

**Conservation and innovation – the evolution of
the metazoan microRNA landscape and its
contribution to reproduction and development**

Shan Quah

(OSS Student ID: 565729)

Merton College

Department of Zoology

Thesis submitted in partial fulfilment of the requirements for the
degree of Doctor of Philosophy.

University of Oxford

2015

Declaration and Statement of Authorship

I declare that the work presented in this thesis was conducted by myself under the supervision of Prof. Peter W. H. Holland (University of Oxford), except in those instances where the contribution of others has been acknowledged. Any material quoted, referenced or paraphrased from other sources has been clearly indicated and its origins cited.

All data chapters have been written up as original scientific research manuscripts lead authored by myself. Chapter 2 has been published in *EvoDevo*. Most of the data in Chapters 3 and 4 have been published together in one manuscript in *Molecular Biology and Evolution* and Chapter 5 has been published in *PLOS ONE*. In each case, I have drafted the original manuscripts and the majority of the work done is attributable to me unless otherwise stated. All co-authors on each of the submitted manuscripts have evaluated the work and contributed to editing the manuscripts prior to submission. A detailed description of author contributions is provided at the start of each data chapter.

I carried out data collection and data analysis by myself for all chapters. *Pararge aegeria* specimens used in Chapters 2, 3 and 4 were provided by Dr. Casper J. Breuker (Oxford Brookes University); all other material obtained from external research groups is clearly stated at the start of each chapter. Laboratory work was mainly carried out in the laboratory of Prof. Peter Holland. Equipment and reagents were also contributed by Dr. Casper Breuker for Chapters 3, 4 and 5, and by Prof. Adrian L. Smith (University of Oxford) for the cell culture work in Chapter 4. Prof. Peter Holland supervised the development of my work for all data chapters, and together with Prof. Jerome Hui (The Chinese University of Hong Kong) assisted in experimental design for Chapters 2 and 3. Funding for data collection and data analyses was secured by Prof. Peter Holland through the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant [268513], with the exception of Chapter 3 where funding for deep sequencing was jointly obtained by myself, Prof. Jerome Hui, and Prof. Peter Holland. Funding for this project was provided by the Elizabeth-Hannah Jenkinson Fund. I also acknowledge support from the A*STAR National Science Scholarship (PhD).

I hereby confirm that this thesis has not been submitted for any other qualification or degree at this university or any other institution.

Acknowledgements

The work presented here would not have been achievable without the support of a number of notable individuals. First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Peter Holland, for his dedicated mentorship and commitment to getting the best out of his graduate students. Peter took an active interest in my skills and career development as a researcher right from the start. In the frustrating initial stages of my degree, he provided the guidance and support I needed to get off the ground and attain the confidence I needed to take ownership over my own research direction. As I progressed, he was always encouraging me to disseminate my work at conferences and to frame all my experimental plans with publication in mind. This advice paid dividends as I was able to have the majority of my DPhil project accepted or submitted for publication prior to submission of my thesis. Even now, as I prepare to take the next step forward, Peter continues to take an active interest in my career progression and is ever ready to provide a listening ear and helpful advice. Peter's passion for science, his discerning eye for detail, and his commitment to nurturing the next generation of scientific talent greatly inspired me as a fresh graduate student and continue to do so today.

I would like to thank our collaborator, Dr Casper Breuker (Oxford Brookes University), for his indispensable assistance and supervision of the aspects of my project involving the Speckled Wood butterfly. Casper worked tirelessly in maintaining the butterfly culture and was always eager to discuss new insights and ideas he came across in the field.

Materials and technical expertise were kindly contributed by a number of researchers around the globe. Prof Adrian Smith, Dr Stephen Preston and Amy Hong (Department of Zoology, University of Oxford) provided access to the required facilities for cell culture work as well as advice on cell culture and luciferase protocols. Prof Stephen Donellan (South Australian Museum, also University of Adelaide) and Dr Mads Bertelsen (Copenhagen Zoo) provided the monotreme and Xenarthran DNA and tissue samples for my survey of miR-615 distribution in mammals. Dr Shankar Srinivas and Antonio Miranda (both at Department of Physiology, Anatomy and Genetics, University of Oxford) supplied materials for the project on mouse miR-615 in Chapter 2. Dr James Hogan (Oxford University Museum of Natural History) granted me permission to photograph lepidopteran specimens for Figure 3.1. Dr Jean-Michel Carter and Luca Livraghi (formerly Oxford Brookes University) assisted with the Speckled Wood culture and experimental protocols used in Chapters 3 and 4, which originally incorporated in

situ hybridisation and imaging of Speckled Wood ovarioles. Prof Ilan Davis (Department of Biochemistry, University of Oxford), Prof Charles Godfray, and Dr Stuart Wigby (Department of Zoology, University of Oxford) all contributed to discussion on the work in Chapter 4. I am particularly grateful to Dr Richard Parton (Department of Biochemistry, University of Oxford) for his assistance with microscopy on this project.

Many past and present members of the Holland laboratory have contributed their time and efforts into helping me learn the ropes, as well as provided stimulating scientific discourse. Of mention are Dr Caroline Bacquet, Dr Thomas Dunwell, Dr Nathalie Feiner, Dr Laura Ferguson, Dr Adam Hargreaves, Dr Ignacio Maeso, Dr Ferdinand Marlétaz, Kyle Martin and Dr Jordi Paps-Montserrat. Particular mention goes to Professor Jerome Hui who supervised me in the initial stages of my DPhil.

I must also thank the fantastic community of people I have come to know at Oxford who helped me ride out the rougher patches in this epic journey and discover myself along the way. Special thanks go to the folks at Oxford Academicals Rowing Club, for the thrill of the race and the humbling honour to have been your captain.

Finally, I wish to express my deepest gratitude to my parents, Choon Poh and Audrey, and my brother James for the long years they have waited patiently for me to come home.

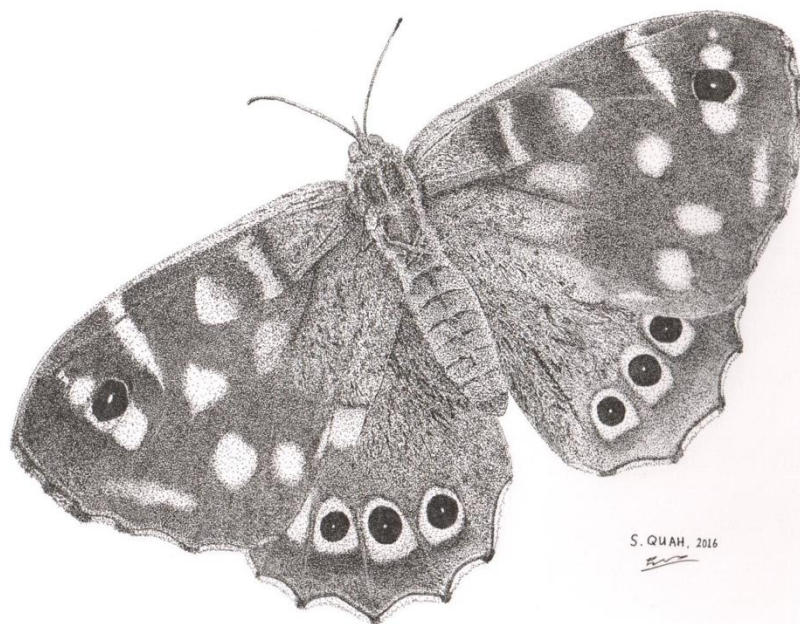


Table of Contents

List of Figures	v
Thesis Abstract	1
Chapter 1: General Introduction	3
1 Overview of metazoan miRNA biology	5
1.1 What are microRNAs?	5
1.2 The canonical metazoan miRNA biogenesis pathway	5
1.3 Mechanisms of miRNA targeting and miRNA action	8
1.4 Requirement for 3' UTR localization of miRNA target sites	9
1.5 miRNAs within a host transcription unit	9
1.6 Distinction between miRNAs and siRNAs	11
2 miRNA annotation and novel miRNA discovery	14
2.1 The need for characterization of miRNA transcriptomes	14
2.2 Early advances in miRNA discovery	14
2.3 Experimental and computational approaches to miRNA annotation	15
2.4 Identification of miRNAs in deep sequencing datasets	16
3 Evolution of miRNAs and the miRNA processing pathway	20
3.1 Acquisition of the miRNA processing pathway	20
3.2 Mechanisms for novel miRNA emergence in the Metazoa	25
3.3 Models for miRNA evolution within a clade	28
4 The role of miRNAs in the evolution of development	30
4.1 miRNAs and morphological complexity	30
4.2 miRNA interactions in genetic regulatory circuits	31
4.3 miRNAs and the canalization of development	32
4.4 The role of miRNAs in early embryogenesis	33
5 Project aims and contributions	35
Chapter 2: miR-615 – A model for intronic microRNA evolution	37
Statement of Author Contributions	39
1 Abstract	40
2 Introduction	41
3 Materials and Methods	43

3.1	Conservation analysis for <i>mir-615</i>	43
3.2	Bioinformatics and target prediction	44
3.3	<i>mir-615</i> promoter prediction	45
3.4	Mouse ovary RT-PCR	45
4	Results	
4.1	A sequence with high similarity to <i>mir-615</i> is present in therian mammals	47
4.2	Candidate <i>miR-615</i> precursors fulfil structural and energetic criteria only in Eutheria	48
4.3	<i>miR-615</i> expression in mammalian cell lines and adult tissues	53
4.4	Expression of <i>miR-615</i> can occur independently from <i>Hoxc5</i>	55
4.5	Ovarian expression of <i>mir-615</i> and <i>HoxC</i> genes	57
4.6	Candidate targets for <i>miR-615-3p</i>	60
5	Discussion	62
5.1	Origins of eutherian <i>mir-615</i>	62
5.2	<i>mir-615</i> as a model for understanding intronic miRNA evolution	63
5.3	Putative functions for <i>miR-615</i>	66
6	Supplementary Information	69
	Chapter 3: MicroRNA innovation within the Lepidoptera	77
	Statement of Author Contributions	79
1	Abstract	80
2	Introduction	81
3	Materials and Methods	84
3.1	Sample collection	84
3.2	RNA extraction and miRNA sequencing	84
3.3	Computational analysis	85
3.4	Phylogenetic placement of lepidopteran miRNAs	85
4	Results	88
4.1	Small RNA sampling in <i>Cameraria ohridella</i> and <i>Pararge aegeria</i>	88
4.2	Deep sequencing and miRNA identification	89
4.3	Conserved miRNAs in <i>Cameraria</i> and <i>Pararge</i>	89
4.4	Novel miRNA discovery in <i>Cameraria</i> and <i>Pararge</i>	89
4.5	Phylogenetic placement of lepidopteran miRNAs	92
4.6	Expression level analysis	95
5	Discussion	97
5.1	MicroRNA discovery in <i>C. ohridella</i> and <i>P. aegeria</i>	97
5.2	Birth and death of miRNAs in the Lepidoptera	98
5.3	miRNA capture – A new model for the evolution of miRNAs within a clade	99

6	Supplementary Information	101
Chapter 4: Functional characterization of the lepidopteran novelty, miR-2768		115
	Statement of Author Contributions	117
1	Abstract	118
2	Introduction	119
3	Materials and Methods	121
3.1	miRNA target predictions	121
3.2	Sequence alignments	121
3.3	Vector construction	121
3.4	Luciferase assay protocol	122
4	Results	124
4.1	<i>Cubitus interruptus</i> as a potential target for miR-2768	124
4.2	miR-2768 is capable of exerting regulation over the ci candidate site	126
4.3	miR-2768 is capable of downregulating the ci candidate site within a coding sequence	128
5	Discussion	129
6	Supplementary Information	133
Chapter 5: Ovarian microRNAs of the Speckled Wood butterfly		147
	Statement of Author Contributions	149
1	Abstract	150
2	Introduction	151
3	Materials and Methods	153
3.1	Sample collection and sequencing	153
3.2	miRNA identification	153
3.3	Transcriptome assembly	154
3.4	miRNA target prediction	155
4	Results	156
4.1	Small RNA sequencing and data processing	156
4.2	Effect of miRNA age on expression level	159
4.3	Older miRNAs are enriched in the ovary	161
4.4	Ovarian-specific and ovarian-upregulated miRNAs	162
4.5	Identification of putative ovarian miRNA targets	164

5	Discussion	169
5.1	Diversity of miRNAs in <i>Pararge aegeria</i>	169
5.2	miRNA age, ovary expression and expression level	169
5.3	'Ovary set' miRNAs may contribute to robustness in ovarian transcriptional networks	171
5.4	Potential conserved ovarian functions for insect miRNAs	172
5.5	Candidate targets and their involvement in butterfly ovarian biology	174
6	Supplementary Information	177
Chapter 6: General Discussion		207
1	Synopsis of research	209
2	A new model for metazoan miRNA evolution	209
2.1	The miRNA capture model	209
2.2	Applying the capture model beyond miRNA evolution	210
3	Intronic exaptation and miRNA evolution	212
4	miRNAs, biological robustness, and the canalization of development	214
5	Evolutionary conservation of miRNAs and its effect on expression levels and tissue specificity	215
Concluding Remarks		217
Appendix A: Scripts		219
A1	fastqtrim.py	221
A2	size-select.py	224
A3	contig-stats.pl	225
Appendix B: List of abbreviations used		227
Appendix C: Resulting publications		229
C1	A burst of microRNA innovation in the early evolution of butterflies and moths	231
C2	The Hox cluster microRNA miR-615: A case study of intronic microRNA evolution	245
C3	A diversity of conserved and novel ovarian miRNAs in the Speckled Wood (<i>Pararge aegeria</i>)	257
Bibliography		273

List of Figures

1	General Introduction	
1.1	Metazoan miRNA biogenesis	7
1.2	Structure of a typical metazoan pri-miRNA	7
1.3	Hypothetical outcomes from a protein-coding locus containing an intronic miRNA	10
1.4	Mirtron structure	13
1.5	Known distribution of miRNAs and miRNA pathway components in basally-divergent metazoan lineages	21
1.6	Alternative scenarios for acquisition of the miRNA biogenesis pathway in basally branching Metazoa	24
1.7	<i>De novo</i> miRNA evolution from existing transcribed RNAs with approximate precursor structure.	27
1.8	Regulatory motifs containing miRNAs	32
2	miR-615 – A model for intronic miRNA evolution	
2.1	Section of <i>Hoxc5</i> intronic alignments corresponding to mir-615 across tetrapods	49
2.2	Distribution of pre-miR-615 in mammals	50
2.3	Predicted structures for mammalian miR-615 precursors	51
2.4	miR-615-3p read counts in mouse small RNA libraries	54
2.5	Expression of <i>mir-615</i> and <i>HOXC5</i> in example human cell lines	55
2.6	Inferred position of the <i>mir-615</i> promoter	56
2.7	RT-PCR on genes flanking <i>Hoxc5</i> in mouse ovaries	59
2.8	GO terms identified in the miR-615-3p target prediction dataset	61
3	MicroRNA innovation within the Lepidoptera	
3.1	Lepidopteran phylogeny and distribution of miRNA studies	83
3.2	Data processing pipeline for lepidopteran miRNA annotation	87
3.3	miRNA distribution in the Lepidoptera	94
3.4	miRNA expression levels by age of origin	96
3.5	Model for the miRNA capture hypothesis	100
4	Functional characterization of the lepidopteran novelty, miR-2768	
4.1	Conservation of the peptide and nucleotide sequences of <i>cubitus interruptus</i> across the Lepidoptera	125
4.2	Plasmid constructs used in the functional testing of miR-2768	127
4.3	Luciferase assay data for functional validation of miR-2768	128
4.4	Genetic regulatory interactions of <i>ci</i> , <i>inv</i> and miR-2768	132
5	Ovarian microRNAs of the Speckled Wood butterfly	
5.1	Annotation of <i>P. aegeria</i> ovarian miRNAs	157

5.2	Box plot showing expression levels for conserved and lineage-specific miRNAs in <i>P. aegeria</i>	160
5.3	Distribution of conserved and lineage-specific miRNAs in ovarian and whole-body miRNA transcriptomes of <i>P. aegeria</i>	161
5.4	GO term distribution of predicted target contigs for Pae-miR-989-3p	168
6	General Discussion	
6.1	Model for the fate of emergent intronic miRNAs	213

Thesis Abstract

There are major gaps in our understanding of the evolution of microRNA (miRNA) genes and how this relates to evolutionary changes in metazoan development. miRNAs function as *trans*-acting post-transcriptional modulators of gene expression. They are implicated in the canalization of development and the consequent provision of evolvability. Here I present data and analyses on miRNA evolution in insects and mammals, both on a phylogenetic scale as well as at the level of individual miRNA genes. I also perform case studies cross-comparing the expression and functions of intronic miRNAs to those of their host genes, and compare tissue-specific enrichment of miRNAs in a specific tissue (butterfly ovary) relative to the whole body.

A broad overview of miRNA evolution across a clade is presented for the Lepidoptera using a combination of small RNA sequencing in two lepidopteran species coupled with bioinformatic analysis of available genomic and miRNA data across the clade. This analysis reveals a burst of rapid miRNA gain on the stem lineage of the Lepidoptera prior to the radiation of the diverse butterfly and moth lineages. I also propose that the uneven distribution of miRNA gain across the lepidopteran phylogeny, coupled with observations of high numbers of newly-emergent miRNAs on terminal branches, is evidence for a model of miRNA evolution comprising high rates of turnover for newly-acquired miRNAs with variable rates of stable miRNA retention. In depth examination is carried out on two miRNA genes, both of which occur within homeobox gene introns - eutherian-specific *mir-615* and lepidopteran-specific *mir-2768*. Phylogenetic distribution and theories for the acquisition of *mir-615* in Eutheria are considered. A combination of publicly available expression data and epigenetic modification data is used to demonstrate the existence of transcriptional regulation for *mir-615* independent from that of its host gene. The study of *mir-2768* reveals the first experimental evidence for an intronic miRNA performing a complementary function to that of its host gene. I also present the first miRNA transcriptome from a butterfly ovary and compare this to whole-body miRNA expression in the same species. This provides insights into the roles performed by miRNAs in the butterfly ovary, including conserved functions, and enables comparison of ovarian miRNA tissue specificity relative to time of origin.

Overall, this work provides a multi-faceted analysis of miRNA evolution in animals and presents an integrated approach to our understanding of the roles of miRNAs in evo-devo.

Chapter 1:

General Introduction

- Overview of metazoan miRNA biology
- miRNA annotation and novel miRNA discovery
- Evolution of miRNAs and the miRNA processing pathway
- The role of miRNAs in evolution and development
- Project aims and contributions

General Introduction

During development, the information in an organism's genome is used to coordinate the generation of morphology. Evolutionary modification of this developmental process in different animal lineages has given rise to the vast diversity of body form and function observed across the Metazoa today. Many evolutionary changes to genes have been documented including gene duplications (Brooke, Garcia-Fernández, and Holland 1998; Putnam et al. 2008), coding sequence alterations (Fondon & Garner, 2004) and modifications to gene expression (Jeong, Rokas, & Carroll, 2006). While much focus has been on the evolution of *cis*-regulatory elements that affect gene expression (Putnam et al. 2008; Chan et al. 2010), there are also *trans*-acting factors that should not be overlooked. In particular, there is emerging evidence that evolutionary changes to microRNAs (miRNAs) could have a role to play in the evolution of development (Griffiths-Jones, Hui, Marco, & Ronshaugen, 2011; Hui et al., 2013; Kosik, 2009; Moran et al., 2014; Ninova, Ronshaugen, & Griffiths-Jones, 2014a).

1. Overview of metazoan miRNA biology

1.1 What are microRNAs?

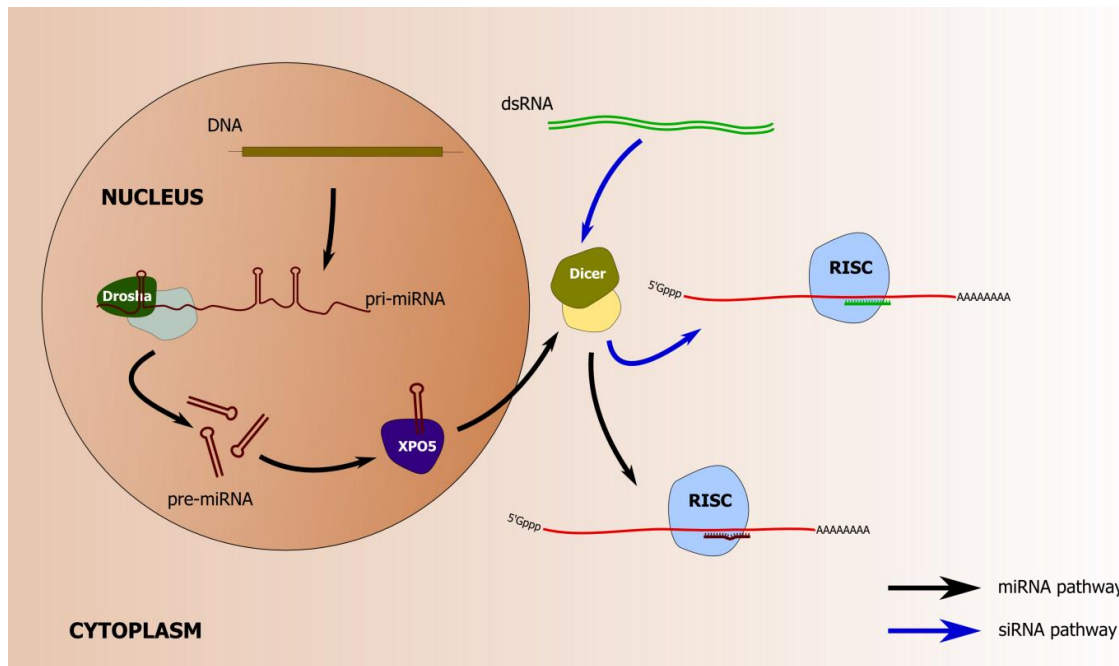
MicroRNAs (miRNAs) are small, non-coding RNAs of endogenous origin which, in their mature form, are approximately 22 nt in length and exert post-transcriptional regulation over gene expression. They are present in several eukaryotic taxa, most notably plants and animals, and characterized by a dsRNA hairpin intermediate which undergoes further processing to generate the mature product (Axtell, Westholm, & Lai, 2011).

1.2 The canonical metazoan miRNA biogenesis pathway

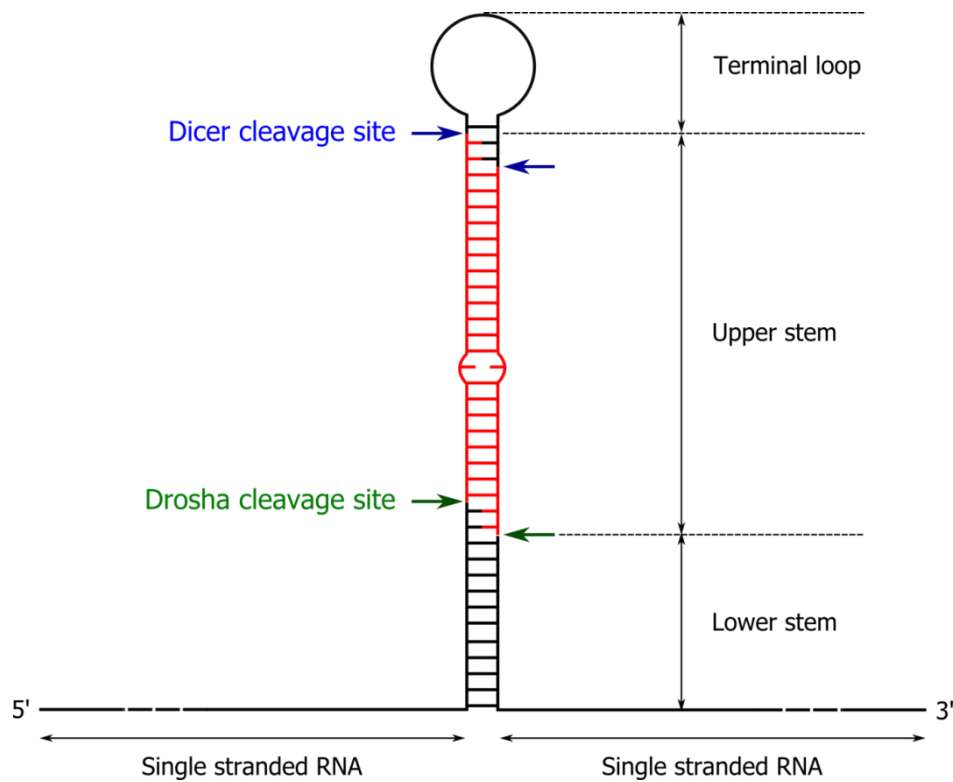
miRNAs are initially transcribed as longer primary transcripts (Figure 1.1). These primary transcripts, or pri-miRNAs, may be several kilobases long (Saini, Griffiths-Jones, & Enright, 2007) and are usually capped and polyadenylated, suggesting that

transcription is carried out by RNA polymerase II (Cai, Hagedorn, & Cullen, 2004; Y. Lee, Jeon, Lee, Kim, & Kim, 2002). Within the nucleus, primary transcripts are processed by Drosha, a nuclear RNase III enzyme, which excises the hairpins to generate precursor miRNAs (Lee et al., 2003). This cropping reaction requires a protein cofactor, Pasha (also known as DGCR8 in vertebrates), which possesses tandem dsRNA-binding domains that enable Drosha to bind to and interact with pri-miRNAs (Cullen, 2004; Tomari & Zamore, 2005). Substrate recognition by the Drosha-DGCR8 complex requires a stereotyped stem loop structure with approximately 33 bases on the stem, a terminal loop, and flanking ssRNA segments (Figure 1.2). The flanking ssRNA segments appear to be essential for pri-miRNA processing in a non-sequence specific manner, and it is suggested that the junction between the stem and the ssRNA may represent a recognition site and anchoring point for DGCR8 (Han et al., 2006). Following binding of the complex, Drosha initiates pri-miRNA cleavage ~11 bp away from the junction. This results in a precursor miRNA (pre-miRNA) hairpin of around 65 – 70 nt in length (Lee et al., 2003) and with a 3' overhang of 2 nt at the Drosha cleavage site (Figure 1.2).

The 3' overhang generated by RNase III cleavage is recognized and bound by Exportin 5 (Tomari & Zamore, 2005), which mediates pre-miRNA export from the nucleus in a RanGTP-dependent manner (Bohnsack, 2004). In the cytoplasm, pre-miRNAs undergo further processing by the cytoplasmic RNase III enzyme, Dicer (Ketting et al., 2001). As with Drosha, Dicer also requires a dsRNA-binding protein cofactor. In flies, this function is performed by Loquacious (Förstemann et al., 2005); the TAR RNA-binding protein (TRBP) and the PKR-activating protein (PACT) serve this role in mammals (Kok, Ng, Ching, & Jin, 2007). Dicer is thought to bind to the overhang generated by Drosha cleavage at the base of the hairpin (Cullen, 2004), and initiates a second cut at a fixed distance of ~22 bp from its binding site. The resultant RNA duplex has 2 nt 3' overhangs at both ends. One strand of this duplex is preferentially incorporated into an Argonaute (AGO)-containing protein complex – the RNA-induced silencing complex (RISC). Within RISC, the mature single-stranded miRNA binds directly to a member of the AGO protein family and guides the complex to specific target transcripts via complementary base pairing (Jonas & Izaurralde, 2015; Meister, 2013).

**Figure 1.1**

Metazoan miRNA biogenesis

**Figure 1.2**

Structure of a typical metazoan pri-miRNA

1.3 Mechanisms of miRNA targeting and miRNA action

While plant miRNAs tend to have targets with perfect or near-perfect complementarity, animal miRNAs typically exhibit imperfect pairing to their targets (Axtell et al., 2011). A short 'seed' sequence comprising nucleotides 2-7 of the miRNA is often fully paired and essential for miRNA targeting. High complementarity is rare, but known to exist (Moran et al., 2014; Yekta, Shih, & Bartel, 2004).

The regulatory effects of miRNA targeting may be one of several outcomes. In the few cases where perfect or near-perfect pairing is attained, the mRNA target is cleaved (Meister, 2013; Moran et al., 2014). With imperfect complementarity, the mRNA undergoes translational silencing, deadenylation and/or degradation as a consequence of miRNA binding. Studies done on the dynamics of miRNA-mediated gene repression in several systems seem to suggest that target transcripts are first subjected to translational repression, followed by deadenylation and decay (Béthune, Artus-Revel, & Filipowicz, 2012; Djuranovic, Nahvi, & Green, 2012).

AGO proteins work in conjunction with a number of other proteins to bring about miRNA-guided gene regulation. Members of the GW protein family are characterized by the presence of an N-terminal AGO-binding domain containing glycine-tryptophan (GW), which directly interacts with AGO proteins. The C-terminal half contains motifs enabling the interaction of the GW protein with the poly(A) tail on the mRNA through poly(A)-binding protein C (PABPC). This interaction is thought to promote translational silencing (Meister, 2013). Additionally, the C-terminal half is also capable of interacting with the CCR4-NOT complex, which promotes deadenylation of the target mRNA. Deadenylation is followed by decapping and subsequent degradation by the XRN1 exonuclease (Fabian et al., 2011).

Sequestration of target transcripts away from the translational apparatus may be another means by which miRNAs achieve repression. Target mRNAs repressed by miRNA have been reported to accumulate in discrete cytoplasmic loci, termed 'P-bodies', which are known to be sites of translational repression and subsequent mRNA decay (Pillai, Bhattacharyya, & Filipowicz, 2007). AGO and GW proteins are also known to localize to P-bodies (Eulalio, Huntzinger, & Izaurralde, 2008).

1.4 Requirement for 3' UTR localization of miRNA target sites

It is widely accepted that metazoan miRNAs bind to their target transcripts predominantly in 3' UTRs. Gu and colleagues proposed a model to explain the biological basis for this apparent restriction, suggesting that active translation through open reading frames interferes with the binding of the RISC complex to the mRNA and thereby reduces efficiency of miRNA-mediated target repression (Gu, Jin, Zhang, Sarnow, & Kay, 2009). The authors of this study discovered that modifying the stop codon to allow translation to extend beyond the miRNA target sequence in the 3' UTR abolished repression. Inducing ribosome pausing through the insertion of rare codons upstream of the target site in the extended open reading frame restored miRNA-mediated knockdown, suggesting that slowing down translational progress permits RISC to attach to the target and induce repression.

While there may indeed be mechanistic reasons for animal miRNA target sites to be predominantly located in 3' UTRs, it is likely that this effect has been overstated for a number of reasons related to *in silico* predictions for miRNA binding sites. The first animal miRNAs to be discovered, *C. elegans* lin-4 and let-7, were found to have binding sites in the 3' UTRs of their respective targets which had previously been identified through genetic analysis (Lee, Feinbaum, & Ambros, 1993; Reinhart et al., 2000). This is likely to have set a precedent for future research into animal miRNA targets (Brodersen & Voinnet, 2009). Many miRNA target prediction algorithms focus exclusively on 3' UTRs (Lewis, Burge, & Bartel, 2005; Maragkakis et al., 2009; Wang & El Naqa, 2008), which may have led to the underreporting of miRNA binding sites found elsewhere on target transcripts. Functional miRNA target sites have been identified in the open reading frames of metazoan mRNAs (Duursma, Kedde, Schrier, le Sage, & Agami, 2008; Moran et al., 2014), as well as in 5' UTRs (Lytle, Yario, & Steitz, 2007).

1.5 miRNAs within a host transcription unit

In 2004, Rodriguez and colleagues presented their findings on the annotation of the genomic position and context of miRNAs in the human and mouse genomes (Rodriguez, Griffiths-Jones, Ashurst, & Bradley, 2004). It was discovered that a large proportion of miRNA genes are located within introns of other genes, both in protein-

coding loci as well as in genes encoding other non-coding RNAs. A later study reported that, at least in humans, the proportion of intronic miRNAs (63%) is higher than expected by chance given the percentage of the genome (30%) consisting of introns (Campo-Paysaa, Sémon, Cameron, Peterson, & Schubert, 2011). In both the human and amphioxus genomes, the majority of intronic miRNAs were encoded in the same orientation as their host genes (77% in humans, 85% in amphioxus). Based on this observation, the authors suggest that intronic miRNAs are processed from their host gene transcripts rather than possessing their own transcriptional control sequences.

Hypothetical outcomes from a given locus containing an intronic miRNA are summarized in Figure 1.3. Assuming a single transcriptional unit under control of the host gene promoter, final mature RNA products generated from the locus would ultimately depend on the balance between miRNA cropping and pre-mRNA splicing. The consensus opinion on cropping and pre-mRNA splicing is that they are not mutually exclusive events (Shomron & Levy, 2009). Several lines of evidence point to the origination of intronic miRNAs and their host mRNAs from the same transcript, including correlated expression patterns and the discovery of partially-spliced EST fragments with either their 5' or 3' ends matching Drosha cleavage sites (Baskerville & Bartel, 2005).

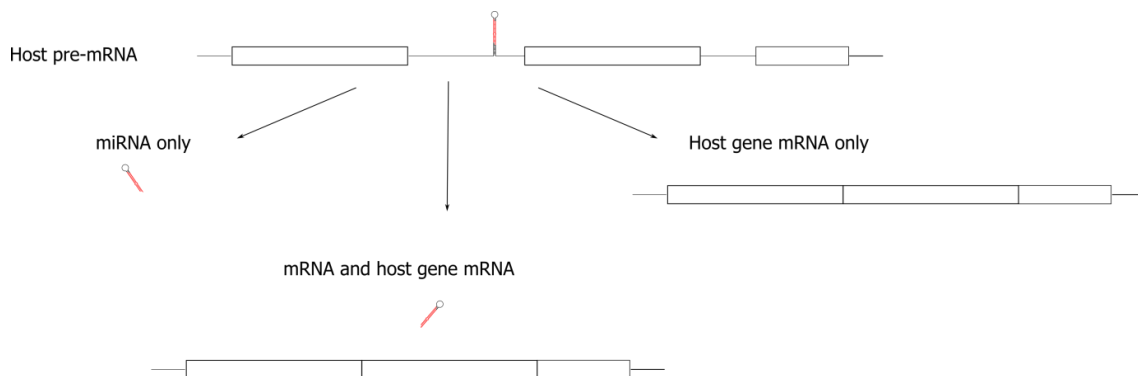


Figure 1.3

Hypothetical outcomes from transcription at a protein-coding locus hosting an intronic miRNA

The literature is somewhat sparser on the potential for exclusive production of either the miRNA or its host transcript. The most straightforward means of obtaining only the miRNA without its host gene would be for the intronic miRNA to possess its own transcriptional control sequences, which may or may not be situated within the host intron. In support of this idea, there have been documented cases whereby host and

miRNA expression are poorly correlated (Baskerville & Bartel, 2005; Liang, Ridzon, Wong, & Chen, 2007; Mi et al., 2007). Promoter prediction analysis suggests that about 35% of intronic miRNAs possess their own regulatory sequences (Monteys et al., 2010). A case study in point is that of the *mir-128-2* gene, which is located within an intron of *ARPP21*. Expression of *mir-128-2* is significantly higher in acute lymphoblastic leukaemia but not acute myeloid leukaemia, and discordant with that of its host gene (Mi et al., 2007). ChIP analysis suggests that this miRNA is expressed under dual control by an intronic Pol III promoter as well as the host gene Pol II promoter (Monteys et al., 2010).

In the converse scenario, it is not immediately clear whether or not production of the mature host mRNA is possible without simultaneous production of the miRNA. Both microprocessing and the splicing reaction are thought to occur co-transcriptionally (Agranat-Tamir, Shomron, Sperling, & Sperling, 2014; Morlando et al., 2008), and it has been suggested that the microprocessing reaction may even promote splicing of introns containing pre-miRNAs (Agranat-Tamir et al., 2014). If it is indeed the case that newly-evolved intronic miRNAs are necessarily coexpressed with their host genes, they would have to provide either a neutral or beneficial contribution to host gene function in order to avoid being eliminated by selection.

Another potentially overlooked area in the study of miRNAs encoded within host transcription units is the possibility that miRNA genes may occur elsewhere within the host locus apart from intronic regions (Berezikov et al., 2011). Most miRNA annotation pipelines will omit predictions corresponding to any mature mRNA sequences (Kang & Friedländer, 2015). While it is not implausible that miRNA genes may overlap with untranslated regions or coding regions, the discovery and annotation of such sequences would pose a bioinformatic challenge and should also ideally involve rigorous experimental validation of any putative miRNA genes identified.

1.6 Distinction between miRNAs and siRNAs

The mature products of miRNAs are virtually indistinguishable from those of small interfering RNAs (siRNAs). The biogenesis of both classes of small non-coding RNAs utilises overlapping components of the RNA interference pathway, most notably Dicer. Both also exhibit similar sequence-based targeting and post-transcriptional repression

of target transcripts (Carthew & Sontheimer, 2009). Key similarities and differences between miRNAs and siRNAs are summarized in Table 1.1.

The main distinction between miRNAs and siRNAs lies in their mode of biogenesis. A hallmark characteristic of miRNAs is the presence of a stem-loop precursor of endogenous origin, the length of which varies between taxa (Axtell et al., 2011). siRNAs are derived from endogenous or exogenous long, double-stranded RNAs and begin their biogenesis pathway with Dicer cleavage in the cytoplasm (Figure 1.1). Dicer is involved in precursor cleavage and maturation of all plant and animal siRNAs and miRNAs, but the Drosha-DGCR8 complex appears to be unique to the metazoan miRNA pathway (Shabalina & Koonin, 2008).

Table 1.1: Similarities and differences between miRNAs and siRNAs

	miRNA	siRNA
Origin	Endogenous genes	Endogenous or exogenous double stranded RNAs
Precursors	Foldback structure containing a hairpin intermediate, often with mismatches in animals	Long, fully complementary double-stranded RNAs
Target complementarity	Partial (in animals), high complementarity (plants)	Full complementarity
Effect	Translational repression, deadenylation, exonuclease degradation, cleavage	Cleavage of target RNAs
Primary function	Regulation of endogenous gene expression	Genome defence against viruses and transposons
Regulation	Posttranscriptional regulation of target gene expression through association with RISC complex	
Mature products	Small single-stranded RNAs, approximately 20-25 nt in length	
Involvement of RNAi pathway components	Dicer required for maturation of both miRNAs and siRNAs, Ago proteins coordinate silencing effector functions	

1.7 Mirtrons and other non-canonical miRNA processing pathways

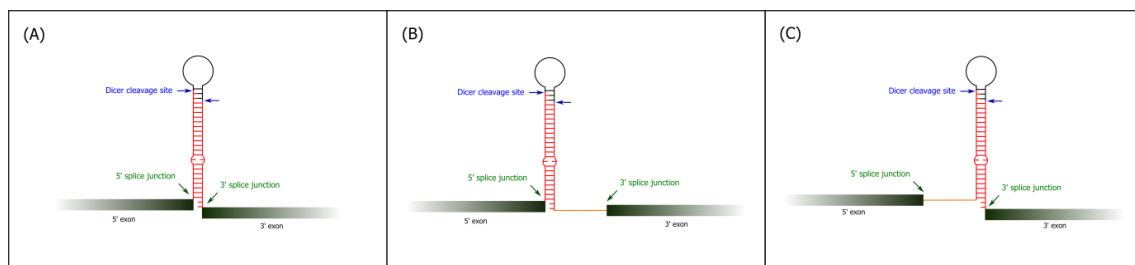
A number of variations on miRNA processing have been reported, which bypass one or more steps in the canonical miRNA biogenesis pathway. A detailed discussion of alternative miRNA production is provided by (Miyoshi, Miyoshi, & Siomi, 2010) and summarized in Table 1.2. This discourse will focus specifically on the mirtron pathway, which is the best characterized of the alternative miRNA biogenesis pathways, and was the first to be identified.

Table 1.2: Non-canonical miRNA biogenesis pathways

Drosha-independent	Dicer-independent
Mirtron pathway snoRNA-derived pathway shRNA-derived pathway tRNA-derived pathway tRNaseZ-derived pathway pre-esiRNA-derived pathway	AGO-dependent pathway tRNaseZ-dependent pathway

Mirtrons were first identified in *Drosophila melanogaster* and *Caenorhabditis elegans* with the discovery of small RNAs arising from miRNA-sized introns capable of adopting pre-miRNA folds, where the mature products mapped precisely to the 3' and 5' splice junctions (Okamura, Hagen, Duan, Tyler, & Lai, 2007; Ruby, Jan, & Bartel, 2007). A critical difference between mirtrons and intronic miRNAs is that mirtrons bypass Drosha entirely, while intronic miRNAs are processed along the canonical miRNA pathway. The mature products of mirtrons function as typical miRNAs, but their precursor stems are shorter than those of canonical miRNAs due to the lack of a lower stem region responsible for recruitment and cleavage by the Drosha/DGCR8 complex (Figure 1.4). Splicing of the host transcript, followed by lariat debranching, allows the intron to adopt a pre-miRNA fold and enter the miRNA processing pathway. The splicing reaction generates both ends of the precursor such that the mature miRNAs correspond exactly to the splice sites, suggesting that splicing effectively substitutes for Drosha-mediated cropping in generating the pre-miRNA intermediate.

Mirtron-like loci where the hairpin resides at either end of the intron, incorporating a short 3' or 5' tail, have also been documented (Westholm & Lai, 2011). These tailed mirtrons are likely to require trimming after splicing and prior to export from the nucleus.

**Figure 1.4**

(A) Typical mirtron, (B) 3' tailed mirtron, (C) 5' tailed mirtron

2. miRNA annotation and novel miRNA discovery

2.1 The need for characterization of miRNA transcriptomes

miRNAs are an essential component of genetic regulatory networks (GRNs) involved in the control of processes essential to organismal function, including cell differentiation (Kanellopoulou et al., 2005), embryogenesis (Pasquinelli, Hunter, & Bracht, 2005; Yekta et al., 2004; Yekta, Tabin, & Bartel, 2008), post-embryonic development (Kenny, Quah, Holland, Tobe, & Hui, 2013; Lozano, Montañez, & Belles, 2015; Reinhart et al., 2000) and reproductive function (McBride et al., 2012; Zovoilis et al., 2010). The understanding and characterization of the roles played by miRNAs in such networks is therefore of importance in gaining a big picture perspective of the biological processes which are indispensable to life.

In this respect, gaining an appreciation for the involvement of miRNAs in developmental GRNs across different taxa would shed light on the evolution of the architecture that shapes animal body plans. This would complement our current understanding on other forces guiding the evolution of development, such as gene duplications (Brooke et al., 1998), protein modifications (Fondon & Garner, 2004) and the evolution of *cis*-regulatory elements (Wittkopp & Kalay, 2012).

2.2 Early advances in miRNA discovery

With hindsight, the significance of animal miRNAs in the regulation of development was foreshadowed by the earliest work in the field. Research into metazoan miRNAs has its roots intimately entwined in the study of animal developmental regulation. The first animal miRNA to be reported was *lin-4* in *C. elegans* (Lee et al., 1993; Wightman, Ha, & Ruvkun, 1993), which was originally identified genetically as a post-transcriptional negative regulator of the *Lin-14* gene involved in the heterochronic pathway which governs developmental timing. The *lin-4* gene was cloned and found to encode two different transcripts of 22 nt and 61 nt with antisense complementarity to the 3' UTR of the *lin-14* RNA. Both the sequence complementarity as well as the regulatory relationship was conserved in other *Caenorhabditis* nematode species studied. This report was followed by the discovery of a second *C. elegans* small regulatory RNA, *let-7*, several years later (Reinhart et al., 2000). At the time, these small regulatory RNAs

were thought of as being a peculiarity of the nematode heterochronic pathway. The first realization that this phenomenon was much more widespread and pervasive across the Metazoa than previously suspected came with three related studies in 2001, which identified more small RNAs of a similar class in *C. elegans*, some of which had potential homologues in *D. melanogaster* and in human cells (Lagos-Quintana, Rauhut, Lendeckel, & Tuschl, 2001; Lau, Lim, Weinstein, & Bartel, 2001; Lee & Ambros, 2001). The term 'microRNA' was coined in agreement by the authors of all three reports to describe these small RNAs with as yet undetermined regulatory roles in a variety of biological processes.

2.3 Experimental and computational approaches to miRNA annotation

The two broad categories of approaches to miRNA discovery used in early studies within the field, computational prediction and sequencing, are still relevant to miRNA research today. Using a computational approach, potential miRNAs in a given genome can be predicted using appropriate bioinformatic tools to screen for features such as hairpin potential, folding energy and conservation. Alternatively, miRNAs can be studied experimentally by isolating total RNA fractions from tissues of interest and subjecting these to size-selection and sequencing (Lagos-Quintana et al., 2001).

Both bioinformatics-based and sequencing strategies have their advantages and drawbacks. A significant strength of computational-based miRNA prediction is that any detection bias due to expression differences between different miRNAs is avoided. However, a major challenge presented to any computational prediction algorithm is the avoidance of false positives due to the large number of genomic sequences which adopt pre-miRNA-like secondary structures by chance alone, not all of which necessarily enter the miRNA processing pathway (Kang & Friedländer, 2015; Liu et al., 2008). The requirement for a predicted hairpin precursor is universal across miRNA prediction algorithms as this is a defining characteristic of miRNAs, on top of which additional criteria such as thermodynamic stability and conservation may be implemented (Berezikov, Cuppen, & Plasterk, 2006). The problem with utilising conservation data is the loss of the ability to identify potential evolutionary novelties in a given genome.

