well plates, 96×10^4 cells per well in 7.2 ml of non-selective media. Wells were transfected after 3 hours. 12 μ l of the siRNA (20 μ M) was added to 240 μ l OptiMEM I (Gibco). 7.2 μ l of Lipofectamine 2000 (Invitrogen) was added to further 24 μ l of OptiMEM and incubated for 5 minutes at room temperature. Both solutions were mixed gently by pipetting the total volume prior to incubation. The two mixtures were combined and mixed as above. This mixture was incubated for a further 25 minutes at room temperature. The media was gently aspirated from the cells and 2.4 ml of fresh, non-selective media was added to each well. The siRNA-lipid complexes in the OptiMEM mixture were then transferred to this fresh media and the wells were mixed by pipetting the total volume very gently and by rocking the plate 20x. 5 hours later, the media was aspirated from the cells and replaced with a further 7.2 ml of non-selective media.

20 hours after the siRNA-lipid complexes were added to the cells, the cells were Xgal stained (see section 2.5.1).

2.11.1.3 LacZ siRNA transfection and slides

2.11.1.3.1 Slide preparation

Dgcr8^{gt1/tm1} cells were transfected as above, but all reagents and cell numbers were scaled proportionately according to the surface area of the wells (1/12).

Slides were fixed and stained as above (see section 2.5.2.2). In addition to *Dgcr8^{gt1/tm1}* cells, *Dgcr8^{tm1,gt1/+}* cells prepared for Oct4 immuno-staining (see section 2.5.2) were stained alongside the transfected cells as a control. Primary antibodies were added together: LacZ (5 Prime-3 Prime, 1/100) and Oct4 (Santa Cruz – sc8628, 1/50). Secondary antibodies were also added together: Alexa Fluor 488 (Donkey anti-goat - Invitrogen) and Alexa Fluor 594 (Donkey anti-rabbit - Invitrogen). A number of control wells were included stained with either primary or secondary antibodies alone.

2.11.1.3.2 Imaging slides siRNA transfected slides

Images were taken essentially as described above (see section 2.5.2.3), however, settings were used as follows: Sequential setting 1 - Laser Line UV (405): 64%, PMT1 - ~412-489 nm, Sequential setting 2 – Laser line Visible (594): 40%, Laser line Visible (488): 100%, PMT2 - ~492-566, PMT3 - ~608-690 nm.

2.11.2 miRNA mimic transfection

2.11.2.1 miRNA mimics

miRIDIAN Negative Control #2 (Dharmacon CN-002000-01-05)

miRIDIAN mmu-miR-291-3p mimic (Dharmacon C-310470-01)

miRIDIAN mmu-miR-25 mimic (Dharmacon C-310564-01)

2.11.2.2 miRNA transfection protocol

Dgcr8^{gt1/tm1} cells were transfected with miRNA mimics in a 6 well plate, according to the Optimised RNAi protocol (see section 2.11.1.2). 240 pmoles of miRNA mimic were added per well. Post-transfection media was replaced with fresh non-selective media each day. 10, 20 and 44 hours after the initiation of transfection cells were lysed for RNA. The cells were washed twice with DPBS (-CaCl₂ and MgCl₂) (Gibco) and lysed with 1ml of Trizol (see section 2.7.1). Trizol lysate was subsequently stored at -70°C.

2.12 Optimisation of transfection conditions

The optimization experiments for the siRNA transfection of *Dgcr8^{gt1/tm1}* cells were conducted in 24 well tissue culture plates and then scaled proportionately to a method suitable for 6-well transfection (See section 2.11.1.2). The details of the early pre-optimised transfection protocols are beyond the scope of this thesis, however, several methods were used to judge the efficiency of the transfection of siRNAs in these experiments. These methods are described below.

2.12.1 siGLO siRNA transfection

Briefly, *Dgcr8^{gt1/tm1}* cells were transfected in suspension. Cells in a volume of 500µl of non-selective media were combined with 100µl of an OptiMEM, Lipofectamine 2000 and siGLO mix (Lamin A/C siRNA – Human, Dharmacon - D-001620-02-05). 240µl aliquots of the combined mixture were plated to individual gelatinized wells of an 8 well culture slide (Nunc). 4 hours and 40 minutes later the media was replaced with fresh non-selective media. The cells were cultured overnight, washed in 1x DPBS (-CaCl₂ and MgCl₂) (Gibco), and fixed with a few drops of 4%

paraformaldehyde solution. The cells were fixed for 20 minutes at room temperature and then the paraformaldehyde solution was removed by aspiration and replaced with DPBS (-CaCl₂ and MgCl₂) (Gibco). Slides were stored at 4°C. Slides were subsequently washed 4x PBS and quenched with a few drops of 50 mM NH₄Cl in PBS for 10 minutes. The slide wells were then washed a further 3x with PBS and finally the chambers were removed from the slide, a few drops of Vectorshield with DAPI (Vector Laboratories) were added to each slide and a cover slip was applied and fixed to the slide with nail varnish. The slides were examined with a Zeiss fluorescence microscope with Cy3 (siGLO) and DAPI filters.

2.12.2 KIF11 siRNA transfection

Dgcr8^{gt1/tm1} cells in a 24 well format were transfected with a KIF11 siRNA (Silencer KIF11(Eg5) siRNA – Ambion –AM4639), a control siRNA (Negative Control #1 siRNA – Ambion – 4611G), or plated without transfection. 48 hours after transfection initiation the non-selective media in which the cells are cultured was replaced with 400 µl of Alamar Blue media. Samples were subsequently incubated in this media for 2-3 hours at 37°C, 7% CO₂). An aliquot of 100 µl of media from each well was transferred to a well of a BD Bioscience plate reader plate (BD Bioscience– 353947). The plate was centrifuged at 3000 rpm for 1 min at room temperature to remove bubbles from the wells. A plate reader was used to detect colorimetric changes in the media from each well (Excitation wavelength: 544nm, Emission wavelength: 590 nm, Reads: 6, Temperature: 37°C). Transfection efficiency was judged by a ratio of the fluorescence reading of the media from a KIF11 siRNA transfected sample divided by the fluorescence reading from a control-transfected sample. The effect of transfection on cell survival was judged by dividing the fluorescence value of media from a well

of control siRNA transfected cells by the value gained from a sample from a well of non-transfected cells.

2.13 Solutions

2.13.1 General laboratory solutions

3x Sucrose/Cresol Solution

28% Sucrose

0.008% Cresol Red

1x T_{0.1}E

Luria-Bertani (LB) Broth

10 mg/ml Bacto-Tryptone

5 mg/ml Yeast Extract

10 mg/ml NaCl

pH 7.4

Propidium Iodide (PI) Solution

0.1% Triton X-100

50 µg/ml RNase A

20 µg/ml Propidium Iodide

10x MOPS Buffer

0.2 M MOPS

50 mM Sodium Acetate

10 mM EDTA (pH 8.0)

pH 7.0

20x Saline sodium citrate (SSC)

3 M NaCl

0.3 M Trisodium Citrate

1x Phosphate Buffered Saline (PBS)

137 mM NaCl

10 mM Phosphate Buffer

2.7 mM KCl

pH 7.4

1x Tris Borate EDTA (TBE)

89 mM Tris-borate

2 mM EDTA (pH 8)

T_{0.1}E

10 mM Tris-HCl (pH 8)

0.1 mM EDTA

10x NEB Buffer

670 mM Tris-HCl (pH 8.8)

166 mM $(\text{NH}_4)_2\text{SO}_4$

67mM MgCl_2

4% Paraformaldehyde Solution

4 % w/v Paraformaldehyde

1x PBS

pH 7.4

2.13.2 Xgal staining solutions

0.1M Phosphate Buffer

27.1 mM Monobasic Sodium Phosphate

72.9 mM Dibasic Sodium Phosphate

pH 7.3

Fix Buffer

0.1 M Phosphate Buffer (pH 7.3)

5 mM EGTA

2 mM MgCl_2

0.2% Gluteraldehyde

Wash Buffer

0.1 M Phosphate Buffer (pH 7.3)

2 mM MgCl_2

Xgal Staining Buffer

0.1 M Phosphate Buffer (pH 7.3)

2 mM MgCl₂

5 mM Potassium Hexacyanoferrate(II) Trihydrate

5 mM Potassium Hexacyanoferrate(III)

1 mg/ml Xgal (Stock = 500 mg in 10 ml Dimethylformamide)

2.13.3 Western blot solutions

Protein Lysis Buffer

50 mM Tris HCl (pH 7.5)

0.5 M NaCl

1% IGEPAL CA-630

1% Sodium Deoxycholate

0.1% SDS

2 mM EDTA

COMPLETE Protease Inhibitors (Roche) (1 tablet per 10ml)

10x TBS

200 mM Tris HCl

1.37 M NaCl

pH 7.6

Western Blocking Solution

0.1% Tween20

10% w/v Powder Milk

1x TBS

Western Washing Solution

0.1% Tween20

1x TBS

2.13.4 miRNA Northern blot solutions

Denhardt's Solution (100x)

2% Ficoll 400

2% Polyvinylpyrrolidone

2% BSA

Filter with 0.2 µm syringe filter

Pre-Hybridisation/Hybridisation Solution

5x SSC

20 mM Na₂HPO₄ (pH 7.2)

7% SDS

2x Denhardt's Solution

2 mg of Sheared Salmon Sperm DNA (Sigma) preheated to 100°C for 5 minutes

Primary Wash Solution

3x SSC

25 mM Na₂HPO₄ (pH 7.2)

5% SDS

10x Denhardt's Solution

Secondary Wash Solution

1x SSC

1% SDS

Strip Solution

0.015 M NaCl

0.1x SSC

1% SDS

2.13.5 Cell culture solutions

β- mercaptoethanol 1000x stock solution

Dilute 70 µl β-mercaptoethanol (Sigma) in 20 ml Distilled H₂O (Gibco)

Filter sterilize (0.22 µm)

L-Glutamine/Sodium Pyruvate (G/P) Solution

Mix 100 ml L-Glutamine (200 mM, 100x Gibco) with 100ml Sodium Pyruvate (100 mM, 100x Gibco)

Filter sterilize (0.22 µm)

ES Cell Culture medium

500 ml GMEM (Sigma)

50 ml Foetal Bovine Serum (FBS) (Gibco)

10 ml G/P solution

5 ml Non-Essential Amino Acids (Gibco)

0.56 ml β - mercaptoethanol solution

LIF (Quantity determined empirically by the Skarnes laboratory)

Alamar Blue Media

8.87 ml GMEM (Sigma)

1 ml FBS (Gibco)

0.2 ml G/P solution

0.1 ml Non-essential amino acids (Gibco)

11.2 μ l 1000x β - mercaptoethanol solution

0.6 μ l LIF (Quantity determined empirically by laboratory of Dr. William Skarnes)

1.130ml Alamar Blue (Biosource)

1x Trypsin

Add 0.1 g EDTA (Sigma) to 500ml DPBS -CaCl₂ and MgCl₂ (Gibco)

Filter sterilize (0.22 μ m)

Add 5 ml Chicken serum (Gibco)

Add 10 ml 2.5% Trypsin (Gibco)

Aliquot and store at -20°C

(2x Trypsin, as above, but double volume of Trypsin added)

0.1% Gelatin

Add 25 ml 2% Bovine gelatin solution to 500 ml DPBS -CaCl₂ and MgCl₂ (Gibco)

Chapter 3: Disrupting the *Dgcr8* locus

3.1 Aim

In order to establish a cell based system with which to identify miRNA target genes in the absence of endogenous miRNAs expression, I wished to generate cell lines that were impaired in their ability to process miRNAs. Therefore the aim of this chapter is to describe the creation of mouse ES cell lines bearing a gene trap construct in each allele of the *Dgcr8* gene truncating *Dgcr8* transcripts and abrogating DGCR8 function. These lines were characterised at the molecular and phenotypic level and compared to similar lines described by previous studies.

3.2 Introduction

To create a cellular system for the identification of miRNA targets in a fashion resembling the work of Giraldez *et al.* in the zebrafish (Giraldez *et al.*, 2006), the first step is to disrupt the maturation of endogenous miRNAs. If successful, this will allow the re-introduction of miRNAs individually into a background that will not be conducive to combinatorial regulation or functional redundancy.

miRNAs are embedded within longer RNA molecules, pri-miRNAs, that are processed by two rounds of RNase III digestion. The first enzyme, Drosha, operating in concert with a dsRNA binding protein, DGCR8, releases the miRNA precursor molecule as a hairpin. This hairpin is exported to the cytoplasm where it is further processed by the second enzyme, Dicer, removing the loop of the hairpin and liberating the mature miRNA as one strand of the

hairpin stem. At the time of project inception, *Dgcr8* seemed to be the most likely candidate gene for the generation of a cell line with a specific interruption to the miRNA-processing pathway, because its function appeared to be restricted to miRNA processing. In contrast, in addition to its role in miRNA processing Dicer is known to be involved in the cleavage pathway required for the production of siRNAs (Bernstein et al., 2001; Hutvagner et al., 2001), while Drosha was initially proposed as an enzyme involved in rRNA processing and was only later ascribed a role in miRNA biogenesis (Wu et al., 2000). The identification of DGCR8 as a probable miRNA specific processing enzyme has since been supported by others working in the field (Wang et al., 2007).

Dgcr8 is required for the canonical processing of miRNAs (Gregory et al., 2004). DGCR8 contains 2 dsRNA binding domains and a WW domain (Fig.3.1). By deleting subregions, a structural analysis of the DGCR8 protein has been conducted (Yeom et al., 2006). This analysis revealed the C-terminus (Residues 739-750) is required for its association with Drosha. Point mutations and further deletion experiments found that both dsRNA binding domains are required to effectively bind the pri-miRNA. Finally the N-terminus of the protein was found to be responsible for nuclear localisation.

Gene trap mutagenesis is reviewed by Stanford *et al.* (Stanford et al., 2001). In brief, the principle is to insert an exon cassette (the “gene trap”) containing a splice acceptor, a selectable marker and polyadenylation site randomly into the genome. Insertions into an intron of an expressed gene, in the correct orientation, will activate the selectable marker through splicing into the host transcript. The polyadenylation site of the gene trap will cause the truncation of the host transcript and lead to the expression of a truncated host protein

fused to the selection marker. In most cases, this will abrogate the host gene's function. The host gene into which the gene trap has inserted can be identified by the sequencing of Rapid amplification of 5' cDNA ends (5'RACE) products, generated with primers specific to the gene trap exon.

There has been an ongoing international effort to disrupt a sizeable fraction of the genes in the mouse genome by gene trap mutagenesis (International Gene Trap Consortium (IGTC)). Currently, this resource covers ~36% of the genes annotated in v52 of Ensembl. Cell lines bearing these mutations in a single allele are available as a public resource, identifiable through both the Ensembl genome browser and publicly accessible websites (Nord et al., 2006; Stryke et al., 2003). In order to disrupt the *Dgcr8* gene, I opted to select mouse ES cell lines with a gene trap in the *Dgcr8* locus from BayGenomics, a contributor to the IGTC (Nord et al., 2006; Stryke et al., 2003), and to disrupt the second *Dgcr8* allele with a targeted trap.

The targeted trapping approach that I intended to use to disrupt the second allele has been previously described by Friedel *et al.* (Friedel et al., 2005). Gene trap cassettes are targeted to a region of interest through homologous recombination, no longer relying on the random insertion of the trap into gene structures. This approach is highly efficient for genes expressed in ES cells and allowed me to mutate both *Dgcr8* alleles.

As I will be adding miRNAs back into this system following the disruption of the *Dgcr8* locus in order to identify ES cell miRNA targets, it is important that the mutant ES cells are not only depleted of mature miRNAs, but also do not differentiate as a result of the loss of

mature miRNA expression and hence retain some ES cell properties. Using an undifferentiated ES cell line would limit the secondary effects of miRNA depletion and increase the likelihood of ES cell miRNA targets retaining their expression in the developed system. Broad differentiation would bring with it a broad change in the cellular mRNA expression profile and a reduced likelihood that the ES cell miRNA targets would still be expressed. As previous studies have successfully knocked out Dicer and depleted miRNAs without triggering differentiation, I was confident that this combination of properties was achievable (Kanellopoulou et al., 2005; Murchison et al., 2005). In this chapter, in addition to generating a cell line depleted in DGCR8, I also attempt to address whether ES cell properties are retained.

3.3 Disrupting the *Dgcr8* locus

3.3.1 Experimental design

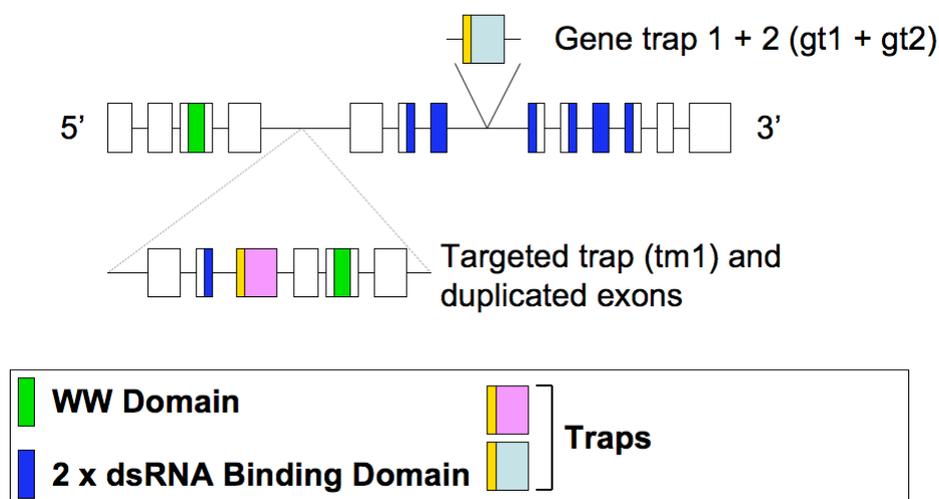
The first step to disrupting *Dgcr8* function in mouse ES cells was to select ES cell lines from the IGTC gene trap resource that contain a random gene trap insertion in one allele of the *Dgcr8* gene. The position of the gene trap in each cell line was confirmed by RT-PCR. The cell lines were subsequently sub-cloned to ensure their homogeneity. Next, I constructed a targeting vector that would insert a gene trap cassette specifically into the second *Dgcr8* allele in these cell lines. The process involves the cloning of a PCR amplified genomic fragment of the *Dgcr8* locus into a vector containing Gateway sites. This fragment was subsequently transferred to a Gateway modified insertion type targeting vector using L/R clonase. The gene trap was designed to integrate into the homologous target sequence by a gap repair mechanism. A gap was introduced into the homologous region and the linearised plasmid construct was then electroporated into each cell line. Additionally this insertion

should lead to the duplication of the target sequence. Cell lines were screened to identify clones that contain the randomly inserted gene trap in the first allele and a targeted trap in the second allele, which were then selected. If successful, this strategy should lead to the truncation of both *Dgcr8* transcripts and the production of a *Dgcr8* hypomorphic or null cell line.

A

Ensembl (v52) ORF

ENSMUST00000115633:



B

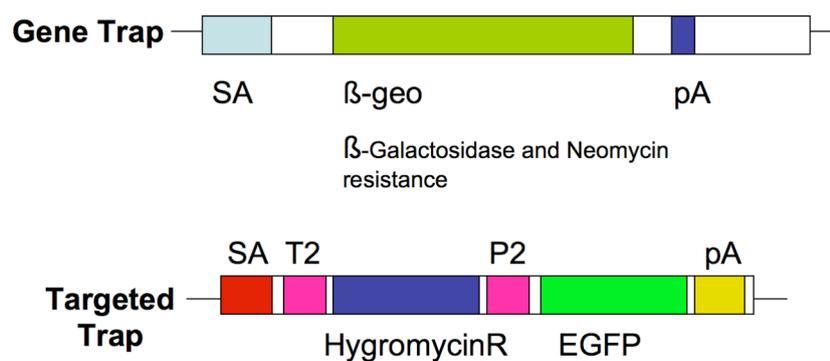


Fig.3.1: A) The structure of the ORF of the *Dgcr8* gene and the positions of inserted traps relative to protein domains. The boxes represent exons and the lines introns. Exons coding for the various protein domains of the DGCR8 protein are shaded. The positions within this gene structure that the BayGenomics gene traps are inserted (gt1 + gt2, light blue) and the intended target site of the targeted trap (pink) are indicated. In addition the exons duplicated by the targeted trap insertion are depicted alongside this trap. Following the insertion of the gene traps the DGCR8 peptide will be truncated and the 3' domains will not be expressed.

B) The structure of the gene trap cassettes used to disrupt the *Dgcr8* locus. Both traps contain a splice acceptor, to ensure splicing into the *Dgcr8* transcript, and a poly-adenylation site to truncate the transcript once inserted. The BayGenomics gene trap (top) contains a region coding for a β geo peptide. When translated this will form a fusion protein with the remaining upstream DGCR8 domains, thus conferring G418 resistance to the

cell line, and staining the cells blue with the addition of Xgal reagents. The targeted trap (bottom) contains T2 and P2 regions. These are of viral origin and cause the elongating ribosomes to skip during translation of the fusion transcript, releasing the upstream peptide and beginning a new peptide from this point. Subsequently independent hygromycin resistance peptides and enhanced green fluorescence protein (EGFP) peptides are translated from this trap. If the two traps were to be inserted into the same allele of *Dgcr8* it would be expected that the expression of selectable markers would be limited to those coded by the upstream trap as the transcript would no longer splice into the downstream cassette. This allows this targeted insertion event to be distinguished from one in which traps are inserted within separate alleles using Xgal staining, which require β -galactosidase expression.

	Position in ENSMUST00000115633	Position in ENSMUST00000009321	Primer Orientation
Features			
Bay Genomics Gene Trap	Between Exons 9 and 10	Between Exons 8 and 9	NA
Duplicated Region	Exon 4 to Exon 8	Exon 3 to Exon 7	NA
Targeted Trap Cassette	Between Exons 8 and 9	Between Exons 7 and 8	NA
Primers			
1	Exon 9	Exon 8	5' -> 3'
3	Exon 8	Exon 7	5' -> 3'
21	Exon 3	Exon 2	5' -> 3'
22	Exon 3	Exon 2	5' -> 3'
2	BayGenomics Gene Trap	BayGenomics Gene Trap	3' -> 5'
4	BayGenomics Gene Trap	BayGenomics Gene Trap	3' -> 5'
23	BayGenomics Gene Trap AND Targeted Trap Cassette	BayGenomics Gene Trap AND Targeted Trap Cassette	3' -> 5'
24	BayGenomics Gene Trap AND Targeted Trap Cassette	BayGenomics Gene Trap AND Targeted Trap Cassette	3' -> 5'
5	Exon 11	Exon 10	3' -> 5'
6	Exon 11/12	Exon 10/11	3' -> 5'

Table 3.1: A description of the relative positions of important features within the structure of the *Dgcr8* transcript. Noted are the positions of the gene trap cassettes and the region duplicated as a consequence of the insertion of the second targeted trap vector. In addition, the relative positions and orientations of the primers used to check the gene structure by RT-PCR are also included.

Cell line genotype	Nomenclature when considering cells with a similar genotype	Schematic of <i>Dgcr8</i> transcript structures
<i>Dgcr8</i> ^{+/+}	NA	
<i>Dgcr8</i> ^{gt1/+}	<i>Dgcr8</i> ^{gt/+}	
<i>Dgcr8</i> ^{gt2/+}		
<i>Dgcr8</i> ^{tm1,gt1/+}	<i>Dgcr8</i> ^{tm1,gt/+}	
<i>Dgcr8</i> ^{tm1,gt2/+}		
<i>Dgcr8</i> ^{gt1/tm1}	<i>Dgcr8</i> ^{gt/tm1}	
<i>Dgcr8</i> ^{gt2/tm1}		

Table 3.2 A summary of the genotype nomenclature used throughout this thesis. The first column lists the genotype of each cell line used in this study. The cell lines that are derived from the same original BayGenomics cell line contain the same initial gene trap (“gt1” OR “gt2”). “tm1” refers to the targeted trap insertion. The second column lists the nomenclature used to refer to sets of cell lines with similar genotypes when analyses are performed that consider two independently derived cell lines as biological replicates and in which their data is combined. In these cases the number is removed from the gene trap names. The third column provides a schematic representation of the gene structures at the *Dgcr8* locus. Red boxes represent exons, light blue boxes are UTRs, lines represent introns, dark blue boxes refer to gene traps and green boxes to targeted traps.

3.3.2 Selecting *Dgcr8* trapped cell lines

Initially, 2 mouse ES cell lines containing a gene trap within the *Dgcr8* locus were selected from the IGTC website (<http://www.genetrapped.org>). These would subsequently form the basis of the second allele targeting experiments. All available IGTC gene trapped cell lines are annotated in Ensembl as a DAS track, with the trap in each cell line mapped to the genome through the sequencing of 5'RACE products, to identify exons upstream of the intron into which the gene trap is inserted. To maximise the possibility that these gene traps create a null allele, cell lines were selected that contain gene traps positioned as near as possible to the 5' end of the *Dgcr8* transcript. Consequently, truncated fusion protein products are unlikely to retain wild type function. The two cell lines selected (XG058 and XH157) are independent gene trapped cell lines derived from separate trapping experiments. However, both of these cell lines contain a gene trap within the same intron of *Dgcr8* (Fig.3.1A).

To ensure that the position of the traps had been annotated correctly, the 5'RACE sequences provided by BayGenomics were analysed by BLAST against mouse cDNA sequences in Ensembl (v31, m33) using default settings. As expected, the RACE sequences from each gene trap aligned to exons 6-8 of the *Dgcr8* Ensembl transcript (ENSMUST00000009321), confirming their correct annotation by the BayGenomics pipeline.

Next a manual comparison of the gene trap insertion site with the Pfam protein domain structure of the *Dgcr8* gene (Ensembl v31, m33) revealed that both gene traps are inserted between exons coding for the first of DGCR8's two double stranded RNA binding domains. As a result, it is expected that truncating the endogenous protein at this point and creating a

fusion protein should have a catastrophic effect on the function of DGCR8 protein coded by mutant transcripts from these trapped alleles (Yeom et al., 2006).

3.3.3 Confirmation of *Dgcr8* gene trapped cell lines.

To confirm that the selected gene trap lines were generating the expected fusion transcript from the disrupted *Dgcr8* allele and a wild type transcript from the unaffected allele, a RT-PCR approach was adopted. Nested primers were designed for RT-PCR to amplify cDNAs expressed from the gene trap and wild type alleles (Fig.3.2). A set of 5' primers that anneal to an exon upstream of the gene trap was partnered to a set of primers within the gene trap cassette and to a further set of primers that anneal to an exon downstream of the gene trap insertion site. As a positive control, a further set of primers were designed to bind to exons either side of a 187bp intron in the *Arsa* housekeeping gene. These primers would allow me to perform a control amplification to judge genomic contamination of the cDNA used in the RT-PCR reaction, as contamination would result in a PCR product of a larger than expected size.

RNA was purified from the two independently trapped BayGenomics cell lines (henceforth known as *Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+} (Table 3.2)). The RNA was quantified and subsequently reverse transcribed into cDNA. PCR reactions using the primers described above were used to confirm the identity of each cell line and the position of the gene trap (Fig.3.2). All products were of the expected size, implying that the gene traps were indeed within intron 8 of the Ensembl transcript ENSMUST0000009321, although there was evidence of a significant second amplification product of a larger than expected size in the wild type transcript lane (see below).

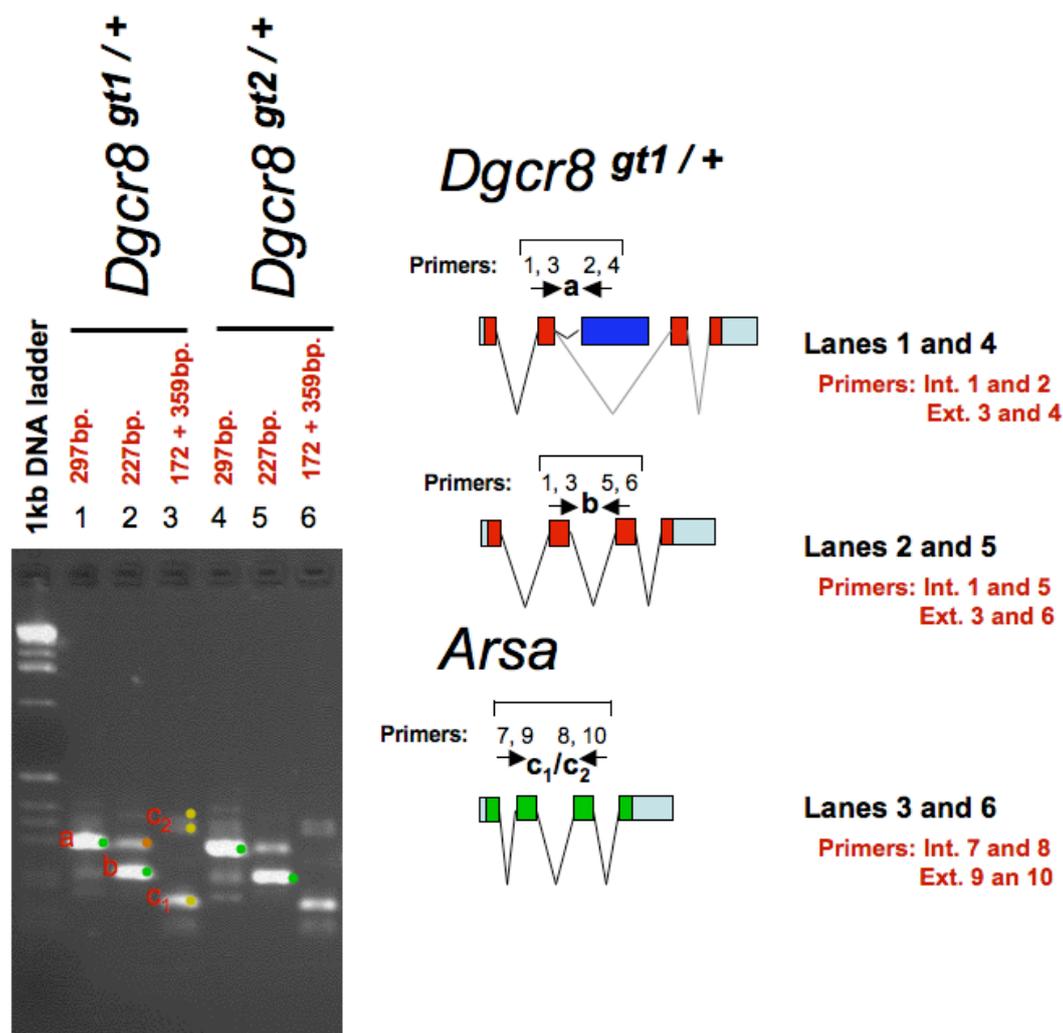


Fig 3.2: Nested RT-PCR to confirm the gene trap position within two BayGenomics *Dgcr8*^{gt1/+} cell lines: Nested primers were designed to amplify fragments across exon-exon boundaries. Set 'a' (Primers 1, 2, 3 and 4) anneal to an exon upstream of the gene trap and to the gene trap itself (Table 3.1), amplifying an expected fragment of 297bp. Set 'b' contains primers 1 and 3 which anneal to the upstream exon, but also two primers that anneal to an exon downstream of the gene trap insertion site (Primers 5 and 6) (Table 3.1). These are expected to amplify a PCR product corresponding to wild type splice events, 227bp in length. Primer set 'c' is a control set of primers. These anneal either side of a short intron 187bp in length within *Arsa* cDNA. Fragments amplified from reverse transcribed, spliced RNA will be 172bp in length (c₁). Fragments amplified from genomic DNA contamination of the cDNA sample will be 359bp in length (c₂). The PCR reactions run in lanes 1, 2 and 3 are amplified from the *Dgcr8*^{gt1/+} cDNA template. Lanes 4, 5 and 6 are amplified from the *Dgcr8*^{gt2/+} template cDNA. The bands in Lanes 1 and 4 are amplified with the gene-trap to exon primers. The bands in Lanes 2 and 5 are amplified with the exon to exon primers. Lanes 3 and 6 contain the *Arsa* control amplification fragments. Green spots represent transcripts of the expected length subsequently re-amplified and sequenced. The orange spot represents an unknown band, also re-amplified and sequenced. Yellow spots in lane 3 are bands later re-amplified and discussed in section 3.3.6.

Products amplified from trapped and wild type transcripts (Fig.3.2, marked in green) were re-amplified, and sequenced. Where possible, non-specific bands were also purified and sequenced in order to characterise them (Fig.3.2, marked in orange). Each RT-PCR product was sequenced from both ends using the internal PCR primers with which it was amplified. The sequence results were compared to the expected sequence in Gap4.

The PCR products equivalent to the green bands in lanes 1 and 4 (Fig.3.2) were as expected. The sequences from both ends of the products overlapped, providing representative sequence for the entire length of the fragment. The sequence alignment also allowed me to confirm position of the splice junction between the *Dgcr8* transcript and gene trap. Likewise, the sequences from the PCR products of the expected sizes amplified from the WT transcripts (Lanes 2 and 5, marked in green (Fig.3.2)) matched their templates without any significant discrepancies and once again the primer sequences were evident at each end of the sequenced PCR products.

The ~300bp, non-specific band marked in orange (Fig.3.2) was also sequenced. Gap4 analysis revealed a 70bp repeat sequence that was probably derived from an RT-PCR artefact, as, when the consensus sequence was folded with RNAfold (Default settings) (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>), it contained a substantial secondary structure.

The control fragments amplified from *Arsa* cDNA produced three bands. Two of these were of the expected sizes. The intermediate band implies that there may be some genomic contamination of my reverse transcribed cDNA. The largest band was of unknown origin. A

more in depth assessment of the identities of these bands was conducted at a later date (see section 3.3.6). Any genomic contamination of the cDNA did not affect the identification of the correctly amplified spliced gene trap products.

3.3.4 Subcloning *Dgcr8* cell lines

Dgcr8^{gt1/+} and *Dgcr8*^{gt2/+} were sub-cloned to ensure that they were homogenous cell populations and did not contain wild type cell contamination. Following subcloning the RT-PCR procedure explained above was repeated in order to confirm the identity of each sub-clone. An E14 wild type control was included in these reactions and as expected the PCR performed on the cDNA template generated from these cell lines only generated a product between the primers pairs which annealed to the 5' and 3' exons and not between the pairs designed to amplify between the upstream exon and the gene trap.

3.3.5 Generating a targeting vector

The *Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+} cell lines originate from the E14 ES cell line, which is derived from the 129P2 mouse strain. Previous studies have demonstrated that the isogenicity of the homologous fragment within a targeting vector can have a profound effect on the efficiency of targeted insertion (te Riele et al., 1992). Therefore it is necessary to ensure that the targeting fragment is amplified from the same strain template DNA to reduce the occurrence of single nucleotide polymorphisms (SNPs) that may interfere with the efficiency of homologous recombination. To this end, a BAC of 129 origin (129S7/AB2.2 BAC clone (bMQ-62C21) (Adams et al., 2005)) that spans the *Dgcr8* gene was identified from Ensembl. The end sequences, by which bMQ-62C21 was mapped, were checked to ensure the clone

had been annotated correctly. Subsequently bMQ-62C21 was used as the template for the amplification of the homologous region of the targeting vector.

The intended placement of the targeted trap within the *Dgcr8* locus placed several constraints upon the design of the primers used to amplify the homologous fragment (Primers 11 and 12, Fig.3.4). Therefore following the repeat masking of the template region, primers were designed such that:

- Primers were restricted to a region downstream of ENSMUST00000009321 exon 2 and upstream of exon 8. As a consequence there would be sufficient sequence 5' of the homologous region within which to design primers to confirm the insertion of a second allele by RT-PCR. In addition the second trap would also be targeted upstream of the gene trap currently inserted in the *Dgcr8* locus, which would cause a loss of marker expression from the first gene trap if the second trap inserted into the same allele rather than into the wild type allele. Therefore it would be simple to distinguish these two events (Fig.3.1).
- Primers were designed within intronic sequences. This ensures that the trapped allele continues to splice as expected and that the splice acceptor of the trap resides in an intron and is available to the splicing mechanisms. The amplified region's ends were at least 120bp from any splice junctions. The primers were positioned between exons 2 and 3 and 7 and 8 of ENSMUST00000009321.

- The amplified fragment was greater than 5000bp in length (6083bp) to allow efficient targeting of the trap.
- Following primer design, an *AscI* restriction site was added to the 5' end of each primer followed by a short "AATT" sequence. This restriction site would later be used in order to clone the fragment into a Gateway vector.

The chloramphenicol resistance gene and *ccdB* cassette were removed from between the attR3 and attR4 Gateway sites of the pR3R4AsiSI Gateway vector (Fig.3.3) by *AscI* restriction digest and the plasmid backbone was gel purified and de-phosphorylated. The fragment for cloning was amplified by Long PCR from *NotI* digested bMQ-62C21 DNA using primers 11 and 12 (Fig.3.4) and gel purified. The *AscI* digested targeting fragment was then ligated into the pR3R4AsiSI vector and transformed into MACH1 *E. coli*.

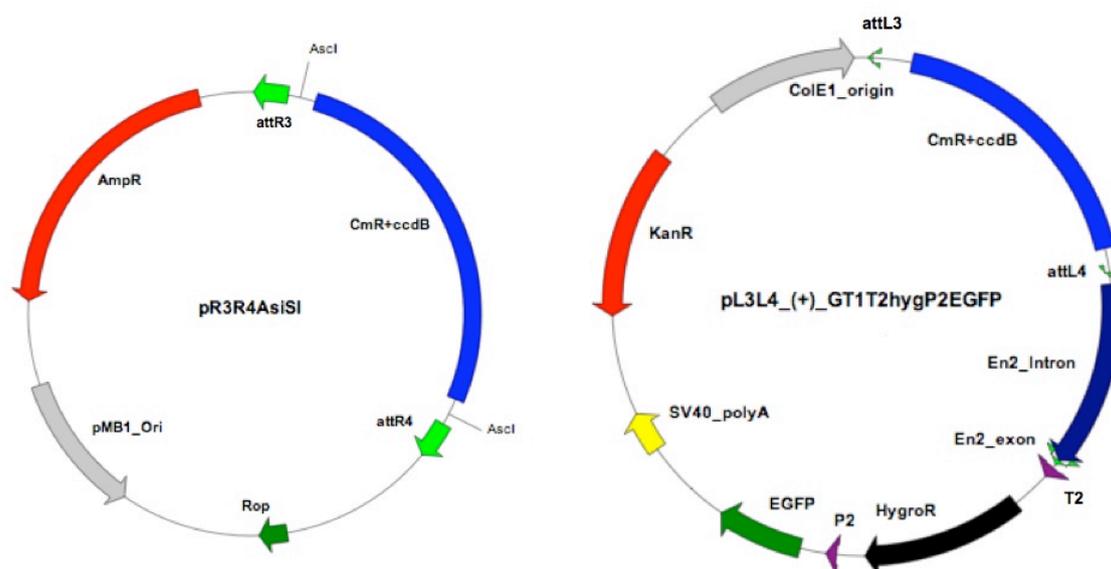
MACH1 cell colonies containing plasmids with the correctly ligated fragments were identified by colony PCR with a set of primers that amplify a region from within the ligated fragment (Primers 13, 14, 15 and 16 (See Fig.3.4)). It is essential for trap function that its splice acceptor be situated upstream of the trap within the *Dgcr8* transcript. As the orientation of the genomic fragment within the pR3R4AsiSI plasmid would determine the orientation within the trap targeting vector following Gateway transfer, a further colony PCR was conducted with primers from both within the fragment and within the flanking regions (Primers 17 and 18, See Fig.3.4). A single colony was identified containing the correctly inserted fragment and the ends of the fragment were sequenced primed by primers 14, 15, 17 and 18. BLASTN was used to compare the consensus of the resulting sequence contigs

against the mouse genomic sequence. As expected the sequences matched *Dgcr8* with a very high percentage alignment (98.96-99.79%) and were positioned as expected.

The targeting fragment was transferred from the cloning vector into the trap vector (pL3/L4_(+)_GT1T2hygroP2EGFP (Fig.3.3)) through an *in vitro* L/R clonase reaction. Once inserted into the genome, it is essential that the phase of the reading frame of the trap is the same as that of the upstream exon of the targeted transcript, so that once transcribed and spliced the trap gene is translated such that the selectable markers are expressed in frame. The phase of the splice junction of the upstream exon (Exon 7, ENSMUST00000009321) was “1”, hence the corresponding phase trap vector was used.

After transformation of the L/R reaction, kanamycin resistant colonies were tested by colony PCR for the presence of the amplified *Dgcr8* fragment (Primers 14, 17, 15, 18). Subsequently, a second colony PCR confirmed the orientation of the inserted fragment within the targeting vector in a subset of the positive colonies (Primers 14, 19, 15, 20) (Fig.3.4).

One of the correct colonies was selected and prepared for electroporation. Prior to electroporation, a unique *HindIII* restriction site approximately in the middle of the homologous fragment, was used to linearise the plasmid. This insertion-type vector will recombine into the ES cell genome and duplicate the target sequence.



Plasmid and Gene Trap Characteristics:	
Plasmid	Features
pGT1Ixf Gene Trap	En2-intron / En2-exon (Splice acceptor) β-geo (Fusion with endogenous peptide - confers neomycin resistance and stains blue with X-gal staining reagents) SV40 pA (Polyadenylation site)
pR3/R4AsiSI	R3-Ascl-CmR-ccdB-Ascl-R4 (Gateway R3-4 Chloramphenicol resistance and ccdB cassette) pMB1 origin rop (Reduces copy No.) AmpR (Ampicillin resistance)
pL3/L4(+)GT1T2hygP2EGFP	L3-CmR-ccdB-L4 (Gateway L3-4 Chloramphenicol resistance and ccdB cassette) En2-intron / En2-exon (Splice acceptor) T2A (Causes ribosome to skip) HygroR (Hygromycin resistance) P2A (Causes ribosome to skip) EGFP (Enhanced green fluorescent protein) SV40 pA (Polyadenylation site) KanR (Kanamycin resistance) ColE1 origin

Fig.3.3: Gene traps and vectors used to disrupt the *Dgcr8* locus. Plasmids generously provided by the Skarnes laboratory. Vector maps were drawn with Savvy v0.1 (<http://www.bioinformatics.org/savvy/>). Red text within the plasmid description refers to elements inserted into the *Dgcr8* transcript.

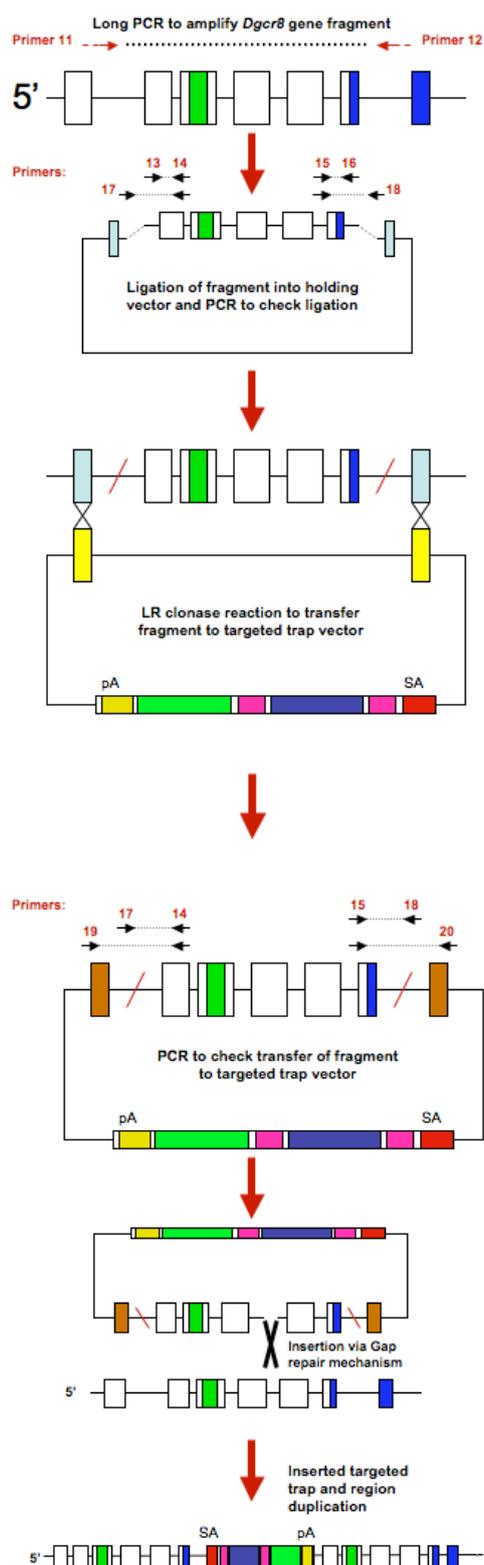


Fig.3.4: Principles of targeting vector construction: The fragment required to target the vector to the *Dgcr8* locus is amplified by long PCR from a BAC clone derived from the *Dgcr8* region (Primers 11 and 12). These

primers contain *AscI* restriction sites, which are used to clone the fragment into the pR3/R4AsiSI plasmid in place of the CmR-ccdB cassette. The fragment is transferred to the pL3/L4(+)_{GT1T2}hygP2EGFP targeting vector, which contains the trap cassette, by an L/R clonase reaction. The target vector is linearised through the introduction of a gap into the centre of the homologous region by restriction digest. This will allow the vector to insert into the *Dgcr8* locus through a gap repair mechanism. The targeting vector is introduced to mouse ES cells by electroporation. The vector then inserts into the region homologous to the amplified fragment, causing a duplication of the region itself. At various stages primers were designed to amplify fragments with which to confirm fragment insertion, transfer and orientation within the plasmids. These are illustrated as red numbers above the schematic vectors.

3.3.6 Identification of cells with a successfully targeted *Dgcr8* locus

Confirmed sub-clones of *Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+} were electroporated with the linearised plasmid pL3/L4_(+)_GT1T2hygroP2EGFP, containing the cloned fragment of *Dgcr8*. Hygromycin resistant colonies from each cell line were picked, expanded, archived and lysed for the preparation of RNA.

A

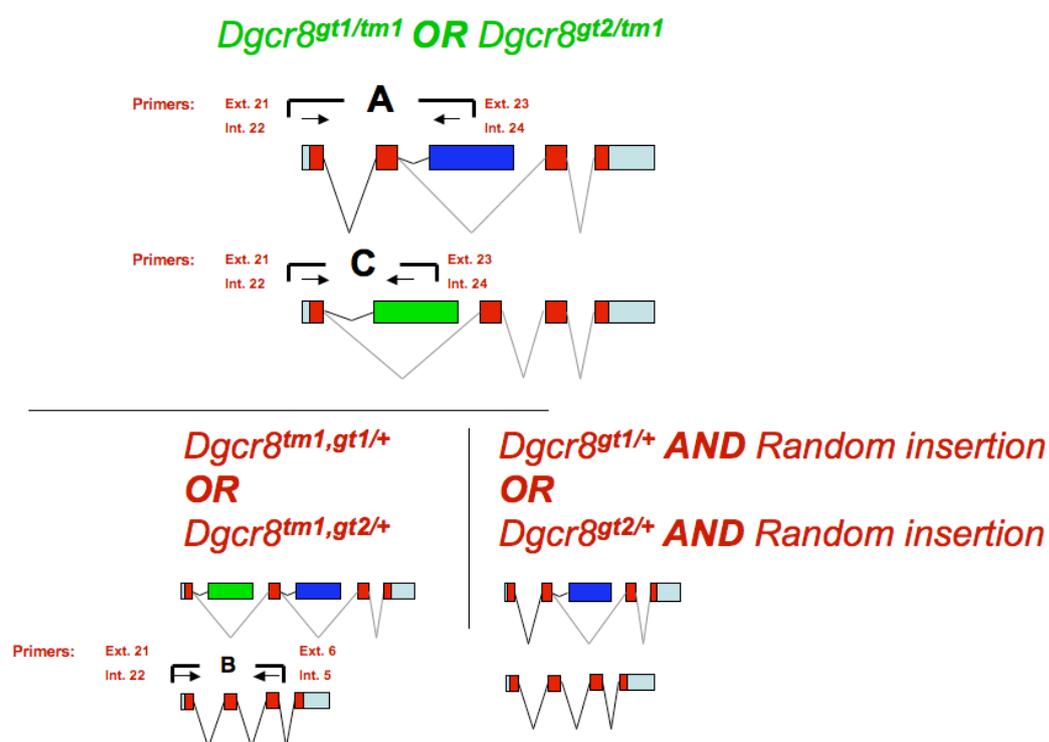


Fig.3.5: A) A schematic of the expected outcomes of the targeted vector insertion and the primers designed to distinguish these events. The targeted trap can insert as intended into the second allele, resulting in the disruption of both alleles (*Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* (Table 3.2)) (top), it can insert into the same allele as the initial gene trap (*Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* (Table 3.2)) (bottom left), which will disrupt the expression of selectable markers from the initial gene trap, or it can insert at a random point in the genome (*Dgcr8^{gt1/+}* and *Dgcr8^{gt2/+}* AND Random insertion) (bottom right). Primers were designed to distinguish these three genotypes as marked. A single primer set (Primers 21, 22, 23 and 24) will amplify fragments from cDNAs containing either gene trap. Primer 21 and 22 are designed to anneal to exons upstream of the region duplicated by the targeted trap insertion while primers 23 and 24 are an overlapping nested set designed to anneal to sequences common to both traps' splice acceptors. These fragments will be of a different size depending upon the traps expressed in the cell line (larger if amplifying from cDNA containing the downstream gene trap and smaller if amplifying from the targeted trap transcripts). Primers were also selected to amplify fragments from cDNAs demonstrating wild type splicing events, with forward primers annealing upstream of the traps and downstream primers annealing 3' of both traps (Primers 21, 22, 5 and 6).