Traditional approaches to the experimental identification of miRNAs were based on a protocol originally derived for cloning siRNAs (Berezikov, Cuppen, & Plasterk, 2006). This involved size-fractionation of cDNA libraries on polyacrylamide gels, recovery of the appropriate size fraction (20-25 nt), reverse transcription and generation of cDNA libraries. Individual clones could then be Sanger sequenced and analysed. This allowed for a higher degree of certainty that reported miRNAs are present in the transcriptome. Sequence data was typically filtered against known non-miRNA annotations and predicted loci checked for appropriate secondary structure in order to ensure accuracy of predictions (Kang & Friedländer, 2015). The main problem with this approach was its lack of sequencing depth; only a single clone was recovered for many of the miRNAs identified in one of the pioneering 2001 *C. elegans* miRNA studies (Lau et al., 2001), suggesting that a large proportion of the miRNA transcriptome was not being sampled. Inadequate depth biases miRNA discovery towards highly expressed sequences, making it difficult to identify miRNAs with low expression levels as well as those which might be cell type- or developmental stage-specific. The development of deep sequencing technologies such as 454 (Rothberg & Leamon, 2008) and Illumina sequencing (Metzker, 2010) have helped overcome these issues to a large extent, permitting greatly increased read depth to be obtained in a much shorter timeframe at much lower cost compared with the traditional cloning strategy. While this has permitted the detection of miRNAs with ever lower expression levels, the large volume of data generated by next-generation sequencing technologies presents a new computational challenge in distinguishing genuine miRNAs from the enormous variety of RNA species present in cells.

2.4 Identification of miRNAs in deep sequencing datasets

The unprecedented sensitivity with which high-throughput deep sequencing technology permits the detection of very lowly expressed miRNAs also means that short RNAs other than miRNAs present at similar levels are picked up in a sequencing run, giving rise to the potential for large numbers of false positives (Kang & Friedländer, 2015). These may be other short non-coding RNAs such as siRNAs and Piwi-interacting RNAs (piRNAs), or degradation products of longer coding and non-coding RNAs. As a result, the raw output from next-generation sequencing runs typically requires additional post-filtering steps on top of those used in traditional Sanger sequencing-based strategies. A number of software tools have been developed to assist with

miRNA annotation from deep sequencing data (Table 1.3). In order to ensure robustness of miRNA predictions, output from these tools should be validated. A summary of selection criteria used to identify genuine miRNA precursors is shown in Table 1.4.

Table 1.3: Software tools for analysis of miRNA deep sequencing data

Program	Genome Dataset	Description	Year
DSAP	List of 142 animal, plant and virus genomes	Offers a suite of data processing features including adaptor trimming, classification of ncRNAs, permits cross-sample analysis	2010
miRanalyzer	Hs, Mm, Rn, Dm, Ce, Dr, Cf*	Detects known miRNAs, filters against other RNA species, predicts new miRNAs	2009
miRDeep, mirDeep2	Any	Predicts miRNA precursors based on position of read mapping to genome, scores precursors for position of mapped reads, read count and predicted secondary structure	2008, 2011
miRDeep*	Any	miRNA prediction strategy from miRDeep2 modified to improve novel miRNA prediction, target prediction functionality, launched as standalone application	2013
miRDeep-P	Plants	miRDeep algorithm modified to account for plant miRNA biogenesis	2011
miRExpress	Any	Aligns sequencing reads against known miRNAs to generate expression profile data. Genomic data not required.	2009
miRNAkey	Not required	Trims adaptors, maps reads to known miRNA databases, allows for differential expression analysis	2010
miRTrap	Any	Predicts miRNA precursors based on position of read mapping to genome, factoring in structure, relative positions of 5' and 3' products and minimum read count	2010

*Hs = *Homo sapiens*, Mm = *Mus musculus*, Rn = *Rattus norvegicus*, Dm = *Drosophila melanogaster*, Ce = *Caenorhabditis elegans*, Dr = *Danio rerio*, Cf = *Canis familiaris*

Table 1.4: Annotation criteria for animal miRNAs, based on Ambros et al (2003) and Kozomara & Griffiths-Jones (2010)

Expression criteria	
Detection of a small RNA (18-24 nt) transcript by sequencing.	Star sequences should be detectable with deep sequencing.
Multiple reads for mature product	Preferably from multiple independent experiments
Biogenesis criteria	
Prediction of a stem loop precursor containing the small RNA sequence	Hairpin should be the folding alternative with the lowest minimum free energy Hairpin should contain at least 18 paired bases on the main stem No large internal loops or bulges (>4 nt) should be present Reads should map to the precursor with 5' homogeneity on both arms Predicted mature products should overlap with 3' overhangs
Phylogenetic conservation	Both the precursor hairpin and the mature miRNA product show phylogenetic conservation
No overlap of non-miRNA annotations	miRNA predictions should be screened against RepBase and Rfam to remove sequences related to repetitive elements and other ncRNAs

3. Evolution of miRNAs and the miRNA processing pathway

3.1 Acquisition of the miRNA processing pathway

The miRNA processing pathway is a subset of a large network of eukaryotic RNA interference (RNAi) systems (Shabalina & Koonin, 2008), three core components of which are Ago-Piwi, Dicer and RNA-dependent RNA polymerase (RdRP). Conservation of the core RNAi proteins across the majority of the eukaryote lineages studied (with the possible exception of Excavata) suggests that the last common eukaryotic ancestor possessed some form of RNAi system, possibly for use in genome defence. Although prokaryotes have a functionally analogous system, it is thought that this is not related to eukaryotic RNAi.

It is unlikely that a functional miRNA biogenesis pathway was in place in the last common eukaryotic ancestor. Current evidence points to several independent acquisitions of the miRNA pathway in eukaryotes through exaptation of the shared RNAi machinery (Tarver, Donoghue, & Peterson, 2012). There is consensus on the existence of a functional miRNA pathway in eumetazoans, embryophyte plants, and the green and brown algae represented by *Chlamydomonas reinhardtii* and *Ectocarpus siliculosus* (Shabalina & Koonin, 2008; Tarver et al., 2015). Although miRNAs have also been reported to exist in other lineages (Braun et al., 2010; Chen, Collins, Biggs, & Penny, 2009; Hinas et al., 2007; Huang, He, & Wang, 2011; Huang et al., 2012; Lin et al., 2009), these claims have been disputed due to lack of convincing evidence that sequencing reads correspond to actual miRNAs (Tarver et al., 2012). Each of the lineages in which miRNAs are known to exist has a distinct set of miRNAs which does not overlap with that found in any of the other lineages, which may be indicative of multiple independent origins of the eukaryotic miRNA pathway.

Further evidence for the hypothesis that miRNAs have more than one distinct origin within the eukaryotes comes from the comparison of their biogenesis and mechanism of action between animals and plants (Table 1.5), the two groups in which they have been most extensively characterized (Axtell et al., 2011; Shabalina & Koonin, 2008; Tarver et al., 2012). Notably, plants perform both pri-miRNA cropping and pre-miRNA dicing steps in the nucleus using the same protein (DCL1). Plant pre-miRNAs are also much more variable in length compared to those in eumetazoans, and mature plant miRNAs typically bind with extensive complementarity to their targets. In recent years the separate origin hypothesis has been challenged, citing the discovery of orthologues

of the plant Dicer binding partner HYL1 in sponges and ctenophores (Moran, Praher, Fredman, & Technau, 2013), along with frequent observation of cnidarian miRNAs binding their targets with high complementarity (Moran et al., 2014). This evidence is likely to be inconclusive, given the very different biology of miRNAs in plants and animals and the lack of strong support for the existence of functional miRNAs in intermediate lineages. It is more likely that the miRNA pathway in basally branching metazoans is slightly different from that of eumetazoans due to greater flexibility in miRNA regulation at the base of the metazoan tree, or due to multiple origins of miRNAs within the Metazoa.

It is generally accepted that the miRNA machinery was in place in the last common eumetazoan ancestor (Wheeler et al., 2009). However, the origins of miRNA-mediated regulation at the base of the metazoan tree are somewhat obscure, as are the actual relationships between the basally branching metazoan lineages (Edgecombe et al., 2011). A summary of our current knowledge on the distribution of miRNAs and components of the miRNA processing machinery at the base of the Metazoa is illustrated in Figure 1.5.

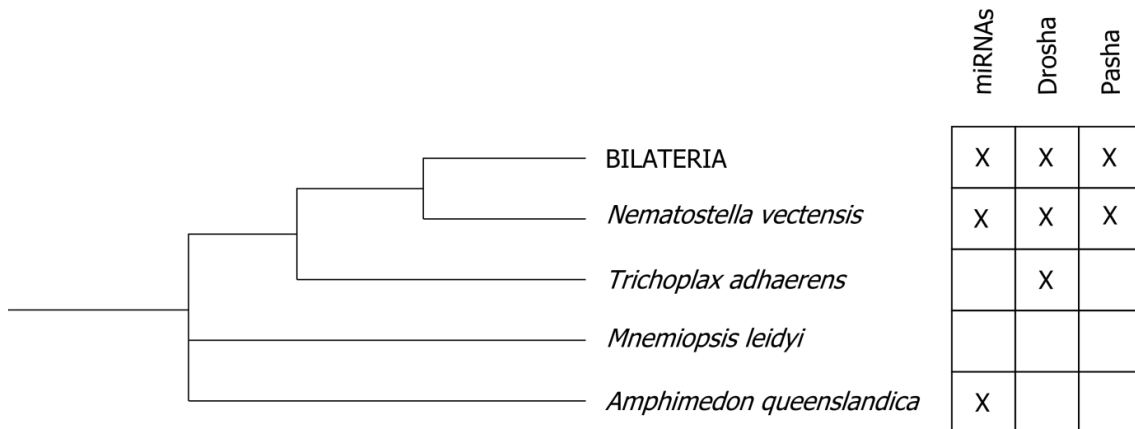


Figure 1.5

Known distribution of miRNAs and miRNA pathway components in basally-divergent metazoan lineages

Table 1.5: Main differences between plant and animal miRNAs

Plants		Animals
Biogenesis		
Location	Pri-miRNA cropping and pre-miRNA dicing both occur in nucleus	Pri-miRNAs cropped in nucleus; pre-miRNAs exported for dicing in cytoplasm
Key components	Cropping and dicing both performed by Dicer-like 1 (DCL1), Drosha not found in plant genomes	Cropping performed by Drosha, pre-miRNA processing by Dicer
Precursor size	Variable between about 60 nt to >400 nt	Approximately 60-75 nt
Clustering	Less prevalent than in animals	More common than in plants; clusters may incorporate unrelated miRNA families
Function		
Target complementarity	Typically high	Often contains mismatches, particularly outside of seed region
Target specificity	One to a few targets	Typically broad regulation of multiple genes
Repression	Primarily by target cleavage	May occur by translational repression, deadenylation, or occasionally target cleavage
Number of target sites	Usually only one per target transcript	Variable (often multiple) number of sites per transcript
Position of binding sites	Almost exclusively in open reading frames	Most documented cases in 3' UTRs, binding sites in coding sequences also documented (usually with higher complementarity)

It has been argued that the use of Drosha and Pasha orthologues for miRNA biogenesis in all animal lineages possessing miRNAs, and their absence in all non-metazoans studied to date, is indicative of a single origin for metazoan miRNAs. No convincing evidence was found for the presence of miRNAs in Placozoa (Grimson et al., 2008) or Ctenophora (Maxwell, Ryan, Schnitzler, Browne, & Baxeavanis, 2012) from small RNA sequencing in *Trichoplax adhaerens* and *Mnemiopsis leidyi* respectively. The Drosha and Pasha proteins are also missing from the *Mnemiopsis* genome (Maxwell et al., 2012). Pasha is absent in *Trichoplax* (Grimson et al., 2008), although Drosha has been identified. The authors of this study have suggested that Drosha may be involved in ribosomal RNA maturation while Pasha has no known roles outside of miRNA biogenesis. Within Porifera, miRNAs have been reported in demosponges (*Amphimedon queenslandica* and *Haliclona* sp., Wheeler et al., 2009) as well as the *Sycon* and *Leucosolenia* calcisponges (Robinson et al., 2013). It is currently unclear if homoscleromorphs possess functional miRNAs (Robinson et al., 2013). However, there appears to be no overlap in the miRNA complements between demosponges and calcisponges, and no overlap between the complements of either group and Eumetazoa. The miRNA precursors of *Amphimedon* and calcisponges also differ in structure; the few calcisponge pre-miRNAs reported resemble eumetazoan miRNAs in structure (Robinson et al., 2013), while *Amphimedon* pre-miRNAs are more similar to those of plants (Grimson et al., 2008). These findings were presented as being suggestive of multiple instances of miRNA acquisition at the base of the Metazoa (Robinson et al., 2013). The phylogenetic status of Porifera is controversial; while there is strong support for the monophyly of sponges (Gazave et al., 2011; Philippe et al., 2009), some authors suggest a possible paraphyletic origin for the clade (Sperling, Peterson, & Pisani, 2009). Under the latter scenario, it is plausible that there were two separate origins of metazoan miRNAs – once in demosponges and once in the common lineage leading to calcisponges and the rest of the animals (Wheeler et al., 2009).

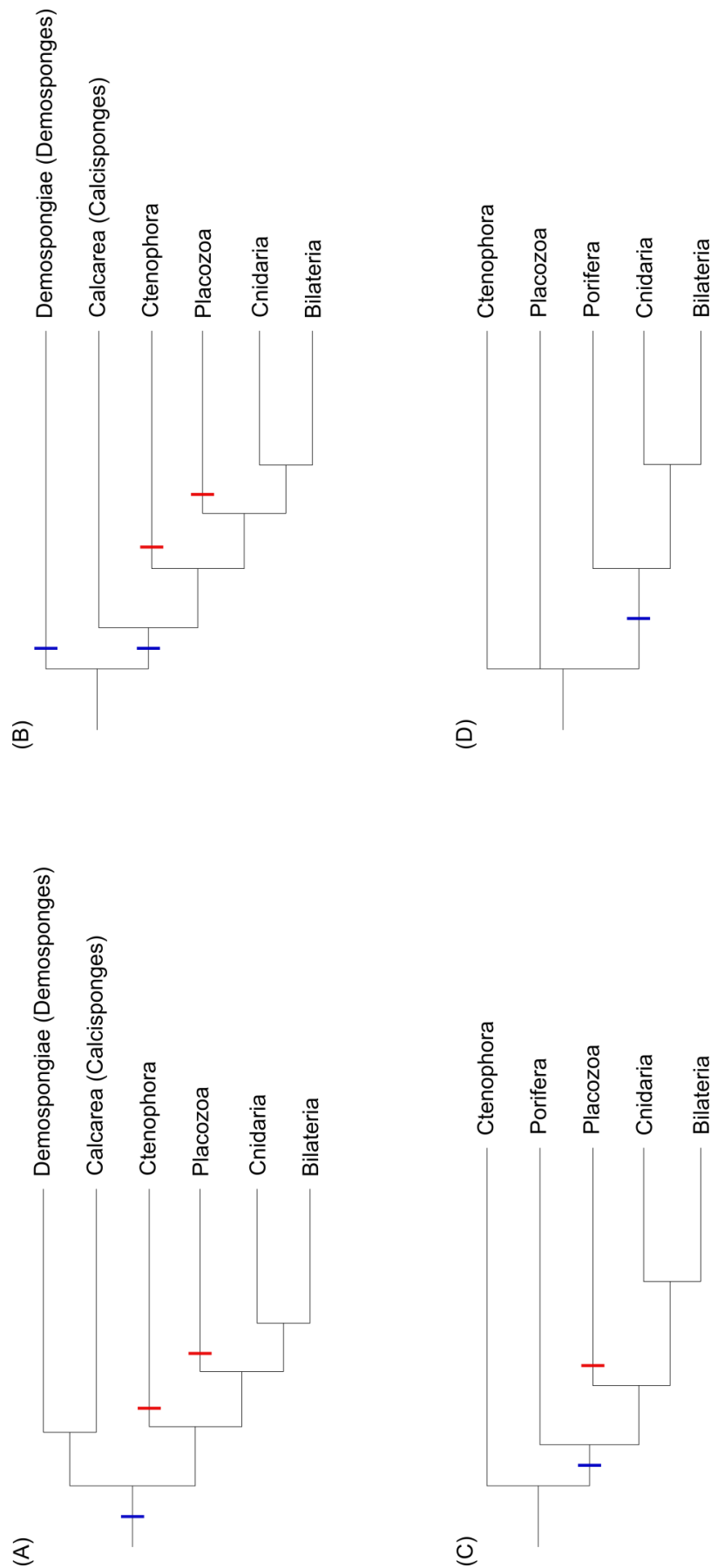


Figure 1.6

Alternative scenarios for acquisition of the miRNA biogenesis pathway in basally branching Metazoa. Blue bars represent gain of miRNA functionality while red bars represent loss. (A) Single origin of miRNAs in Metazoa. Differences in miRNA structure and functionality in basally divergent lineages are due to deep divergence. (B) Sponge paraphyly and separate origin of miRNAs in demosponges explains observed differences from calcisponge/eumetazoan miRNAs. (C) and (D) indicate miRNA acquisition in the Metazoa after some lineages have branched off.

Ultimately, evolutionary relationships amongst basally-branching metazoan lineages will need to be resolved prior to making any conclusions regarding the number of separate origins of the miRNA pathway early on in animal evolution. However, the current data outlines a scenario whereby miRNA-mediated regulation of gene expression was not strictly essential to organismal function prior to the origin of the Eumetazoa. Indeed, it has been argued that the lack of any conserved miRNAs across sponges suggests that poriferan miRNAs lack a conserved role in specifying cell type identity and morphology (Robinson et al., 2013). This may generally be true in basally branching metazoan lineages, which tend to have relatively simple body plans and few tissue types, and if true could explain the putative secondary losses of the miRNA pathway in Ctenophora and Placozoa.

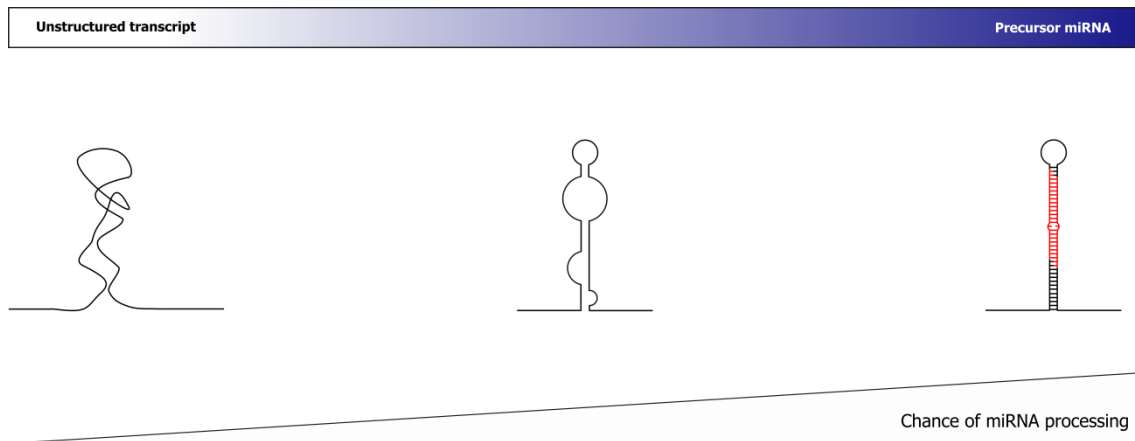
3.2 Mechanisms for novel miRNA emergence in the Metazoa

New miRNA genes, like their protein-coding counterparts, may arise through duplication and subsequent modification from existing genes. However, a crucial difference between the evolution of protein-coding genes and that of protein-coding genes is the relative ease at which a new miRNA may emerge *de novo* from existing transcribed RNAs (Berezikov, 2011).

The *de novo* emergence of miRNA hairpins from previously unstructured transcribed RNAs is likely to be the dominant means of miRNA innovation in metazoans, and may be intimately tied in with the reduced complementarity required for animal miRNA regulation (Axtell et al., 2011). Plant miRNAs usually target one to a few targets for downregulation due to the requirement for extensive pairing, which would allow for any potential detrimental effect of a novel miRNA being compensated for by increasing expression of the target. Due to reduced requirements for complementary base pairing in animals, a novel miRNA may have broad-ranging detrimental effects which would not be so easily compensated for. It is therefore more likely for new animal miRNAs to emerge with low expression levels and strict tissue-specificity (Chen & Rajewsky, 2007). One means of achieving low expression levels of newly emergent miRNAs would be through inefficient entry of precursors into the miRNA biogenesis pathway, perhaps due to suboptimal secondary structures which are only occasionally processed to generate functional miRNAs (Axtell et al., 2011). Such structures are likely to occur fairly frequently in animal transcriptomes by chance alone. In this scenario, an

occasional beneficial regulatory effect conferred by a novel miRNA might be selected for. This would occur together with the selective loss of target sites mediating detrimental regulation. Selection may also occur for improved secondary structure or increased transcription in order to boost production of the mature miRNA (Axtell et al., 2011). It is likely that the transition from an unstructured RNA to a fully-functional miRNA precursor is a gradual progression (Figure 1.7) along which the RNA slowly acquires increased competence for processing as a miRNA (Berezikov, 2011).

Intronic sequences, both from coding and non-coding transcripts, are likely to represent a major source of pre-existing transcribed RNAs for *de novo* miRNA acquisition (Campo-Paysaa et al., 2011; Meunier et al., 2013). Comparative analyses carried out on the genomic locations of miRNAs in *Saccoglossus kowalevskii* (acorn worm), *Strongylocentrotus purpuratus* (sea urchin), *Branchiostoma floridae* (lancelet), *Petromyzon marinus* (lamprey) and *Homo sapiens* reveal that in humans, 59% of species-specific miRNAs are intronic, while only 38% of miRNAs conserved across all five species occur in introns (Campo-Paysaa et al., 2011). This suggests that there is a tendency for newly-acquired miRNAs to be located in introns. Similar findings were presented using a later study on mammals (Meunier et al., 2013), where introns in general displayed an overrepresentation of miRNA loci. This effect was more pronounced for miRNAs emerging after the bird-mammal split (42-71%), than for miRNAs emerging before the split (30-54%). This preferential tendency for new miRNAs to arise in introns, termed 'intronic exaptation' (Campo-Paysaa et al., 2011), is thought to be a significant contributor to miRNA innovation due to the ease of acquiring the appropriate secondary structure within an existent transcription unit without the need for evolving new transcriptional control sequences. Subsequent loss of the host gene may occur together with maintenance of the formerly intronic miRNA, which may account for low proportions of conserved intronic miRNAs observed (Campo-Paysaa et al., 2011).

**Figure 1.7**

De novo miRNA evolution from existing transcribed RNAs with approximate precursor structure.

De novo emergence of new miRNAs may also occur through the formation of new hairpins in existing pri-miRNA transcripts, leading to the formation of miRNA clusters (Marco, Ninova, Ronshaugen, & Griffiths-Jones, 2013). Such clusters may also arise due to the tandem duplication of existing miRNA genes. An example which illustrates both of these modes of cluster formation is the evolution of *mir-2* and its related miRNA genes. It is likely that miR-2 was acquired via *de novo* hairpin formation within the primary transcript of miR-71. The miR-2 family underwent subsequent expansion through tandem duplication, giving rise to the related miR-13 subfamily (Marco, Hooks, & Griffiths-Jones, 2012).

Another mechanism of novel miRNA acquisition, thought to occur more commonly in plants than in animals, is through intragenomic duplication of an existing gene (Axtell et al., 2011). This would result in the production of a 'proto-miRNA' as a long foldback structure which upon transcription would be capable of producing small RNAs exactly complementary to the original transcript. In plants, it is thought that such 'proto-miRNAs' pass through a stage where they are capable of being processed as siRNAs by different Dicers, during which multiple small RNA species can be generated from the same precursor. Eventually, random mutations or selection for a single product would degrade extensive complementarity such that only one region retains target pairing and corresponds to the mature miRNA. It is suggested that this mechanism for miRNA innovation is uncommon in metazoans due to the greater chance of off-target effects. Local duplication was postulated, but not proven, for vertebrate *mir-196* and *Hoxb8*.

(Axtell et al., 2011), which are in close proximity and exhibit extensive sequence complementarity (Yekta et al., 2004).

Transposable elements (TEs) and repetitive elements in the genome represent a third potential source for miRNA innovation (Daniel, Behm, & Öhman, 2015; Smalheiser & Torvik, 2005). For example, *Alu* elements are highly abundant in the human genome and represent a potential source of novel miRNAs, including intronic miRNAs (Daniel et al., 2015). They are unevenly distributed in the human genome, displaying a strong preference for integration within gene-rich regions (Sela et al., 2007). *Alus* have been identified within the introns of protein-coding genes, where they may increase the likelihood of novel miRNA emergence; it has been suggested that transcription across repeat elements integrated in opposite orientations can result in the formation of a suitable miRNA precursor (Daniel et al., 2015).

3.3 Models for miRNA evolution within a clade

The miRNA complement in any given lineage is constantly in flux over geologic time, balanced out by the opposing forces of miRNA gene gain and gene loss. New precursors are continually arising in the genome (Lyu et al., 2014), supplying ample substrate on which evolutionary forces can act to shape the miRNA transcriptome. Existing miRNAs may also be secondarily lost, perhaps through gene deletion or through mutations which disrupt precursor secondary structure. The relative effect of these counteracting forces has been the subject of debate. Some authors have argued that miRNAs are continuously added to metazoan genomes over time, with very low rates of secondary loss (Peterson, Dietrich, & McPeck, 2009; Tarver et al., 2013; Wheeler et al., 2009). Others suggest that miRNA birth occurs in excess of miRNA death and that a large proportion of newly emergent miRNA genes is eventually lost due to deleterious or neutral regulatory effects (Lyu et al., 2014; Meunier et al., 2013).

It is clear from the literature that there is a general tendency within the Metazoa for the net gain of miRNA genes over evolutionary time (Berezikov, 2011; Heimberg, Sempere, Moy, Donoghue, & Peterson, 2008; Meunier et al., 2013; Wheeler et al., 2009). Studies done in mammals (Meunier et al., 2013) and flies (Berezikov et al., 2010) suggest a constant process of miRNA innovation through the identification of many young, lineage-specific miRNAs found on terminal branches. Most of these are

likely to be neutral or deleterious and will subsequently be eliminated through drift or negative selection. The occasional miRNA which provides a beneficial effect may be retained by positive selection and eventually become integrated into genetic regulatory pathways, following which secondary loss would become more difficult due to disruption of gene regulation. Claims that miRNAs are rarely lost from a lineage once acquired (Tarver et al., 2013; Wheeler et al., 2009) seem to be largely based on the analysis of highly conserved miRNA genes, which are likely to be heavily integrated into essential genetic regulatory pathways and hence unlikely to be secondarily lost. Thus, it is likely that the observed trend of miRNA gene gain over evolutionary time is due to a slight predominance of miRNA birth rates over miRNA death rates rather than very low rates of secondary loss.

Despite the general trend for net miRNA gain, conserved miRNA families can occasionally be lost in specific lineages, usually in association with morphological simplification. This has been documented in flatworms (Fromm, Worren, Hahn, Hovig, & Bachmann, 2013), which have undergone significant losses of conserved bilaterian, protostomian and lophotrochozoan miRNAs in tandem with secondary morphological simplification. The loss of conserved miRNA families is more pronounced in endoparasitic trematode and cestode lineages than in free-living flatworms. It is suggested that this is related not only to morphological simplification, but also to metabolic changes associated with endoparasitism.

The observed distribution of miRNA gains over the metazoan phylogeny suggests that the net gain of miRNAs over geologic time is not a uniform process. miRNA acquisition rates are known to accelerate along particular lineages, such as in therian mammals following the split from monotremes (Meunier et al., 2013). Several cases have also been documented where bursts of miRNA innovation occur, most notably on the bilaterian stem following divergence from Cnidaria (Guerra-Assunção & Enright, 2012; Prochnik, Rokhsar, & Aboobaker, 2007) and on the vertebrate stem (Heimberg et al., 2008). Such bursts would result from a relative increase in the rate of gain of miRNAs relative to their rate of loss. This could be achieved either by increasing the rate at which new miRNAs arise, by decreasing the rate at which they are purged from the genome, or both. An increased number of substrate precursors could hypothetically be attained through increasing general transcriptional activity in the genome (Heimberg et al., 2008). However, it is perhaps more likely that these observed bursts result from an increase in the rate at which new miRNAs that arise are fixed in the genome. Such could plausibly occur in the event that an internal or external event opens up new niches and relaxes constraint on the transcriptome, increasing the likelihood that a

newly emergent miRNA is able to confer some beneficial function and undergo positive selection whilst decreasing the likelihood of negative selection against new sequences.

4. The role of miRNAs in the evolution of development

4.1 miRNAs and morphological complexity

A broad perspective of miRNA evolution patterns across the Metazoa is suggestive that, in general, the size of the miRNA complement in a given lineage correlates with morphological complexity. Increases in the miRNA complement due to bursts of fixation of new miRNAs, such as those that occurred along the bilaterian stem lineage (Prochnik, Rokhsar, & Aboobaker, 2007) and in the vertebrates following divergence from the rest of the chordates (Heimberg, Sempere, Moy, Donoghue, & Peterson, 2008a), tend to correlate with advances in body complexity. Likewise, decreases in miRNA complement size due to losses of previously-conserved miRNAs have been linked to the simplification of body plans (Fromm, Worren, Hahn, Hovig, & Bachmann, 2013).

Although it is not possible to accurately compare the miRNA repertoires of all metazoans due to confounding factors such as differing sequencing depths (e.g. more extensive sampling of humans and model species) and the annotation of novel miRNAs with different levels of stringency and different criteria by different authors, it appears to be broadly true that bilaterians have many more miRNAs than non-bilaterians. The vast majority of conserved bilaterian miRNAs emerged following the bilaterian split from Cnidaria, and has been postulated to contribute to increased regulatory complexity which would support the increased number of cell and tissue types in the Bilateria (Christodoulou et al., 2010; Prochnik et al., 2007). Evidence for an ancestral role in the specification of tissue identities for miRNAs acquired on the bilaterian stem was presented by Christodoulou et al., 2010. This study identified 34 miRNAs conserved between protostomes and deuterostomes, of which many show conserved tissue-specific expression patterns in modern bilaterians. The oldest conserved animal miRNA is miR-100, which is also present in Cnidaria. In the cnidarian *Nematostella vectensis*, as well as the annelids *Platynereis dumerilii* and *Capitella*, and the sea urchin *Strongylocentrotus*, miR-100 is expressed in a specific population of foregut cells. This expression pattern is also common to the let-7 and miR-125 miRNAs

which are related to miR-100, but has expanded into other tissues in flies and vertebrates. Other sets of miRNAs showing conserved tissue-specific expression patterns include miR-9 and miR-131 in the brain, miR-29, miR-34 and miR-92 in protostome and deuterostome larval ciliary bands, miR-124 as a neuronal marker and miR-8 as a marker for differentiating sensory organs. The authors propose that the conserved bilaterian miRNAs evolved in a highly tissue-specific context, lending weight to the concept that miRNAs are key players in tissue specification and morphological complexity.

4.2 miRNA interactions in genetic regulatory circuits

miRNAs are an integral component of many of the genetic regulatory circuits which shape cell fate and tissue identity during development (Ebert & Sharp, 2012). The potential involvement of miRNAs in the regulatory circuitry governing development are summarized in Figure 1.8.

A miRNA involved in an incoherent feed-forward loop will typically show spatiotemporally overlapping domains of expression with its target. The miRNA is activated by the same transcription factors that activate its target, and consequently miRNA-mediated target repression opposes the action of the upstream system (Ebert & Sharp, 2012; Moran et al., 2014; Shkumatava, Stark, Sive, & Bartel, 2009). Incoherent feed-forward loops are likely to be involved where the miRNA plays a tuning or buffering role in modulating target expression (Ebert & Sharp, 2012; Peterson et al., 2009).

A coherent feed-forward loop has the miRNA activated by the same transcription factors which also repress expression of its target. In this scenario, the miRNA and its target often have near-mutually exclusive expression domains (Shkumatava et al., 2009). Coherent regulation has been documented in insects (Stark, Brennecke, Bushati, Russell, & Cohen, 2005), fish (Shkumatava et al., 2009) and mammals (Tsang, Zhu, & van Oudenaarden, 2007). It is suggested to be involved in maintaining cell and tissue identity, particularly in tissues derived from common progenitors. This is possibly achieved through the reinforcement of the transcriptional gene expression program through repression of leaky transcription (Ebert & Sharp, 2012). Coherent feed-forward motifs, along with positive and negative feedback loops (Figure 1.8), are

often used for lineage commitment (Ebert & Sharp, 2012). It is likely that coherent regulation is the prevalent means of miRNA-mediated regulation, at least in vertebrates, although incoherent regulation has also been documented (Shkumatava et al., 2009).

Motif	Schematic	Example functions
Coherent feed-forward loop		Lineage commitment Establishment and maintenance of tissue identity
Incoherent feed-forward loop		Tuning of gene expression levels

Figure 1.8

Regulatory motifs containing miRNAs. Ovals represent two hypothetical tissues. Dark yellow shading indicates high target expression; pale yellow shading indicates low target expression. Dots represent the miRNA expression domain.

4.3 miRNAs and the canalization of development

Robustness refers to the ability of a system to sustain its functions despite internal or external perturbations (Kitano, 2007). A biological system needs to be robust in order to function in unpredictable environments. Robustness is also closely tied in with evolvability (Pigliucci, 2008), as it allows for existent genetic diversity to be acted on by selection through the minimization of genetic noise (Ebert & Sharp, 2012). This reduces phenotypic variation and increases the precision of genotype to phenotype links, enabling selection on phenotypes to act more efficiently in shaping the genome (Peterson et al., 2009).

miRNAs reduce the amount of noise in the system through both coherent and incoherent regulation (Peterson et al., 2009). In a coherently regulated network, expression of the miRNA in domains where target gene expression is to be excluded may act as a fail-safe mechanism to prevent spurious mRNA transcription from giving rise to biologically relevant amounts of protein (Shkumatava et al., 2009). With

incoherent regulation, induction of both the miRNA and its target by similar upstream regulatory factors would allow for fine-tuning of target gene levels (Ebert & Sharp, 2012). It has been suggested that a greater role for miRNA-mediated regulation over geologic time in the Metazoa has allowed for increased canalization of development, reducing phenotypic variation while simultaneously increasing developmental precision and permitting the evolution of increasingly complex body plans (Peterson et al., 2009). The role of miRNAs in noise reduction may be especially important for transcripts that are lowly expressed. A combination of mathematical modelling and single-cell reporter assays were used to demonstrate that increased transcription coupled with miRNA regulation reduces protein expression noise in such transcripts (Schmiedel et al., 2015). This same study also proposed that combinatorial targeting of genes by more than one miRNA contributes to overall noise reduction.

4.4 The role of miRNAs in early embryogenesis

In order to fully appreciate the roles of miRNAs in shaping morphological complexity in animals, it is vital that we obtain an understanding of how they behave during embryonic development. It has been argued that the conservation of miRNA expression during development follows the 'hourglass' model for developmental evolution (Ninova, Ronshaugen, & Griffiths-Jones, 2014a), where early and late stages of embryonic development are more evolutionarily labile relative to a central conserved 'phylotypic' stage. This model is known to be adopted by conserved developmental regulatory genes across a wide variety of animals including molluscs (Paps, Xu, Zhang, & Holland, 2015), nematodes (Levin, Hashimshony, Wagner, & Yanai, 2012), flies (Kalinka et al., 2010) and vertebrates (Irie & Kuratani, 2011). Comparisons made between two drosophilid species (*Drosophila melanogaster* and *Drosophila virilis*, which diverged about 63 million years ago), reveal that miRNA expression profiles are well correlated between homologous developmental stages, and that mid-developmental stages such as gastrulation, germband elongation and germband shortening display the highest similarity in miRNA expression between both species (Ninova et al., 2014a).

Emerging evidence also hints that, at least in flies, the early embryo represents a permissive environment whereby newly emergent miRNAs can be expressed and exert their effects (Ninova et al., 2014b). Young, fast-evolving miRNAs were found to be

highly expressed in the early embryos of *D. melanogaster* and *D. virilis*. This was not true for later embryonic, larval or adult stages. The authors of this study have suggested that different developmental stages are not equally robust to change, and that early development is more buffered from functional consequences of change relative to later development. It may be the case that new miRNAs are presented with the opportunity to be selected for and integrated into genetic regulatory circuits within the permissive environment of the early embryo, following which such networks can be exapted to serving other functions beyond embryogenesis. However, we should also consider the possibility of the opposite scenario, where it is the abundant expression of new miRNAs which contributes to the robustness of early embryogenesis.

The specific functions of miRNAs during embryogenesis, along with the relative contributions of maternally inherited miRNAs compared with zygotically expressed miRNAs, have yet to be elucidated. Although an early study suggested that maternal miRNAs are essential for normal oocyte maturation and zygotic development in the mouse (Tang et al., 2007), this conclusion was based on observations that Dicer depletion in growing oocytes led to failure of embryogenesis to progress beyond the first cell division. However, Dicer is not unique to the miRNA pathway. Subsequent work has demonstrated that these effects may be due to the effects of Dicer loss on the processing of endogenous siRNAs instead (Suh et al., 2010), and that miRNA function may be suppressed during oogenesis in order for reprogramming to occur (Ma et al., 2010; Suh et al., 2010). miRNAs are likely to be required slightly later in embryogenesis at the maternal-to-zygotic transition, where the use of zygotic miRNAs to deplete specific maternal messenger RNAs has been observed in flies (Bushati, Stark, Brennecke, & Cohen, 2008) and zebrafish (Giraldez et al., 2006). It is currently unclear whether this is an example of convergent evolution or whether this phenomenon is conserved between insects and vertebrates as different miRNAs are involved in both species (Bushati et al., 2008). While there has yet to be any evidence presented, it may be possible that maternal and zygotic miRNAs perform reciprocal functions whereby maternal miRNAs suppress any leaky zygotic transcription prior to the maternal-to-zygotic transition, with zygotic miRNAs coming into play to clear up unwanted maternal transcripts after this point.

5. Project Aims and Contributions

The work presented in this thesis aims to further our understanding of how miRNAs contribute to evolution and development. miRNAs have the potential to shape animal body form and function not only through the modulation of developmental patterning pathways during embryogenesis, but also through their actions on post-embryonic development as well as ovarian function and oogenesis.

In Chapter 2, I perform a case study of the intronic miRNA gene, *mir-615*. This Hox cluster miRNA gene is located within the transcription unit of *Hoxc5*. Alignment and structural predictions suggest that the corresponding intronic region is only efficiently processed as a miRNA in eutherians; marsupials possess a sequence with high similarity but markedly different secondary structure that does not correspond to that of a pre-miRNA hairpin. No similar sequence is found outside the Mammalia. Using high-throughput sequencing data available in the public domain, I show that mature miR-615 is found wherever transcripts of its host gene are detected, indicating co-regulation of miRNA and host transcription. Additionally, miR-615 is also detected in human tissues and cell lines where *Hoxc5* is not known to be expressed, hinting at the possibility of independent regulation. Through the comparison of EST, chromatin and CAGE tag data, I predict the likely location of the miRNA-specific promoter within the *HOXC5* locus. I also present likely scenarios for the evolution of *mir-615* through intronic exaptation, and hypothesise on the role of the CpG island spanning the miRNA gene and its predicted promoter in the acquisition of independent regulation. Target prediction implicates miR-615 in the regulation of growth and developmental processes, and I speculate that this miRNA may be involved in the canalization of such processes during embryonic development.

A broader perspective on the evolution of miRNAs across a clade is presented in Chapter 3, with an emphasis on the acquisition of novel miRNAs over evolutionary time. In this chapter, I focus on the Lepidoptera (the insect clade consisting of butterflies and moths) and present a characterization of the miRNA transcriptomes via deep sequencing and comparative analysis in two species – the Speckled Wood butterfly *Pararge aegeria* and the Horse chestnut Leafminer moth *Cameraria ohridella*. This data was combined with existing miRNA and genomic resources within the Lepidoptera and its sister group, Trichoptera (caddisflies), in order to plot the likely point of origin of each miRNA acquired during the course of lepidopteran evolution. Two main findings result from this analysis, the first being an abundance of young, lineage-specific miRNAs on terminal branches of the lepidopteran phylogeny. The

second finding is that the majority of miRNA innovation retained across the course of lepidopteran evolution is concentrated at the base of the tree, suggesting that a burst of miRNA gain preceded radiation and diversification of the Lepidoptera.

The potential regulatory effects of one of these lepidopteran miRNA innovations, miR-2768, are functionally characterized in Chapter 4. Target prediction analyses suggest that miR-2768 binds with almost perfect complementarity to a target site in the coding sequence of the *cubitus interruptus (ci)* gene, which is known to be involved in several developmental patterning pathways including segmentation of the embryo, wing development and lepidopteran eyespot patterning. The observation that the *mir-2768* gene lies within an intron of *invected*, itself involved in similar pathways, suggests a possible role for this miRNA in complementing the biological function of its host gene. Using a cell culture-based luciferase assay system, I demonstrate that miR-2768 is capable of binding to and downregulating transcripts containing the candidate target site, regardless of whether this sequence is present in the 3' untranslated region or in a coding region.

Chapter 5 explores the ovarian miRNA transcriptomes in two separate populations of the Speckled Wood butterfly. Cross-comparison of the miRNA repertoires in both populations reveals a common set of 55 miRNAs which are expressed in both samples. I report that a higher proportion of conserved miRNAs are expressed in the ovary compared to that of lineage-specific miRNAs, and that there is no difference in expression levels between newly-emergent and conserved miRNAs. Seven miRNAs are either expressed at a considerably higher level in the ovary relative to the pooled miRNA transcriptome reported in Chapter 3, or only found in the ovary. Using a combined transcriptome assembly generated from the same RNA samples used for miRNA deep sequencing, I present target prediction data for these miRNAs, identifying those targets corresponding to a list of genes known to be involved in ovary and egg function as described in the literature (Carter et al., 2013). I also compare gene ontology (GO) term distribution between predicted targets of these miRNAs and the combined transcriptome as a whole, and discuss the potential roles of these miRNAs in conferring robustness to ovarian transcriptional networks.

The research presented here makes a significant contribution to our understanding of miRNA evolution and its consequences on the evolution of development, as well as to the miRNA-based regulation of both lepidopteran and mammalian reproduction.

Chapter 2:

miR-615 – A model for intronic miRNA evolution

Statement of Author Contributions

The majority of the material in this chapter has been written up and submitted to *EvoDevo* (submitted August 31, 2015; Manuscript ID EVOD-D-15-00018). It was accepted for publication on September 25, 2015. The authors on this manuscript are (in this order): Shan Quah and Peter W. H. Holland. The resultant paper is attached in Appendix C.

The original idea for the project was conceived by discussion between SQ and PWHH. Jerome H. L. Hui (The Chinese University of Hong Kong) provided scientific advice in the initial stages of the work. Tissue samples for monotreme and Xenarthran genomic DNA extraction were supplied by Stephen Donnellan (South Australian Museum) and Mads Bertelsen (Copenhagen Zoo) respectively. Mouse ovarian tissue was provided by Antonio Miranda (University of Oxford) and Shankar Srinivas (University of Oxford). SQ performed computational and experimental work and wrote the original draft manuscript. PWHH supervised the work of SQ throughout the length of this study. Both authors read and revised the manuscript.

1. Abstract

Introns are thought to represent a rich source of existing transcription for the evolution of novel miRNAs. Here I present a case study of an intronic miRNA gene, *mir-615*, located within the intron of the homeobox gene *Hoxc5*. Alignment and structural analysis show that this sequence is highly conserved across eutherian mammals and absent in non-mammalian tetrapods. Marsupials possess a similar sequence which is not predicted to be efficiently processed as a miRNA. Transcription of *HOXC5* in humans is accompanied by expression of *mir-615* in all observed cases. The miRNA can also be transcribed independently of its host gene, most likely through the use of an intragenic promoter. I present alternative scenarios for the evolution of *mir-615* through intronic exaptation and speculate on the secondary acquisition of independent transcriptional regulation. Target prediction analyses suggest that the dominant product of *mir-615* is involved in the regulation of growth and a variety of developmental processes. This hypothesised role for *mir-615* is consistent with its expression during mouse embryogenesis.

2. Introduction

Intronic miRNAs may give insights into *de novo* miRNA innovation and its contribution to the evolution of development. Pre-existing transcribed RNAs, particularly in the form of introns, are thought to be a major source from which new miRNAs evolve (Campo-Paysaa et al., 2011; Meunier et al., 2013). Here I present a case study of *mir-615*, a mammalian intronic miRNA located within a Hox cluster. I describe the regulation, expression pattern and likely biological functions of *mir-615*, and use these to draw hypotheses on its role in the context of eutherian evolution and development.

Homeobox genes form a large, diverse superclass with a wide range of roles in metazoan developmental biology. Most encode transcription factors which exert regulation over the expression of other genes during embryogenesis, thereby having a direct influence on the resultant adult morphology (Holland, 2012). Perhaps the best characterized of the homeobox genes are the Hox genes, whose primary roles are in the specification of anteroposterior patterning in the developing bilaterian embryo. Hox genes also have other important functions during animal morphogenesis such as patterning of the limbs (Tabin, 1992), embryonic kidney (Patterson, Pembaur, & Potter, 2001; Taguchi et al., 2014) and reproductive tract (Taylor, Vanden Heuvel, & Igarashi, 1997). Metazoan Hox clusters contain miRNA genes of varying ages and evolutionary origin. The *mir-10* gene is conserved between arthropods and mammals and occurs in the same position in both insect and vertebrate Hox clusters (Tanzer, Amemiya, Kim, & Stadler, 2005). The *mir-196* and the bidirectionally transcribed *mir-iab-4* genes are chordate-specific (Candiani, 2012) and arthropod-specific (Hui et al., 2013) respectively. Another Hox cluster miRNA gene, *mir-615*, is less well-characterized and reported to be restricted to either mammals or amniotes (Woltering & Durston, 2008). A detailed analysis of the phylogenetic distribution of *mir-615* is currently lacking. Unlike *mir-10*, *mir-196* and *mir-iab-4*, which are intergenic, *mir-615* is nested within an intron of the *Hoxc5* gene in human and mouse. This positioning makes *mir-615* an ideal candidate for study as Hox genes are highly conserved, facilitating alignment of the intronic region through alignments of the flanking exons.

As *mir-615* occurs in the same orientation as *HoxC5*, it is likely to be co-transcribed with its host gene and may complement host protein function. Studying the transcriptional regulation and expression of *mir-615*, particularly in relation to that of *Hoxc5*, would contribute towards our understanding of the relationship between the

miRNA and its host gene. The development and improvement of next-generation sequencing (NGS) technologies over the past decade has greatly facilitated high-throughput analysis of genomic and transcriptomic data at relatively low cost (Metzker, 2010). Projects such as the Encyclopedia of DNA Elements (ENCODE) and the Human Protein Atlas (Uhlen et al., 2015) have provided large amounts of data in the public domain which can be harnessed to answer questions regarding the expression and regulation of *Hoxc5* and *mir-615*.

Using sequence data, both derived from the public domain as well as obtained through my own sequencing efforts, I investigate the phylogenetic distribution of *mir-615* across tetrapods. I also perform structural predictions and evaluate minimum folding energies of the corresponding region in the *Hoxc5* intron to evaluate whether candidate pre-miR-615 transcripts are efficiently processed into mature miRNAs, and report that this is likely to occur only in eutherian mammals. I also analyse publicly released RNA-seq data in combination with published studies on chromatin modification and expressed sequence tag (EST) studies to argue for the existence of a promoter specific to *mir-615* located within the transcription unit of *HOXC5*. Finally, I carry out target predictions for the dominant 3p product of miR-615 and cross-compare these to its expression domains in order to work out likely *in vivo* functions of this miRNA.

3. Materials and Methods

3.1 Conservation analysis for *mir-615*

Eutherian *Hoxc5* intronic sequences, along with those from *Xenopus tropicalis* and *Anolis carolinensis* were obtained from the UCSC Genome Browser (Kent et al., 2002) and the Ensembl genome database (Flicek et al., 2012). The *Sarcophilus harrisii* sequence was obtained from Ensembl while the *Macropus eugenii* sequence data used was obtained from GenBank (accession: JN378720.1). Monotreme DNA tested was obtained from Stephen Donnellan in the South Australian Museum via the Oxford University Museum of Natural History (CITES registration number: GB026). Three species were sampled: Platypus (*Ornithorhynchus anatinus*), Short-beaked echidna (*Tachyglossus aculeatus*) and Long-beaked echidna (*Zaglossus bruijnii*). Xenarthran samples were provided by Mads Bertelsen at the Copenhagen Zoo. Xenarthran genomic DNA was obtained from sloth (*Choeloeopus didactylus*) heart and anteater (*Myrmecophaga tridactyla*) liver by phenol-chloroform extraction. Data for *Ochotona princeps* (American pika), *Jaculus jaculus* (lesser Egyptian jerboa) and *Sorex araneus* (common shrew) *mir-615* was obtained by running BLASTn using the hsa-miR-615 sequence against draft genome assemblies for these species. BLAST parameters used are: word size 4, reward 5, penalty -4, gapopen 8 and gapextend 6. The contigs on which the predicted *Hoxc5* sequences were verified to contain the flanking *Hoxc5* exons by sequence alignment. Alignment was carried out using MAFFT (Katoh, 2002).

Primer sequences used for amplification of all *Hoxc5* intronic sequences are: 5'-TTGGACTTAAGCATCACTTTCCCACCG-3' and 5'-CCAGAGTCTGGTAGCGCGTGTAAGTGG-3'. These primers were designed to bind a region in the flanking CDS exons which is conserved between most tetrapods (Supplement S2.1). The annealing temperature used was 60°C for anteater and 55°C for sloth. PCR products were gel purified using the illustra GFX PCR and DNA Gel Band Purification Kit (GE Healthcare, 28-9034-71). The sloth sequence was cloned into the pGEM-T Easy Vector System (Promega) and sequenced using the M13F and M13R primers. The anteater amplicon was directly sequenced following gel purification with the same primers used for amplification. Forward and reverse sequencing reactions were performed using the BigDye Terminator v3.1 Cycle Sequencing Kit. 25ng of template was used with 2 µl of sequencing buffer, 0.5 µl of Big Dye Mix and 1 µl of primers. To overcome secondary structure in the template, samples were heated

at 98 °C for 10 minutes prior to addition of Big Dye Mix. Samples were sent to the Department of Zoology sequencing facility for electrophoresis.

Structural criteria for the annotation of functional miR-615 are as follows: stable hairpin structure with < -18 kcal/mol free energy, at least 18 paired bases on the main stem and the absence of large internal loops and bulges in mature regions. Minimum free energy and predicted secondary structure was calculated for each putative miR-615 sequence using RNAeval.

3.2 Bioinformatics and target prediction

Data used to determine expression of *mir-615* and *HOXC5* in human cell lines was obtained from the ENCODE project and visualised on the UCSC Genome Browser (<http://genome.ucsc.edu/ENCODE/>).

A list of publicly released datasets used in the bioinformatic analysis of *mir-615* and mRNA expression is provided in Table 2.1. These data were obtained from the Gene Expression Omnibus (Barrett et al., 2013) and the ENCODE project. In the microarray studies, a gene was determined to be expressed based on the ABS_CALL value (P = present, A = absent, if provided). If this was not provided, a cutoff p-value of <0.05 was used to determine expression. Raw data for the transcriptome assemblies was quality checked using FastQC, followed by trimming of the first 15 bases using the fastqtrim.py script (Appendix A1). Trimmed reads were assembled using Trinity and resultant contig FPKMs determined with RSEM. BLASTn with standard parameters was used to determine Hox gene expression in each transcriptome, with appropriate sequences from the corresponding species obtained from GenBank. For the pig, human *Hoxc* cDNAs were used to query the transcriptome reported in Nie et al., (2011). Small RNAseq data was put through the miRDeep2 pipeline using the quantifier.pl script to map reads onto known human miRNAs. Handling of raw reads prior to mapping was carried out as described in Chapter 3. Read counts for mouse developmental stages as presented by Chiang et al. (2010) were obtained from miRBase (Kozomara & Griffiths-Jones, 2011).

Target prediction for hsa-miR-615 was run using the downloadable executable for the PITA prediction algorithm (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007) on all

human cDNAs downloaded from Ensembl BioMart (genome assembly: GRCh38.p2). Predicted targets were ranked by score and a cutoff score of -20 was applied. Documented functions for predicted targets were obtained from UniProtKB/Swiss-Prot (<http://www.uniprot.org/>, The UniProt Consortium, 2014). RPKM values for tissue-specific expression of each gene were obtained from the Human Protein Atlas (Uhlén et al., 2015), except for cerebellum where RPKM values were obtained from gene quantifications available through the ENCODE project (accession: ENCFF917ANZ).

Accession number	Tissue/Cell line	Species	Type
ENCFF917ANZ	Cerebellum (foetal, 37 weeks and 19 weeks)	<i>Homo sapiens</i>	RNA-seq
ENCSR046XHI	Ovary	<i>Homo sapiens</i>	RNA-seq
ENCSR516UNF	Ovary (age = 10 weeks)	<i>Mus musculus</i>	RNA-seq
GSM874239	Ovary (adult)	<i>Mus musculus</i>	Microarray
GSM874241	Ovary (neonatal)	<i>Mus musculus</i>	Microarray
GSM35553	Ovary (commercial RNA)	<i>Mus musculus</i>	Microarray
GSE22776	Ovarian follicles	<i>Macaca mulatta</i>	Microarray
ENCSR109IQO	K562	<i>Homo sapiens</i>	RNA-seq
ENCSR000AET	K562	<i>Homo sapiens</i>	Small RNA-seq
ENCSR000AJG	GM12878	<i>Homo sapiens</i>	Small RNA-seq

Table 2.1

Datasets used in analysis of *mir-615* and *Hoxc5* expression

3.3 *mir-615* promoter prediction

Histone methylation, DNase I and CAGE data used to find the position of the *mir-615* promoter were obtained using ENCODE data tracks available through the UCSC Genome Browser (<http://genome.ucsc.edu/ENCODE/>). The EST data track ('Human ESTs Including Unspliced') was generated at UCSC from publicly available EST sequencing data.

3.4 Mouse ovary RT-PCR

Ovaries from 6 week old adult mice were donated by Shankar Srinivas for three different physiological states – unmated, mated but non-pregnant and pregnant. RNA was obtained by TRIzol (Life Technologies, #15596-026) extraction, followed by DNase I digestion using the TURBO DNA-free kit (Life Technologies, #AM1907) to remove

contaminating genomic DNA. The RNA was checked for quality and lack of genomic DNA contamination by running 1 µl on a gel under denaturing conditions. Reverse transcription was performed using the SuperScript III First Strand Synthesis System (Life Technologies, #18080051) and an oligo(dT) primer. Primers used to validate ovarian *Hox* gene expression were validated on genomic DNA before being used on ovarian cDNA. Primer sequences are provided in Table 2.2.

Oligo name	Sequence
mHoxc5_RT_F	ACATGCAAACCTGTGGGAACATG
mHoxc5_RT_R	AGTGAGGTAGCGTTAAAATGGAA
mHoxc6_RT_F	ACCAGAAAGCCAGTATCCAGATTTA
mHoxc6_RT_R	CTCTTCTTCTGTCTCTTCTCGCTT
mHoxc8_RT_F	AATATCCCGACTGTAAATCCTCCG
mHoxc8_RT_R	ATTCCCTTCTTCTCCACCTTCTC
mHoxc10_RT_F	CAGTCCCAATGAGATCAAGACAGA
mHoxc10_RT_R	AATTCTTTCTCCAATTCCAGCGTC
mHoxc9_RT_F2	CATGCGGACTTGGCTCGAGCCGCTGTCCGG
mHoxc9_RT_R2	TCCTTCTCCAGTTCAGCGTCTGGTACTTGG

Table 2.2

Primers used to test ovarian *Hoxc* gene expression

4. Results

4.1 A sequence with high similarity to *mir-615* is present in therian mammals

miR-615 is reported as being restricted to either mammals or amniotes (Woltering & Durston, 2008) but current literature lacks a detailed analysis of its evolutionary history. I sought to characterise its phylogenetic distribution in order to determine its most likely time of origin.

Available entries for *mir-615* on miRBase (Kozomara & Griffiths-Jones, 2011) are limited to two of the four main clades of eutherian mammals (Laurasiatheria and Euarchontoglires). In order to determine whether or not miR-615 was present in other amniotes, I searched for its precursor gene sequence in the genomes of a marsupial (*Monodelphis domestica*), a monotreme (*Ornithorhynchus anatinus*) and three non-mammalian vertebrates (the chicken *Gallus gallus*, the anole lizard *Anolis carolinensis* and the zebrafish *Danio rerio*). None of these species produced convincing hits, while a positive control BLASTn run using the same parameters successfully recovered *mir-615* in mouse (*Mus musculus*). Within the Mammalia, no miRBase entries exist for the eutherian clades Afrotheria and Xenarthra, or for any non-eutherian lineages. Although I was able to identify the *mir-615* gene by BLASTn in the elephant (*Loxodonta africana*; Afrotheria), this gene was not found in the genome assembly for a Xenarthran - the sloth (*Choloepus hoffmanni*).

The syntenic position of *mir-615* within a highly conserved Hox gene allows for direct comparisons of the corresponding region to be made between vertebrate species. In human (*Homo sapiens*) and mouse, *mir-615* is located within the intron of *Hoxc5*. I performed alignments of this intron and its flanking coding exons across the four main eutherian lineages, non-eutherian mammals and other tetrapods to check for conservation of the syntenic position of *mir-615* (Figure 2.1). As no sequence data for *Hoxc5* were available for Xenarthra, I sequenced the corresponding region from sloth (*Choloepus didactylus*) and anteater (*Myrmecophaga tridactyla*) using primers designed against highly conserved regions within vertebrate *Hoxc5* coding regions (Supplement S2.1). The alignment between the two Xenarthran sequences and the other eutherian *mir-615* loci is strongly suggestive of a functional *mir-615* gene in the Xenarthra. In marsupials, both the Tasmanian devil (*Sarcophilus harrisii*) and the tammar wallaby (*Macropus eugenii*) possess a sequence with strong similarity to

eutherian *mir-615* within *Hoxc5* (Figure 2.2). However, the regions corresponding to the mature miR-615 products differ slightly from the corresponding eutherian versions. Neither *mir-615* nor any of the flanking exonic sequences of *Hoxc5* are present in the genome assembly for the opossum (*Monodelphis domestica*). I was unable to amplify the *Hoxc5* gene from any monotremes with the same primers. This genomic region is also missing in the platypus (*Ornithorhynchus anatinus*) genome assembly.

These alignment data suggest that *mir-615* is conserved at nucleotide sequence level within the Eutheria, and confirm the lack of sequence similarity between mammals and non-mammalian tetrapods (Figure 2.1).

4.2 Candidate miR-615 precursors fulfil structural and energetic criteria only in Eutheria

Having established that all therian mammals (eutherians and marsupials) studied possess a sequence very similar to the annotated *mir-615* genes of human and mouse (Figure 2.1), I tested these sequences for fulfilment of published structural and energetic criteria used in miRNA annotation. This analysis reveals that all the eutherian *mir-615* orthologues are likely to produce functional pre-miR-615 (Table 2.3). However, in the two marsupials for which sequence data is available, minimum folding energy (MFE) for the putative pre-miR-615 homologues (-39.40 kcal/mol in *M. eugenii* and -42.60 kcal/mol in *S. harrisi*) are higher than those observed for all eutherian miR-615 precursors (mean MFE -59.40 kcal/mol). Both marsupial sequences also have large internal loops, bulges and mismatches which are likely to preclude efficient entry into the miRNA pathway (Figure 2.3). These predicted structures differ between *M. eugenii* and *S. harrisi*, suggesting a lack of selection on the secondary structure of this region in marsupials. It is therefore likely that marsupials lack the means to generate mature miR-615.

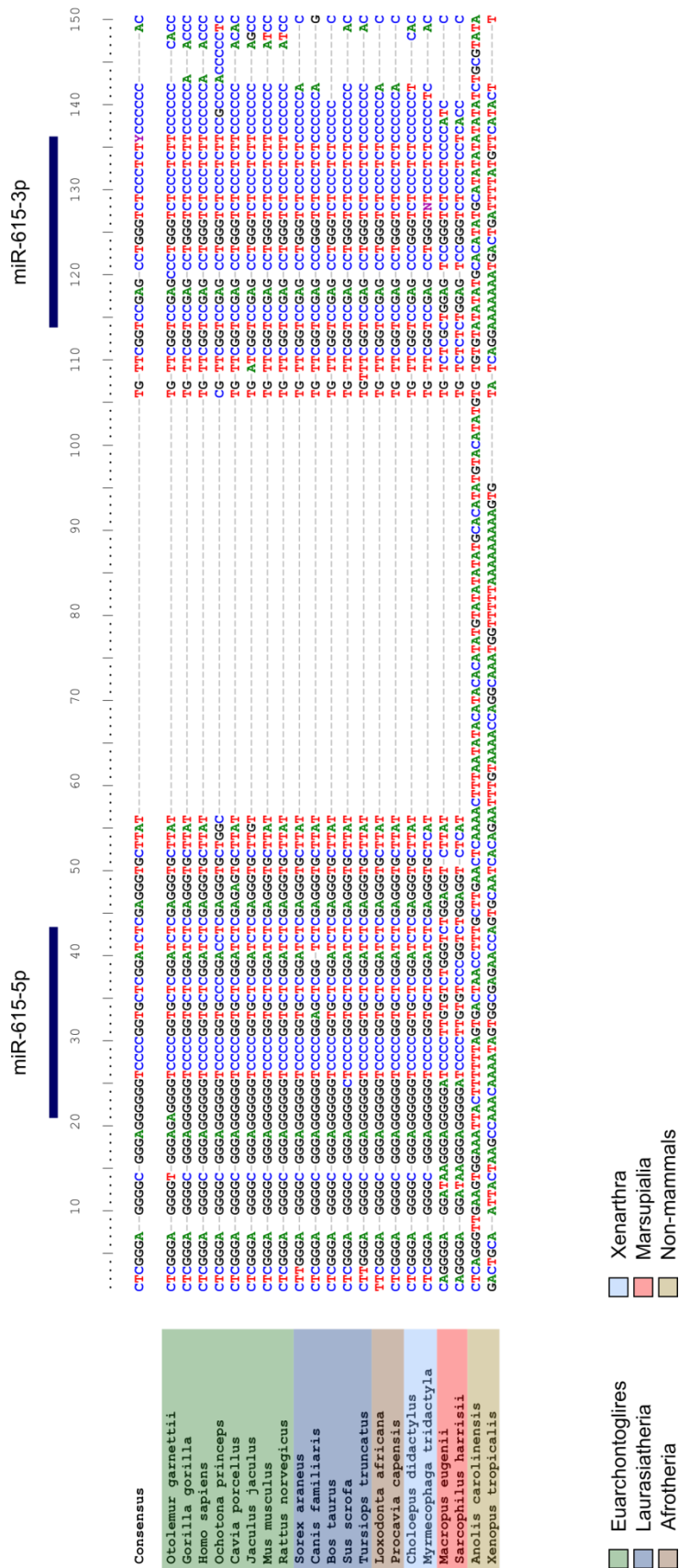


Figure 2.1
Section of *Hoxc5* intronic alignments corresponding to *mir-615* across tetrapods

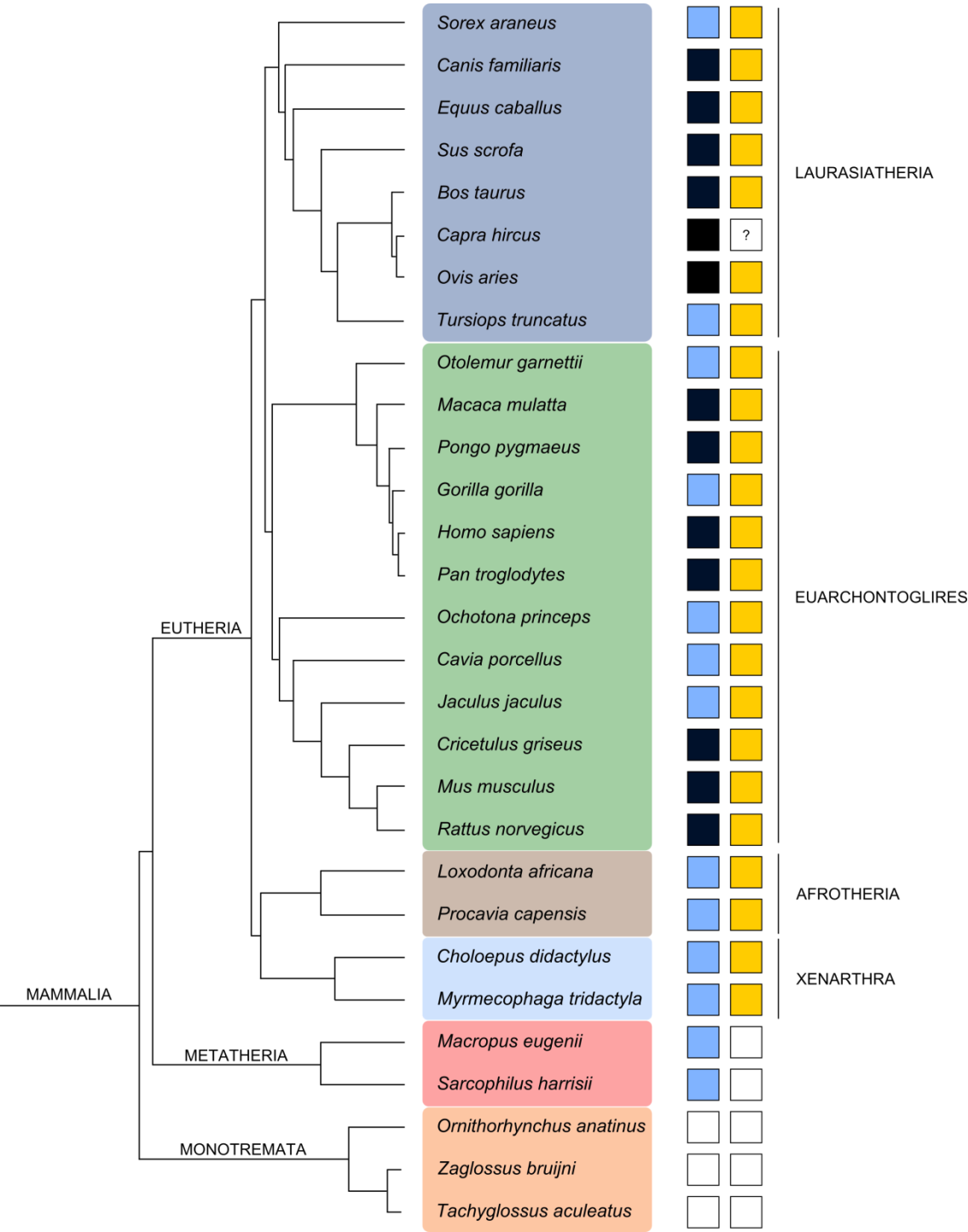
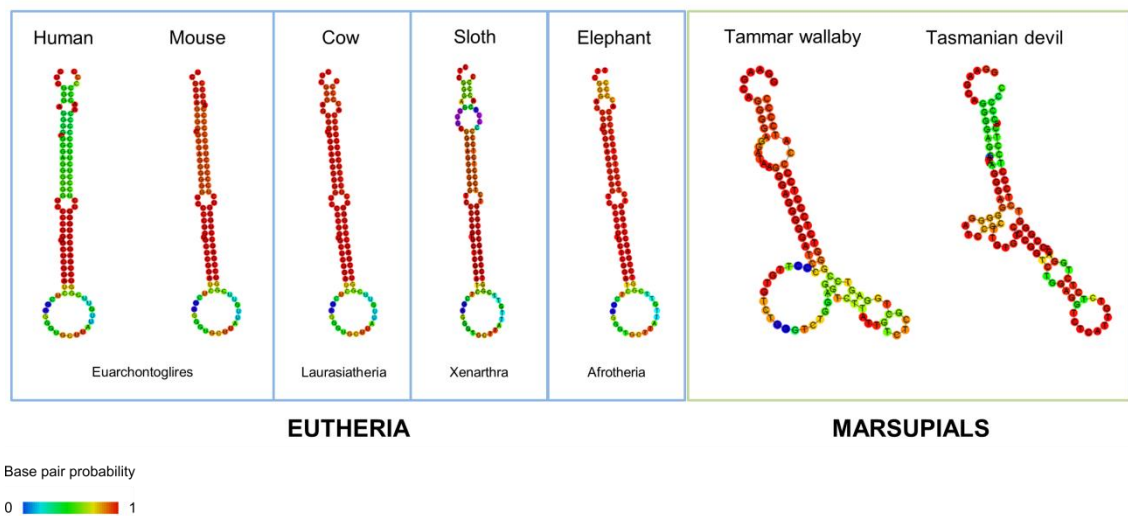


Figure 2.2

Distribution of *mir-615* in mammals. Black squares represent miRBase entries or published deep sequencing data for *mir-615*. Blue squares indicate BLASTn predictions from genome data or PCR products obtained in this study. Yellow squares indicate strong secondary structure predictions. No structural prediction data exists for *Capra hircus* as I was unable to identify the *mir-615* precursor by BLASTn in the corresponding genome assembly.

**Figure 2.3**

Predicted secondary structures for miR-615 precursors in marsupials and the four major eutherian clades.

[illegible]

Table 2.3

Predicted secondary structures and minimum free energies for the region corresponding to *mir-615* across therian mammals.

4.3 miR-615 expression in mammalian cell lines and adult tissues

Using small RNA deep sequencing data from the ENCODE project (Katalin et al., 2009) obtained through the UCSC Genome Browser (Kent et al., 2002), I analysed miR-615 expression in 36 human cell lines and tissues (Supplements S2.2 and S2.3). miR-615 is expressed in the majority of these datasets. A minority of cell lines lacked expression of miR-615, notably those with immune-related functions (CD14+, CD34+ and CD20+) as well as H1 embryonic stem cells and H1-neurons. There is no evidence for antisense transcription of miR-615 in any of the samples. In all cases where miR-615 is expressed, the miR-615-3p product is dominant.

Since the above browser included only ENCODE data released prior to 2013, I extended these analyses using available raw sequence reads from more recent small RNA deep sequencing projects available through the ENCODE Project Portal using the miRDeep2 quantifier module. I first confirmed that this approach gives comparable results to the UCSC browser information by analysing the raw sequence reads from the human K562 and GM12878 cell lines (Supplement S2.4).

While there are relatively few tissue-specific small RNA sequencing studies available through the ENCODE Consortium, some additional data are available for select tissues and developmental stages in published literature. I studied the expression of miR-615 in published data for five organs (cerebellum, cerebral cortex, heart, testis and kidney) in human and mouse (Meunier et al., 2013), as well as ovary in goats (Zhang et al., 2013) and sheep (Torley et al., 2011).

Species	Normalized miRNA read counts							
	Cerebellum	Brain	Testis	Heart	Kidney	Sertoli	Spermatocyte	Spermatid
<i>H. sapiens</i>	1.83	0	7	0	12.8	-	-	-
<i>M. musculus</i>	1.50	0	0	0	8.9	100.1	2	0

Table 2.4

Data from Zhang et al. (2013) showing normalized tissue-specific read counts for the dominant product of miR-615 (miR-615-3p). Sertoli cells, spermatocytes and spermatids were only sampled in mouse.

These data indicate that miR-615 is expressed in kidneys, cerebellum and in both testis and ovary (Tables 2.4 and 2.5). Although miR-615 was not detected in whole mouse testis, it is present in the Sertoli cells and spermatocytes (Meunier et al., 2013, Table 2.4). In the adult goat ovary, expression of both mature products of *mir-615* is increased during pregnancy (Zhang et al., 2013, Table 2.5).

miRNA	Ovary (pregnant)	Ovary (non-pregnant)
miR-615-3p	406	3
miR-615-5p	587	8

Table 2.5

Read counts for both mature products of miR-615 in pooled ovarian sample from six pregnant and six non-pregnant goats. The dataset used for this analysis was obtained from Zhang et al. (2013)

My analysis of existing small RNA sequencing data in the mouse (data obtained from Chiang et al., 2010) reveals that miR-615 expression is also detected in whole mouse embryos (Figure 2.4). Expression of the miRNA appears to be variable between replicates at each developmental stage, but displays a general trend whereby it increases throughout embryogenesis and persists in the newborn.

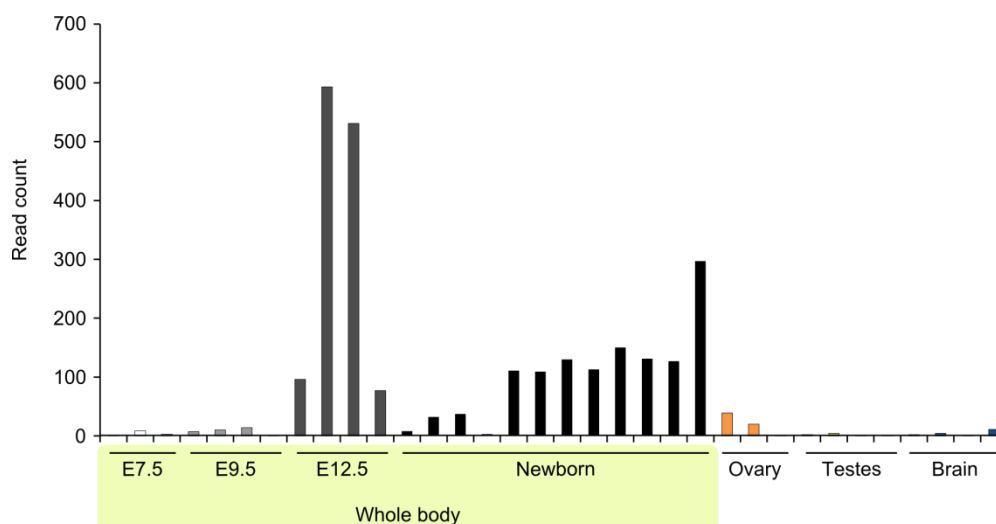


Figure 2.4

miR-615-3p read counts in mouse small RNA libraries as reported in the dataset of Chiang et al. (2010). Data was obtained through miRBase.

4.4 Expression of miR-615 can occur independently from Hoxc5

Comparing the expression patterns of *mir-615* and those of *Hoxc5*, would facilitate an understanding of whether or not expression of *mir-615* is solely driven by its host transcription unit, or whether there is independent regulation of this intronic miRNA. Out of the 36 human cell lines and tissue types used to survey miR-615 expression, 34 also had available RNA-seq data (Supplement S2.2), which I used to study expression of *HOXC5*. Examples of the comparison between miR-615 expression and expression of *HOXC5* are shown in Figure 2.5. A full comparison of *mir-615* and *HOXC5* is provided in Supplement S2.5. I found evidence for simultaneous expression of both miR-615 and *HOXC5* in several datasets, such as hMSC-AT. Some cell lines, such as H1 embryonic stem cells, express neither the miRNA nor its host gene. However, it is notable that in many of the cell lines where miR-615 is found, including GM12878 and K562, there is little to no expression of *Hoxc5*. To validate the observation that *HOXC5* is not expressed in K562 cells, I constructed a transcriptome assembly for this cell line using raw data from the ENCODE Consortium (Dunham et al., 2012) and did not find any evidence for *HOXC5* expression (Table 2.6). There are no instances of *HOXC5* expression without concurrent expression of miR-615 in all the cell lines surveyed.

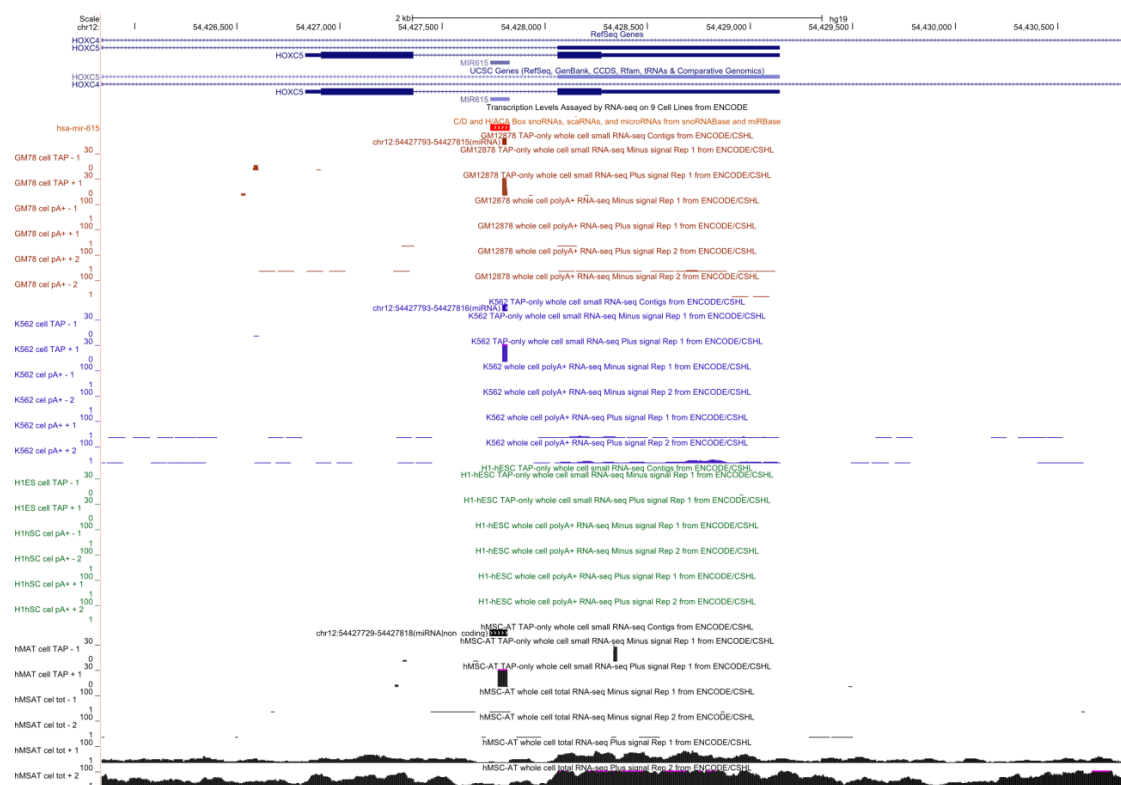


Figure 2.5

Expression of *mir-615* and *Hoxc5* in example human cell lines

An analysis of the chromatin landscape surrounding *mir-615* in ENCODE human cell lines provides evidence for the existence and location of a miRNA-specific promoter (Figure 2.6). Trimethylation of histone H3 on lysine 4 (H3K4me3), a marker of active chromatin associated with transcriptional start sites (TSSs) is distributed throughout the first exon and the intron of *HOXC5* in cell lines that express *mir-615*. Cell lines which express neither *HOXC5* nor *mir-615*, represented by H1-hESC and AG044540, have minimal H3K4 trimethylation at the *HOXC5* locus. DNase I hypersensitivity data also provides a measure of chromatin accessibility and is concentrated in two regions, one ~150 bp upstream of the annotated TSS for *HOXC5*, and the other distributed more broadly across the first exon and the 5' end of the intron. DNase I hypersensitivity sites are also seen at the 3' end of the intron and in the second exon in NHDF cells only. 5' cap analysis of gene expression (CAGE) data identifies potential TSSs within the *HOXC5* locus. These sites map to both the known *HOXC5* TSS and a region about ~200bp upstream of the exon-intron boundary which corresponds well with the 5' ends of several expressed sequence tags (ESTs). They also align with the DNase I hypersensitivity sites. Collectively, these analyses of chromatin accessibility and transcription at the *HOXC5* locus, coupled with evidence for expression of *mir-615* in the absence of *HOXC5*, are strongly suggestive of the existence of an internal miRNA-specific promoter.

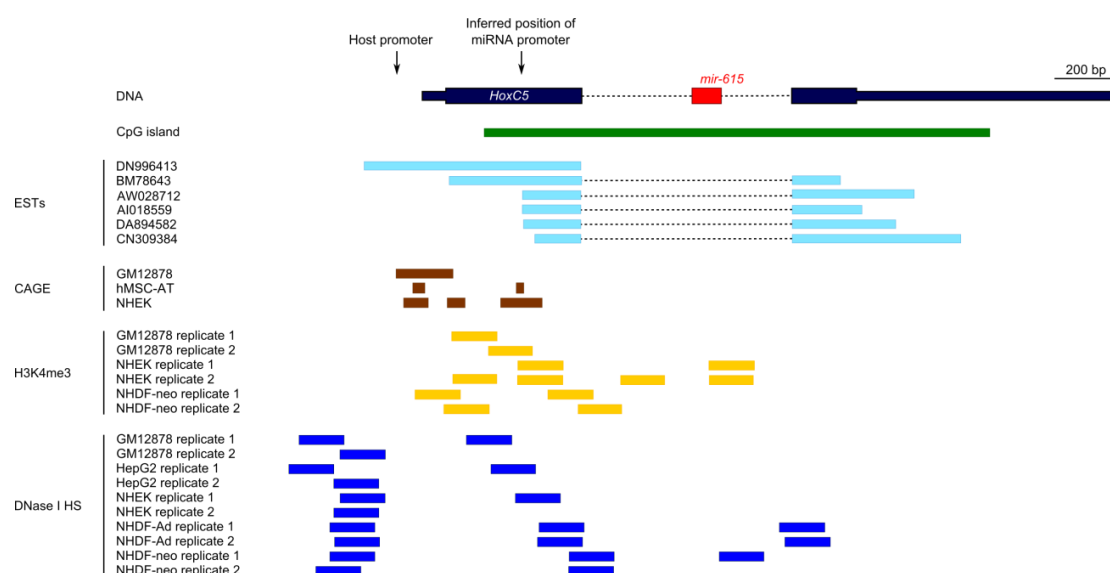


Figure 2.6

Inferred position of the *mir-615* promoter based on EST, CAGE, H3K4 trimethylation and DNase I hypersensitivity data.

<i>HOXC4</i>	<i>HOXC5</i>	<i>HOXC6</i>	<i>HOXC7</i>	<i>HOXC8</i>	<i>HOXC10</i>	<i>HOXC11</i>	<i>HOXC12</i>	<i>HOXC13</i>
1.08	0	0.73	0.72	4.19	0	0	0	0

Table 2.6

RPKM values for *HOXC* genes in a K562 transcriptome assembly

4.5 Ovarian expression of *mir-615* and *HoxC* genes

Utilising the observation that *HOXC5* expression in human cell lines is a good indicator of *mir-615* expression, I looked for known expression of *HOXC5* in human organs and tissues. This would allow for identification of potential *mir-615* expression in tissues where small RNA sequencing data is not available. Data from the Human Protein Atlas (Uhlén et al., 2015) suggests that *HOXC5* is expressed at a low level in both male and female gonads as well as fallopian tubes, stomach, smooth muscle and adrenal glands, but not in any of the other tissues known to express miR-615.

To validate the expression of *HOXC5* in the human ovary and obtain a better understanding of gene expression across the cluster, I used raw data from the ENCODE Consortium (Dunham et al., 2012) to construct an ovarian transcriptome assembly. I searched this for *HOXC5* and other *HoxC* genes (Table 2.7). This analysis confirms that *HOXC5* is expressed in the human ovary. Most of the *HoxC* cluster appears to be accessible to transcription to some degree except for the posterior genes (*HoxC11*, *HoxC12* and *HoxC13*).

Species	Description	<i>Hoxc4</i>	<i>Hoxc5</i>	<i>Hoxc6</i>	<i>Hoxc8</i>	<i>Hoxc9</i>	<i>Hoxc10</i>	<i>Hoxc11</i>	<i>Hoxc12</i>	<i>Hoxc13</i>
<i>Mus musculus</i>	Ovary, age = 10 weeks.	13.8	5.05	78.56	32.38	5.1	14.9	0	0	0
<i>Homo sapiens</i>	Ovary, age = 47 years	8.78	131.88	19.01	12.12	7.06	7.09	0	0	0

Table 2.7

RPKM values as determined by RSEM on ovarian transcriptome assemblies for *M. musculus* and *H. sapiens*. Raw data used was obtained from the ENCODE project.

The conservation of miR-615 across the Eutheria is strongly suggestive of adaptive function. Identification of a conserved tissue-specific expression pattern would provide further evidence for putative function. *mir-615* is co-expressed with *Hoxc5* in the mammalian ovary, and upregulation of *mir-615* expression occurs during pregnancy in goats (Zhang et al., 2013). This is suggestive of biological function. I screened for expression of *Hoxc* genes in mouse ovaries by PCR (Figure 2.7). Three replicates were sampled for both pregnant and non-pregnant females to examine any potential effect of pregnancy on ovarian *Hoxc* gene regulation. I also sampled virgin females of the same age to control for any effects of pseudopregnancy. These data suggests that there are no striking differences between *Hoxc* gene expression in the ovaries of pregnant females, non-pregnant mated females and virgin females. *Hoxc6* and *Hoxc8* are expressed in the ovaries while expression of *Hoxc5*, *Hoxc9* and *Hoxc10* is very low or undetectable in most of the samples. No bands were detected for *Hoxc4*, *Hoxc11*, *Hoxc12* and *Hoxc13*, but I am unable to conclusively state that these genes are not expressed in the ovary as the same primers did not amplify their corresponding targets in mouse genomic DNA. Available microarray (Kogo et al., 2010) and RNA-seq data for mouse ovaries (Table 2.7) support the PCR data in demonstrating that *Hoxc6* and *Hoxc8* are expressed in adult ovaries, and also indicate that transcripts of *Hoxc4* and *Hoxc5* are present in the ovarian transcriptome. Of the posteriorly-located *Hoxc* genes, *Hoxc9* was only detected in one of the three microarray experiments, *Hoxc10* was present in all three, and *Hoxc13* is not expressed. The microarrays did not include probes for *Hoxc11* and *Hoxc12*.

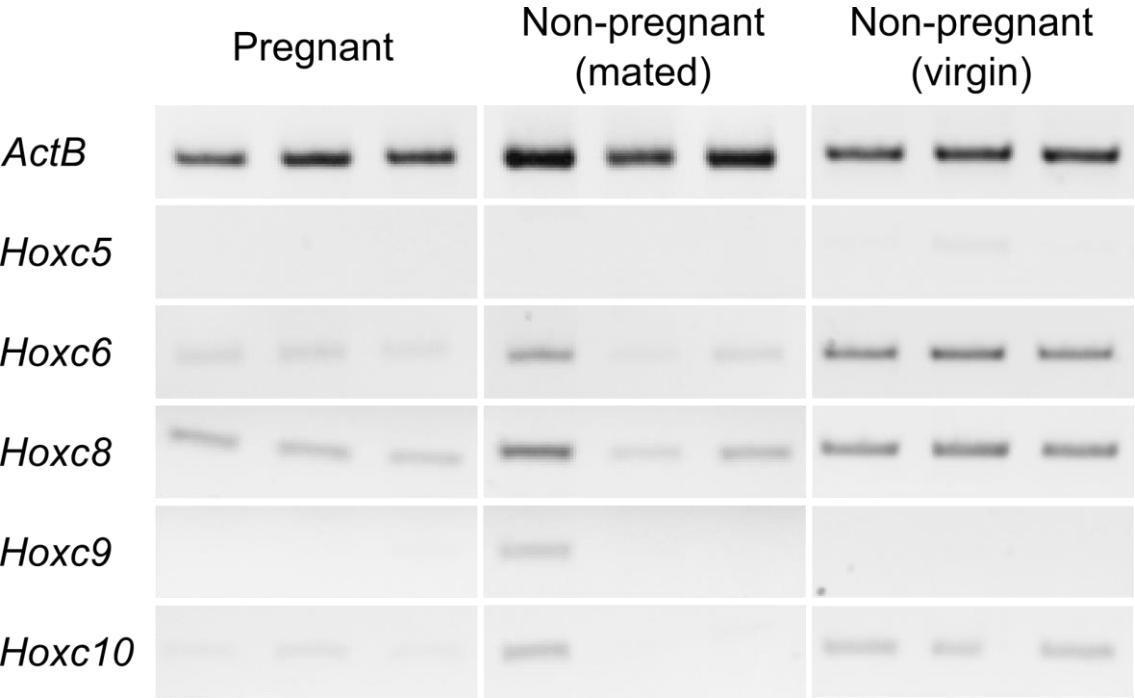


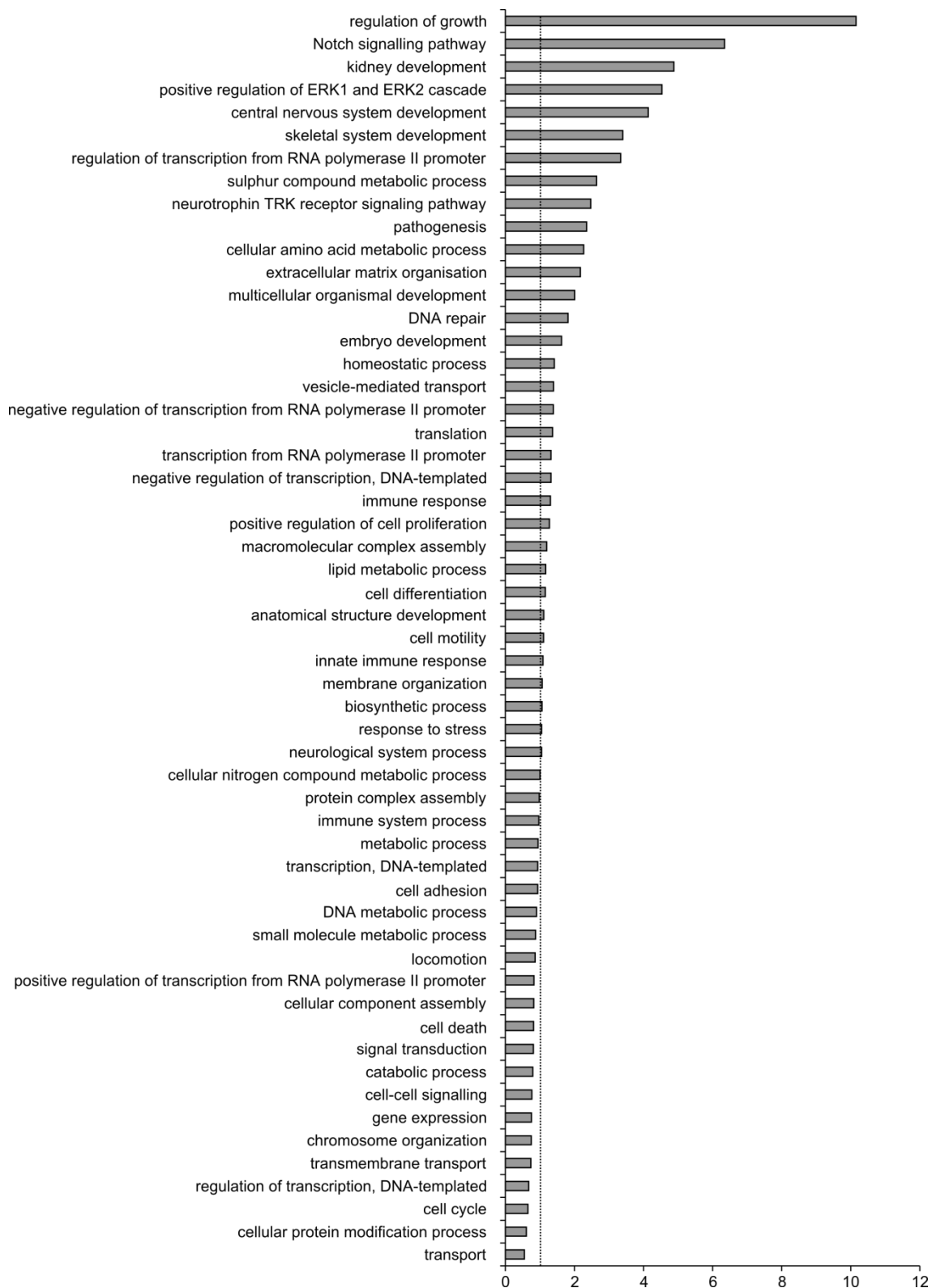
Figure 2.7
RT-PCR on mouse ovarian cDNAs. *ActB* is included as a positive control. Three biological replicates were performed for each physiological state.