B

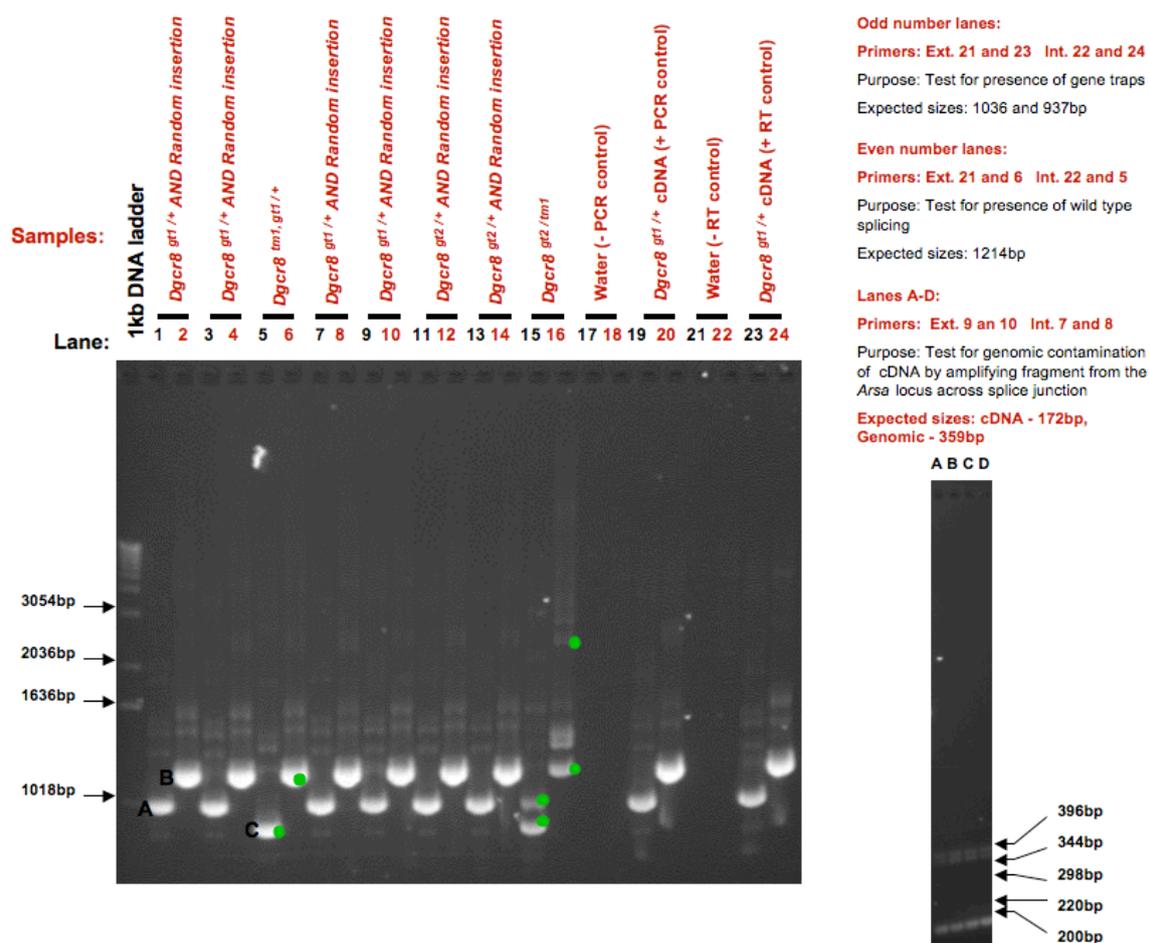


Fig.3.5: B) RT-PCR was used to distinguish the genotypes of ES cell lines resulting from the electroporation of the gene trap targeting vector. Nested RT-PCR was conducted with the primer combinations described in Fig.3.5A with RNA derived from each of the hygromycin resistant colonies collected from the electroporation of the targeting construct. Odd number lanes are expected to contain either a single fragment (Fig.3.5A (top) fragments A or C) or both of these fragments. If both templates are present in the lysate, and hence both fragments are amplified in the PCR reaction, the cell line is expected to contain a trap in each allele (*Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}*). The presence of only a single fragment in this lane indicates that either the targeted trap has inserted upstream of the gene trap at this locus (smaller fragment) or the targeted trap has randomly inserted elsewhere in the genome (larger fragment). Therefore only a single gene trap is expressed from the *Dgcr8* locus. As explained previously (see section 3.3.3), primers 7, 8, 9 and 10 amplify a fragment of the *Arsa* cDNA in order to judge the presence of genomic contamination of the cDNA. As previously noted, this last set of primers amplifies an unknown band in addition to the two expected bands.

Correctly targeted clones were identified by RT-PCR performed on cDNA produced from a set of 25 *Dgcr8*^{gt1/+} and 23 *Dgcr8*^{gt2/+} derived hygromycin resistant colonies (Table 3.3) (Fig.3.5B). Nested primer sets were designed to distinguish between cell lines with a correctly inserted targeted trap from those in which the traps have inserted randomly or into the same allele as the BayGenomics gene trap (Fig.3.5A). This process identified 4-5 colonies associated with RT-PCR results consistent with the event of a targeted trap inserting into non-gene-trapped allele of *Dgcr8* (Table 3.3), of which 2-3 derived from each parent cell line. The bands amplified in one of the 5 cases were weak, so the PCR was deemed inconclusive.

	Random Insertion	Gene trapped allele insertion	2nd Allele Insertion	Failed PCR
<i>Dgcr8</i> ^{gt1/+}	17	4	2 (+1)	1
<i>Dgcr8</i> ^{gt2/+}	16	5	2	NA
Total	33	9	4 (+1)	1

Table.3.3: A summary of the number of each kind of insertion event resulting from the targeted gene trap electroporation experiment. The genotype of each cell line at the *Dgcr8* locus was determined by RT-PCR.

It might be expected that in the case of potential null cell lines, the primers designed to amplify fragments from wild type cDNA (Primers 21, 22 and 5, 6) would not amplify products; as transcripts from both alleles should be truncated by traps. However, these products were seen (Fig.3.5B). It should be noted that this aberrant splicing through the traps does not necessarily imply that the experiment has been unsuccessful as such transcripts may not be full length or translated. RT-PCR is also notoriously non-quantitative with these products perhaps being derived from an insignificant quantity of residual wild type transcript. However, this does raise the possibility that the cell lines may be hypomorphic for *Dgcr8*

rather than null. This issue was addressed by Northern blot as described later (see section 3.4.1).

Once again the control RT-PCR reaction specific for the *Arsa* gene transcript showed products derived from both cDNA and genomic DNA templates for all samples. The presence of genomic DNA in the reactions is not a problem as all PCR primers used were designed to amplify products across multiple exon-exon splice junctions and any products derived from this contamination would be of a larger than expected size.

Interestingly, there are three bands present in the lane corresponding to the *Arsa* primer PCR reaction, the third being slightly larger than the expected genomic DNA size. This reaction was repeated on cDNA derived from DNase treated RNA from wild type cells. Again there were three bands evident in the lane. A PCR reaction performed on simultaneously prepared cDNA, which omitted the reverse transcriptase failed to produce any products. These results imply that none of the bands are amplified from genomic DNA, even those that correspond to the genomic band size. It is therefore likely that the template for this product is generated from unspliced RNA molecules. All three bands were sequenced. As expected, sequences derived from the smaller of the two bands corresponded to the expected splice form of the cDNA. The band that corresponded to the unspliced cDNA also aligned to the genomic template as expected. Finally, sequences from the largest of the three bands did not align to each other easily and were not of a high enough quality to ascertain their origin. However, since these products are not amplified from non-reverse transcribed template, it would suggest that they originate from an RNA source and hence the exact origin of the sequences

is not essential in order to fulfil the purpose of the control PCR. In future a non-reverse transcribed template should prove a better control for genomic DNA contamination.

Subsequently, the RT-PCR was repeated for a representative sample of the cell lines originally tested in this PCR. The cDNA for these tests was prepared from Trizol purified, DNase treated RNA. Amplified PCR products were purified for both primers 21, 23 and 22, 24 and primers 21, 6 and 22, 5 equivalent to those marked green in Fig.3.5B. On re-amplification, the PCR product for the wild type transcript primers amplified a larger than expected product from a template based on cDNA from one of the cell lines believed to contain traps in both alleles. This larger band corresponded to the size product that might be expected if the cDNA template skipped the gene trap but contained all of the duplicated exons. This band was also purified.

All purified products were sequenced and the sequences aligned to the ENSMUST00000009321 sequence as expected. As noted above, in the case of cell lines containing a trap in each allele of *Dgcr8*, two fragments are expected to be amplified in a single PCR reaction. It is important therefore to ensure that the larger of the two fragments sequenced from these lanes is indeed amplified from cDNA originating from transcripts containing the first gene trap and is not a fragment amplified by the external nested primers from targeted trap transcripts. Importantly, the sequence contig for the larger of the two fragments extended into exon 8 of ENSMUST00000009321, and thus confirmed the identity of the band as this exon is downstream of the intron within which the targeted trap has inserted.

Sequence derived from the large band amplified by primers 21, 6 and 22, 5, that was suspected to contain the duplicated exons caused by the targeted insertion, aligned to the relevant exons of *Dgcr8*. However, the sequences obtained were not long enough to determine if the product was indeed generated from duplicated exons, although the fact that the sequences derive from *Dgcr8*, in addition to the length of the PCR product, supports this hypothesis.

3.3.7 Determining the functional significance of trap insertions through the use of miRNA Northern blots

The purpose of disrupting the *Dgcr8* locus is to cause an interruption in miRNA processing and a reduction of mature miRNAs in mouse embryonic stem cells in order to provide a resource for miRNA target identification in the absence of endogenous miRNA expression. To determine whether the cell lines that appeared to contain a trap in each *Dgcr8* allele should be pursued further, the functional significance of the RT-PCR results was examined through a series of Northern blots for miRNAs known to be expressed in mouse ES cells (Houbaviy et al., 2003).

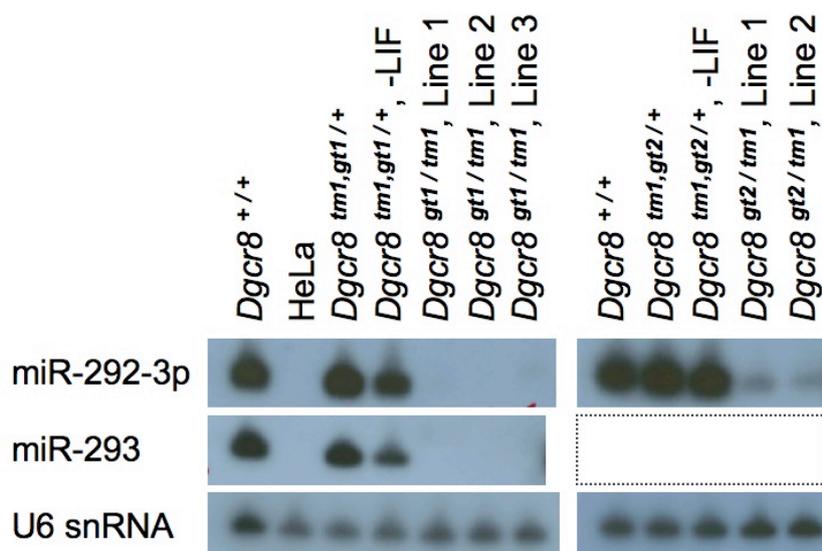


Fig.3.6: Northern blot to judge miRNA expression within a range of cell lines. At this early stage the cells were cultured and RNA was prepared in a fashion that differed slightly from that described in the Methods section. Cells were cultured under hygromycin selection, where suitable. *Dgcr8*^{+/+} cells were cultured in non-selective media. HeLa S3 cells were cultured under specific conditions (see section 2.7.4). The hygromycin selection was removed as cells were plated for RNA preparation. Cells were lysed 2 days after plating. Cell lines plated without LIF were seeded at half the cell number and cultured for 3-4 days prior to lysis. RNA was prepared with Trizol reagent. At this stage the confluence of the cells on RNA lysis was also not tightly controlled. Generally 20ug of total RNA was run per lane. However, where this was not possible, as much RNA was used as feasible (*Dgcr8*^{+/+} - 17.1ug, *Dgcr8*^{tm1,gt1/+} - 15.1ug, *Dgcr8*^{tm1,gt1/+} - 16.3ug). These Northern blots were probed with oligos complementary to miR293, miR-292-3p and U6 loading control. The missing panel was not probed.

The mutant cells were maintained under hygromycin selection throughout their culture to prevent homologous recombination between the duplicated regions that flank the targeted trap, which could potentially restore DGCR8 function (an event which may produce a selective advantage in untreated cells). For each parent cell line (*Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+}), hygromycin resistant colonies with two traps in the same allele were chosen as the controls, as these cells would also control for the effect of hygromycin treatment and for other potential effects caused by the point of insertion (*Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} (Table

3.2)). These cells were cultured under the same selective conditions as the homozygous mutant cells with a trap in each allele (*Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} (Table 3.2)). An apparent growth defect was observed in the 5 homozygous mutant cell lines (*Dgcr8*^{gt/tm1}) as they appeared to replicate more slowly than the control cell lines. At this early stage these cell lines were generally split at a much lower ratio than the heterozygous cell lines and were initially not fed every day (although these conditions were changed as I became more accustomed to their growth patterns). The 2 heterozygous cell lines (*Dgcr8*^{gtm1,gt1/+} and *Dgcr8*^{gtm1,gt2/+}) were also grown up in the absence of LIF to trigger differentiation. This control was included in case the hygromycin selection was having a more profound effect on the differentiation status of the homozygous mutant cells. Inducing significant differentiation in the heterozygous cells through the removal of LIF from the media and then comparing the ES cell specific miRNA expression to that of the homozygous cells should help to determine whether low level differentiation is the root cause of any expression changes seen between the heterozygous and homozygous cells. As a further positive control for the ES cell miRNA expression, RNA derived from wild type E14 cells was included on the Northern blots. In addition RNA derived from HeLaS3 cells was also included as a negative control for the miRNAs profiled.

For Northern blotting, RNA was separated by 15% TBE-Urea gel electrophoresis suitable for discriminating small RNA molecules and transferred to a nylon membrane. The membrane was hybridised with radiolabelled DNA oligos complementary to the mature form of the miRNA of interest. The results clearly demonstrate a significant reduction of the ES cell specific miRNAs tested in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines (Fig.3.6), to the point at which the miRNAs cannot be clearly discerned on the blot without developing the blot for an

extended period. Subsequently, the miRNA signal from the *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{+/+} samples was so strong that it is difficult to know if the faint signal in the *Dgcr8*^{gt2/tm1} lane was produced by residual mature miRNA expressed in these cells or by slight contamination of these wells during the gel loading process. The extent of the mature miRNA depletion is addressed in section 3.4.2 and further in Chapter 4. The HeLaS3 sample is devoid of miRNA signal. This is expected, as the miRNAs selected are thought to be specifically ES cell expressed, thereby increasing the confidence in the specificity of the probes. There also appears to be a reduction in the level of these ES cell miRNAs upon differentiation (in the absence of LIF) although the reduction in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} is far greater than that seen in the cells cultured without LIF. This implies that the reduction seen in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells is beyond that expected from a relative increase in the proportion of differentiated cells in the culture. In all these results suggested that the introduction of the traps into the *Dgcr8* locus was having a significant effect on the processing of miRNAs in the stem cells.

At this point the choice of miRNA deficient cell lines for further study was refined to a single representative line for each of the original clones (*Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1}). Each of these mutants was paired with a heterozygous control cell line with a second trap inserted upstream of the first gene trap within the same allele, (*Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+}).

3.3.8 Confirmation of the position of inserted targeted traps by gene-trap specific staining

Due to the position of the traps within each of the cell lines, it should be possible to distinguish individual insertion events by Xgal staining. As shown in Fig.3.3, the initial gene trap within each line (gt1 and gt2) will code for a β -geo fusion protein that will cause cells that express it to stain blue upon staining with X-gal. The insertion of the second trap upstream of this trap in the same allele, as is the case for *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+}, should cause the loss of positive Xgal staining, while *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} should retain their Xgal staining phenotype as the trap should be inserted into the previously wild type allele. As a control for this experiment, a cell line was selected that was derived from *Dgcr8*^{gt1/+} in the electroporation experiment but with a second, expressed random insertion of the targeting vector elsewhere in the genome. Once again this line would be expected to stain blue. Wild type E14 cells (*Dgcr8*^{+/+}) should stain white following the Xgal staining procedure.

All of the cell lines stained in the expected manner (Fig.3.7). This reinforces the veracity of the original genotyping performed by RT-PCR. Notably, the *Dgcr8*^{gt1/+} cell line stained slightly less strongly than *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1}. This pattern replicates data gathered in other X-gal staining experiments (data not shown). It may be that some form of *Dgcr8* auto-regulation is disrupted when *Dgcr8* is knocked down within the system, resulting in an up-regulation of the *Dgcr8* promoter. Indeed, a slightly different form of *Dgcr8*-Drosha co-regulation has recently been described that may account for these changes in *Dgcr8* expression level (Han et al., 2009). *Dgcr8* mRNA contains embedded RNA hairpins within the 5'UTR and the 5' end of the ORF. The mechanism proposed by Han *et al.* would predict

that a decrease in DGCR8 protein would cause a destabilisation of Drosha. The loss of microprocessor activity would lead to a reduction in the cleavage of these embedded hairpins and an up-regulation of *Dgcr8* mRNA, and hence fusion proteins, when compared to control cells.

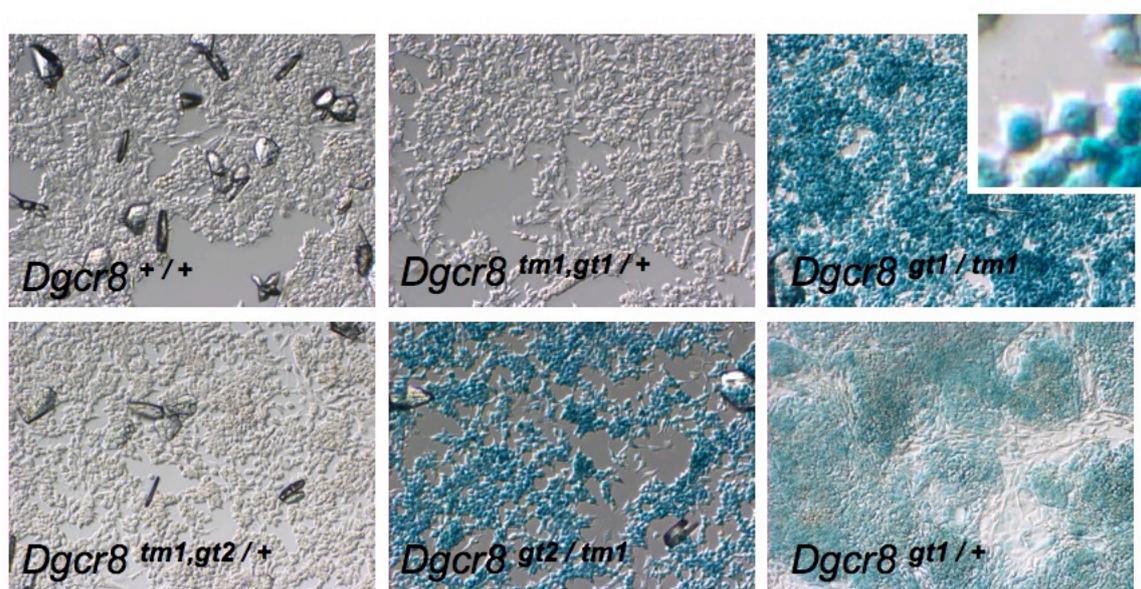


Fig.3.7: Xgal staining of cell lines to determine β -geo (β -galactosidase) activity. *Dgcr8^{gt1/tm1}*, *Dgcr8^{gt2/tm1}*, *Dgcr8^{tm1,gt1/+}*, *Dgcr8^{tm1,gt2/+}* and *Dgcr8^{gt1/+}* (+ Random-insert), were all maintained under hygromycin selection until stained. *Dgcr8^{+/+}* cells were cultured in non-selective media. The insert shows nuclear localisation of β -galactosidase activity of the β -geo fusion protein. Note crystals were formed in some wells during staining. However, these were seen in both control and *Dgcr8^{gt1/tm1}* / *Dgcr8^{gt2/tm1}* wells.

3.4 Characterisation of *Dgcr8* expression and a broader spectrum of miRNAs

3.4.1 Detection of wild type and truncated *Dgcr8* transcripts

Given that the results of the RT-PCR suggest the presence of wild type splicing events in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1}, it is important to use a more quantitative approach to determine whether there is significant reduction in *Dgcr8* wild type transcripts in these cell lines. I designed probes to bind upstream of the duplicated region and downstream of the two trap insertions (411bp and 423bp respectively) within the *Dgcr8* transcript (ENSMUST00000115633 /ENSMUST00000009321) and cloned them into pGEM-T-Easy, confirming them by sequencing. It could be expected that the probes designed to bind upstream should reveal changes in the *Dgcr8* transcript sizes between the cell lines, caused by the insertion of traps into the gene; the resulting splice events truncate the wild type coding sequence and add gene trap sequence in its place. The downstream probes should demonstrate a depletion of the wild type transcripts in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines.

7 cell lines were analysed by Northern blot: *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+} and *Dgcr8*^{+/+} cells were selected. In addition, samples of RNA from the parental gene trap cell lines (*Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+}) were also used. These parental cell lines had been grown under slightly different growth conditions but this issue will be addressed in Chapter 5 (see section 5.3.1).

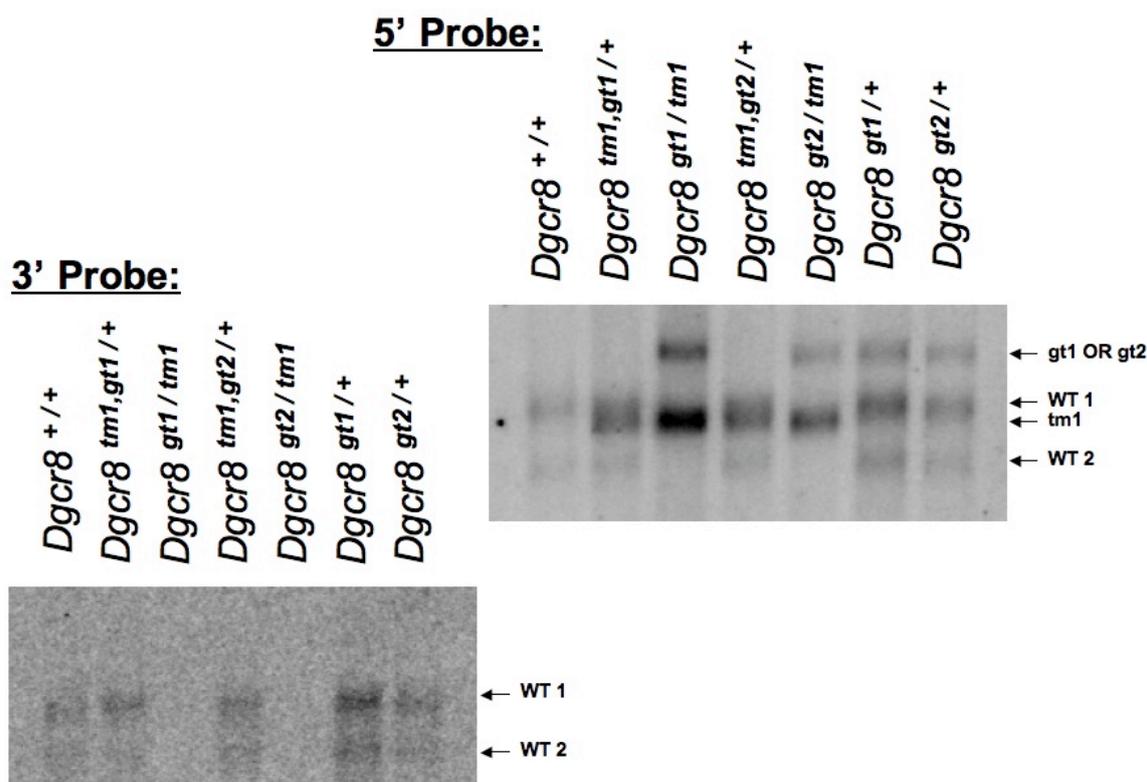


Fig.3.8: Northern blot to demonstrate the loss of wild type *Dgcr8* transcript in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines and the effect of the traps on transcript length. RNA from each cell line was size separated, transferred to a membrane and hybridised with either a radiolabeled-probe specific to the 5' end of the *Dgcr8* transcript (top) or the 3' end (bottom). The expected transcript lengths are: *gt1* and *gt2* - ~6.4kb, *tm1* - ~4.2kb, WT1 - ~4.3kb (ENSMUST0000009321) or 4.5kb (Shiohama et al., 2003), WT2 - ~3.5kb (Shiohama et al., 2003). Each of these transcripts is marked with an arrow on the blots. Note the truncated transcripts are only labeled when the 5' probe is used.

The Northern blot was probed with both the 5' and 3' probe by Dr Claudia Kutter in Dr Duncan Odom's laboratory in the Cambridge Research Institute (CRI). Transcript sizes were consistent with those predicted. Most significant is the lack of wild type transcript in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} lanes when probed with the 3' probe, in contrast to the strong fusion transcripts seen when the 5' probe is used on the same blot (Fig.3.8). All other samples expressed detectable levels of wild type transcript identifiable with both the 5' and 3' probes.

The 5' probe also clearly identifies the transcripts for the gene trap and the targeted trap events in the *Dgcr8*^{gt1/+}, *Dgcr8*^{gt2/+} and *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+} samples respectively. I was unable to identify a clone representing the 3.5kb wild type transcript reported in a previous study (Shiohama et al., 2003). However, Gregory *et al.* identified a carboxyl truncated version of the DGCR8 protein in HEK-293 cells (Gregory et al., 2004). This may be coded for by the smaller transcript, which would imply a 3' alteration to the transcript structure. Such an alteration may explain why the wild type transcripts appear to be of two alternative sizes, while trapped transcripts are of a single length, as both wild type transcripts would share the 5' exons.

It is clear from this blot that there is a significant reduction in wild type *Dgcr8* transcripts in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells when compared to the control cell lines. It follows that there is minimal splicing across the inserted traps to derive these functional transcripts, and demonstrates that the RT-PCR result identifying potential wild type transcripts in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells was not quantitative. Ultimately, this result supports the conclusion that the *Dgcr8* locus has been disrupted by the insertion of traps to a point at which DGCR8 function is impaired, as confirmed by the absence of processed miRNAs.

3.4.2 The expression of ES cell miRNAs

In parallel to establishing the allelic structure of *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1}, *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+}, further miRNA Northern blots were performed for these cells to investigate the functional significance of the allelic alterations in more detail.

As I grew more accustomed to the growth properties of *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells the growth conditions altered slightly from those used for the previous miRNA Northern blot analyses. Hygromycin selection was removed from the cells 2 passages (4 days) prior to Trizol lysis, to allow the cells to recover from the effects of the drug. At the stage at which the cells were plated for RNA they were plated in quantities that allowed sub-confluent cells to be harvested 2 days later from all of the samples. Due to the slower growth rate apparent amongst *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells the quantity of cells plated was higher than in the case of *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+}. At this stage a protocol was established at which *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cells were plated at a higher density than *Dgcr8*^{+/+} cells. Although no significant differences were seen between the properties of the *Dgcr8*^{tm1,gt/+} cells and wild type cells throughout my research, synchronising cell lines that are maintained with or without selection, that grow at different rates and which have different phenotypes makes it unlikely that perfectly matched cell densities between the cell lines used in these experiments is achievable. However, by following these plating ratios the cells reached similar degrees of confluence by the time of lysis so these plating densities were maintained throughout for consistency.

The set of miRNAs to be tested were previously identified as ES cell expressed (miR-292as, miR-293, miR-16, miR-130, miR-21 and miR-92a) (Houbaviy et al., 2003). In addition, let-7a was selected as the let-7 family (Landgraf et al., 2007) is specifically repressed in ES cells but is widely expressed throughout differentiated tissues. The U6 small nuclear RNA (snRNA) probe was used as a loading control for all the Northern blots. (Fig.3.9) The results clearly demonstrate the loss of miRNA expression in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells.

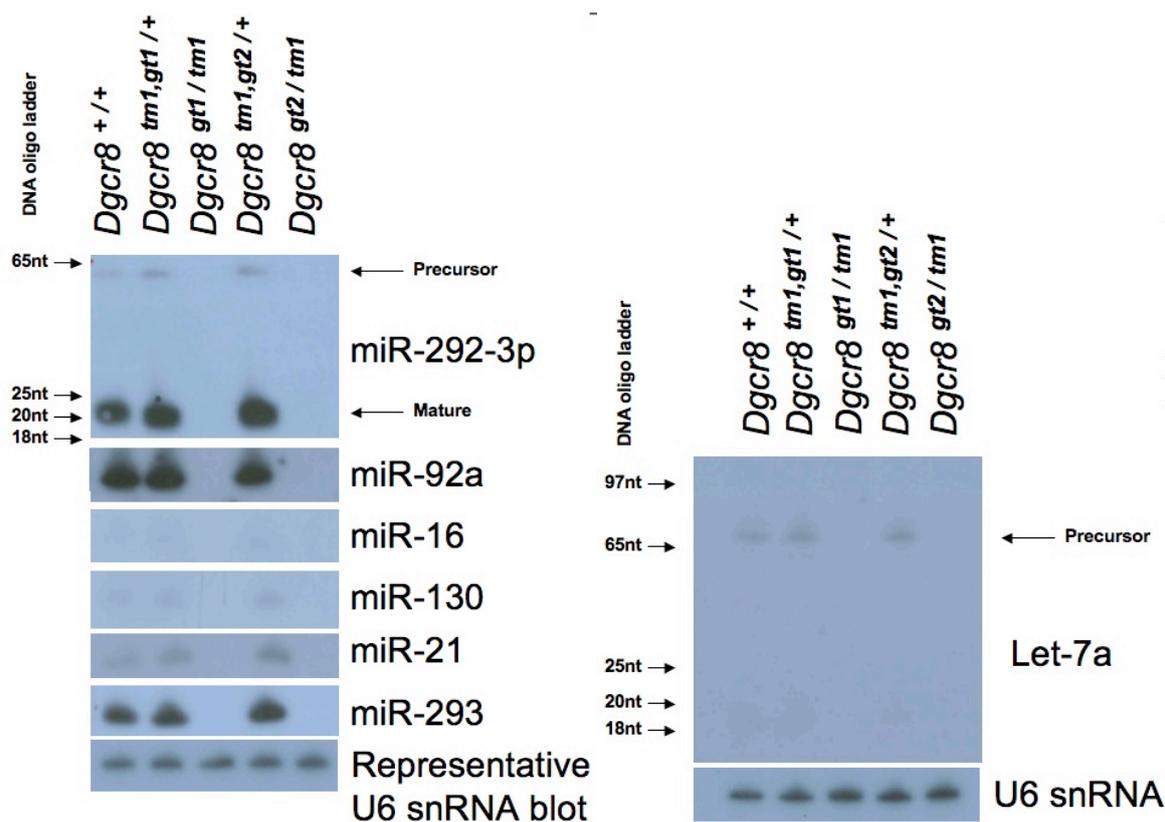


Fig.3.9: miRNA Northern blots demonstrating the loss of ES cell miRNA expression in *Dgcr8*^{gt1/tm} and *Dgcr8*^{gt2/tm1} cells. 12.7µg to 20µg of total RNA from each cell line was size separated per gel and transferred to a membrane before being hybridised with radio-labelled DNA oligos complementary to the miRNA sequences. All samples per gel consisted of the same quantity of RNA. All filters were hybridised with U6 snRNA as a loading control to ensure that all samples were loaded in comparable quantities. The upper blot on each panel is labelled with a ladder to the left. To the right of each panel are the names of each miRNA/snRNA tested and the labelled mature and precursor sequences where appropriate.

An intriguing feature demonstrated by these Northern blots is that the precursor let-7a band is apparent in the wild type, *Dgcr8*^{gt1, gt1/+} and *Dgcr8*^{gt1, gt2/+} samples. This is to be expected, as the let-7 family of miRNAs is expressed in ES cells but remains unprocessed at either the primary transcript or precursor stages (Heo et al., 2008; Newman et al., 2008), due to a let-7 specific block instigated by the Lin-28 protein. These results would seem to be consistent with a block at the stage of precursor processing. Let-7a expression is not evident in the homozygous mutant *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines. As let-7 is a widely expressed

miRNA family, this argues against the possibility that differentiation of the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines may account for the changes seen in miRNA expression when compared to the control cell lines.

3.5 Investigation of ES cell properties in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines

A possible explanation for the lack of ES cell miRNA expression in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines could be that the disruption of the canonical miRNA-processing pathway and the associated protein expression changes could trigger partial differentiation of the ES cells resulting in a change in the population of miRNAs expressed. As a result the expression profile of miRNAs in these cells could be expected to be altered not because of the lack of a functioning DGCR8, but due to the triggering of a new set of cell specific promoters. Taking into account the intention to use these cells for ES cell miRNA target identification, it is preferable that the cells retain ES cell qualities post-electroporation and *Dgcr8* disruption.

3.5.1 Expression of core ES cell transcription factors

3.5.1.1 Oct4 expression measured by Western blot

Stem cells, by definition, are capable of continuous self-renewal and are pluripotent (capable of differentiating into many, specialised cell types). At the centre of the transcriptional network responsible for maintaining this cell state is a set of core TF, including Oct4, *Sox2* and *Nanog* (Avilion et al., 2003; Kopp et al., 2008; Mitsui et al., 2003; Niwa et al., 2000). To investigate whether the expression of a core TF in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines

is altered, protein was purified from *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1}, *Dgcr8*^{tm1,gt1/+} *Dgcr8*^{tm1,gt2/+} and wild type cells and a Western blot was performed for the transcription factor Oct4 (Fig.3.10). The results are consistent with the expectation that *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines maintain their ES cell identity despite the significant reduction of *Dgcr8*, as the expression of Oct4 appears unchanged between cell lines. α -tubulin was used as a loading control.

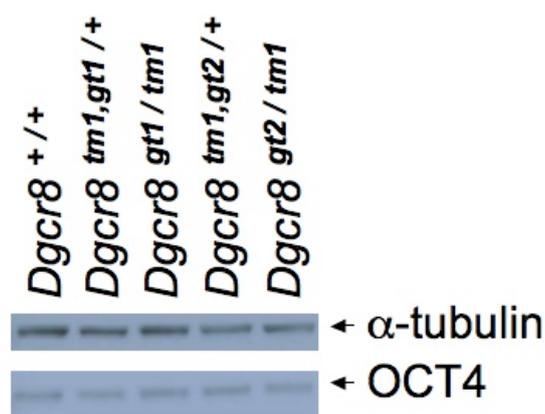


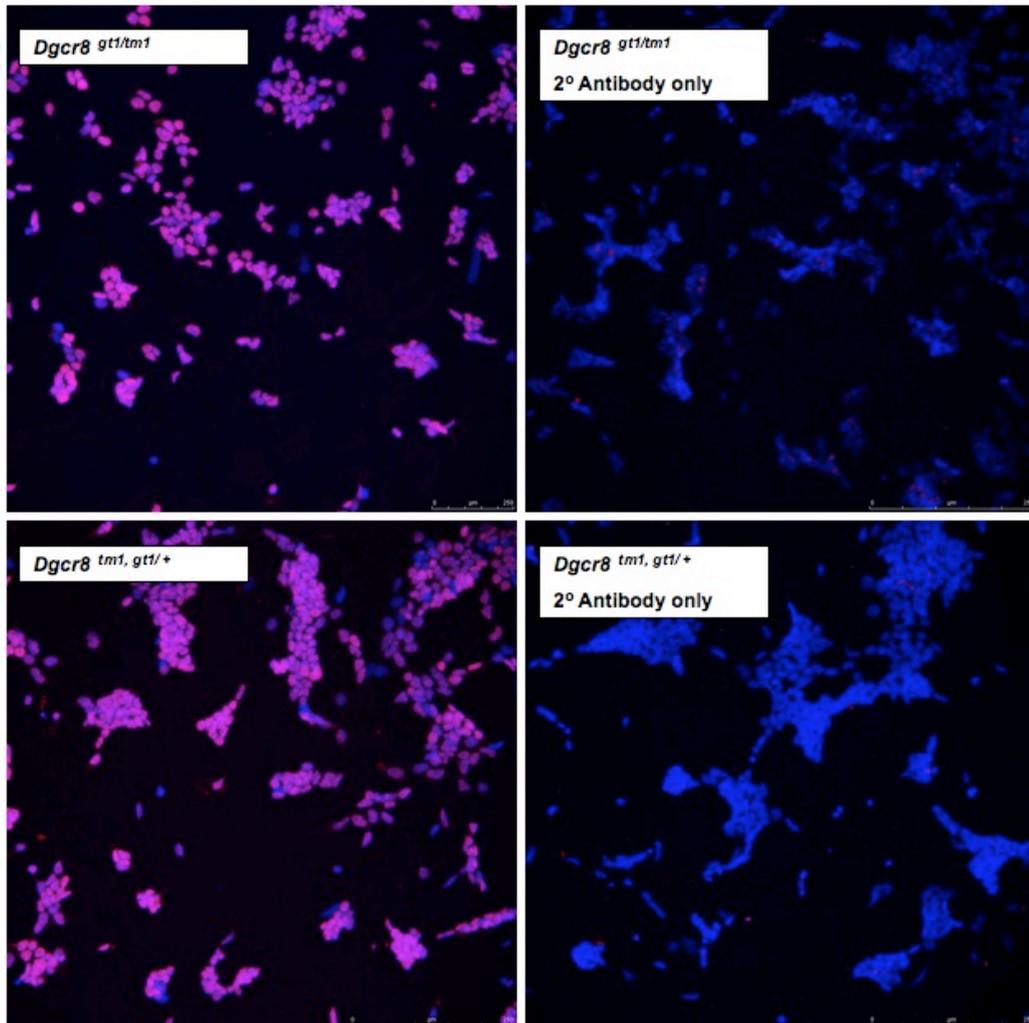
Fig.3.10: Western blot to demonstrate comparable Oct4 expression between cell lines. The same protein blot was probed with both an Oct-4 antibody and an α -tubulin antibody as a loading control.

3.5.1.2 Comparing Oct4 and SOX2 expression by the immuno-staining of cell cultures

The above analysis of the core transcriptional network was extended by culturing *Dgcr8*^{gt1/tm1} and the *Dgcr8*^{tm1,gt1/+} cells on slides and immuno-staining with antibodies to Oct4 and SOX2. One important consideration is that the excitation spectrum of enhanced green fluorescent protein (EGFP), potentially expressed by both cell lines from the targeted trap, overlaps with that of Alexa-488 used to detect the SOX2 antibody. However, earlier experiments with fluorescent microscopy suggested that EGFP is expressed at very low levels in these lines (data not shown). In order to determine if background fluorescence interfered with the antibody specific fluorescence signals, fluorescence sections were taken through control

wells containing cells stained solely with secondary antibodies at the same experimental settings. None of these wells produced significant signals. This would suggest that although EGFP may be expressed from this trap, the level is insufficient to interfere with the results of the cell staining.

Oct4:



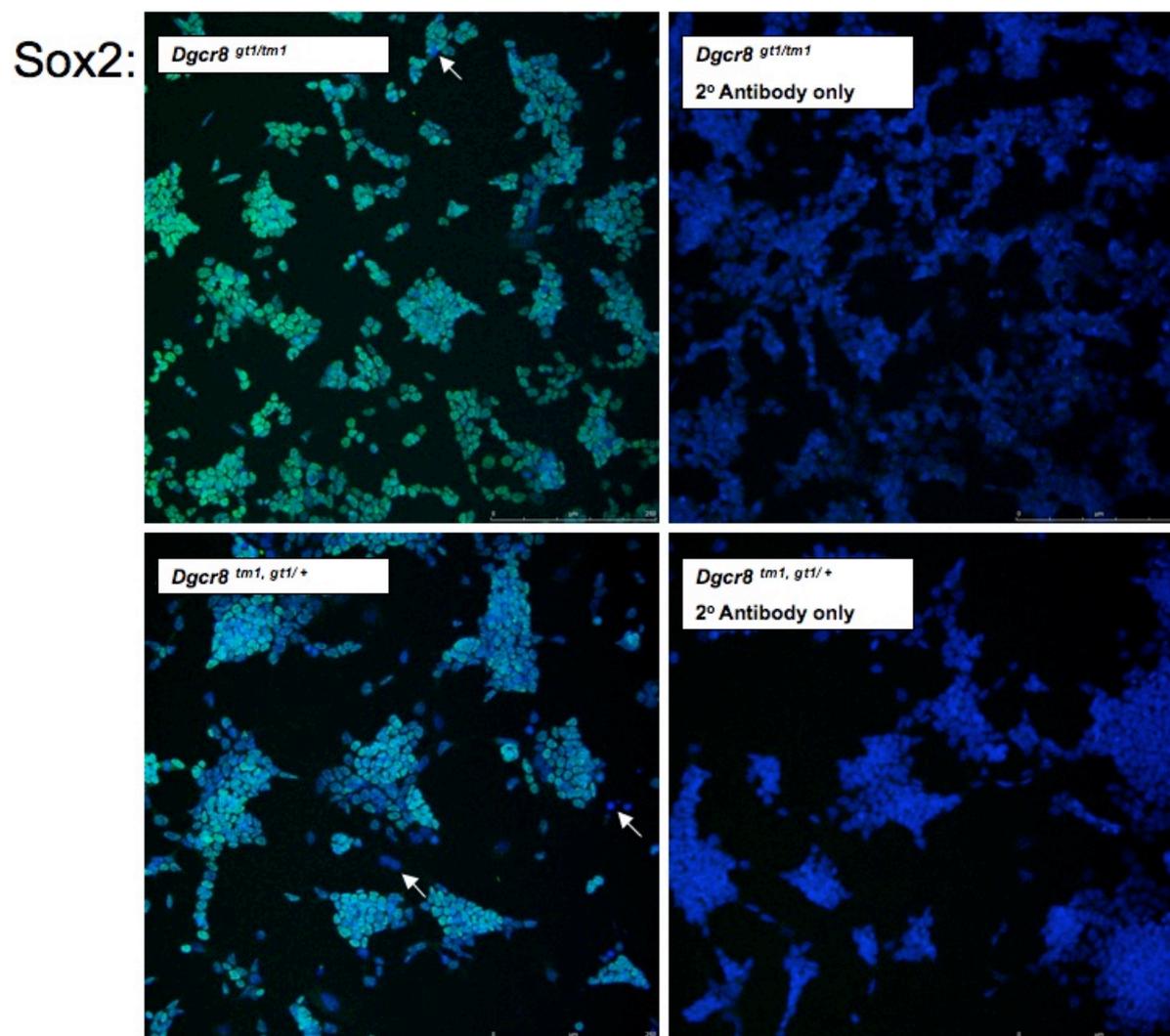


Fig.3.11: Immuno-staining of *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{gt1/tm1} cells to demonstrate consistent expression of both Oct4 and Sox2 within the cultures. Slides of *Dgcr8*^{tm1, gt1/+} and *Dgcr8*^{gt1/tm1} were prepared. Slide wells were stained with primary antibodies specific to Oct4 or SOX2 and Alexa 594 (Oct4, red) or Alexa 488 (SOX2, green) conjugated secondary antibodies. A z-series of ten images was captured for test samples and the maximum intensity at each x,y coordinate was projected to produce a flat image. This image was overlaid with DAPI staining (blue). Control images are a maximum projection of a series of 5 slices taken through any visible fluorescence, again overlaid with DAPI staining. White arrows indicate cells that appear not to express SOX2.

The results suggest that there is not a substantial difference in the expression of Oct4 and Sox2 between the *Dgcr8*^{gt1/tm1} and the *Dgcr8*^{tm1,gt1/+} cell lines, implying that these cells do broadly maintain their ES cell identity as judged by the expression of pluripotency critical transcription factors (Fig.3.11). It should be noted that although both cell lines do appear to

express these proteins, it is very difficult to quantitate the absolute level of expression by this method and hence subtle differences in protein levels cannot be excluded (as may be the case with *Sox2*). The apparent low-level variation could be the result of differential staining in different regions of the well, or residual fluorescence from non-specific antibody that may vary to a small extent between wells. However, one benefit of the use of immuno-staining in this way, rather than by Western blots is that it is possible to see that the expression of these proteins remains consistent throughout the cell populations and that there are not significant populations of differentiated cells within these cultures.

3.5.2 Morphological phenotype of *Dgcr8*^{gt/tm1} cells

During the course of this study, it became apparent that a subset of cells in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cultures had a subtly different morphology to those of the control cell lines (Fig.3.12). This phenotype has either not been seen or not described by previous studies with impaired miRNA processing pathways (Kanellopoulou et al., 2005; Murchison et al., 2005; Sinkkonen et al., 2008; Wang et al., 2007). However, in some of these studies ES cell cultures were maintained on a supportive layer of mouse embryonic fibroblasts (MEFs) (Wang et al., 2007). Indeed one study noted that these supportive cells improved the phenotypes they saw upon the disruption of the miRNA processing pathway (Murchison et al., 2005). Nevertheless, I continued to culture the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells in the absence of MEFs as neither the *Dgcr8*^{gt1/tm1} or *Dgcr8*^{gt2/tm1} cells showed the loss of ES cell marker expression after an extended period in culture. In addition, it was my intention to use these cells for the generation of ES cell specific miRNA target lists. For these experiments, a MEF free culture would be desirable to obtain the most specific signal possible. In the future, it would be interesting to culture these cells with MEFs in order to see if they would

ameliorate this morphological phenotype. ES cells lines cultured for extended periods and mice of different strain backgrounds can also harbour copy number variations or other mutational changes that may cause phenotypic differences. These effects could account for differing phenotypes and changes in expression between different knockout cell lines (Liang et al., 2008; Sibilina and Wagner, 1995).