As similar experimental testing of ovarian *Hoxc* expression by PCR is difficult in mammalian species apart from mice due to ease of access to material, I looked into publicly available datasets for this data (Supplements S2.5). A study using controlled ovulation in *Macaca mulatta* (Xu et al., 2011), provides a microarray-based analysis of ovarian gene expression in response to exposure to an ovulatory stimulus (exogenously administered chorionic gonadotropin). Although probes for *Hoxc5* and *Hoxc6* are missing from this dataset, the general trend for *Hoxc* gene expression in the rhesus ovary appears to be broadly similar to that observed in human and mouse. Expression of posterior *Hoxc* genes appears to be variable between replicates and between samples and is often entirely undetected, while *Hoxc4* and *Hoxc8* are expressed in all samples regardless of treatment. A separate study on *M. mulatta*, identified mRNAs with 2-fold or greater change in expression levels during simulated early pregnancy (Bishop, Satterwhite, Xu, Hennebold, & Stouffer, 2012); none of the *Hoxc* genes were included in this list. I was able to identify *Hoxc8*, but not any of the other *Hoxc* transcripts, in an ovarian transcriptome from the pig, *Sus scrofa* (Nie et al., 2011). However, this transcriptome was sequenced at a much shallower depth compared to the human and mouse ovarian transcriptomes.

It is therefore likely that, at least in rodents and primates, adult ovarian *Hoxc* gene expression shows a similar pattern with *Hoxc4* – *Hoxc8* consistently expressed and low or variable expression of *Hoxc9* – *Hoxc13*.

4.6 Candidate targets for miR-615-3p

The description of likely targets for miR-615, integrated with the expression pattern data presented above, would contribute to our understanding of how this miRNA fits in with the story of eutherian biology and evolution. Dominance of the 3' arm is observed in all cases where I have observed expression of *mir-615*, hence I have focused efforts in obtaining a target prediction dataset for miR-615-3p against all known human cDNAs (Supplement S2.6). This was generated using the PITA target prediction algorithm (Kertesz et al., 2007) with a high stringency cutoff (-20) and includes transcripts with roles in development and function of the nervous system (e.g. *KIF1A* and *PAX6*), transcriptional regulation (e.g. *YY1AP1* and *PBX3*) and cell cycle control (e.g. *RSL1D1*). For each target above the cutoff, I ascertained its expression level in tissues known to express miR-615 (kidney, testis, ovary and cerebellum; Supplement S2.6). Some targets, such as *AES*, are co-expressed with *mir-615* in all tissues examined. Others are expressed with the miRNA in some tissues but not in others; *PAX6*, for example, is expressed in the cerebellum but not in the kidney or gonads. GO term distribution was compared between the 72 predicted targets and the input set of human cDNAs (Figure 2.8). Included amongst those GO terms with >2-fold overrepresentation in predicted targets are terms associated with growth and development, as well as those involved with signalling pathways (e.g. Notch) known to regulate development (Artavanis-Tsakonas, 1999).

**Figure 2.8**

GO terms identified in the miR-615-3p target prediction dataset. The normalised frequency of each term is plotted relative to its frequency in a set of all human cDNAs. The dotted line indicates equal frequency of expression in both datasets.

5. Discussion

5.1 Origins of eutherian *mir-615*

My sequence alignment and structural prediction data suggests that *mir-615* generates functional mature products in all eutherian mammals. A similar sequence exists in the *Hoxc5* intron of marsupials (Figure 2.1) and adopts a very approximate hairpin structure (Figure 2.3). This secondary structure, while not a completely random fold, does not conform to the energetically stable stem loop required for efficient entry into the miRNA processing pathway. It is perhaps occasionally captured by the miRNA processing machinery, but is unlikely to be processed as efficiently as the eutherian precursor (see Figure 1.7 in the General Introduction).

The presence of functional miR-615 in eutherians and its absence from marsupials could be explained by either the acquisition of *mir-615* on the lineage leading to the Eutheria following its divergence from marsupials, or through acquisition of the miRNA in a common therian ancestor followed by its lineage-specific loss in marsupials. While an approximation of the ancestral therian condition might be inferred through examination of the corresponding region in monotremes, I was unable to recover any convincing hits to *Hoxc5* by BLASTn against the platypus (*Ornithorhynchus anatinus*) genome assembly, or amplify its intron through PCR in three monotreme species (*O. anatinus*, short-beaked echidna *Tachyglossus aculeatus*, long beaked echidna *Zaglossus bruijnii*). It is plausible that monotremes have undergone lineage-specific loss of part of the *Hoxc* cluster; posterior *Hoxc* genes (*Hoxc10* – *Hoxc13*) can be found in the platypus draft genome (Warren et al., 2008) while *Hoxc4*, *Hoxc5*, *Hoxc6*, *Hoxc8* and *Hoxc9* are all absent. A hypothetical monotreme homolog of *mir-615* would therefore either have been lost with the *Hoxc5* gene, or exist in some other genomic region. I was unable to locate any sequence resembling *mir-615* in the *O. anatinus* genome, suggesting that it is unlikely to be present in this lineage. Given that that none of the 17 eutherian species studied across the four major eutherian lineages have lost this miRNA or significantly altered either of its mature products, the scenario whereby *mir-615* is a eutherian novelty is probably more likely than the alternative scenario where it arose on the common mammalian lineage and was subsequently lost in monotremes and marsupials.

Whichever situation represents the present-day phylogenetic distribution of *mir-615*, the underlying inference is identical – selective pressure for the maintenance of this miRNA occurred only in the Eutheria. While it is tempting to speculate that this necessitates an involvement of the miRNA in some eutherian-specific biological process (for example, the highly developed system for maternal provisioning during gestation), it is also possible that *mir-615* has become incorporated into regulatory loops governing existing conserved biological functions.

5.2 miR-615 as a model for understanding intronic miRNA evolution

Intronic exaptation has been proposed as a major contributor to miRNA innovation (Campo-Paysaa et al., 2011; Meunier et al., 2013). Under this model, an intronic origin for a novel miRNA eliminates the need to evolve a separate promoter driving miRNA transcription, at least initially. This reduces the requirement for novel miRNA emergence to the acquisition of a suitable secondary structure by the existing transcript.

A possible model for the origin of miR-615, describing a eutherian-specific origin via intronic exaptation, is shown in Figure 2.9A. The ancestral sequence was not processed very efficiently as a miRNA or stably integrated into any genetic regulatory pathways. Following the divergence of the Eutheria from the Metatheria, different evolutionary forces shaped the evolution of the intron in each lineage. Modification of the eutherian sequence, perhaps initially through drift and subsequently through positive selection, resulted in the gradual emergence of a precise stem-loop structure which could efficiently enter the miRNA processing pathway (Axtell et al., 2011). Under this scenario the lack of either the right mutations, or the lack of suitable selection pressure after the emergence of such a structure, would explain the absence of functional miR-615 in marsupials. The observed sequence-level similarity between the corresponding *Hoxc5* intronic regions in marsupials and eutherians, with a precise stem-loop structure present in eutherians but not marsupials, tied in with the observation that no sequence resembling *mir-615* exists elsewhere in any mammalian genome, presents a strong case for the evolution of miR-615 through intronic exaptation.

Given the observation that the structure of the corresponding region in the *Hoxc5* intron of modern marsupials resembles more of an approximate hairpin than a completely random fold, we should also consider the possibility of secondary loss or degradation in the Metatheria (Figure 2.9B). Terminal branches of phylogenies tend to be rich in young miRNAs, most of which undergo rapid turnover and will eventually be lost (Lyu et al., 2014; Meunier et al., 2013; Quah, Hui, & Holland, 2015, also see Chapter 3). It is therefore not implausible that miR-615, newly emergent in the ancestral therian lineage, was captured by selection and retained in only in the eutherian lineage but not in Metatheria. Under this model, a young *mir-615* gene would not have been stably integrated into any genetic regulatory networks at the time of the eutherian-metatherian split. Positive selection and drift represented two counteracting forces governing the fate of new miRNA genes in each lineage. In the Eutheria, the effect of positive selection in maintaining *mir-615* in the genome outweighed the degradation of the secondary structure through drift, while the opposite would have been true for the Metatheria. This model would push back the origin of *mir-615* to the common therian lineage.

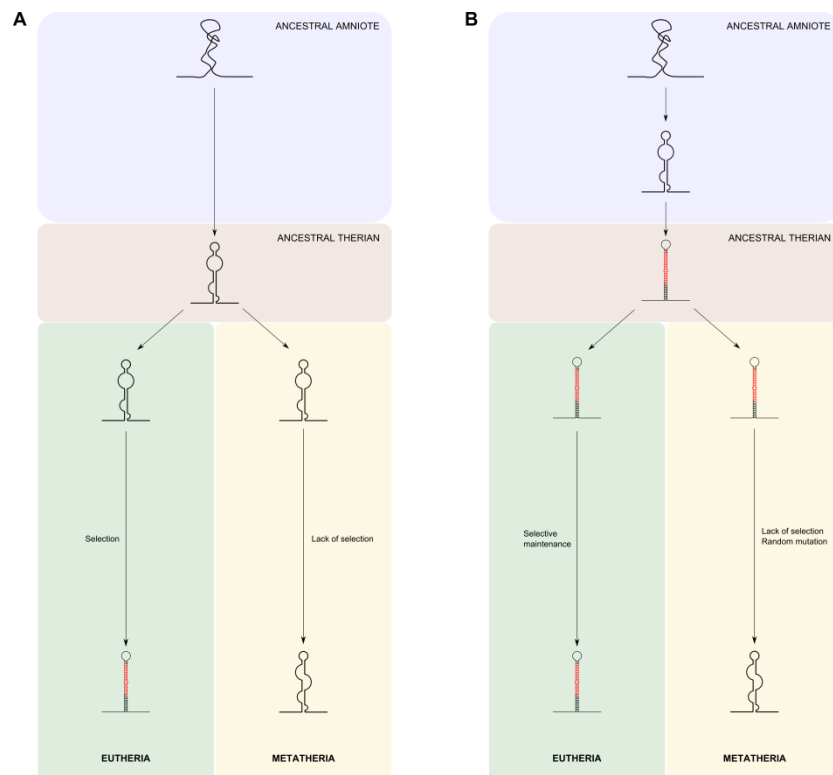


Figure 2.9

Hypothetical scenarios for the origin of miR-615. (A) Eutherian-specific origin. (B) Origin in the ancestral therian lineage followed by secondary loss in Metatheria.

The study of *mir-615* and its relationship with its host gene also sheds light on the evolution of intronic miRNA regulation. Under the intronic exaptation model (Campo-Paysaa et al., 2011), transcription of the *mir-615* gene would have been solely regulated through its host promoter upon its initial emergence. If the miRNA was to persist in the genome, it must have had functions that complemented, or at least did not interfere with, those of its host gene. In modern eutherians, all tissues and cell lines known to express *Hoxc5* also express *mir-615*. This observation suggests either a mechanistic constraint where pre-miR-615 is regulated mainly at the transcriptional level and must complete processing following its production from the host promoter, or an adaptive reason to retain miRNA expression with host expression rather than subjecting pre-miR-615 to post-transcriptional modulation. Here, I argue for the latter scenario. It is known that mature miRNA production can be modulated post-transcriptionally through differential processing and maturation (Ambros, Lee, Lavanway, Williams, & Jewell, 2003; Obernosterer, Leuschner, Alenius, & Martinez, 2006; Suzuki et al., 2009). This suggests that the retention of mature miR-615 production where *Hoxc5* is expressed is the result of selective maintenance for complementary host and miRNA functions rather than an inherent mechanistic constraint. In Chapter 4, I present functional evidence that this is the case for another intronic miRNA, miR-2768 (see also Quah et al., 2015).

Multiple lines of evidence collectively present a strong case for independent regulation of *mir-615* from its host gene in modern eutherians. RNA-seq data indicate that mature products of *mir-615* are detected in many human cell lines and tissues with no *Hoxc5* expression. This independent transcription of *mir-615* is likely to occur through the use of separate transcriptional control sequences which do not result in production of a functional host transcript (Figure 2.5). Most miRNA promoters are likely to be located within ~500 bp upstream of the pre-miRNA sequence (Zhou, Ruan, Wang, & Zhang, 2007), which would place it within either the intron or first coding exon of *Hoxc5*. The region upstream of human *mir-615*, within the limits of the large CpG island spanning most of the *HOXC5* locus (Figure 2.6), is known to undergo DNA methylation in pancreatic cancer cell lines and is reported as a promoter for *mir-615* (Gao et al., 2014) although no precise promoter elements were identified. Histone methylation, DNase I hypersensitivity and CAGE tag data, combined with analysis of ESTs points to a probable transcription initiation site within the first exon, approximately 550 bp upstream of *mir-615* (Figure 2.6). As the majority of the *HOXC5* locus including the candidate miRNA promoter lies within the CpG island, it is likely that transcription is

driven over a broad range of TSSs rather than a precise initiation site (Carninci et al., 2006; Frith et al., 2008). Also, the observed upregulation of miR-615 production in goat ovaries during pregnancy may be driven separately from *Hoxc5*; ovarian *Hoxc* expression in other mammals (rodents and primates) appears to be fairly constant across physiological states. Future work elucidating whether this constancy in *Hoxc* expression is also applicable in goats, and whether *mir-615* upregulation is observed during pregnancy in other mammals, will assist in evaluating the importance of an miRNA-specific promoter to ovarian *mir-615* expression during pregnancy.

I hypothesise that the acquisition of the CpG island in eutherian *Hoxc5* could have facilitated the evolution of independent promoter sequences for *mir-615*. The existence of TSSs located within exons has been noted (Brodsky et al., 2005), and has been hypothesised as a significant contributor to truncated non-coding mRNAs (Sandelin et al., 2007). Internal initiation leading to the production of such transcripts may have enabled *mir-615* to be expressed independently of its host gene. While this would have occurred initially by chance, it would afford subsequent selection to increase expression levels, perhaps in a tissue-specific manner. CpG island-associated promoters are known to evolve rapidly (Carninci et al., 2006), which would promote the evolution of *mir-615* regulation. The CpG island covering the putative TSS for *mir-615* along with much of the *Hoxc5* locus in eutherians is absent in marsupials. In the Tasmanian devil (*Sarcophilus harrisi*), the corresponding region is not annotated as a CpG island, and the GC content of the *Hoxc5* intron is lower in marsupials than in eutherians (52% in Tasmanian devil and tammar wallaby compared to 60% in human, 60% in mouse and 62% in cow). It is more likely that CpG islands were gained in eutherians, rather than present in a common ancestor and lost in marsupials (S. Suzuki, Shaw, Kaneko-Ishino, Ishino, & Renfree, 2011).

5.3 Putative functions for miR-615

Several studies have explored the roles of miR-615 in the context of human pathologies (Backes, Meese, Lenhof, & Keller, 2010; El Tayebi, Hosny, Esmat, Breuhahn, & Abdelaziz, 2012; Gao et al., 2014; Hulf et al., 2011). The miRNA is known to be highly expressed in the prostate cancer cell line LNCaP relative to normal prostate epithelial cells (Hulf et al., 2011), as well as in cirrhotic and cancerous liver tissues and cell lines but not healthy liver (El Tayebi et al., 2012). It is likely that in

these contexts, miR-615-5p is acting as a tumour suppressor (El Tayebi et al., 2012; Gao et al., 2014) through the inhibition of insulin-like growth factor 2 (IGF2). However, the involvement of miR-615 in non-pathologic biological processes is currently unknown.

Integrating the data on miR-615 in this analysis as well as in the context of published literature can provide clues to the potential function of this miRNA. The 3p product is dominant in every known instance where the miRNA is expressed, which is suggestive of biological function. My analysis of data from Chiang et al. (2010) indicates that whole-body expression of miR-615 appears to increase throughout mouse embryonic development and persists in the newborn (Figure 2.4). Although expression levels across replicates in this study were somewhat variable, the greatest increase in expression seems to occur between E9.5 and E12.5. miR-615 is also amongst the 10 miRNAs which are most upregulated upon differentiation of human embryonic stem cells (Bar et al., 2008). Known tissues in which miR-615 expression is conserved in more than one eutherian species include the cerebellum, kidney and gonads. While the observation that miR-615 is strongly upregulated in goat ovaries during pregnancy (Zhang et al., 2013) is striking, this data comes from a single biological replicate and should be interpreted with caution.

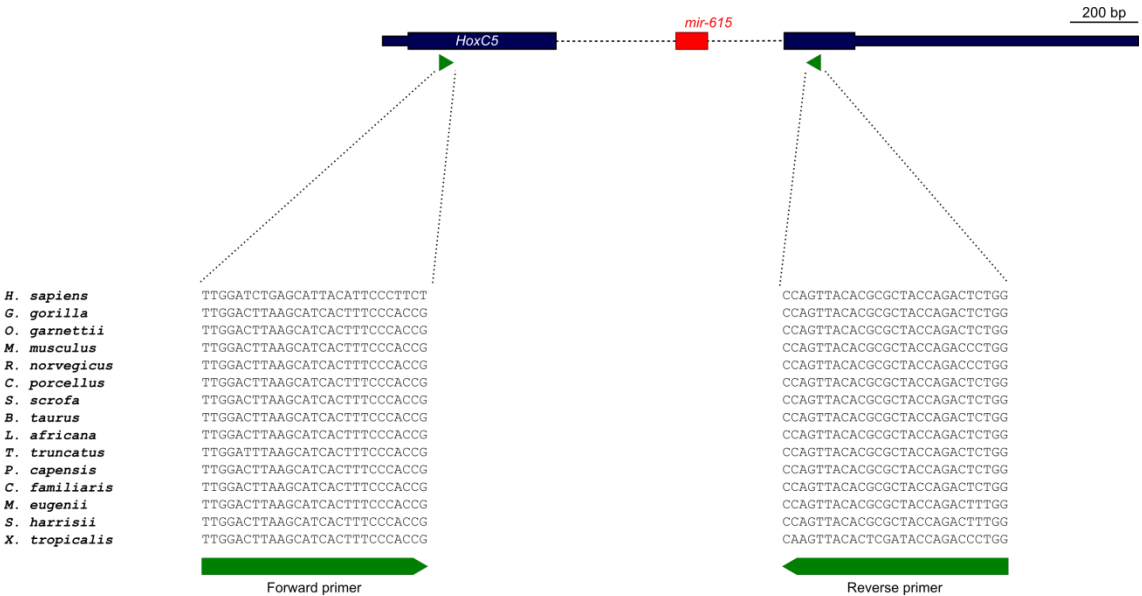
Here I demonstrate that GO terms associated with regulation of growth and development are overrepresented in a high stringency target prediction dataset for miR-615-3p. This observation, coupled with the high expression of miR-615-3p in the mouse embryo, hints at a potential role for miR-615-3p in the regulation of growth and development during embryogenesis. Although the expression of *Hoxc5* during axial patterning is likely to result in production of miR-615-3p and may account for some of the observed expression, axial Hox gene expression is initiated prior to E9.5 (Deschamps & van Nes, 2005) and is unlikely to explain the apparent upregulation of the miRNA in between E9.5 and E12.5. I hypothesize that the observed increase in expression results from either upregulation through the host promoter elsewhere in the developing embryo, or through the miRNA-specific promoter. Major developmental events occurring between E9.5 and E12.5 include organogenesis of the kidney (Vainio & Lin, 2002; Wellik, 2011), sex-specific differentiation of the gonads (Ross & Capel, 2005) and outgrowth of the limb primordia (Wanek, Muneoka, Holler-Dinsmore, Burton, & Bryant, 1989). The observation that ovaries from newborn mice express miR-615-3p at a fairly high level (Ahn et al., 2010) suggests that some of this may result from ovarian expression of miR-615. A study done on the embryonic gonads of the sheep

(*Ovis aries*) lends some weight to the hypothesis that miR-615 may be involved in gonadal differentiation (Torley et al., 2011). miR-615 is one of the miRNAs identified as being expressed at significantly higher levels (fold change ≥ 2) in ovaries at early gestation (42 days) than in mid gestation. This difference is not observed in the foetal testis. However, caution should be applied in interpretation of data across different species. Another possibility is that miR-615 is involved in the modulation of embryonic growth; studies on human cancer cell lines have implicated this miRNA in the regulation of cell growth, proliferation and migration (El Tayebi et al., 2012; Gao et al., 2014).

6. Supplementary Information

Supplement S2.1

Position of forward and reverse primers used to amplify the *HOXC5* intronic region in *Xenarthra*. Primers were designed against highly conserved regions within the *HoxC5* coding sequence. Corresponding sequences for each vertebrate species are shown.



Supplement S2.2

Cell lines available through the ENCODE Consortium (2003-2012) for which small RNA sequencing data is available.

Cell line/Tissue	Description	Sex	RNAseq data available?
A549	Epithelial cell line, lung carcinoma	Male	Yes
AG04450	Foetal lung fibroblast	Male	Yes
BJ	Skin fibroblast	Male	Yes
CD14+	Monocytes	Female	Yes
B cells CD20+	B cells	Female	Yes
CD34+	Haematopoietic progenitor cells - mobilized	Male	Yes
GM12878	Lymphoblastoid	Female	Yes
H1-hESC	Embryonic stem cell	Male	Yes
H1-neurons	Neurons derived from H1 ESCs	Male	No
HAoAF	Aortic adventitial fibroblasts	Both	Yes
HAoEC	Aortic endothelial cells	Both	Yes
HCH	Undifferentiated chondrocytes	Female	Yes
HeLa-S3	Cervical carcinoma	Female	Yes
HepG2	Liver carcinoma	Male	Yes
HFDPc	Follicle dermal papilla cells	Female	Yes
HMEpC	Mammary epithelial cells	Female	Yes
hMSC-AT	Undifferentiated mesenchymal stem cells from abdomen adipose tissue	Female	Yes
hMSC-BM	Undifferentiated mesenchymal stem cells from femoral head	Female	Yes
hMSC-UC	Undifferentiated mesenchymal stem cells from matrix (umbilical cord)	Female	Yes
HOB	Undifferentiated osteoblasts	Female	Yes
HPC-PL	Undifferentiated pericytes	Female	Yes
HPIEpC	Placental epithelial cells amniotic membrane	Female	Yes
HSaVEC	Saphenous vein endothelial cells	Male	Yes
HVMF	Villous mesenchymal fibroblasts	Male	Yes
HWP	Undifferentiated white preadipocytes	Male	Yes
IMR90	Foetal lung fibroblast	Female	Yes
K562	Chronic myelogenous leukaemia	Female	Yes
MCF-7	Mammary gland, adenocarcinoma	Female	Yes
NHDF	Dermal fibroblasts	Female	Yes
NHEK	Epidermal keratinocytes	Unknown	Yes
NHEM.f_M2	Epidermal melanocytes	Male	Yes
NHEM_M2	Epidermal melanocytes	Both	Yes
Prostate	Prostate tissue	Male	No
SkMC	Skeletal striated muscle cells	Both	Yes
SK-N-SH	Neuroblastoma	Female	Yes
SK-N-SH_RA	Neuroblastoma cell line, treated with retinoic acid	Female	Yes

Supplement S2.3

Summary of data representing small RNA and mRNA sequencing in ENCODE cell lines

Cell line/Tissue	miR-615 expression	HOXC5 expression
A549	+	-
AG04450	Low	-
BJ	Low	-
CD14+	-	-
B cells CD20+	-	-
CD34+	-	-
GM12878	+	-
H1-hESC	-	-
H1-neurons	-	(No data available)
HAoAF	+	Low
HAoEC	Low	-
HCH	+	+
HeLa-S3	+	-
HepG2	Low	-
HFDPC	+	-
HMEpC	+	-
hMSC-AT	+	+
hMSC-BM	+	+
hMSC-UC	+	+
HOB	+	+
HPC-PL	+	Low
HPIEpC	+	-
HSaVEC	+	-
HVMF	+	Low
HWP	+	-
IMR90	+	-
K562	+	-
MCF-7	+	-
NHDF	+	+
NHEK	+	Low
NHEM.f_M2	+	Low
NHEM_M2	+	-
Prostate	+	(No data available)
SkMC	+	+
SK-N-SH	+	+
SK-N-SH_RA	+	+

Supplement S2.4

Read counts for miR-615 mature products in the GM12878 and K562 cell lines, as generated by the miRDeep2 quantifier.pl script. Raw small RNA sequencing data was obtained from the ENCODE Consortium.

Cell line	ID	miRNA	Read count
GM12878	ENCFF001RDJ	hsa-miR-615-5p	0
		hsa-miR-615-3p	5
	ENCFF001RDK	hsa-miR-615-5p	1
		hsa-miR-615-3p	9
K562	ENCFF001RDL	hsa-miR-615-5p	0
		hsa-miR-615-3p	132
	ENCFF001RDM	hsa-miR-615-5p	0
		hsa-miR-615-3p	44

Supplement S2.5

Microarray data on ovarian *HoxC* gene expression, obtained via the Gene Expression Omnibus.

Dataset	Type	Species	Description	Expression								
				Hoxc4	Hoxc5	Hoxc6	Hoxc8	Hoxc9	Hoxc10	Hoxc11	Hoxc12	Hoxc13
GSM874239	Microarray	<i>M. musculus</i>	Ovary, age = 12 weeks	Yes	Yes	Yes	Yes	No	Yes	?	?	No
GSM874241	Microarray	<i>M. musculus</i>	Ovary, neonatal day 1	Yes	Yes	No	Yes	No	Yes	?	?	No
GSM35553	Microarray	<i>M. musculus</i>	Ovary (commercially obtained RNA)	Yes	Yes	Yes	Yes	Yes	Yes	?	?	No
GSM563219	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 1	Yes	?	?	Yes	No	No	No	Yes	Yes
GSM563220	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 2	Yes	?	?	Yes	No	No	No	No	Yes
GSM563221	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 3	Yes	?	?	Yes	No	Yes	Yes	Yes	No
GSM563222	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 4	Yes	?	?	Yes	No	Yes	Yes	Yes	No
GSM563223	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 5	Yes	?	?	Yes	No	No	No	Yes	Yes
GSM563224	Microarray	<i>M. mulatta</i>	Ovarian follicle, 0h stimulation - replicate 6	Yes	?	?	Yes	No	No	No	No	Yes
GSM563225	Microarray	<i>M. mulatta</i>	Ovarian follicle, 12h post-hCG - replicate 1	Yes	?	?	Yes	No	No	No	No	Yes
GSM563226	Microarray	<i>M. mulatta</i>	Ovarian follicle, 12h post-hCG - replicate 2	Yes	?	?	Yes	No	No	No	No	Yes
GSM563227	Microarray	<i>M. mulatta</i>	Ovarian follicle, 12h post-hCG - replicate 3	Yes	?	?	Yes	No	No	No	Yes	Yes
GSM563228	Microarray	<i>M. mulatta</i>	Ovarian follicle, 12h post-hCG - replicate 4	Yes	?	?	Yes	No	Yes	No	No	No
GSM563229	Microarray	<i>M. mulatta</i>	Ovarian follicle, 24h post-hCG - replicate 1	Yes	?	?	Yes	No	Yes	No	No	Yes
GSM563230	Microarray	<i>M. mulatta</i>	Ovarian follicle, 24h post-hCG - replicate 2	Yes	?	?	Yes	No	Yes	No	No	No
GSM563231	Microarray	<i>M. mulatta</i>	Ovarian follicle, 24h post-hCG - replicate 3	Yes	?	?	Yes	No	No	No	No	No
GSM563232	Microarray	<i>M. mulatta</i>	Ovarian follicle, 24h post-hCG - replicate 4	Yes	?	?	Yes	No	No	No	Yes	No
GSM563233	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 1	Yes	?	?	Yes	No	No	No	No	Yes
GSM563234	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 2	Yes	?	?	Yes	No	No	No	No	Yes
GSM563235	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 3	Yes	?	?	Yes	No	No	No	No	Yes
GSM563236	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 4	Yes	?	?	Yes	No	Yes	Yes	Yes	Yes
GSM563237	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 5	Yes	?	?	Yes	No	No	No	Yes	Yes
GSM563238	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 6	Yes	?	?	Yes	No	Yes	No	No	Yes
GSM563239	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 7	Yes	?	?	Yes	No	Yes	No	Yes	Yes
GSM563240	Microarray	<i>M. mulatta</i>	Ovarian follicle, 36h post hCG - replicate 8	Yes	?	?	Yes	No	No	No	No	Yes

Supplement S2.6

PITA prediction targets for miR-615-3p

Gene	Function	Kidney	Testis	Ovary	Cerebellum
ACACB	Regulation of fatty acid oxidation	20	12	21	5.12
AES	Corepressor for NF-kappaB	178	98	165	39.03
AGO3	RISC complex component	4	12	5	19.19
AMER1	Inhibition of Wnt signalling through promoting β -catenin degradation	2	1	6	4.43
ARHGAP35	Regulation over formation of the cytokinetic furrow during cell division	20	59	17	33.51
BATF2	Possible role in regulation of Wnt signalling pathway	4	1	3	0.08
BIN2	Promotes cell motility and migration	2	12	2	0.48
BLOC1S3	Component of the BLOC-1 complex, required for normal biogenesis of lysosome-related organelles	3	5	5	1.07
C12orf65	Mitochondrial matrix protein, possibly involved in mitochondrial translation	14	16	19	6.37
CACNA1A	Component of voltage-gated calcium channel, predominantly neuronal expression	0	3	1	28.03
CAMK2D	Serine/threonine kinase important in cardiac and skeletal muscle function	16	24	27	41.93
CARNS1	Synthesis of carnosine and homocarnosine	1	1	1	0.19
CD14	Cell surface antigen involved in innate immune response to bacterial lipopolysaccharide	20	31	15	7.53
CENPP	Kinetochore component	1	1	1	1.64
CES1	Triglyceride and cholesterol metabolism	3	8	20	0.33
CFC1B	Unknown	0	0	0	0
CHST12	Catalyzes the transfer of sulphate to position 4 on N-acetylgalactosamine	5	4	3	3.57
CHST6	Catalyzes the transfer of sulphate to position 6 on N-acetylgalactosamine	0	1	0	1.37
CIC	Transcriptional repressor, possibly involved in development of the central nervous system	15	39	28	10.15
CYSLTR2	Receptor for cysteinyl leukotrienes	1	1	1	0.52
DCHS1	Calcium-dependent cell adhesion protein, may be involved in neurogenesis	3	4	8	26.4
DCLRE1B	5'-3' exonuclease, involved in telomere protection and maintenance	2	3	3	3.6
DLL3	Notch signalling pathway, potential roles in axial patterning and inhibition of neurogenesis	0	1	0	7.13
EIF3B	Required for initiation of protein synthesis	39	73	56	24.68
ENSA	Regulation of ATP-sensitive potassium channels	89	40	71	43.31
EPN1	Regulation of receptor-mediated endocytosis, required for activation of Notch signalling in mammals	38	32	22	10.41
FAH	Amino acid degradation	56	19	25	4.18

FAM134B	Required for long-term survival of nociceptive and autonomic ganglion neurons	38	42	3	10.6
FURIN	Ubiquitous endoprotease activity within constitutive secretory pathways	25	8	10	4.93
FXD3	Regulator of Na-K ATPase activity	9	0	1	0.32
GAB2	Adapter protein acting downstream of several membrane receptors	5	6	19	7.07
GCN1L1	Translational activator	15	23	19	31.32
GDAP2	Unknown	6	12	4	18.22
HAPLN4 (BRAL2)	Synaptic stabilization, involved in central nervous system development	0	0	0	0.63
HSD17B12	Fatty acid elongation	72	43	48	37.15
IQGAP3	Rho family GTPase, promotes axon and neurite outgrowth in brain	0	4	0	2.67
KIF1A	Motor protein for anterograde axonal transport of synaptic vesicle precursors	0	14	2	90.37
LMNTD2	Unknown	3.7	12.3	0	0.06
LPIN3	Regulates fatty acid metabolism	11	4	12	0.15
LSP1	Cytoskeletal organisation and motility in neutrophils	5	17	5	0.34
MADCAM1	Immunoglobulin superfamily adhesion molecule expressed by lymphocytes	0	1	0	0.77
MED22	Mediator complex component, involved in transcriptional regulation of Pol II-dependent genes	9	7	13	8.81
NARS	Asparagine--tRNA ligase	103	43	48	33.5
NDN	Growth suppressor, promotes cell cycle arrest particularly in postmitotic neurons in the brain	10	21	64	27.84
NFRKB	Component of the INO80 complex involved in chromatin remodelling, transcriptional regulation and DNA repair	9	23	20	9.86
NODAL	Morphogen required for mesoderm induction and axial patterning	0	1	0	0.43
NPEPL1	Unknown	24	25	19	5.06
ORMDL3	Endoplasmic reticulum transmembrane protein, mediates sphingolipid homeostasis	39	37	36	18.63
PAX6	Transcription factor, involved in development of the eye, nose, forebrain and pancreas	1	2	0	63.86
PBX3	Transcriptional activator	7	13	76	53.12
PES1	Ribosomal RNA processing	25	29	30	12.91
PHACTR3 (SCAPININ)	Regulation of the actin cytoskeleton involved in cell motility	2	1	5	8.54
PIF1	Inhibition of telomerase activity	1	3	0	2.23
PLA2G2C	Unknown, possibly fertilization. Expressed in mouse testis.	0	0	0	0
PRAME	Dominant repressor of retinoic acid signalling	1	78	3	0.04
PRG2 (LPPR3)	Integral membrane protein, modulates bioactive lipid phosphates in the context of cell migration, neurite retraction and mitogenesis	0	1	0	0.16

PTP4A3	Promotes cell proliferation	9	12	5	13.33
RBFA	Ribosomal RNA processing	12	11	11	26.72
RSL1D1 (CSIG)	Regulation of cell cycle, cell senescence and apoptosis	59	53	249	24.01
SESN2	Modulator of peroxide signalling, antioxidant defence, regulation of cell growth	15	8	13	1.36
SKA3	Component of Ska complex which mediates kinetochore-microtubule interactions during mitosis	0	8	0	5.26
SLC2A11 (GLUT11)	Glucose transporter	25	8	13	13.95
SPECC1	Unknown, highly expressed in testicular tissue	6	35	23	28.82
STRA6	Receptor for retinol binding protein, required for retinol transport to specific sites e.g. eye	13	7	1	5.53
STRBP	Microtubule-associated RNA binding protein, highly expressed in testis and brain and possibly important to development of these structures	12	49	4	34.87
STX16-NPEPL1	Naturally occurring read-through transcript between STX16 and NPEPL1, unlikely to be functional	4	2	3	5.21
TMEM161A	Unknown	12	6	12	4.52
TRAPPC4	Component of TRAPP complex involved in vesicle trafficking	62	57	39	13.96
TRAPPC5	Component of TRAPP complex involved in vesicle trafficking	27	52	16	19.42
TRIM31	TRIM family member, possibly involved in cell compartment specification	0	2	0	0.02
UBA1	Ubiquitin conjugation, required for ubiquitylation-dependent signalling of double-stranded breaks and replication stress in cells	83	84	80	47.62
YY1AP1	Coactivator for transcriptional regulator YY1	21	71	42	92.15

Chapter 3:
miRNA innovation within the
Lepidoptera

Statement of Author Contributions

The main body of this chapter was submitted as a manuscript formatted as a Research Article for *Molecular Biology and Evolution*, in conjunction with material from Chapter 4 (submitted September 16, 2014; Manuscript ID MBE-14-0982). It was accepted for publication on December 30, 2014 after minor revision and first published online on January 8, 2015. The authors on this manuscript are in order: Shan Quah, Jerome H. L. Hui and Peter W. H. Holland (corresponding author).

The original idea for the project was conceived through discussion amongst all authors. Material from the Speckled Wood butterfly was supplied by Casper J. Breuker (Oxford Brookes University). SQ performed sample collection for the Horse chestnut Leafminer moth, all laboratory-based work, computational analyses, and drafted the manuscript. All authors discussed the results and revised the manuscript. JHLH supervised the work in the initial stages of this study; PWHH supervised the work throughout the duration of this study. Ferdinand Marlétaz (University of Oxford), Jordi Paps (University of Oxford) and James Tarver (University of Bristol) provided scientific and bioinformatics advice. Jean-Michel Carter (University of Oxford) and Luca Livraghi (Oxford Brookes University) assisted with maintaining the laboratory-based Speckled Wood stocks.

Sections 4.6 and 5.3 in this chapter were not part of the submitted manuscript. The miRNA evolution model presented in Section 5.3 is the subject of ongoing research between SQ, PWHH and Michael B. Bonsall (University of Oxford). We aim to submit this as a mathematical model in a hypothesis paper.

1. Abstract

Lepidoptera (butterflies and moths) is a species-rich clade with a well-established phylogeny and abundant genomic resources, thereby representing an ideal system in which to study miRNA evolution. I sequenced small RNA libraries from developmental stages of two divergent lepidopterans, *Cameraria ohridella* (Horse chestnut Leafminer) and *Pararge aegeria* (Speckled Wood butterfly), discovering 90 and 82 conserved miRNAs respectively, and many species-specific miRNA sequences. Mapping miRNAs onto the lepidopteran phylogeny reveals rapid miRNA turnover and an episode of miRNA fixation early in lepidopteran evolution, implying that miRNA acquisition accompanied the early radiation of the Lepidoptera. I propose that this observed burst of miRNA acquisition on the lepidopteran stem, coupled with the presence of large numbers of young, lineage-specific miRNAs is evidence for a constant, high turnover of miRNA genes with variable rates of miRNA retention in the genome.

2. Introduction

The miRNA repertoire in a given evolutionary lineage is constantly shifting under the opposing influences of miRNA birth and miRNA death (the gain and loss of miRNA genes). Previous work in vertebrates (Heimberg et al., 2008), mammals (Meunier et al., 2013), drosophilids (Lu et al., 2008; Lyu et al., 2014; Nozawa, Miura, & Nei, 2010) and flatworms (Fromm et al., 2013) have suggested that miRNA birth rates can differ substantially in different lineages, while loss is thought to be less frequent than gain. For example, there was a burst of miRNA innovation on the stem lineage leading to the vertebrates following divergence from other chordates, and this has been postulated to be a contributory factor in the evolution of vertebrate complexity (Heimberg et al., 2008). Endoparasitic platyhelminths have undergone significant losses of conserved miRNA families which have not been observed to the same extent in other bilaterian phyla studied. This is proposed to be correlated with the evolution of reduced morphological and metabolic complexity (Fromm et al., 2013).

Lepidoptera represents an ideal clade to study the relationship between miRNAs and morphological evolution (Figure 3.1). This group comprises over 150,000 species of moths and butterflies and is one of the most species-rich insect orders. Lepidopterans share a large number of morphological apomorphies and underwent a major evolutionary radiation during the Cretaceous (Regier et al. 2013; Wahlberg, Wheat, and Peña 2013; Kawahara and Breinholt 2014) following their divergence from their sister group Trichoptera (caddisflies). The phylogeny of the Lepidoptera is well-studied, with clearly supported superfamilies identified through both morphological and molecular studies (Regier et al. 2013; Wahlberg, Wheat, and Peña 2013). In outline, the vast majority of species diversity is found in the Ditrysia, which makes up 99% of extant lepidopteran species (Wahlberg et al., 2013). Within this clade, the Yponomeutoidea (e.g. *Plutella xylostella*) and Gracillarioidea (e.g. *Cameraria ohridella*) are distant sister groups that diverged early from the rest of the Ditrysia (Bazinet, Cummings, Mitter, & Mitter, 2013; Regier et al., 2013). Most of the remaining ditrysians form a strongly supported clade – the Apoditrysia (Bazinet et al., 2013), which includes the Bombycoidea (including silkmoths and hawkmoths) and the Papilionoidea (butterflies, including the highly diverse nymphalids). One major advantage of using the Lepidoptera for the comparative analysis of miRNA complements is the plethora of genome resources available in the public domain, including sequences from the Domestic Silkmoth *Bombyx mori* (Mita, 2004; Xia et al., 2004), the butterflies

Heliconius melpomene (The Heliconius Genome Consortium, 2012) and *Danaus plexippus* (Zhan, Merlin, Boore, & Reppert, 2011), the Diamondback moth *Plutella xylostella* (You et al., 2013) and the Tobacco Hornworm hawkmoth *Manduca sexta*. Our group has also recently sequenced and assembled draft genomes for the Horse chestnut Leafminer moth *Cameraria ohridella*, the Speckled Wood butterfly *Pararge aegeria*, the Comma butterfly *Polygonia c-album* and the Orange Swift moth *Hepialus sylvina* (Ferguson et al., 2014). The same study also reports a draft genome assembly for the caddisfly *Glyptotendipes pallidus* as a representative of the Trichoptera (Figure 3.1).

However, while genome sequences can be used to search for known miRNA homologues, sampling of actual small RNAs is required for the identification of novel sequences and hence the determination of the miRNA repertoire in a given species. Prior to a recent report (Etebari, Hussain, & Asgari, 2013) on small RNA sequencing in *P. xylostella* (Yponomeutoidea), miRNA data were limited to *B. mori* and *M. sexta* (Bombycoidea; Liu et al., 2009, 2010) with limited information from wing discs of *H. melpomene* (Surridge et al., 2011). To enlarge the comparative analysis, ensuring the inclusion of distant lineages, the gracillarid *C. ohridella* and the nymphalid butterfly *P. aegeria* were selected for small RNA sequencing. My analyses reveal that a burst of miRNA fixation occurred early in the evolution of the clade. Few miRNA fixations occur on more recent internal branches, even in speciose groups such as the Nymphalidae. I also find high numbers of evolutionarily young, species-specific miRNA sequences, suggesting that high miRNA birth rates are matched by high death rates with few novel genes persisting to integrate stably into developmental programs.

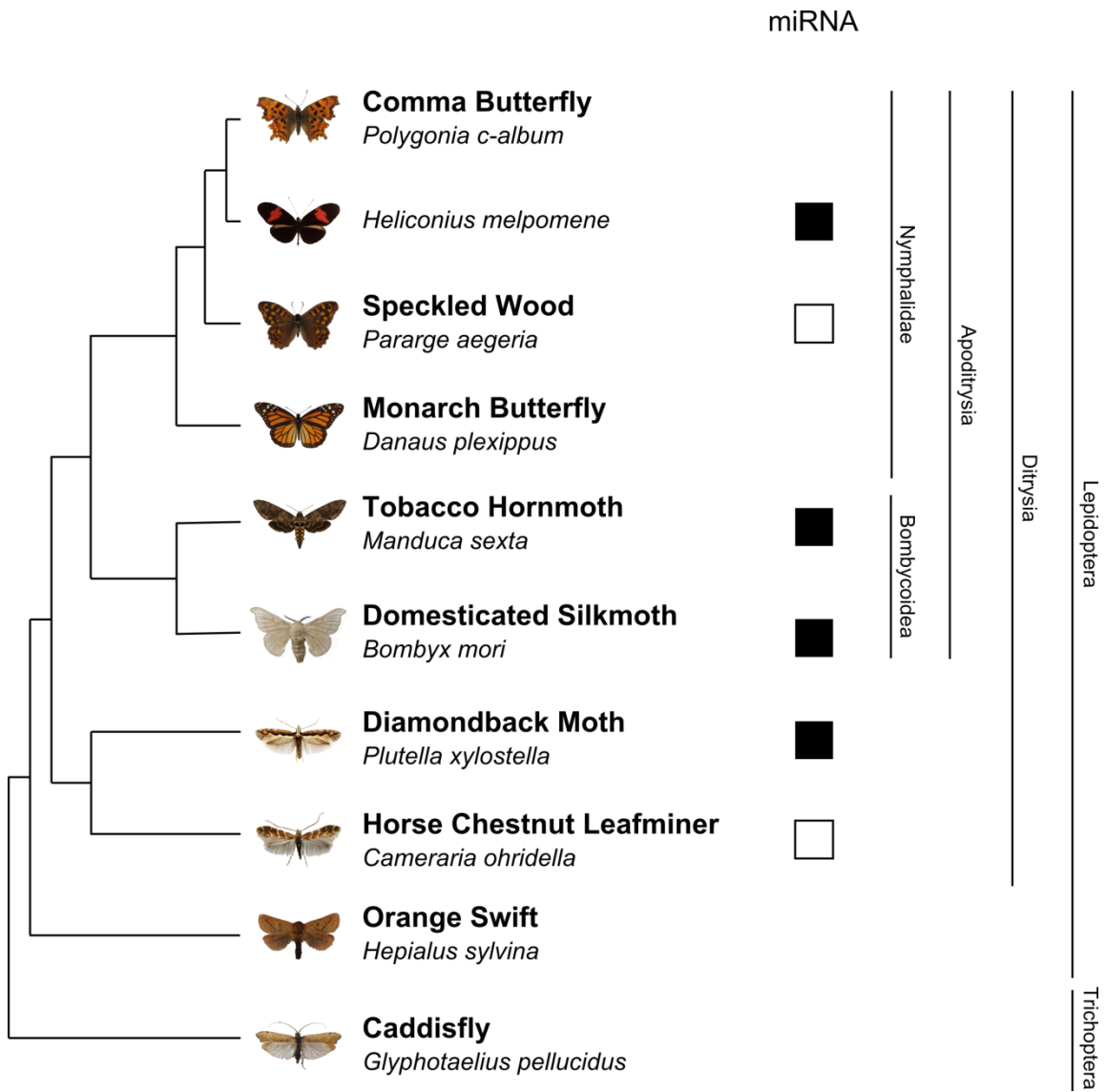


Figure 3.1

Phylogeny of the lepidopteran species for which genome sequence data is available. The orange swift genome is of insufficient coverage and was not used for assessment of miRNA distribution. Black squares indicate published miRNA sequencing projects and white squares represent the data presented in this study.