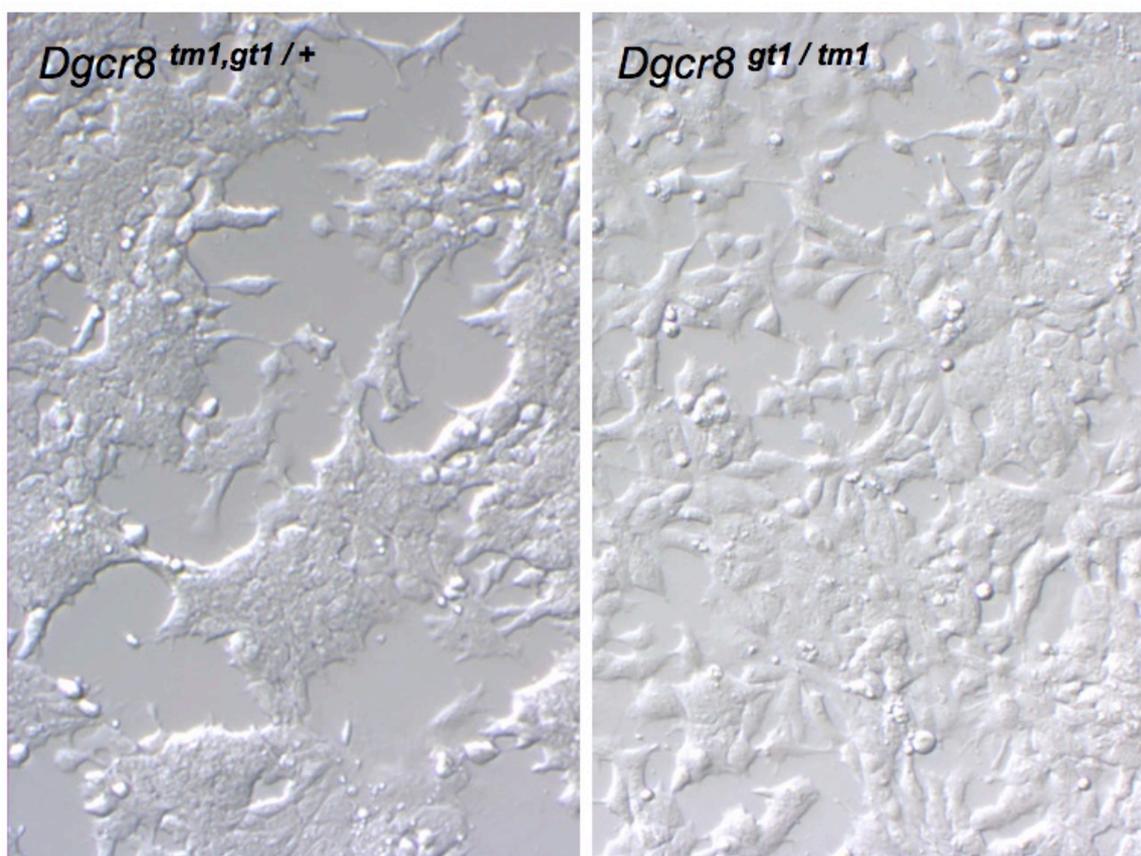


Fig.3.12: An example of the morphological phenotype seen among *Dgcr8*^{gt1/tm1} cells when compared to *Dgcr8*^{tm1,gt1/+} controls. Cells were cultured for 4 days with no selection prior to images being captured (10x relief contrast). A proportion of the *Dgcr8*^{gt1/tm1} cells have a broader and slightly flatter morphology than the *Dgcr8*^{tm1,gt1/+} cells, while the *Dgcr8*^{tm1,gt1/+} cell morphology resembles that of wild type cells.

3.5.3 Investigation of *Dgcr8*^{gt1/tm1} and *Dgcr8*^{tm1,gt1/+} cell pluripotency

Differentiation via embryoid bodies (EBs) is an *in vitro* technique used to simulate the early embryonic differentiation of embryonic stem cells (Desbaillets et al., 2000). Within the balls

of cells that develop, ES cells differentiate into cells derived from all three embryonic germ layers.

Upon plating into a non-tissue culture treated, round bottom plate, in the absence of LIF, wild type cells aggregate into balls of cells. When transferred into low attachment plates these EBs will continue to grow and after 7-8 days they will begin to beat and contract due to the development of cardiomyocytes within these aggregates. When transferred to gelatinised plates after 8 days, the embryoid bodies attach and begin to spread across the plate, revealing a myriad of morphologically distinct cell types. Patches of these cells continue to beat, again betraying their lineage. The addition of retinoic acid (RA) can induce a greater proportion of cells within the aggregates to follow the neuronal lineage. EBs cultured with RA at the correct stage of differentiation and in the correct concentration will no longer beat and a large number of neuritic outgrowths are evident when the aggregates spread on gelatinised plates (Guan et al., 2001).

I attempted to optimise the technique for *Dgcr8^{gt1/tm1}* differentiation. First I plated a cell number equal to and double that used to differentiate wild type cells in order to compensate for their slower rate of growth. It rapidly became apparent that the *Dgcr8^{gt1/tm1}* cells formed EBs less successfully than wild type cells. At the time of transfer to 24 wells, the *Dgcr8^{gt1/tm1}* EBs had formed as multiple foci in each well as opposed to forming a single large EB. Each EB was tiny and was not spherical. The phenotype was worse in EBs formed from 1×10^3 cells and these EBs were discarded. By day 8 the EBs formed from 2×10^3 cells had begun to disintegrate. When transferred to standard wells the mutant EBs did not stick to the plates. Next, I increased the number of *Dgcr8^{gt1/tm1}* cells added to each 96 well to 3×10^3 and 4×10^3

cells, reasoning that the formation of tiny EBs may be due to a growth deficit in the *Dgcr8*^{gt1/tm1} cells. These EBs still seemed to form as multiple foci in each well. At the stage of transfer the EBs formed from 3x10³ cells were still small with a few rare examples of reasonably sized EBs. The 4x10³ wells seemed to follow the same trend of improved phenotype with increased cell number, but still formed as multiple foci of various sizes. The *Dgcr8*^{gt1/tm1} EBs were still not spherical and were “hairy”. By day 8 the RA treated EBs appeared to be less coherent than those without RA. By day 12 a very small proportion of the EBs formed from 4x10⁴ cells had begun to stick to the well and spread. The EBs without RA had spread to a greater extent as is seen with wild type EBs. No contraction of *Dgcr8*^{gt1/tm1} cells without RA was seen. The majority of the remaining EBs were loosely attached, although some had attached more firmly than others. Clearly this was a limited set of experiments and the specific phenotypes relating to –LIF/+RA and –LIF EBs would have to be followed up by a far more comprehensive study in order to define them more confidently.

Following the optimisation of the *Dgcr8*^{gt1/tm1} differentiation and with a slight alteration in the protocol by which the cells were spun out of trypsin and resuspended in media without LIF prior to plating to 96 wells, in order to remove any traces of LIF in the cultured cell media, *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{gt1/tm1} cells were all differentiated via EBs, both with and without RA. Both the control cell lines to which I compared to the *Dgcr8*^{gt1/tm1} cells (*Dgcr8*^{+/+} and *Dgcr8*^{tm1,gt1/+}) were plated at densities of 1x10³ cells per 96 well. These control cell lines developed as expected. Following the removal of LIF, a large proportion of beating EBs were observed, indicating the presence of differentiated cardiac muscle. This differentiation pathway was broadly negated by the addition of RA and once plated these EBs spread to reveal a greater presence of tight bundles of neuronal cells (Table 3.4). The single beating

colony amongst those of RA treated *Dgcr8^{tm1,gt1/+}* EBs suggests that future experiments should be repeated with a higher concentration of RA to achieve more significant commitment to neuronal development to the detriment of mesodermal development. The *Dgcr8^{gt1/tm1}* cells were plated at a density of 4×10^3 cells per 96 well as this was the concentration that resulted in the largest proportion of EBs spreading on the standard gelatinised tissue culture plates once the EBs were plated. Once again the *Dgcr8^{gt1/tm1}* cells formed as multiple foci of many varying sizes within the wells of the 96 well plates. A rough average of the number of foci per well, taken across ten 96 wells, was 26, as opposed to the single foci seen in the case of the wild type and *Dgcr8^{tm1,gt1/+}* cells.

<i>Dgcr8^{+/+}</i>		
Conditions	Beating EBs	Total EBs
-LIF	18	30
-LIF +RA	0	36

<i>Dgcr8^{tm1,gt1/+}</i>		
Conditions	Beating EBs	Total EBs
-LIF	26	29
-LIF +RA	1	38

<i>Dgcr8^{gt1/tm1}</i>		
Conditions	Loose EBs	Attached EBs
-LIF	~361	~12.75
-LIF +RA	~300	~3.25
	Average of 2 wells	Average of 4 wells

Table 3.4: Summary of the phenotypes seen amongst EB colonies after 12 days of culture in the absence of LIF. The vastly increased number of smaller EBs in the case of the *Dgcr8^{gt1/tm1}* cells caused me to only count the number of EBs in a sample of wells, whereas all the control EBs plated were considered. All of the *Dgcr8^{+/+}* and *Dgcr8^{tm1,gt1/+}* EBs attached to the plates successfully, while none of the *Dgcr8^{gt1/tm1}* EBs began to beat as cells differentiated into cardiomyocytes.

Once plated to low attachment plates, there was evidence of EBs fusing both amongst the control EBs and the *Dgcr8^{gt1/tm1}* EBs as expected. By day 8, the control cells had formed large round EBs (Fig.3.13). There was clear evidence of twitching cardiac muscle cells amongst

the *Dgcr8*^{+/+} and *Dgcr8*^{tm1,gt1/+} EBs without LIF or RA. The *Dgcr8*^{gt1/tm1} EBs were much smaller and poorly defined (Fig.3.13). In some cases the balls of cells were so small that it was difficult to discern the EBs from cellular debris that had accumulated in the wells.

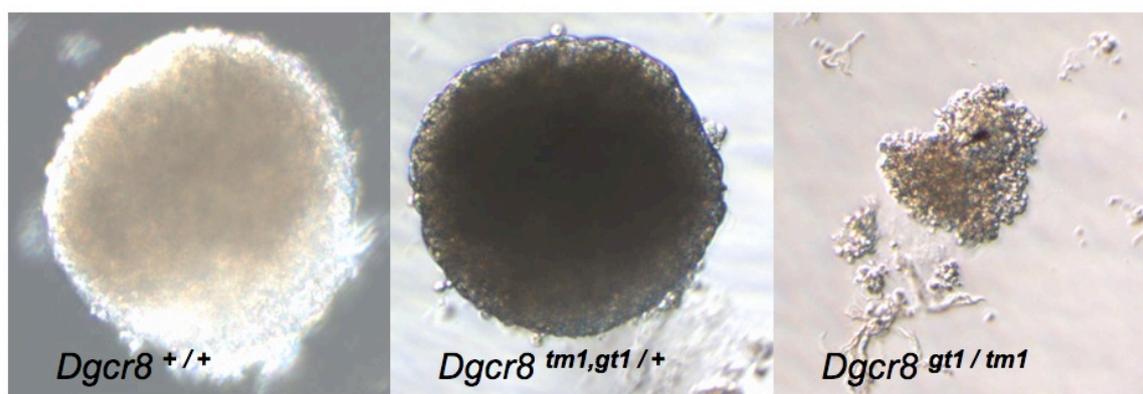


Fig.3.13: Images of EBs cultured on low attachment plates after 8 days of culture in the absence of LIF. All three images are taken at 10x relief contrast.

At day 12, the EBs of each cell line were counted and loosely phenotyped (Table 3.4). The attached and spread EBs of all cell lines were also fixed and representative images taken (Fig.3.14).

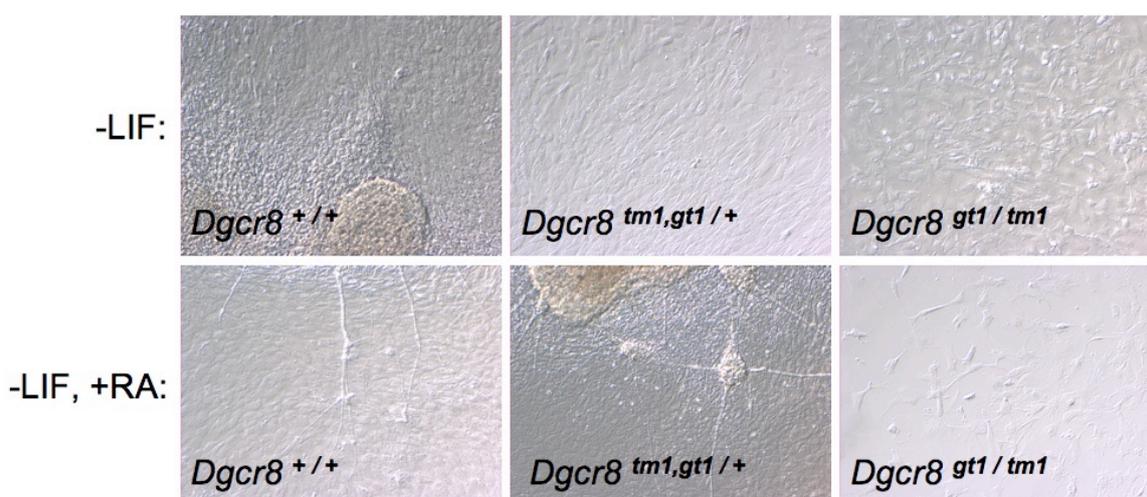


Fig.3.14: Representative images of cell morphologies seen amongst EBs spreading on gelatinised tissue culture plates. Examples are presented of EBs cultured in both the presence and absence of RA. All images are taken at 10x relief contrast. Due to the size difference between the *Dgcr8*^{gt1/tm1} EBs and the control EBs, the spread *Dgcr8*^{gt1/tm1} cells are more sparsely packed on the gelatinised plates.

It is not possible to ascribe a cause to the phenotype seen in the case of the *Dgcr8*^{gt1/tm1} cells. The number of pathways involved in differentiation is vast and any or all of these pathways could be affected in some way by the reduced levels of *Dgcr8* and miRNAs within the cells. Contributing factors could range from deficiencies in cell adhesion to a reduced rate of cellular proliferation. *Dgcr8*^{gt1/tm1} cells are clearly incapable of forming fully functional EBs. They do, however, seem capable of differentiating into a range of morphologically different cells.

3.5.4 Flow sort for cell cycle

It has been noted that knocking out both *Dicer* and *Dgcr8* in mouse embryonic stem cells leads to a slower cell cycle and an accumulation of cells in the G1 phase (Murchison et al., 2005; Wang et al., 2007). *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells appeared to grow more slowly than *Dgcr8*^{gtm1,gt1/+} and *Dgcr8*^{gtm1,gt2/+} cells. I therefore performed a cell cycle analysis by fluorescence activated cell sorting (FACS) to ascertain whether *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells were progressing through the cell cycle in the normal fashion.

Cells were grown for 2 days without selection and then plated in the same quantities as had been previously used for RNA lysis, into non-selective media. *Dgcr8*^{gt1/tm1} cells were plated at two quantities. Throughout it had been difficult to ensure *Dgcr8*^{gtm1,gt1/+} and *Dgcr8*^{gtm1,gt2/+} cells and *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells were equally confluent after a further 2 days in culture. As confluence may have an effect on the cell cycle experiments, I decided to plate more than one cell quantity for this analysis in the case of *Dgcr8*^{gt1/tm1}. I would then be able to compare the results to see if cell density had a profound effect on cell cycle distribution. For future analysis I would first generate a growth curve for each cell line to be used, to be

sure that the cell cycle profile for each cell line is generated while the cells are in the exponential phase of growth.

Despite these caveats, this experiment does concur with previous cell cycle analyses conducted on Dicer and *Dgcr8* knockout cell lines, with *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells accumulating in G1 phase (34-37%) compared to *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+} and *Dgcr8*^{+/+} cells (18-22%) (Fig.3.15). It is also clear that the increase in cell density from 304 x 10⁴ cells plated to 380 x 10⁴ cells had a minimal effect on the cell cycle profile of the *Dgcr8*^{gt1/tm1} cells. These results along with those of previously published studies concur that miRNAs play an important role in the regulation of the cell cycle in mouse ES cells, and more specifically in the regulation of the G1 to S phase transition.

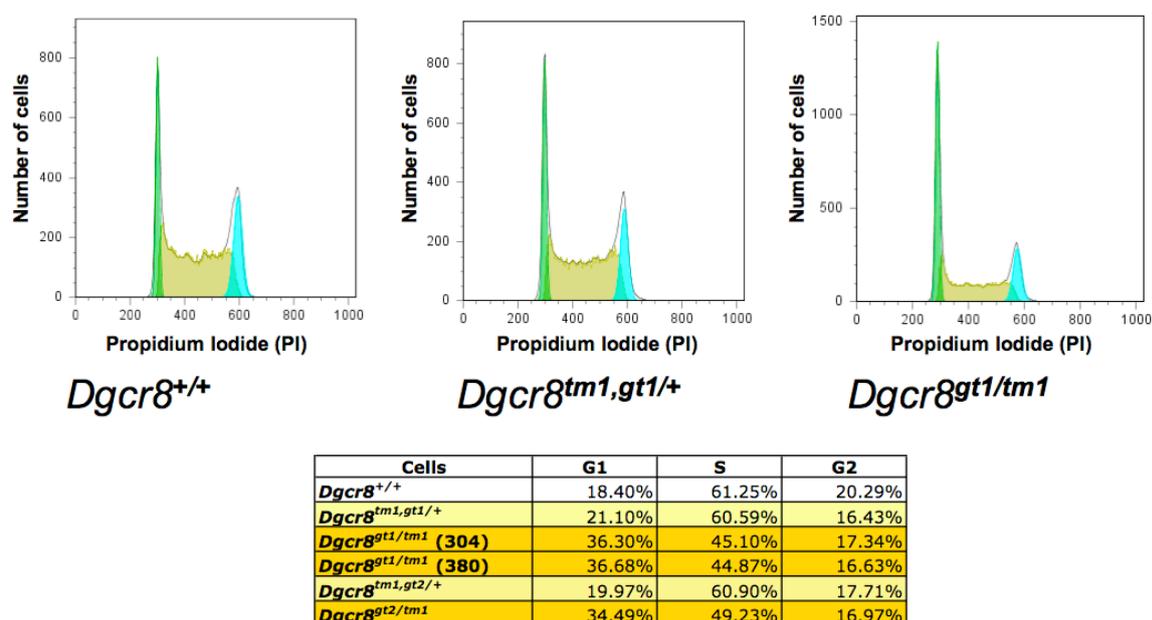


Fig.3.15: ES cells depleted in the expression of *Dgcr8* and mature miRNAs accumulate in the G1 phase of the cell cycle. The upper three panels are a representative sample of the FlowJo cell cycle histograms. The x-axis is a linear measure of the propidium iodide fluorescence associated with each cell, used to quantify cell DNA content. The green, beige and blue segments represent the portion of cells in the G1, S and G2 phases respectively. The lower panel provides a summary of the proportion of cells from each cell line in each phase of the cell cycle.

3.6 Discussion

Using a targeted trap to target the second allele in heterozygous gene trapped cell lines, I derived multiple cell lines in which both alleles of the *Dgcr8* locus were disrupted. I confirmed the genotypes of these cell lines by RT-PCR to amplify regions from trap specific transcripts, by X-gal staining to reveal the expression of trap specific markers and by Northern blot identifying the presence of trap disrupted transcripts specific to the *Dgcr8* locus. The Northern blot also revealed a severe reduction in wild type *Dgcr8* transcript levels in these cells.

Subsequently, despite subtle changes in morphology and an altered cell cycle profile in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} lines, I confirmed that they maintained their expression of essential ES cell associated TFs. The *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells express Oct4 at a comparable level not only to *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cells but also to wild type cell lines, as assessed by Western blot. In addition, both Oct4 and *Sox2* expression is maintained throughout the culture, as seen with the immuno-stained cultures.

Although it is possible to maintain the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells in culture for extended periods without the loss of ES cell identity, there are fundamental alterations in ES cell phenotype of *Dgcr8* homozygous mutant cells. Most significant is the loss of ES cell miRNA expression, presumably due to reduced microprocessor activity. Through the use of miRNA Northern blots I demonstrated the loss of miR-292-3p, miR-293, miR-130, miR-21 and miR-92a expression in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells. miRNAs are expected to regulate a large proportion of the mouse transcriptome (Friedman et al., 2009) and the loss of a large number of miRNAs from the system could lead to considerable alterations in the

cellular transcriptional profile. It is not surprising therefore that there are some fundamental changes in both the cell cycle and EB formation, when these cells are compared to *Dgcr8^{tm1,gt1/+}*, *Dgcr8^{tm1,gt2/+}* and *Dgcr8^{+/+}* cells. EBs from *Dgcr8^{gt1/tm1}* cells were smaller and less coherent than *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{+/+}* aggregates. *Dgcr8^{gt1/tm1}* EBs do not adhere to gelatinised plates consistently although when they can adhere and spread, a variety of differentiated cellular morphologies are identifiable. In order to interpret the inability to form healthy EBs and understand the mechanistic implications, considerable further work would be required to identify candidate genes affected at the core of this process. This has recently been achieved by Sinkkonen *et al.* who demonstrated the role of the miR-290 cluster in the regulation of the methylation of the Oct4 promoter during the course of differentiation (Sinkkonen *et al.*, 2008). The re-addition of this cluster to differentiating Dicer deficient ES cells went some way to rescuing the Oct4 transcriptional silencing defect identified in these cells.

In order to continue this avenue of research, a wide range of experiments could be attempted. Initially lineage marker immuno-staining would provide an insight into the exact lineages to which the *Dgcr8^{gt1/tm1}* cells are capable of contributing. This study should be inclusive of both EBs that adhere to plates and those that remain loose after 12 days of culture. It may also be informative to differentiate *Dgcr8^{gt1/tm1}* cells as a monolayer, removing the more complex aspects of EB differentiation. In parallel, an *in vivo* analysis of *Dgcr8^{gt1/tm1}* cell contribution to lineages in a mouse embryo should be considered. The β -geo marker expressed by the gene trap in *Dgcr8^{gt1/tm1}* cells would allow these cells to be traced following microinjection into blastocysts. The presence of a β -geo marker is a distinct advantage over previous microinjection studies of similar cell lines (Kanellopoulou *et al.*, 2005). These experiments

would lead to a more complete understanding of the developmental potential of mutant *Dgcr8* cells.

In addition, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells seem to accumulate in the G1 phase of the cell cycle, 34-37% compared to 18-22% for control cells. Members of the miR-290 cluster have been shown to regulate the cell cycle inhibitor p21 (*Cdkn1a*) and appear to play an active role in increasing the rate of proliferation when reintroduced into *Dgcr8* knock out mouse ES cells (Wang et al., 2008). Other miRNAs have also been found to regulate the cell cycle (Liu et al., 2008a; Petrocca et al., 2008a). It seems that a number of miRNAs may converge in the regulation of the cell cycle in mouse ES cells, thus piecing together the exact nature of the interplay between these factors will require considerable future effort.

The phenotype of the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells is broadly consistent with previous attempts to disrupt the miRNA-processing pathway in mouse ES cells (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007). I have confirmed the disruption of the *Dgcr8* locus, the loss of ES cell mature miRNAs from the system and the continued expression of stem cell markers. I therefore conclude that I have derived a suitable system into which I can reintroduce miRNAs to determine mouse ES cell specific miRNA targets, in a background within which endogenous miRNA maturation will be depleted.

Chapter 4: The miRNA expression profile of ES cells depleted of functional *Dgcr8*

4.1 Aim

Having derived two ES cell lines depleted in wild type *Dgcr8* transcript through the insertion of two traps into the *Dgcr8* locus, I wished to demonstrate the abrogation of *Dgcr8* function in these cell lines. In order to do this I conducted a series of experiments to ascertain the extent of the effect of these mutations on the miRNA-processing pathway, by measuring mature miRNA expression in these cell lines and controls.

4.2 Introduction

It is widely accepted and logically expected that the disruption of the miRNA-processing pathway will lead to the broad depletion of mature miRNAs. In the last few years a number of studies have knocked out both *Dicer* and *Dgcr8* in mouse embryonic stem cells (Calabrese et al., 2007; Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007). The miRNA expression profiles of cell lines with either a disrupted *Dicer* locus or a disrupted *Dgcr8* locus have been determined and both cell types are depleted for the vast majority of miRNAs confirming that the successful targeting of these two loci will cause a major reorganization of the miRNA expression profile (Babiarz et al., 2008; Calabrese et al., 2007; Wang et al., 2007).

The preferred methods with which to measure broad miRNA expression have changed in the recent past. miRNA microarrays have been designed with a wide

selection of miRNA complementary probes (Babak et al., 2004; Barad et al., 2004; Miska et al., 2004). It is clear, however, that miRNA microarrays are subject to a variety of flaws relating to the short length of the miRNA sequence against which to design specific probes. As a consequence there is little choice when it comes to designing probes for a microarray platform with a uniform melting temperature (T_m) across the complete array, thus probe to probe T_m varies and miRNA microarrays have suffered from problems regarding miRNA target specificity (Miska et al., 2004). To some extent these problems have been reconciled by the advent of locked nucleic acid (LNA) technology. LNAs are nucleic acid analogues, which can be used to increase the thermostability of duplexes via their inclusion within oligonucleotides. It is thus possible to generate arrays upon which probes possess a uniform T_m . Ultimately these arrays can be hybridised with confidence at a high temperature without compromising the sensitivity or specificity of any of the miRNA probes (Castoldi et al., 2006).

An alternative approach to improving miRNA expression profiling was demonstrated by Lu *et al.* (Lu et al., 2005). The authors used a bead-based miRNA profiling system, in which miRNA probes are attached to multicoloured beads, to allow hybridization to be conducted in an environment that “might more closely approximate hybridization in solution”. The authors supposed that this might improve the specificity of hybridization of the miRNAs to the probes. Following hybridization of fluorescently labeled, amplified miRNA libraries to the beads, they were flow sorted and the bead colour (associated with a particular miRNA specific probe) and the bead fluorescence (dependent upon the quantity of hybridized miRNA) was used as an indication of the expression level of the miRNA. The authors demonstrated this

increase in accuracy by comparing the specificity of synthetic miRNA detection in both a bead and array format.

The advent of high-throughput sequencing methods now allows the small RNA content of cells to be profiled with extremely high sensitivity and without the requirement of pre-designed, complementary probes. This allows the identification of not only known miRNAs and small RNAs but also the sequences of previously unknown small RNAs. As a consequence this technology is being rapidly applied to the study of miRNAs. The small-RNA complement of both Dicer and *Dgcr8* knockout mouse ES cells have been sequenced using this method and compared to the profiles of wild type and heterozygous cell lines (Babiarz et al., 2008; Calabrese et al., 2007). As a consequence of these efforts, accurate and deeply sequenced mouse ES cell miRNA profiles have been generated. In addition a number of DGCR8-independent miRNAs have been identified and a series of Dicer-dependent small RNAs that were previously not recorded in ES cells, including endogenous siRNAs.

In this chapter, I describe the use of both bead-based miRNA profiling and high-throughput Illumina/Solexa sequencing to determine the miRNA expression profiles of *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells. I demonstrate the broad depletion of mature miRNAs in cells with a trap inserted into each allele of *Dgcr8*, while the cells with a single trapped allele and the wild type cells have comparable miRNA expression profiles. This demonstrates the functional significance of the traps at the *Dgcr8* locus. As a consequence of the sequencing experiment, I also confirm the identity of a number of DGCR8 independent miRNAs

previously described by Babiarz *et al.* in addition to proposing a number of other candidates.

4.3 Results

4.3.1 Use of the Luminex platform to profile mature miRNA expression

Initially the miRNA expression profile of each cell line (*Dgcr8*^{+/+}, *Dgcr8*^{tm1.gt1/+}, *Dgcr8*^{tm1.gt2/+}, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1}) was determined using a Luminex bead-based miRNA profiling system (Blenkiron *et al.*, 2007; Lu *et al.*, 2005). Briefly, each RNA sample is spiked with a predefined set of control probes (pre-controls). This allows the efficiency of the miRNA cloning reactions to be determined. Subsequently, the miRNAs and controls are purified and cloned in a series of ligation reactions to adapter oligos. The samples are then reverse transcribed, and PCR amplified using biotinylated primers. Finally, each cloned sample is spiked with a further set of control oligos. These post-controls allow an assessment of whether the subsequent hybridization is successful. Each sample is hybridised to beads bearing miRNA specific oligo probes. The bead sets are exposed to streptavidin-phycoerythrin and flow sorted. The bead sets contain beads with a myriad of colours. Each colour corresponds to a miRNA specific probe. Therefore by measuring the colour and fluorescence of each bead it is possible to determine the relative expression of each miRNA between samples. In order to maximize the number of miRNAs that can be profiled in a single experiment, given the limited coloured bead selection available, the samples are hybridized to four bead sets. Each set contains a different set of probes corresponding to different miRNAs associated with each of the beads. Spiked n controls are profiled by all 4 bead sets. The flow sorter returns a set of raw median

fluorescence intensity (MFI) values, each corresponding to a specific miRNA or control.

I conducted the labeling and hybridization reactions on two separate occasions. Initially, I grew up a single flask of each cell line to be tested, purified its RNA and labeled and hybridized these samples as two technical replicates simultaneously (referred to as replicates 1A and 1B below). Subsequently, I grew up two new sets of cells for each cell line, simultaneously (replicates 2 and 3). I purified the RNA from these fresh samples and labeled and hybridized these samples at the same time. Upon examining the results obtained from these two separate experiments it is clear that there are some fundamental differences seen between the two sets of samples (Fig.4.1).

The most striking difference is seen in the levels of the pre-control probe MFIs both between the two sample sets and within each set. It is clear that for at least 5 of the samples prepared for replicates 2 and 3 there is a loss of the spiked-in pre-control signal, this is despite an approximately constant signal from each of the post-controls. This would suggest that a significant proportion of the miRNA fraction from these samples was lost in the sample preparation process. In contrast, samples from replicates 1A and 1B all retain approximately equal levels of pre-control signal, implying a roughly equal retention of the miRNA population between samples. There is also a discrepancy in the signal gleaned from each of the individual pre-controls and post-controls in the two independent experiments. When a comparison is made between each of the controls relative to the other controls it seems that there may be a difference in the quantity of each individual spiked in pre-control and post-control on

the two separate occasions. As the pre-controls were to be used to normalise the miRNA signal between the samples, this discrepancy makes it difficult to combine the two experiments into a single analysis.

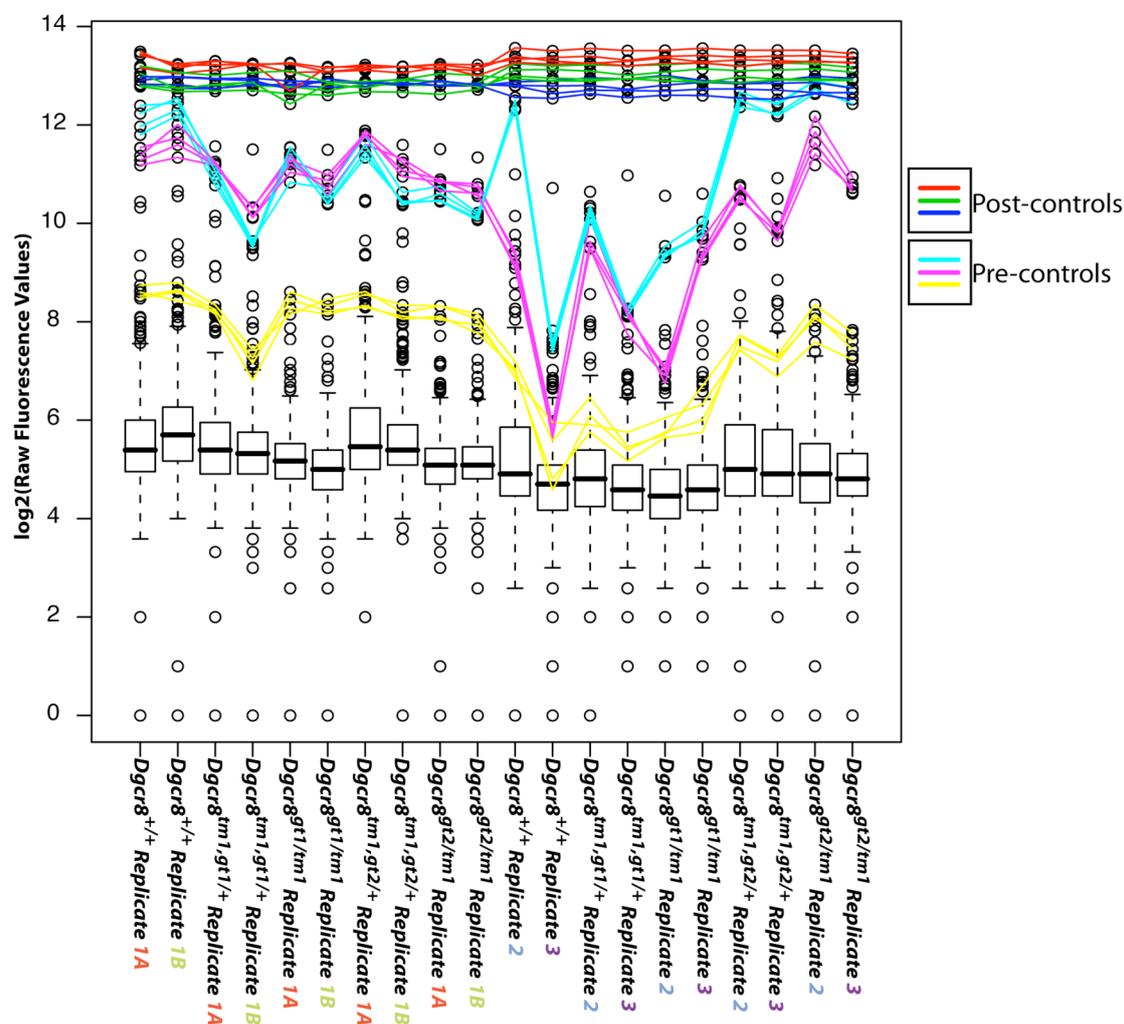
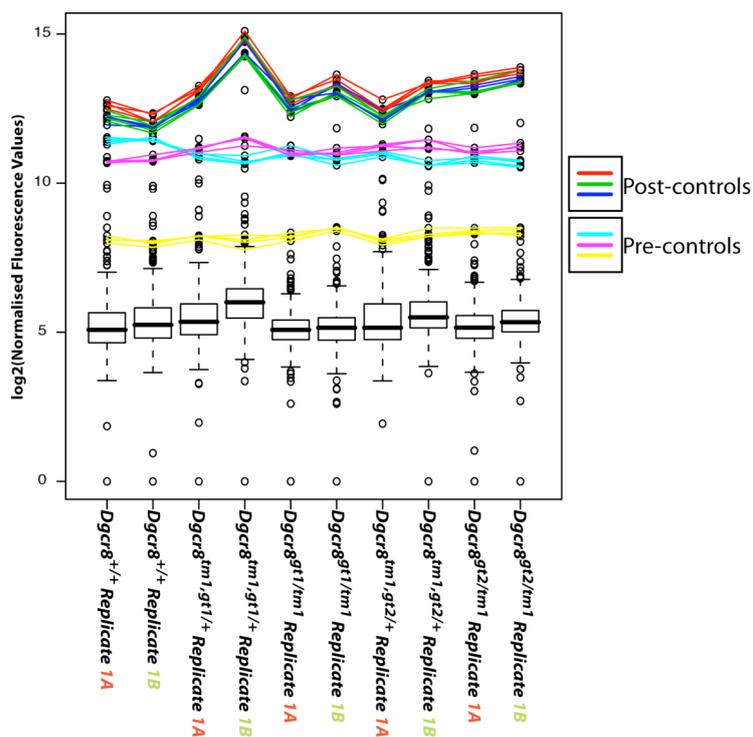


Fig.4.1: A box plot of the raw fluorescence values for each miRNA specific bead derived from the Luminex miRNA profiling system across multiple samples. The mature miRNA expression profiles of the *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/+} and *Dgcr8*^{gt2/+} cell lines were measured using the Luminex bead-based system. Replicates 1A and 1B were measured on the same day from the same RNA samples. Replicates 2 and 3 were grown in parallel and RNA was analysed on a second day. Coloured lines represent the fluorescence values for each of the spiked control sequences as indicated (see section 2.10.2.1).

As the initial experiment contained a more complete set of efficiently cloned samples, it was the samples from this experiment that were selected for further analysis. I attempted to normalise replicates 2 and 3 according to the levels of the pre-controls in order to make use of these samples. However, it seems that in the cases where the pre-control levels appear to be relatively depleted, normalising the bulk of the data according to these spiked in oligos leads to the distortion of the sample expression signals (Fig.4.2B). This is seen most clearly in the case of the *Dgcr8*^{+/+} replicate 3. Following normalization, the bulk of the fluorescence values for this sample are much higher than those of the other, more successfully cloned samples. As a consequence of this distortion of signal I chose to exclude these replicates from further analysis.

A



B

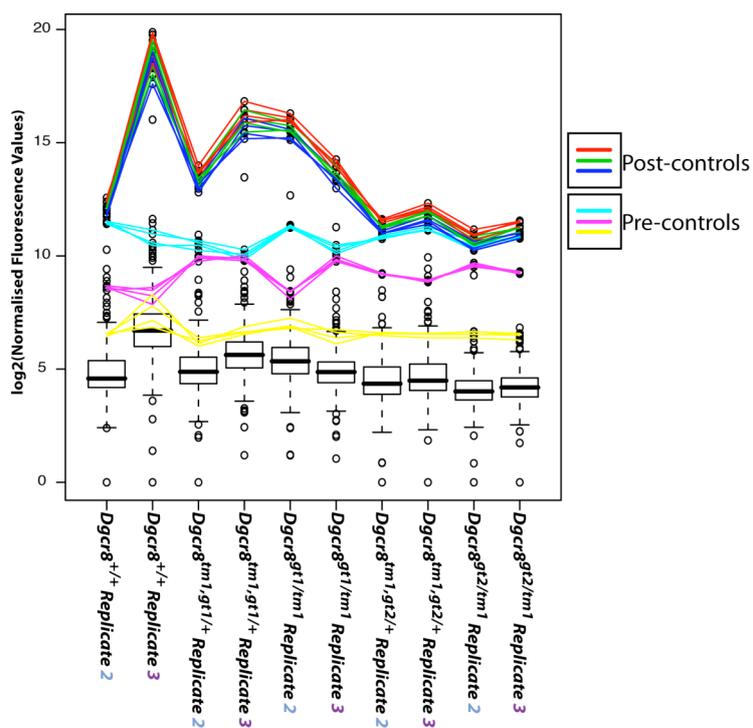


Fig.4.2 Box plots of bead/miRNA-associated MFIs following normalization. Replicate sets 1A and 1B (Fig.4.2A) and sets 2 and 3 (Fig.4.2B) were normalised as separate groups. Samples were

normalised according to the pre-control fluorescence levels in each sample. Coloured lines represent the fluorescence values of spiked in pre and post-controls.

A comparison of *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} miRNA expression and the expression values of the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells (Fig 4.3) demonstrates a broad depletion of a large subset of miRNAs. For this subset, the MFI associated with each miRNA specific probe demonstrates a negative log fold change in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells compared to the control cell lines. Among this subset are a large number of miRNAs known to be expressed in wild type mouse ES cells (Houbaviy et al., 2003). These are miRNAs for which a clear reduction of bead fluorescence would be expected in this comparison given the reduced *Dgcr8* expression.

However, it seems that the fluorescence detected by other miRNA specific beads remains relatively constant between the cells with these differing phenotypes. It is very difficult, using this system, to determine which of the probes may be detecting noise rather than a clear miRNA signal, and there is no obvious fluorescence level at which to place a detection threshold cut off to remove noisy expression data. It seems obvious that the signal from a large bulk of the probes with an average log₂ expression level of approximately 5.5 will be below the accurate detection threshold of the method and may not be expressed in these cell lines at all. It is also clear that cross-hybridisation between probes and non-specific miRNAs or other interfering nucleic acids can be misinterpreted as a specific miRNA induced signal as would be expected by any oligo based detection method. The miR-191 probe is a case in point, which is known to generate a high background signal “in the absence of any prepared target” (Lu et al., 2005).

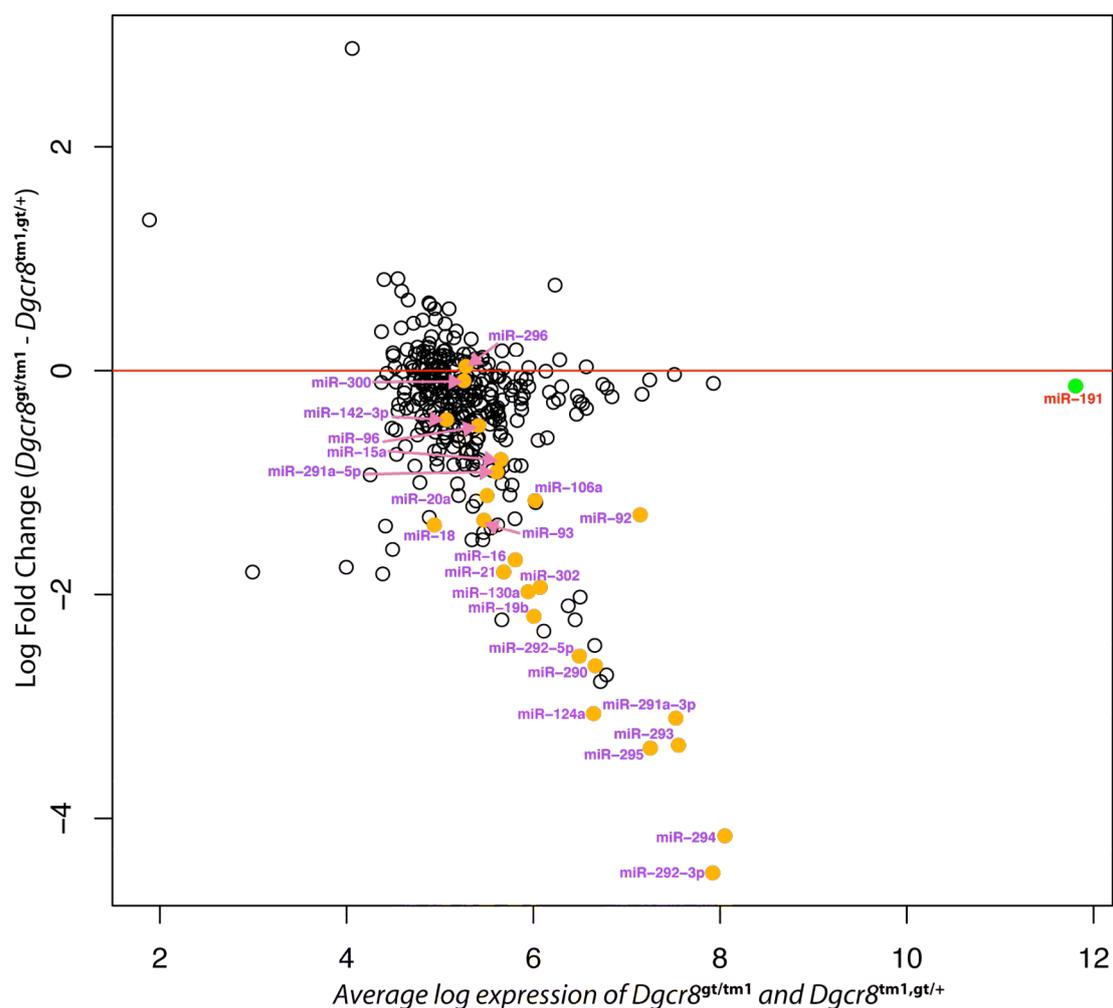


Fig.4.3: A comparison of the expression of miRNAs in $Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$ cells compared to $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ cells. The averaged log fluorescence of beads from of $Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$ cells and $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ cells from replicates 1A and 1B were compared to each other in two sets according to broad genotype ($Dgcr8^{gt/tm1}$ vs. $Dgcr8^{tm1,gt/+}$). The expression of each probe was compared between these sets with each dot representing a miRNA. Orange dots are those miRNAs shown to be expressed in mouse ES cells cultured in feeder free conditions by Houbaviy *et al.* (Houbaviy *et al.*, 2003), with sequences that correlate between the aforementioned study and the probes of the Luminex set. The miR-191 probe (green) is known to register high background readings (Lu *et al.*, 2005).

As a consequence it is unclear whether any miRNAs are still expressed in the *Dgcr8* depleted cell lines and to what extent. It is clear, however, that those mature miRNAs known to be expressed in ES cells are strongly down regulated in these knockout

cells. These results concur with the results of my Northern blots in Chapter 3 and extend them to a much broader set of ES cells expressed miRNAs.

4.3.2 miRNA expression profiling of cell lines with Illumina/Solexa high throughput sequencing

In order to avoid the problems inherent to array analysis which is limited by probe design, probe specificity and sensitivity, I prepared small RNA libraries for each of the *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines for Illumina/Solexa massively parallel sequencing (the sequencing itself was conducted by the Sanger Institute Sequencing Core). Following the sequencing procedure the resultant reads were quality controlled and mapped by Dr. Cei Abreu-Goodger.

Initially all of the reads for each library were mapped to reference sequences comprising of a range of non-coding RNAs (including miRNAs, tRNAs, snoRNAs, snRNAs, rRNAs, mitochondrial tRNAs, mitochondrial rRNAs and miscellaneous other ncRNAs) all derived from the Ensembl database (for non-miRNA ncRNAs) (Flicek et al., 2008) or miRBase (miRNA hairpins) (Griffiths-Jones et al., 2008) (Table 4.1). Subsequently, for each base of each of these reference sequences the number of Illumina/Solexa reads representing that base was calculated. The maximum number of reads representing a single base within a ncRNA sequence was then used to represent the expression level of that RNA. The use of this value means that the length of the RNA sequence does not affect its representative value on any subsequent plot. As the sequences associated with the ncRNAs did not appear to be randomly distributed across the length of the sequences, this would seem to produce the fairest reflection of expression.

Sample	Filtered read depth of library	Depth of sequence that maps to ncRNAs
Dgcr8^{+/+}	1865369	1409029
Dgcr8^{tm1,gt1/+}	2581471	1842631
Dgcr8^{gt1/tm1}	1569525	728577
Dgcr8^{tm1,gt2/+}	2389852	1756044
Dgcr8^{gt2/tm1}	1073911	483343

Table 4.1: The filtered read depth of each of the small RNA libraries and the number of reads that subsequently map to the Ensembl and miRBase derived non-coding RNAs. For each sample the first column represents the number of reads remaining following the removal of the 3' adapter sequence from each read and the removal of reads that were comprised of a single nucleotide for >75% of their length or that were less than 16 bp long following the removal of the adapter sequence. The second column represents the number of the remaining reads that subsequently mapped to the ncRNA sequences derived from Ensembl and miRBase (See section 2.10.3.2 for details). These initial stages of library processing were performed by Dr. Cei Abreu Goodger.

Of the reads which were mapped against this ncRNA set, 71-76% of the reads from libraries derived from the *Dgcr8^{+/+}*, *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* cell lines matched a library sequence and passed all of the associated thresholds imposed (see section 2.10.3.2). By contrast only 45-46% of the reads from libraries derived from the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell lines matched to the libraries. As can be seen in Fig.4.4, of the reads that do match to the library sequences a dramatically reduced proportion of the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell line libraries can be considered as miRNA derived sequences. It seems that the reduced proportion of ncRNA matching reads in these libraries may be accounted for by a reduction in the number of miRNA matching reads in the library (Fig.4.4).

In parallel, an Illumina/Solexa miRNA sequence library generated for wild-type mouse ES cell lines in a previous study (Babiarz et al., 2008) was remapped to the same ncRNA set, following the same criteria as presented here. In order to

demonstrate consistency between the experimental approaches used in these two studies, I calculated the Spearman correlation (0.731) between the maximum sequence depths for all of the miRNAs with a sequence depth greater than 8 in both the wild type ES cell set of Babiarz *et al.* and that which is presented in this thesis for the *Dgcr8*^{+/+} cell line. This comparison confirms that the expression profile for miRNAs generated for mouse ES cells in this study is consistent with that seen in previous studies, bearing in mind variation introduced by experimental procedure. A closer comparison of the expression of individual miRNAs within each library confirmed that there were also no startling expression differences between miRNAs at this level (Data not shown).