3. Materials and Methods

3.1 Sample collection

C. ohridella larvae and pupae were obtained by dissection from leaf mines on Horse chestnut (*Aesculus hippocastanum*) leaves collected in Oxford University Parks, UK, in July 2013. Adults were collected by incubating leaves in a damp environment in the laboratory at room temperature. Twenty individuals were sampled for each developmental stage. Adults were not sexed.

P. aegeria specimens were obtained from a colony at Oxford Brookes University originally founded with specimens from Belgium (Carter et al., 2013). Eggs were collected within 20 minutes of oviposition and homogenised 1 h, 15 h, 24 h and 48 h after oviposition. For 1st instar larvae, a mixture of whole larvae was homogenised for a total mass of 0.10 g. Two final instar larvae (total mass = 0.11 g) were used. Single whole pupae were sampled at 0 days, 3 days and 5 days post-pupation. One adult male and one adult female were homogenised separately following removal of the wings and legs.

3.2 RNA extraction and miRNA sequencing

Small RNA extraction was carried out using the miRVana miRNA Isolation Kit (Life Technologies, AM1560) following the manufacturer's protocol. Tissues were homogenised in miRVana Lysis Buffer using a single bead in the TissueLyser II (Qiagen). All tissues were homogenised fresh.

Quantity and quality of the extracted RNA was assessed using the Experion Automated Electrophoresis System (Bio-Rad) on an Experion HighSens chip. Sample pooling was carried out for each species such that equal amounts of RNA were used from each sampling timepoint. Library preparation and deep sequencing were carried out at the Beijing Genomics Institute (Hong Kong). The two libraries were barcoded and run on a single lane of the Illumina HiSeq 2000 platform. Raw read and processed data are available on the Gene Expression Omnibus (accession number GSE63644).

3.3 Computational analysis

I developed a conservative pipeline for identification of true miRNA sequences in the small RNA sequencing reads (Figure 3.2). The FastQC quality control tool (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) was first used to validate read quality. Raw reads were processed by converting FASTQ to FASTA file formats and removing adapter sequences. Reads were collapsed to remove redundancy; collapsed reads <18 nt and >24 nt were discarded. Size selection was carried out using the size-select.py script (Appendix A2) provided by Ferdinand Marlétaz. Remaining reads were mapped to Bowtie-indexed genome assemblies and those that failed to align were discarded.

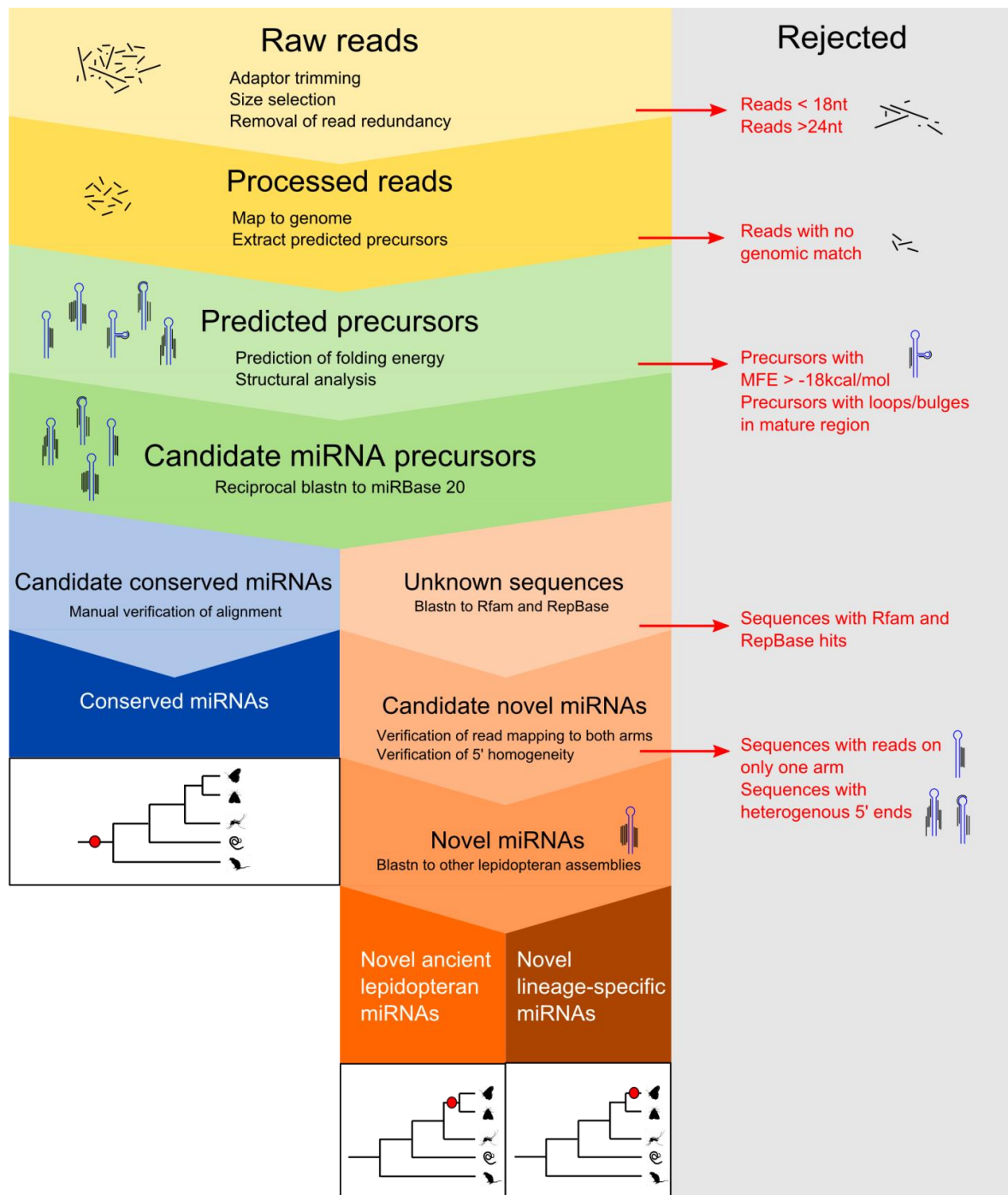
The miRDeep2 algorithm (Friedländer, Mackowiak, Li, Chen, & Rajewsky, 2012) was used to predict precursor hairpins at genomic locations where the reads mapped. Each predicted precursor was then manually inspected for the following criteria before being accepted as a true putative miRNA: stable hairpin structure with <-18 kcal/mol free energy, at least 18 paired bases on the main stem and the absence of large internal loops and bulges in mature regions. Free energy of the predicted precursors was calculated using RNAeval.

For conserved miRNA identification, all precursors passing the structural criteria were checked against known precursors in miRBase 20 using the following BLASTn parameters: word_size= 4, reward = 5, penalty -4, gapopen = 8 and gapextend = 6. BLAST hits were assessed and verified by reciprocal BLASTn. Verified sequences were annotated as conserved miRNAs. Sequences without BLASTn hits to miRBase 20 were subjected to further processing and retained only if they passed the following criteria: homogeneity of 5' ends in mature products, read mapping to both arms of the hairpin and no BLASTn hits to RepBase, Rfam or NCBI bacterial sequences (taxid:2). The positions where each candidate precursor mapped onto the genome were screened to remove duplicate, overlapping predictions.

3.4 Phylogenetic placement of lepidopteran miRNAs

All documented miRNA precursor sequences were downloaded from miRBase 20 (Kozomara & Griffiths-Jones, 2011) and filtered to remove sequences present outside

of the Lepidoptera (represented by *B. mori*, *M. sexta* and *H. melpomene*). Lepidopteran-specific miRBase entries, along with sequences reported here, were then compared by BLASTn against genome assemblies for *Heliconius melpomene* (www.butterflygenome.org), *Danaus plexippus* (<http://monarchbase.umassmed.edu/>), *Manduca sexta* (<https://www.hgsc.bcm.edu/arthropods/tobacco-hornworm-genome-project>), *Bombyx mori* (<http://sgp.dna.affrc.go.jp/KAIKObase/> and <http://silkworm.genomics.org.cn/silkdb/>), *Plutella xylostella* (<http://dbm.dna.affrc.go.jp/px/>), *Glyphotaelius pellucidus*, *Hepialus sylvina*, *Cameraria ohridella*, *Pararge aegeria* and *Polygonia c-album* using the BLASTn parameters described above. Genome assembly data for *G. pellucidus*, *H. sylvina*, *C. ohridella* and *P. aegeria* are also publicly available in the Oxford University Research Data archive (doi: 10.5287/bodleian.dury.3) and raw sequence data is available in the NCBI BioProject database (accession: PRJNA241175), as reported in Ferguson et al. 2014. Each species was scored for presence or absence of individual miRNA families. Where possible, syntenic conservation was verified by comparing the positions of the miRNA and its flanking genes between species.

**Figure 3.2**

Data processing pipeline for annotation of miRNAs in *C. ohridella* and *P. aegeria* from raw Illumina sequencing output.

4. Results

4.1 Small RNA sampling in *Cameraria ohridella* and *Pararge aegeria*

Selection of informative species for miRNA sampling to help elucidate the pattern of miRNA evolution across the Lepidoptera was carried out using three essential criteria: phylogenetic placement, accessibility of different life-cycle stages to afford deep sampling of miRNA species, and availability of genomic sequence of sufficient coverage to enable novel miRNA precursor prediction and detection. The phylogenetic criteria focussed my attention on basally-branching Lepidoptera lineages and on representation of the speciose clade Nymphalidae, particularly since the published nymphalid miRNA data derive from a restricted tissue type (*H. melpomene* wing disc, Surridge et al., 2011).

Species from basally branching lineages with genomic data include the Orange Swift moth *Hepialus sylvina* and the Horse-chestnut Leafminer *Cameraria ohridella* (Figure 3.1). To assess suitability for verification of miRNA sequences, I first used a BLASTn search with 20 conserved miRNAs originating in the common hexapod ancestor or earlier (Supplement S3.1). This identified only 12 out of 20 query miRNA sequences in the former genome, but 19 out of 20 in the latter. Furthermore, *C. ohridella* represents a basally divergent lineage within the Ditrysia and a distant sister group to *P. xylostella* (Bazinet et al., 2013; Regier et al., 2013) and provides an excellent basis for comparative analysis of miRNAs originating at the base of the Ditrysian tree. This invasive species is now extraordinarily common in the UK and most life-cycle stages can be readily collected.

The second species chosen for deep-sequencing of small RNAs was the Speckled Wood butterfly (*Pararge aegeria*), a nymphalid. This species also has fairly complete genome coverage (19 out of 20 sequences recovered), and life-cycle stages can be collected from a laboratory-reared population.

4.2 Deep sequencing and miRNA identification

Small RNA sequencing was undertaken for multiple developmental stages to maximise representation of miRNAs expressed. In *P. aegeria* I sampled four time points during embryogenesis (1 h, 15 h, 24 h and 48 h post-oviposition), two larval time points (first and last instar), three pupal time points (0 days, 3 days and 5 days post-pupation) and a single adult male and female. *C. ohridella* RNA samples were from larvae, pupae and adults. In both species, whole bodies, eggs or pupae were used at each time point for RNA extraction. RNA extracted from individual stages was pooled within each species.

A total of 80,414,031 raw Illumina HiSeq reads were obtained (Table 3.1) and processed according to the schematic in Figure 3.2, designed to ensure removal of spurious reads, degradation products and other small RNAs. Since small RNA sequencing yields the mature miRNA products (22 nt), not precursor sequences (~80 nt), the miRDeep2 algorithm was used to extract predicted miRNA precursor sequences based on the positions of mapped reads in the genome. All predictions generated by miRDeep2 were first subjected to a free energy filtering step to select stable hairpins. Only genomic sequences fulfilling stringent stability criteria were considered to be candidate miRNA precursors.

4.3 Conserved miRNAs in *Cameraria* and *Pararge*

Of the candidate precursors identified, I first identified evolutionarily conserved sequences occurring in more than one species. A total of 87 and 80 'conserved miRNAs' of varying evolutionary ages were identified in *C. ohridella* and *P. aegeria* respectively (Supplement S3.2). *C. ohridella* has undergone lineage-specific duplications of several miRNA genes including miR-14, miR-2763 and miR-2796 of which only single copies exist in other Lepidoptera.

4.4 Novel miRNA discovery in *Cameraria* and *Pararge*

Following identification of previously-documented miRNAs, I filtered the remaining candidates to retain only genuine miRNA genes and classify these into two categories.

The first category consists of novel lineage-specific miRNAs found thus far only in the species sampled; the second comprises newly-discovered sequences of older origin which may have been missed in previous miRNA sequencing projects.

To ensure removal of non-miRNA gene sequences from my dataset, I subjected all remaining sequences to additional validation (Figure 3.2). In particular, sequences were checked for 5' homogeneity in the small RNA reads. Sequences with highly heterogeneous 5' ends in their mature reads are unlikely to represent genuine miRNAs as the 5' end determines the miRNA seed (Kozomara & Griffiths-Jones, 2011) and is likely to have effects on target specificity. I discarded these precursors as they could represent degradation products of longer transcripts. As a final conservative criterion, I stipulated that a candidate must have reads mapped to both arms of the hairpin in order to be described as a miRNA.

Comparison to available lepidopteran genome assemblies revealed that my dataset includes 23 and 57 lineage-specific miRNAs in *C. ohridella* and *P. aegeria* respectively (Supplement S3.2). Only one sequence, 'Par-471', was identified by BLASTn in several other ditrysian species but has not been reported in miRBase (Supplement S3.3). The recent publication of a *P. xylostella* miRNA sequencing study (Etebari et al., 2013) has identified a predicted precursor (PC-3p-82883_6) which is likely to be Par-471. Conservation of this sequence was verified by alignment and synteny checks as well as structural analysis of predicted precursors in all species where this sequence is found (Supplement S3.4).

	<i>Cameraria ohridella</i>	<i>Pararge aegeria</i>
Stages sampled	Larvae, pupae, adults	Embryos, larvae, pupae, adults
Raw reads	41621320	38792711
Processed reads	2338028	2813742
Total predicted miRNAs	113	139
Conserved miRNAs	87 + 3*	80 + 2*
Lineage-specific miRNAs	23	57

Table 3.1

Summary of small RNA sequencing data. In *C. ohridella*, I identified 87 miRNAs with entries in miRBase 20 as well as 3 newly-reported sequences (Cam-042/Par-110, Cam-110 and Cam-154) shared with other lepidopterans. In *P. aegeria*, there were 80 miRNAs with entries in miRBase 20 as well as 2 newly-reported sequences (Cam-042/Par-110 and Par-471). Cam-042 and Par-110 were discovered separately in *C. ohridella* and *P. aegeria* and are likely to be homologous miRNAs.

4.5 Phylogenetic placement of lepidopteran miRNAs

Using the information revealed by my deep sequencing together with previous studies on lepidopteran miRNAs, I determined the presence or absence of each sequence in each species (Figure 3.3A) and compared these to a phylogeny of the Lepidoptera (Regier et al., 2013). This revealed the latest possible node of origin for each sequence (Figure 3.3B). The caddisfly (*Glyphotaelius pellucidus*, Trichoptera) was used as an outgroup to the Lepidoptera. I validated the completeness of this genome by running BLASTn using the same set of 20 conserved miRNAs described previously (Supplement S3.1). 19 out of 20 of the query sequences were retrieved, suggesting that genome coverage of *G. pellucidus* is adequate and that the likelihood of miRNA genes reported absent as an artefact of genome completeness is low.

As expected, many miRNA precursors had origins prior to the last common ancestor of Lepidoptera and Trichoptera; these are not the focus of this study. A total of 459 sequences were identified which were unique to the Lepidoptera and/or Trichoptera. The majority of these are lineage-specific: 315 from *Bombyx mori*, 10 from *Heliconius melpomene*, 8 from *Manduca sexta*, 23 from *Cameraria ohridella* and 57 from *Pararge aegeria*. Of the novel sequences reported in *Plutella xylostella* (Etebari et al., 2013), I identified 21 sequences with reads mapping to both arms which are likely to be miRNA precursors (Supplement S3.4).

A total of 25 potential miRNA precursors were found to be present in Lepidoptera and Trichoptera, and shared by at least two species. Six of these (identified in previous studies) were removed from further analysis as I am not confident that they represent genuine miRNAs under my stringent criteria. These are miR-2739, miR-3389, miR-2733, miR-6495, miR-6496 and miR-6497. The miR-2738 sequence has a predicted secondary structure with an asymmetric bulge in its mature region; its pattern of read mapping in *B. mori* is suggestive of a degradation product. miR-3389 displays a heterogeneous pattern of read mapping on its 5' arm, again suggestive of a degradation product, and no reads from its 3' arm have been reported. miR-2733 maps to several different genomic regions in *B. mori* and may be derived from a repetitive element. The miR-6495, miR-6496 and miR-6497 sequences were found to be derived from ribosomal RNA fragments.

I searched for the remaining 19 sequences in all available lepidopteran and trichopteran genome assemblies using BLASTn. The miR-2779 sequence was

removed from further analysis as I am unable to confirm that the predicted precursors for this sequence identified in *B. mori*, *M. sexta* and *P. xylostella* are homologous to each other by structure or synteny.

I plotted the latest possible origin for each miRNA onto the lepidopteran phylogeny (Figure 3.3B). Only one sequence (miR-2755) is shared between the Lepidoptera and its sister group Trichoptera. Strikingly, I find that the majority of novel miRNA acquisition is concentrated at the base of the lepidopteran tree, with 11 of these novel sequences likely to have been present in the last common lepidopteran ancestor. Two other sequences (Cam-110 and Cam-154) unite the distantly-related sister species *C. ohrdella* and *P. xylostella*. There were three further innovations on the Apoditrysian stem. I find very little miRNA acquisition on more recent internodes, with a single gain (miR-6303) in a common ancestor of the Heliconiinae and Nymphalinae. The current miRNA distribution also suggests that several lineage-specific losses may have occurred, though I have less certainty about miRNA loss due to the limitations of genome assemblies.

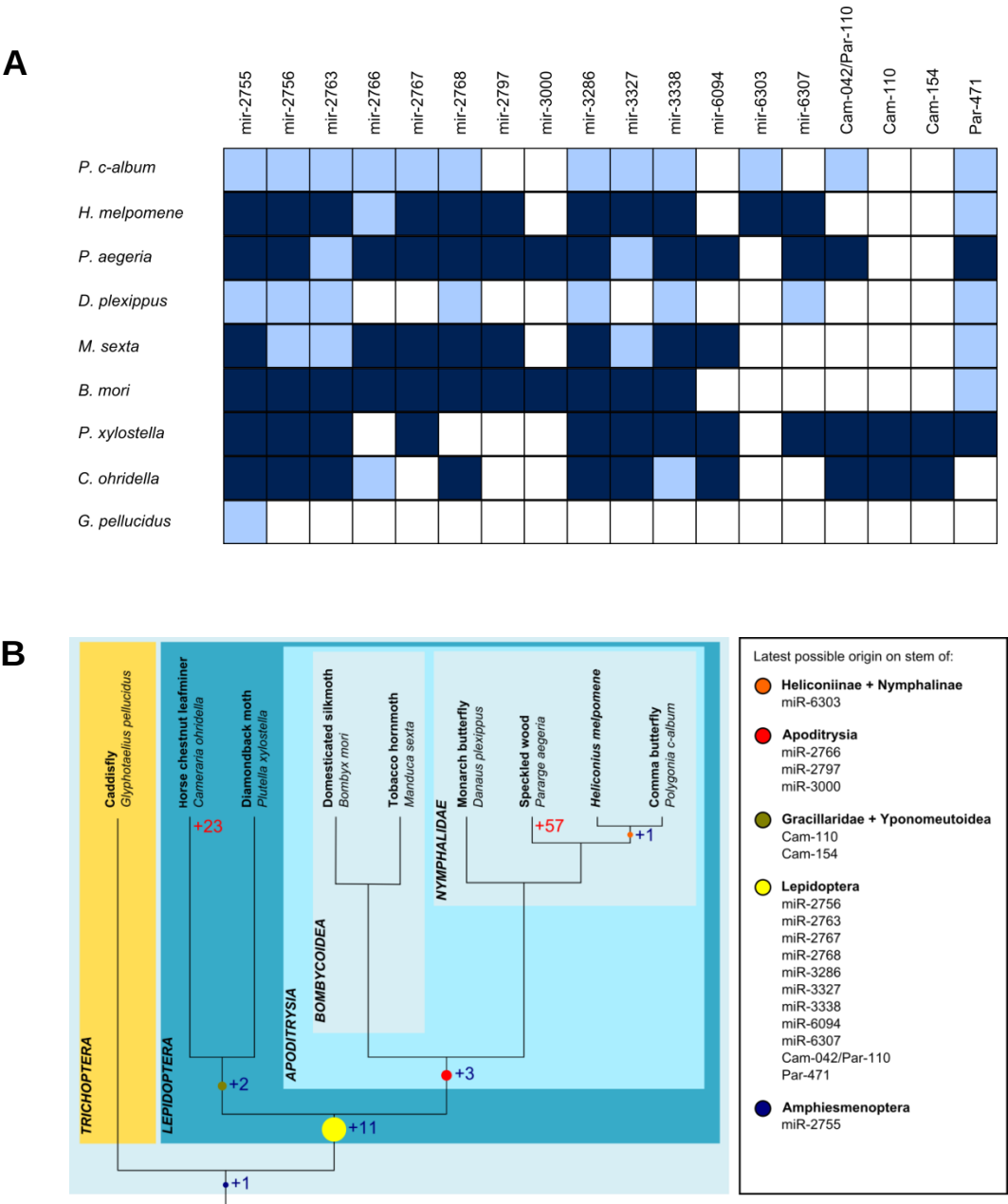


Figure 3.3

- (A) Presence/absence matrix for each Lepidopteran-specific miRNA family. Light blue squares indicated BLASTn and structural predictions of presence. Dark blue squares indicate evidence from deep sequencing.
- (B) Phylogenetic distribution of miRNAs in the Lepidoptera and its sister group Trichoptera showing inferred miRNA gains at each node. Lineage-specific gains are only plotted for *C. ohridella* and *P. aegeria*.

4.6 Expression level analysis

I asked whether evolutionary age correlated with expression level, since I hypothesized that older miRNAs might be more deeply integrated into developmental pathways and more widely utilised, and existing studies have provided evidence for a correlation between expression level and miRNA family age (Meunier et al., 2013). To simplify this analysis, I divided miRNAs into three categories depending on evolutionary age. Sequences with their origins dating back to before the last common ancestor of Lepidoptera and Trichoptera, including sequences widespread in animal evolution, are defined in Category A. Expression levels for 10 randomly-selected sequences from category A, each dating back to the last common bilaterian ancestor, are plotted in Figure 3.4. Category B contains the 18 lepidopteran novelties shared by more than one species, plus miR-2755 which is present in Lepidoptera and Trichoptera. Category C sequences are the most recently evolved: these are the lineage-specific miRNAs occurring only within the species of interest. I used dominant arm read counts as a measure of miRNA expression.

Both *P. aegeria* and *C. ohridella* display considerable variation in miRNA expression level within each category (Figure 3.4). The mean expression levels for dominant products is significantly different between young and conserved miRNAs in *P. aegeria* (t-test; $P = 0.024$), but not in *C. ohridella* (t-test; $P = 0.35$). Using the total number of reads for both mature products as an indicator of expression level gives no significant difference between expression levels of young and conserved miRNAs for either species (t-test; $P = 0.11$ and 0.35 for *P. aegeria* and *C. ohridella* respectively, see also Chapter 5). The variation in expression levels within each category results in overlap in miRNA expression levels between the three categories. Recently evolved miRNAs can be expressed as highly as many miRNAs of much older evolutionary origin. However, the most extreme cases of high expression are of older origin. miR-1 is the most highly expressed miRNA in both species, with over 100,000 reads per million in *C. ohridella* and over 40,000 reads per million in *P. aegeria*.

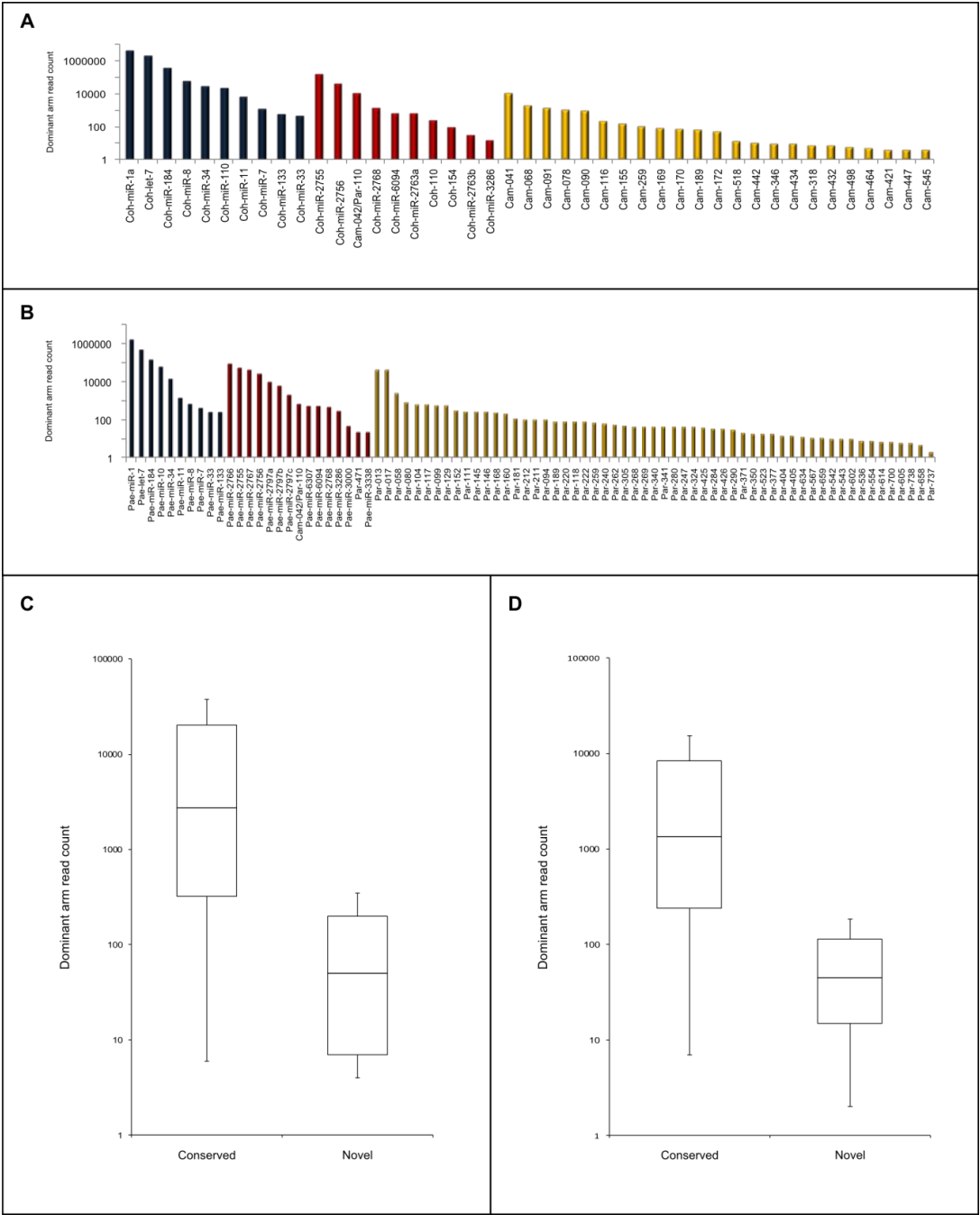


Figure 3.4

Dominant arm expression read counts for 10 randomly selected miRNAs of origin prior to the last common ancestor of Lepidoptera and Trichoptera (blue), all lepidopteran-specific miRNAs (red) and lineage-specific miRNAs (yellow) in (A) *Cameraria ohridella* and (B) *Pararge aegeria*. Boxplots show the distribution of dominant arm read counts for all conserved miRNAs, including lepidopteran-specific miRNAs, as well as lineage-specific novel sequences in (C) *Cameraria ohridella* and (D) *Pararge aegeria*.

5. Discussion

5.1 MicroRNA discovery in *C. ohridella* and *P. aegeria*

Using small RNA sequencing, I identified 113 and 139 miRNA genes in the Horse chestnut Leafminer moth *C. ohridella* and the Speckled Wood butterfly *P. aegeria* respectively. It is likely that these are not fully comprehensive figures, since some conserved ancient miRNAs known from other animals were not detected, perhaps due to highly restricted spatiotemporal expression. While the number of sequences shared with at least one other species is similar in both organisms (90 in *C. ohridella*; 82 in *P. aegeria*), the number of lineage-specific sequences identified in *P. aegeria* (57) is more than detected in *C. ohridella* (23) despite equal sequencing depth in both species. I suggest this results from the difference in the range of developmental stages sampled in each species. *P. aegeria* was extensively sampled at different timepoints for each developmental stage, including embryos; in contrast, I lack data on small RNAs expressed during embryonic development in *C. ohridella*. Consistent with this, research in drosophilid fruitflies has shown that many young, fast-evolving miRNAs are highly expressed during embryogenesis and are downregulated during later development (Ninova et al., 2014b). It is possible that the early embryo provides a more permissive environment for gene expression changes; this suggestion is compatible with ‘phylogenetic hourglass’ models that identify early developmental stages as prone to evolutionary change by incorporation of newly evolved genes (Domazet-Lošo & Tautz, 2010).

Most of the evolutionarily ‘older’ miRNAs have been reported before. This is not surprising since conserved miRNAs have been the focus of much study. I did identify one additional conserved miRNA, Par-471, from *P. aegeria* which I suggest is widely distributed across Lepidoptera, since its precursor sequence could be identified in most lepidopteran genome sequences. This miRNA is also present in a small RNA dataset from the Diamondback moth *P. xylostella* (Etebari et al., 2013). While the precursor has been conserved in sequence and structure across other lepidopterans (Supplement S3.3), it has not been reported in deep sequencing projects in the Apoditrysia. The mature Par-471 miRNA is not highly represented in my pooled *P. aegeria* library suggesting it may have a spatiotemporally restricted expression pattern.

5.2 Birth and death of miRNAs in the Lepidoptera

Assigning miRNA genes to nodes on a phylogenetic tree of Lepidoptera yielded two seemingly paradoxical conclusions. First, there is extensive birth of new miRNAs along terminal branches in the phylogeny generating a plethora of young miRNAs. Second, most of the 'conserved' miRNAs, shared by more than one species of Lepidoptera, originated early in phylogeny, with little evidence for integration of new miRNAs in the 100 million years (Wahlberg et al., 2013) following the initial radiation of the Lepidoptera. How can these two conclusions be reconciled?

The present-day miRNA complement of any given lineage is the net result of the opposing forces of miRNA birth and miRNA death. The birth of new miRNAs may occur through *de novo* generation of new precursors from existing hairpin-like segments of existing transcripts (Campo-Paysaa et al., 2011) or through the duplication and divergence of existing miRNA genes. MicroRNAs may also be lost through mutations that disrupt the formation of a stable stem loop structure. The balance between these forces in shaping miRNA evolution throughout the Metazoa has been a subject of debate. Some authors have argued for a continuous addition of miRNAs to metazoan genomes over time with very low rates of secondary loss (Tarver et al 2013); others have presented evidence for a more fluid model with continuous gains and losses operating on each stem in a clade (Meunier et al., 2013). What is clear, and also supported by my data, is that terminal nodes in a phylogeny tend to have an abundance of young, lineage-specific miRNAs (Lu et al. 2008; Nozawa, Miura, and Nei 2010; Meunier et al. 2013; Lyu et al. 2014). This was reported by Lu and colleagues in drosophilids (Lu et al., 2008) and although this study was subsequently criticised for overestimating lineage-specific miRNA gains through sampling of possible degradation products (Berezikov et al., 2010), evidence for a high miRNA birth rate on terminal branches was also later presented for mammals (Meunier et al., 2013) and confirmed in drosophilids (Lyu et al., 2014; Nozawa et al., 2010). In Lepidoptera, the number of lineage-specific miRNAs identified with available miRNA deep sequencing data are 23 in *C. ohridella*, 21 in *P. xylostella*, 315 in *B. mori*, 8 in *M. sexta*, 57 in *P. aegeria* and 10 in *H. melpomene*. While the exact numbers quoted for each lineage vary (probably due to depth of transcriptome sequencing and rigor of miRNA annotation), this is still indicative of the continuous birth of new miRNA sequences on terminal branches within the lepidopteran phylogeny.

It is estimated that a large proportion of new miRNAs (>94% in drosophilids) will rapidly be lost (Lyu et al., 2014). The quick turnover phase in miRNA evolution is estimated to last for about 30 million years in drosophilids, after which the remaining sequences will have been integrated into gene regulatory pathways. It is widely thought that miRNAs incorporated into such pathways are then unlikely to be lost, although there are exceptions. For example, secondary loss of conserved miRNAs have been noted in tapeworms, which may be linked to morphological simplification (Fromm et al., 2013), and loss of conserved miRNAs has also been documented in mammals (Meunier et al., 2013).

It is likely that the constant generation of new miRNA precursors has been going on throughout lepidopteran evolution. Birth rates of novel miRNAs are unlikely to drop substantially given the low requirements for transcripts to adopt a suitable secondary structure and become integrated into the miRNA processing pathway (Campo-Paysaa et al., 2011; Lyu et al., 2014), though it has been postulated that they may sometimes be increased through increased transcriptional activity in the genome (Heimberg et al., 2008). Despite this constant production of new material for selection to act upon, I find that hardly any new miRNAs have been integrated into lepidopteran miRNA repertoires following the establishment of the deeper branches in lepidopteran evolution. The distribution of lepidopteran-specific miRNAs shows the majority of innovation occurring along the Lepidopteran stem following divergence from the Trichoptera. Only a single sequence discovered so far, miR-2755, was acquired along the common lineage leading to the Lepidoptera and Trichoptera, and few or no conserved sequences were acquired in nodes such as Nymphalidae or Bombycoidea. Hence, I suggest that the concentration of new miRNA sequences close to the base of Lepidoptera, around 100 million years ago (Wahlberg et al., 2013), was not a burst of accelerated miRNA birth, but rather a decrease in miRNA death rate. I argue that early in lepidopteran evolution, a set of new miRNA genes were incorporated into genetic regulatory pathways. Such pathways might include those governing lepidopteran-specific characters such as the patterning of wing and body scales, colour vision or interaction with angiosperms.

5.3 miRNA capture – A new model for the evolution of miRNAs within a clade

Based on the findings I have presented above, I propose a model to describe the turnover of miRNAs across geologic time within a clade. Several authors have already

reported on the tendency for large numbers of novel miRNAs to emerge on terminal branches in a phylogeny (Lyu et al., 2014; Meunier et al., 2013), most of which will subsequently be lost through drift or negative selection. I suggest that under certain circumstances, perhaps due to the influence of an internal or external event resulting in the opening of new niches and the concomitant relaxation of selection pressures, the rate of miRNA death along specific lineages is significantly decreased. Reduction in negative selection against novel miRNAs would allow more alleles to be either neutral in effect, increasing the chances of a neutral miRNA drifting to fixation, or beneficial in effect and positively selected for (Figure 3.5).

The basic tenets of this ‘miRNA capture model’ I am proposing are, firstly, a constant high rate of novel miRNA generation on terminal branches. In most cases, the majority of these new precursors are rapidly eliminated from the genome by a combination of drift and negative selection. On occasional lineages, the relaxation of negative selection permits the persistence of more newly-born miRNAs, effectively ‘capturing’ them and fixing them in the genome. This phenomenon would lead to the perceived bursts of miRNA innovation observed at the base of the Lepidoptera in this study, as well as in vertebrates (Heimberg et al., 2008) and at the base of the Bilateria (Christodoulou et al., 2010; Prochnik et al., 2007).

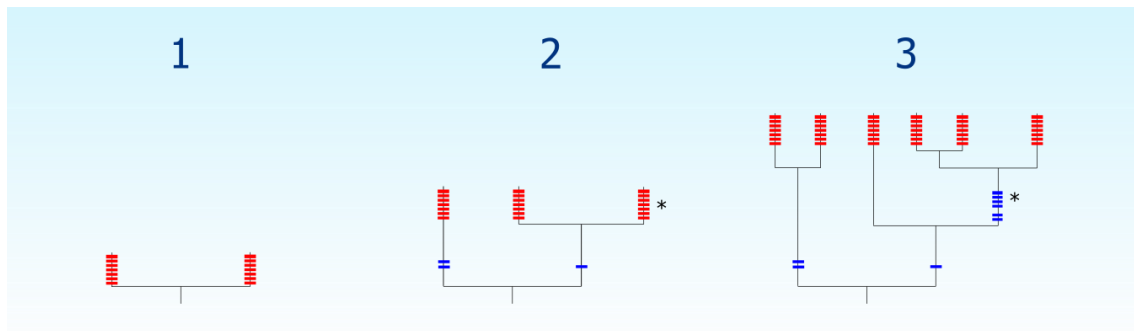


Figure 3.5

Model for miRNA capture. Evolution of the clade proceeds through 1, 2 and 3. Red bars indicate young, lineage-specific miRNAs at the tips of the phylogeny. Blue bars indicate miRNAs which have been fixed by selection over geologic time. On the lineage marked with an asterisk, selection pressures have been reduced, permitting the persistence of a larger proportion of the miRNAs born along that lineage.

6. Supplementary information

Supplement S3.1

List of miRNAs of ancient origin used to verify genome assemblies and miRNA datasets

miRNA name	Origin	Recovered in						
		<i>Glyphotaelius pellucidus</i>	<i>Hepialus sylvina</i>	<i>Cameraria ohridella</i>		<i>Pararge aegeria</i>		<i>Polygonia c-album</i>
				Genome	miRNA-seq	Genome	miRNA-seq	
bmo-let-7	Bilateria	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-bantam	Protostomia	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-lab-4	Arthropoda	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-1a	Bilateria	No	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-2a-1	Protostomia	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-7	Bilateria	Yes	No	Yes	Yes	Yes	Yes	Yes
bmo-mir-8	Bilateria	Yes	Yes	Yes	Yes	No	Yes	Yes
bmo-mir-9a	Bilateria	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-10	Eumetazoa	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-11	Hexapoda	Yes	No	Yes	Yes	Yes	Yes	Yes
bmo-mir-12	Protostomia	Yes	No	Yes	No	Yes	Yes	No
bmo-mir-14	Hexapoda	Yes	Yes	Yes	Yes	Yes	No	Yes
bmo-mir-33	Bilateria	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-34	Bilateria	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-92a	Bilateria	Yes	No	No	Yes	Yes	Yes	No
bmo-mir-124	Bilateria	Yes	Yes	Yes	Yes	Yes	Yes	Yes
bmo-mir-133	Bilateria	Yes	No	Yes	Yes	Yes	Yes	Yes
bmo-mir-184	Bilateria	Yes	No	Yes	Yes	Yes	Yes	Yes
bmo-mir-279a	Protostomia	Yes	No	Yes	Yes	Yes	Yes	Yes
bmo-mir-285	Bilateria	Yes	No	Yes	Yes	Yes	Yes	Yes

(A) List of conserved miRNAs identified in *Cameraria ohridella*

[illegible]

[illegible]

[illegible]

(B) List of novel miRNAs identified in *Cameraria ohridella*

Precursor sequence			5p	3p
Cam-041	gaagaaagaaagaauaauugcuuaauuggguguguaucuaaaucauacuaucuaaggguaagccagcagcauca	gcuugagucuaacuaucuaaaua	cuaguguguaauagggugucacagcgga	
Cam-068	gcuugaagaaacgucuaagaaugucuaauuacuaagggucuguaacccuaacuaucuaaagaaugcuguaaaucauacaa	gcuugugacuaacuaucuaaaua	cuaguguguaauagggugucagagcgga	
Cam-078	aagagaggaagaaagaaugucuaagaaagagaaugucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	agucuaucuaucuaucuaaggguguguaa	ucacugugagagaaugaaaucauauag	
Cam-090	accugucuaucuaagagccuacacacagcgccgucuaagaaagcuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	aagcucucgcccugucuaacuaucua	aguaagugucgagcagcagcuaucua	
Cam-091	auugucugacagcuguaagaaugcuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	gcuuguguaagcugucuaugugugua	caacacuaugacgcuuacacagcgga	
Cam-116	uuucagaaauaauugcauugggccauugucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	caacucuaagaaugaaauag	gaugucuaaauagcugucuaagcagcg	
Cam-155	cuaauuguaucuaagcagaaugaaugcuguaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	caguaagaaugaaugcugucuaucuaucua	gaaauaucugucuaucuaucuaucua	
Cam-169	gaaagaaauaagaaugaaugcuguaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	auugaguuuguaucuaucuaucuaucua	uaaauaucgagcuaacuaucuaucua	
Cam-170	uccuuguuuuuaagaaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	cguuaagaaugaaugcuaucuaucuaucua	aaugucacagcuaucuaucuaucua	
Cam-172	agaaauaucguguaagaaugcuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	ugcugucguguaagaaugcuaucuaucua	gaagggugucuaucuaucuaucua	
Cam-189	uauguccagcaguguguaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	uaaagcuguaucuaucuaucuaucuaucua	caucauaauuugcagcuaucuaucua	
Cam-269	uaacgaagaaagccauugcuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	uccauugaaucugucuaucuaucuaucua	acggaugaaucguguguaugaa	
Cam-318	cgaagccagcugcuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	aucuaucuaucuaucuaucuaucuaucua	agcuguaagaaugcuaucuaucua	
Cam-346	aaacaucuaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	uuugcaacagcuaucuaucuaucuaucua	cuguaucuaagaaugcuaucuaucua	
Cam-421	aaauaaucaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	caacaauuuaaagcuaucuaucuaucua	agaaauaucgucuaucuaucua	
Cam-432	aucaugcuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	ggcgaugugugucuaucuaucuaucuaucua	ggcggguguguguaucuaucuaucua	
Cam-434	uuucuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	ggcgcgucuguguaucuaucuaucuaucua	acugagccacuaucuaucuaucua	
Cam-442	aaauaaucaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	agcugcuaucuaucuaucuaucuaucua	aaugucuaucuaucuaucuaucua	
Cam-447	uuucuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	agcugcuaucuaucuaucuaucuaucua	ggcugucugugucuaucuaucuaucua	
Cam-464	uuucuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	agcugcuaucuaucuaucuaucuaucua	ggcuaucuguguaucuaucuaucua	
Cam-498	uuucuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	aaauaauuugcuaucuaucuaucuaucua	uaagaaugcuaucuaucuaucuaucua	
Cam-518	uaaauaagcagcuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	uaagcagacacgguguaucuaucua	gaaacggcugucuaucuaucuaucua	
Cam-545	ccuuguaucuaucuaucuaucuaucuaucuaaggguguguaagaaugcuguaaaucauacaa	agcugcuaucuaucuaucuaucuaucua	ggcgaugcuaucuaucuaucuaucua	

(C) List of conserved miRNAs identified in *Pararge aegeria*

[illegible]

[illegible]

[illegible]

(D) List of novel miRNAs identified in *Parage aegeria*

[illegible]

[illegible]

Chapter 4:

Functional characterization of the lepidopteran novelty, miR- 2768

Statement of Author Contributions

The work presented in this chapter was submitted together with the main body of material in Chapter 3 as a Research Article in *Molecular Biology and Evolution* (submitted September 16, 2014; Manuscript ID MBE-14-0982). It was accepted for publication on December 30, 2014 after minor revision and first published online on January 8, 2015. The authors on this manuscript are in order: Shan Quah, Jerome H. L. Hui and Peter W. H. Holland (corresponding author).

The original idea for the project, involving the functional testing of the miR-2768:*ci* interaction, was conceived by discussion between all three authors. SQ suggested the addition of incorporating the candidate site into a coding sequence through generation of a fusion protein, performed all experimentation, interpreted the data and drafted the manuscript. PWHH and Adrian L. Smith (University of Oxford) provided equipment, reagents and laboratory space for this study. Ignacio Maeso (Universidad Pablo de Olavide, Spain) provided advice on building the fusion constructs. All authors discussed the results of the study and edited the manuscript for improvement. PWHH supervised the work of SQ throughout the duration of this study.

1. Abstract

A burst of miRNA gain occurred on the stem lineage of the Lepidoptera. Amongst the 11 miRNA genes acquired during this period of evolutionary innovation is *mir-2768*, which is located within an intron of the homeobox gene *invected*. Together with the product of the closely related *engrailed* gene, *Invected* functions in the Hedgehog signalling pathway as a negative regulator of the Gli homologue *cubitus interruptus (ci)*. I show that the dominant product of *mir-2768*, miR-2768-3p, is predicted to bind with almost perfect complementarity within the coding region of the *ci* transcript, and demonstrate that this binding site is conserved across the Lepidoptera. I also experimentally validate this interaction using a luciferase assay system and demonstrate that miR-2768-3p is capable of exerting downregulation on transcripts containing the candidate site regardless of whether the site occurs in the 3' untranslated region or within the coding sequence. This finding represents the first functional evidence for an intronic miRNA performing a complementary function to that of its host gene.