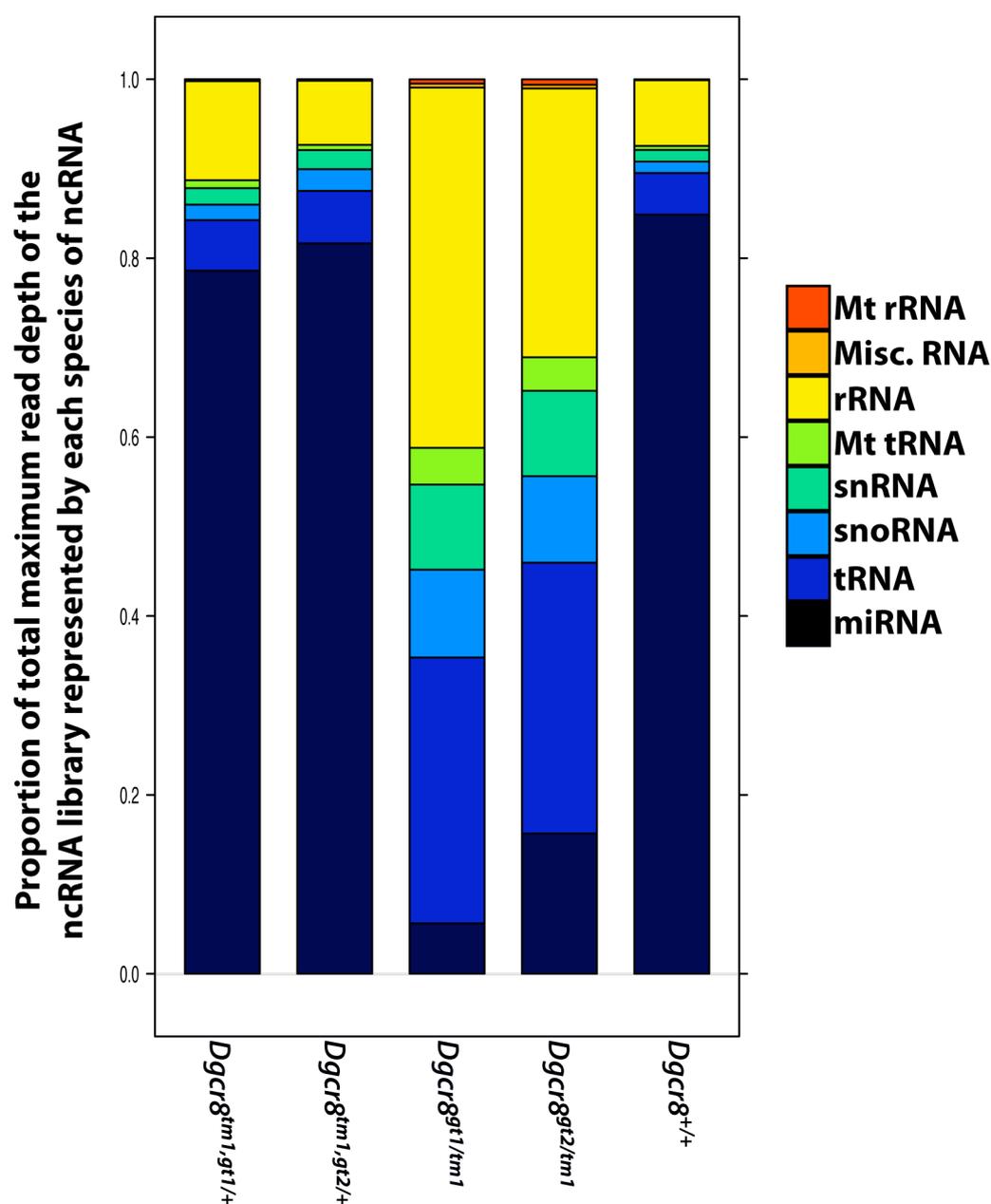


Fig.4.4: The proportion of ncRNA matching reads in each library that are accounted for by each ncRNA type.

In order to confirm that the two *Dgcr8^{tm1,gt/+}* and *Dgcr8^{gt/tm1}* cell lines behave consistently, in a manner which not dependent on their cell line of origin, the maximum sequence depths for each of the ncRNAs was compared between the two heterozygotes and the homozygous mutant cell lines (Fig.4.5A). As can be seen in these plots the cell lines with corresponding genotypes behave very similarly. However, it should be noted there is a degree of variation in both cases. This is

particularly obvious in the case of the *Dgcr8^{gt/tm1}* cell line comparison, where *Dgcr8^{gt1/tm1}* is consistently attributed with slightly lower counts for miRNAs relative to the other non-coding RNA species. It is unclear whether this difference is attributable to a genuine difference in the effectiveness of the disruption of the *Dgcr8* locus or is a consequence of a difference in the library preparation in each case. Further replicates of each library would allow an effective estimate of the variance in results due to the experimental procedure and would help to clarify the source of these differences.

When the ncRNA populations for the *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* libraries were compared to the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell line libraries pairwise it was clear that there is a linear relationship between the max sequence depths for the non-miRNA sequences in each case (Fig.4.5B), in addition to a substantial reduction in miRNA associated reads in the case of the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell line libraries (Discussed below). It therefore seems likely that none of these non-miRNA ncRNAs species are DGCR8 dependent. It is of particular note that although Drosha has been proposed to play a role in rRNA processing (Wu et al., 2000), there was no clear dependence on wild type *Dgcr8* for the expression rRNAs in these cells. Therefore all of these DGCR8 independent RNAs were used to normalise the read depths for each sample. This normalization will allow fair cross sample comparisons to judge the effect of the inserted traps on *Dgcr8* function.

A

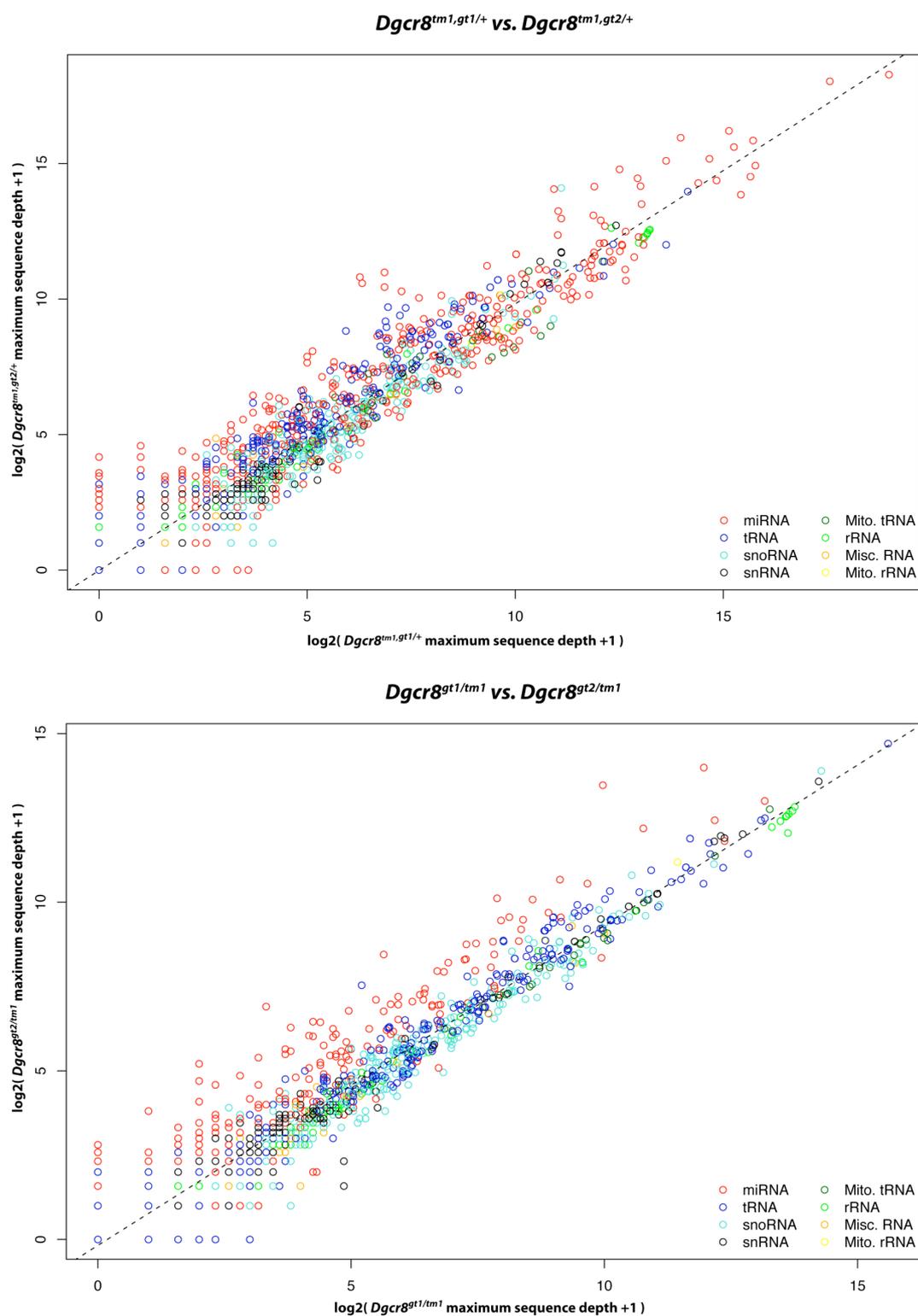


Fig.4.5: A) A Comparison of the small RNA libraries derived from the replicate cell lines. For each cell line maximum sequence depth for individual miRNAs, tRNAs, snoRNAs, rRNAs, snRNAs, mitochondrial tRNAs and rRNAs and the Ensembl class of miscellaneous RNAs are plotted. The dotted line represents a linear model fitted to the non-miRNA sample maximum read depths, based on ncRNAs with a maximum read depth greater than 8 in both samples.

B

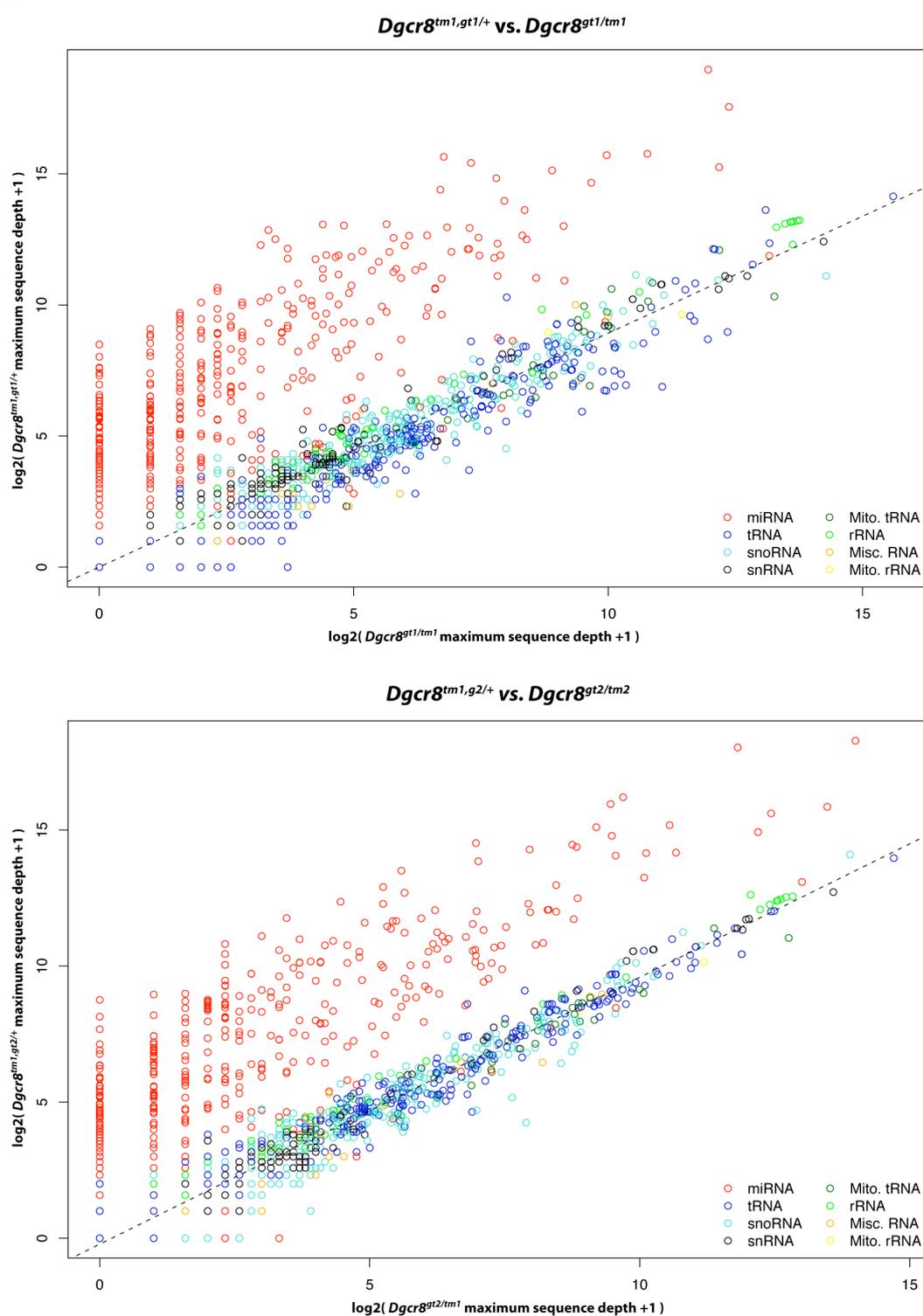


Fig.4.5: B) A wide selection of ncRNAs are not affected by the depletion of functional *Dgcr8* and are thus used to normalise the samples. In a comparison between the maximum read depth (pre-normalisation) of a set of ncRNAs between the *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* cells and the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell lines, tRNAs, snoRNAs, rRNAs, snRNAs, mitochondrial tRNAs and rRNAs and the Ensembl class of miscellaneous RNAs all demonstrate a clustered linear relationship between the samples. The dotted line represents a linear model fitted to the non-miRNA sample maximum read

depths, based on ncRNAs with a maximum read depth greater than 8 in both samples. It is widely accepted that DGCR8 does not play a role in the processing of the majority of these RNA classes. This analysis demonstrates that DGCR8 independent processing extends to a much larger set of ncRNAs, demonstrating that all of these non-miRNA ncRNA species behave in a similar fashion upon the depletion of DGCR8.

Having normalised the maximum sequence depth values for all of the ncRNAs from each cell line it is clear that there is a fundamental depletion of miRNAs in the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell lines when compared to the control lines (Fig.4.6) (Appendix A (CD)). In addition, following normalization, comparisons were made between the *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* libraries and the *Dgcr8^{+/+}* library (Fig.4.7A) in addition to the comparisons between the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cells and their respective control cell line (Fig.4.7B). From these comparisons it is evident that there is a broad reduction in the level of expression of miRNAs in both the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cells as would be expected when compared to the expression levels in control cell lines. It also seems apparent that *Dgcr8* is not functionally limiting in the processing of miRNAs as there seems to be a limited change in miRNA expression between the *Dgcr8^{tm1,gt1/+}* and *Dgcr8^{tm1,gt2/+}* cells and the *Dgcr8^{+/+}* control.

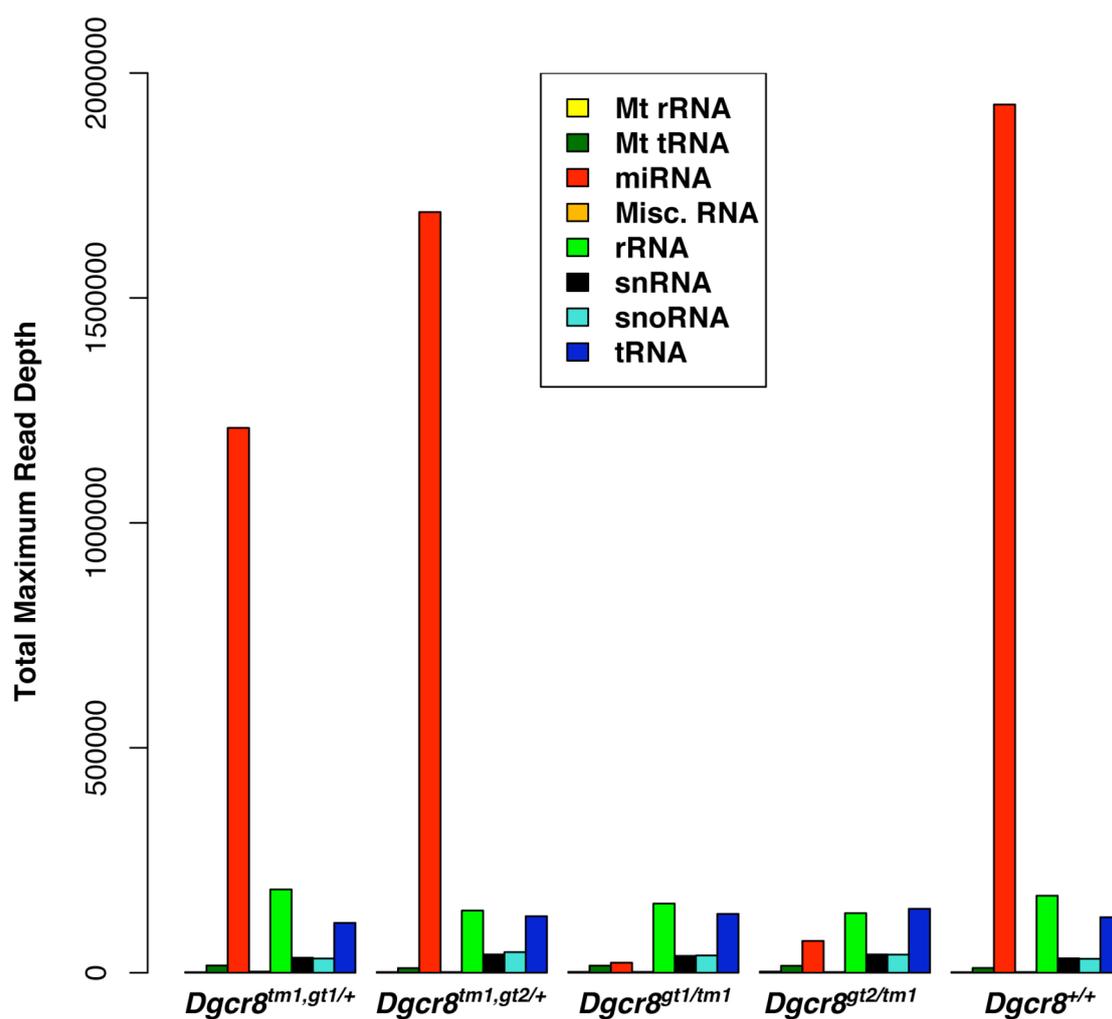
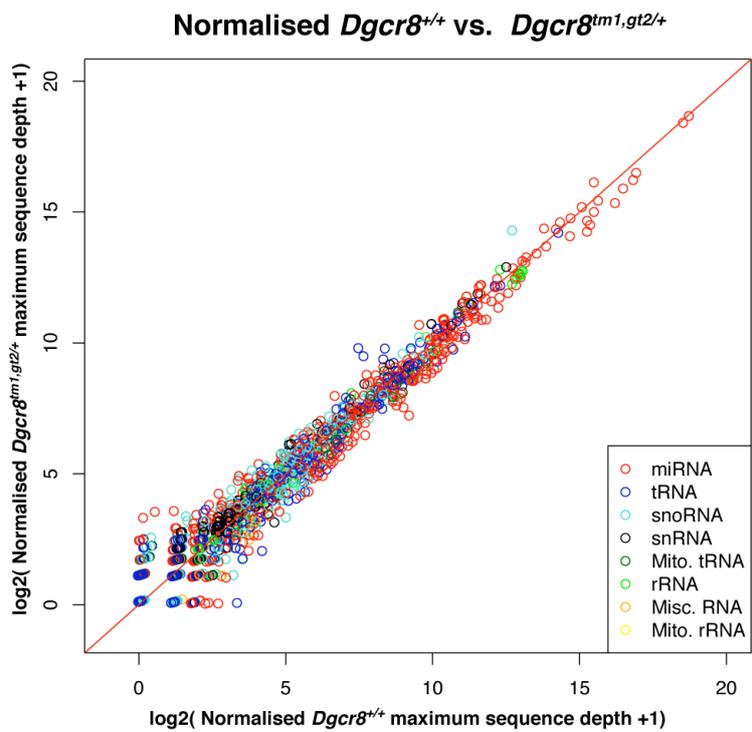
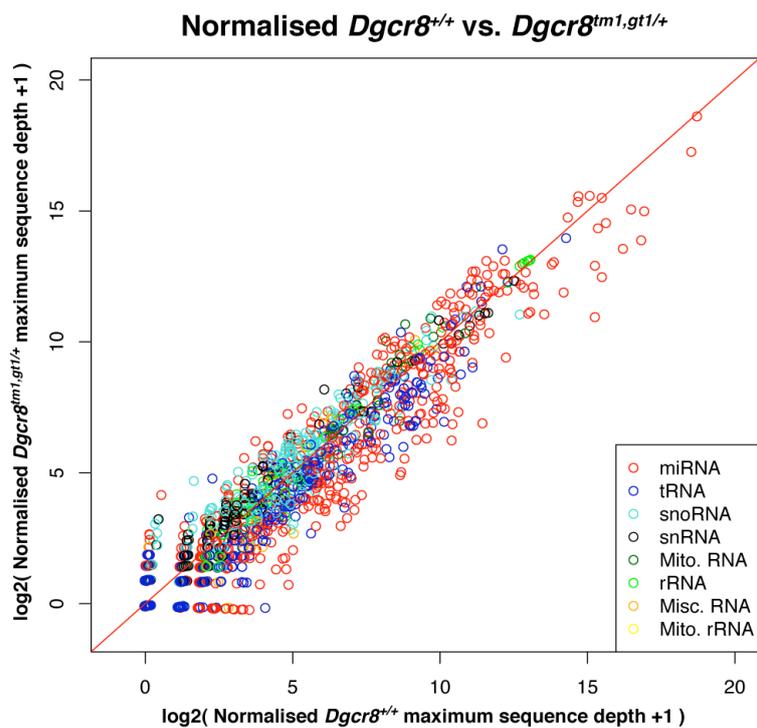


Fig.4.6: Normalised total maximum sequence depths for each ncRNA species in each cell line. Each column represents the total of the maximum sequence depths for a ncRNA species following the loess normalization of the data based on non-miRNA ncRNA species.

A



B

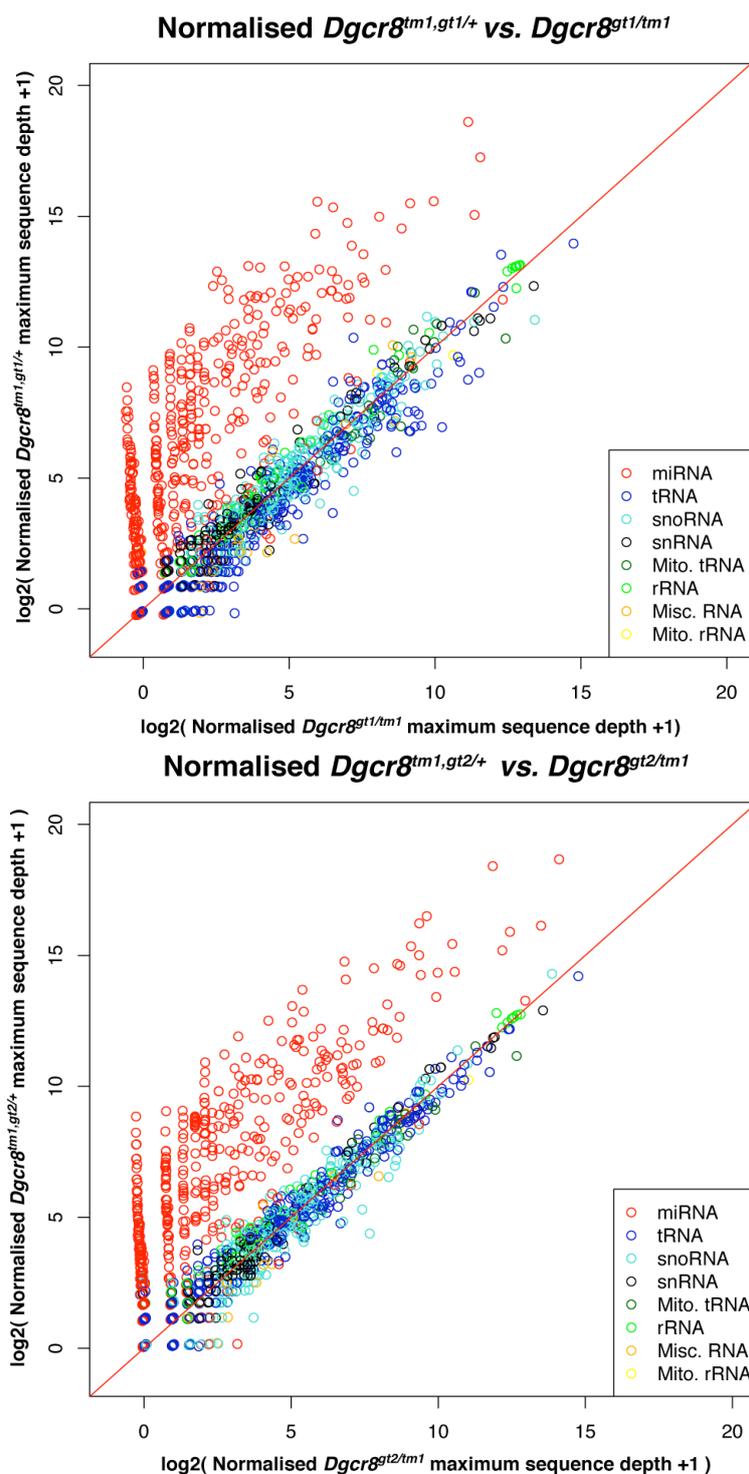


Fig.4.7: The normalised maximum sequence depth for the library of ncRNAs was compared between samples. Following normalisation the maximum sequence depths for each of the ncRNAs in the assembled sequence library were compared in a pairwise fashion. A) The ncRNA read counts were compared between the *Dgcr8*^{+/+} cells and the heterozygous *Dgcr8*^{tm1.gt1/+} and *Dgcr8*^{tm1.gt2/+} control cells.

B) A comparison between the expression of ncRNAs in *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} control cells and *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells depleted in functional DGCR8. The red line represents a 1:1 relationship between the samples.

In order to develop a clearer picture as to the significance of the miRNA reduction induced by the insertion of a trap into each *Dgcr8* allele, I considered the *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cells and the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells as biological replicates of each other (*Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{gt1/tm1}) and conducted a statistical analysis of the expression changes between these two sets (Fig.4.8). The vast bulk of the mature miRNAs are clearly down regulated upon the depletion of DGCR8 (Shifted left, Fig.4.8). This analysis revealed that 414/580 miRBase annotated mature miRNAs are significantly down regulated (adjusted *P*-value < 0.05, LFC = log₂(2)) (Appendix A (CD)), while no miRNAs seem to be significantly up regulated upon DGCR8 depletion. The average maximum read depth included a wide range of expression levels varying from a maximum read depth of greater than 10,000 reads per miRNA (Fig.4.8, orange) down to ~10 reads (Fig.4.8 black). Across this entire expression spectrum the same pattern of down regulation in the *Dgcr8*^{gt1/tm1} is evident.

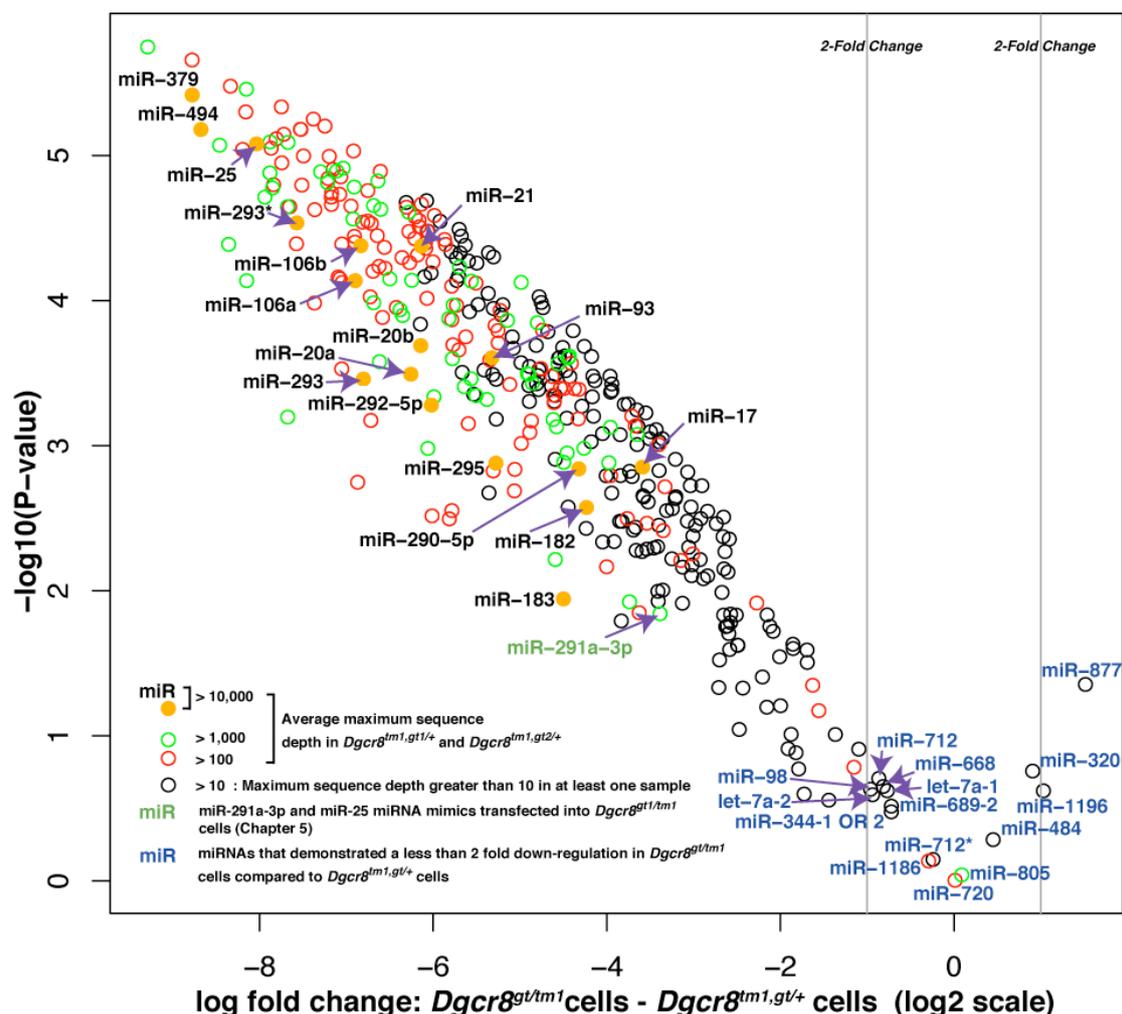


Fig.4.8: A demonstration of the significance of the change in expression of miRBase annotated miRNAs between *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{gt/tm1} genotyped cells. A volcano plot comparing the log fold change of miRNA expression, as assessed by maximum read depth, between the *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{gt/tm1} cell lines and the statistical significance of these changes. To be plotted miRNAs must have a maximum depth of at least 10 in one cell line library.

It is interesting to consider the prospect that some miRNAs may be processed in a DGCR8 independent way. Indeed there is a clear precedent for this with the discovery of mirtrons (Okamura et al., 2007; Ruby et al., 2007). These miRNAs reside in short introns and are excised from their host transcript through splicing rather than RNase III excision. In addition, further examples of DGCR8 independent miRNAs have also been proposed by Babiarz *et al.* (Babiarz et al., 2008); a study also conducted with *Dgcr8* knockout ES cell lines. Here, considering only miRNAs with a post-

normalisation maximum read depth greater than 10 in any of the samples to remove noise and with the same loose criteria as Babiarz *et al.*, considering miRNAs whose expression changes less than 2-fold following *Dgcr8* disruption as DGCR8 independent, 16 miRNAs maintain a relatively constant expression level in cell lines of both genotypes. Of these miR-320, miR-344, miR-668, miR-877 and miR-484 were all identified as DGCR8 independent by Babiarz *et al.* (Babiarz *et al.*, 2008) (Table 4.2).

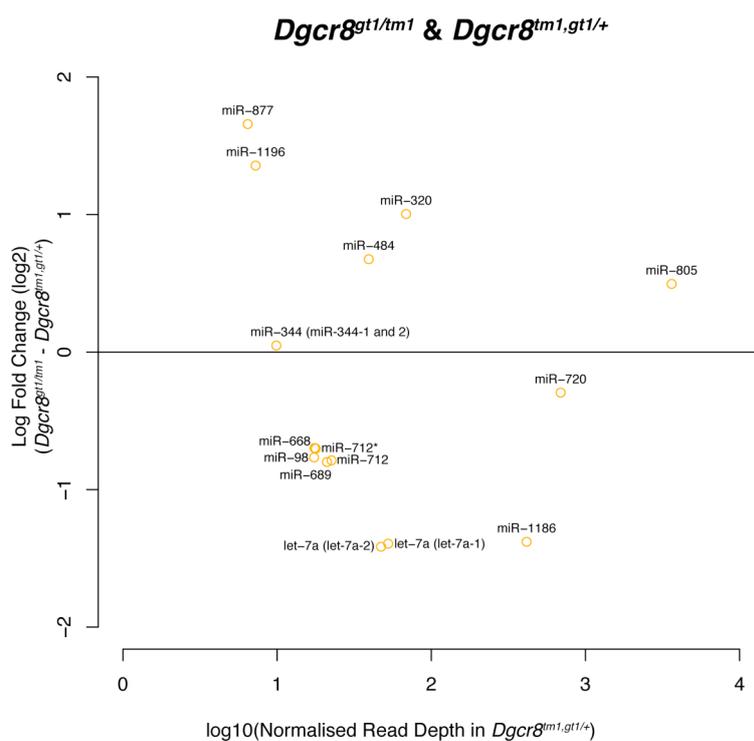
miRNA hairpin	Intronic/Intergenic	Associated Gene	Suggested mirtrons	Suggested DGCR8 independent
miR-320	Intergenic			Babiarz <i>et al.</i>
miR-720	Intergenic			
miR-877	Intronic (sense)	<i>Abcf1</i>	Berezikov <i>et al.</i>	Babiarz <i>et al.</i>
miR-344-1	Intergenic			Babiarz <i>et al.</i> ?
miR-344-2	Intergenic			Babiarz <i>et al.</i> ?
let-7a-1	Intergenic			
miR-689-2	Intronic (antisense)	<i>Zc3h7a</i>		
miR-1196	Intergenic			
miR-1186	Intergenic			
let-7a-2	Intergenic			
miR-712 (miR-712*)	Not Mapped			
miR-484	Intergenic			Babiarz <i>et al.</i>
miR-98	Intronic (sense)	<i>Huwe1</i>		
miR-668	Intergenic			Babiarz <i>et al.</i>
miR-805	Intergenic			

Table 4.2: miRNAs that maintain relatively constant expression between the *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{gt/tm1} genotyped cell lines. miRNAs with a maximum read depth greater than 10 in any sample and with a negative fold change less than 2 between the *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{gt/tm1} cell lines are considered as processed in a potentially DGCR8 independent manner and their details are given here. In addition, intronic status and corroborative information from previous studies is also noted.

In order to determine whether the miRNAs identified as potentially processed in a DGCR8 independent manner in this combined analysis are consistently identified as

such by both *Dgcr8*^{tm1,gt/+} and *Dgcr8*^{gt/tm1} cell line pairs and to gain a clearer indication of the sequence depth of these miRNAs, I conducted a comparison of the expression changes associated with these 16 miRNAs in each case (Fig.4.9). As expected, the majority of these miRNAs appear to comply with the criteria I have used to define this set in both cases. However, some fall outside the cutoff in one of the cases. These include miR-344 (miR-344-1 and miR-344-2), let-7a (hairpin let-7a-1 and let-7a-2) and miR-1186. More stringent criteria could be used to remove these from the set of interest, however miR-344 has also been identified by Babriaz *et al.* as processed in a DGCR8 independent fashion. Interestingly the read depth associated with these miRNAs spans several orders of magnitude (Fig.4.9 and Appendix A (CD)).

A



B

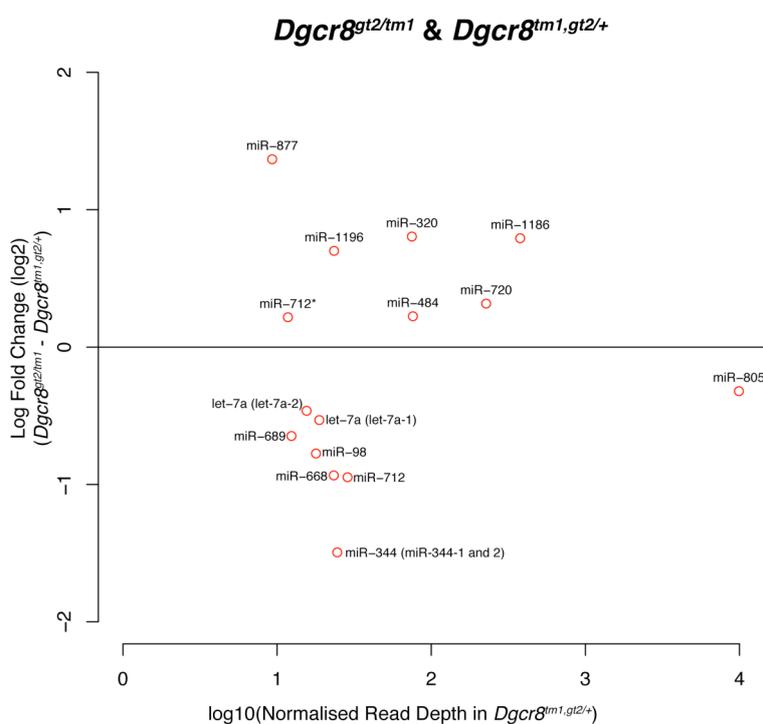


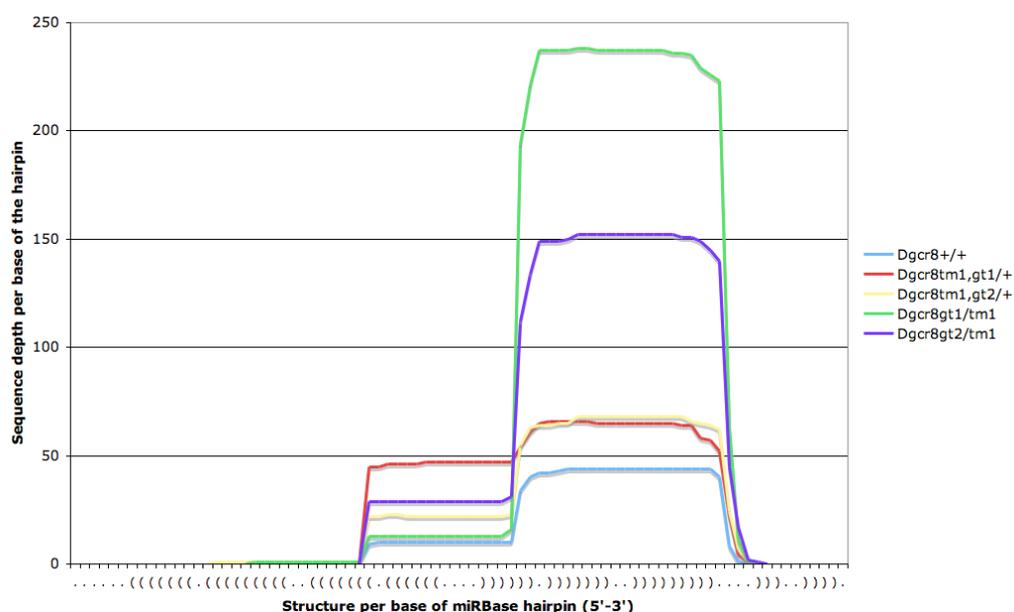
Fig.4.9: The maximum read depth and log fold changes of miRNAs that are potentially processed in a DGCR8 independent manner in pairwise comparisons between each *Dgcr8^{gt/tm1}* cell line and their respective heterozygous control. Each spot represents the expression of an individual miRNA. Where a miRNA can be derived from multiple hairpins the hairpin number is included in brackets. A) A comparison between the *Dgcr8^{gt1/tm1}* and *Dgcr8^{tm1,gt1/+}* cell lines. B) A comparison between the *Dgcr8^{gt2/tm1}* and *Dgcr8^{tm1,gt2/+}* cell lines.

Previously, *miR-877* has been proposed as a mirtron (Berezikov et al., 2007) and resides in the intron of *Abcf1*. Together with the work of Babiarz *et al.* here I confirm that this miRNA does appear to be expressed via a mechanism that does not require *Dgcr8*, confirming its identity as a mirtron. Although not significantly up regulated, this miRNA had the greatest positive log fold change following the depletion of DGCR8 activity. This increase in expression when DGCR8 is depleted may be due to a lack of competition for the remainder of the processing machinery from canonically processed miRNAs precursors leading to more efficient release of *miR-877* mature miRNAs. It could also be caused by a lack of competition from other miRNAs for the miRNP complexes that may protect mature miRNA sequences from nucleolytic degradation.

Babiarz *et al.* noticed that the vast majority of reads mapping to the *miR-320* hairpin were associated with the 3' arm (Babiarz et al., 2008). The sequence reads which map to this hairpin from my libraries do not refute this observation, with the greatest sequence depth clearly associated with the 3' hairpin arm (Fig.4.10A). However, the read depth generated from my experiments is not as great as that published by Babiarz *et al.* so it is difficult to be as confident of the strand bias. The strand bias is much clearer in the Babiarz *et al.* libraries (Fig.4.10B). Babiarz *et al.* hypothesise that the concentration of miR-320 hairpin matching sequence reads to the 3' end side of the miR-320 hairpin may be the result of a lack of the 5' phosphate group at the 5' end of the hairpin, which is left by RNase III like cleavage of the primary sequence and which is required for successful cloning. This supports the notion of an alternative mechanism for processing the pri-miRNA. A Northern Blot confirmed the expression

of miR-320 in *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines (Fig.4.11). Despite the faint signal it seems that there is greater expression of this miRNA in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines. This further confirms the use of Illumina high through put sequencing as a quantitative indication of expression.

A



B

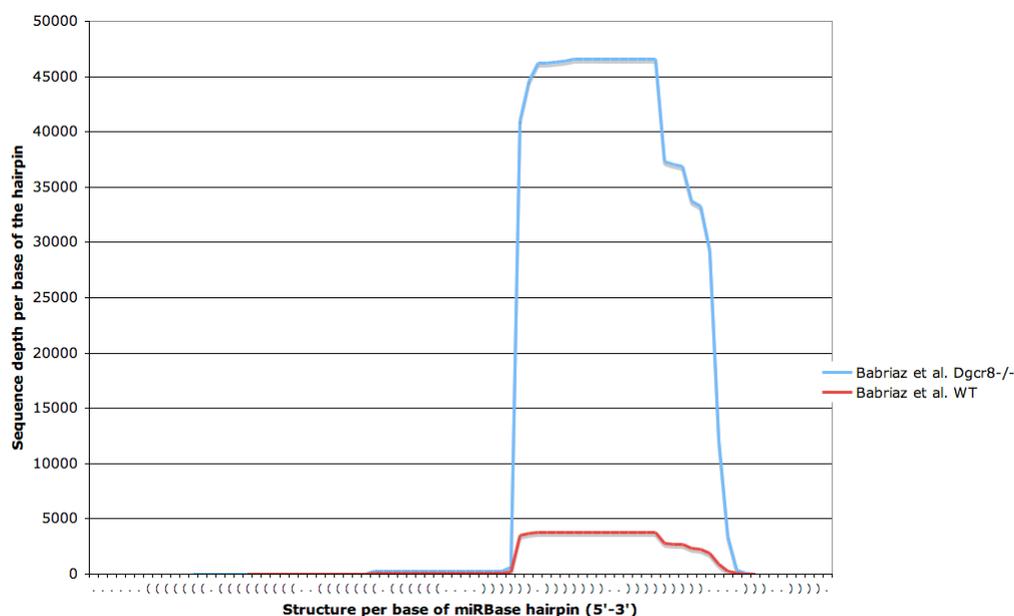


Fig.4.10: The maximum depth of reads which match the miR-320 hairpin sequence. A) The depth of reads across the miR-320 miRBase annotated hairpin from libraries generated as part of this study. B) The depth of sequence reads across the miR-320 miRBase annotated hairpin from libraries generated by Babiarz *et al.* (Babiarz *et al.*, 2008).

Babiarz *et al.* also saw a similar strand bias associated with the *miR-484*. It is difficult to compare my results for this miRNA currently as Babiarz *et al.* used a miRNA precursor hairpin that differs from the miRBase annotation; basing their new version

on local secondary structure. In order to address these differences I would need to map all the sequence reads against the genome or a different hairpin structure. This is beyond the scope of the current study.

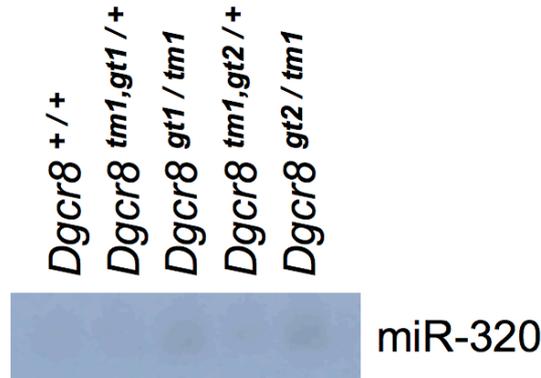


Fig.4.11: A Northern blot for miR-320 expression in *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines.

A further set of miRNAs, previously unidentified as DGCR8 independent were also selected by these loose criteria as potentially unaffected by the loss of *Dgcr8* function, in my experiments. These are let-7a-1, let-7a-2, miR-98, miR-720, miR-689-2, miR-712, miR-1196, miR-712*, miR-1186, miR-805 and miR-344-1 or 2 (Babiarz *et al.* did not define which loci of miR-344 they suspected to be DGCR8 independent (Babiarz *et al.*, 2008)) (Table 4.2). Due to the nature of Illumina/Solexa sequencing and miRNA families it is not always possible to determine the locus from which a miRNA sequence will have been derived as multiple miRNA loci can comprise a set of highly related precursors and as a result sequence reads can match each equally (e.g. *let-7a-1* and *-2*). Where this is the case we halved the number of identical sequences that map to each loci and add this number to the sequence depth at each. While this is not an ideal solution to the problem it seemed like the simplest and fairest method by which to divide the identical reads between highly similar loci. As a consequence, however, it is not entirely clear as to which locus the let-7a miRNA

reads will have necessarily derived from and which primary sequences, if either, will likely be processed in a DGCR8 independent manner.

Of this remaining set of miRNAs that fall within my criteria for DGCR8 independence, the majority are intergenic (7/11) and hence are not expected to be mirtronic in their origin. The two miRNAs miR-689-2 and miR-98 are both intronic, but the miR-689-2 miRNA is antisense with relation to the *Zc3h7a* gene and the miR-98 locus is within a large intron of the *Huwe-1* so, again, neither are expected to be processed as mirtrons. Finally the locus associated with the miR-712 hairpin is not annotated in miRBase. Further investigation of these loci and continued experimentation would be required to confirm the enzymatic dependencies of these miRNAs and to identify alternative routes of miRNA processing.

4.4 Discussion

Through the use of the Luminex bead-based miRNA expression-profiling system I have demonstrated the broad depletion of ES cell associated mature miRNAs in the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell lines. I have then confirmed this knockdown through the generation and sequencing of small RNA cloned libraries with Illumina/Solexa high through put sequencing technology. The latter method offers several improvements over the former probe based technology including a greatly increased sensitivity and fewer specificity-associated issues. The sequencing results therefore clarified the differential expression of a far larger set of miRNAs, in a fashion that was not possible with the bead experiment alone due to the inherent noise associated with the probe based Luminex method.

The sequence results demonstrate a broad and inclusive reduction of mature miRNA expression in the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cell lines. The vast majority of miRNAs are significantly depleted in these cell lines confirming that the insertion of a trap into each *Dgcr8* allele has depleted DGCR8 function in these cells. These results concur with the results of the miRNA Northern blots presented in Chapter 3 which demonstrate a reduction in the expression of a more select group of ES cell expressed miRNAs (Fig.4.12 reference to Fig.3.6 and Fig.3.9). However, sequence reads from canonically processed miRNAs are not totally absent despite the depletion of DGCR8 and there is a low base level of miRNA expression in the *Dgcr8*^{gt/tm1} cells. It is possible that there is a small amount of residual wild type *Dgcr8* expression in these cells, undetected by mRNA Northern blots, from transcripts that splice over the traps from upstream exons to downstream exons. This may explain why there is not a complete absence of mature canonical miRNA expression in these cells, as judged by library sequencing, as residual DGCR8 may process some canonical miRNAs. It is also plausible that there may be alternative processing pathways or dsRNA binding proteins that are able to compensate for the loss of DGCR8 to some extent. Finally due to the sensitivity of the method it is not inconceivable that low level cross-library contamination could lead to residual miRNA associated sequences in the *Dgcr8*^{gt/tm1} samples.

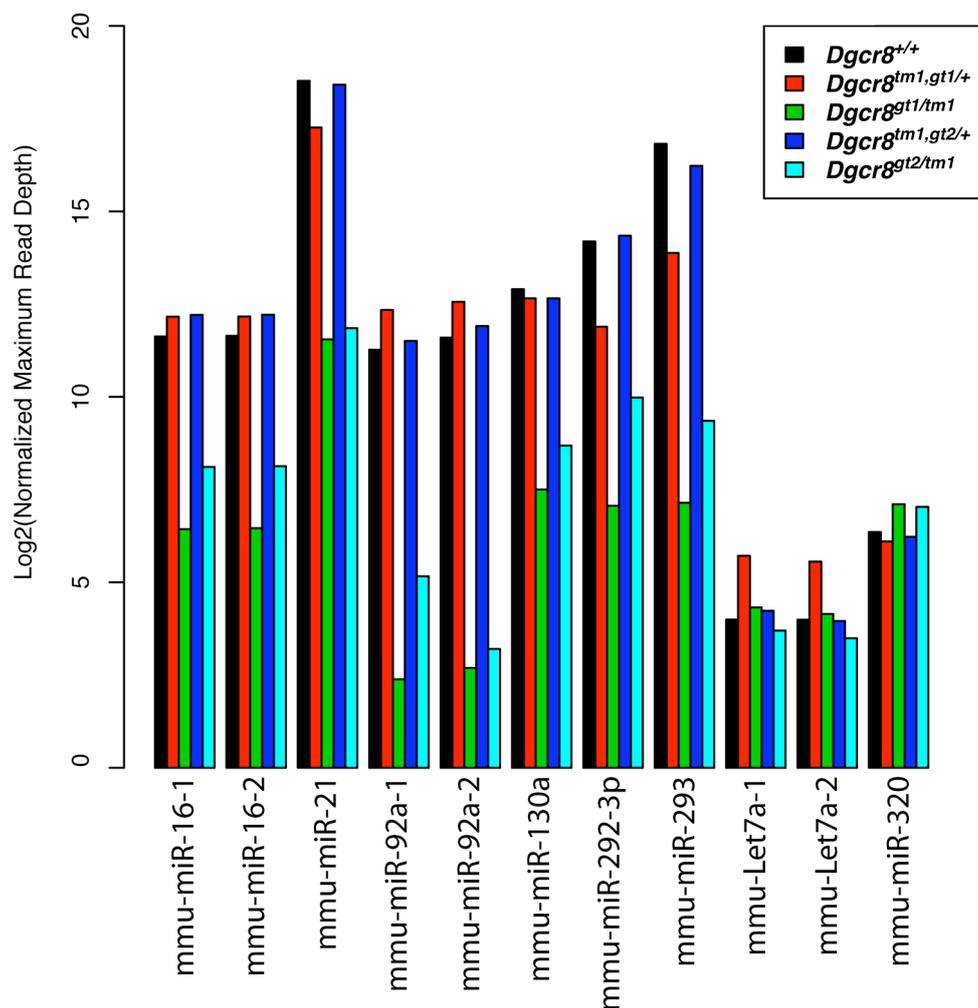


Fig.4.12: Maximum read depth for all miRNAs for which the expression pattern has been demonstrated by Northern blot in this thesis. Normalised maximum read depth for each of the miRNAs tested by Northern Blot (Fig.4.11, Fig.3.6 and Fig.3.9) in *Dgcr8*^{+/+}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{gt1/tm1}, *Dgcr8*^{tm1,gt2/+} and *Dgcr8*^{gt2/tm1} cells.

In addition to demonstrating the substantial depletion of the majority of miRNAs, the sequencing experiment seems to suggest that both the vast majority of other ncRNAs amongst the set in the sequence libraries used to map Solexa data, and a small number of miRNAs are processed in a DGCR8 independent manner. With reference to the broader set of ncRNAs, this result is potentially significant as a dsRNA binding protein may be expected to be involved in further RNA processing pathways, especially since Drosha (its RNase III partner protein) has been implicated in rRNA processing (Wu et al., 2000).

The DGCR8 independent miRNAs include miRNAs that have been identified in a previous study as processed in the absence of this protein in mouse ES cells. These include miR-320, miR-344, miR-668, miR-877 and miR-484. Of these miR-877 is likely to be a mirtron, released from its primary transcript by the mRNA splicing mechanism (Babiarz et al., 2008; Berezikov et al., 2007). It is also intriguing to note that recently miR-320 has been proposed as capable of directing transcriptional repression (Kim et al., 2008a). Consequently it would seem that this miRNA does not conform to a number of canonical rules for miRNA processing and function and as such is deserving of further research.

In addition to these miRNAs, let-7a, miR-98, miR-720, miR-689, miR-712, miR-1196, miR-712*, miR-1186 and miR-805 are also potential DGCR8 independent candidates. These miRNAs were identified with a loose set of criteria (Greater than a maximum sequence depth of 10 in at least one cell line and less than a 2 fold down-regulation of mature miRNA levels in the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cell lines). More stringent criteria would refine this list, but it may be interesting to initially consider this complete set of miRNAs for further study.

In this chapter I have confirmed the functional significance of the traps inserted into the *Dgcr8* locus in Chapter 3. Subsequently it is possible to use these cells for miRNA re-addition experiments in an environment depleted of endogenously expressed miRNAs. On conducting such experiments I can be confident that, to a very large extent, interference in post re-introduction expression profiles from miRNA-target saturation and complicated expression networks caused by miRNAs which target the same transcripts will have been dramatically reduced.

Chapter 5: The derivation of ES cell miRNA candidate target lists in a background depleted of endogenous miRNA expression

5.1 Aim

The aim of this chapter is to use the *Dgcr8^{gt1/tm1}* knock out cell line to identify mRNA targets of endogenous miRNAs in the context of the ES cell transcriptome. These cells have a broad depletion of endogenous miRNA levels and therefore derived targets will be identified in the absence of functional redundancy and the effects of co-regulation. In order to do this I optimize the reintroduction of miRNA mimics into the *Dgcr8^{gt1/tm1}* cell line and subsequently introduce a set of ES cell expressed miRNAs, producing candidate target lists.

5.2 Introduction

The original large-scale analysis of miRNA target perturbations was conducted by Lim *et al.* (Lim et al., 2005). Following the reintroduction of miR-1 or miR-124 into HeLa cells the authors judged changes in the cells' expression profiles through the use of mRNA expression arrays. As a consequence they noted that by transfecting miRNAs not normally expressed in HeLa cells but highly expressed in the heart and the brain respectively, the mRNA profiles of the transfected cells were affected in such a way as to become more akin to the tissue in which the miRNAs are normally expressed. Although the expression changes witnessed in these experiments did significantly overlap with computational miRNA target predictions, it is logical that the system will be somewhat restricted to the identification of the *in vivo*

targets of those miRNAs that are incidentally also expressed in HeLa cells. This is a clear limitation if the intention was to better understand the significance of non-artificial miRNA-target interactions. Grimson *et al.* took advantage of this relatively simple cellular system to generate mRNA profiles from HeLa cells transfected with 11 different miRNA duplexes (Grimson *et al.*, 2007). Again the miRNAs used were not HeLa cell specific and the experiment would thus suffer from the same limitations as those of the Lim study. However, while the limited expression of each miRNA's *in vivo* targets in HeLa cells will interfere with the identification of *in vivo* miRNA-target interactions these gene lists suit the purpose of better understanding the rules underlying miRNA target associations.

Mutant zebrafish embryos lacking both maternal and zygotic Dicer (*MZDicer*) activity display a severe morphological phenotype (Giraldez *et al.*, 2005). The reintroduction of miR-430 into these embryos by microinjection went some way to rescuing the structural abnormalities in the embryonic brain. Subsequently, microarrays were used to identify the putative targets of the miR-430, through a comparison of the mRNA expression in embryos with no Dicer activity to either wild type embryos or embryos with miR-430 reintroduced, in which the expression of miR-430 targets will be down-regulated (Giraldez *et al.*, 2006). The intersection of these two comparisons identified 328 genes down regulated (≥ 1.5 fold) in the presence of miR-430 with experimentally annotated 3'UTRs. Of these, 203 had sequences complementary to miR-430 seed sequence (Lewis *et al.*, 2005) within their 3'UTR. Reporter constructs with the putative wild type and mutant target 3'UTRs fused downstream of a GFP reporter gene were used to validate predicted miRNA target UTRs, through injection into embryos of a wild type or *MZDicer* background. Of the 3'UTRs tested with a single 6mer (2-7) or 7mer target site, 5/7 were validated by these assays (>2 -fold up regulation in *MZDicer*

background). An even greater proportion of targets were validated, if the 3'UTRs contained either a greater number of sites or an 8mer target site. Only ~25% of the proposed targets were conserved in both Tetradon and Fugu. This would mean that any predictive algorithm that required target conservation as predictive criteria would miss these targets.

Although miR-430 is very highly expressed in the early zebrafish embryo, in the *MZDicer* fish the lack of the usual spectrum of miRNAs expressed in the wild type embryos makes the role of individual miRNAs easier to discern. There will be no combinatorial regulation or functional redundancy in the system directed by other miRNAs. As a result the function of miR-430 is potentially even more clear and startling than if individual miR-430 duplexes were knocked out or over expressed in an otherwise wild type system.

Recently similar systems have been developed and used in mouse embryonic stem cells. *Dgcr8* knock out cells have been used to identify miRNAs responsible for the regulation of the cell cycle (Wang et al., 2008). Mouse embryonic stem cells with a disrupted miRNA-processing pathway accumulate in the G1 phase of the cell cycle (Murchison et al., 2005; Wang et al., 2007). Wang *et al.* performed a screen of mouse miRNAs to identify miRNAs that improved this proliferation phenotype (Wang et al., 2008), subsequently identifying members of the miR-290 cluster and miRNAs with similar sequences as important regulators of the p21 (*Cdkn1a*) CDK inhibitor, and central to cell cycle regulation. The knockout background was fundamental to the identification of p21 as a target of these miRNAs as both over-expression and knock down of individual miRNAs in wild-type cells had minimal effect on the ES cell proliferation. This demonstrates the advantages attributable to the removal of

functional redundancy and target saturation from the system prior to screening for miRNA targets.

Whereas Wang *et al.* initially relied on TargetScan predictions to identify p21 as a candidate miRNA target, Sinkkonen *et al.* took an alternative approach and extended the system used by Giraldez *et al.* (Giraldez et al., 2006) to an embryonic stem cell system (Sinkkonen et al., 2008). They compared the expression profile of *Dicer*^{-/-} cells to heterozygous counterparts and then identified from among those genes up-regulated in the *Dicer*^{-/-} cells mRNAs that are down-regulated upon the re-addition of the miR-290 cluster by electroporation. This transfer of the Giraldez approach to a simpler cellular system produced a list of 253 predicted targets of this cluster (including *Cdkn1a*). These genes provided the authors with a functional shortlist enriched with miRNA targets ultimately leading to the identification of *Rbl2* as a potential target of the miR-290 cluster, which could go some way to accounting for the lack of *de novo* DNA methylation at the Oct4 promoter upon differentiation. Simultaneously a second independent study reached the same conclusion concerning the miR-290 cluster orchestrating methylation control via the *Rbl2* gene (Benetti et al., 2008).

In this chapter I assess the expression profiles of the cell lines derived in Chapter 3 and use these profiles to investigate the broad roles of miRNAs in embryonic stem cells. I then explain the optimization of the re-addition of miRNAs into the *Dgcr8*^{gt1/tm1} cell line. By using the expression profiles of the cell lines and expression data from *Dgcr8*^{gt1/tm1} cells transfected with miRNA mimics, I was able to produce lists of genes enriched in miRNA targets for mmu-miR-25 and mmu-miR-291a-3p, using a system of intersecting gene sets similar to that used by both Giraldez *et al.* and Sinkkonen *et al.* (Giraldez et al., 2006; Sinkkonen et al.,

2008). Based on these gene lists I am able to construct hypotheses and functional relationships for future investigation.

5.3 Results

5.3.1 A comparison of growth conditions and their effect on cell phenotype

Initially I performed a comparison between the mRNA expression profiles of *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1}, *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+} and *Dgcr8*^{+/+} cells. RNA was prepared from each cell line in triplicate and the expression profiles were assessed by Illumina Mouse-6 V1.1 expression chips. However as has been mentioned previously it was necessary to alter the cell culture conditions during the course of the work presented in this thesis. The centrifugation step was initially included in the cell splitting protocol to remove the residual trypsin from the cells prior to plating to fresh plates, given the low split ratios used to divide the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells (see section 2.4.1.1). It also initially seemed that as short a period in trypsin as possible might help to limit the morphological phenotype I was seeing amongst the *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells. However, the initial culture conditions used were causing a greater than optimal proportion of the *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+} and wild type cells to differentiate. It was judged that this could be the result of the mechanistic stress of the centrifugation step, used to remove the cells from the trypsin. Subsequently it was deemed more appropriate to culture the cells in the conditions most optimal for the propagation of the control cell lines, as any subsequent differences seen between the growth of the control and mutant cells could still be attributed to the depletion of DGCR8 (see section 2.4.1.2). As a result I regrew the cells in triplicate using the new set of growth conditions and repeated the arrays. It was considered important that this experiment be repeated in full as the results would not only provide the basis for judging likely miRNA

controlled genes, but comparison between the two array sets prepared with RNA grown under the two conditions could be used to judge whether the growth condition were having a fundamental effect on the cells and the results gathered under the two conditions.

In addition to this alteration in protocol between the two array sets, I prepared and labeled the RNA for analysis under the first set of conditions personally. The RNA prepared under the fresh growth conditions was labeled for the array by Dr Peter Ellis in the Sanger Institute microarray facility.

In order to determine the magnitude of the effect of the growth conditions on the expression data subsequently gathered by array, I performed cluster analysis based on the correlation in normalized expression between samples and arrays. Initially, if the samples are clustered based on the entire normalised expression set for each sample, the separation of the clustered structure appears to be based on growth conditions and the identity of the person who conducted the labeling of the samples (Fig.5.1A), (although there is evidence that some of the *Dgcr8^{gt1/tm1}* miRNA transfected samples, grown under the new growth conditions also cluster with the older expression array samples (See section 5.3.3.4)). It is worth bearing in mind, however, that when considering the entire normalized expression set, a large proportion of the data contributing to the inter-sample correlation, will be derived from probes that vary by an insignificant amount, due to small changes attributable to the growth condition or array and experimental noise. As an alternative I refined the selection of probes used to cluster the arrays. First I removed probes from the set based on the detection score attributed to probes by the Illumina Beadstudio programme, which is a measure of the likelihood that the probe target is expressed above background (detection score < 0.05 in more than 5 of the samples).

Subsequently the probe set was further narrowed by using only those probes whose intensity varied significantly between the set of 6 combined $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ arrays and the six $Dgcr8^{tm1,gt1/+}$, $Dgcr8^{tm1,gt2/+}$ arrays, all grown without spinning the cells out of the trypsin (Section 2.4.1.2) (P -value < 0.05 , LFC $> \log_2(1.2)$). The resultant probe set used to separate the samples contained 3489 probes. When only considering these probes the samples clustered as would be expected if the clustering was attributable to genotype instead of growth condition (Fig.5.1B). All of the $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ samples cluster together irrespective of the culture method as do the control cells. Also of note, when the clustering is based on the intensity of this limited set of probes, the $Dgcr8^{gt1/tm1}$ cells transfected with miRNA mimics (Section 5.3.3.4) also cluster with the other $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ samples. By limiting the set of probes used to only those that change most significantly in a single growth condition, the selection of this probe set is not biased by those probes which change significantly between cells with differing genotypes grown under the alternative growth conditions. The choice of probes therefore does not assume that cells with the same genotypes grown under each condition will necessarily behave in the same manner. However, by limiting the probe set in this way, the clustering is now based on the expression changes evident within a set of the most biologically relevant genes, removing a large fraction of the noisy probes whose expression varied by small amounts.

B

Expression Changes between *Dgcr8^{tm1,gt/+}* and *Dgcr8^{gt/tm1}*

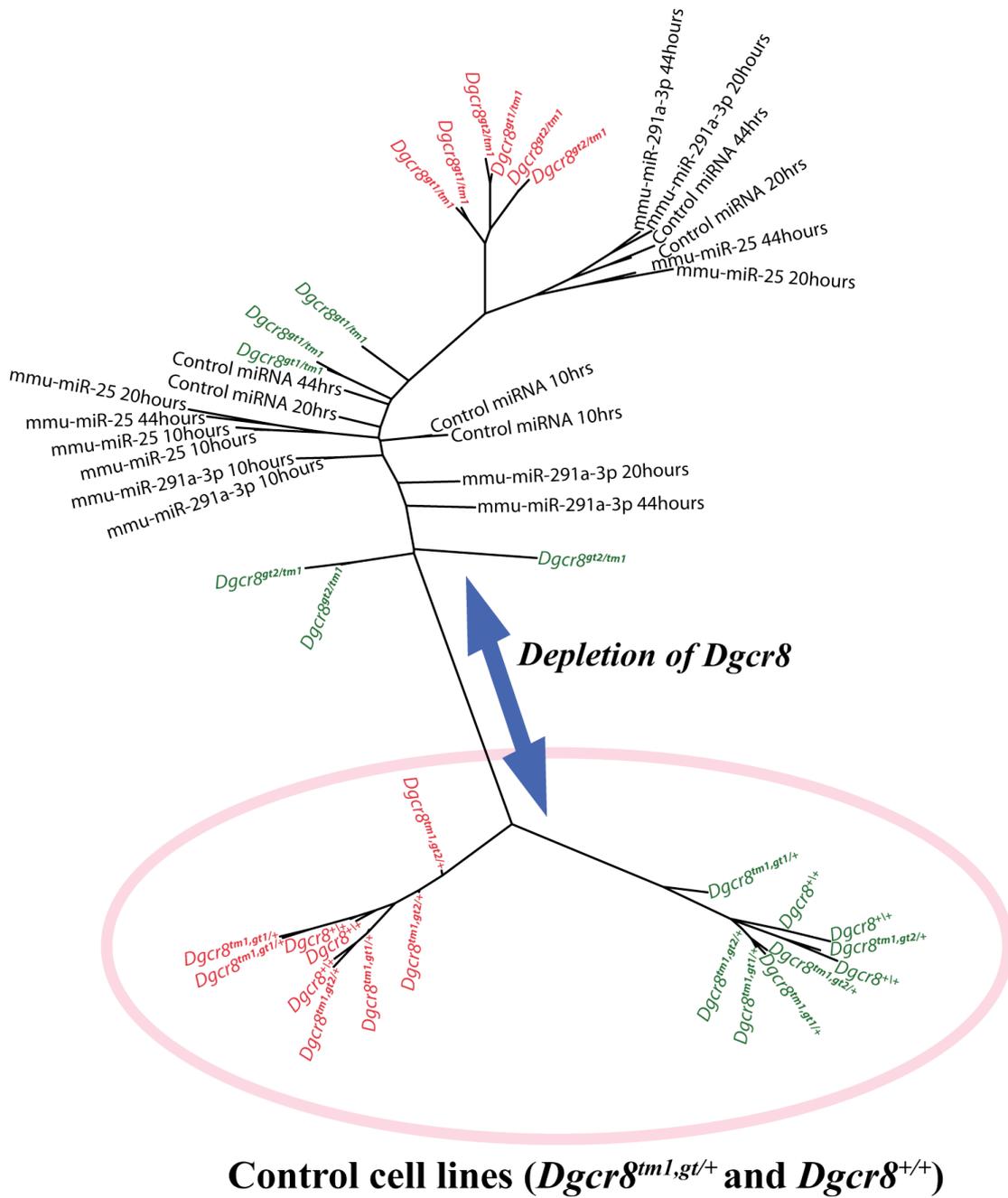


Fig.5.1B: A rootless tree depicting the degree of correlation between all of the mRNA array profiles on the basis of a refined set of probe intensities. The probes used to calculate the correlation between samples were selected as those which were registered as “present” in more than 5 of the array samples as judged by the

lumiBatch detection calls (P -value < 0.05) and which demonstrated significant differential expression between all the of the arrays for cells with two disrupted alleles compared to arrays with a single disrupted allele. All the cells in the comparison were grown under the most recent culture conditions (2.4.1.2). Sample relationships were calculated by Spearman's correlation.

If the arrays grown under the older set of conditions are considered and the same expression comparison is conducted (e.g. $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ vs. $Dgcr8^{tm1,gt1/+}$, $Dgcr8^{tm1,gt2/+}$) 3738 probe intensities change significantly. The overlap between the two sets of significant probes accounts for 2697 probes in this set. Again it is apparent that the two sets of array experiments both identify similar gene sets as significantly altered following the depletion of functional DGCR8 irrespective of the culture conditions.

To further ensure that the cells cultured by the two methods behave in an approximately equivalent fashion, I decided to determine whether there was an enrichment of miRNA seed sequences within the 3'UTRs of genes up regulated upon the depletion of DGCR8, considering the cells grown by each method independently. Sylamer is a program that searches ordered lists of sequences for the enrichment of defined motifs (van Dongen et al., 2008). The program begins at one end of the sequence list and progresses through the list by considering increasingly large sets of sequences for an enrichment of a given motif above what would be considered usual given the number of motifs within the remainder of the sequences in the list. If provided with a list of UTRs ordered by their log fold change in expression following the removal of DGCR8, the program is able to search the list to identify miRNA seed sequences that are enriched within the UTRs of the genes whose expression has either up-regulated or down-regulated by the depletion.

For each set of culture conditions I again compared the 6 expression arrays for *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} against the six arrays for *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} (3 of each genotype) to produce two lists of probes; ordered by their log fold change following the disruption of both alleles of *Dgcr8*. These ordered lists of probes were converted into lists of associated 3' UTRs and analysed (by Dr. Cei Abreu-Goodger) using the Sylamer program for the enrichment of any miRNA seed sequence motifs within either the 3'UTRs of up-regulated or down-regulated genes. As can be seen in Fig.5.2, the most significantly enriched miRNA seed sequences in the genes up-regulated upon the disruption of both *Dgcr8* alleles are essentially the same for both culture conditions. This again implies that fundamentally the two sets of arrays are in agreement as to the most significant expression changes despite practical alterations to the experimental procedure. It is also of note that seed sequences that appear to be amongst the most enriched correspond with miRNAs among those most highly expressed within the *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cells, judged by Solexa/Illumina sequencing (See Chapter 4, Fig.4.8), and whose expression was subsequently reduced in *Dgcr8*^{gt1/tm1} and *Dgcr8*^{gt2/tm1} cells. This enrichment of these miRNA specific seeds implies that a significant proportion of the genes up regulated following miRNA depletion are subject to ES cell miRNA mediated regulation, which is disrupted following miRNA depletion. There are two miRNAs for whom the target seed sequences appear to be significantly depleted within a portion of the ordered gene list (miR-879-7mer-2/m8 and miR-127-7mer-1A), with their line falling below 0 on the y-axis of the chart, although these *P*-values seem to be marginal. The relevance of the enriched seed sequences will be considered more thoroughly later in the chapter, where I will consider the changes in gene expression following DGCR8 depletion, by combining the array data for cells of differing genotypes, grown by either culture method.

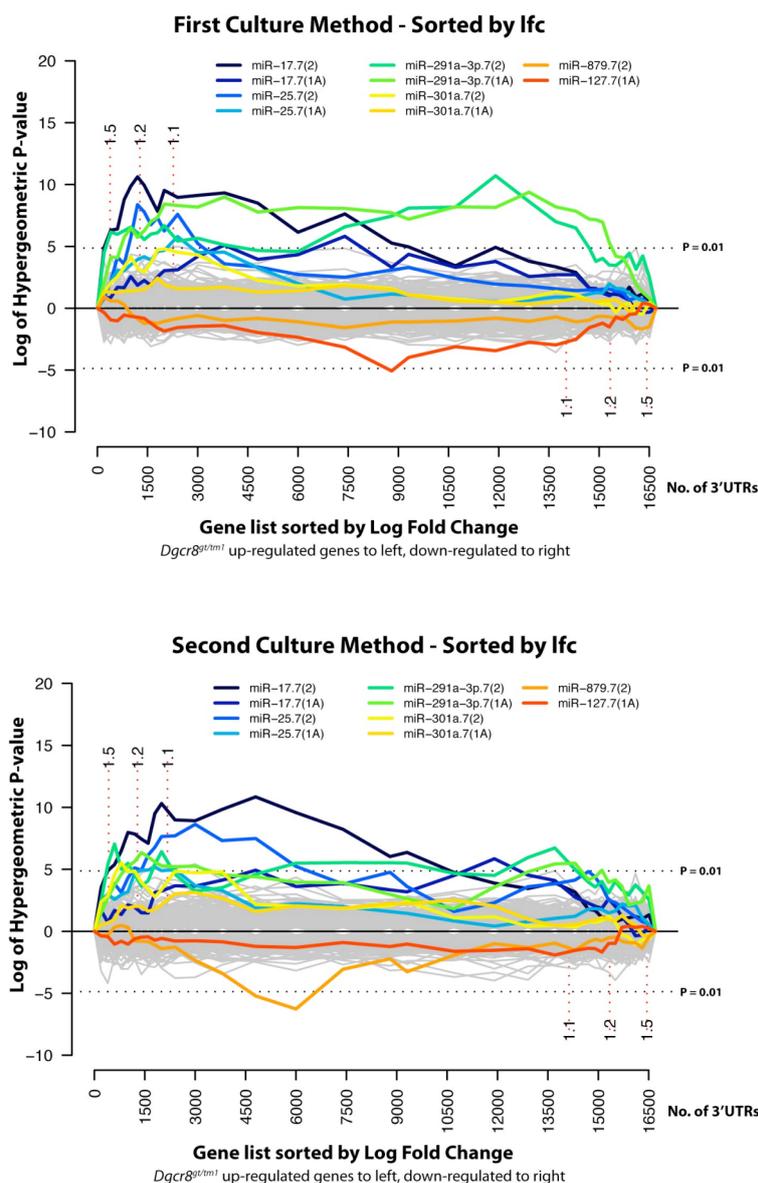


Fig 5.2: Sylamer plots to identify miRNA seed sequence enrichment or depletion within up-regulated or down-regulated genes following the depletion of DGCR8, conducted separately using arrays derived from cells cultured under the older culture conditions (2.4.1.1) (top) and the more recent culture methods (2.4.1.2) (bottom). Genes were ordered according to their log fold change following the depletion of DGCR8 (*Dgcr8^{gt/tm1}* vs. *Dgcr8^{tm1,gt/+}*). These gene lists are plotted along the x-axis, with up-regulated genes on the left and down-regulated genes on the right. Subsequently the 3'UTRs of the ordered list were assessed in sets anchored at 0 on the x-axis, but of increasing size, for enrichment of all miRNA 7mer seed sequences (7mer-1A and 7mer-m8 (7mer-2) seeds based on mirBase v12 miRNA sequences) via a series of hypergeometric tests. The test results for each miRNA are represented by a line on the plot. miRNA target enrichment within a portion

of the list is associated with either a peak to the left of the graph (if enriched in upregulated genes) or trough to the right (if enriched in down-regulated genes). *P*-values cutoffs of 0.01 are represented as dotted lines stemming from the y-axis. On the Sylamer plot, red dotted lines represent the positions of various log-fold-change cutoffs. Sylamer analysis of the gene lists was conducted by Dr. Cei Abreu-Goodger.

5.3.2 Expression profiles of *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1} and *Dgcr8*^{+/+} cells

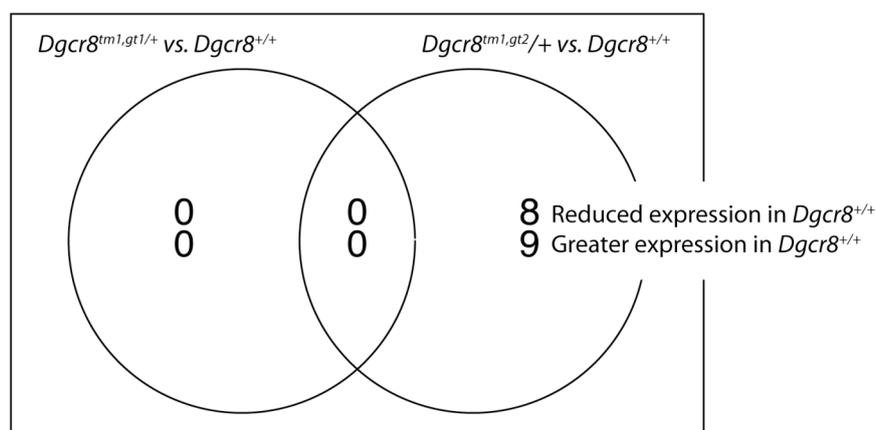
5.3.2.1 Comparison of the expression profiles of *Dgcr8*^{tm1,gt1/+}, *Dgcr8*^{tm1,gt2/+}, *Dgcr8*^{gt1/tm1}, *Dgcr8*^{gt2/tm1} and *Dgcr8*^{+/+} cells

As the expression profiles for the cells cultured by the two methods seem broadly comparable, I decided to combine the array results for each genotype from each culture method into a single analysis. As a result I have 6 expression arrays corresponding to each genotype, which will increase the statistical confidence of any derived expression changes.

First I compared the expression of *Dgcr8*^{tm1,gt1/+} and *Dgcr8*^{tm1,gt2/+} cells to the *Dgcr8*^{+/+} cells to ensure that the single disrupted allele wasn't having profound effects on the gene expression in the cells (Fig.5.3A). As can be seen, there are very few significant differences in the expression profiles of these cells. This suggests that *Dgcr8* isn't a limiting factor in the miRNA mediated gene regulation as there are few significant alterations to expression when it is partially depleted. Alternatively the partial reduction in functional DGCR8 within the cell could trigger an autoregulatory feed back loop to compensate with an increase in the stability of *Dgcr8* mRNA, as described by Han *et al.* (Han et al., 2009). Indeed they noted that although a 50% reduction in *Dgcr8* mRNA levels is expected in heterozygous *Dgcr8* knockout ES cells and MEFs, these cells appear to retain 90% of the wild type mRNA

expression and neither cell line demonstrated a significant reduction in DGCR8 or Drosha protein levels.

A



B

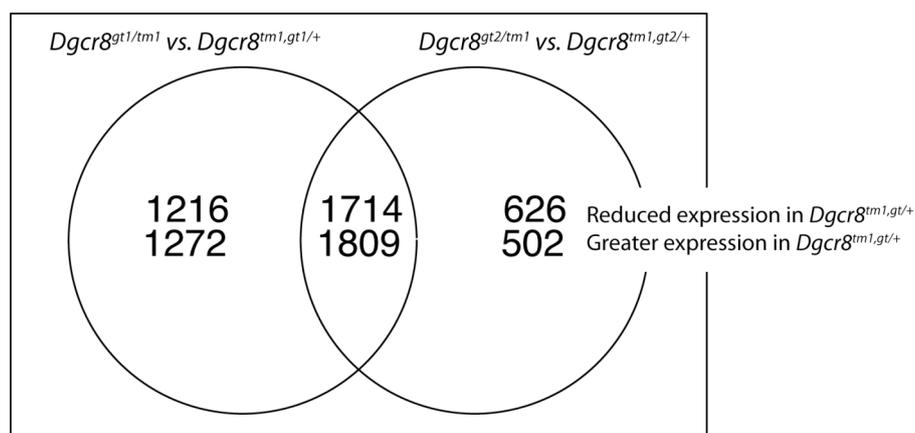


Fig.5.3: A Venn diagram depicting the number of probes registering significant expression changes between cell types. A) Comparisons between wild type cells ($Dgcr8^{+/+}$) and cells with a single disrupted allele ($Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$). B) Comparisons between cells with a single disrupted allele ($Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$) and their paired DGCR8 depleted cell lines, each with a trap within both alleles of $Dgcr8$ ($Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$). Log fold change $> \log_2(1.1)$, P -value < 0.05 .

A comparison between the expression profiles of the $Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$ cells and the $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$ cell lines identified a considerable overlap between the

genes with significantly altered expression when the two biological replicate comparisons were made separately (Fig.5.3B). Therefore subsequently I combined the expression array data for the independent replicate cell lines ($Dgcr8^{tm1,gt1/+}$ and $Dgcr8^{tm1,gt2/+}$, $Dgcr8^{gt1/tm1}$ and $Dgcr8^{gt2/tm1}$) to give me data from 12 arrays corresponding to each broad genotype; $Dgcr8^{tm1,gt/+}$ and $Dgcr8^{gt/tm1}$. By using more arrays I would increase the statistical robustness of any comparison and therefore be able to discern significant expression changes more accurately. Data from these two sets of 12 arrays were compared to determine the expression changes that result from the depletion of DGCR8, stemming from the insertion of a gene trap cassette into both alleles. A total of 6695 probes altered significantly between these two sets (3251 probes were up regulated in the $Dgcr8^{gt/tm1}$ cells compared to $Dgcr8^{tm1,gt/+}$ and 3444 probes were down regulated. P -value < 0.05 , LFC $> \log_2(1.1)$ (Fig.5.4A)). In addition the array probes for this joint analysis were arranged in order according to their LFC and differential expression related t-statistic that resulted from this comparison, and again Sylamer was used (Dr. Cei Abreu-Goodger) to determine if there were any miRNA seed enrichments associated with these expression changes as explained above (Fig.5.2). A similar set of seed sequences were found to be enriched in the genes up-regulated upon the insertion of a trap into both alleles of $Dgcr8$ in this combined analysis as were found when the two growth conditions were considered separately (Fig.5.4B). However, when all the arrays are combined there are no longer any significant miRNA seed depletion signals seen below 0 on the y-axis. The list of miRNAs for whom the seed sequences are enriched amongst those genes up regulated closely resembles the list of the most highly expressed miRNAs in the $Dgcr8^{tm1,gt/+}$ cells (Fig.5.4C, Fig4.8). This is to be expected as it would seem reasonable that these highly expressed miRNAs would be playing an influential role in post-transcriptional regulation in mouse ES cells. What is not necessarily as obvious is that this influential role

would encompass a broad set of target genes, to the extent at which a strong enrichment signal would be seen following the miRNA depletion, rather than influencing a small refined set of target genes. Also of note is the number of highly expressed miRNAs sharing the same enriched seed sequences (Fig.5.4C). In particular the miR-17 7mer seeds are shared by 4 of the top 17 most highly expressed miRNAs in the *Dgcr8^{tm1,gt/+}* cells. In this case the enrichment of the miR-17 seed could likely result from the deregulation of the targets of all of these miRNAs. Conversely all of these miRNAs are likely to overlap in their target lists.

A

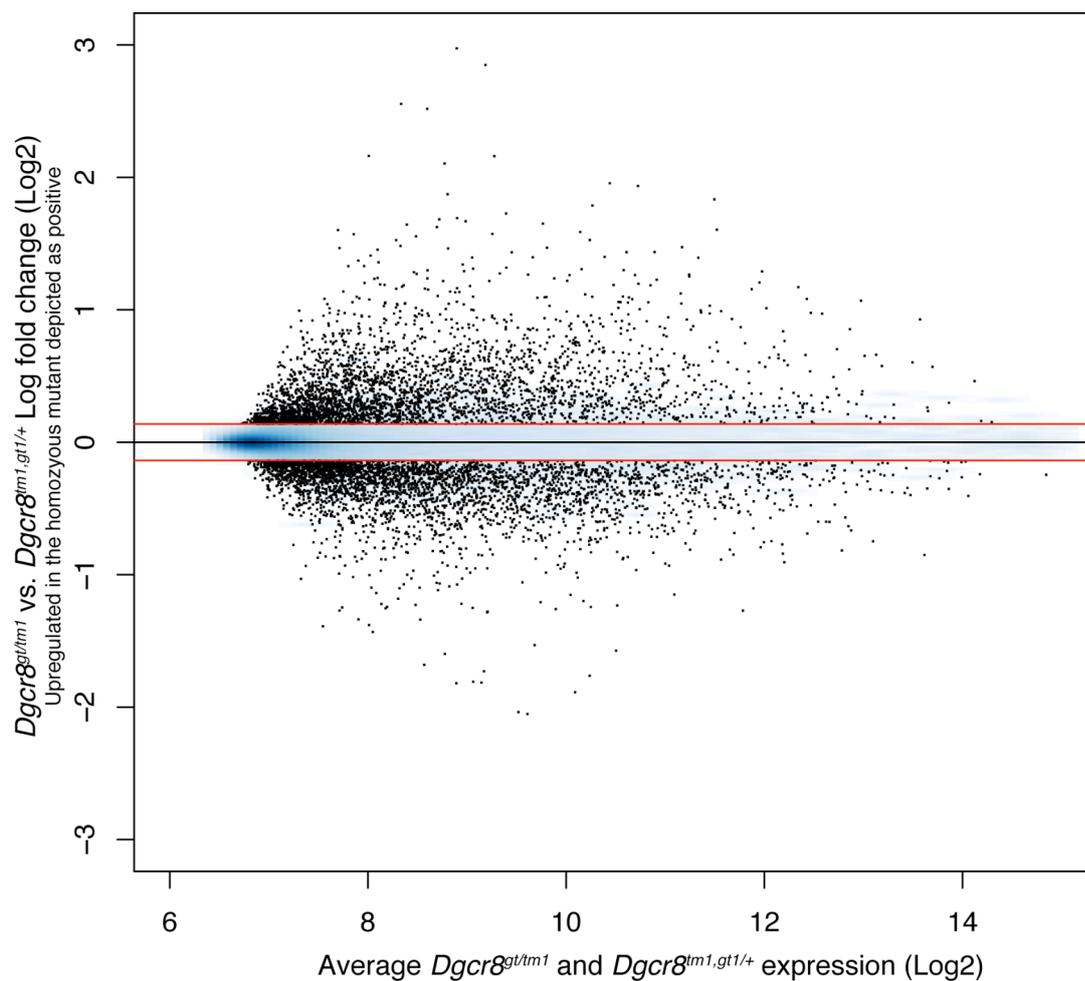


Fig.5.4: A) A plot depicting the average expression of each probe in the heterozygote and homozygous mutant cell lines against the log fold change of each probe following the depletion of DGCR8 from the ES cells. The red lines define a log fold cut off of $\log_2(1.1)$. Probes deemed insignificant given an adjusted p-value cut off of 0.05 and a LFC cut off of $\log_2(1.1)$ are plotted as a density gradient in blue. Probes whose expression alters significantly between the cell lines are plotted as black spots.

B

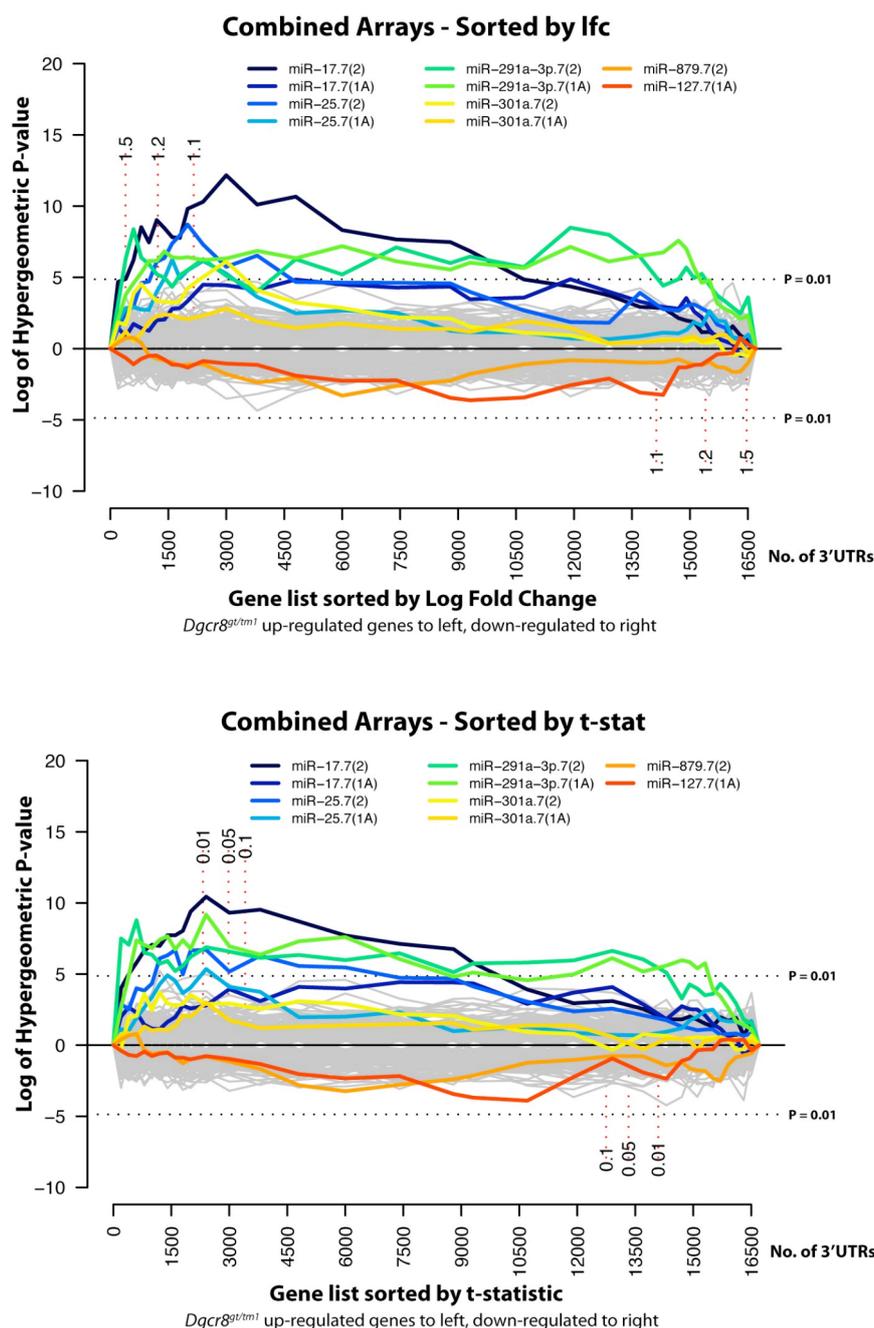


Fig.5.4: B) Sylamer plots to identify miRNA seed sequence enrichment or depletion within up-regulated or down-regulated genes following the depletion of DGCR8 based on expression data derived from cells grown under both the older and the most recent culture conditions For a description of these plots see Fig.5.2. In this case the gene lists were ordered according to their log fold change (LFC) and t-statistic. Sylamer analysis of the gene lists was conducted by Dr. Cei Abreu-Goodger.

C

miRNA Seed	miRNA Target Seed Sequence	miRNAs with a common seed
miR-17.7(2)	GCACUUU	miR-20a.7(2) , miR-20b.7(2), miR-93.7(2) , miR-106a.7(2) , miR-106b.7(2)
miR-17.7(1A)	CACUUUA	miR-20a.7(1A) , miR-20b.7(1A), miR-93.7(1A) , miR-106a.7(1A) , miR-106b.7(1A)
miR-25.7(2)	GUGCAAU	miR-32.7(2), miR-92a.7(2), miR-92b.7(2), miR-363.7(2), miR-367.7(2)
miR-25.7(1A)	UGCAAUA	miR-32.7(1A), miR-92a.7(1A), miR-92b.7(1A), miR-363.7(1A), miR-367.7(1A)
miR-291a-3p.7(2)	AGCACUU	miR-294.7(2) miR-295.7(2) , miR-302a.7(2), miR-302b.7(2), miR-302d.7(2)
miR-291a-3p.7(1A)	GCACUUA	miR-290-3p.7(1A), miR-291b-3p.7(1A), miR-292-3p.7(1A), miR-294.7(1A), miR-295.7(1A) , miR-302a.7(1A), miR-302b.7(1A), miR-302d.7(1A), miR-467a.7(1A), miR-467b.7(2), miR-467c.7(1A), miR-467d.7(1A)
miR-301a.7(2)	UGCACUA	miR-130a.7(1A), miR-130b.7(1A), miR-301b.7(1A), miR-721.7(1A)
miR-301a.7(1A)	UUGCACU	miR-130a.7(2), miR-130b.7(2), miR-301b.7(2), miR-721.7(2)
miR-879.7(2)	AAGCCUC	
miR-127.7(1A)	GAUCCGA	

Fig.5.4: C) The miRNA seed sequences highlighted as enriched or depleted by Sylamer analysis in gene lists ordered by LFC or t-statistic following the disruption of both alleles of the *Dgcr8* locus. The right hand column lists miRNAs that share the same target seed, which is given in the second column. Highlighted in red are miRNAs with an average maximum read depth >10,000 between the *Dgcr8*^{tm1.gi/+} cells (Fig.4.8).

5.3.2.2 Functional analysis of genes up regulated upon DGCR8 depletion

To gain an insight into the broad cellular functions that may be subject to miRNA regulation in ES cells, the group of genes that were up regulated upon the depletion of DGCR8 and miRNAs were subjected to GO analysis and KEGG pathway analysis (3251 probes, P -value < 0.05, LFC > $\log_2(1.1)$) (Table 5.1, Appendix B (CD)). Perhaps the most interesting trend is seen amongst those “Biological Process” GO terms over-represented amongst this set of genes. The most enriched terms include a large number of terms relating to “development”.

This implies that miRNAs may play an important role in the dampening of developmental genes in pluripotent stem cells, perhaps stabilizing this undifferentiated and pluripotent state (Table 5.1).

Also of interest is the enrichment of ‘ECM-receptor interaction’ genes identified by the KEGG pathway analysis. Cellular interactions with the extra-cellular matrix are known to affect cellular differentiation and proliferation. miRNA involvement in the regulation of these pathways may be in an interesting avenue for future research (Table 5.1).

Chapter 5: The derivation of ES cell miRNA candidate target lists in a background depleted of endogenous miRNA expression

KEGG ID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
4512	6.50E-05	3.3	9.7	22	53	ECM-receptor interaction
4510	7.70E-04	2	24.1	39	131	Focal adhesion
5222	1.90E-03	2.3	11.9	22	65	Small cell lung cancer
10	2.30E-03	3.4	5.1	12	28	Glycolysis / Gluconeogenesis
280	3.20E-03	3.2	5.3	12	29	Valine, leucine and isoleucine degradation
5212	3.60E-03	2.2	11.8	21	64	Pancreatic cancer
720	4.00E-03	6.7	1.8	6	10	Reductive carboxylate cycle (CO2 fixation)
640	5.40E-03	3.7	3.7	9	20	Propanoate metabolism
4070	9.00E-03	2.3	8.8	16	48	Phosphatidylinositol signaling system

GO BP IDs	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0001525	9.60E-06	2.6	18.6	37	98	angiogenesis
GO:0016044	5.50E-05	2	31.1	52	164	membrane organization and biogenesis
GO:0009887	8.40E-05	1.7	60.6	88	321	organ morphogenesis
GO:0006897	9.40E-05	2.2	21.8	39	115	endocytosis
GO:0009790	2.00E-04	1.6	63.6	90	337	embryonic development
GO:0001568	2.10E-04	2	27.9	46	147	blood vessel development
GO:0045941	2.90E-04	1.7	45.9	68	242	positive regulation of transcription
GO:0031325	8.30E-04	1.6	56	78	295	positive regulation of cellular metabolic process
GO:0048518	9.00E-04	1.4	121.2	152	639	positive regulation of biological process
GO:0045944	9.20E-04	1.7	35.3	53	186	positive regulation of transcription from RNA polymerase II promoter
GO:0001889	9.30E-04	4	4.7	12	25	liver development
GO:0009891	1.10E-03	1.6	50.5	71	266	positive regulation of biosynthetic process
GO:0051254	1.10E-03	1.7	39.6	58	209	positive regulation of RNA metabolic process
GO:0008285	1.20E-03	2	19	32	100	negative regulation of cell proliferation
GO:0045595	1.50E-03	1.8	28.6	44	151	regulation of cell differentiation
GO:0032774	1.50E-03	1.3	216.8	254	1143	RNA biosynthetic process
GO:0010604	1.50E-03	1.5	53.7	74	283	positive regulation of macromolecule metabolic process
GO:0006732	1.90E-03	2.3	11.9	22	63	coenzyme metabolic process
GO:0045892	2.50E-03	1.7	30.2	45	159	negative regulation of transcription, DNA-dependent
GO:0048519	2.60E-03	1.3	126.9	155	669	negative regulation of biological process
GO:0045670	2.80E-03	6	2.3	7	12	regulation of osteoclast differentiation
GO:0051188	3.40E-03	2.4	9.5	18	50	cofactor biosynthetic process
GO:0043009	3.90E-03	1.6	41.7	58	220	chordate embryonic development
GO:0019219	4.10E-03	1.2	229.1	263	1208	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0007160	4.30E-03	3	5.5	12	29	cell-matrix adhesion
GO:0044262	5.00E-03	1.7	25.4	38	134	cellular carbohydrate metabolic process
GO:0048704	5.10E-03	2.8	6.3	13	33	embryonic skeletal morphogenesis
GO:0021575	5.10E-03	5	2.5	7	13	hindbrain morphogenesis
GO:0048731	5.60E-03	1.2	205.8	237	1085	system development
GO:0009880	6.10E-03	2.9	5.7	12	30	embryonic pattern specification
GO:0006790	6.10E-03	2.5	7.8	15	41	sulfur metabolic process
GO:0003002	6.30E-03	1.7	26.6	39	140	regionalization
GO:0010629	6.60E-03	1.6	35.9	50	189	negative regulation of gene expression
GO:0006733	6.90E-03	3.5	3.8	9	20	oxidoreduction coenzyme metabolic process
GO:0006766	7.40E-03	2.5	7.2	14	38	vitamin metabolic process
GO:0030888	7.80E-03	3.8	3.2	8	17	regulation of B cell proliferation
GO:0032318	7.80E-03	2.4	8	15	42	regulation of Ras GTPase activity
GO:0007157	8.60E-03	4.3	2.7	7	14	heterophilic cell adhesion

Table 5.1: GO term and KEGG pathway analysis of genes up regulated upon the depletion of DGCR8.