2. Introduction

The persistence of the common cohort of miRNAs gained at the base of the lepidopteran tree in multiple lineages, as described in Chapter 3, is strongly suggestive of selective forces acting to preserve these sequences in the genome. It is likely that these miRNAs have been successfully integrated into essential genetic regulatory circuits in the Lepidoptera, such that any loss due to drift would be selectively disadvantageous. An understanding of the involvement of this shared set of lepidopteran miRNA innovations in developmental genetic pathways would enable greater appreciation of their importance to lepidopteran development, and at the same time offer new perspectives into miRNA gain and its contribution to the evolution of development.

One often overlooked aspect of miRNA biology is the genomic context in which a miRNA gene is situated. As noted in the General Introduction, miRNA genes situated within a host transcription unit are known to show correlated expression patterns with their host transcripts (Baskerville & Bartel, 2005). While similar expression domains open up the possibility for shared regulation of downstream targets between the miRNA and its host gene, experimentally validated examples for this phenomenon have yet to be documented in the literature.

In order to address this question using the lepidopteran miRNA data, I studied the surrounding genomic context for each of the 11 miRNA genes which originated on the common lepidopteran stem. One of these genes, *mir-2768*, is located within an intron of the *invected* (*inv*) homeobox gene. The *inv* gene is closely related to *engrailed* (*en*), and both are likely to have redundant functions (Gustavson, Goldsborough, Ali, & Kornberg, 1996). While the observation that *inv* was host to *mir-2768* was initially reported in the domesticated silkworm (*Bombyx mori*) through a Solexa sequencing study (Liu et al., 2010), I found that this association was conserved in the Tobacco Hornmoth (*Manduca sexta*), the Monarch butterfly (*Danaus plexippus*) and the Speckled Wood butterfly (*Pararge aegeria*). Target prediction analyses carried out for miR-2768 reveal the presence of a candidate site with near-perfect complementary base pairing to miR-2768-3p located within the coding sequence of the *cubitus interruptus* (*ci*) gene. Using luciferase assays, I demonstrate that miR-2768 is capable of downregulating mRNAs containing the candidate target site, both when the site is present in the 3' UTR and when it is incorporated into coding sequence. As both *ci* and

inv are involved in the Hedgehog signalling pathway (Motzny & Holmgren, 1995), I postulate that miR-2768 contributes to this interaction, perhaps through reinforcement of repression by *enlinv* on *ci*.

3. Materials and Methods

3.1 miRNA target predictions

Transcriptome data for *H. melpomene* was downloaded from Ensembl Metazoa. The PITA target prediction algorithm (Kertesz et al., 2007) with default parameters was run on all contigs using a cutoff score of -20. This algorithm computes the difference in free energy gained upon pairing of the miRNA to its target and free energy lost in opening up RNA secondary structure in order to make the site accessible to miRNA binding.

3.2 Sequence alignments

Alignment of *ci* nucleotide and protein sequences was carried out using Clustal Omega (Sievers et al., 2011) as well as the Geneious R6 MAFFT plugin (Kato, 2002).

3.3 Vector construction

A synthetic oligonucleotide duplex was constructed consisting of the predicted miR-2768 target site (22 bp) from the *P. aegeria ci* gene flanked by restriction sites and inserted into the multiple cloning site of the psiCHECK-2 vector (Promega, C8021) downstream of the *hRluc* stop codon to make the psiCHECK-2:ci-UTR construct. To generate a construct, psiCHECK-2:ci-CDS(WT), with the putative *ci* target site within its coding region, a 369 bp region of the *ci* exon containing the site was cloned and inserted immediately preceding the stop codon of *hRluc*. A mutated version with synonymous transition mutations within the candidate site was generated by PCR to give the psiCHECK-2:ci-CDS(M) construct. All constructs were validated by sequencing. For construction of the miR-2768 expression vector, a 339 bp segment containing the miRNA precursor was amplified by PCR from *P. aegeria* genomic DNA and inserted into the multiple cloning site of pCS2+.

All primers and oligonucleotides used in vector construction are listed in Table 4.1. Details of the protocol used for generation of the *hRluc:ci* fusion constructs are provided in Supplement 4.1.

3.4 Luciferase assay protocol

HeLa cells (ATCC CCL-2) were cultured in 25 cm² flasks at 37°C, 5% CO₂ in Dulbecco's modified Eagle's medium (Sigma-Aldrich, D5671) supplemented with 10% fetal calf serum, 2 mM L-glutamine and penicillin/streptomycin. Cells were passaged and transferred to 96-well assay plates (Sigma-Aldrich, CLS3603) 36 h prior to transfection. The time between plating and transfection was optimised experimentally for maximal transfection efficiency. Transfection was carried out using FuGENE HD (Promega, E2311). Each transfection reaction was set up according to the manufacturer's protocol in 100 µl using a total of 2µg of plasmid DNA and 6 µl of FuGENE HD. Luciferase assays were performed 36 h after transfection with the Dual-Glo Luciferase Assay System (Promega, E2920). Firefly and Renilla luciferase expression was quantified by measuring luminescence with no emission filter on the FLUOstar Omega microplate reader (BMG LABTECH). For each assay, three biological replicates (two for psiCHECK-2:3'-UTR without miR-2768) on separate plates were performed at separate times and relative Renilla and firefly luciferase activities were recorded. Each biological replicate consisted of three or five technical replicates. The mean Renilla:firefly ratio for pCS2+:2768 co-transfected with blank psiCHECK-2 was used to normalise readings across plates.

Table 4.1

List of primers used to generate plasmid constructs used in the luciferase assay to test the interaction between miR-2768 and its candidate target site in *ci*

Oligo name	Sequence	Remarks
Pae-ci-2768.F	GTGCATCTCGAGATGATGCAATATCTTAACCAATGCGCCGCTACGTG	Hybridised with Pae-ci-2768.R to generate the <i>ci</i> target site
Pae-ci-2768.R	CACGTAGCGCGCCGATTGGTTAAGATATTGCATCATCTCGAGATGCAC	Hybridised with Pae-ci-2768.F to generate the <i>ci</i> target site
Pae-2768_F1	TGAAAAACACTTATTGGTCTTCCCT	Forward primer used to amplify pre-miR-2768 for cloning into pCS2+
Pae-2768_R1	ACAGAAATAGCAACACAAATCAGCC	Reverse primer used to amplify pre-miR-2768 for cloning into pCS2+
CiIF_1	TGATCTGATCGGAATGGGTAAAGTC	Primer used to generate <i>hRluc:ci</i> fusion construct
CiIF_2	ACCATCTCTCCTTCAGCGACCTGCTCGTTCTTCAGCACGCGCTCCACGAA	Primer used to generate <i>hRluc:ci</i> fusion construct
CiIF_3	GCGTGCTGAAGAACGAGCAGGTCGCTGAAGGAGAGATGGTGGAGAACAAAG	Primer used to generate <i>hRluc:ci</i> fusion construct
CiIF_4	GTGCATCTCGAGTTAATCTAAAGACTTGGAGGCTAATAGAGAGCC	Primer used to generate <i>hRluc:ci</i> fusion construct
CiIF_5	GACTGATTGAGGTACTGCATCATTTTCATCTGGTATCACCAGCTTG	Primer used to mutate <i>hRluc:ci</i> fusion construct
CiIF_6	ATGATGCAGTACCTCAATCAGTCAATATTGGGAACCGACTCGGTAAACACCAGC	Primer used to mutate <i>hRluc:ci</i> fusion construct

4 Results

4.1 *Cubitus interruptus* as a potential target for miR-2768

Given the acquisition of new miRNAs early in lepidopteran evolution, I sought to identify target genes. This analysis is hampered by the lack of well-annotated lepidopteran transcriptomes with documented 3' UTRs, the target of many animal miRNAs. However, applying the PITA target prediction algorithm (Kertesz et al., 2007) to annotated coding sequences from the *H. melpomene* transcriptome (Hines et al., 2012) revealed an almost perfectly paired match between miR-2768 and the *cubitus interruptus* (*ci*) transcript. Subsequent analyses of the corresponding region in other lepidopterans reveal that this site is highly conserved between butterflies and moths (Figure 4.1), with 21 of the 22 bases (including the seed sequence) in miR-2768-3p capable of pairing with the consensus candidate site. This complementarity is greater than that observed between most bilaterian miRNAs and their targets. The miR-2768-3p sequence itself is also completely conserved in all lepidopteran species studied (Supplement S4.2) and its precursor gene lies in an intron of the *invected* (*inv*) gene in both moths and butterflies (Supplement S4.3).

To gauge if the candidate site in *ci* is identical between species by chance, or as a by-product of conservation of amino acid sequence, I assessed the overall conservation of the *ci* gene across Lepidoptera. Using genome data for three lepidopterans (*B. mori*, *D. plexippus* and *H. melpomene*), I aligned predicted *ci* coding sequences and compared nucleotide and deduced peptide sequences (Figure 4.1). This revealed that while the peptide sequence of Ci is divergent between different insect orders (Supplement S4.4) and only aligns well in the region containing the zinc finger domains, in Lepidoptera it is more conserved across much of its length (Figure 4.1, Supplement S4.5). Consistent with selection acting at the amino acid level, the corresponding nucleotide sequence is weakly conserved with the majority of the coding sequence showing evidence of third codon position wobble. However, there are two clear regions in the predicted *ci* cDNAs where the nucleotide sequence is almost completely conserved, and strikingly one of these contains the putative miR-2768-3p target site. This implies there is a reason other than amino acid constraints that has forced conservation of local sequence, consistent with conservation as a target site for interaction with miR-2768-3p.

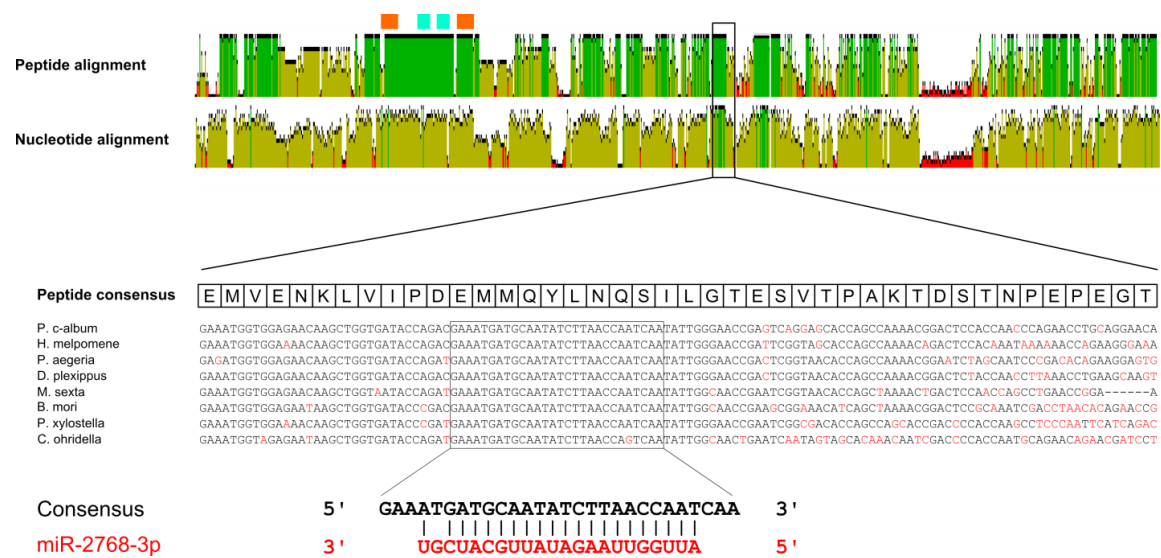


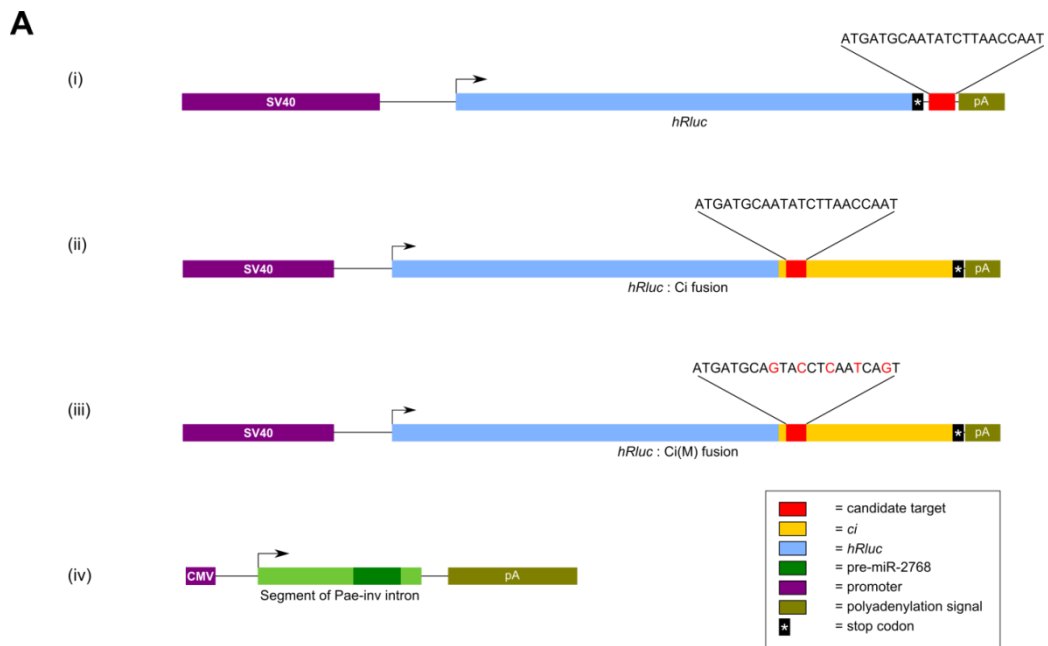
Figure 4.1

Conservation of the peptide (top graph) and nucleotide (lower graph) sequence across the length of the ci gene in 8 lepidopteran species. Positions with 100% mean pairwise identity between all pairs in the column are coloured in green. Brown indicates positions with identity under 100% but above 30%. Red indicates <30% identity. The orange boxes show the positions of zinc finger C2H2 domains and blue boxes indicate the position of the zinc finger double domains. A closeup is shown of the nucleotide sequence alignment in the region of the putative miR-2768 binding site. Residues differing from the consensus are coloured in red.

4.2 miR-2768 is capable of exerting regulation over the *ci* candidate site

In order to functionally test the interaction of miR-2768-3p with its predicted target, I cloned the consensus 22 bp candidate site from lepidopteran *ci* into a luciferase expression vector downstream of the *hRluc* Renilla luciferase stop codon (Figure 4.2A(i); construct psiCHECK-2:ci-UTR). The vector also contains a synthetic firefly luciferase gene (*hluc+*) as a control for transfection efficiency. I also generated an expression construct for miR-2768 by cloning a 339 bp fragment of the *invected* intron containing the miR-2768 precursor sequence under a constitutive promoter (Figure 4.2A(iv); construct pCS2+:2768). Following cotransfection, the ratio of Renilla:firefly luciferase expression is used as a measure of miRNA activity (Figure 4.3).

I detect a very strong (~37-fold) downregulation of Renilla luciferase expression in the presence of both miR-2768 and its candidate site from *ci*. In the absence of the miRNA, no downregulation is observed. This confirms that the miR-2768 precursor region is capable of being processed into a functional miRNA in cells, and that this downregulates transcripts containing the predicted binding site, such as lepidopteran *ci* gene.

**B**

Peptide:

M	M	Q	Y	L	N	Q
---	---	---	---	---	---	---

Wild type: ATG ATG CAA TAT CTT AAC CAA

Mutant: ATG ATG CAGTACCTCAATCAG

Figure 4.2

- (A) Constructs used for functional validation of *Pae-miR-2768* binding to its predicted site in *Pae-ci*. (i) psiCHECK-2:*Ci*-UTR, (ii) psiCHECK-2:*Ci*-CDS(WT) (iii) psiCHECK-2:*Ci*-CDS (M), (iv) pCS2+:2768
- (B) Synonymous mutations used to alter the predicted *miR-2768* binding site in the *hRluc:ci* fusion construct.

4.3 miR-2768 is capable of downregulating the *ci* candidate site within a coding sequence

As the identified candidate site lies within the coding sequence, I further tested the ability of miR-2768 to downregulate a target when the site occurs within a coding region. To do this, I modified the *hRluc* gene on the psiCHECK-2 vector to encode a fusion protein between *hRluc* and the *ci* exon containing the target site (Figure 4.2A(ii); construct psiCHECK-2:Ci-CDS). A mutant version of this construct was also generated with five synonymous transition mutations introduced into the candidate target site (Figure 4.2A(iii); construct Ci-CDS(M), Figure 4.2B). These mutations should ensure that the protein sequence is not affected, while simultaneously disrupting miRNA binding. Both constructs were assayed for Renilla:firefly luciferase activity in the presence and absence of miR-2768. I detect downregulation of expression from the wild type construct in the presence of miR-2768, but there is no downregulation of the mutant fusion construct (Figure 4.3). A slight increase in Renilla:firefly luciferase activity was observed in the mutant construct relative to the wild type irrespective of the presence of miR-2768. Overall, these experiments support my hypothesis that miR-2768 is capable of binding to the target sequence from the *ci* gene and exerting downregulation within coding sequences.

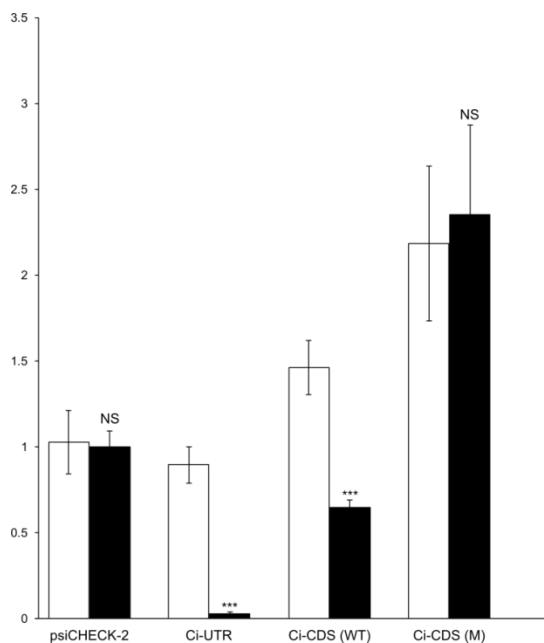


Figure 4.3

Pooled luciferase assay data for the candidate miR-2768 target site in the presence (black) or absence (white) of miR-2768, and for wild type and mutant *hRluc:ci* fusion constructs in the presence or absence of miR-2768. Error bars represent standard deviation. *** $P < 0.001$; NS = not significant.

5 Discussion

To understand how novel lepidopteran miRNAs were incorporated into genetic regulatory pathways, it is necessary to identify candidate target genes which these miRNAs regulate. The majority of identified target sites for metazoan miRNAs lie within 3' UTRs, possibly due to functional constraints (Grimson et al., 2007; Gu et al., 2009). This phenomenon may be somewhat exaggerated, however, as many miRNA target prediction programs consider only 3' UTRs in predictions of miRNA binding (Lewis et al., 2005; Maragkakis et al., 2009; Wang & El Naqa, 2008). Indeed, miRNA target sites located within coding sequences have been documented in animals (Forman, Legesse-Miller, & Coller, 2008; Tay, Zhang, Thomson, Lim, & Rigoutsos, 2008), including humans (Forman et al., 2008), and have recently been suggested to be a common occurrence in cnidarians (Moran et al., 2014). There is some evidence suggesting that miRNA binding in coding regions may require higher sequence complementarity between the miRNA and its target, and proceed via RISC-mediated cleavage (Duursma et al., 2008).

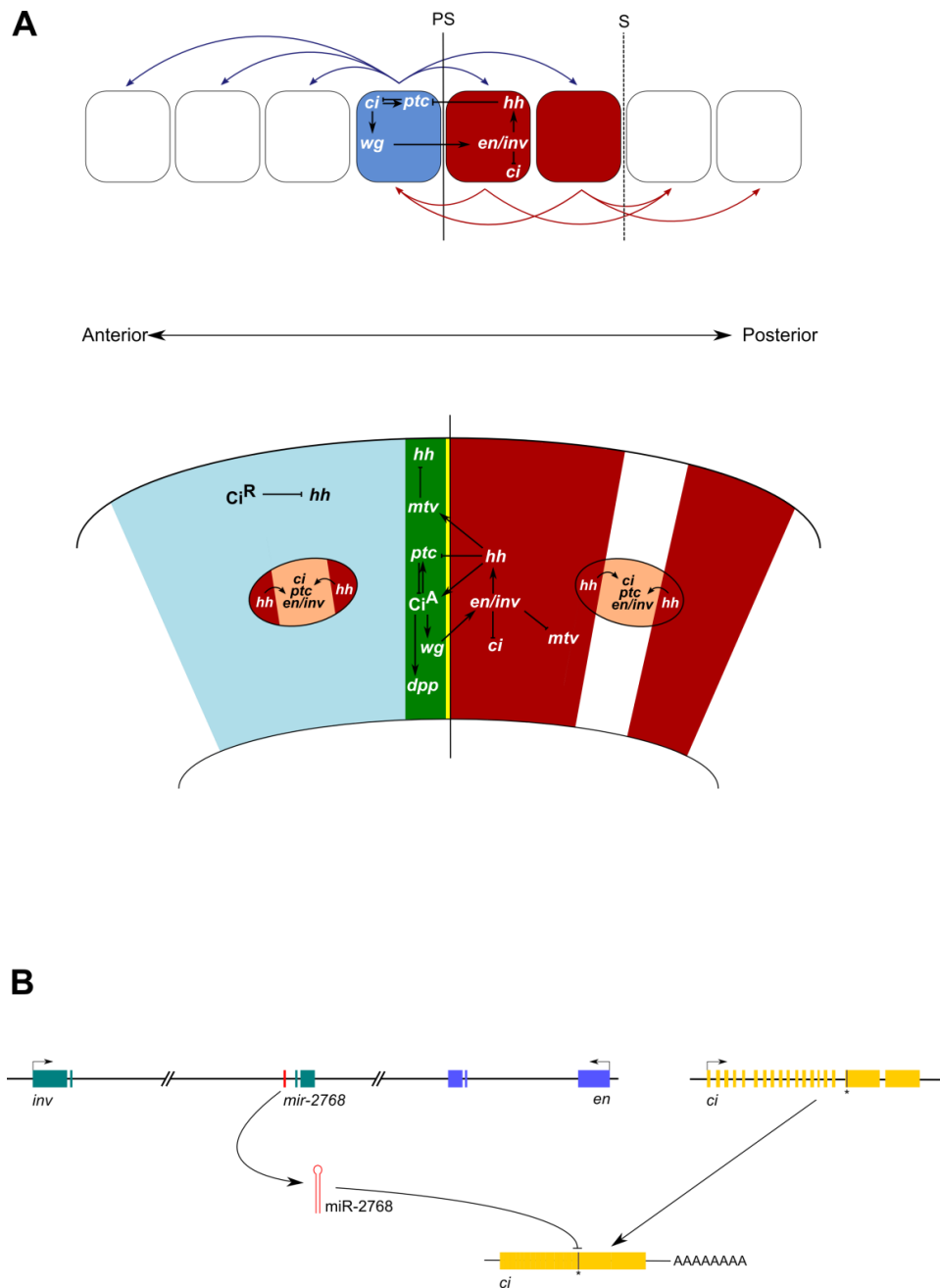
I found evidence that the transcript from the *ci* (*cubitus interruptus*) gene is a target for downregulation by one of the novel lepidopteran miRNAs, miR-2768, within its coding sequence. All lepidopteran *ci* cDNAs contain a sequence which is an almost perfect complement to miR-2768-3p. This candidate target is strikingly well-conserved at the nucleotide level in all species analysed, even though the majority of the *ci* coding sequence is not, suggestive of selection for function on its nucleotide sequence rather than only on its peptide sequence. Luciferase assays conducted in cell culture also indicate that this target site is capable of being targeted and downregulated by miR-2768, both when located within the 3' UTR as well as when present as part of the coding sequence. I suggest this miRNA-target interaction arose early in lepidopteran evolution, since I find that the miR-2768 precursor sequence is present in all lepidopteran genomes examined, but absent in the caddisfly *G. pellucidus*. I also discovered that the *mir-2768* gene itself is located in an intron of the homeobox gene *invected* (*inv*), in the same transcriptional orientation.

I postulate that miR-2768 serves to downregulate the expression of *ci* during butterfly and moth morphogenesis, possibly during segmentation of the embryo or during postembryonic patterning of the wing primordia, or indeed both processes. *Ci* is the insect homologue of the vertebrate Gli protein family and functions as an effector of

Hedgehog (Hh) signalling. The Hh signalling pathway is essential in the establishment of compartment boundaries during anteroposterior segmentation of the insect embryo (Ingham & McMahon, 2001), subsequent anteroposterior patterning of the cuticle in *Drosophila* (Sanson, Alexandre, Fascetti, & Vincent, 1999), as well as in the wing imaginal disc during late larval and pupal stages in insects (Ingham & McMahon, 2001; Keys, 1999). Common to these systems is a regulatory circuit (Figure 4.4) whereby the Hh ligand, expressed in posterior compartment cells under the influence of Engrailed (En) and its paralogue Invected (Inv), signals anteriorly across the compartment boundary to activate *wg* expression. The transcription factor acting downstream of the Hh signal to induce expression of *wg* and other downstream genes is Ci. Wg signals back across the compartment boundary to maintain expression of *en/inv*. In the *en/inv* domain, *ci* is not expressed due to repression by En/Inv. Regulation of Ci activity is essential as it mediates many of the downstream actions of Hh. This is achieved both at the transcriptional level through the actions of En/Inv in repressing *ci* transcription, as well as through post-translational modifications to the Ci protein which regulate its localisation and activity. I propose that in the Lepidoptera, miR-2768 functions to add another level of regulation over the Ci activity at a post-transcriptional level. As miR-2768 is located within the transcription unit of *inv*, it is likely to be coexpressed with *inv* (Baskerville & Bartel, 2005; Rodriguez et al., 2004) and may reinforce the repression by *en/inv* on *ci*. This would represent a coherent regulatory loop (refer to Figure 1.8 in the General Introduction) and would confer robustness to the system through promoting efficient repression of *ci*.

There is another system where Hh signalling has been co-opted for use in patterning where the gene interactions described above have been modified. Many butterflies and moths have large, brightly coloured eyespots on their wings. Eyespot morphology and genes involved in determining eyespot colour and pattern may differ between species (Brunetti et al., 2001; Oliver, Tong, Gall, Piel, & Monteiro, 2012), and the regulatory networks behind eyespot focal specification have only been studied in detail only in Nymphalidae (Oliver et al., 2012). In their report, Monteiro and colleagues propose that a network co-option event occurred in the evolution of eyespot focal patterning and was subsequently modified in different lineages to generate novel patterns. In *Bicyclus anynana* and *Junonia coenia*, focal development first involves an accumulation of *hh* in cells flanking developing foci. Hh signalling subsequently induces expression of *ptc*, *ci* and *en/inv* in cells which will form the focus (Keys, 1999; Figure 4.4). This process of eyespot specification occurs regardless of whether the focus develops in the anterior or posterior compartment, implying a relaxation of compartmental restrictions. Also

relaxed must be the transcriptional repression of *ci* by *en/inv* as both are coexpressed in the same cells. It has been noted that some miRNAs are coexpressed with their target genes, and it has been postulated that their role may be to subtly modulate target mRNA levels.

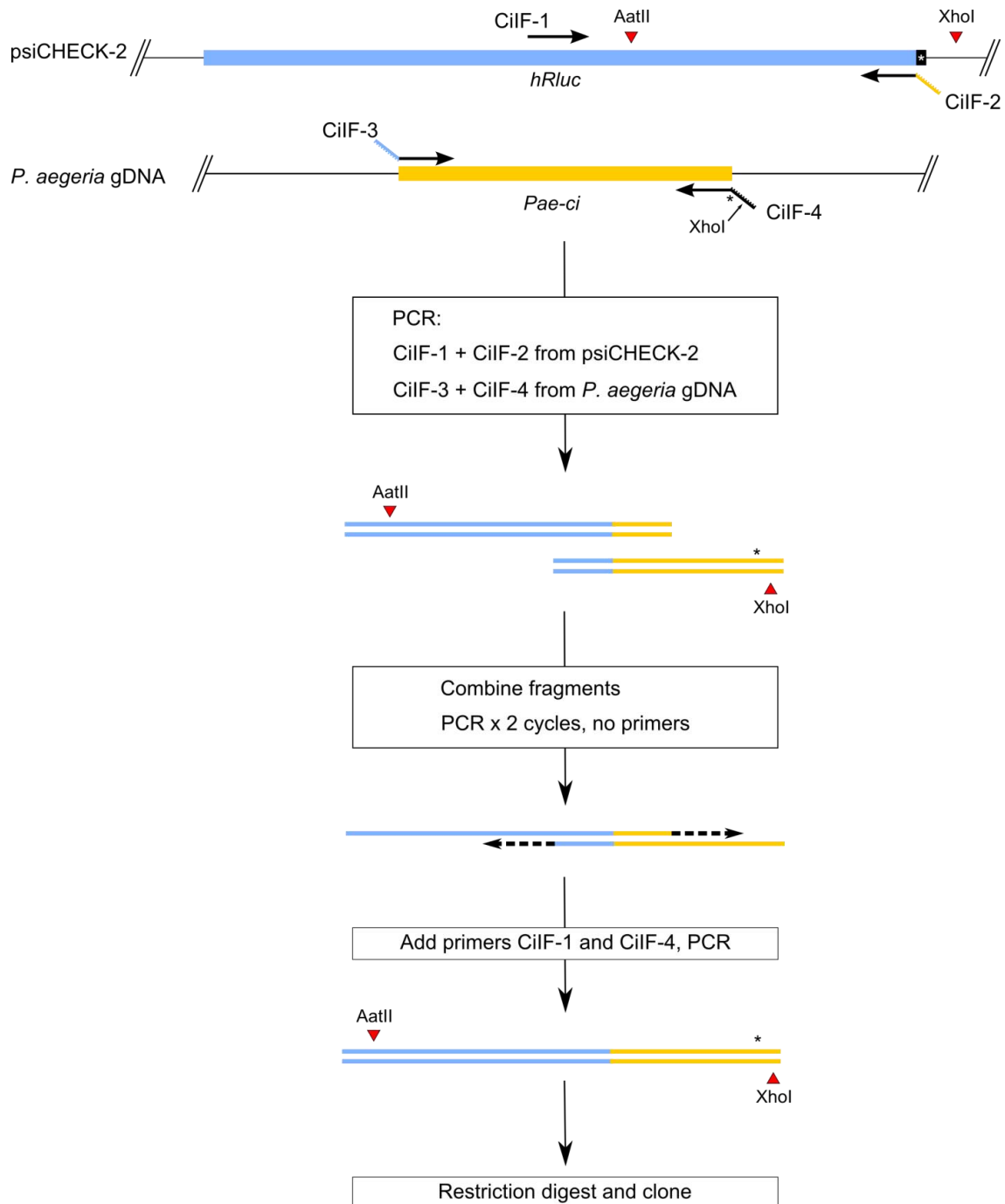
**Figure 4.4**

- (A) Involvement of *ci* in the Hedgehog signalling pathway during anteroposterior patterning of the *Drosophila melanogaster* embryo (above) and wing imaginal disc patterning in *Precis coenia* (below). PS = parasegment boundary. S = segment boundary.
- (B) Model for the contribution of miR-2768 to the repression of *ci*. The asterisk indicates the position of the miR-2768-3p binding site in the coding region of the *ci* messenger RNA.

6 Supplementary Information

Supplement S4.1

(A) Schematic for generation of the *hRluc:ci* wild type fusion construct. Asterisks indicate positions of stop codons. Red triangles indicate restriction sites.



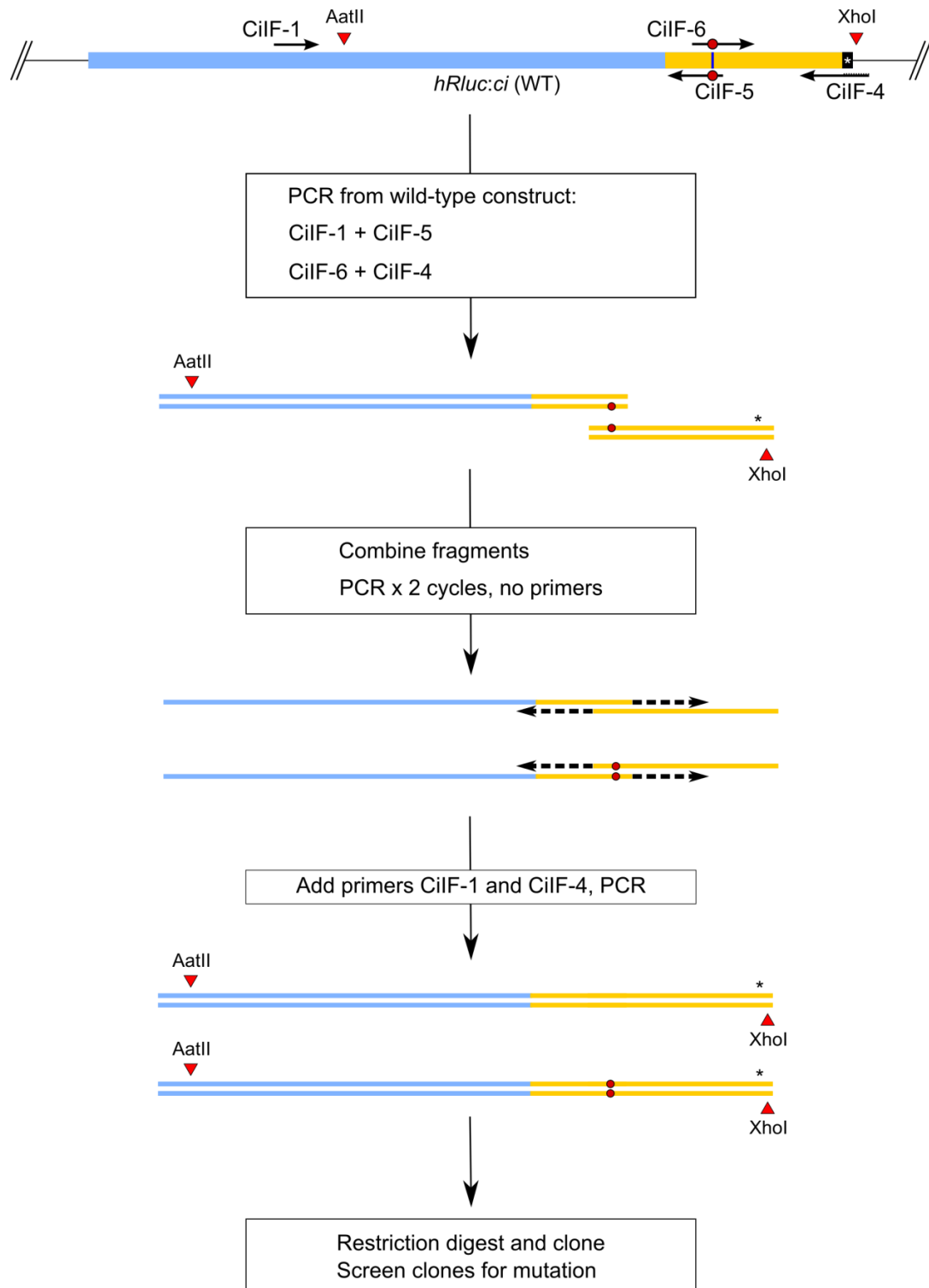
Protocol for generation of the *hRluc:ci* wild type fusion construct

1. Amplification of a 706 bp segment from the 3' end of the *hRluc* gene in the psiCHECK-2 plasmid, excluding the stop codon
 - Primer CiIF-1 is located 5' of an AatII restriction site within *hRluc*
 - Primer CiIF-2 contains a 20 bp extension corresponding to the first 20 bp on *Pae-ci*
2. Amplification of a 369 bp segment of the *Pae-ci* exon containing the target site to be tested
 - Primer CiIF-3 contains a 20 bp extension corresponding to the last 20 bp on *hRluc*, excluding the stop codon
 - Primer CiIF-4 contains an extension carrying a stop codon (TTA), an XhoI restriction site, and flanking sequence
3. Gel purification of products from step (1) and (2) followed by 1:100 dilution in dH₂O
4. 60 µl PCR reaction set up using 5 µl of each diluted product. No primers used.
 - PCR parameters: 2 cycles, 45 s annealing time, 5 min extension time

Component	Concentration	Volume
RedTaq ReadyMix™ reaction mix	2X	30 µl
Product (1)	Diluted 1:100 from gel extraction	5 µl
Product (2)	Diluted 1:100 from gel extraction	5 µl
dH₂O		12 µl

5. 3 µl of each of primers CiIF-1 and CiIF-4 added, standard PCR protocol run for 35 cycles at T_m = 58°C
6. PCR product gel-purified and cloned into pGEM-T Easy Vector System for subcloning into psiCHECK-2

- (B) Schematic for generation of the *hRluc:ci* mutant fusion construct. Asterisks indicate positions of stop codons. Red triangles indicate restriction sites. Red circles indicate the position of the five synonymous transitions introduced to mutate the candidate site.



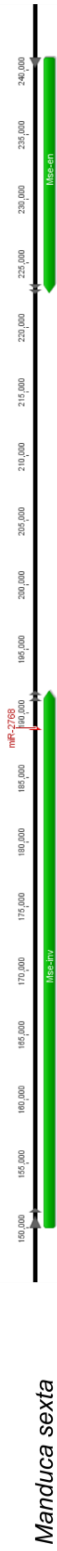
Protocol for generation of the *hRluc:ci* mutant fusion construct

1. Amplification of a 777 bp segment from the *hRluc:ci* fusion gene containing the candidate site at its 3' end
 - Primer CiIF-5 contains the five transition mutations at its 5' end
2. Amplification of a 324 bp segment from the *hRluc:ci* fusion gene containing the candidate site at its 5' end
 - Primer CiIF-6 contains the five transition mutations at its 5' end
3. Gel purification of products from step (1) and (2) followed by 1:100 dilution in dH₂O
4. 60 µl PCR reaction set up as described for generation of the wild type construct
5. Primers CiIF-1 and CiIF-4 added as described for generation of the wild type construct
6. PCR product gel-purified and cloned into pGEM-T Easy Vector System. Clones are screened for acquisition of the five transitions in the candidate site prior to selection for subcloning into psiCHECK-2

Alignment of lepidopteran miR-2768 sequences

[illegible]

Supplement S4.3
Position of *mir-2768* relative to *inv*



Supplement S4.4

Clustal Omega alignment of Ci peptide sequences across insect orders

CLUSTAL O(1.2.0) multiple sequence alignment

```

Manduca sexta          -----
Drosophila melanogaster -----MDAYALPTYFPLAYSELQFLASRRAAVAAAATVLPGSPCINQHHPTDVS
Tribolium castaneum     MPEKEADFQAAMSAGLPLQFPFAAFHAP-----LPV---DQ---RTHEG

Manduca sexta          -----
Drosophila melanogaster SSVTVPSIIPTGGTSDSIKTSIQPQICNENTLLGNAGHQHNHPQHVH---NINVTGQPH
Tribolium castaneum     RYVWDPRMQPPGAP-----FHHPTPPGVSGSPGSLQLLGR

Manduca sexta          -----GLEYLS
Drosophila melanogaster DFHPAYRIPGYMEQLYSLQRTNSASSFHDPIVNCASAFHLAGLGLGSADFLGSRGLSSLG
Tribolium castaneum     ELHPAYRLPPHMEYLYSLQHSTAN-----SIHGLGLS-PEYLSARGLSDLH
                                **.

Manduca sexta          AARSLHPELHAGSTLASQDFQLSLDGKKHSSGSRIASPTRLRSLSGGAIGASANRKRASVSW
Drosophila melanogaster ELHNAVAAAAAGSLASTDFHFSVDGNRRLLGS-----PRPPGGSIRASISRKRALLSS
Tribolium castaneum     -----PSSALASSEFPFSIDGSR-LNS-----PRP--GSIR--QSRKRALLSS
                                .:*** : * :*:*. : . * * :***:

Manduca sexta          SPYSAESLDLAAVIRASPASLAVRA---PSAASTGSYGHLSAGAISPALSLSHASLAQQ
Drosophila melanogaster SPYS-DSFDINSMIRFSPNSLATIMNGSRGSSAASGSYGHISATALNPMSHVHS-TRLQQ
Tribolium castaneum     SPYS-DSFDINS-IRYSNSLVLV-NGSR-SSSASGSYGHLSAGAISPAFNMHPHLLHQ
                                **** :*: : * * * * . *:::*****: * :. * :

Manduca sexta          LLARGGVVGSSGVLGGLVLDPAHQAAAAA---AHHAHAHLVAGI-HRHISSTPQLL
Drosophila melanogaster IQAHL-LRASAGL----L-NPMTQQVAASGFSIGHMPTSASLRVNDVHPNLSDSHIQI
Tribolium castaneum     LQAHL-MRSA-GS-----L-LPLPPATPTHSMYSLGHHPLHVSSPHNISKTD-----
                                : * : :. * * * : : . * . :

Manduca sexta          MGAP-----VDVRPG--L---GLDGT----PPQHL--QQPPQQSEVTSIMEADRRKSPQ
Drosophila melanogaster TTSPTVTVDVSVQVPAAFSLKNLDDAREKKGPFKDVVPEQPSSTSGGVQVEADSSQL
Tribolium castaneum     -----ALN-----DKKKSETGTLTHLSDAQARKI
                                * : : : : : : : : : :

Manduca sexta          -----SLMSHRDN---MHSNKPLSAAAESTVHDGLDSKDEPGDFIE
Drosophila melanogaster SDRCYNNVNNITGIPGDVKVNSRLDEYINCGSISIPSEYDCANADTTDIKDEPGDFIE
Tribolium castaneum     -----KIKK-----EPSNAVSGTINEGQNDQTDLKDDEPGDFIE
                                . : . . * : * * *****

Manduca sexta          TNCHWVDCLEFPTQDDLKVINNDHIHASKKAFVCRWVGCSRDEKPFKAQYMLLVHMR
Drosophila melanogaster TNCHWRSCEIFITQDELVKHINNDHIQTNKKAFCRWEDCTRGEKPFKAQYMLLVHMR
Tribolium castaneum     TNCHWKDCGTEFQTDDELVKHINNDHIHANKKSFCRWGDCSRAEKPFKAQYMLLVHMR
                                ***** . * * * * :*****.***:..*:***** *.* *****

Manduca sexta          HTGEKPHKCTFEGCCKAYSRLNLKTHLRSHTGEKPYTCEYPGCAKAFSNASDRAKHQNR
Drosophila melanogaster HTGEKPHKCTFEGCFKAYSRLNLKTHLRSHTGEKPYTCEYPGCSKAFSNASDRAKHQNR
Tribolium castaneum     HTGEKPHKCTFEGCVKAYSRLNLKTHLRSHTGEKPYTCEYPGCSKAFSNASDRAKHQNR
                                *****

Manduca sexta          THSNEYCSQKPLVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHKGCSRGDDSAESG
Drosophila melanogaster THSNE---KPLVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYANKKHGKPLNDANSRLQ
Tribolium castaneum     THSNE---KPLVCKAPGCTKRYTDPSSLRKHVKTVHGADEFYANKKHGKIDGGSDEG---
                                ***** ** :*****.***.***** ..

Manduca sexta          -----GGGAGSSPRSEEGVPGTARGHASSASVKSESPASPLPHGL---HTPAHL
Drosophila melanogaster QNNSRHNLEHNIIDSSPCSEDHSLGKML--GTSSPSIKSESDISSNNHVLVNGVRASDSL
Tribolium castaneum     -----AAGLDSSPRSEDMQSHKTA--SLSSPSIKSESDVNSPGHQQQGSPGLGATQL
                                ** * * : ** * :*** . *

Manduca sexta          SAQCGGELD---FGGAALGGFGDDNGAPYFRLDGEVEQEVVGEVQGLPLMLRAMVAIGEP
Drosophila melanogaster LTYSRDDLAEN-----LNLDDGWNC-----DDVDVADLPVLRAMVNIGNG
Tribolium castaneum     AGGCNDDFPGDMTTGVPRNGPMDDPAWPYDA-----EDLEIDLPVVLRAVGGGVA
                                . : : * . : :*:*****

Manduca sexta          RAHHHTP-----RFGNK-----MGVGR---RSSDISRKS-
Drosophila melanogaster NASASTIGGSVLARQFRGRLOTKGINSSTIMLCNIPESNRTFGISELNQRITELKMEPG
Tribolium castaneum     VSE-----RNPRTRMKPRLOTKPPS-----NIMSPAQRKNIGITDLNRRINDLMEGT
                                : * : : * : : :

Manduca sexta          -----SQASVVSQAVT-----
Drosophila melanogaster TDAEIKIPKLPNTTIGGYTEDPLQNTSFRNTVSNKQGTVSGSIQGFRRDSQNSTASTY
Tribolium castaneum     VTAQNKT-----QFTEL--QRLQPPGNTQAQIRRD-SNSTVSSY
                                . : *
```

```

Manduca sexta      -----GAAVAGPQRLVVSQHQTVDQLSPDSSRRSSQVSCVGYA-----
Drosophila melanogaster YGSMQSRSSQSSQVS-SIPTMRPNPSCNSTASFYDPISPGCSRSSQMSNGANCNSFTS
Tribolium castaneum    YGSMRSADMSRKSLLASQASSMRPG---MAPGSFYDPI SAGSSRRSSQLSTATTG-----
                        *                :.* * : * .*****.*

Manduca sexta      -----PAPSSALAAVQAVRTSQGNQAVL
Drosophila melanogaster TSGGLPVLNKESNKSLNACINKPNIGVQGVGIYNSSLPPPPSSHLIATNLKRLQRKDS-E-
Tribolium castaneum    -----GTCIPSPSPSHLLAGQLQRLQNNPN-K-
                        * * * * * : * . . .

Manduca sexta      LRGVTCSEVR--AEELALE-----LDPNVQVKEEARRLSEQSNLSD-QAQTYPYP
Drosophila melanogaster YHNFTSGRFSVPSPYMSHLHIKNNKPVGENEFDKAIASNARRQTDVPVNNLPLTNISRF
Tribolium castaneum    ITGGN-GNLVLQQTQNASLQQTALQQT-WMSSNTTASTDARRMSEPCHTLT-----DRK
                        . . . : : * . : : * * :

Manduca sexta      GDAPFFFKTESDQET---VTYKESRSNSTNTV---VITTAQVHHPNQEVNLEQVQVAE
Drosophila melanogaster STTPHSFDINVGKTNNIASSINKDNLRKDLFTVSIKADMAMTSDQHPNERINLDE--V--
Tribolium castaneum    SPPPRPA-----SVTLSPKGTVSSSELHPNQAVVLDE--VGE
                        . * : : . . * * : : * : *

Manduca sexta      GEMVENKLVIPDEMOYLNQSLATESVTPAKTD-SNQPEPENNDKDSATKDNASEIKNT
Drosophila melanogaster -----EELILPDDEMLQYLNLVKDDTNHLEKE-----
Tribolium castaneum    GEMVENKLVIPDEM VHYLNQVADNQNNDFSTAMSWSQKPIPSF-----
                        :.*:****:*** :

Manduca sexta      NNSRDKISDVATSDDSLKLNLAGISDLNISDIQVDLRLSLDVSMGNSGSLLAAKTPDDK
Drosophila melanogaster -----HQAVPVGSNVSETIASN-----
Tribolium castaneum    -----SQTLPSPSSLSHILPSP-----
                        : : :.*. .

Manduca sexta      STPLQEQVIPEVTAEKEKEFKPVTSAVTTNPLQSLQTMATANQTEQSNRIRVSLPQKQNSA
Drosophila melanogaster -----
Tribolium castaneum    -----

Manduca sexta      SPKTVISQNVLSFQNLPHSMLSPQSLPHSAMSQSIMSPHNIPHTVM-SPPSIYNVMS-P
Drosophila melanogaster ----HYR---EQSNIY-----YTN-----KQILT-PPSNVDIQPNT
Tribolium castaneum    ----PSTNVLSFPGNLNQMMPSPSNQA-----NQMMPSPTSNNM-MPSP
                        . * : : : * *

Manduca sexta      QSVMSVMS PQQAMSPQSMQSLMSPQM-PNQMMMSPRHNNIGSPMSQNMASPMVNMA SPM
Drosophila melanogaster TK-FT-----VQDKFA--MTAV---GGSFS-QR---EL-----
Tribolium castaneum    GNQVTQMMPSPGNQLNQMM---PSPSSNFNQIMPSP---SSNFG-AIQNPMLSPAPSM
                        . . : . * : . . :

Manduca sexta      TQNIASPM SHGMPSPMHGQLQSPMRSQNVQQYPVMQQYNQNMMPMMHPMNQNMVYNNQ
Drosophila melanogaster -----S-----TLAVPNEHGHA KESFH
Tribolium castaneum    TQMVPSPMSARCHS-----NQMPMQSPNAQVVPQ-----NCMMPHNQNMHCYNNR
                        . * : : . :

```

```

Manduca sexta      MMNYQQMNPQNQSNQMHP LQMSRSSMM-----SVDNSANMSRGAMNNYCEPQNQ
Drosophila melanogaster -----HQSQKYMNTDIGSKQQSALPSAHQRQTEKSNYNQIIDSSMT
Tribolium castaneum  PMCW-----DNN--QNRPCQANGGYMGNCATHYHNGYVPNG-CASNKPMNYNTYQNPAYN
                        : . * . . * :

Manduca sexta      CPPMQN-PFYNPNMQYPQPPPYNSVVNNTNVMGPPPKNNHQYNQTMNNNQYNNHQRPY
Drosophila melanogaster SLPELNVDSIYP-----R
Tribolium castaneum  CLPSLNEPLPSPAMAAPAPA-----DL-----MNQPHYNQTCYQA--NYNQCVPPQ
                        . * * *

Manduca sexta      NQWDYFGNQFNKHNMQKSMQNSVNMSTGSQKPINGTRQMNCNQIKTDQTDSCMNSLRSQN
Drosophila melanogaster N-----ETENIF-----KVHGDHDNEIQCGIISQSQMSPTNL--NND
Tribolium castaneum  NNCNCNCRN-----PPHKCYH-----QCSRGGNAEIQCKDISQSQMSPGVVA--N--
                        * : : : * * . * . .

Manduca sexta      NQTSEVQVWDISQSQIE-----ASNGRKKNQNTMRQETYQRTLEYVENCENWKSSEMV
Drosophila melanogaster GQFSTVNMQPIITTSKLFPEPQKIVCDTQASNTSVMHLDTYQRTLEYVQSQNWME TNNT
Tribolium castaneum  -----GVQPLGMRQDTYQRTLEYVQNCQSWVSNTTEM
                        : * :*****:.*.* ..

Manduca sexta      SSST-----HP-----LQGGDNMVVNDLQTSLSSEFYEENQYLQMIQ
Drosophila melanogaster STNQIQSLPGMPVNNTLFPDVSSSTHPYHGTNMVINDMTTSLTSLLEENRYLQMMQ
Tribolium castaneum  VTSTTT-----HPL-----AKCGDGTSSNMVNDMTSSLSLLEENRYLQLIQ
                        :. * . ***:**: **:**: ***:**: *

```

Key:

Zinc finger double domain

Zinc finger C2H2 domain

Peptide corresponding to candidate target site

Supplement S4.5**Clustal Omega alignment of Ci peptide sequences across the Dityrisia**

CLUSTAL O(1.2.1) multiple sequence alignment

```

H. melpomene      -----MFGSP
C. ohridella      -----
P. aegeria        -----
P. xylostella     -----
P. c-album        -----FGSR
B. mori           -----
D. plexippus      MPDRESVGGPGTSGGGFLPLQFPSAFAFHASTPPGAATTAMHHATHYHHHAQLAAAAA
M. sexta          -----

H. melpomene      AARQTDHMQYFGRFS-----ESARLLIGLEYLSAARSLHPELHAGSTLASQDF
C. ohridella      -----GLEYLNAARSLHPELHVGSTLASQEF
P. aegeria        -----GLEYLNAARSLHPELHAGSTLASQEF
P. xylostella     -----GLEYLNAARSLHPELHAGSTLASQDF
P. c-album        AARQTDHMQY-----FRRFSESLFTGLEYLSAARSLHPELHAGSTLASQDF
B. mori           -----GLEYLNAARSLHPELHAGSTLASQDF
D. plexippus      AAGATSELSYLAALHPAYRPVPHYDPLYGANTLRGLEYLSAARSLHPELHAGSTLASQDF
M. sexta          -----GLEYLNAARSLHPELHAGSTLASQDF
                      *****.*****.*****.*

H. melpomene      QFSLDG-----SRIASPNRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
C. ohridella      QLSLEGSFETSL---PGSRIASPTRLRLSGGALSASVNRKRAVSWSPYSAESLDLAAVI
P. aegeria        QLSLEG-----SGSRIASQNRRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
P. xylostella     QLSLEG-----GSRIASPTRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
P. c-album        QLSLEGDFTFNIFGIFSGSRIASPNRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
B. mori           QLSLDG-----GSRIASPTRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
D. plexippus      QLSLEG-----SRIASPNRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
M. sexta          QLSLDGKKH-----SSGSRIASPTRLRLSGGAIGASANRKRAVSWSPYSAESLDLAAVI
                      *.**.*          *****.***.*****

H. melpomene      RASPASLAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVVGGSGVLPAGV
C. ohridella      RASPVS LAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVGG-SGVLPAGV
P. aegeria        RASPASLAVRAPSAAEISSFFF-----
P. xylostella     RASPVS LAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGG----SGVLGGGV
P. c-album        RASPASLAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVVGGSGVLPAGV
B. mori           RASPASLAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVVGGSGVLAGGV
D. plexippus      RASPASLAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVGG-SSVLSGGV
M. sexta          RASPASLAVRAPSAASTGSYGHLSAGAI SPALSLSHASLAQQLLARGGVVGGSGVLAGGV
                      ****.*****.***.

H. melpomene      LLDPAHQQAAAAAAHHAHAHLVAGIHR---SHISSPTQLLIGAPVDVRPGLGLDGTTPPQ
C. ohridella      LLDPAHQQAAAAAAHHAHAHLVAGIHR---QHISSPTQLLMG-PLDVKPPYGLDGAGFPQ
P. aegeria        -----
P. xylostella     LLDPAHQQAAAAAAHHAHAHLVAGIH---YRSHISSPTQLLMG-PVDVRPSLNLDTGTPPQ
P. c-album        LLDPAHQQAAAAAAHHAHAHLVAGIHR---SHISSPTQLLIGGPVDVRPGLG---TPPQ
B. mori           LLDPAHQQAAAAAAHHAHAHLVAGIHR---SHISSPTQLLMG-PVDVRQGLGLDGTTPPQ
D. plexippus      LLDPAHQQAAAAAAHHAHAHLVAGIHR---SHISSPTQLLIGGPVDVRNGLGLDGTTPP-
M. sexta          LLDPAHQQAAAAAAHHAHAHLVAGIHR---HISSPTQLLMGAPVDVRPGLGLDGTTPPQ

```

```

H. melpomene      HMQQPP-QQPEITSVMEAD-----RKSPPLISHRD---INKPLSAAAEESTVHDGLD
C. ohridella      HLQQPPPPQQHEVTSIMEAD-----RRKSPQRL-----GNKPLSAAAEESTVHDGLD
P. aegeria        -----SASTALNQRKSPQVLMSHRDNMNSNKPLSAAAEESTVHDGLD
P. xylostella     HLQQPP-QQPEITSIMEAD-----RRKSPQSLMSHRESMHSNKPLSAAAEESTVHDGLD
P. c-album        HMQQPP-QQPEITSVMEAD-----RRKSPPLSAHRENMHSNKPLSAAAEESTVHDGLD
B. mori           HLQQPS-QQPEITSVMEADSAT-GLMQRKSPQGILTHRDNV-SNKPLSAAAEESTVHDGLD
D. plexippus      HMQQPP-QQPEITSVMEADSASTALTQRKSPQVLSHRDNMHGNKPLSAAAEESTVHDGLD
M. sexta          HLQQPP-QQSEVTSIMEAD-----RRKSPQSLMSHRDNMHSNKPLSAAAEESTVHDGLD
                  ****      :      *****

H. melpomene      SKDEPGDFIEQ-----VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
C. ohridella      CKDEPGDFIEQ-----VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
P. aegeria        SKDEPGDFIETNCHWDQVVDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
P. xylostella     CKDEPGDFIETNCHW--VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
P. c-album        SKDEPGDFIEQ-----VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
B. mori           SKDEPGDFIETNCHW--VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
D. plexippus      SKDEPGDFIETNCHW--VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
M. sexta          SKDEPGDFIETNCHW--VDCKLEFPTQDDLVKHINTDHIHASKKAFVCRWVGCSRDEKPF
                  .*****      *****

H. melpomene      KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
C. ohridella      KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
P. aegeria        KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
P. xylostella     KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
P. c-album        KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
B. mori           KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
D. plexippus      KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
M. sexta          KAQYMLVVHMRRTGEKPHKCTFEGCCKAYSRLNLKTHLSHTGEKPYTCEYPGCAKAF
                  *****

H. melpomene      SNASDRAKHQNRTHSN---EKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
C. ohridella      SNASDRAKHQNRTHSNE---QKP YVCKAPGCTKRYTDPSSLRKHVKTVHGPDPFYASKKHK
P. aegeria        SNASDRAKHQNRTHSNE---QKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
P. xylostella     SNASDRAKHQNRTHSNE---QKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
P. c-album        SNASDRAKHQNRTHSNE---QKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
B. mori           SNASDRAKHQNRTHSN---EKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
D. plexippus      SNASDRAKHQNRTHSN---EKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
M. sexta          SNASDRAKHQNRTHSNEYCSQKP YVCKAPGCTKRYTDPSSLRKHVKTVHGAEFYASKKHK
                  *****      :      *****

H. melpomene      GCSRGDDSAESGGGGAGSSPRSEEGV--PLVRGHTSSASVKSESPASPLPMGLHTNAH--
C. ohridella      G-----
P. aegeria        GCSRGDDSAESGGGGAGSSPRSEEGGISMVGRGHTSSASVKSESPASPLPHGLHTPAHQV
P. xylostella     GCNRGDDSAESGGGGAGSSPRSEEGVPMGIRGHTSSASVKSESPASPLNGLHIS-A--
P. c-album        G-----
B. mori           GCTRGDDSAESGGGGAGSSPRSEEGVPPAIRGHTSSASVKSESPASPIPHGLHTSAH--
D. plexippus      GCSRGDDSAESGGGCAGSSPRSEEGGA--LIRGHTSSASVKSESPASPLPLMHTA-AH--
M. sexta          GCSRGDDSAESGGGGAGSSPRSEEGVPGTARGHASSASVKSESPASPLPHGLHT-PA--
                  *

H. melpomene      QLSAQCGGELDFGGSLGGFSDENGAPYFRLDGEVEQEVVGEVQGQLPLMLRAMVAIGEPR
C. ohridella      -LSAHCGGELDYGGSRLLGGFGDDHGGPYFPLDGEVDQEVVE-VG--QLMLRAMVAIGEPR
P. aegeria        QLSAQCGGELDFGVSLGGFSDENGAPYFRLDGEVNLFSLSHFILLPLMLRAMVAIGEPR
P. xylostella     HLSAQCGGELDFGGSLSGFNDENGAPYFRFDGEVEQEVVGEVQGQLPLMLRAMVAIGEPR
P. c-album        -LSAQCGGDLDFGGSLGGFSDENGAPYFRLDGEVEQEVVGEVQGQLPLMLRAMVAIGEPR
B. mori           QLSAQCGGELDFGG-GLGGYRDENGLPYFRLDGEIEQEVVGEVQGQLPLMLRAMVAIGEPR
D. plexippus      QLSAQCGGDLDFGGSLGGFSDENGAPYFRLDGDVEQEVVGEVQGQLPLMLRAMVAIGEPR
M. sexta          HLSAQCGGELDFGGAALGGFGDDNGAPYFRLDGEVEQEVVGEVQGQLPLMLRAMVAIGEPR
                  ***:***:*:*      *.: *. * * * :*:::      :      .      *****

```

```

H. melpomene      AHHPAPRFGNKMVG-----GRTELGG--TNVAVELKSGLPNT
C. ohridella      VQHPAPRLNNKMGP-----RTDLGT--TNI TGD LKVAIPNT
P. aegeria        AHHPMPRFGNKMGLG-----GRTDIGG--TNVAVELKSGVPNT
P. xylostella     AHHTTPRFGNKMGLGR-----TELGTTTGTGTGELKVAIPNT
P. c-album        AH-----GRTEIGG--TNVTVELKTGHPNT
B. mori           AAPHAPRLGNKMALARMPPPHADLGGGELHFRICGVTGRTELGS--TNVGVELKTGASNT
D. plexippus      ---HGPRFGNKMALGRMLPSPVHDMG-----AVQGRTELGS--TNVAVELKTGLPNT
M. sexta          AHHTTPRFGNKMVGGR-----GRTELG--TNVAVELKTGLPNT
                  *.:*   * :   :** .  **

H. melpomene      RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVAITTA---P-RLVTHATVYDQLSPD
C. ohridella      RRDSGISSGSSLYSARSSDVSRKSSQASVISGAVATTGTG VAGQRLLSQHTPVYDQLSPD
P. aegeria        RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVVTTGTG VAGQQLV SQHTAIYDQLSPD
P. xylostella     RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVATTGTG VAGVQVRVMSHHTAVYDQLSPD
P. c-album        RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVATTGTIAGPQRLLSHHSTVYDQLSPD
B. mori           RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVVGG---VGAQRLVAHHTAA YDQLSPD
D. plexippus      RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVATTIGIAGPQRLASQHTNVYDQLSPD
M. sexta          RRDSGISSGSSLYSARSSDISRKSSQASVVS GAVT-GAAVAGPQRLV SQHQT VYDQLSPD
                  *****.****.*****.****.          *      *****

H. melpomene      SSRR-----SSQVSCVGYAPPPSSALAAVQAVRTSQGNQ-AVLLRGVTCSEVRAEE
C. ohridella      SSRR-----SSQVSCVGYAPT PSSALAAVQAVRNSQGNQ-AVFLRGLTCSEVTAGE
P. aegeria        SSRR-----SSQVSCVGYAPPPSSALAAVQAVRTSQGNQ-AVLLRGVTCSEVRAEE
P. xylostella     SSRR-----SSQVSCVGYAPT PSSALAAVQAVRTSQGNQAVLLRGVTCSEVRAEE
P. c-album        SSRR-----SSQVSCVGYAPPPSSALAAVQAVRTSQGNQ-AVLLRGVTCSEVRAEE
B. mori           SSRR-----SSQVSCAGYAPPPSSALAAVQAVRTSQGNQ-----
D. plexippus      SSRRQVFIYYLYKSSQVSCVGYAPPPSSALAAVQAVRTSQGHQ-AVLLRGVTCSEVRAEE
M. sexta          SSRR-----SSQVSCVGYAPAPSSALAAVQAVRTSQGNQ-AVLLRGVTCSEVRAEE
                  ****              *****.****.*****.****.***.*

H. melpomene      LALELDPNVQIKEEARLSEQSNLSDQ-AQGYQPYPCNDDV---GDAPFPFKTED---
C. ohridella      LTLELDPNIQKKEEARLSEQSNLSDH---YRPYP-----GDAPFQFKTEQLDS
P. aegeria        LALELDPNVQIKEEARLSEQSNLSDQ-AQGYQPYPCNDDVSTPGDAPFPFKTED---
P. xylostella     LALELDPNVQKKEEARLSEQSNVSDPAAQGYQPYPCNTDDV---GDAPFPFKTEDVDS
P. c-album        LALELDPNVQKKEEARLSEQSNLSDQ-AQGYQPYPCNDDV---GDAPFPFKTED--D
B. mori           ---ELDPNVQKKEEARLSEQSNLSDQ-TQGYQPYP-----GDAPFPFKTEC--D
D. plexippus      LALELDPNVQKKEEARLSEQSNLSDQ-AQSYQPYPSNDDV---GETPFPFKTED--D
M. sexta          LALELDPNVQKKEEARLSEQSNLSDQ-AQTYQPYP-----GDAPFPFKTES--D
                  *****.*****.*****.*****.*****.*****.*****.*****

H. melpomene      DHEVNYKESRSNSTNTVVITTAQVHHPNQEVNLEQ---VAEGEMVENKLVIPDEMMQYTL
C. ohridella      H-EVQYKDSRSNSTSTVVVTTAQIHHPNQEVNLEQVFSQVAEGEMVENKLVIPDEMMQYTL
P. aegeria        DHEITYKESRSNSTNTVVITTAQVHHPNQEVNLEQV--IXXX-XXXXXXXXXXMMQYTL
P. xylostella     HDAVSYKESRSNSTNTVVITTAQVHHPNQEVNLEQV--QVAEGEMVENKLVIPDEMMQYTL
P. c-album        H-EVNYKESRSNSTTTVVITTAQVHHPNQEVNLEQV---VAEGEMVENKLVIPDEMMQYTL
B. mori           QETVTYKESRSNSTSTVVVTTAQIHHPNQEVNLEQ---VAEGEMVENKLVIPDEMMQYTL
D. plexippus      H-EIAYKESRSNSTNTVVITTAQVHHPNQEVNLEQ---VAEGEMVENKLVIPDEMMQYTL
M. sexta          QETVTYKESRSNSTNTVVITTAQVHHPNQEVNLEQV--QVAEGEMVENKLVIPDEMMQYTL
                  . : ** :*****.****.****.****.*****.*****

H. melpomene      NQSIIGTDSVAPAKTDSTNKKPEGKSEDDNPTKDILSNETQNNNDSDKISDVATSDDS
C. ohridella      NQSIILATESIVAQTIDPTNAEQNDPKRDNKEISANKDRNEIDNENIPSEKISDVATSDDS
P. aegeria        NQSIIGTDSVTPAKTESSNPDETVNNEKDTTAKDILTSEIHNTNNSDDKISDVATSDDS
P. xylostella     NQSIIGTESATPASTDPTKPPNSSDPKDTIDNSATKDKGTSDDVNNSGDKISDVATSDDS
P. c-album        NQSIIGTESGAPAKTDSTNPEPAGTSEEKDTTTKDTTNSEIQNTNNSDDKISDVATSDDS
B. mori           NQSIILATEAETS AKTDSANRPNTEPSLDKDSATKDNTNEQQNTNNSDDKISDVATSDDS
D. plexippus      NQSIIGTDSVTPAKTDSTNL---KPEASKDNDTTKDILTSETQTNNSDDKISDVATSDDS
M. sexta          NQSIILATESVTPAKTDSNQF---EPENNDKDSATKDNASEIKNTNNSRDKISDVATSDDS
                  *****.*: : . . : : . : : * :*****

```


H. melpomene LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLLVPKSPNEKNSPIQEQVIEPEVI-EK
C. ohridella LLKNLGAIGSDLNISDIPVDLRSLDVSMGNSGSLASKSPDEKNTSLQDQVIAEVKTD
P. aegeria LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLASKSLDDKNPPLQEQVIEVTEK
P. xylostella LLKNLGAIGSDLNISDIPVDLRSLDVSMGNSGSLAAKSPDDKNPSLQEQVIEAATDQ
P. c-album LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLVAKSPDDKNTPLEQEQVIEVTEK
B. mori LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLAVKTPEDK-TPLQEQTPEVKAEN
D. plexippus LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLAAKSPDEKNPPLQEQVIEVANEK
M. sexta LLKNLGAIGSDLNISDIQVDLRSLDVSMGNSGSLAAKTPDDKSTPLQEQVIEVTAEK
***** . *: ::* :*:** * . :

H. melpomene CEQEYTKPPTS--NVSTNNPLQSLQTMANQTEQ-SNRMRIINNPL-----PQKP
C. ohridella NEQEYTYNR-SDSTPNQRNPLQSLQCMASQNMKPRRGSS---ENQKPNVSRSPNNM
P. aegeria CEQEFSKPPTS--NIVTNNPLQSLQTMANQTEQ-STRMRVNNPL--PQ-----KP
P. xylostella CEQEKYVKPVKNNTNNQNNPLQSLQSMAGNQTMDSNRMNNHNTIVTQNNQS-ASPMTM
P. c-album CEQDFSKPKS--NASANNPLQSLQTMENQTEQ-SNRMNMNSH--TQKP-A-MSPKTI
B. mori KERE-----T--KKSDPNPLQSLQTMANQTEQ-SNKIRANA-----QKQ-N-KNTK-
D. plexippus CEDYLK--PTA--NNAASNPLQSLQTMANQTEQ-SNRMIRNNSL--PQKP-A-ISPKT-
M. sexta EKE-FK--PVT--SAVTNPLQSLQTMANQTEQ-SNRIRVS--L--PQKQ-N-SASPK-
: . ***** *: .* . : : . .

H. melpomene AMSPKTVVMTPTIMSPNLVHSMSPQSLPHSSMSQSVRSPQHMPHGIMSPPSVYNVMS
C. ohridella VKSPQ--SVPSNVMSQMLPQNMMSPRSLPHSAMSPQSVMSPHHSPNVMSPPNVYNVMS
P. aegeria AMSPKTVVMT-QTIMSPSLAHSMSPQSLPHSSMSQSVRSPQHMPQSIMSPASVYNVMS
P. xylostella VMSPQT-QIPSNVMSQSLAHSMMSPRSLPQSAMSPQSVMS-PHMPHNVMSPPSVYNVMS
P. c-album VMSPKTIVMSPTVMSPNLVHSMSPQSLPHSSMSQSVRSPQHMPHGIMSPPSIYNVMS
B. mori -----TIMSQNILSPQNLPHSMSPQSLPHSAMSPQSVMSPHNIPQNMISPPSAYNVMS
D. plexippus -----IIMTPQTVMSPNLVHSMSPQSLPHSSMSQSVRSPQHMPHGIMSPPSIYNVMS
M. sexta -----TVISQNVLSQNLPHSMSPQSLPHSAMSPQSIMSPHNIPHTVMSPPSIYNVMS
: : : * :.*:***:*.*****: ** . :*** . *****

H. melpomene PQSVMSVMSPQHNASPQSMQSLMSPQ--MPNQM---TMSPRHNNIGSPISQNM-----
C. ohridella PQSAMSVMSPQHNASPQSMQSMMSPOHPMQNQVQSQIMSSRNNNMVSPMSQNMSSPMVN
P. aegeria PQSVMSVMSPQHNASPQSMQSLMSPQ--MPNQM---MMSPRHNNIGSPHNSQNIASPMMN
P. xylostella PQSVMSVMSPQHNASPQSMQSLMSPQHQMNM---MMSPRHNNMGSPMSQNMTPMVN
P. c-album PQSVMSVMSPQHNASPQSMQSLMSPQ--MPNQM---MMSPRHNNIGSPNSQNIASPMMN
B. mori PQSAMSVMSPQHNASPQSMQSLMSPQ--IPNQI---MLSPRHNNVSPMSQNMASPMMN
D. plexippus PQSVMSVMSPQHNASPQSMPLMSPQ--MPNQM---MSPRNNNIASPIITQNMGSPMMN
M. sexta PQSVMSVMSPQONAMSPQSMQSLMSPQ--MPNQM---MMSPRHNNIGSPMSQNMASPMVN
.**:***** *:*** : **: :* *.** : ** :**:

H. melpomene -----ASPMHGMASPMHPGLQSPVTNPMV----QNMNTMTMNVQPVQNMNM
C. ohridella MGSPMTQNVASPLSHGMPSPHNTNLQSPQMHPIQ-----
P. aegeria MASPMTQNIASPMHGMSPSPHHPGLQSPQIQMM-----PNNHQNIY--NN-----
P. xylostella LASPMTQNMPSPMHGMSPSPHHPGMQSPM-----
P. c-album MASPMAQNIASPMHGMSPSPHHPGLQSPY-----
B. mori MASPMGQTIASPMHGMSPSPHHPGMQSPMGQMASPMIQNMMPVAMNGFVN--PNLQNM
D. plexippus ITSPNLQNIASPMHGMSPSPHHPGLQSPNPMVQN---LSNMPRNSQAIQ--NQNQHHMM
M. sexta MASPMTQNIASPMHGMSPSPHHPGLQSPM-----
:* **:* :*

H. melpomene NT---PALSAPPYNNRQNCNKNVKNFNNVNQFQN-QNYNQAPPYPAQN-QNFNMNRN
C. ohridella -----
P. aegeria -----QMTNYVQSMNYPNQNAQ-----
P. xylostella -----RN
P. c-album -----
B. mori NVNQQQ----MALSYPNCH-RMPARENMPMPNQYTNQNNNQPPQYPMQ-NQNVNVHH
D. plexippus NMNQTTQQLAQQAAYNNHQNCC--KLPSKNYNNVNPQYQNT-YTQAAAYPTQNMNINMRH
M. sexta -----RS

```

H. melpomene      QNMQPYQMMAQFNPQPQMAMMPANHQNVMVYNNQ--MNYI--QPMAYMNQTNQLHPMQMS
C. ohridella      -----QQI-----NQNPGNMVYNN-QMINYQHQQHMPYPNQGNPPNPLQSS
P. aegeria        --MH-----QVQMMFPNNHQNIIYNN-QMANYV--QSMNYPNQNAQMHPMLQLS
P. xylostella     QNMQQYQMMQQFNQNMAMMAA-NQANMSYNNQMMNYQ--QGVNYPNQNAQMHPMLQMS
P. c-album        -----QIMQHYNQNHQTQMP--VMAQNQLVYNN-QMINYA--QPMNYSNQSNAQMHPMQMS
B. mori           QNVQNYQIMQSYNQNAQPTVNQMVQNVQNVHYPN-QPMINY--QQMNF-NQNNRAHPMQSS
D. plexippus      QNLQQYQMMQQFNQNSQIPIPPPNNSNMAYGN-QMINYA--QHGSYTNQSNQLHPMQIS
M. sexta          QNVQQYFVMQQYNQNMQPMHMPMNQNMVYNN-QMMNYQ--Q-MNYPNQSNQMHPMLQMS
                  *: * *           :   : **   .:* *

H. melpomene      RSSVMSVDNSGNMSRGALNSYCEQQSMCPPAQNARYNQNMQYPQPPP--YNTVVNNSNV
C. ohridella      RSSGMSVDNSASMGRGTMSYCYERQY---S-QMQNQYNQNMYSYPQQPPPSYNAAV-NNQF
P. aegeria        RSSVMSVDNSGNMSRGAMNSYCEQQNVCPP-MQNVQYNQNIQYPQPPP--YNSVANSVNV
P. xylostella     RSSVMSVDNSASIGRTMSYCEQQY---P-QQQSQYNQPMYPHPPPP--YNSVNV-NNV
P. c-album        RSSMMSVDNSGNMSRGALNSYCEQQPQCQP-VQNAQYNQNMQYPQPPP--YNSVN-AANV
B. mori           RSSMMSVDNSGNMTRGVMTSYCDTQHQCQP-AQTNHYNQNVQYPQPPP--YNAVNV-NNV
D. plexippus      RSSMMSVDNSGNMNRGSLNGYEQ-TMSPS-AQNQYNQNVQYPAPPP--YNAVNAANV
M. sexta          RSSMMSVDNSANMSRGAMNNYCEPQNQCQP-MQNPPYNPNMQYPQPPP--YNSVNVNTNV
                  *** *****.: ** :. * :           ** : ** ** *:. :.

H. melpomene      MGPPPPKNNHQYNQAMMNNNQYNNQQRSYNQWDYPGNQFNKHN--MQKSVQNSVNMSTG
C. ohridella      MGPPPPKNNHQYNQTMNNNQYNNHQRFPYSQWDYPGNQFNKHN--VQKSMQESVNMSTG
P. aegeria        MGPPPPKNNHQYNQAMMNNNQYNNHQRSYNQWDYPGNQFNKHN--MQKSGQNSVNMSTG
P. xylostella     MGPPPPKNNHQYNQTMNNNQYNNHQRFPYNQWDYPGNQFNKHNNGNPAQKSMQNSVNMSTG
P. c-album        MGPPPPKNNHQYNQAMMNNNQYYSHQRSYNQWDYPGNQFNKHN--MQKSIQNSVNMSTG
B. mori           MGPPPPKNNHQYNQAMMNNNQYNNHQRFPYNQWDYPGNQFNKHN--PQKSTPNSVNMSTG
D. plexippus      MGPPPPKNNHQYNQAMMNNNQYNNQQRSYNQWDYPGNQFNKHN--MQKSTQNSVNMSTG
M. sexta          MGPPPPKNNHQYNQTMNNNQYNNHQRFPYNQWDYPGNQFNKHN--MQKSMQNSVNMSTG
                  *****.:*****.* * .*****          *** :*****

H. melpomene      SQKPNGA--RMSINCNQIIKNGGEQQTDCSMNSLRSQNQ--ADVQVWDISQSQIEATNGR
C. ohridella      SQKPNGTRPQPVNPING---LKNDQADC SMNSLKGEQNQPSDVQVYDISQSQIEAKNGR
P. aegeria        SQKPTNG-ARVPMNCNQMAKNNGEQQADC SMNSLRSQTNQ-ADVQVWDISQSQIEATNGR
P. xylostella     SQKPNNGRITIV---NQMMK---NAETDC SMNSLKSQGNKPSDVQVWDISQSQIEATNGR
P. c-album        SQKPVNG-TRIPMHCNQIVKN-GEQQTDC SMNSLRSQNNQ-ADVQVWDISQSQIEATNGR
B. mori           SQKQVKCN--RLTNCNQIK---NDQADC SMNSLRSQNNQPSQEVQVWDISQSQIEANNGR
D. plexippus      SQKPVNMGTRISMNCNQIMKN-PGEQTDC SMNSLRSQNNQ-ADVQVWDISQSQIEATNGR
M. sexta          SQKPINGT--RQMNCNQIK---TDQTDC SMNSLRSQNNQTSQEVQVWDISQSQIEASNGR
                  ***           *           :*****.: : : :*:*****.* **

H. melpomene      KKNQ-----NNMRQETYQRTLEYVENCENWKSSEMVSSTHPLQGGDNMVVNDLQTSLS
C. ohridella      KKNQN-----TAMQETYQRTLEYVENCENWKSSEMVSSTHPLQAGDNMVVNDLQTSLS
P. aegeria        KKNQ-----NTMRQETYQRTLEYVENCENWKSSEMVSSTHPLQGGDNMVVNDLQTSLS
P. xylostella     KKNQNNNNNNNMVRQETYQRTLEYVESCENWKSSEMVSSTHPLPGADNMVNDLQTSLS
P. c-album        KKNQ-----NTMRQETYQRTLEYVENCENWKSSEMVSSTHPLQGGDNMVVNDLQTSLS
B. mori           RKTQN-----NSMRQETYQRTLEYVENCENWKSSEMVSSTHPLQG-DNMVNDLQTSLS
D. plexippus      KKNQ-----NTMRQETYQRTLEYVESCENWKSSEIVSSSTHPLQGGDNMVVNDLRTSLS
M. sexta          KKNQN-----NTMRQETYQRTLEYVENCENWKSSEMVSSTHPLQGGDNMVVNDLQTSLS
                  :*. *           . :*****.* *****:***:*** . *****:***:

H. melpomene      SFYEENQYLQMIQ
C. ohridella      SFYEENQYLQMIQ
P. aegeria        SFYEENQYLQMIQ
P. xylostella     SFYEENQYLQMIQ
P. c-album        SFYEENQYLQMIQ
B. mori           SFYEENQYLQMIQ
D. plexippus      SFYEENQYLQMIQ
M. sexta          SFYEENQYLQMIQ
                  *****

```

Key:

Zinc finger double domain

Zinc finger C2H2 domain

Peptide corresponding to candidate target site

Chapter 5:

Ovarian miRNAs of the Speckled Wood butterfly

Statement of Author Contributions

The material in this chapter has been written up as a manuscript and submitted to *PLOS ONE* (submitted August 19, 2015; Manuscript ID: PONE-D-15-36496). It was accepted for publication on October 18, 2015 and published online on November 10, 2015. The authors on this manuscript are (in this order): Shan Quah, Casper J. Breuker and Peter W. H. Holland (corresponding author). This material has undergone revision after initial submission of this thesis and prior to acceptance of the paper; the final paper is attached in Appendix C.

The original idea for the project was developed by SQ and PWHH. CJB provided the material used in the study as well as expertise on the Speckled Wood ovarian transcriptome. SQ performed the laboratory work and data analysis and wrote the original draft manuscript. All authors discussed the implications of the study in detail and carried out revision of the draft manuscript. CJB supervised the annotation of the Speckled Wood mRNA transcriptome. CJB and PWHH supervised the work of SQ throughout the length of this study.

1. Abstract

miRNAs are important regulators of animal development and other processes, and impart robustness to living systems through post-transcriptional regulation of specific mRNA transcripts. To investigate the role of miRNAs in ovarian function and formation of the butterfly egg, I sequenced the ovarian miRNA transcriptome of Speckled Wood butterflies, *Pararge aegeria*, from two geographically distant populations. A total of 74 miRNAs were identified, including 11 newly discovered miRNAs, bringing the total of miRNA genes known from *P. aegeria* up to 150. A common set of 55 miRNAs are expressed in both populations. From this set, I identify seven that are consistently either ovary-specific or highly upregulated in ovaries relative to other tissues. This 'ovary set' includes miRNAs with known contributions to ovarian function in other insect species with similar ovaries and mode of oogenesis, including miR-989 and miR-2763, plus new candidates for ovarian function. Target predictions against an expanded *P. aegeria* ovary mRNA transcriptome reveal that several ovary set miRNAs have large numbers of predicted targets in genes with roles critical to ovary function. I suggest these miRNAs act to fine tune expression levels of their targets in the ovary. Other miRNAs have fewer predicted targets and may function to suppress spurious production of undesired transcripts in the ovary. These findings imply that miRNAs act in diverse ways to add robustness to genetic regulatory loops in butterfly ovarian function. I also note that conserved miRNAs are overrepresented in the ovary relative to the whole body. Contrary to prediction, I found no significant difference in expression levels of old and young miRNAs in our datasets.

2. Introduction

MicroRNAs are of significance at all stages during an animal's life cycle from embryonic development (Jones & Newbury, 2010; Lasko, 2011; Pasquinelli et al., 2005; Yekta et al., 2004) to post-embryonic development (Kenny et al., 2013; Lozano et al., 2015) and reproduction (Lasko, 2011; McBride et al., 2012). However, the precise function of specific miRNAs at each of these stages remains poorly understood. Here I focus on miRNAs expressed during oogenesis, which are likely regulators of egg production and ovary function.

Oogenesis in insects must generate a viable egg capable of initiating and sustaining embryonic development after fertilisation. Adequate provisioning of the egg is essential (Telfer, 2009). In addition to sustaining the energetic requirements for embryogenesis, the egg must also be pre-patterned with a polarity, which in insects occurs largely in the form of localised maternal determinants that drive zygotic spatiotemporal gene expression patterns essential for establishment of body axes (Tadros & Lipshitz, 2009). Much is known about the identity and function of maternally provisioned mRNA molecules in insect eggs, but far less about other factors that influence the activity of such mRNAs, such as miRNAs. Furthermore, most of these studies have been performed in drosophilids (Barckmann & Simonelig, 2013; Kok et al., 2007), mosquitoes (Mead & Tu, 2008), cockroaches (Chen et al., 2013) and beetles (Lozano et al., 2015; Preuss, Lopez, Colbourne, & Wade, 2012). Far less is known about the molecular control of oogenesis in butterflies and moths, despite Lepidoptera being one of the largest insect orders. Carter et al., (2013) used transcriptome sequencing to examine gene expression during oogenesis and the diversity of transcripts incorporated into freshly laid eggs of the Speckled Wood butterfly, *Pararge aegeria* (L.). Ferguson et al. (2014) described spatiotemporal expression of maternally-expressed homeobox genes in the same species. However, neither study examined ovarian miRNA expression. A diversity of lepidopteran miRNA genes is known, from several species, an evolutionary analysis of which is presented in Chapter 2 (see also Quah et al., 2015). For *P. aegeria*, this work used a pooled RNA library combining whole body RNAs from various life cycle stages, and did not look specifically at ovarian miRNAs. It was, therefore, not possible in the previous study to identify butterfly miRNAs specifically of ovarian origin from amidst other miRNAs.

Here I present the first analysis of miRNAs in relation to ovary function in a butterfly. I sequenced the ovarian miRNA transcriptomes of Speckled Woods from two different populations, and identified 74 ovary-expressed miRNAs, including a common ovary-

specific set of 7 miRNAs highly upregulated in ovaries of both populations. Target predictions for ovary-enriched miRNAs against a combined ovarian transcriptome assembly generated from the same individuals gave insights into potential functions of miRNAs in butterfly ovaries.

3. Materials and Methods

3.1 Sample collection and sequencing

I sequenced the ovarian small RNA and mRNA transcriptomes of two replicate females from each of two *P. aegeria* populations kept as lab stocks: (1) Zonza on Corsica, France, and (2) St Hubert, from the south of Belgium. The St Hubert population is the same population for which phenotypic data has previously been collected on maternal effects and egg production (Gibbs, Breuker, & Van Dyck, 2010) as well as ovarian transcriptomic data (Carter et al., 2013) and for which a draft genome has been characterized (Ferguson et al., 2014). It is also the population surveyed in the whole-body miRNA transcriptomic study presented in Chapter 3. The Zonza stock was newly established in 2014; the St Hubert stock has been in the lab since 2007 (see Carter et al., 2013 for rearing *P. aegeria* details in general and for St Hubert specifically).

Fully eclosed females were allowed to mate, and were placed individually in netted cages (0.5 m³) along with a potted *Poa trivialis* grassplant for oviposition and an artificial flower containing a 10% honey solution. Females (kept at 21±0.3°C, LD 16:8) were sacrificed 6 days after mating, when they reached the peak of egg production (Gibbs et al., 2010). Ovaries were dissected and cleared of fat and surrounding tissue (cf. Carter et al., 2013). Total RNA was extracted using TRIzol (Life Technologies). Total RNA samples and analysed by Illumina HiSeq 2000 sequencing at the Beijing Genomics Institute. Each sample was used to generate both a mRNA transcriptome library as well as a small RNA library. The mRNA transcriptome sequencing used paired-end reads of 2 x 90 nucleotides from a 160 bp insert library.

3.2 miRNA identification

Small RNA data processing was carried out as previously described in Chapter 3, except that instead of annotating known miRNAs by comparison to miRBase, precursors were scanned by BLASTn against the set of previously reported *P. aegeria* miRNAs (Quah et al., 2015, see also Chapter 3). Sequences without hits against this dataset were subject to additional processing to remove hits against repetitive elements and non-coding RNAs, as well as precursors with reads mapping only to one arm and those with <95% of reads on either arm with the same 5' starting position.

Only sequences with >50 dominant arm reads in Zona or >18 dominant arms in St Hubert were retained for annotation. The introduction of a minimal read count is because in this study I am focussing on miRNAs with putative ovarian function, whereas in Chapter 3, I was searching for all miRNAs; the minimal read counts chosen for the two samples are equivalent after compensating for relative miRNA sequencing depth. All shortlisted candidates were then compared against miRBase 20 to identify known miRNAs which were not have been found in the original *P. aegeria* miRNA survey in Chapter 3. The same cutoff values described above were used to determine if a known miRNA was expressed in both ovarian samples.

Read counts for each mature product, as generated by miRDeep2 (Friedländer et al., 2012), were cross-checked by running the quantifier.pl script in the miRDeep2 package, using known *P. aegeria* miRNA precursors for read mapping. Both approaches generated similar read count data. Raw read counts for the Zona and St Hubert samples, as well as for the original pooled dataset, were normalised against the total number of miRNA reads in each sample.

3.3 Transcriptome assembly

To assemble a reference ovarian mRNA transcriptome, raw RNAseq reads were processed at source for removal of low quality reads and adaptor trimming. Read quality was then validated by checking FASTQ data using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and the first 15 bases trimmed off all reads using the fastqtrim.py script (Appendix A1). Reads from Zona and St Hubert were pooled and assembled using the Trinity assembler (version: trinityrnaseq_r20140717) (Grabherr et al., 2011) to generate a combined *P. aegeria* ovarian transcriptome assembly containing 54,861 contigs. Contig FPKM (Fragments Per Kilobase of exon per Million of fragments mapped) values for each sample were determined using RSEM (Li & Dewey, 2011). In order to retain only biologically interpretable contigs, I employed a multi-step approach to annotation. Contigs were first screened by BLASTn against the nr database on NCBI to reduce the number of misassembled contigs in the dataset. Remaining contigs were then compared by BLASTn against the *P. aegeria* ovarian transcriptome reported by Carter et al., 2013. A referenced list of proteins with known roles in ovarian and egg function (Carter et al., 2013) was also used to screen the combined transcriptome assembly by tBLASTn. All

BLAST searches were subject to an E-value cutoff of $1e-05$. Only contigs with hits to one or more of these datasets were retained, leaving 17636 annotated contigs which were further characterized using Blast2GO (Conesa et al., 2005). This involved screening contigs by BLASTx followed by mapping and annotation. InterProScan and ANNEX were used to augment annotation of GO terms. GO terms were condensed using the generic GO Slim subset.

3.4 miRNA target prediction

miRNA target prediction was carried out using the PITA algorithm (Kertesz et al., 2007) with default parameters. Only predictions with a score of -10 or below were accepted as predicted hits.

For the purpose of comparing GO term distribution between PITA prediction datasets and the combined transcriptome assembly, the number of GO terms in each dataset was normalised against the total number of contigs in that dataset.

4. Results

4.1 Small RNA sequencing and data processing

I constructed and sequenced small RNA libraries from ovaries of the Speckled Wood butterfly *P. aegeria* to identify miRNAs expressed in the lepidopteran ovary. To compensate for possible differences dependent on geography, climate or diet, I chose to analyse females from two distinct populations from different parts of Europe: St Hubert in Belgium and Zonza in Corsica. Total RNA was extracted from ovarioles and used to construct small RNA libraries for each sample. A total of 46,481,949 Illumina HiSeq reads were obtained (22,604,869 for Corsica and 23,877,080 for Belgium). After filtering reads to remove possible degradation products and non-miRNAs, these were compared to a published dataset of *P. aegeria* miRNAs (Quah et al., 2015, see also Chapter 3).