Genes significantly up regulated (P -value < 0.05 , $LFC > \log_2(1.1)$) between cells with a trap in each allele of *Dgcr8* when compared with cells with a single disrupted allele (*Dgcr8^{gt/m1}* vs. *Dgcr8^{m1,gt/+}*) were subjected to GO and KEGG analysis for over and under-represented terms. Displayed above are the results of the over-represented “Biological Process” GO terms and over-represented KEGG pathway terms. Additional analyses are available in Appendix B (CD).

5.3.2.3 Identifying DGCR8 dependent alterations to the targets of the ES cell core transcriptional network

The mouse ES cell core transcriptional network (including Oct4, *Sox2*, *Nanog*, *Klf4* and c-Myc), plays a vital role in the maintenance of pluripotency and ES cell identity. Given the strongly enriched miRNA seed sequences identified amongst genes up regulated upon DGCR8 depletion (Fig.5.4B) and given the enrichment of “development” and “differentiation” GO terms amongst the significantly up regulated genes (Table.5.1) it seems reasonable to examine the potential influence of miRNA-coordinated regulation on the targets of this core transcriptional network. To achieve this I downloaded ChIP-Seq and ChIP-chip data for two studies that identified targets of transcription factors within this network (Chen et al., 2008; Kim et al., 2008b). There can be considerable variation between the targets predicted within each data set, for example Kim *et al.* predict 753 target genes for Oct4, while Chen *et al.* predict 6851 targets. Of these 420 are predicted by both studies. In order to ensure that there is the minimum noise associated with the data as possible I refined the target lists for each transcription factor to include only those target genes identified by both studies.

Subsequently, for these five transcription factors listed above, I compared the distribution of the fold changes of their target genes between the *Dgcr8^{tm1.gt/+}* and *Dgcr8^{gt/tm1}* cell lines, to the distribution of the other genes within the profile. All of the target distributions were significantly different to the background distribution (Fig.5.5). This suggests a significant perturbation of the targets of the core transcriptional network following the depletion of DGCR8 and miRNAs within ES cells. Interestingly, while the targets of the other 4

transcription factors all seem to be up regulated upon the removal of miRNAs from the system, the targets of c-Myc appear to be down regulated.

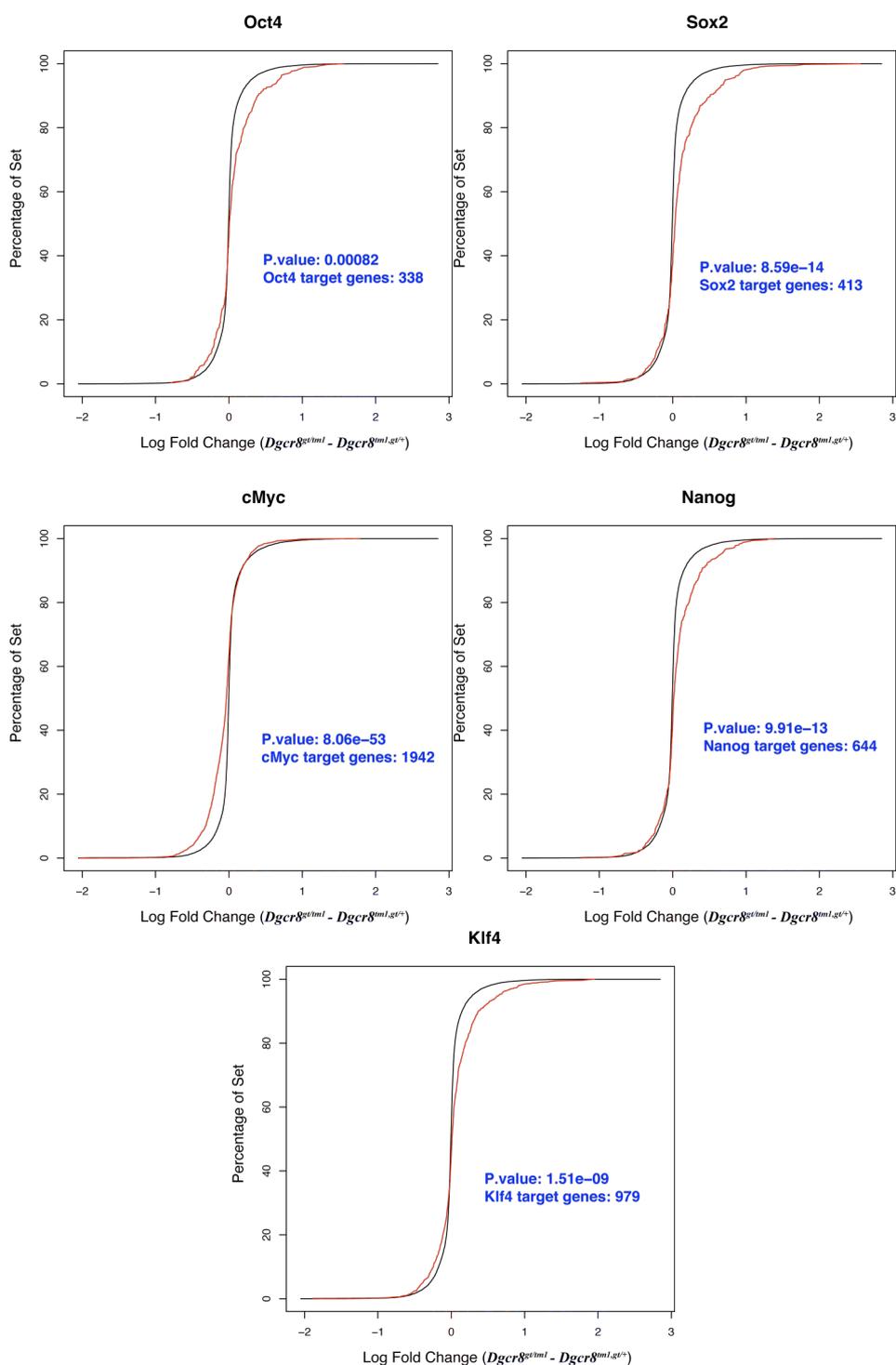


Fig.5.5: Plots to assess the change in the expression of the transcriptional targets of various core transcription factors relative to the other genes following the depletion of DGCR8 ($Dgcr8^{gt/ml}$ vs. $Dgcr8^{ml,gt/+}$).

Dgcr8^{gt/tm1}). The transcriptional targets of each transcription factor were ordered according to their LFC as were the remainder of the genes on the array. These sets were plotted with the black line representing the background data and the red line representing the transcriptional targets. In each case the log fold change distributions of these two sets of genes were subjected to a Wilcoxon test to determine if they differed significantly, with the *P*-values given in blue.

There are three reasons why the target gene expression may have been broadly altered. Either the expression of the core transcription factor itself is affected by the loss of miRNAs or the targets themselves are broadly regulated by ES cell expressed miRNAs either directly or indirectly. To further address the first possibility, beyond the Western blots and immunostaining presented in Chapter 3, I searched the probes whose expression altered significantly following the depletion of DGCR8, for the probes corresponding to these five transcription factors. Perhaps surprisingly, given my initial Western blots and immuno-staining (Fig.3.10 and Fig.3.11), probes corresponding to three of these TFs were indeed within these differentially expressed sets (Table 5.2). Microarrays are a notoriously noisy method, not suited to judging accurately expression changes for single genes and further confirmation by qRT-PCR or Western Blot would be necessary to confirm these alterations in gene expression.

Gene Symbol	Entrez IDs	Illumina Probe	Log Fold Change	Adjusted P-value
Klf4	16600	rpUHFdf15SFI5LRC1U	0.397	0.0011
Klf4	16600	35LRC1Xd1PNCJ05Ras	0.536	0.0012
Myc (c-Myc)	17869	T39XFZ0GEknZUDXsXY	-0.143	0.0029
Klf4	16600	0khLe85Huv0juQw.sQ	0.576	0.0123
Sox2	20674	ZSdKhcTgadIRcSTtcc	0.394	0.0198

Table 5.2: Probes associated with Oct4, Sox2, Nanog, Klf4 and c-Myc that exhibit a significant expression change between *Dgcr8^{tm1,gt/+}* and *Dgcr8^{gt/tm1}* cell lines; (LFC > log₂(1.1), *P*-value < 0.05). Negative LFCs are associated with a reduced expression in *Dgcr8^{gt/tm1}* cells, while positive LFCs are associated with increased expression upon DGCR8 depletion.

In addition, the c-Myc probe seemed to only detect an approximately 10% change in c-Myc transcript levels, while the *Sox2* probe was not significant at the 1% *P*-value cutoff. Despite this, it is intriguing to consider the possibility that some of the phenotypic effects of DGCR8 depletion may reflect the disruption of these core transcription factors. Indeed, a recent study in mouse ES cells noted that the transfection of miR-21 into the cells caused a reduction in cellular self-renewal. Based on computational predictions of miRNA target sites the authors proposed that the miRNA may target *Sox2* and *Nanog* directly (Singh et al., 2008). The Solexa/Illumina sequencing of the miRNA populations in *Dgcr8^{tm1,gt/+}* cells revealed miR-21 as one of the most highly expressed miRNAs to show a significant reduction in expression upon the depletion of DGCR8 in *Dgcr8^{gt/tm1}* cell lines. This reduction correlates with the potential increase in *Sox2* expression. If *Sox2* were indeed to be a miR-21 target this may be a route by which DGCR8 reduction can lead to the disruption of this important transcriptional network. However, this remains a hypothesis until further experimental evidence is presented to support the theory.

Three *Klf4* probes all seem to suggest significant changes in *Klf4* levels. TargetScanS (v.4.2) predicts two miR-25 targets in the 3'UTR of *Klf4*. miR-25 is one of the most significantly depleted miRNAs in the *Dgcr8^{gt/tm1}* cells and if these were to be true miR-25 target sites, miRNAs could be directly manipulating this core transcription factor. It should be noted that *Klf4* wasn't within my miR-25 target-enriched list described later in this chapter (Section 5.3.3.4), so this target relationship again remains speculation.

Oct4, *Sox2*, *Nanog*, *Klf4* and to some extent c-Myc cooperate on their regulation of target genes (Chen et al., 2008; Kim et al., 2008b) and are also involved in the regulation of each other, so it would not be surprising if the disruption of a single member in the network would lead to alterations in the expression levels of the targets of other members of the system. Oct4, SOX2, NANOG and KLF4 bind to the promoters of genes that are either active or repressed in non-differentiated stem cells (Kim et al., 2008b; Liu et al., 2008b). c-Myc on the other hand has a much clearer tendency to bind to the promoters of genes that are active in undifferentiated stem cells (Kim et al., 2008b) and also seems to generally target a distinct set of genes with distinct functions (Kim et al., 2008b; Liu et al., 2008b). The trend for Oct4, SOX2, NANOG and KLF4 targets all to be up regulated in *Dgcr8^{gt/tm1}* cells while c-Myc targets are down regulated is therefore not counterintuitive. In addition the down regulation of c-Myc, upon the depletion of DGCR8, implied by the microarray probes may conceivably account for a proportion of the repression of its target genes.

In order to address the possibility that the deregulation of these TF target gene sets following the depletion of DGCR8 may be due to an enrichment for the targets of ES cell expressed miRNAs amongst these genes (particularly amongst the up regulated gene sets) the genes from each set were mapped to Ensembl transcript IDs and the 3'UTRs of each set (annotated as described in section 2.9.8) were scanned for 7mer seed enrichments with Sylamer (Dr. Stijn van Dongen). No significant 7mer miRNA seed enrichments were identified in the target set. Consequently it is less likely that a broad deregulation of these target sets is due to a depletion of direct miRNA mediated target degradation and is more likely the result of changes at the level of transcription, although miRNA regulation of these sets not detectable by this method can not be ruled out.

5.3.3 Reintroducing miRNA mimics to *Dgcr8^{gt1/tm1}* cells

In order to generate miRNA target lists I planned to reintroduce miRNA mimics into *Dgcr8* deficient ES cells and then interrogate the cells' expression profiles for miRNA dependent expression changes, with mRNA expression arrays. These ES cells do not express endogenous miRNAs and would not be saturated for miRNA mediated target regulation. They would also be devoid of any inter-miRNA redundancy caused by shared seed sequences and should be depleted of endogenous combinatorial miRNA mediated target regulation. This should provide a simplified system for understanding the roles of individual miRNAs. For these reasons I hoped that this system would allow me to generate accurate miRNA candidate target lists with a high degree of sensitivity.

5.3.3.1 Optimisation of the conditions for miRNA reintroduction

I decided to use the *Dgcr8^{gt1/tm1}* cell line as the basis for the miRNA reintroduction experiments. In order to reintroduce miRNAs into the *Dgcr8^{gt1/tm1}* cells they would have to be transfected as mature miRNA duplexes as it was difficult to identify commercially available miRNA mimics that could enter the miRNA processing pathway at any other stage and that could be guaranteed not to require a functional microprocessor in order to be processed to a mature form. To optimize the transfection conditions (Lipofectamine 2000 quantity, siRNA concentration and cell number) for the reintroduction of these miRNAs into the mutant ES cells, siRNAs were used that would have identifiable cellular effects when transfected successfully.

Initially the conditions for siRNA transfection were adapted from the experimental procedures presented in two papers for ES cell siRNA transfection and the recommended

Invitrogen protocol (http://tools.invitrogen.com/content/sfs/manuals/stealth_sirna_tsf_lf2k_man.pdf) (Chen et al., 2007; Takahashi and Yamanaka, 2006). These conditions were used to transfect *Dgcr8^{gt1/ml1}* cells at various cellular densities and with various quantities of siGLO conjugated siRNAs and Lipofectamine 2000 (Section 2.12) (Data not shown). However, although it was apparent that the cells were successfully being transfected under these conditions it was very difficult to quantitate transfection efficiency as there was a relatively high background in each well of what appeared to be untransfected siGLO associated lipid complexes. The intense and punctate siGLO staining also made it more difficult to discern which cells, amongst those in groups, were successfully transfected and which just lay in close proximity to those transfected or an untransfected lipid complex.

Subsequently I altered the method of optimization. As an alternative I transfected the cells with an siRNA targeting *Kif11*, a gene encoding a motor protein. KIF11 depletion causes growth arrest and triggers apoptosis (Ambion). The effect on cell number and viability was measured through the use of an Alamar blue redox assay. This method was originally provided by Dr. Ian Sudbery and had been optimized for working with HeLa cells. Subsequently alterations were made to suit my purposes. In principle metabolically active cells cause a reduction of the Alamar Blue indicator resulting in a colour and fluorescence change. This change is easily quantifiable (Section 2.12) (Data not shown). However, after initial experiments, an alternative method was used as it became clear that this method would require further optimization if it were to be used to effectively quantitate the effectiveness of siRNA transfection, as there was an apparently non-linear relationship between the number of untransfected cells plated in a well and the subsequent Alamar Blue associated fluorescence reading. It was unclear if this non-linear relationship was caused by differential growth rates

at different cell densities or non-linear reduction of the Alamar blue as the cell numbers increased. In addition, due to the nature of mouse ES cells, with this cellular assay it is difficult to know exactly what is the cause of growth changes, whether it is cell cycle arrest or triggered differentiation.

Ultimately, a LacZ siRNA, which would target the β -geo fusion transcripts derived from the gene-trapped allele, was used to further optimise the transfection method. Successful transfection of this siRNA could be visualised by fixing and Xgal staining the cells (Section 2.11.1.2). Successfully transfected cells exhibited reduced blue staining (Fig.5.6) when compared to non-transfected cells or cells transfected with a control siRNA. It became clear from a time course experiment (Data not shown) that these siRNA transfections were very fast acting with clear differences in LacZ activity apparent ~21 hours after the transfection was begun. In addition, this method also revealed the importance of ensuring the cells are spread as evenly as possible across the well prior to transfection. Plating the cells and co-transfecting them caused cells to aggregate at the centre of the well due to the small volume of transfection medium. The aggregated cells seemed to be transfected far less efficiently than the more sparse populations. Subsequently it was necessary to plate the ES cells 3 hours prior to transfection. This allowed the cells time to settle and prevented them from aggregating at the centre of the well during the transfection.

Using this method transfection efficiency could be estimated by comparing the number of white to blue cells in representative photographs of the LacZ siRNA transfected population. An optimization of the cell numbers with this new method (10×10^4 cells to 40×10^4 per 24 well) resulted in 'blue/total cell-number' ratios varying from ~20-36%.

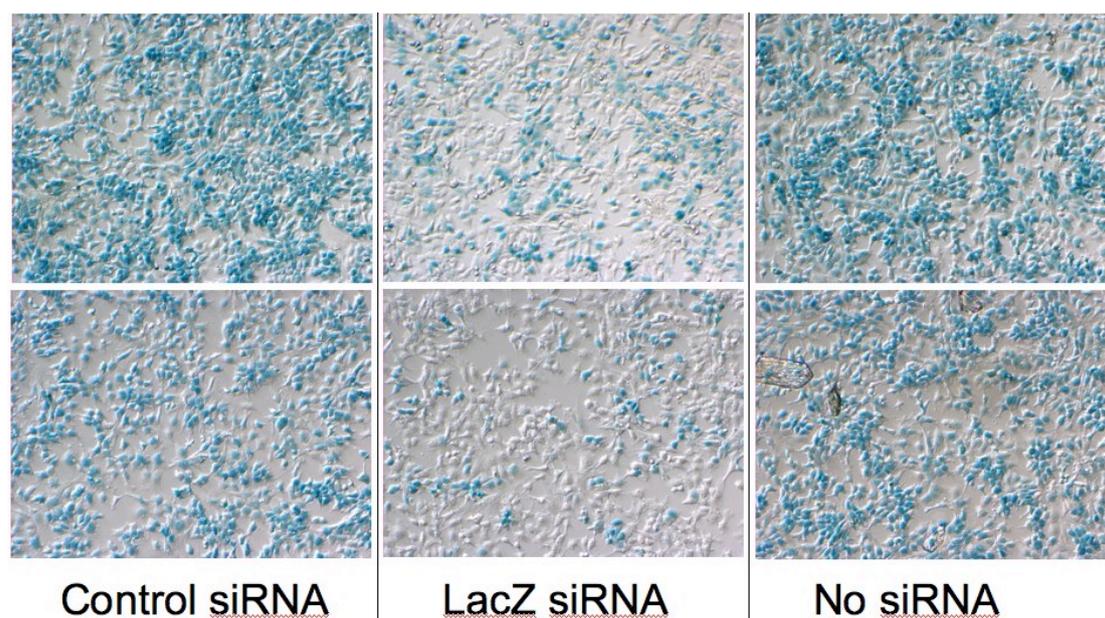


Fig.5.6: Xgal staining results of LacZ siRNA transfected *Dgcr8^{gt1/tm1}* cells in the 6 well format to be used for miRNA mimic reintroduction experiments. Photos demonstrating the depletion of β -geo activity in cells transfected with a LacZ targeting siRNA. Images taken following Xgal staining which stains LacZ expressing cells blue. Taken at 10x relief contrast. Also depicted are cells transfected with a control siRNA or No siRNA. In 5 photos taken of the LacZ siRNA transfected cells, the proportion of blue cells to total cells varies between approximately 23% to 48% with an average ~34% staining positive for LacZ activity. By contrast, in 4 pictures of the control siRNA transfected cells, between 95-97% of the cells stained blue.

Finally, the method was then scaled up in order to transfect 6 well plates rather than 24 well plates to ensure sufficient RNA for the array analyses of the subsequent miRNA-transfected cells (for optimised method see section 2.11), (Fig.5.6).

5.3.3.2 Demonstration of Oct4 expression in transfected cells

In order to ensure that the transfection of the *Dgcr8^{gt1/tm1}* cells with short RNAs wasn't causing the depletion of the ES cell marker expressing cells in the culture (either through cell death or by triggering differentiation) I tested the expression of Oct4, a factor essential in order to maintain the ES cells pluripotent state (Niwa et al., 2000), in transfected cells. I repeated the LacZ siRNA transfection in a chamber slide format, scaling the reagents to suit

the different size culture. Alongside cells transfected with the LacZ siRNA, I transfected an equal number of wells with the control siRNA and cultured four wells without transfection as negative controls. Slides were stained with Oct4 and LacZ primary antibodies, which were bound by Alexa 594 and 488 secondary antibodies respectively. 2 LacZ siRNA and 2 Control siRNA transfected wells were stained with these antibodies to replicate results. Simultaneously non-transfected cells were stained with combinations of these antibodies as controls. As a further control *Dgcr8^{tm1.gt1/+}* cells (grown at a different density and under different conditions) were also stained with either both primary and secondary antibodies or simply secondary antibodies. *Dgcr8^{tm1.gt1/+}* cells do not exhibit LacZ activity (Fig.3.7) and so would act as negative LacZ controls. Again it is important to bear in mind that the Alexa 488 and EGFP expressed from the targeted trap could potentially interfere with each other's signal. However after the cell staining experiments in Chapter 3 and further fluorescence microscopy experiments (data not shown), the EGFP appears to be expressed at very low levels in these cells and was not expected to interfere with the results.

All of the staining-control cells gave the expected background fluorescence for each of the antibodies used. Although there is diffuse fluorescence it is at a far lower level than that seen in the wells with both primary and secondary antibodies included in the staining. Only *Dgcr8^{gt1/tm1}* cells known to express LacZ (β -geo) fusion protein gave a clear, nuclear-staining Alexa 488 signal when stained with both primary and secondary antibodies (Fig.5.7). This coincides with the nuclear localization of the LacZ activity seen in the Xgal staining experiments in Chapter 3. However, it is clear that the *Dgcr8^{tm1.gt1/+}* cells also fluoresce when probed with the LacZ antibody, although this is a cytoplasmic stain rather than a nuclear stain. This does not coincide with Xgal staining activity seen in Chapter 3 for this cell line.

This stain is not seen if only the secondary antibodies are included in the staining procedure, so seems to be a non-specific binding caused by the primary antibody rather than EGFP expressed from the targeted trap cassette interfering with the interpretation of the results.

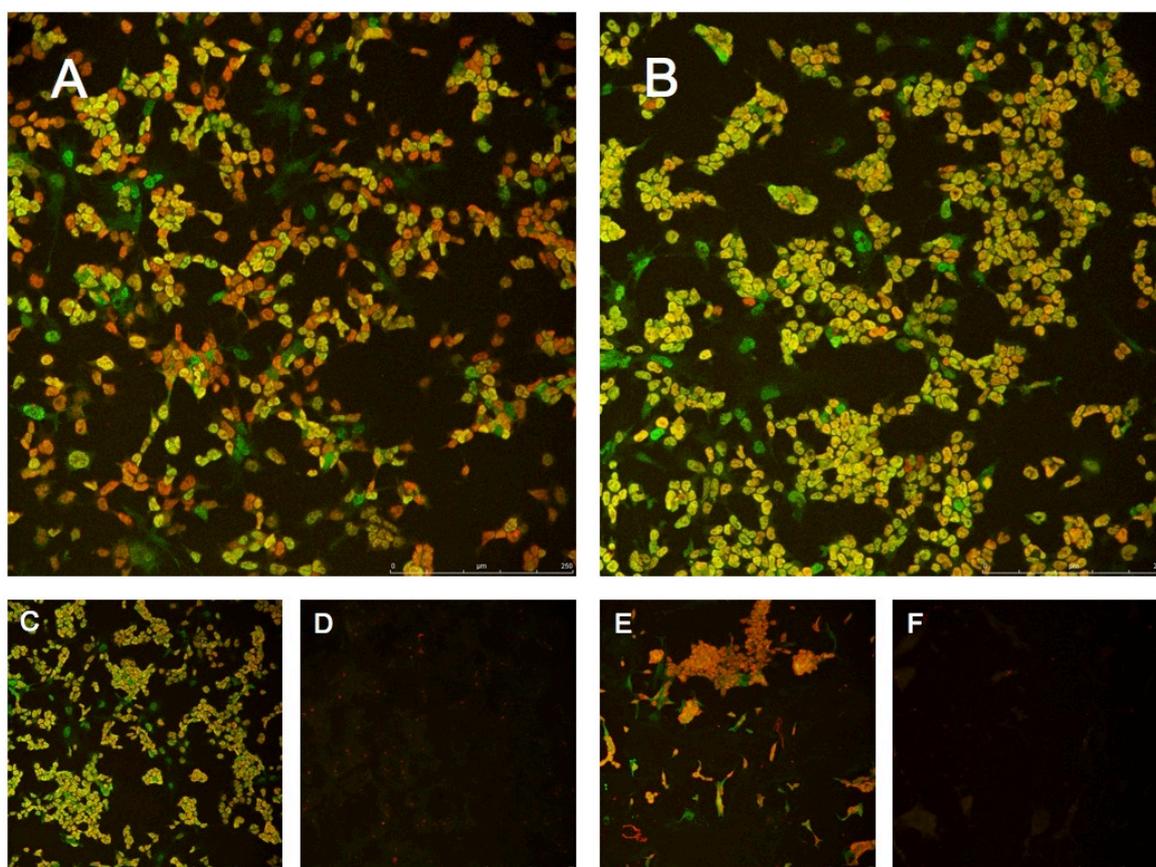


Fig.5.7: Immunostaining of transfected $Dgcr8^{gt1/tm1}$ cells with LacZ and Oct4 specific antibodies to demonstrate siRNA transfection does not influence Oct4 expression. LacZ stained green. Oct4 stained red. A) $Dgcr8^{gt1/tm1}$ cells transfected with LacZ specific siRNA. B) $Dgcr8^{gt1/tm1}$ cells transfected with control siRNA. C) Non-transfected $Dgcr8^{gt1/tm1}$ cells. D) Non-transfected $Dgcr8^{gt1/tm1}$ cells with secondary antibodies only E) Non-transfected $Dgcr8^{tm1,gt1/+}$ cells F) Non-transfected $Dgcr8^{tm1,gt1/+}$ cells with secondary antibodies only. A,B,C and E – 10 sections through stained cells – Maximum projection. D and F – 5 sections through visible fluorescence – Maximum projection.

Transfection with the siRNA targeting the LacZ (β -geo) fusion transcript causes a subsequent reduction in nuclear 488 staining in a large proportion of the $Dgcr8^{gt1/tm1}$ cells when

compared to those cells subjected to the control transfection, which is expected (Fig 5.7A). However, there is still clear Oct4, 594 staining in the affected cells. I conclude that the successful transfection of pluripotent ES cells expressing marker proteins does not cause subsequent differentiation in the timeframe within which the cells are likely to be lysed and they are not depleted through cell death caused by lipid transfection.

5.3.3.3 Selecting miRNA mimics to reintroduce into *Dgcr8^{gt1/tm1}* cells

mmu-miR-291a-3p and mmu-miR-25 miRNA mimics were selected as the two miRNA-mimics I would transfect to optimize and prove the system for the identification of miRNA targets. Mmu-miR-291a is a member of the miR-290 miRNA cluster. The functional study of this cluster is popular in the field. It is known to be highly expressed in mouse ES cells (Fig.4.8) and contains miRNAs with seed sequences common among ES cell expressed miRNAs (Fig.5.4C) (Calabrese et al., 2007). As a consequence of these studies miR-291a-3p has been associated with a number of target genes. These can therefore be used as positive controls for my target selection. Functional studies of the role of mmu-miR-25 are scarcer in ES cells. This makes the generation of mmu-miR-25 target lists an interesting prospect. Mmu-miR-25 was a miRNA highly expressed in the control cell lines as judged by Solexa sequencing (Fig.4.8) and exhibited a large and significant expression change when both alleles of *Dgcr8* are disrupted, identified when *Dgcr8^{tm1.gt/+}* miRNA expression was compared to that of *Dgcr8^{gt/tm1}* cells. The miR-291a-3p and miR-25 seed sequences were also among the most enriched seed sequences in the 3'UTRs of genes up regulated in *Dgcr8^{gt/tm1}* cells when compared to *Dgcr8^{tm1.gt/+}* cells (Fig.5.4B) adding to the evidence that these are important miRNAs in ES cells.

5.3.3.4 Transfection time course to identify the optimal time point for cell lysis

In order to identify mRNAs specifically degraded as a consequence of the presence of miRNAs, miRNA mimics would be transfected in parallel to control mimics (Dharmacon). The changes in mRNA expression resulting from the miRNA transfection would then be judged by a comparison to the mRNAs expression profile of the control transfected cells as estimated by Illumina expression microarray.

To determine the optimal period post-miRNA transfection for which to leave the cells prior to RNA lysis, I conducted a time series, lysing the cells 10 hours, 20 hours and 44 hours after the miRNA transfection was begun. It seemed reasonable to hypothesise that this period should be kept to a minimum since ES cells transfected with siRNAs seem to react rapidly to the transfection (Takada et al., 2005b). Assuming that a system devoid of competing small RNAs may respond still more rapidly, the response may be very quick and fairly brief if the siRNAs are cleared with little competition. It is also reasonable to assume that the shorter the delay between transfection and lysis the more likely that the genes that are affected are primary miRNA targets rather than downstream effects. However, it would be necessary to leave the cells for a sufficient time to allow the delivered miRNAs to enter the system and affect mRNA levels.

Both miR-25 and miR-291a-3p mimics were transfected as part of this time series. All transfections were duplicated (although both 10 hour replicates were conducted upon the same day, while the 20 hour and 44 hour replicates were conducted on two separate

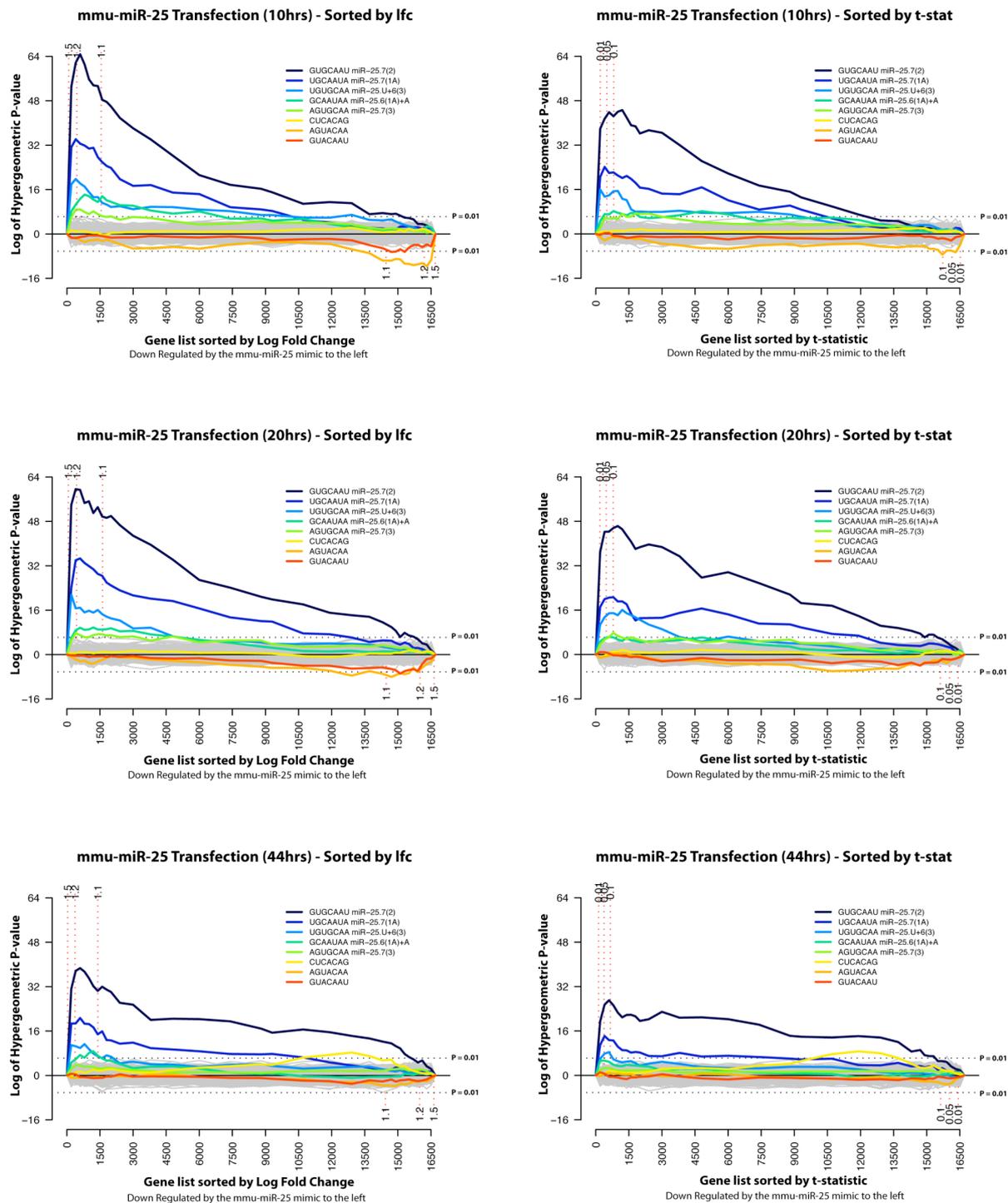
occasions). As described above, for each replicate a set of cells was also transfected with a control miRNA to control the derived expression data for the effects of the transfection itself.

At each time point the Illumina mRNA expression microarray expression profiles for each miRNA were compared to the microarray expression profiles for the control miRNA. Once again, following this comparison, for each miRNA and time point the probes from the array were ordered according to their LFC and their associated t-statistic. Sylamer was used to examine these lists for the enrichment of miRNA seed sequences (Fig.5.8) (Dr. Cei Abreu-Goodger) in mRNAs that were up or down regulated. As expected the most significant seed sequences, seen amongst the genes down regulated by either miR-25 or miR-291a-3p relative to the control, were those for the miRNA that was transfected in each case. These enrichments were particularly significant at the 10 hour time point and subsequently became less significant as time increased. This loss of seed of enrichment amongst the down-regulated genes is particularly obvious in the case of miR291a-3p. At 44 hours the down-regulated genes do not appear to be particularly enriched for the miR-291a-3p seeds. This is despite the fact that at 44 hours there are approximately the same number of probes demonstrating a log fold change of at least '1.1' as was seen at 10 hours (Fig.5.8). It follows that the loss of enrichment is not entirely caused by a return to normal/control expression levels, but may perhaps be compounded by changes in the expression of indirect targets, that do not contain miRNA seed sequences, which dilute the enrichment signal.

At the other end of the x-axis there are also enriched seed signals. Unsurprisingly these seed sequences correspond to the first 8 bases of the control miRNA used (Based on cel-miR-239 - UUGUACUACACAAAAGUACUG) or the 6mer with an extra U at the 3' end of the target

site. It is worth noting that the current miRBase annotation and the Dharmacon annotation for this miRNA differ, although as the miRNA negative controls are selected for having the minimal sequence similarity with other mouse miRNAs this is not important. Although the control mimic does appear to be affecting the expression of certain genes the enrichment signal is far weaker than that seen in the case of the transfected endogenous miRNA. This is probably because the targets and *C. elegans* miRNA mimics will not have co-evolved to optimize target selection, so that the signal will be due to random off target effects.

A



B

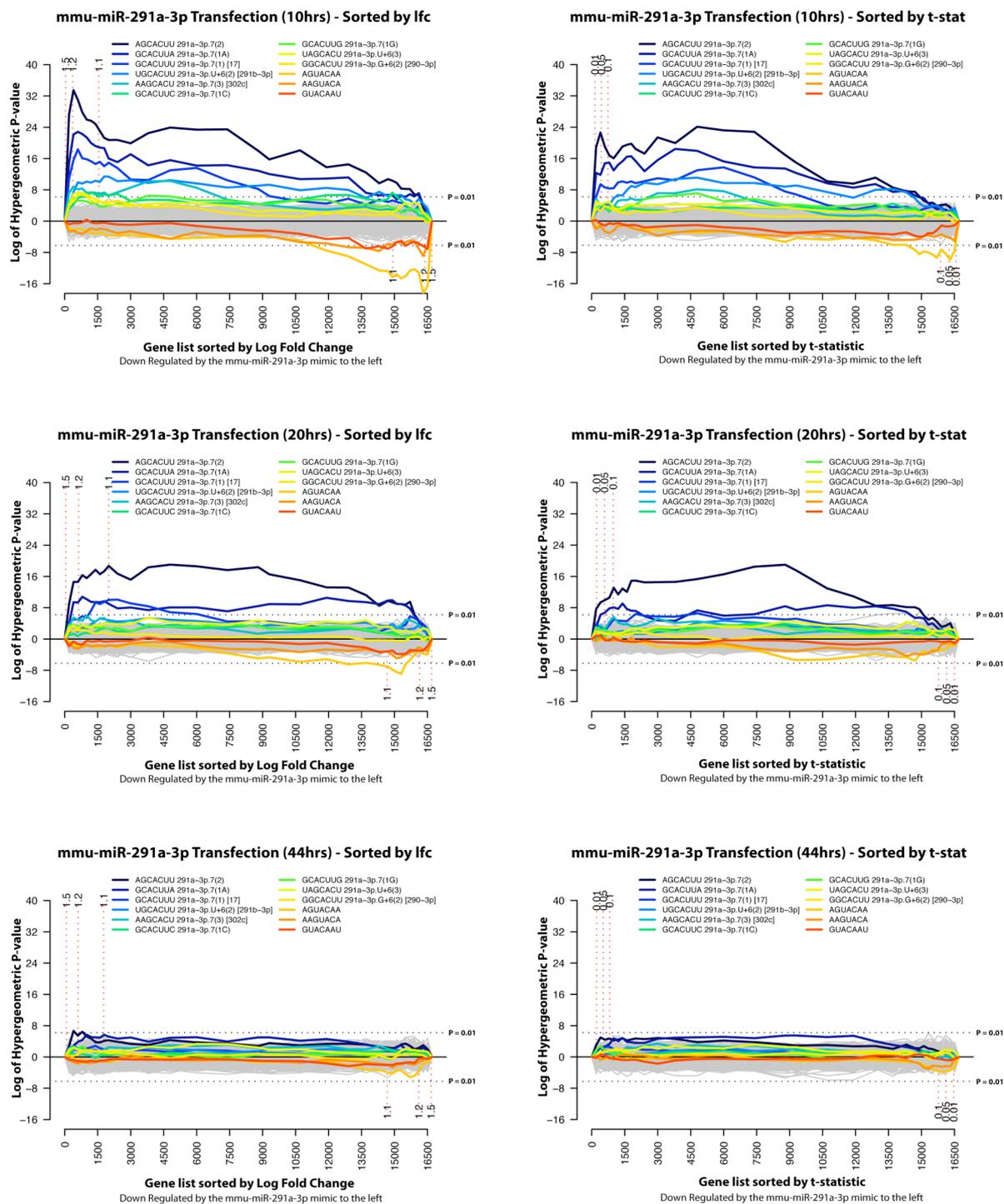


Fig.5.8: Sylamer plots querying gene lists ordered according to LFC and t-statistic following the transfection of DGCR8 depleted cells (*Dgcr8^{gt1/tm1}*) with miRNA mimics and lysed according to a time series. Two time series were conducted to find the optimal time point post-transfection at which to lyse cells and purify their RNA. Cells were transfected with mmu-miR-25 mimic (A) or mmu-miR-291a-3p mimic (B) and lysed after 10 hours, 20 hours or 44 hours. Subsequently differential gene expression was determined

relative to cells transfected with a control miRNA mimic. This differential expression was used to order the gene list according to LFC or t-statistic and Sylamer was used to query these gene lists for the enrichment of any 7mer sequences in any regions of these gene lists (down-regulated by miRNA mimic relative to control to the left of axis, up regulated to the right). See Fig.5.2 for the principle of Sylamer analysis and Sylamer plots. Significantly enriched 7mer seeds are represented by a coloured line on the plots. Sylamer analysis of these gene lists was conducted by Dr. Cei Abreu-Goodger.

5.3.3.5 miR-25 and miR-291a-3p potential target lists

I used the comparisons between the expression profiles of miRNA transfected cells and control transfected cells at the 10 hour time point as the basis for identifying relevant targets for miR-25 and miR-291a-3p. Working on the principle that the enrichment of miRNA seed sequences among those genes down regulated by the miRNA transfection was a consequence of the miRNA mimic causing the degradation of miRNA sensitive targets, I tried to select *P*-value and LFC cut offs for these comparisons that would provide a gene set most enriched for these down regulated seed bearing genes in each case. To do this I referred to the Sylamer plots of the transfection time course and selected *P*-value and LFC limits that encompassed the gene set with the greatest *P*-value for the enrichment of miR-25 or miR-291a-3p seed sequences. At the same time I bore in mind the importance of maintaining stringent criteria for the identification of significantly affected targets. As a result I selected probes that were down regulated in the presence of miRNA mimics as opposed to a control mimic transfection with a *P*-value cut off of 0.1 and a LFC cut off of $\log_2(1.2)$. This identified 83 probes demonstrating down-regulation in the case of mu-miR-25 and 31 probes in the case of mmu-miR-291a-3p.

These probes that reacted to the presence of each miRNA were further filtered by intersection with those probes that were identified as up regulated in the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* vs.

Dgcr8^{tm1.gt1/+} and *Dgcr8^{tm1.gt2/+}* expression analysis (see section 5.3.2.1 above). This requirement for the probes to be in both sets in order to be considered as targets will tend to remove any probes that may have changed due to the over expression of the transfected miRNA and may not be relevant under normal cellular conditions (as they remain relatively unchanged when miRNAs are depleted in the stem cell system). On the other hand, the transfection experiment allows the probes that change upon a broad miRNA depletion to be subsetted into those that are controlled by each stem cell expressed miRNA. Of the original 83 and 31 probes identifying down regulated genes upon miRNA transfection, 47 miR-25 affected probes and 28 miR-291a-3p probes remained after this filtering step. These correspond to 40 independent genes with annotated Entrez IDs in the case of miR-25 and 25 genes with annotated Entrez IDs in the case of miR-291a-3p. These probes and their associated annotation were amalgamated into HTML tables (Table 5.3/Appendix C (CD)) in order to aid the interpretation of these results. Some of the probes did not correspond to an annotated gene and transcript, but the probe ID is included for completeness.

Where possible the 3'UTRs of annotated transcripts associated with each of the microarray probes within the set of predicted targets was searched for the presence of target seed sequences that may account for the targeted degradation of the transcript by the transfected miRNA. Of the miR-25 potential targets, 35 out of 38 genes with annotated Havana or Ensembl transcripts and all but ENSMUST00000057037 of the miR-291a-3p annotated target transcripts contain at least one of the respective miRNA target seed 6mers within its 3'UTR. ENSMUST00000057037 itself has no associated 3'UTR and thus seed counts could not be conducted. The presence of seed sequences that may identify active miRNA target

sites reinforces the suggestion that these genes may be primary miRNA targets rather than secondary effects of miRNA mimic transfection.

There could be several reasons why the *49331047G18Rik*, *Gli2* and *Rgma* transcripts are within the miR-25 predicted target set and yet do not appear to have a seed sequence within their 3'UTR. Although both *Gli2* and *49331047G18Rik* only possess a single Ensembl annotated transcript there may be additional non-annotated alternative splices, with alternative 3'UTRs that may contain miRNA target sites not found here. *Rgma* has four transcripts annotated within Ensembl (v52). Differences in the 3'UTRs annotated for these transcripts will obviously affect seed sequences identified within these 3'UTRs. Seeds within alternative 3'UTRs could account for miRNA induced expression changes. Alternatively there may be miRNA targets within the ORFs that may be susceptible to miRNA targeting. Indeed a simple text search of the coding sequences for the transcripts for these genes annotated in Table 5.3 identified 4 GTGCAA seed sequences within the coding region of *Rgma* and 2 independent 6mer seeds (GTGCAA and GCAATA) in the *Gli2* ORF. A further possibility could be that these genes are not in fact direct targets of the miR-25 miRNA and are secondary targets affected downstream of other targets within or outside the list.

In contrast to the pervasive presence of seed sequences in the 3'UTRs of genes in these sets, TargetScan and the miRBase Targets version of the miRanda algorithm were not comprehensive in their prediction of these genes as targets of these miRNAs (Table 5.3). The databases were queried through the target genes' EntrezIDs. In the case of miRBase targets these were mapped via Ensembl IDs using the database's own tables (Dr. Stijn van Dongen). Of the probes with associated Entrez ID annotation, TargetScanS v4.2 identified 15/40 of the

miR-25 target gene set as miR-25 targets and 19/25 of the miR-291a-3p genes as targets. MiRBase Targets identified 9/40 of the miR-25 targets and 2/25 of the miR-291a-3p targets. Obviously the fact that these methods do not identify a significant proportion of my candidate targets is a cause for concern if these computational predictions are to be used as a basis for hypothesising further experimental investigation. As a consequence, these lists of candidate targets are a very useful supplement to these predictive methods. In addition, the list is a refinement on the broad and inclusive target lists derived from these computational predictions, which do not consider co-expression as a necessity for target prediction. Although it could be useful to use these candidate lists in parallel to produce lists of the most likely candidates predicted by all of the methods, before planning further research, it is also worth bearing in mind that the lists provided here are based on experimental data rather than computational prediction and as such should be more reliable when considering the role of miRNAs in ES cells.

Table 5.3: Tables summarizing the genes that are the suspected targets of mmu-miR-25 and mmu-miR-291a-3p as determined by transfection analysis. Information concerning each of the genes that was identified as a predicted target of mmu-miR-25 or mmu-miR-291a was summarised within HTML tables (See below and Appendix C (CD)). These tables contain the probe IDs that successfully fulfilled the criteria required to be associated with a miRNA target. The “Gene Symbols”, “Gene Name” and “Entrez ID” for each probe were annotated according to the *lumiMouseAll* Bioconductor library. However, where these annotations were not available, “Gene Symbols” and “Entrez ID” were annotated according to the annotations of Dr. Cei Abreu-Goodger (derived from Ensembl, Nov 2008) and marked with an asterisk. Also included are the LFCs and *P*-values associated with each array experiment used to define the set, associated GO terms and KEGG pathway information, relevant miRNA seed counts within the 3’UTRs of associated transcripts and TargetScan (v4.2) and miRBase Target predictions for each gene for the respective miRNA.

Mmu-miR-25 Probable MiRNA Targets

Illumina Probe IDs	Gene Symbols	Gene Name	Entrez ID	Log Fold Change upon Dgcr8 Depletion	P-value of change upon Dgcr8 Depletion	Log Fold Change upon miR-25 Reintroduction	P-value of change upon miR-25 Reintroduction	GO Terms	KEGG Pathways	Transcript used for 3'UTR seed counts	6mer(1A)-GCAATA	6mer(3)-GTGCAA	6mer(2)-TGCAAT	7mer(2)(m8)-GTGCAAT	7mer(1A)-TGCAATA	8mer(1A)-GTGCAATA	Target Scan Conserved Target Context Score	Target Scan Not-conserved Target Context Score	miRBase Targets P-value
iXuHIO61FICWdz.rd4	Tmem184b	transmembrane protein 184b	223693	1.23	1.0e-18	0.43	6.4e-05			ENSMUST00000074991	1	1	1	1	1	1	-0.1188	Not Found	Not Found
Bp5e4GiPe018Rk3.L4	Adam23	a disintegrin and metallopeptidase domain 23	23792	1.26	4.2e-16	0.71	1.2e-05	cell adhesion (GO:0007155) proteolysis (GO:0006508) integrin-mediated signaling pathway (GO:0007229)		OTTMUST00000004901	2	2	5	2	2	1	-0.3949	Not Found	Not Found
ikuSKIB557us.j4.5I	Plekhl1	pleckstrin homology domain containing, family M (with RUN domain) member 1	353047	0.94	5.5e-16	0.98	3.2e-05	intracellular signaling cascade (GO:0007242)		OTTMUST00000005762	2	2	2	1	2	1	-0.1376	-0.3862	Not Found
Ej5ewT_J98eO7nnqEo	Whsc111	Wolf-Hirschhorn syndrome candidate 1-like 1 (human)	234135	0.66	3.9e-15	0.56	2.0e-05	transcription (GO:0006350) regulation of transcription, DNA-dependent (GO:0006355) chromatin modification (GO:0016568)		ENSMUST000000084026	1	3	3	1	1	0	Not Found	Not Found	Not Found
OXuL8aOr_34e.kHjnk ul.p8e6gnXuJL.OheU7o	Rab8b	RAB8B, member RAS oncogene family	235442	0.61	4.7e-15	0.89	7.3e-08	transport (GO:0006810) small GTPase mediated signal transduction (GO:0007264) protein transport (GO:0015031)		ENSMUST000000041139	1	2	1	1	1	1	Not Found	-0.342	Not Found
oKJsk.kySo_5ezce7U	Gcnt2	glucosaminyl (N-acetyl) transferase 2, 1-branching enzyme	14538	0.76	1.1e-13	0.45	1.2e-04		Glycosphingolipid biosynthesis - neolactoseries (00602) Glycan structures - biosynthesis 2 (01031)	ENSMUST000000110191	1	0	2	0	1	0	Not Found	-0.046	Not Found
3uKIKTKCulcniH3Ok	Bicd2	bicaudal D homolog 2 (Drosophila)	76895	0.63	9.8e-12	0.39	1.6e-04	microtubule-based movement (GO:0007018)		ENSMUST000000048544	0	1	2	0	0	0	Not Found	Not Found	Not Found
EeMSH_Qure_2ep4910	4931407G18Rik	RIKEN cDNA 4931407G18 gene	70977	0.75	9.2e-11	0.57	3.3e-05			OTTMUST000000052375	0	0	0	0	0	0	Not Found	Not Found	Not Found
um2NaXFmc0Od13BQqg				0.55	1.2e-10	0.77	6.5e-06												
ldWmLHEDmUk16L30	Pcaf *		18519	0.55	1.5e-10	0.48	1.8e-04			ENSMUST00000000724	2	1	1	1	1	1	-0.5192	Not Found	Not Found
cuznc0PHj0yUv0Xw	Ccnc	cyclin C	51813	0.40	1.6e-10	0.93	1.2e-09	transcription (GO:0006350) regulation of transcription, DNA-dependent (GO:0006355)		OTTMUST000000010608	1	2	1	1	1	1	-0.4342	Not Found	Not Found
c0rHdgr7fpXuuvIdo	Plekkg3	pleckstrin homology domain containing, family G (with RhoGef domain) member 3	263406	0.62	1.0e-09	0.48	1.1e-04	intracellular signaling cascade (GO:0007242) regulation of Rho protein signal transduction (GO:0035023)		ENSMUST000000101279	1	0	0	0	0	0	Not Found	Not Found	Not Found
Hg_3_j_0j70SgsstdV0	Adam19 *		11492	0.65	1.6e-09	0.58	1.2e-04			OTTMUST00000012216	3	1	3	1	3	1	-0.3185	-0.0716	Not Found
WfXy3tUWSN58XTpnv4	Rnf38	ring finger protein 38	73469	0.62	2.2e-09	0.94	3.2e-05			OTTMUST000000015974	2	3	3	2	1	1	-0.3052	Not Found	Not Found
0Xgh4_noK_vkoCFoU	Acbd4	acyl-Coenzyme A binding domain containing 4	67131	0.43	3.9e-09	0.46	6.7e-05	biological_process (GO:0008150)		OTTMUST00000006883	0	1	1	1	0	0	Not Found	-0.2064	0.0318833
KmNSu8OCceiikq5ag	4930519N13Rik	RIKEN cDNA 4930519N13 gene	78177	0.45	4.8e-09	0.40	1.2e-04			OTTMUST000000037530	0	1	1	1	0	0	Not Found	-0.2048	0.00937581
ZZw4u1kd1fn5rUKI	Leprot	leptin receptor overlapping transcript	230514	0.47	4.9e-09	0.87	9.3e-07			OTTMUST000000019180	0	2	1	1	0	0	Not Found	Not Found	0.0246051
ohUKUK40unAvgCq_v0	Ric8b	resistance to inhibitors of cholinesterase 8 homolog B (C. elegans)	237422	0.47	6.7e-09	0.70	1.7e-05			ENSMUST000000038523	0	1	1	1	0	0	Not Found	-0.1647	Not Found

lenfuu6HT5eEyFenIQ	Gli2	GLI-Kruppel family member GLI2	14633	0.68	1.5e-06	0.47	6.3e-05	<p>formation(GO:0009954) floor plate formation(GO:0021508) spinal cord dorsal/ventral patterning(GO:0021513) ventral spinal cord development(GO:0021517) cerebellar cortex morphogenesis(GO:0021696) smoothened signaling pathway involved in ventral spinal cord interneuron specification(GO:0021775) smoothened signaling pathway involved in spinal cord motor neuron cell fate specification(GO:0021776) dorsoventral neural tube patterning(GO:0021904) smoothened signaling pathway in regulation of granule cell precursor cell proliferation(GO:0021938) spinal cord ventral commissure morphogenesis(GO:0021963) pituitary gland development(GO:0021983) cell differentiation(GO:0030154) lung development(GO:0030324) mammary gland development(GO:0030879) hindbrain development(GO:0030902) tube development(GO:0035295) odontogenesis of dentine-containing tooth(GO:0042475) positive regulation of neuron differentiation(GO:0045666) positive regulation of transcription(GO:0045941) positive regulation of transcription from RNA polymerase II promoter(GO:0045944) embryonic gut development(GO:0048566) developmental growth(GO:0048589) anatomical structure formation(GO:0048646) neuron development(GO:0048666) branching morphogenesis of a tube(GO:0048754) anatomical structure development(GO:0048856) notochord regression(GO:0060032)</p>	Hedgehog signaling pathway(H3340) Basal cell carcinoma(05217)	ENSMUST00000062483	0	0	0	0	0	0	0	0	Not Found	Not Found	Not Found
fuf5P1KK.Vzv6sA78k	Znrf2	zinc and ring finger 2	387524	0.29	3.3e-06	0.90	6.2e-05	biological_process(GO:0008150)		OTTMUST00000056945	1	1	1	1	1	1	Not Found	0.2721	Not Found		
HLKRZF_ztEIS698r3s	Ttc39b	tetrapeptide repeat domain 39B	69863	0.51	6.4e-06	0.68	1.8e-04			OTTMUST00000000091	2	3	3	1	2	0	Not Found	Not Found	Not Found		
TblXcNVESKpfvK1Mul	Dbt	dihydrolipamide branched chain transacylase E2	13171	0.24	2.2e-05	0.75	2.4e-06	metabolic process(GO:0008152)	Valine, leucine and isoleucine degradation(00280)	OTTMUST00000066974	0	1	1	1	0	0	Not Found	0.2365	0.000120217		
ECymiJ06TKt568y6l	Decbd1	discoidin, CUB and LCCL domain containing 1	66686	0.45	5.2e-05	1.04	4.5e-05										Not Found	Not Found	0.00284702		
xQMgij9qDrurjHh1Kk	Sfxn3	sideroflexin 3	94280	0.49	6.6e-05	0.45	7.6e-05	transport(GO:0006810) ion transport(GO:0006811) cation transport(GO:0006812) iron ion transport(GO:0006826)		ENSMUST00000062213	0	0	2	0	0	0	0	Not Found	Not Found	Not Found	
fi_CR_P_vuJrvefol	Rgma	RGM domain family, member 1A	244058	0.26	7.3e-05	0.41	3.4e-05	neuronal tube closure(GO:0001843) positive regulation of protein amino acid phosphorylation(GO:0001924)		OTTMUST00000077776	0	0	0	0	0	0	Not Found	Not Found	Not Found		

Mmu-miR-291a-3p Probable MiRNA Targets

Illumina Probe IDs	Gene Symbols	Gene Name	Entrez ID	Log Fold Change upon Dicer8 Depletion	P-value of change upon Dicer8 Depletion	Log Fold Change upon miR-291a-3p Reintroduction	P-value of change upon miR-291a-3p Reintroduction	GO Terms	KEGG Pathways	Transcript used for 3'UTR seed counts	6mer(3)-AGCACT	6mer(1A)-CACTTA	6mer(2)-GCACTT	7mer(2)(m8)-AGCACTT	7mer(1A)-GCACCTA	8mer(1A)-AGCACTTA	Target Scan Conserved Target Context Score	Target Scan Non-conserved Target Context Score	miRBase Targets P-value
Nb0eo_u_ToKS39e_w	Hp1bp3	heterochromatin protein 1, binding protein 3	15441	1.47	2.9e-21	0.65	6.4e-05	nucleosome assembly(GO:0006334)		OTTMUST00000023193	3	3	3	1	3	1	Not Found	-0.2342	Not Found
HKTHeDfR2_w4k118	Detn4	dynactin 4	67665	1.10	5.8e-21	0.92	1.0e-05			ENSMUST00000025505	1	0	4	1	0	0	Not Found	-0.208	Not Found
sXFXjc.RJjK16QSo	Arhgef18	rho/rac guanine nucleotide exchange factor (GEF) 18	102098	1.20	2.6e-19	0.56	1.9e-05	actin cytoskeleton organization and biogenesis(GO:003036) small GTPase mediated signal transduction(GO:0007264) regulation of cell shape(GO:008360) regulation of Rho protein signal transduction(GO:0035023)		OTTMUST00000076126	2	0	2	1	0	0	-0.1761	Not Found	Not Found
Zsu87Jc6Dg9d9Vei11	Chic1 *		122125	1.24	2.2e-18	0.67	1.0e-05			OTTMUST00000044104	4	5	5	1	1	0	Not Found	-0.1875	Not Found
BuBy_Ub24UCMUp76E4	Tmem87b	transmembrane protein 87B	22477	0.67	1.3e-17	0.71	1.2e-05			OTTMUST00000036395	3	0	3	1	0	0	Not Found	-0.1248	Not Found
6SVcgv15yvkUr3.d5E	Mgat4a	mannoside acetylglucosaminyltransferase 4, isoenzyme A	269181	0.66	1.6e-17	0.48	2.5e-05	carbohydrate metabolic process(GO:0006975)	N-Glycan biosynthesis(00510) Glycan structures - biosynthesis(01030)								Not Found	-0.4217	Not Found
KFiKoch3SifxL7dcS8	Tbccl	tubulin folding cofactor E-like	272589	0.79	1.4e-15	0.66	1.4e-06	protein modification process(GO:0006464)		ENSMUST00000066148	2	1	3	1	0	0	Not Found	Not Found	Not Found
Ej5ewT_J98cO7mqEo	Whsc111	Wolf-Hirschhorn syndrome candidate 1-like 1 (human)	234135	0.66	3.9e-15	0.54	3.1e-05	transcription(GO:0006350) regulation of transcription, DNA-dependent(GO:0006355) chromatin modification(GO:0016598)		ENSMUST00000084026	7	2	7	5	2	1	Not Found	-0.1427	0.0012896
kKcJNVbH4qQBK_U9ck	Gnb5	guanine nucleotide binding protein (G protein), beta 5	14697	0.54	7.0e-15	0.69	2.1e-06	signal transduction(GO:0007165) G-protein coupled receptor protein signaling pathway(GO:0007186)		ENSMUST00000076889	1	0	2	1	0	0	-0.1571	Not Found	Not Found
BXu_516p5BRyHJf14	D030056L22Rik	RIKEN cDNA D030056L22 gene	225995	0.62	9.7e-15	0.50	8.7e-06			ENSMUST00000062753	1	0	2	1	0	0	Not Found	-0.2362	Not Found
QXS1Ku6d7erks5LTug				0.73	3.1e-14	0.76	1.3e-05												
Holi8c.uF76nVr.e.E	Tor1b	torsin family 1, member B	30934	0.56	3.9e-14	0.52	2.9e-05	biological_process(GO:0008180) chaperone cofactor-dependent protein folding(GO:0051085)		OTTMUST00000027995	3	1	2	2	0	0	Not Found	-0.1008	Not Found
oyM1RN1xu0f7Xiges	Il7	interleukin 7	16196	0.67	3.0e-12	0.74	1.0e-06	T cell lineage commitment(GO:0002260) anti-apoptosis(GO:0006916) immune response(GO:0006955) regulation of gene expression(GO:0010468) positive regulation of B cell proliferation(GO:0030890) negative regulation of apoptosis(GO:0043066) negative regulation of catalytic activity(GO:0043086) bone resorption(GO:0045153) positive regulation of T cell differentiation(GO:0045582) positive regulation of organ growth(GO:0046622) homeostasis of number of cells within a tissue(GO:0048873)	Cytokine-cytokine receptor interaction(04060) Jak-STAT signaling pathway(04630) Hematopoietic cell lineage(04640)								Not Found	-0.2369	Not Found
Nf_4d307578y3pW6oQ	Bat5	HLA-B associated transcript 5	193742	0.65	4.6e-12	0.52	2.8e-06			ENSMUSEST00000055410	2	1	2	2	1	1	Not Found	-0.3073	0.00226352
								regulation of cell cycle(GO:0051726) apoptosis(GO:0006915) transcription(GO:0006350) regulation of transcription, DNA-dependent(GO:0006355)	Cell cycle(04110) Pancreatic cancer(05212) Glioma(05214)										

KeRj_XBL3vV15.j8	E2f1	E2F transcription factor 1	13555	0.73	5.0e-12	0.49	3.2e-05	cell cycle(GO:007042) forebrain development(GO:0030900) regulation of transcription(GO:0045449) positive regulation of transcription, DNA-dependent(GO:0045893) positive regulation of transcription from RNA polymerase II promoter(GO:0045944)	positive regulation of transcription(GO:0045449) Melanoma(G5218) Bladder cancer(G5219) Chronic myeloid leukemia(G5220) Small cell lung cancer(G5222) Non-small cell lung cancer(G5223)	OTTMUST00000037809	1	0	2	0	0	0	0	Not Found	Not Found	Not Found
3uKIKTKCulcniH3iOk	Bicd2	bicaudal D homolog 2 (Drosophila)	76895	0.63	9.8e-12	0.49	2.0e-05	microtubule-based movement(GO:007018)		ENSMUST00000048544	1	2	3	1	0	0	Not Found	Not Found	Not Found	
6YOkcXyoT3LPn3Pv6k	Pfn2	profilin 2	18645	1.00	6.8e-11	0.72	7.1e-05	actin cytoskeleton organization and biogenesis(GO:0030036) cytoskeleton organization and biogenesis(GO:007010)	Regulation of actin cytoskeleton(G4810)	OTTMUST00000065997	1	0	4	1	0	0	-0.1369	Not Found	Not Found	
QPvVV_Lwb_selc3E4i	Cep170	centrosomal protein 170	543389	0.56	6.0e-10	0.57	1.1e-05			ENSMUST00000057037							Not Found	Not Found	Not Found	
Nn_v556_i6DXiLhVc				0.43	1.8e-09	0.79	4.2e-06													
NT8NPGki aPjFXE9.8	Entpd7	ectonucleoside triphosphate diphosphohydrolase 7	93685	0.58	3.8e-09	0.65	3.5e-06	ribonucleoside diphosphate catabolic process(GO:0009191) ribonucleoside triphosphate catabolic process(GO:0009203)	Starch and sucrose metabolism(G0500) Folate biosynthesis(G0790)	OTTMUST00000065276	1	1	1	1	1	1	Not Found	Not Found	Not Found	
9o66jF5AB7up_T14e4	Rab11a	RAB11a, member RAS oncogene family	53869	0.46	1.6e-08	0.52	4.7e-06	transport(GO:0006810) small GTPase mediated signal transduction(GO:007264) protein transport(GO:0015031)									-0.3858	Not Found	Not Found	
QKBEi8HyKHip5q5K11	Irf35	interferon-induced protein 35	70110	0.28	4.7e-08	0.39	2.1e-05			OTTMUST00000005867	1	2	1	1	1	1	Not Found	Not Found	Not Found	
xekoCDuV.HExa3S_c	Tnfrsf1	tumor necrosis factor, alpha-induced protein 1 (endothelial)	21927	0.34	2.6e-07	0.90	6.0e-05	potassium ion transport(GO:0006813) embryonic development(GO:0009790)		OTTMUST00000000462	2	0	2	2	0	0	-0.2121	Not Found	Not Found	
o6cIXSzdUskfej_D0	4933434E20Rikas		99650	0.51	8.2e-06	0.65	1.1e-06			ENSMUST00000068798	2	0	2	1	0	0	Not Found	-0.2088	Not Found	
lq_eOICErKvpo8ESI	Zbtb41	zinc finger and BTB domain containing 41 homolog	226470	0.20	3.3e-05	0.68	4.5e-05	transcription(GO:0006350) regulation of transcription, DNA-dependent(GO:0006355)		OTTMUST000000051928	4	5	5	3	2	2	-0.2492	Not Found	Not Found	
6X9TIVN87_f_OKfgo				0.42	1.1e-04	0.38	6.5e-05													
lnt_R3rt4ml4l9Vq4	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	12575	0.68	5.0e-04	0.94	2.6e-05	regulation of cell cycle(GO:0051726) response to DNA damage stimulus(GO:0006974) cell cycle arrest(GO:0007050) negative regulation of cell proliferation(GO:0008285) response to UV(GO:0009411) positive regulation of B cell proliferation(GO:0038890) negative regulation of apoptosis(GO:0043066) positive regulation of non-apoptotic programmed cell death(GO:0043071) negative regulation of cyclin-dependent protein kinase activity(GO:0045736)	ErbB signaling pathway(G4012) Cell cycle(G4110) p53 signaling pathway(G4115) Glioma(G5214) Prostate cancer(G5215) Melanoma(G5218) Bladder cancer(G5219) Chronic myeloid leukemia(G5220)	OTTMUST00000078425	2	1	3	2	1	1	Not Found	-0.3089	Not Found	
ud77JUnSnu2p_l_gcM	Gtf3c2	general transcription factor IIIc, polypeptide 2, beta	71252	0.18	1.8e-03	0.58	5.5e-06	transcription(GO:0006350)		OTTMUST000000053730	4	1	2	2	1	1	Not Found	-0.1006	Not Found	

5.3.3.6 Comparison of genes in the miR-291a-3p candidate target list to a previously published experimentally predicted miR-290 cluster target list

Sinkkonen *et al.* recently conducted a similar study in Dicer-deficient mouse ES cells (Sinkkonen et al., 2008). They compared the transcriptional profiles of *Dicer1*^{-/-} cells to *Dicer1*^{+/-} cells and then searched the 3'UTR sequences of the up regulated transcripts for the enrichment of 7mer motifs. They identified a list of 7mer sequences very similar to the list of enriched seeds that were identified in the up-regulated transcripts from the *Dgcr8*^{gtm1,gt/+} and *Dgcr8*^{gt/tm1} comparison with Sylamer (Fig.5.4C). They found 7mer motifs GCACUUU, AGCACUU, GCACUUA, UGCACUU and AAGCACU to be enriched, which correspond to miR-17 (2) (and miR-291a-3p (bases 1-7)), miR-291a-3p (2), miR-291a-3p (1A), miR-291b-3p (2) and miR-302b (2) respectively. GCACUUU, AGCACUU and GCACUUA were all amongst the top motifs enriched within my comparison, although there was an additional enrichment of the miR-25 seeds and a miR-301a seed.

Next they transfected the *Dicer1*^{-/-} cells with miRNAs of miR-290 cluster (mmu-miR-290-5p, 291a-3p, 292-3p, 293, 294 and 295) all of which, except mmu-miR-290-5p, share similar seed sequences. They therefore conducted an experiment like that described in this chapter. Upon comparison of the expression profiles of the miRNA transfected cells to those of the control transfections Sinkkonen *et al.* again found an enrichment of miR-291a-3p seed sequences, this time amongst those genes down regulated, in agreement with my results described above. They then intersected those genes up regulated in the knockout cells with those down regulated upon the reintroduction of the miRNAs and filtered this list for genes that contain the hexamer

GCACUU seed within their 3'UTR. The remaining list of predicted targets contained 253 genes.

A comparison between my list of miR-291a-3p predicted targets and the Sinkkonen *et al.* miR-290 cluster target predictions reveals an overlap of 9 genes (*Hplbp3*, *Dctn4*, *Tbcel*, *D030056L22Rik*, *Il7*, *Pfn2*, *Tnfaip1*, *Zbtb41* and *Cdkn1a*). Significantly, both lists predict *Cdkn1a* to be a target of these miRNAs. *Cdkn1a* has indeed been recently demonstrated to be a miR-291a-3p target (Wang *et al.*, 2008).

There could be many reasons for these discrepancies in both the number and identity of predicted targets. First of all there are several major differences between these experiments. The most important is that I transfected single miRNAs in order to generate my predicted target lists as opposed to a set of miRNAs. Although many of the miRNAs transfected by Sinkkonen *et al.* do share similar seed regions, mmu-miR-293-3p has a 2nt shift in its seed and mmu-miR-290-5p has an entirely different seed sequence. These two alterations could be expected to result in the targeting of entirely different mRNA transcripts. As miRNAs can co-regulate genes, combinatorial regulation could cause increased repression and degradation of some targets beyond that which would be expected if they were to be targeted by a single miRNA complementary to a single seed target site within the same UTR. Furthermore Sinkkonen *et al.* harvested the RNA from their cells 24 hours post-transfection as opposed to the 10 hours for which I cultured my cells. Assuming that both cell lines react with the same efficiency to the transfected duplexes and given the results of my optimization experiments, it could be that this period of 24 hours results in the

enrichment of their target lists with a greater number of secondary targets than lists generated by the method described here.

However, it is worth noting that Sinkkonen *et al.* identified potential targets with *P*-value cutoffs of 0.001, as opposed to 0.1 – 0.05 used by my analyses. The more stringent *P*-value criteria were in part made possible by the use of triplicate, rather than duplicate transfections. It is possible that further replicates of miR-291a-3p transfections would allow me to expand my miRNA target lists, as *P*-values could be expected to become more significant. Sinkkonen *et al.* also worked without LFC criteria, which may have expanded their lists of deregulated genes.

There may have also been differences in the transfection efficiencies between the two experiments. The greater over expression of a miRNA duplex within a cell could lead to far greater perturbations of target genes and may deregulate genes with far from optimal target sites in their transcripts, although to some extent off target effects would be remedied by the intersection of the array experiment results. On the other hand, a lower transfection efficiency could mean that the signal from down regulated transcripts may be dilute within the RNA pool by the transcripts from non-transfected cells whose RNA levels remain unaltered.

Finally, as described in Chapter 3, ES cells of different origins and cultured for different periods of time can accumulate genetic differences and these themselves may affect gene expression and hence predicted target lists. In future it may be informative to repeat the transfection experiments described in this chapter with the

Dgcr8^{gt2/tm1} cell line in order to replicate results in an independent gene trap background that has been cultured in parallel.

5.3.3.7 Candidate targets of particular interest

The annotated GO terms and KEGG pathways amongst these predicted target gene sets contain some interesting trends. There are 4 genes associated with the cell cycle (*Pols* and *Hus1* for miR-25; *E2f1* and *Cdkn1a* for miR-291a-3p). The implication is that although the miR-290 cluster's regulation of *Cdkn1a* (p21) has recently been demonstrated as a fundamental cause of the G1/S phase transitional block following DGCR8 depletion, I would concur with the authors of that study that the deregulation of this transition is unlikely to be caused solely by a single gene (Wang et al., 2008).

Hus1, a potential miR-25 target, is involved in the recognition of DNA damage and the activation of a chain of reactions that can culminate in the phosphorylation of p53 (*Trp53*) and the subsequent transcriptional up-regulation of p21 and a block at the G1/S phase transition (Niida and Nakanishi, 2006). However, while miRNA mediated degradation of the transcripts of these two proteins in wild type cells might be expected to maintain cell cycling, *E2f1*, which encodes a transcription factor associated with the positive regulation of the G1 to S transition is also within the list of predicted miR-291a-3p targets (Cam and Dynlacht, 2003; Trimarchi and Lees, 2002). This at least appears to be a demonstration of the same miRNAs bearing a responsibility for the tightening of the regulation of a range of factors involved in the same cellular process, rather than as a molecular switch required to trigger a transition from one cellular state to another. It is worth bearing in mind, however, that the cell cycle involves highly complex interactions between a large number of proteins.

Obviously in such a background the influence of a miRNA and its targets can be affected by many outside factors, and as a consequence a gene list is insufficient to prescribe the effect of changes in miRNA and target levels.

As a further demonstration of the complexity involved in discerning the extent of miRNA control in G1/S phase transition, p21 has also been shown to be a transcriptional target of *Klf4* (Chen et al., 2003). As described above this marker is up regulated in the *Dgcr8* deficient cells. Whether this is a direct result of the removal of miRNA induced regulation of *Klf4* or a consequence of downstream cellular changes is uncertain, but it is clear that there are further pathways that impinge on the regulation of the cell cycle that will be influenced by miRNAs.

In addition to its role in cell cycle regulation, E2F1 can also trigger apoptosis. The recruitment of E2F1 to the promoter of p73 (*Trp73*), a pro-apoptotic gene, is enhanced by its acetylation. E2F1 stability is also enhanced by genotoxic stress induced acetylation (Ianari et al., 2004). This increase in stability is PCAF (KAT2B) dependent. PCAF is a histone acetyl-transferase also involved in transcriptional activation and also appears in the miR-25 predicted target lists presented here (Marmorstein, 2001).

The miR-106b-25 miRNA cluster is regulated by E2F1 (as they are embedded within the intron of the *Mcm7* mRNA). As a result the miR-290 cluster potentially plays a role in the regulation of the miR-106b-25 cluster via E2F1, adding to the connectivity of the miRNA network. Members of the latter cluster are also known to auto-regulate their own expression by targeting the *E2f1* mRNA and have been demonstrated to

target the p21 mRNA (Petrocca et al., 2008a; Petrocca et al., 2008b). The prospect of miR-291a-3p and miR-106b and miR-93 sharing targets is not that unusual as the first 8 bases of miR-291a-3p correspond to bases 2-9 of the other miRNAs. If these targets are indeed subsequently demonstrated to be true *in vivo* miRNA targets it would therefore appear that the regulation of *E2f1* by miRNAs occurs through both direct and indirect mechanisms all of which are tightly coordinated.

A homozygous deletion of *ADAM23*, a member of the ADAM family of proteins, which regulates cell-cell and cell-extracellular matrix interactions, has been identified in a gastric cancer cell line (Takada et al., 2005a). Subsequently a broader down regulation of *ADAM23* was seen in both gastric cancer cell lines and gastric primary tumours, although these were not necessarily associated with genomic deletions. Methylation of the gene's CpG island was attributed with a role in this down-regulation. However, a recent study has demonstrated miR-25 to be up regulated in gastric primary tumours (Petrocca et al., 2008b). *ADAM23* was a member of the miR-25 candidate target list generated in my study and given the number of miR-25 seed sequences in the 3'UTR of the gene it is intriguing to speculate that miR-25 may play a role in the down-regulation of this gene in gastric cancer. In the same study the reintroduction of *ADAM23* into gastric cancer cells seemed to suppress their growth in a colony formation assay.

A number of other tumour suppressor genes also appear within these target gene lists. Most notable is the *Fbxw7* gene that may be a target of miR-25. FBXW7 forms a part of the SCF type ubiquitin ligase complex, which targets a number of oncogenes (including c-Myc and cyclin E) for degradation (Welcker and Clurman, 2008) via the

26S proteasome. It is intriguing to speculate that the up regulation of *Fbxw7* in *Dgcr8^{gt/tm1}* indirectly cause the disruption of c-Myc and in turn lead to the disruption of c-Myc regulated targets, as seen earlier in this chapter.

Wwp2 is another ubiquitin ligase. It seems to be ubiquitously expressed but is expressed at higher levels in undifferentiated ES cells. It seems that Oct4 may be among the targets of *Wwp2* and ectopic expression of Oct4 and *Wwp2* along with Oct4 sensitive expression constructs demonstrates that *Wwp2* negatively regulates Oct4 transcriptional activity in ES cells (Xu et al., 2004). Although Oct4 levels do not seem to change significantly in cells depleted in DGCR8 it is plausible that more subtle changes could lead to a general disruption of multiple targets of the ES cell core transcriptional network which may again go some way to explaining both the differences described between *Dgcr8^{gt/tm1}* and *Dgcr8^{tm1,gt/+}* expression profiles and the phenotypes described in Chapter 3.

In addition to tumour suppressors among the predicted targets there are also oncogenes. *Whsc1L1* is one such gene that appears in both target lists with a large number of seed sequences within its 3'UTR, as well as being a miR-291a-3p predicted target by both TargetScan and miRBase Targets. It is within amplified chromosomal regions in a number of cancer types, while siRNAs directed against it lead to a 50% fewer soft-agar colonies in anchorage-independent growth assays (Tonon et al., 2005).

Finally, within the target lists there are a large number of genes involved in signal-transduction in some way, as judged by associated GO terms and KEGG pathway

terms. These include *Adam23*, *Plekhm1*, *Rab8b*, *Plekhg3*, *Gnaq*, *Gli2*, *Rgma* and *Prkar1a* within the miR-25 list and *Arhgef18*, *Gnb5*, *Il7*, *Rab11a* and *Cdkn1a* in the miR-291a-3p list. *Rgma* is potentially one of the most interesting genes in this list. BMP signaling is important in the maintenance of mouse embryonic stem cells in culture. *Rgma* is involved in the regulation of this signaling pathway by increasing BMP signaling, allowing BMP2 and BMP4 to bind the ACTR2A receptors more effectively in addition to the BMPR2 receptor (Xia et al., 2007). Clearly alterations within this signaling pathway may have a knock on effect on downstream gene regulation and phenotype.

Undoubtedly there are a number of other associations and conclusions that can be garnered from these gene lists of potential miRNA targets. It is worth noting that all conclusions are speculation and would require further experiments in order to support and confirm any subsequent hypotheses. However, these are miRNA targets supported by experimental evidence and are no longer solely computational predictions. Therefore as a resource for constructing these hypotheses it is superior to purely predictive methods.

5.4 Discussion

The depletion of miRNAs in ES cells due to the disruption of the miRNA-processing pathway leads to significant disruption of the mRNA expression profile of these cells. GO analysis identified an enrichment of genes with GO terms associated with development and morphogenesis amongst the genes up regulated in *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cells leading me to conclude that miRNAs probably play a fundamental role in the stabilization of the undifferentiated stem cell state. Indeed, by comparing

the differential mRNA expression changes identified upon the insertion of a trap into both alleles of *Dgcr8*, to the transcriptional targets of transcription factors integral to the determination of the ES cell fate, it appears that there are significant perturbations of these TF target gene sets induced by the depletion of miRNAs. In the context of these alterations in gene expression it is perhaps less surprising that the *Dgcr8^{gt1/tm1}* and *Dgcr8^{gt2/tm1}* cells exhibit some of the phenotypes presented in Chapter 3. The exact nature of the miRNA mediated control of these core transcriptional networks warrants considerable further study as it may prove fundamental to understanding ES cell pluripotency. It is worth bearing in mind, however, that there is considerable scope with which to improve the ChIP-Chip and ChIP-Seq data. As noted in Liu *et al.* there are major differences seen between available ChIP-Chip and ChIP-Seq datasets and further repeats and replications of these results may be required to refine transcriptional target sets (Liu et al., 2008b).

Subsequently I reintroduced two miRNAs into *Dgcr8^{gt1/tm1}* cells. These miRNAs are highly expressed in wild type E14 cells, have seed sequences enriched amongst those genes up-regulated upon DGCR8 depletion and consequential disruption of miRNA processing and do not share similar seed regions. By introducing these miRNAs into cells that have a reduced complement of endogenous miRNAs, I was able to generate a list of predicted miRNA target genes for each miRNA that I may not have been able to generate in the context of knockdown or over expression experiment in wild type cells. This is because within wild type cells such experiments are susceptible to the saturation of endogenous targets by expressed miRNAs or functional redundancy by miRNAs with similar seed sequences, which may mask the effect of the loss of the expression of a single, or limited set, of miRNAs.

These lists enriched for true miRNA targets do not necessarily concur with computational target predictions. This is not necessarily surprising as the computational algorithms accept that the lists they present will likely be neither comprehensive nor noise free. The discrepancies seen between the predictions of different algorithms are an illustration of the problems inherent with computational extrapolation of experimentally derived targeting rules (Sethupathy et al., 2006). Indeed I believe it is widely accepted that the rules used to predict miRNA targets are not necessarily comprehensive as demonstrated by the continuous improvements that are made to these predictive methods as the data available improves and expands. Despite this there are a number of previously described miRNA targets that do not appear in my target lists such as *Bim* for miR-25 and *Lats2* and *Rbl2* for miR-291a-3p (Wang et al., 2008) as well as a set of predicted targets for the miR-290 cluster presented Sinkkonen *et al.* identified by a method similar to that presented here (Sinkkonen et al., 2008). Although there could be a number of reasons why various targets are not identified within my target lists, including an absence of mapped probes on the array and expression and functional differences between cell lines, it is also likely that the target lists presented here are not comprehensive. It is probable that the number of genes identified as targets would increase if further miRNA transfection replicates were conducted, which would improve the statistical power of the significant probe identification. Although it is possible that, as a consequence of this limitation, these lists are enriched for the most significant miRNA targets, if longer and less noisy lists were desired, I would recommend conducting these replicates at a later date.

However, despite these differences the method presented in this chapter also successfully confirmed a previously identified miRNA target gene as a probable miR-291a-3p target (*Cdkn1a*) as well as independently identifying 9 genes from the Sinkkonen *et al.* predicted miR-290 cluster targets as miR-291a-3p potential targets (*Hplbp3*, *Dctn4*, *Tbc1l*, *D030056L22Rik*, *Il7*, *Pfn2*, *Tnfrsf1*, *Zbtb41* and *Cdkn1a*) (Sinkkonen *et al.*, 2008). In addition the gene lists presented here predict a large number of additional, functionally relevant, miRNA target relationships and allow the description of a large number of hypotheses that warrant further investigation or confirmation.

Ultimately, the system presented here is simple and inherently scalable to generate a large number of gene lists enriched for a specific miRNA's target genes. This would provide a useful resource for generating target lists based on experimental evidence from which to draw patterns and hypotheses for further study.

Chapter 6: Discussion

The aim of this project was to develop a cell based system for the identification of miRNA targets on a large scale in a cellular system. The system was intended to resemble that of Giraldez *et al.*, which was used to investigate the role of miRNAs in zebrafish (Giraldez et al., 2006). The principle of the method used is that miRNAs introduced into a biological system will elicit expression changes in target mRNAs detectable by microarray. However, use of an *MZDicer* mutant fish allowed the derivation of candidate miRNA target lists in a cellular background that was depleted of endogenous miRNA expression. As a consequence the identification of candidate target mRNAs, through the reintroduction of duplexes by injection, would not be affected by target site saturation by endogenous miRNAs or complicated by the combinatorial relationships inherent in the miRNA mediated regulatory network. In addition, unlike the knockdown of individual miRNAs the system is not affected by the functional redundancy between miRNA family members that share the same seed regions, which are expected to target the same mRNAs.

6.1 Disruption of the *Dgcr8* locus to deplete mature miRNAs in mouse ES cells

To this end, using mouse ES cells with a gene trap inserted within a single allele at the *Dgcr8* locus, which I obtained from Bay Genomics (Nord et al., 2006; Stryke et al., 2003), I have disrupted the second allele with a targeted trap inserted by homologous recombination. Consequently I used the expression of marker genes on each of the traps, Northern blots and RT-PCR to confirm the configuration of these traps within the locus and in the case of the Northern Blot I demonstrated a dramatic

reduction of the expression of wild type *Dgcr8* in the *Dgcr8^{gt/tm1}* cells. I also identified a substantial reduction in ES cell miRNA expression (Houbaviy et al., 2003) through the use of Northern blots, which confirmed the functional significance of the mutations at the *Dgcr8* locus. Western blots and immuno-staining confirmed that the *Dgcr8^{gt/tm1}* cells were still expressing ES markers; transcription factors at the centre of the regulatory network which modulates pluripotency and self renewal.

The phenotypic effects of the disruption of the miRNA processing pathway in ES cells broadly resembles the phenotypes seen by others both published before and since this project was begun (Fukagawa et al., 2004; Murchison et al., 2005; Wang et al., 2007). Notably that *Dgcr8^{gt/tm1}* cells appear to accumulate in the G1 phase of the cell cycle and are unable to form wild type EBs, although like the *Dgcr8* knock out cell line, they do appear to differentiate with a degree of success, forming cells with various, conspicuous morphologies. How this change in morphology is reflected with respect to the expression of markers of differentiation and the silencing of ES cell markers in the *Dgcr8^{gt/tm1}* cells is a subject for future research.

6.2 The small RNA profile of *Dgcr8^{gt/tm1}* and *Dgcr8^{tm1,gt/+}* cells

I used the highly sensitive Illumina/Solexa RNA sequencing platform to sequence small RNA libraries to confirm the depletion of miRNA expression in the *Dgcr8^{gt/tm1}* cells. This was an important step, since ES cells are pluripotent and capable of differentiation. As a consequence the disruption of *Dgcr8* may have caused the cells to change or adapt, altering their expression profile. In this case, an incomplete depletion of DGCR8 may have been sufficient to trigger cellular alterations without significantly depleting miRNA processing. As a result a limited miRNA profile by

Northern blot may have simply missed miRNAs still expressed or newly expressed in these cells. The Solexa/Illumina sequencing of the miRNA profiles of the *Dgcr8^{gt/tm1}* and *Dgcr8^{tm1,gt/+}* cells and the comparison of these profiles confirmed that this was not the case and demonstrated a significant depletion of the expression of the vast majority of miRNAs in these cells. In addition, by comparing the raw maximum mapped sequence depths across a range of ncRNA species annotated in Ensembl it was clear that these species seemed to be proportionally represented in both the *Dgcr8^{gt/tm1}* and *Dgcr8^{tm1,gt/+}* cells. This adds weight to the observation that DGCR8 does not appear to play a significant role in the processing of other non-coding RNA species, including rRNAs, for which Drosha has been implicated as a processing enzyme (Wu et al., 2000).

6.2.1 DGCR8 independent miRNA processing

Finally, the Solexa/Illumina sequencing of the small RNA populations allowed me to identify a number of miRNAs that appear to be processed, at least in part, in a DGCR8 independent fashion; let-7a-1, let-7a-2, miR-98, miR-720, miR-689-2, miR-712, miR-1196, miR-712*, miR-1186, miR-805, miR-344-1 or 2, miR-320, miR-668, miR-877 and miR-484. These include a number of DGCR8 independent miRNAs proposed by Babiarz et al. (miR-320, miR-344, miR-668, miR-877 and miR-484) (Babiarz et al., 2008). Of these *miR-877* has been suggested as a mammalian mirtron through bioinformatics analysis (Berezikov et al., 2007). To this end both the results presented in this thesis and the work of Babiarz *et al.* confirm that this miRNA does not require DGCR8 to be processed. As our understanding of miRNAs and their underlying functional mechanisms progresses it is increasingly clear that a number of the canonical rules previously believed to apply to all miRNAs are prone to

exceptions. The microprocessor dependent release of pre-miRNAs from their pri-miRNAs is clearly one such rule and is deserving of further research.

6.3 Generating miRNA candidate target lists

In order to identify miRNA candidate target lists I selected probes by intersecting those up regulated upon DGCR8 depletion with those down regulated upon the re-addition of individual miRNAs to *Dgcr8^{gt1/tm1}* cells. The intersection with the up regulated probe lists would constrain the candidate targets to those relevant to the ES cell system and remove those demonstrating a significant change due to the potential over-expression of the added miRNA. Conversely, by restricting the list to those down regulated upon the re-addition of specific miRNAs, the list would be limited to a large extent to specific miRNA primary targets. This would remove any genes that contain predicted but non-functional target sites whose expression is altered upon miRNA depletion due to secondary effects as a consequence of the expression and phenotype of the cells adapting to their new miRNA transcriptome.

6.3.1 The influence of miRNAs on the ES cell transcriptome

The comparison of the expression profiles of the *Dgcr8^{gt1/tm1}* and *Dgcr8^{tm1,gt/+}* clearly revealed an enrichment of miRNA seed sequences within the 3'UTRs of genes up regulated upon the depletion of functional *Dgcr8*. This enrichment includes the seed sequences of miRNAs from amongst those most highly expressed in the *Dgcr8^{tm1,gt/+}* cells but depleted in the *Dgcr8^{gt1/tm1}* cells, as judged by Sylamer analysis (van Dongen et al., 2008). Although this enrichment of miRNA seed sequences amongst the genes up regulated upon miRNA depletion implies that the miRNAs in these sets regulate a broad spectrum of mRNAs in wild type ES cells, it does not exclude other miRNAs

from playing equally important roles in ES cell transcriptome regulation. Indeed it seems likely that other miRNAs can regulate smaller sets of highly influential genes that may be more difficult to detect through miRNA seed sequence enrichment analysis.

With respect to the large gene set significantly up regulated following the depletion of *Dgcr8* from the ES cell system (3251 probes were up regulated, P -value < 0.05 , LFC $> \log_2(1.1)$), it would appear that miRNAs, between them, do indeed play a broad regulatory role in ES cells, which is to be expected given the number of miRNA targets predicted by computational algorithms (Friedman et al., 2009; Griffiths-Jones et al., 2008). What is perhaps more surprising therefore is the maintained expression of ES cell specific markers in these mutant cells rather than the catastrophic disruption of the transcriptional network required to maintain ES cell identity. It does seem however that miRNAs do influence the targets of this core transcriptional network (including *Oct4*, *Sox2*, *Nanog*, *Klf4* and *c-Myc*), as determined by an examination of the differential expression of their targets. What is less clear is whether this altered expression results from miRNAs directly interacting with the transcriptional targets of the TF, with the mRNAs of the TFs themselves or with factors upstream of the TFs that may alter their expression. Given the profligate nature of miRNA:target relationships, it is reasonable to predict that all three may be of an influence. Although there was no apparent enrichment of miRNA seed sequences amongst the TF target gene mRNA 3'UTRs, this does not rule out miRNA target enrichments on a scale below the sensitivity of this method. In addition there were apparent changes in the expression of several of the TFs themselves as assessed by Illumina array. These changes would require confirmation by a more sensitive and

specific method, but potential miRNA regulation of these TFs clearly does warrant further research to aid the understanding of the regulation of ES cell pluripotency. Indeed such relationships have previously be implied in non-differentiated ES cells although, to my knowledge, they are yet to be confirmed (Singh et al., 2008).

6.3.2 The re-addition of miRNAs to the DGCR8 deficient ES cells and miRNA candidate target lists

Finally I optimized the protocol for the re-addition of miRNA mimics to the *Dgcr8^{gt1/tm1}* cells and reintroduced miR-25 and miR-291a-3p miRNAs into the cells by transfection. Subsequently I used Illumina mRNA expression microarrays to judge which genes are down regulated in the presence of these miRNAs with reference to cells transfected with a control duplex. As described above I then used these expression changes to determine candidate target mRNAs by intersecting the probe list with those probes up regulated upon DGCR8 depletion. This intersection followed by the removal of probes identifying the same gene and probes not annotated with an Entrez ID left 40 miR-25 candidate target genes and 25 miR-291a-3p candidates. The miR-291a-3p target candidates and a target candidate list for the miR-290 cluster generated by Sinkkonen *et al.* had an overlap of 9 genes (9/40 of my candidates and 9/253 of the Sinkkonen list) (Sinkkonen et al., 2008). The limited overlap may be a consequence of fundamental differences between the experiments, the most critical of which includes their transfection of a cluster of miRNAs, compared to my transfection of a single miRNA. Although the miRNAs of Sinkkonen *et al.* do predominantly share similar seed sequences (excluding miR-290-5p), the differences in seeds combined with combinatorial regulation could result large differences in the mRNA target identified by their study. Despite this both studies identified *Cdkn1a* as

a target of these miRNAs; a target that has been confirmed in independent study (Wang et al., 2008).

All but 3 of the candidate genes identified in my study, for which I derived annotated Ensembl or Vega transcripts with an associated 3'UTR, contained at least a single 6mer seed sequence corresponding to the relevant transfected miRNA that may identify the miRNA target site responsible for miRNA induced transcript degradation (Lim et al., 2005). Of the remaining three genes that do not contain a seed in the 3'UTR, two of the selected transcripts contained relevant 6mer seed in their ORF, which again may be prognostic of target sites (Grimson et al., 2007; Tay et al., 2008). The final gene's transcript contained no obvious seed sequences in the ORF or the 3'UTR. This gene may have alternative non-annotated transcripts with differing exons and 3'UTRs not searched here. Another possibility is that this gene is not a true miRNA target, but a secondary target of upstream genes affected by miRNA regulation.

These lists should prove invaluable for guiding research as they are experimentally derived and relevant in an ES cell context. As such they contain a number of interesting target genes worthy of further research for the roles that they may play in stem cell physiology. While the goal of this research was to generate large, inclusive target list for miRNAs, it is important to bear in mind that some of the genes in these target candidate lists may not be true targets and may have been selected as a result of experimental noise. If definitive relationships and pathways are to be concluded it would be necessary to conduct additional experiments.

In order to generate the optimal gene lists for further study it may be worth considering those genes identified as targets by multiple, independent methods. One approach would be to follow up those genes identified both here and by Sinkkonen *et al* (Sinkkonen et al., 2008). An alternative would be to use the experimentally identified gene lists presented in this thesis to refine the candidate genes that are predicted by target prediction programmes such as TargetScan (Friedman et al., 2009). 34/65 genes I identified as candidate miRNA targets were also predicted by TargetScan version 4.2. A final alternative would be to follow up those genes demonstrating the most significant expression changes upon miRNA removal and re-addition in preference to other targets. These shorter lists of genes would be a useful starting point for future work.

6.4 Future work

6.4.1 Expanding the system for the generation of candidate target lists

I have developed a system that is easily scalable in order to generate candidate target lists for a large number of miRNAs in a relatively short period of time. As such it would be interesting to generate target lists for a broader set of miRNAs. I would envisage these including miR-17, miR-20a, miR-93, miR-106a or miR-106b as the seed sequence of these miRNAs was also clearly enriched amongst the 3'UTRs of genes up regulated in *Dgcr8^{gt/tm1}* cells when compared to *Dgcr8^{tm1,gt/+}* cells. Whether more than one of these miRNAs should be included in a further set of experiments would be interesting to consider. As they share seed sequences, they are likely to share the majority of their targets (Friedman et al., 2009). Consequently it is difficult to know whether the methods presented here would be sufficiently sensitive to discern

whether members of the resulting candidate targets are likely to be specific to single miRNA. It is possible that through the expansion of the study in this way, features that discern targets beyond the seed sequence could be identified in order to discriminate between the targets of miRNAs with shared seeds. Other miRNAs worthy of transfection would include those most highly expressed in the *Dgcr8^{tm1.gt/+}* cells. This large set of miRNAs could also be expected to regulate genes important for maintaining the undifferentiated state of ES cells. Finally it would be interesting to co-transfect a number of these miRNAs with differing seed sequences. This would allow an investigation of the extent and significance of coordinate regulation of miRNA targets by multiple miRNAs.

Other refinements could be adopted to improve the candidate gene lists. Further transfection replicates of miR-25 and miR-291a-3p would be expected to increase the statistical significance of any replicated expression changes, potentially improving the sensitivity of the method. Furthermore, repeating the transfections in the independently derived *Dgcr8^{gt2/tm1}* cell line would again add a further refinement to the prediction lists. Ultimately an approximation of the number of false positives present in each of these lists could be gleaned through reporter gene assays the principle of which has been demonstrated in many published studies (Giraldez et al., 2006; Lewis et al., 2003). In brief, 3'UTRs containing predicted target sites would be inserted into a plasmid downstream of a reporter gene ORF. This plasmid is then co-transfected into HeLa cells with a miRNA mimic. The reporter gene expression is then compared to that of a similar experiment but in which the predicted target site contains point mutations within the target seed region. A significant increase in

reporter gene expression in cells transfected with the mutated plasmid can be used as an indication of the presence of a true miRNA target site.

An adaptation of the pSILAC method used to assess expression changes at the level of the proteome would allow me to assess the contribution of translational inhibition to target regulation for each transfected miRNA in a system with all of the advantages associated with the depletion of endogenous miRNAs (Baek et al., 2008; Selbach et al., 2008). The use of this method in the place of, or in addition to the Illumina microarray platform could prove an interesting addition to this study and it would be fascinating to determine whether the targets of particular miRNAs are more prone to translational inhibition than the targets of others and whether miRNA sequence or other factors may influence this.

6.4.2 Novel miRNAs and siRNAs expressed in mouse ES cells

The Solexa/Illumina sequenced small RNA libraries generated in this study would also allow me to further investigate the small RNA populations of both wild type and DGCR8 depleted mouse ES cells in ways that are beyond the scope of this thesis. Deep sequencing of small RNA populations have recently allowed the identification of both novel miRNAs in mouse ES cells (Calabrese et al., 2007) and endogenously expressed siRNAs in mouse ES cells (Babiarz et al., 2008) and in mouse oocytes (Tam et al., 2008; Watanabe et al., 2008). In order to investigate the expression of these novel short RNAs more completely, the sequence libraries derived from this study should be mapped against the complete mouse genome. Regions within the genome in which short sequence reads are found to accumulate but which are not annotated as a known miRNA locus could be investigated further in order to

determine a mechanism by which they may originate. The local secondary structure of a transcript from the region could be used to identify whether the reads may originate from a canonical miRNA hairpin. In addition genomic features such as inverted repeat elements, diverging or converging gene structures or pseudogenes mapped at alternative loci all potentially suggest the generation of dsRNA molecules that may form Dicer substrates for the production of siRNAs that would map to this location.

6.4.3 The roles of DGCR8 independent miRNAs

Additional experiments could be conducted to ascribe roles to the DGCR8 independent miRNAs. Transfection of the *Dgcr8^{gt1/tm1}* cells with Locked Nucleic Acid (LNA) antagomirs, that bind specific miRNAs and block their function, could be used to block selected miRNAs (Chan et al., 2005). As this experiment would be performed in a cellular background depleted of DGCR8 it would benefit from the same associated advantages concerning a reduction of functional redundancy. Once again microarrays could be used to identify transcripts that are up regulated following this functional block to compile candidate target lists for these miRNAs.

6.4.4 Prospects for improved pri-miRNA annotation

Currently intergenic miRNA primary transcripts are relatively poorly annotated. Although experiments have been conducted to specifically annotate a small set of these miRNAs (for example (Cai et al., 2004; Houbaviy et al., 2005; Lee et al., 2004)) and a recent effort has been made to compile relevant annotation data to annotate the structure of a much broader set of these pri-miRNAs (Saini et al., 2008), there is still requirement for further experimental evidence to support these annotations. One of the side effects of a depletion of functional DGCR8 is expected to be an accumulation

of unprocessed pri-miRNA transcripts (Wang et al., 2007). Previously Illumina/Solexa sequencing of fragmented poly-A selected transcripts from mouse tissues has identified mouse primary transcripts (Mortazavi et al., 2008). By using this technique to sequence the poly-A transcripts of the *Dgcr8^{gt/tm1}* cells it may be possible to derive an RNA-seq transcriptome profile enriched for pri-miRNA sequences to aid their annotation.

6.4.5 Extending the investigation of the cellular phenotype upon mature miRNA depletion

There are several improvements and extensions to the phenotypic profiling of the *Dgcr8^{gt/tm1}* cells that I would wish to include as part of my future studies. As described in Chapter 3 I wish to determine a growth curve for both the *Dgcr8^{tm1,gt/+}* and *Dgcr8^{gt/tm1}* cells prior to further cell cycle analysis. This would allow me to ensure that all of the cell samples subsequently processed for cell cycle profiling are in the exponential phase of their growth curve. It would also provide an additional comparison of the growth rate of the cells to supplement the cycle profile. I also wish to further my analysis of the potential of the *Dgcr8^{gt/tm1}* cells to differentiate. Initially I would like to assess the expression of ES cell markers and differentiation markers in these cells following EB differentiation through qRT-PCR or immunostaining. Wang *et al.* noted that *Dgcr8* knock out ES cells were unable to silence ES cell markers upon the induction of differentiation (Wang et al., 2007). Sinkkonen *et al.* noticed a similar failure to repress Oct4 expression upon the induction of differentiation in Dicer knockout ES cells (Sinkkonen et al., 2008). They later ascribed this inability to silence Oct4 to a failure to methylate the Oct4 promoter during differentiation, a process indirectly modulated by the miR-290 cluster. I would like to confirm a similar

phenotype in my *Dgcr8^{gt1/tm1}* cells. This will allow me to determine if my *Dgcr8^{gt1/tm1}* cells are comparable to knockouts generated in other studies in this respect. In addition it may be informative to generate chimeric embryos from the *Dgcr8^{gt1/tm1}* cells through injection into blastocysts. The *Dgcr8^{gt1/tm1}* cells express lacZ and should therefore be traceable in the mouse embryo making it relatively easy to determine to which tissues, if any, the cells are able to contribute.

6.4.6 An investigation into the role of miRNAs in ES cell adhesion

Considering the enrichment of “ECM-receptor interaction” and “focal adhesion” GO terms among those genes significantly up regulated upon the depletion of DGCR8, which include several integrins, and in light of the morphological phenotype and the inability of the homozygous mutant cell lines to assemble into EBs it would be interesting to explore the adhesive properties of the *Dgcr8^{gt1/tm1}* cells further. Initially, *Dgcr8^{gt1/tm1}* and *Dgcr8^{tm1,gt/+}* cells could be plated onto culture plates coated with different ECM substrates in a fashion similar to that used by Ohnishi *et al.* (Ohnishi *et al.* 2008). This could then be used to identify substrates upon which the cells differentially adhere through fixing attached cells and staining them with crystal violet. Subsequent extraction of the dye and the measurement of its absorbance would provide a reading proportional to the number of the attached cells (Ohnishi *et al.* 2008). A comparison between the adhesive properties of *Dgcr8^{gt1/tm1}* cells transfected with a set of ES cell miRNAs and control miRNAs would also provide an assay that could be used to identify miRNAs that may rectify any adhesion phenotype in a screen similar to that used by Wang *et al.* which assigned miRNAs with roles associated with the regulation of the ES cell cycle (Wang *et al.* 2008). Once miRNA candidates have been selected, arrays could be used to formulate target lists and

ascribe a more explicit role to the miRNAs through the functional investigation of the potential targets.

6.4.7 Use of a conditional mutation in an alternative system

There is a risk when using constitutively mutated cell lines that compensatory mutations or changes in expression may accumulate that are secondary to the initial mutation. This has been suggested as the cause of the recovery of Dicer knock out cells from an initial growth deficit following the mutation of the Dicer locus (Murchison et al., 2005). One method by which irreversible secondary effects could be identified would be to attempt to rescue the cell lines through the re-expression of *Dgcr8* in the mutated cells. This could be achieved through either the transfection of a transiently expressed cDNA or through the insertion of this cDNA at a random or targeted locus.

An alternative approach would be to use the Bay Genomics gene trapped cell lines as a basis from which to derive a conditional knock out cell line. One option would be to flank essential exons of the second allele at the *Dgcr8* locus with loxP sites. The loxP system would allow the excision of the intervening DNA upon the expression of Cre recombinase. Expression of a Cre protein fused to a mutant oestrogen hormone binding domain allows the activity of the Cre recombinase to be controlled through the addition of an oestrogen analogue to the cell culture medium (Hameyer et al., 2007; Vooijs et al., 2001). In this way a mutation can be triggered within the second allele of *Dgcr8* upon demand and the subsequent alterations to the cell phenotype can be monitored over time to discriminate the initial effects of miRNA depletion from the effects of compensatory changes.

6.5 Conclusion

In this thesis I have described a method with which to conduct large-scale analyses of miRNA targeting in mouse ES cells. In addition I have used this methods to derive target candidate lists for 2 of the miRNAs highly expressed in these ES cells. It is becoming increasingly well understood that miRNAs play an essential role in modulating ES cell pluripotency and differentiation (Singh et al., 2008; Sinkkonen et al., 2008; Tay et al., 2008; Wang et al., 2007). Therefore, although it is not feasible to discern the intricate details of the individual miRNA-target relationships within the time-scale of my doctoral thesis, I strongly believe that both the candidate target lists and the underlying method for their generation should be of interest and assistance to a large body of researchers. Therefore I hope that the publication of the data will make it widely available to the scientific community and aid an understanding of the function of miRNAs both in mouse ES cells and in general.

References

- Adams, D. J., Quail, M. A., Cox, T., van der Weyden, L., Gorick, B. D., Su, Q., Chan, W. I., Davies, R., Bonfield, J. K., Law, F. et al.** (2005). A Genome-Wide, End-Sequenced 129Sv BAC Library Resource for Targeting Vector Construction. *Genomics* **86**, 753-758.
- Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N. and Lovell-Badge, R.** (2003). Multipotent Cell Lineages in Early Mouse Development Depend on SOX2 Function. *Genes Dev.* **17**, 126-140.
- Axtell, M. J. and Bowman, J. L.** (2008). Evolution of Plant microRNAs and their Targets. *Trends Plant Sci.* **13**, 343-349.
- Azuma-Mukai, A., Oguri, H., Mituyama, T., Qian, Z. R., Asai, K., Siomi, H. and Siomi, M. C.** (2008). Characterization of Endogenous Human Argonautes and their miRNA Partners in RNA Silencing. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7964-7969.
- Babak, T., Zhang, W., Morris, Q., Blencowe, B. J. and Hughes, T. R.** (2004). Probing microRNAs with Microarrays: Tissue Specificity and Functional Inference. *RNA* **10**, 1813-1819.
- Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. and Blelloch, R.** (2008). Mouse ES Cells Express Endogenous shRNAs, siRNAs, and Other Microprocessor-Independent, Dicer-Dependent Small RNAs. *Genes Dev.* **22**, 2773-2785.
- Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P.** (2008). The Impact of microRNAs on Protein Output. *Nature* **455**, 64-71.
- Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K., Kroh, E. M., Bendoraitis, A., Mitchell, P. S., Nelson, A. M. et al.** (2008). MicroRNA Discovery and Profiling in Human Embryonic Stem Cells by Deep Sequencing of Small RNA Libraries. *Stem Cells* **26**, 2496-2505.
- Barad, O., Meiri, E., Avniel, A., Aharonov, R., Barzilai, A., Bentwich, I., Einav, U., Gilad, S., Hurban, P., Karov, Y. et al.** (2004). MicroRNA Expression Detected by Oligonucleotide Microarrays: System Establishment and Expression Profiling in Human Tissues. *Genome Res.* **14**, 2486-2494.
- Baskerville, S. and Bartel, D. P.** (2005). Microarray Profiling of microRNAs Reveals Frequent Coexpression with Neighboring miRNAs and Host Genes. *RNA* **11**, 241-247.
- Beddington, R. S. and Robertson, E. J.** (1989). An Assessment of the Developmental Potential of Embryonic Stem Cells in the Midgestation Mouse Embryo. *Development* **105**, 733-737.

- Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P. and Izaurralde, E.** (2006). mRNA Degradation by miRNAs and GW182 Requires both CCR4:NOT Deadenylation and DCP1:DCP2 Decapping Complexes. *Genes Dev.* **20**, 1885-1898.
- Benetti, R., Gonzalo, S., Jaco, I., Munoz, P., Gonzalez, S., Schoeftner, S., Murchison, E., Andl, T., Chen, T., Klatt, P. et al.** (2008). A Mammalian microRNA Cluster Controls DNA Methylation and Telomere Recombination Via Rbl2-Dependent Regulation of DNA Methyltransferases. *Nat. Struct. Mol. Biol.* **15**, 268-279.
- Berezikov, E., Chung, W. J., Willis, J., Cuppen, E. and Lai, E. C.** (2007). Mammalian Mirtron Genes. *Mol. Cell* **28**, 328-336.
- Bernstein, E., Caudy, A. A., Hammond, S. M. and Hannon, G. J.** (2001). Role for a Bidentate Ribonuclease in the Initiation Step of RNA Interference. *Nature* **409**, 363-366.
- Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V. and Hannon, G. J.** (2003). Dicer is Essential for Mouse Development. *Nat. Genet.* **35**, 215-217.
- Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S. F., Dunning, M. J., Barbosa-Morais, N. L., Teschendorff, A. E., Green, A. R., Ellis, I. O. et al.** (2007). MicroRNA Expression Profiling of Human Breast Cancer Identifies New Markers of Tumor Subtype. *Genome Biol.* **8**, R214.
- Borchert, G. M., Lanier, W. and Davidson, B. L.** (2006). RNA Polymerase III Transcribes Human microRNAs. *Nat. Struct. Mol. Biol.* **13**, 1097-1101.
- Bradley, A., Evans, M., Kaufman, M. H. and Robertson, E.** (1984). Formation of Germ-Line Chimaeras from Embryo-Derived Teratocarcinoma Cell Lines. *Nature* **309**, 255-256.
- Brennecke, J., Stark, A., Russell, R. B. and Cohen, S. M.** (2005). Principles of microRNA-Target Recognition. *PLoS Biol.* **3**, e85.
- Cai, X., Hagedorn, C. H. and Cullen, B. R.** (2004). Human microRNAs are Processed from Capped, Polyadenylated Transcripts that can also Function as mRNAs. *RNA* **10**, 1957-1966.
- Calabrese, J. M., Seila, A. C., Yeo, G. W. and Sharp, P. A.** (2007). RNA Sequence Analysis Defines Dicer's Role in Mouse Embryonic Stem Cells. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18097-18102.
- Cam, H. and Dynlacht, B. D.** (2003). Emerging Roles for E2F: Beyond the G1/S Transition and DNA Replication. *Cancer. Cell.* **3**, 311-316.

- Castoldi, M., Schmidt, S., Benes, V., Noerholm, M., Kulozik, A. E., Hentze, M. W. and Muckenthaler, M. U.** (2006). A Sensitive Array for microRNA Expression Profiling (miChip) Based on Locked Nucleic Acids (LNA). *RNA* **12**, 913-920.
- Chan, J. A., Krichevsky, A. M. and Kosik, K. S.** (2005). MicroRNA-21 is an Antiapoptotic Factor in Human Glioblastoma Cells. *Cancer Res.* **65**, 6029-6033.
- Chen, S., Choo, A., Wang, N. D., Too, H. P. and Oh, S. K.** (2007). Establishing Efficient siRNA Knockdown in Mouse Embryonic Stem Cells. *Biotechnol. Lett.* **29**, 261-265.
- Chen, X., Whitney, E. M., Gao, S. Y. and Yang, V. W.** (2003). Transcriptional Profiling of Kruppel-Like Factor 4 Reveals a Function in Cell Cycle Regulation and Epithelial Differentiation. *J. Mol. Biol.* **326**, 665-677.
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J. et al.** (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**, 1106-1117.
- Chendrimada, T. P., Finn, K. J., Ji, X., Baillat, D., Gregory, R. I., Liebhaber, S. A., Pasquinelli, A. E. and Shiekhattar, R.** (2007). MicroRNA Silencing through RISC Recruitment of eIF6. *Nature* **447**, 823-828.
- Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K. and Shiekhattar, R.** (2005). TRBP Recruits the Dicer Complex to Ago2 for microRNA Processing and Gene Silencing. *Nature* **436**, 740-744.
- Davis, E., Caiment, F., Tordo, X., Cavaille, J., Ferguson-Smith, A., Cockett, N., Georges, M. and Charlier, C.** (2005). RNAi-Mediated Allelic *trans*-Interaction at the Imprinted *Rtl1/Peg11* locus. *Curr. Biol.* **15**, 743-749.
- Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. and Hannon, G. J.** (2004). Processing of Primary microRNAs by the Microprocessor Complex. *Nature* **432**, 231-235.
- Desbaillets, I., Ziegler, U., Groscurth, P. and Gassmann, M.** (2000). Embryoid Bodies: An in Vitro Model of Mouse Embryogenesis. *Exp. Physiol.* **85**, 645-651.
- Doench, J. G. and Sharp, P. A.** (2004). Specificity of microRNA Target Selection in Translational Repression. *Genes Dev.* **18**, 504-511.
- Dye, M. J., Gromak, N. and Proudfoot, N. J.** (2006). Exon Tethering in Transcription by RNA Polymerase II. *Mol. Cell* **21**, 849-859.
- Ender, C., Krek, A., Friedlander, M. R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N. and Meister, G.** (2008). A Human snoRNA with microRNA-Like Functions. *Mol. Cell* **32**, 519-528.

- Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S.**(2003). MicroRNA Targets in *Drosophila*. *Genome Biol.* **5**, R1.
- Eulalio, A., Behm-Ansmant, I. and Izaurralde, E.** (2007a). P Bodies: At the Crossroads of Post-Transcriptional Pathways. *Nat. Rev. Mol. Cell Biol.* **8**, 9-22.
- Eulalio, A., Behm-Ansmant, I., Schweizer, D. and Izaurralde, E.** (2007b). P-Body Formation is a Consequence, Not the Cause, of RNA-Mediated Gene Silencing. *Mol. Cell. Biol.* **27**, 3970-3981.
- Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M. and Izaurralde, E.** (2009). Deadenylation is a Widespread Effect of miRNA Regulation. *RNA* **15**, 21-32.
- Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S. F., Doerks, T., Dorner, S., Bork, P., Boutros, M. and Izaurralde, E.** (2007). Target-Specific Requirements for Enhancers of Decapping in miRNA-Mediated Gene Silencing. *Genes Dev.* **21**, 2558-2570.
- Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B. and Bartel, D. P.** (2005). The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. *Science* **310**, 1817-1821.
- Filipowicz, W., Bhattacharyya, S. N. and Sonenberg, N.** (2008). Mechanisms of Post-Transcriptional Regulation by microRNAs: Are the Answers in Sight? *Nat. Rev. Genet.* **9**, 102-114.
- Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. et al.** (2008). Ensembl 2008. *Nucleic Acids Res.* **36**, D707-14.
- Forstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E. and Zamore, P. D.** (2005). Normal microRNA Maturation and Germ-Line Stem Cell Maintenance Requires Loquacious, a Double-Stranded RNA-Binding Domain Protein. *PLoS Biol.* **3**, e236.
- Friedel, R. H., Plump, A., Lu, X., Spilker, K., Jolicoeur, C., Wong, K., Venkatesh, T. R., Yaron, A., Hynes, M., Chen, B. et al.** (2005). Gene Targeting using a Promoterless Gene Trap Vector ("Targeted Trapping") is an Efficient Method to Mutate a Large Fraction of Genes. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13188-13193.
- Friedman, R. C., Farh, K. K., Burge, C. B. and Bartel, D. P.** (2009). Most Mammalian mRNAs are Conserved Targets of microRNAs. *Genome Res.* **19**, 92-105.

- Fukagawa, T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T. and Oshimura, M.** (2004). Dicer is Essential for Formation of the Heterochromatin Structure in Vertebrate Cells. *Nat. Cell Biol.* **6**, 784-791.
- Fukuda, T., Yamagata, K., Fujiyama, S., Matsumoto, T., Koshida, I., Yoshimura, K., Mihara, M., Naitou, M., Endoh, H., Nakamura, T. et al.** (2007). DEAD-Box RNA Helicase Subunits of the Drosha Complex are Required for Processing of rRNA and a Subset of microRNAs. *Nat. Cell Biol.* **9**, 604-611.
- Gangaraju, V. K. and Lin, H.** (2009). MicroRNAs: Key Regulators of Stem Cells. *Nat. Rev. Mol. Cell Biol.* **10**, 116-125.
- Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., Hammond, S. M., Bartel, D. P. and Schier, A. F.** (2005). MicroRNAs Regulate Brain Morphogenesis in Zebrafish. *Science* **308**, 833-838.
- Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J. and Schier, A. F.** (2006). Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs. *Science* **312**, 75-79.
- Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N. and Shiekhattar, R.** (2004). The Microprocessor Complex Mediates the Genesis of microRNAs. *Nature* **432**, 235-240.
- Griffiths-Jones, S., Saini, H. K., van Dongen, S. and Enright, A. J.** (2008). MiRBase: Tools for microRNA Genomics. *Nucleic Acids Res.* **36**, D154-8.
- Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and 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. and 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.
- Guan, K., Chang, H., Rolletschek, A. and Wobus, A. M.** (2001). Embryonic Stem Cell-Derived Neurogenesis. Retinoic Acid Induction and Lineage Selection of Neuronal Cells. *Cell Tissue Res.* **305**, 171-176.
- Haase, A. D., Jaskiewicz, L., Zhang, H., Laine, S., Sack, R., Gatignol, A. and Filipowicz, W.** (2005). TRBP, a Regulator of Cellular PKR and HIV-1 Virus Expression, Interacts with Dicer and Functions in RNA Silencing. *EMBO Rep.* **6**, 961-967.

- Hameyer, D., Loonstra, A., Eshkind, L., Schmitt, S., Antunes, C., Groen, A., Bindels, E., Jonkers, J., Krimpenfort, P., Meuwissen, R. et al.** (2007). Toxicity of Ligand-Dependent Cre Recombinases and Generation of a Conditional Cre Deleter Mouse Allowing Mosaic Recombination in Peripheral Tissues. *Physiol. Genomics* **31**, 32-41.
- Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H. and Kim, V. N.** (2004). The Drosha-DGCR8 Complex in Primary microRNA Processing. *Genes Dev.* **18**, 3016-3027.
- Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T. and Kim, V. N.** (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell* **125**, 887-901.
- Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., Yang, W. Y., Haussler, D., Belloch, R. and Kim, V. N.** (2009). Posttranscriptional Crossregulation between Drosha and DGCR8. *Cell* **136**, 75-84.
- He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. et al.** (2005). A microRNA Polycistron as a Potential Human Oncogene. *Nature* **435**, 828-833.
- Heo, I., Joo, C., Cho, J., Ha, M., Han, J. and Kim, V. N.** (2008). Lin28 Mediates the Terminal Uridylation of Let-7 Precursor MicroRNA. *Mol. Cell* **32**, 276-284.
- Hock, J., Weinmann, L., Ender, C., Rudel, S., Kremmer, E., Raabe, M., Urlaub, H. and Meister, G.** (2007). Proteomic and Functional Analysis of Argonaute-Containing mRNA-Protein Complexes in Human Cells. *EMBO Rep.* **8**, 1052-1060.
- Houbaviy, H. B., Dennis, L., Jaenisch, R. and Sharp, P. A.** (2005). Characterization of a Highly Variable Eutherian microRNA Gene. *RNA* **11**, 1245-1257.
- Houbaviy, H. B., Murray, M. F. and Sharp, P. A.** (2003). Embryonic Stem Cell-Specific MicroRNAs. *Dev. Cell.* **5**, 351-358.
- Huang, Q., Gumireddy, K., Schrier, M., le Sage, C., Nagel, R., Nair, S., Egan, D. A., Li, A., Huang, G., Klein-Szanto, A. J. et al.** (2008). The microRNAs miR-373 and miR-520c Promote Tumour Invasion and Metastasis. *Nat. Cell Biol.* **10**, 202-210.
- Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W. and Melton, D. A.** (2008). Induction of Pluripotent Stem Cells from Primary Human Fibroblasts with Only Oct4 and Sox2. *Nat. Biotechnol.* **26**, 1269-1275.

- Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T. and 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**, 834-838.
- Hutvagner, G. and Simard, M. J.** (2008). Argonaute Proteins: Key Players in RNA Silencing. *Nat. Rev. Mol. Cell Biol.* **9**, 22-32.
- Ianari, A., Gallo, R., Palma, M., Alesse, E. and Gulino, A.** (2004). Specific Role for p300/CREB-Binding Protein-Associated Factor Activity in E2F1 Stabilization in Response to DNA Damage. *J. Biol. Chem.* **279**, 30830-30835.
- John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S.** (2004). Human MicroRNA Targets. *PLoS Biol.* **2**, e363.
- Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D. M. and Rajewsky, K.** (2005). Dicer-Deficient Mouse Embryonic Stem Cells are Defective in Differentiation and Centromeric Silencing. *Genes Dev.* **19**, 489-501.
- Kaufman, M. H., Robertson, E. J., Handyside, A. H. and Evans, M. J.** (1983). Establishment of Pluripotential Cell Lines from Haploid Mouse Embryos. *J. Embryol. Exp. Morphol.* **73**, 249-261.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E.** (2007). The Role of Site Accessibility in microRNA Target Recognition. *Nat. Genet.* **39**, 1278-1284.
- Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J. and 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**, 2654-2659.
- Khvorova, A., Reynolds, A. and Jayasena, S. D.** (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell* **115**, 209-216.
- Kim, D. H., Saetrom, P., Snove, O., Jr and Rossi, J. J.** (2008a). MicroRNA-Directed Transcriptional Gene Silencing in Mammalian Cells. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16230-16235.
- Kim, J., Chu, J., Shen, X., Wang, J. and Orkin, S. H.** (2008b). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell* **132**, 1049-1061.
- Kim, Y. K. and Kim, V. N.** (2007). Processing of Intronic microRNAs. *EMBO J.* **26**, 775-783.
- Kiriakidou, M., Tan, G. S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P. T. and Mourelatos, Z.** (2007). An mRNA m7G Cap Binding-Like Motif within Human Ago2 Represses Translation. *Cell* **129**, 1141-1151.

- Knight, S. W. and 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**, 2269-2271.
- Kok, K. H., Ng, M. H., Ching, Y. P. and Jin, D. Y.** (2007). Human TRBP and PACT Directly Interact with each Other and Associate with Dicer to Facilitate the Production of Small Interfering RNA. *J. Biol. Chem.* **282**, 17649-17657.
- Kong, Y. W., Cannell, I. G., de Moor, C. H., Hill, K., Garside, P. G., Hamilton, T. L., Meijer, H. A., Dobbyn, H. C., Stoneley, M., Spriggs, K. A. et al.** (2008). The Mechanism of Micro-RNA-Mediated Translation Repression is Determined by the Promoter of the Target Gene. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8866-8871.
- Kopp, J. L., Ormsbee, B. D., Desler, M. and Rizzino, A.** (2008). Small Increases in the Level of Sox2 Trigger the Differentiation of Mouse Embryonic Stem Cells. *Stem Cells* **26**, 903-911.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M. et al.** (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129**, 1401-1414.
- Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P. Y., Soll, S. J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M. and Tuschl, T.** (2008). Molecular Characterization of Human Argonaute-Containing Ribonucleoprotein Complexes and their Bound Target mRNAs. *RNA* **14**, 2580-2596.
- Laurent, L. C., Chen, J., Ulitsky, I., Mueller, F. J., Lu, C., Shamir, R., Fan, J. B. and Loring, J. F.** (2008). Comprehensive microRNA Profiling Reveals a Unique Human Embryonic Stem Cell Signature Dominated by a Single Seed Sequence. *Stem Cells* **26**, 1506-1516.
- Lee, R. C., Feinbaum, R. L. and Ambros, V.** (1993). The *C. Elegans* Heterochronic Gene *Lin-4* Encodes Small RNAs with Antisense Complementarity to *Lin-14*. *Cell* **75**, 843-854.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. et al.** (2003). The Nuclear RNase III Droscha Initiates microRNA Processing. *Nature* **425**, 415-419.
- Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H. and Kim, V. N.** (2004a). MicroRNA Genes are Transcribed by RNA Polymerase II. *EMBO J.* **23**, 4051-4060.
- Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J. and Carthew, R. W.** (2004b). Distinct Roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell* **117**, 69-81.

- Lewis, B. P., Burge, C. B. and Bartel, D. P.** (2005). Conserved Seed Pairing, often Flanked by Adenosines, Indicates that Thousands of Human Genes are microRNA Targets. *Cell* **120**, 15-20.
- Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B.** (2003). Prediction of Mammalian microRNA Targets. *Cell* **115**, 787-798.
- Liang, Q., Conte, N., Skarnes, W. C. and Bradley, A.** (2008). Extensive Genomic Copy Number Variation in Embryonic Stem Cells. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 17453-17456.
- Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. and Johnson, J. M.** (2005). Microarray Analysis shows that some microRNAs Downregulate Large Numbers of Target mRNAs. *Nature* **433**, 769-773.
- Lin, S. L., Chang, D. C., Chang-Lin, S., Lin, C. H., Wu, D. T., Chen, D. T. and Ying, S. Y.** (2008a). Mir-302 Reprograms Human Skin Cancer Cells into a Pluripotent ES-Cell-Like State. *RNA* **14**, 2115-2124.
- Lin, S. M., Du, P., Huber, W. and Kibbe, W. A.** (2008b). Model-Based Variance-Stabilizing Transformation for Illumina Microarray Data. *Nucleic Acids Res.* **36**, e11.
- Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J. J., Hammond, S. M., Joshua-Tor, L. and Hannon, G. J.** (2004). Argonaute2 is the Catalytic Engine of Mammalian RNAi. *Science* **305**, 1437-1441.
- Liu, J., Valencia-Sanchez, M. A., Hannon, G. J. and Parker, R.** (2005). MicroRNA-Dependent Localization of Targeted mRNAs to Mammalian P-Bodies. *Nat. Cell Biol.* **7**, 719-723.
- Liu, Q., Fu, H., Sun, F., Zhang, H., Tie, Y., Zhu, J., Xing, R., Sun, Z. and Zheng, X.** (2008a). MiR-16 Family Induces Cell Cycle Arrest by Regulating Multiple Cell Cycle Genes. *Nucleic Acids Res.* **36**, 5391-5404.
- Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G. and Kang, J.** (2008b). Yamanaka Factors Critically Regulate the Developmental Signaling Network in Mouse Embryonic Stem Cells. *Cell Res.* **18**, 1177-1189.
- Liu, X., Jiang, F., Kalidas, S., Smith, D. and Liu, Q.** (2006). Dicer-2 and R2D2 Coordinately Bind siRNA to Promote Assembly of the siRISC Complexes. *RNA* **12**, 1514-1520.
- Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A. et al.** (2005). MicroRNA Expression Profiles Classify Human Cancers. *Nature* **435**, 834-838.

- Lu, Z., Liu, M., Stribinskis, V., Klinge, C. M., Ramos, K. S., Colburn, N. H. and Li, Y.** (2008). MicroRNA-21 Promotes Cell Transformation by Targeting the Programmed Cell Death 4 Gene. *Oncogene* **27**, 4373-4379.
- Mallory, A. C. and Bouche, N.** (2008). MicroRNA-Directed Regulation: To Cleave Or Not to Cleave. *Trends Plant Sci.* **13**, 359-367.
- Mansfield, J. H., Harfe, B. D., Nissen, R., Obenauer, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A. E., Ruvkun, G., Sharp, P. A., Tabin, C. J. and McManus, T.** (2004). MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat. Genet.* **36**, 1079-1083.
- Marmorstein, R.** (2001). Structure of Histone Acetyltransferases. *J. Mol. Biol.* **311**, 433-444.
- Maroney, P. A., Yu, Y., Fisher, J. and Nilsen, T. W.** (2006). Evidence that microRNAs are Associated with Translating Messenger RNAs in Human Cells. *Nat. Struct. Mol. Biol.* **13**, 1102-1107.
- Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J. et al.** (2008). Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells. *Cell* **134**, 521-533.
- Martin, G. R.** (1981). Isolation of a Pluripotent Cell Line from Early Mouse Embryos Cultured in Medium Conditioned by Teratocarcinoma Stem Cells. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7634-7638.
- Mathonnet, G., Fabian, M. R., Svitkin, Y. V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W. C., Darzynkiewicz, E., Pillai, R. S. et al.** (2007). MicroRNA Inhibition of Translation Initiation in Vitro by Targeting the Cap-Binding Complex eIF4F. *Science* **317**, 1764-1767.
- Mattick, J. S. and Makunin, I. V.** (2006). Non-Coding RNA. *Hum. Mol. Genet.* **15** Spec No 1, R17-29.
- Medina, P. P. and Slack, F. J.** (2008). MicroRNAs and Cancer: An Overview. *Cell Cycle* **7**, 2485-2492.
- Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G. and Tuschl, T.** (2004). Human Argonaute2 Mediates RNA Cleavage Targeted by miRNAs and siRNAs. *Mol. Cell* **15**, 185-197.
- Miska, E. A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Sestan, N., Rakic, P., Constantine-Paton, M. and Horvitz, H. R.** (2004). Microarray Analysis of microRNA Expression in the Developing Mammalian Brain. *Genome Biol.* **5**, R68.

- Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M. and Yamanaka, S.** (2003). The Homeoprotein Nanog is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. *Cell* **113**, 631-642.
- Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. et al.** (2008). Application of Massively Parallel Sequencing to microRNA Profiling and Discovery in Human Embryonic Stem Cells. *Genome Res.* **18**, 610-621.
- Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I. and Proudfoot, N. J.** (2008). Primary microRNA Transcripts are Processed Co-Transcriptionally. *Nat. Struct. Mol. Biol.*
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B.** (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621-628.
- Murchison, E. P., Partridge, J. F., Tam, O. H., Cheloufi, S. and Hannon, G. J.** (2005). Characterization of Dicer-Deficient Murine Embryonic Stem Cells. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12135-12140.
- Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N. and Yamanaka, S.** (2008). Generation of Induced Pluripotent Stem Cells without Myc from Mouse and Human Fibroblasts. *Nat. Biotechnol.* **26**, 101-106.
- Newman, M. A., Thomson, J. M. and Hammond, S. M.** (2008). Lin-28 Interaction with the Let-7 Precursor Loop Mediates Regulated microRNA Processing. *RNA* **14**, 1539-1549.
- Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H. and Smith, A.** (1998). Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. *Cell* **95**, 379-391.
- Ning, Z., Cox, A. J. and Mullikin, J. C.** (2001). SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res.* **11**, 1725-1729.
- Niida, H. and Nakanishi, M.** (2006). DNA Damage Checkpoints in Mammals. *Mutagenesis* **21**, 3-9.
- Niwa, H., Miyazaki, J. and Smith, A. G.** (2000). Quantitative Expression of Oct-3/4 Defines Differentiation, Dedifferentiation Or Self-Renewal of ES Cells. *Nat. Genet.* **24**, 372-376.
- Nord, A. S., Chang, P. J., Conklin, B. R., Cox, A. V., Harper, C. A., Hicks, G. G., Huang, C. C., Johns, S. J., Kawamoto, M., Liu, S. et al.** (2006). The International Gene Trap Consortium Website: A Portal to all Publicly Available Gene Trap Cell Lines in Mouse. *Nucleic Acids Res.* **34**, D642-8.

- Ohnishi, Y. N., Sakumi, K., Yamazaki, K., Ohnishi, Y. H., Miura, T., Tominaga, Y., Nakabeppu, Y.** (2008). Antagonistic Regulation of Cell-Matrix Adhesion by FosB and Δ FosB/ Δ 2 Δ FosB Encoded by Alternatively Spliced Forms of *fosB* Transcripts. *Mol. Biol. Cell* **19**, 4717-4729.
- Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. and Lai, E. C.** (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in *Drosophila*. *Cell* **130**, 89-100.
- Okamura, K., Phillips, M. D., Tyler, D. M., Duan, H., Chou, Y. T. and Lai, E. C.** (2008). The Regulatory Activity of microRNA* Species has Substantial Influence on microRNA and 3' UTR Evolution. *Nat. Struct. Mol. Biol.* **15**, 354-363.
- Okita, K., Ichisaka, T. and Yamanaka, S.** (2007). Generation of Germline-Competent Induced Pluripotent Stem Cells. *Nature* **448**, 313-317.
- Okita, K. and Yamanaka, S.** (2006). Intracellular Signaling Pathways Regulating Pluripotency of Embryonic Stem Cells. *Curr. Stem Cell. Res. Ther.* **1**, 103-111.
- Olsen, P. H. and Ambros, V.** (1999). The Lin-4 Regulatory RNA Controls Developmental Timing in *Caenorhabditis Elegans* by Blocking LIN-14 Protein Synthesis After the Initiation of Translation. *Dev. Biol.* **216**, 671-680.
- Orford, K. W. and Scadden, D. T.** (2008). Deconstructing Stem Cell Self-Renewal: Genetic Insights into Cell-Cycle Regulation. *Nat. Rev. Genet.* **9**, 115-128.
- Petersen, C. P., Bordeleau, M. E., Pelletier, J. and Sharp, P. A.** (2006). Short RNAs Repress Translation After Initiation in Mammalian Cells. *Mol. Cell* **21**, 533-542.
- Petrocca, F., Vecchione, A. and Croce, C. M.** (2008a). Emerging Role of miR-106b-25/miR-17-92 Clusters in the Control of Transforming Growth Factor Beta Signaling. *Cancer Res.* **68**, 8191-8194.
- Petrocca, F., Visone, R., Onelli, M. R., Shah, M. H., Nicoloso, M. S., de Martino, I., Iliopoulos, D., Pillozzi, E., Liu, C. G., Negrini, M. et al.** (2008b). E2F1-Regulated microRNAs Impair TGFbeta-Dependent Cell-Cycle Arrest and Apoptosis in Gastric Cancer. *Cancer. Cell.* **13**, 272-286.
- Pillai, R. S., Artus, C. G. and Filipowicz, W.** (2004). Tethering of Human Ago Proteins to mRNA Mimics the miRNA-Mediated Repression of Protein Synthesis. *RNA* **10**, 1518-1525.
- Pillai, R. S., Bhattacharyya, S. N., Artus, C. G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E. and Filipowicz, W.** (2005). Inhibition of Translational Initiation by Let-7 MicroRNA in Human Cells. *Science* **309**, 1573-1576.

- Place, R. F., Li, L. C., Pookot, D., Noonan, E. J. and Dahiya, R.** (2008). MicroRNA-373 Induces Expression of Genes with Complementary Promoter Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1608-1613.
- Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. and Bradley, A.** (2004). Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Res.* **14**, 1902-1910.
- Ruby, J. G., Jan, C. H. and Bartel, D. P.** (2007). Intronic microRNA Precursors that Bypass Drosha Processing. *Nature* **448**, 83-86.
- Saini, H. K., Enright, A. J. and Griffiths-Jones, S.** (2008). Annotation of Mammalian Primary microRNAs. *BMC Genomics* **9**, 564.
- Selbach, M., Schwanhauser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N.** (2008). Widespread Changes in Protein Synthesis Induced by microRNAs. *Nature* **455**, 58-63.
- Sethupathy, P., Megraw, M. and Hatzigeorgiou, A. G.** (2006). A Guide through Present Computational Approaches for the Identification of Mammalian microRNA Targets. *Nat. Methods* **3**, 881-886.
- Shiohama, A., Sasaki, T., Noda, S., Minoshima, S. and Shimizu, N.** (2003). Molecular Cloning and Expression Analysis of a Novel Gene DGCR8 Located in the DiGeorge Syndrome Chromosomal Region. *Biochem. Biophys. Res. Commun.* **304**, 184-190.
- Sibilia, M. and Wagner, E. F.** (1995). Strain-Dependent Epithelial Defects in Mice Lacking the EGF Receptor. *Science* **269**, 234-238.
- Silva, J., Chambers, I., Pollard, S. and Smith, A.** (2006). Nanog Promotes Transfer of Pluripotency After Cell Fusion. *Nature* **441**, 997-1001.
- Silva, J. and Smith, A.** (2008). Capturing Pluripotency. *Cell* **132**, 532-536.
- Singh, S. K., Kagalwala, M. N., Parker-Thornburg, J., Adams, H. and Majumder, S.** (2008). REST Maintains Self-Renewal and Pluripotency of Embryonic Stem Cells. *Nature* **453**, 223-227.
- Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., Zavolan, M., Svoboda, P. and Filipowicz, W.** (2008). MicroRNAs Control De Novo DNA Methylation through Regulation of Transcriptional Repressors in Mouse Embryonic Stem Cells. *Nat. Struct. Mol. Biol.* **15**, 259-267.
- Smalheiser, N. R.** (2003). EST Analyses Predict the Existence of a Population of Chimeric microRNA Precursor-mRNA Transcripts Expressed in Normal Human and Mouse Tissues. *Genome Biol.* **4**, 403.

- Stanford, W. L., Cohn, J. B. and Cordes, S. P.** (2001). Gene-Trap Mutagenesis: Past, Present and Beyond. *Nat. Rev. Genet.* **2**, 756-768.
- Stryke, D., Kawamoto, M., Huang, C. C., Johns, S. J., King, L. A., Harper, C. A., Meng, E. C., Lee, R. E., Yee, A., L'Italien, L. et al.** (2003). BayGenomics: A Resource of Insertional Mutations in Mouse Embryonic Stem Cells. *Nucleic Acids Res.* **31**, 278-281.
- Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., Cha, K. Y., Chung, H. M., Yoon, H. S., Moon, S. Y. et al.** (2004). Human Embryonic Stem Cells Express a Unique Set of microRNAs. *Dev. Biol.* **270**, 488-498.
- Takada, H., Imoto, I., Tsuda, H., Nakanishi, Y., Ichikura, T., Mochizuki, H., Mitsufuji, S., Hosoda, F., Hirohashi, S., Ohki, M. et al.** (2005a). ADAM23, a Possible Tumor Suppressor Gene, is Frequently Silenced in Gastric Cancers by Homozygous Deletion Or Aberrant Promoter Hypermethylation. *Oncogene* **24**, 8051-8060.
- Takada, T., Nemoto, K., Yamashita, A., Kato, M., Kondo, Y. and Torii, R.** (2005b). Efficient Gene Silencing and Cell Differentiation using siRNA in Mouse and Monkey ES Cells. *Biochem. Biophys. Res. Commun.* **331**, 1039-1044.
- Takahashi, K. and Yamanaka, S.** (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663-676.
- Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M. et al.** (2008). Pseudogene-Derived Small Interfering RNAs Regulate Gene Expression in Mouse Oocytes. *Nature* **453**, 534-538.
- Tang, F., Hajkova, P., Barton, S. C., Lao, K. and Surani, M. A.** (2006). MicroRNA Expression Profiling of Single Whole Embryonic Stem Cells. *Nucleic Acids Res.* **34**, e9.
- Tay, Y., Zhang, J., Thomson, A. M., Lim, B. and Rigoutsos, I.** (2008). MicroRNAs to Nanog, Oct4 and Sox2 Coding Regions Modulate Embryonic Stem Cell Differentiation. *Nature* **455**, 1124-1128.
- te Riele, H., Maandag, E. R. and Berns, A.** (1992). Highly Efficient Gene Targeting in Embryonic Stem Cells through Homologous Recombination with Isogenic DNA Constructs. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5128-5132.
- Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., Gardner, R. L. and McKay, R. D.** (2007). New Cell Lines from Mouse Epiblast Share Defining Features with Human Embryonic Stem Cells. *Nature* **448**, 196-199.

- Tonon, G., Wong, K. K., Maulik, G., Brennan, C., Feng, B., Zhang, Y., Khatry, D. B., Protopopov, A., You, M. J., Aguirre, A. J. et al.**(2005). High-Resolution Genomic Profiles of Human Lung Cancer. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9625-9630.
- Trimarchi, J. M. and Lees, J. A.** (2002). Sibling Rivalry in the E2F Family. *Nat. Rev. Mol. Cell Biol.* **3**, 11-20.
- van Dongen, S., Abreu-Goodger, C. and Enright, A. J.**(2008). Detecting microRNA Binding and siRNA Off-Target Effects from Expression Data. *Nat. Methods* **5**, 1023-1025.
- Vasudevan, S. and Steitz, J. A.** (2007). AU-Rich-Element-Mediated Upregulation of Translation by FXR1 and Argonaute 2. *Cell* **128**, 1105-1118.
- Vella, M. C., Choi, E. Y., Lin, S. Y., Reinert, K. and Slack, F. J.**(2004). The C. Elegans microRNA Let-7 Binds to Imperfect Let-7 Complementary Sites from the Lin-41 3'UTR. *Genes Dev.* **18**, 132-137.
- Ventura, A., Young, A. G., Winslow, M. M., Lintault, L., Meissner, A., Erkeland, S. J., Newman, J., Bronson, R. T., Crowley, D., Stone, J. R. et al.**(2008). Targeted Deletion Reveals Essential and Overlapping Functions of the miR-17 through 92 Family of miRNA Clusters. *Cell* **132**, 875-886.
- Vooijs, M., Jonkers, J. and Berns, A.** (2001). A Highly Efficient Ligand-Regulated Cre Recombinase Mouse Line shows that LoxP Recombination is Position Dependent. *EMBO Rep.* **2**, 292-297.
- Voorhoeve, P. M., le Sage, C., Schrier, M., Gillis, A. J., Stoop, H., Nagel, R., Liu, Y. P., van Duijse, J., Drost, J., Griekspoor, A. et al.**(2006). A Genetic Screen Implicates miRNA-372 and miRNA-373 as Oncogenes in Testicular Germ Cell Tumors. *Cell* **124**, 1169-1181.
- Wakiyama, M., Takimoto, K., Ohara, O. and Yokoyama, S.** (2007). Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes Dev.* **21**, 1975-1982.
- Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W. and Orkin, S. H.** (2006). A Protein Interaction Network for Pluripotency of Embryonic Stem Cells. *Nature* **444**, 364-368.
- Wang, Y., Baskerville, S., Shenoy, A., Babiarz, J. E., Baehner, L. and Blelloch, R.** (2008). Embryonic Stem Cell-Specific microRNAs Regulate the G1-S Transition and Promote Rapid Proliferation. *Nat. Genet.* **40**, 1478-1483.
- Wang, Y., Medvid, R., Melton, C., Jaenisch, R. and Blelloch, R.**(2007). DGCR8 is Essential for microRNA Biogenesis and Silencing of Embryonic Stem Cell Self-Renewal. *Nat. Genet.* **39**, 380-385.

- Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T. et al.** (2008). Endogenous siRNAs from Naturally Formed dsRNAs Regulate Transcripts in Mouse Oocytes. *Nature* **453**, 539-543.
- Welcker, M. and Clurman, B. E.** (2008). FBW7 Ubiquitin Ligase: A Tumour Suppressor at the Crossroads of Cell Division, Growth and Differentiation. *Nat. Rev. Cancer*. **8**, 83-93.
- Wightman, B., Ha, I. and Ruvkun, G.** (1993). Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern Formation in *C. Elegans*. *Cell* **75**, 855-862.
- Wu, H., Xu, H., Miraglia, L. J. and Crooke, S. T.** (2000). Human RNase III is a 160-kDa Protein Involved in Preribosomal RNA Processing. *J. Biol. Chem.* **275**, 36957-36965.
- Wu, L., Fan, J. and Belasco, J. G.** (2006). MicroRNAs Direct Rapid Deadenylation of mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 4034-4039.
- Xia, Y., Yu, P. B., Sidis, Y., Beppu, H., Bloch, K. D., Schneyer, A. L. and Lin, H. Y.** (2007). Repulsive Guidance Molecule RGMA Alters Utilization of Bone Morphogenetic Protein (BMP) Type II Receptors by BMP2 and BMP4. *J. Biol. Chem.* **282**, 18129-18140.
- Xu, H. M., Liao, B., Zhang, Q. J., Wang, B. B., Li, H., Zhong, X. M., Sheng, H. Z., Zhao, Y. X., Zhao, Y. M. and Jin, Y.** (2004). Wwp2, an E3 Ubiquitin Ligase that Targets Transcription Factor Oct-4 for Ubiquitination. *J. Biol. Chem.* **279**, 23495-23503.
- Yekta, S., Shih, I. H. and Bartel, D. P.** (2004). MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science* **304**, 594-596.
- Yeom, K. H., Lee, Y., Han, J., Suh, M. R. and Kim, V. N.** (2006). Characterization of DGCR8/Pasha, the Essential Cofactor for Drosha in Primary miRNA Processing. *Nucleic Acids Res.* **34**, 4622-4629.
- Yi, R., Qin, Y., Macara, I. G. and Cullen, B. R.** (2003). Exportin-5 Mediates the Nuclear Export of Pre-microRNAs and Short Hairpin RNAs. *Genes Dev.* **17**, 3011-3016.
- Ying, Q. L., Nichols, J., Chambers, I. and Smith, A.** (2003). BMP Induction of Id Proteins Suppresses Differentiation and Sustains Embryonic Stem Cell Self Renewal in Collaboration with STAT3. *Cell* **115**, 281-292.
- Ying, Q. L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P. and Smith, A.** (2008). The Ground State of Embryonic Stem Cell Self-Renewal. *Nature* **453**, 519-523.

- Zeng, Y. and Cullen, B. R.** (2004). Structural Requirements for Pre-microRNA Binding and Nuclear Export by Exportin 5. *Nucleic Acids Res.* **32**, 4776-4785.
- Zeng, Y. and Cullen, B. R.** (2005). Efficient Processing of Primary microRNA Hairpins by Drosha Requires Flanking Nonstructured RNA Sequences. *J. Biol. Chem.* **280**, 27595-27603.
- Zeng, Y., Yi, R. and Cullen, B. R.** (2005). Recognition and Cleavage of Primary microRNA Precursors by the Nuclear Processing Enzyme Drosha. *EMBO J.* **24**, 138-148.
- Zhao, Y. and Srivastava, D.** (2007). A Developmental View of microRNA Function. *Trends Biochem. Sci.* **32**, 189-197.
- Zhou, X., Ruan, J., Wang, G. and Zhang, W.** (2007). Characterization and Identification of microRNA Core Promoters in Four Model Species. *PLoS Comput. Biol.* **3**, e37.