Of the 139 previously known miRNAs in *P. aegeria* (Quah et al., 2015; Chapter 3), I obtained reads for 80 and 76 miRNAs in the ovary-specific samples for St Hubert and Zonza respectively (Supplement S5.1). Some of these were expressed at low levels and should not be considered as candidates for ovarian function; using read-count cut-off thresholds, I reduced this number to 58 and 57 in St Hubert and Zonza respectively. Of these, 52 miRNAs are expressed above cutoff threshold in both populations, implying that 63 of the previously described 139 *P. aegeria* miRNAs are expressed in ovary (Figure 5.1A, Supplement S5.1). In addition, a total of 11 miRNAs that had not been previously described in *P. aegeria* were found; 2 were present in the Zonza sample only, 6 in the St Hubert sample only, and 3 were found in both: B-0070/C-0072, B-0210/C-0160 and B-0401/C-0173; Table 5.1). The total number of miRNAs identified as expressed in *P. aegeria* ovary is therefore 74 (63 previously described miRNAs and 11 newly discovered).

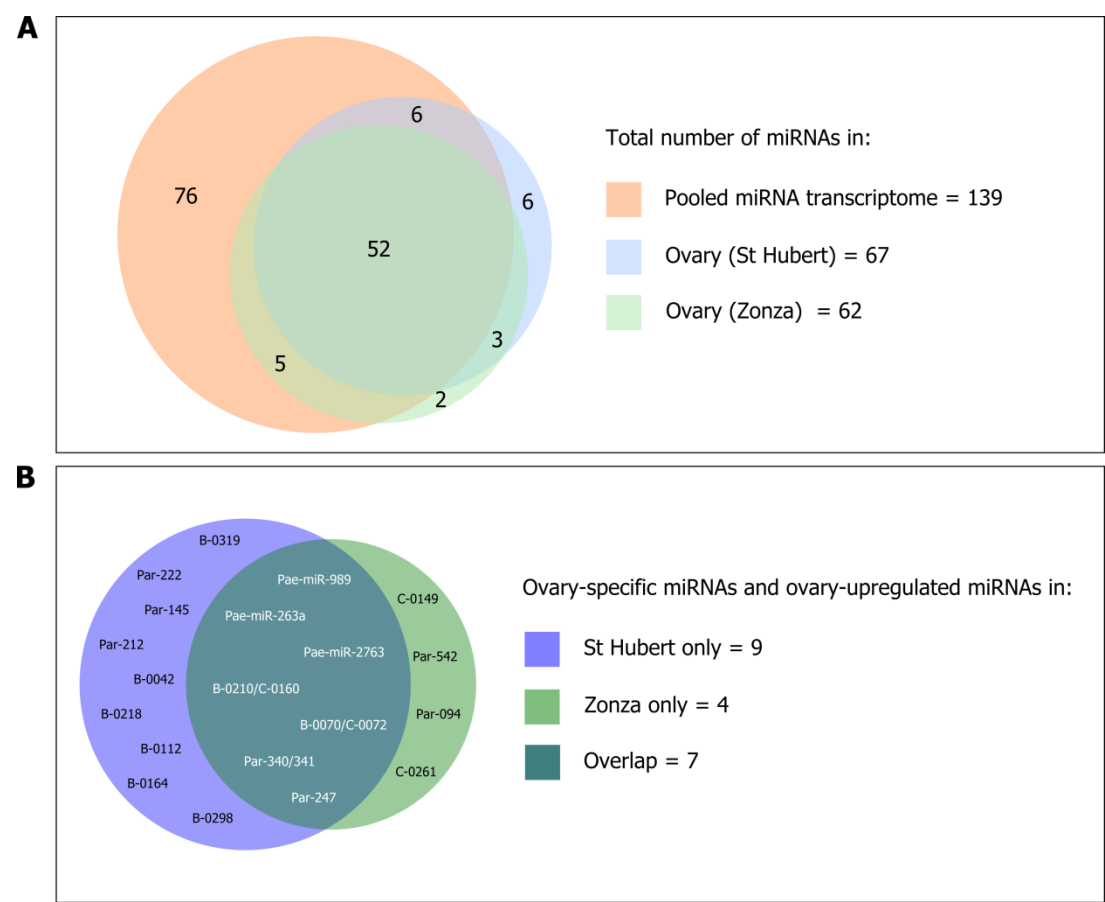


Figure 5.1

(A) Distribution of miRNAs identified in each small RNA library – St Hubert, Zonza and the whole body pooled transcriptome across multiple life stages reported in Chapter 3

(B) miRNAs with higher abundances in the ovarian datasets relative to the pooled miRNA transcriptome

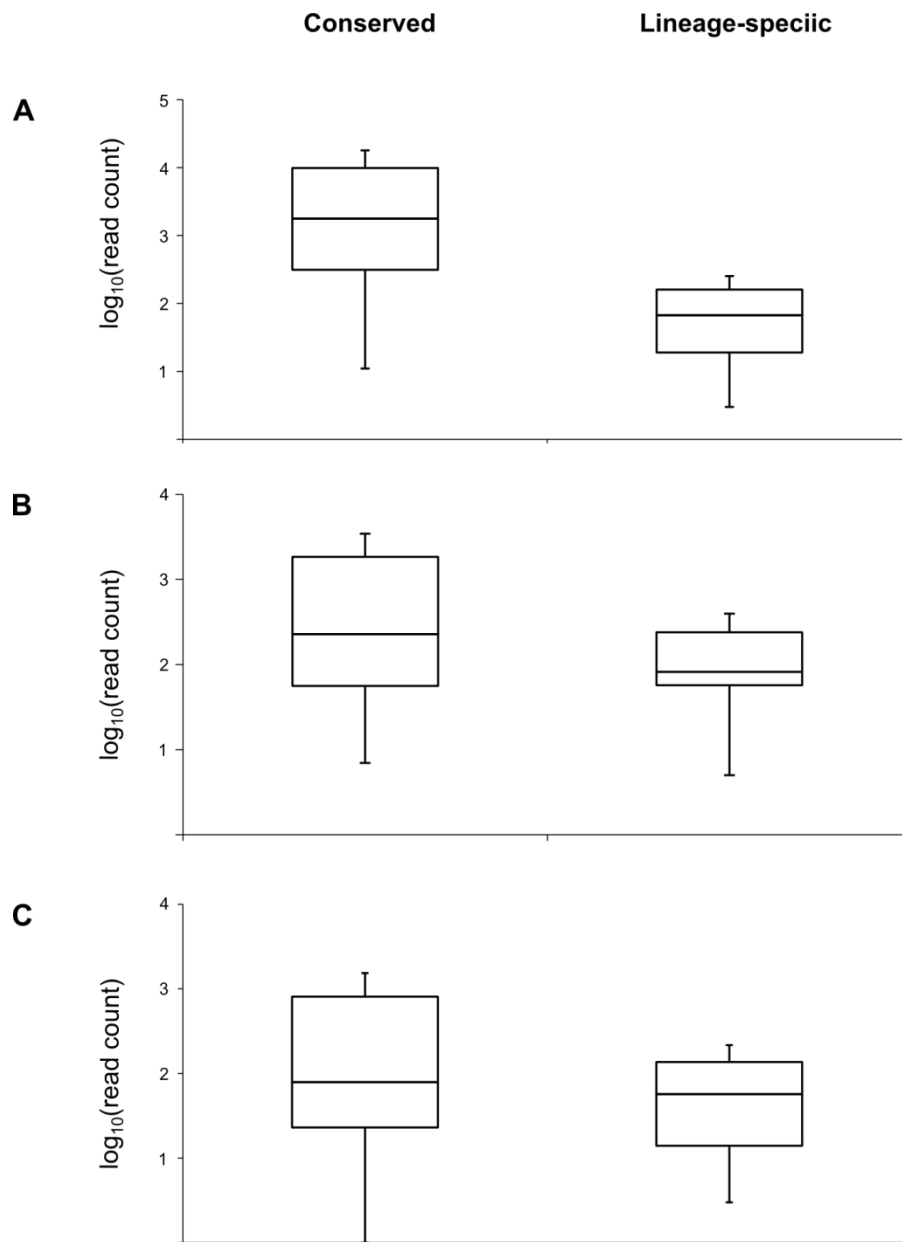
[illegible]

Table 5.1 miRNAs identified in this study which were not found in the whole body pooled transcriptome reported in Chapter 3

4.2 Effect of miRNA age on expression level

I categorised the miRNAs present in the updated *P. aegeria* dataset by age, defining 'conserved' miRNAs as those present in *P. aegeria* and at least one other species and 'lineage-specific' miRNAs as those not found (or predicted in genome sequence) outside of the Speckled Wood. To test if the 11 newly discovered miRNAs are lineage-specific genes, or whether they are evolutionarily older, each was checked using BLASTn against all known miRNA precursors on miRBase and their inferred pre-miRNA sequences were compared by BLASTn against the genomes of other Lepidoptera (*Heliconius melpomene*, *Danaus plexippus*, *Manduca sexta*, *Bombyx mori*, *Plutella xylostella* and *Cameraria ohridella*) as well as caddisfly (*Glyptotendipes pallidus*). One sequence, B-0401/C-0173, was orthologous to miR-2763 (Supplement S5.2), which likely originated in the common lepidopteran ancestor, but was not previously identified in a whole body small RNA library for *P. aegeria* (Chapter 3); the other 10 sequences are thus far specific to *P. aegeria* and inferred to have arisen more recently.

The total number of reads mapped to each miRNA precursor was taken as a measure of expression level and compared between conserved and lineage-specific miRNAs in St Hubert, Zonza, and the pooled miRNA transcriptome reported in Chapter 3. There is a large spread of expression levels detected for both conserved and lineage-specific miRNAs in each of the three datasets. No significant difference in mean expression levels was observed in any of the miRNA transcriptomes (Figure 5.2). However, in all cases it was the conserved miRNAs which exhibited the highest expression levels.

**Figure 5.2**

Box plot showing the expression levels for conserved and lineage-specific miRNAs in (A) pooled miRNA transcriptome, (B) St Hubert and (C) Zonza. Mean expression level is not significantly different between conserved and lineage-specific miRNAs in all transcriptomes ($P = 0.109$, 0.086 and 0.431 respectively; t-test).

4.3 Older miRNAs are enriched in the ovary

I studied relative proportions of conserved and lineage-specific miRNAs expressed in the ovary and compared these to the proportions observed in the pooled transcriptome (Figure 5.3). Ovarian expression of each miRNA was defined by its detection above cut-off levels in both populations. There is a higher proportion of conserved miRNAs (64%) expressed in the ovary compared to that in the pooled miRNA transcriptome (47%, chi-square $P < 0.05$).

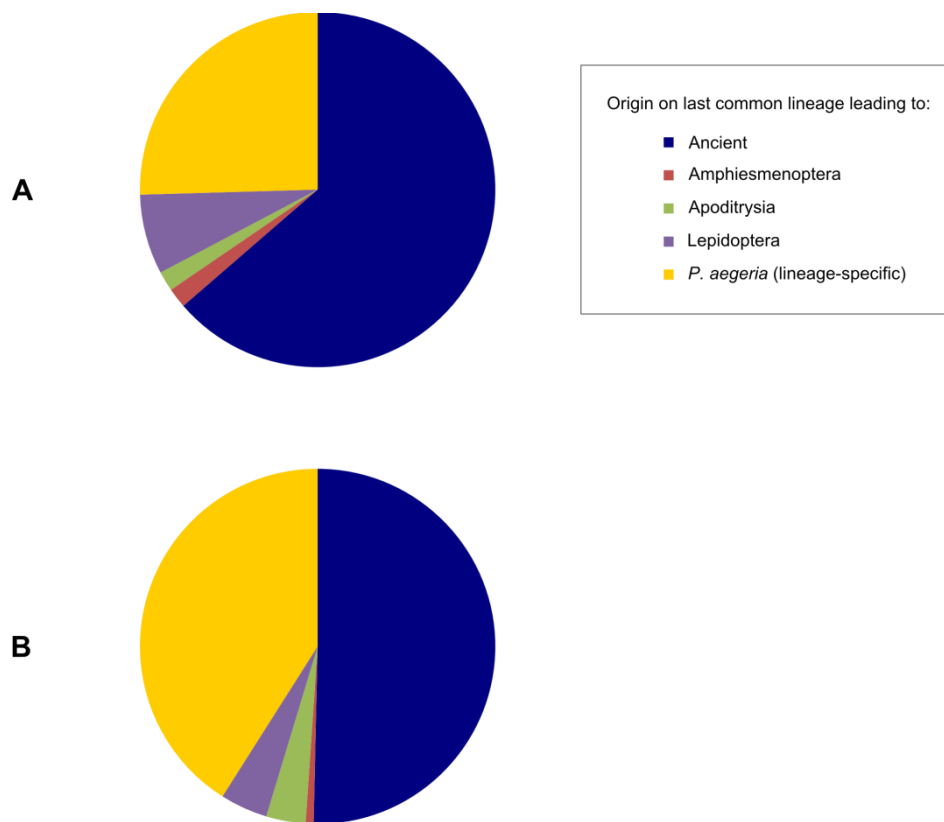


Figure 5.3

Pie chart showing the proportions of conserved miRNAs originating on each lineage which are (A) present in both St Hubert and Zonza ovarian transcriptomes and (B) present in the original pooled transcriptome reported in Chapter 3. 'Ancient' miRNAs are defined as those originating prior to the emergence of the Amphiesmenoptera (clade consisting of Lepidoptera and Trichoptera)

4.4 Ovarian-specific and ovarian-upregulated miRNAs

The most abundant miRNAs identified in the ovary are (in decreasing order of abundance) miR-989, miR-263a and miR-1. This observation is consistent between St Hubert and Zonza. Other consistently highly-expressed ovarian miRNAs include miR-993, miR-2755 and miR-184 (Table 5.2). In the pooled whole-body miRNA transcriptome, miR-1 is the most abundantly expressed sequence, followed by miR-989 and let-7; miR-263a and miR-993 are relatively highly expressed in the pooled transcriptome.

The above analyses reveal which miRNA genes are expressed in ovary, but this list includes miRNAs expressed in a range of tissues, as well as ovarian-upregulated sequences. To gain insight into the role of miRNAs in ovarian function, I wished to identify the latter category of miRNAs. Mature and dominant arm read counts for all miRNAs were normalised against the total number of miRNA reads per sample, and then compared with normalised read counts from the pooled miRNA dataset for *P. aegeria* from Chapter 3. miRNAs with dominant arm read counts below a cutoff value in the Belgian and Corsican samples were deemed to not be expressed and excluded from analysis. These analyses identified five miRNAs with > 4-fold upregulation in both libraries (Table 5.3): Pae-miR-989, Pae-miR-263a, Par-247, Par-340 and Par-341. An additional 13 miRNAs showed up-regulation in just one ovary sample, or were only detected in either ovarian dataset but not in the pooled miRNA transcriptome (Figure 5.1B, Supplement S5.1). As Par-340 and Par-341 share 3' arm reads, I group them together (Par-340/341) for the purpose of analysis. These four 'ovarian-upregulated' miRNA sequences can be combined with the three newly identified 'ovary-specific' miRNAs common to both samples (described above), to form a category of seven 'ovary set miRNAs' for further analysis.

Rank	Whole body (pooled)	Ovary (St Hubert)	Ovary (Zonza)
1	Pae-miR-1	Pae-miR-989	Pae-miR-989
2	Pae-miR-989	Pae-miR-263a	Pae-miR-263a
3	Pae-let-7	Pae-miR-1	Pae-miR-1
4	Pae-miR-184	Pae-miR-184	Pae-miR-2755
5	Pae-miR-2766	Pae-miR-993	Pae-miR-184
6	Pae-miR-10a	Pae-miR-2755	Pae-miR-993
7	Pae-miR-31	Pae-miR-279d	Pae-let-7
8	Pae-miR-2755	Pae-miR-279b	Par-017
9	Pae-miR-263a	Par-017	Par-013
10	Pae-miR-281	Pae-let-7	Pae-miR-279b

Table 5.2

The ten most abundant miRNAs ranked by expression in the whole body pooled transcriptome and each of the St Hubert and Zonza ovarian transcriptomes.

miRNA	Belgium				Corsica			
	Total	5'	3'	Fold change	Total	5'	3'	Fold change
Par-247	125	120	5	27.5	146	142	4	11.5
Par-341	52	51	1	10.5	70	69	1	5.0
Par-340	52	51	1	10.3	70	69	1	4.9
Pae-miR-989	288099	32	288067	5.2	873884	88	873796	5.6
Pae-miR-263a	24821	24819	2	4.5	63593	63591	2	4.1
B-0070/C-0072	291	173	118	-	574	130	444	-
B-0210/C-0160	97	9	88	-	262	19	243	-
B-0401/C-0173 (Pae-miR-2763)	27	6	21	-	137	8	129	-

Table 5.3

The 'ovary set' miRNAs which show greater than 4-fold abundance in both St Hubert (Belgium) and Zonza (Corsica) ovarian transcriptomes relative to the whole body transcriptome. Fold change is calculated from normalized total read counts. An accurate fold change in abundance cannot be given for the last three miRNAs on the table, as these were detected only in ovarian transcriptomes (both populations) and not at all in the pooled transcriptome.

4.5 Identification of putative ovarian miRNA targets

To identify possible mRNAs which may be regulated by the ovary-set miRNAs, it is necessary to compare to *P. aegeria* transcriptomes. The same samples used for construction of small RNA libraries were also used to generate separate transcriptome libraries. The Illumina HiSeq 2000 platform was used to generate 26,332,892 read pairs for Zonza and 26,466,283 read pairs for St Hubert.

Reads from both samples were pooled to construct a combined transcriptome assembly (Table 5.4). The PITA target prediction algorithm (Kertesz et al., 2007) was used to identify potential targets in the transcriptome for both 5' and 3' mature products for each of the 7 'ovary-set' miRNAs.

I compared the distribution of GO terms between the annotated contigs in the combined transcriptome assembly and the predicted targets for each of the 'ovary set' miRNAs to obtain a broad perspective on the actions of each mature product (Supplement S5.3). For example, analysis using the dominant mature product, miR-989-3p, of the *mir-989* gene showed significant bias in representation of biological functions (Figure 5.4). Amongst the predicted targets, contigs associated with the 'metabolic processes' term are vastly underrepresented relative to the combined transcriptome. Contigs associated with GO terms relevant to oogenesis and ovarian function, including 'biosynthetic process', 'anatomical structure development' and 'embryo development' are overrepresented. Similar enrichments were observed for the dominant products of miR-263a and B-0210/C-0160 (miR-263a-5p and B-0210/C-0160-3p respectively (Supplement S5.3). Few predicted targets were found for both arms of Par-340/341 and Par-247. The minor 5' product of miR-2763 has more predicted targets than the dominant 3' product. Unlike the other miRNAs identified in this study which show clear arm dominance consistent between both samples, the mature product of the B-0070/C-0072 miRNA differs between St Hubert and Zonza. Predictions for the 3' mature product do not show the same depletion in 'metabolic process' terms observed for the other miRNAs.

I then sought to identify specific miRNA-mRNA target pairs of interest within the dataset. Predicted target contigs corresponding to genes involved in ovarian and egg function (Carter et al., 2013) are listed in Table 5.5. These contigs cover a variety of functions including regulation of egg chamber development to oocyte provisioning and defence against both pathogens and the external environment. In many cases, multiple "ovary set" miRNAs are predicted to target the same transcript, an example of which is

neuroglian (*nrg*) involved in regulation of border cell migration in flies (Wei, Hortsch, & Goode, 2004). Occasionally, both mature products of the same miRNA gene are predicted to regulate the same target. This is best demonstrated by miR-989, whose 5' and 3' arms both regulate expression levels of the serosal gene *ShxD*, the dynein heavy chain component *Dhc64C* and the transcriptional activator *Mad*.

	Combined assembly	St Hubert	Zonza
Number of contigs	54861	40915	45206
Total length	64060799	47311956	51104606
Maximum contig length	16200	16174	14446
N50	2236	2295	2130
N90	439	423	426

Table 5.4

Trinity assembly statistics for the separate St Hubert and Zonza mRNA transcriptomes as well as a transcriptome combining reads from both populations, as generated by the *contig-stats.pl* script (Appendix A3). Number of contigs is prior to filtering by BLAST to remove potentially misassembled contigs.

Function	miRNA	Targets
Egg polarity and embryo patterning	B-0210/C-0160-3p B-0210_C-0160-3p Pae-miR-989-3p Pae-miR-263a-3p Pae-miR-989-5p	<i>bun, Ets97D, exu, msn, spoonbill</i> <i>Rab11</i> <i>msn, spoonbill</i> <i>exu</i> <i>bun, Rab11, PI4KIIIalpha, gcl</i>
Morphogenesis	Pae-miR-989-3p Pae-miR-989-5p Pae-miR-263a-5p	<i>inx3</i> <i>blistered</i> <i>stmA</i>
Follicle cell function	B-0210/C-0160-3p Pae-miR-989-3p Pae-miR-989-5p	<i>Krn, Tie, sdt</i> <i>crb, sdt</i> <i>mbc, tai</i>
Border cell migration	B-0070/C-0072-5p B-0401/C-0173-3p Pae-miR-989-5p	<i>Nrg</i> <i>Nrg</i> <i>Nrg</i>
Immune defence	Pae-miR-989-3p	<i>hemolin</i>
Stem cell regulation	B-0210/C-0160-3p Pae-miR-989-5p Pae-miR-989-3p Pae-miR-263a-3p Pae-miR-263a-5p	<i>Mad</i> <i>Mad, osa, RanBPM</i> <i>Mad</i> <i>Mad</i> <i>Mad</i>
Oocyte determination	B-0210/C-0160-3p Pae-miR-989-5p Pae-miR-989-3p	<i>egl, Dhc64C</i> <i>Dhc64C</i> <i>egl, Dhc64C</i>
Serosa	Pae-miR-989-5p Pae-miR-989-3p B-0210/C-0160-3p	<i>ShxD</i> <i>ShxD</i> <i>wdb</i>
Lipid storage	Pae-miR-263a-3p Par-247-5p Pae-miR-263a-5p	<i>Akt1</i> <i>Akt1</i> <i>Akt1</i>
Hormone response	B-0210/C-0160-3p Pae-miR-989-5p	<i>E75</i> <i>E75</i>

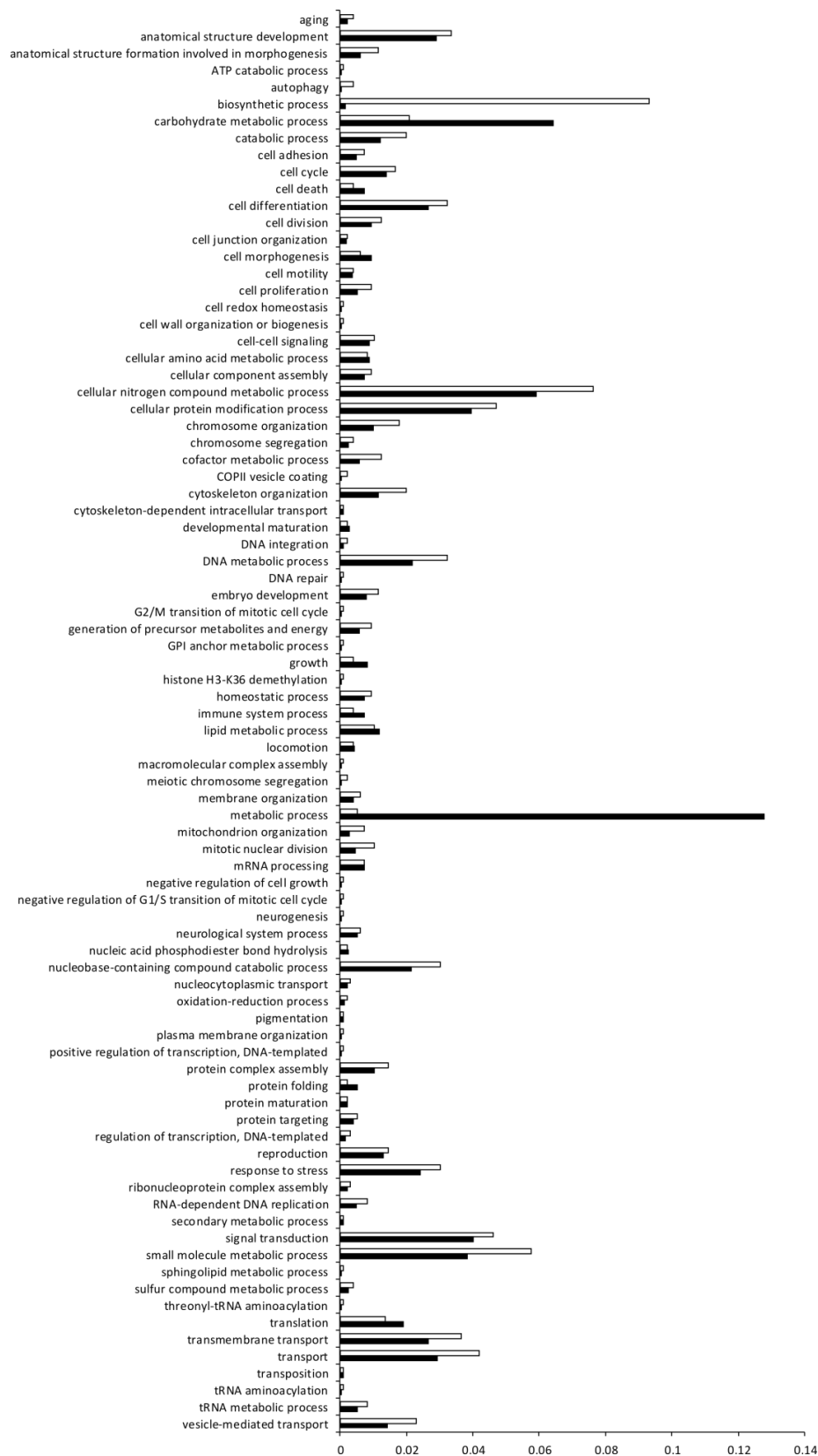
Table 5.5

Predicted targets of 'ovary set' miRNAs with reported roles in ovarian and egg function.

	Dominant products	Star products
Many predicted targets, enriched for ovary function	miR-989-3p miR-263a-3p B-0210/C-0160-3p	miR-989-5p miR-263a-5p
Many predicted targets, no clear enrichment for ovary function	B-0070/C-0072-5p* B-0070/C-0072-3p*	miR-2763-5p
Few targets	miR-2763-3p miR-340/341-5p Par-247-5p	miR-340-3p miR-341-3p B-0210/C-0160-5p Par-247-3p

Table 5.6

Summary of target predictions for mature products of 'ovary set' miRNAs. It was not possible to assess enrichment for ovary function terms where there were few targets predicted.

**Figure 5.4**

GO term distribution of predicted target contigs for Pae-miR-989-3p (white) compared with GO term distribution across the combined ovarian transcriptome (black).

5. Discussion

5.1 Diversity of miRNAs in *Pararge aegeria*

In a previous study, 139 miRNA genes were identified in the Speckled Wood butterfly, *P. aegeria* (Quah et al., 2015). Through deep sequencing of small RNAs in adult ovaries from two geographically-distinct populations, I have now increased this number to 150 expressed miRNA genes. Of the 11 new miRNAs identified in the ovarian small RNA transcriptome, six are uniquely expressed in the St Hubert sample, two are unique to Monza, and three were found in both populations.

A common set of 55 miRNAs are expressed in the ovaries in both St Hubert and Monza. The most highly expressed of these include miRNAs also known to be highly expressed in other insects, such as miR-989 and miR-1. miR-1 is highly conserved across bilaterians (Wheeler et al., 2009) and has documented expression in muscle in vertebrates (Chen et al., 2006; Kloosterman, Wienholds, de Bruijn, Kauppinen, & Plasterk, 2006) and insects (Sokol, 2005). Its expression is also known to be highly abundant in the ovaries of the cockroach *Blattella germanica* (Cristino, Tanaka, Rubio, Piulachs, & Belles, 2011). miR-989 is known to be expressed in the ovary in mosquitoes (Mead & Tu, 2008) and flies (Kugler, Chen, Weng, & Cohen, 2013).

21 of the 150 miRNA genes in the *P. aegeria* miRNA transcriptome are upregulated by at least 4-fold in, or are specific to, the ovary in at least one sample. Seven of these are either upregulated or ovary-specific in both populations. I refer to these seven genes as the 'ovary set' miRNAs.

5.2 miRNA age, ovary expression and expression level

According to the model postulated by Chen & Rajewsky (2007), new miRNAs are expressed at low levels and in restricted spatiotemporal domains. These characteristics would avoid severe fitness penalties due to the minimal requirements for miRNA-target binding and the resultant large number of potential targets in the transcriptome. Over time, deleterious target sites would be purged from the genome and advantageous

target sites preserved, enabling the miRNA to increase its expression level and relax its tissue specificity.

Here I utilise my characterization of the butterfly ovarian miRNA transcriptome, combined with the previous analysis of a pooled whole body miRNA transcriptome (Chapter 3), to examine expression levels of conserved and lineage-specific miRNAs, as well as their relative contribution to the ovarian miRNA transcriptome. I consider separately the two ways in which overall miRNA expression might be increased – breadth of expression and level of transcription. Based on the model described above, I would predict that conserved miRNAs exhibit overall higher expression levels in the pooled transcriptome due to a wider range of expression domains. If evolutionarily-younger miRNAs were more tissue-specific miRNAs, these might be predicted to be under-represented in the pooled transcriptome, and reads might even be missed due to dilution effects.

I observed a larger proportion of conserved miRNAs expressed in the ovary compared to the pooled transcriptome, which is at odds with the predictions made by the model. One possible explanation for this is that the pooled transcriptome includes RNAs from embryonic stages. It has been suggested that the early embryo represents a permissive environment for newly acquired miRNAs to be expressed and exert their effects (Ninova et al., 2014b). Increased representation of young miRNAs in a library comprised of pooled developmental stages may therefore result from such miRNAs being expressed during embryogenesis. An alternative explanation is that the capacity of a particular tissue to tolerate experimentation with novel miRNA expression may vary between tissues; tissues with highly complex transcriptomes such as the primate brain, cited as evidence for Chen & Rajewsky's model of miRNA evolution (Berezikov, 2011; Berezikov, Thuemmler, et al., 2006) may differ in their robustness and resultant response to new miRNAs compared to those with simpler transcriptomes. Thus, it is possible that while some tissues express a large number of lineage-specific (young) miRNAs, new miRNAs arising under restricted transcriptional control in less permissive tissues are more likely to be eliminated by selection.

The second prediction of the model is that conserved miRNAs would display higher expression levels than lineage-specific miRNAs. Evidence for this was presented in flies (Lu et al., 2008), where miRNAs conserved across five or more drosophilid species were found to have much higher expression levels than those conserved across four or fewer species. However, the number of sequences annotated as miRNAs in this study was subsequently found to be exaggerated as a result of the

annotation of degradation products as miRNAs (Berezikov et al., 2010). I observe that both conserved and young miRNAs possess a wide range of expression levels, and that these ranges overlap substantially with each other. There is no significant difference in the mean expression levels for conserved and young miRNAs in my datasets, although in each transcriptome the miRNAs with the highest expression levels are conserved miRNAs. Conserved miRNAs also have a skew towards higher expression levels in the pooled transcriptome. This is possibly due to consistent high expression in a wide range of tissues. My observations and the model of Chen & Rajewsky can be reconciled. Given the constant high turnover of young miRNAs in metazoan genomes (Lyu et al., 2014; Meunier et al., 2013), deep sequencing of a particular tissue presents a snapshot of the transcriptome at a given point in time. It is unlikely that there is any mechanistic bias behind miRNA birth to predispose newly evolved miRNAs to emerge under the control of tissue-specific promoters with low expression levels, but it may be the case that such miRNAs are more likely to subsequently avoid elimination through natural selection.

5.3 'Ovary set' miRNAs may contribute to robustness in ovarian transcriptional networks

A combined transcriptome assembly for both populations was constructed and used for miRNA target prediction on mature 5' and 3' products of all seven 'ovary set' miRNAs. Six of these miRNAs show arm dominance consistently between the two populations; in contrast, B-0070/C-0072 has reads mapping to both arms at a similar level and the more highly expressed product is different between St Hubert and Zonza. An intriguing observation is that while some mature miRNAs have large numbers of predicted targets, some have very few (Table 5.6). Comparison of GO terms between the combined transcriptome and the target prediction datasets ovary set miRNAs gives clues to the general roles of these miRNA genes. In particular, the dominant products of miR-989, miR-263a and B-0210/C-0160 (miR-989-3p, miR-263a-3p and B-0210/C-0160-3p respectively) share certain patterns in their target predictions, as does the 5p mature product of B-0070/C-0072 (although this is the dominant arm in only one population); these miRNAs have large numbers of predicted targets. In each case there is a strong underrepresentation of contigs associated with metabolic processes amongst predicted targets, and an overrepresentation of contigs annotated with GO

terms associated with egg and ovary functions such as development and reproduction. However, a similar trend was not observed for the dominant products Par-340/341-5p, miR-2763-3p and Par-247-5p, which have very few predicted targets. Most of the star products also had very few predicted targets, except for miR-263a-5p and miR-989-5p, which display the same overrepresentation of ovary function-related GO terms as do their corresponding dominant products.

One possible explanation for the observed distribution of GO terms in the miRNA prediction datasets may be the mode of regulation adopted by each miRNA. In 2012, Ebert and Sharp reviewed miRNA expression profiles and their contribution to developmental robustness, suggesting that miRNAs may be involved in either coherent or incoherent feedforward loops (Ebert & Sharp, 2012). In a coherent loop, the miRNA and its target have mutually exclusive expression domains (Shkumatava et al., 2009) and the miRNA is thought to reinforce tissue identity established by gene expression through repression of leaky transcription. In contrast, miRNAs incorporated into incoherent loops are co-expressed with their targets, and function to modulate or buffer levels of target gene expression (Ebert & Sharp, 2012). I hypothesise that the miRNAs for which very few targets are detected in the transcriptome may be involved in coherent regulation. These would be present in the ovary mainly as safeguards, suppressing spurious transcription of transcripts whose expression might otherwise be detrimental to ovary function, or lead to embryonic defects through maternal effects. In contrast, those miRNAs with an abundance of target predictions in my dataset may well be involved in incoherent loops where they act to fine tune quantitative levels of gene expression in the ovary; consistent with this, these target sets are enriched in ovarian functions.

5.4 Potential conserved ovarian functions for insect miRNAs

One of the newly identified miRNAs within the 'ovary set' is the Speckled Wood homologue of miR-2763, previously described in the moths *B. mori* and *Manduca sexta* (Zhang et al., 2012). In silkworm, miR-2763 is more highly expressed in adult moths than in larvae or pupae (Jagadeeswaran et al., 2010). Three potential targets were identified for Bmo-miR-2763 in the *B. mori* genome (Jagadeeswaran et al., 2010). The first encodes the diapause hormone receptor (*DHR*), which is expressed in the

prothoracic glands and may have a role in hormonal signalling. *DHR* is also expressed in the silkworm ovary (Homma et al., 2006) where it is involved in transduction of the diapause hormone signal and exerts regulation over embryonic diapause. The second predicted target, *GRF*, encodes an insect protein related to vertebrate germ cell nuclear factor (GCNF). (Charles, Shinoda, & Chinzei, 1999). Mouse GCNF is expressed in gonads of both sexes and thought to be involved in gametogenesis (Chung & Cooney, 2001). GRF may function in ecdysteroid signalling in moths (Charles et al., 1999). The third putative target, *GATA β* , is implicated in choriogenesis (Jagadeeswaran et al., 2010). Although these target predictions were not replicated in *P. aegeria*, this may be due to partial mRNA predictions in the ovary transcriptome data, so some conservation of function is still possible. I note that *DHR* expression was detected in the Speckled Wood ovary transcriptome, and this species does not undergo embryonic diapause.

Another miRNA with possible conserved ovarian functions across a wide variety of insect species is miR-989: a likely insect-specific miRNA (Kozomara & Griffiths-Jones, 2011) and the most abundant miRNA in the butterfly ovary. I found expression of this mRNA to be upregulated by > 5-fold in both butterfly ovary samples relative to the pooled sample, and similarly it is highly expressed in ovaries of several fly species (Lucas, Myles, & Raikhel, 2013). miR-989 is also upregulated in mosquito ovaries in response to blood feeding (Mead & Tu, 2008). The pupal phase, during which oogenesis occurs, coincides with increased expression miR-989 in *B. mori* (Jagadeeswaran et al., 2010), suggestive of a similar upregulation of this miRNA during egg production in the Lepidoptera. Experimental evidence for the involvement of miR-989 in ovarian function comes from a study in *Drosophila melanogaster* (Kugler et al., 2013), where it was demonstrated that a loss of miR-989 function impairs border cell migration in the egg chamber, leading to sterility.

miR-263a is the third 'ovary set' miRNA found in *P. aegeria* which has likewise been characterized in a variety of other insect species. It has been hypothesised that miR-263a could be involved in trans-generational immune priming in the greater wax moth *Galleria mellonella* (Mukherjee & Vilcinskis, 2014). It is strongly upregulated in *G. mellonella* pupae, and is also induced in the larval midgut, the rest of the body, and the eggs of females fed as larvae on diets contaminated with pathogenic bacteria. Non-pathogenic bacterial contamination did not produce the same effect (Mukherjee & Vilcinskis, 2014). My finding that miR-263a is highly expressed in the ovaries of *P. aegeria* lends support to the idea that such a mechanism could be conserved across Lepidoptera, but more direct evidence is required to address this hypothesis in the

Speckled Wood. Outside of the ovary, miR-263a has documented contributions to developmental robustness in the *Drosophila* eye (Hilgers, Bushati, & Cohen, 2010). Double mutants for *mir-263a* and the related *mir-263b* undergo sporadic loss of mechanosensory interommatidial bristles (IOBs) as a result of elevated activity of the pro-apoptotic gene *hid*. Under normal conditions, *hid* activity is randomly variable but does not typically lead to bristle loss. The authors propose that loss of *mir-263a/b* increases the chance of individual nascent IOB cells succumbing to apoptosis as robustness of the pruning process becomes impaired.

5.5 Candidate targets and their involvement in butterfly ovarian biology

Further insights into the role of each miRNA on ovarian biology came by comparing the putative target mRNA sequences identified bioinformatically with a manually curated list of proteins known to be involved in egg and ovary function (Carter et al., 2013). In each case, I suggest that the miRNA of interest act as fine-tuners of gene activity rather than on/off switches for gene expression. This would grant robustness to the system and buffer it against developmental perturbation (incoherent feedforward loops).

1. Hormone signalling

The interplay between hormone signalling and lepidopteran oogenesis has been studied in the domesticated silkmoth, *B. mori* (Papantonis, Swevers, & Iatrou, 2015; Ramaswamy, Shu, Park, & Zeng, 1997). Silkmoth oogenesis initiates during the pupal stage in response to 20E signalling from the pupal thoracic glands (Swevers & Iatrou, 2003). The nuclear receptor E75 is an early response gene in the ecdysone signalling pathway and is involved in the ovarian response to hormone signalling in both silkmoths (Papantonis et al., 2015) and flies (Ables, Bois, Garcia, & Drummond-Barbosa, 2015). I identified an *E75* mRNA splice variant, possibly *E75B*, as a potential target for two of my ovary set miRNAs (Table 5.5). Both the predicted miR-989 and B-0210/C-0160 target sites are located at the start codon of this transcript. *E75B* is thought to bind to and inhibit HR3, blocking the transactivation of downstream targets such as FTZ-F1. FTZ-F1 is a competence factor required for the initiation of

choriogenesis (Papantonis et al., 2015). This may provide a timing mechanism for the activation of FTZ-F1, which is dependent on clearance of E75B (Swevers, Ito, & Iatrou, 2002; White, Hurban, Watanabe, & Hogness, 1997).

2. Follicle cell function

Follicle cells perform several key functions during insect oogenesis. In *B. mori*, they are responsible for egg provisioning through the transport of vitellogenin from maternal haemolymph into the oocyte, as well as through the secretion of egg-specific protein (Swevers & Iatrou, 2003). Follicle cells also secrete maternal ecdysteroids into the oocyte for use during embryogenesis (Kadono-okuda, Amornsak, & Yamashita, 1994), and lay down the chorion and eggshell. Follicle cells in flies are also known to provide spatial patterning information to the oocyte (Saunders & Cohen, 1999).

In *Drosophila*, a subset of follicle cells detaches from the follicular epithelium and migrates between the nurse cells to the anterior border of the oocyte. These border cells contribute to both anterior patterning as well as the formation of the micropyle through which sperm enter the egg at fertilization (Montell, 2003). Several “ovary set” miRNAs, including miR-989 as mentioned above, are predicted to target genes involved in follicle cell function and border cell migration. Some are related to establishment of cell polarity such as *stardust (sdt)* and *crumbs (crb)*, essential to both integrity of the follicle cell layer and to border cell migration (Pinheiro & Montell, 2004). Other targets include *neuroglian (nrg)* which is required for efficient delamination of border cells prior to migration (Wei et al., 2004) and *taiman (tai)*, a coactivator for the ecdysone receptor (McDonald, Pinheiro, & Montell, 2003).

3. Serosa and immune defence

Once laid, insect eggs require protection against outside elements including temperature extremes, relative humidity and pathogen infection (Panfilio, 2008). One component of this defence mechanism is the serosa, an extraembryonic membrane that protects the embryo from desiccation. The *Shx* genes encode a set of homeobox proteins thought to be involved both in specifying the location of the extra-embryonic serosal tissue domain in developing oocytes, as well as regulating its subsequent development. These genes originated by tandem duplication of *zen* early in lepidopteran evolution. I predict that miR-989 may target the mRNA of *ShxD*, a gene

expressed maternally in ovarioles and zygotically in the serosa (Ferguson et al., 2014). As noted above, miR-263a may also be involved in immune defence, through different targets.

6. Supplementary Information

Supplement S5.1

Read count comparison data for the pooled miRNA transcriptome (Chapter 3) and the ovarian transcriptomes from St Hubert and Zonza. Read counts are normalised to the total number of reads mapped as miRNAs in each transcriptome. FC = fold change relative to pooled transcriptome.

	Pooled		St Hubert (Belgium)			Zonza (Corsica)		
	Total	Norm.	Total	Norm.	FC	Total	Norm.	FC
B-0042	0	0	472	1.26E-03	N.A.	0	0.00E+00	N.A.
B-0070/C-0072	0	0	291	7.77E-04	N.A.	574	5.47E-04	N.A.
B-0112	0	0	173	4.62E-04	N.A.	0	0.00E+00	N.A.
B-0164	0	0	90	2.40E-04	N.A.	0	0.00E+00	N.A.
B-0210/C-0160	0	0	97	2.59E-04	N.A.	262	2.50E-04	N.A.
B-0218	0	0	80	2.14E-04	N.A.	0	0.00E+00	N.A.
B-0298	0	0	91	2.43E-04	N.A.	0	0.00E+00	N.A.
B-0319	0	0	48	1.28E-04	N.A.	0	0.00E+00	N.A.
B-0401/C-0173	0	0	27	7.21E-05	N.A.	137	1.30E-04	N.A.
C-0149	0	0	0	0.00E+00	N.A.	378	3.60E-04	N.A.
C-0261	0	0	0	0.00E+00	N.A.	56	5.33E-05	N.A.
Pae-bantam	6734	1.86E-03	576	1.54E-03	0.8	771	7.34E-04	0.4
Pae-let-7	471778	1.30E-01	1895	5.06E-03	0.0	6133	5.84E-03	0.0
Pae-miR-1	1611661	4.45E-01	21127	5.64E-02	0.1	26920	2.56E-02	0.1
Pae-miR-1000	125	3.45E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-10a	78187	2.16E-02	473	1.26E-03	0.1	658	6.27E-04	0.0
Pae-miR-11	1412	3.90E-04	79	2.11E-04	0.5	227	2.16E-04	0.6
Pae-miR-1175	310	8.56E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-12	849	2.34E-04	47	1.26E-04	0.5	103	9.81E-05	0.4
Pae-miR-124	1265	3.49E-04	4	1.07E-05	0.0	23	2.19E-05	0.1
Pae-miR-133	279	7.70E-05	9	2.40E-05	0.3	0	0.00E+00	0.0
Pae-miR-137	23	6.35E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-13a	1534	4.23E-04	38	1.01E-04	0.2	88	8.38E-05	0.2
Pae-miR-13b	333	9.19E-05	15	4.01E-05	0.4	42	4.00E-05	0.4
Pae-miR-184	136896	3.78E-02	4805	1.28E-02	0.3	10233	9.75E-03	0.3
Pae-miR-190	4251	1.17E-03	585	1.56E-03	1.3	800	7.62E-04	0.6
Pae-miR-193	183	5.05E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-210	513	1.42E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-252	6140	1.69E-03	19	5.07E-05	0.0	115	1.10E-04	0.1
Pae-miR-263a	53155	1.47E-02	24821	6.63E-02	4.5	63593	6.06E-02	4.1
Pae-miR-263b	54	1.49E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-274	64	1.77E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-2755	56422	1.56E-02	4440	1.19E-02	0.8	10312	9.82E-03	0.6
Pae-miR-2756	26489	7.31E-03	1406	3.76E-03	0.5	3516	3.35E-03	0.5

Pae-miR-276	36461	1.01E-02	384	1.03E-03	0.1	580	5.52E-04	0.1
Pae-miR-2765	4745	1.31E-03	1	2.67E-06	0.0	42	4.00E-05	0.0
Pae-miR-2766	88117	2.43E-02	1628	4.35E-03	0.2	4035	3.84E-03	0.2
Pae-miR-2767	39977	1.10E-02	51	1.36E-04	0.0	1224	1.17E-03	0.1
Pae-miR-2768	452	1.25E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-277	4616	1.27E-03	81	2.16E-04	0.2	106	1.01E-04	0.1
Pae-miR-278	9065	2.50E-03	625	1.67E-03	0.7	1536	1.46E-03	0.6
Pae-miR-2788	105	2.90E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-2796	5006	1.38E-03	2	5.34E-06	0.0	21	2.00E-05	0.0
Pae-miR-2797a	9503	2.62E-03	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-2797b	6120	1.69E-03	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-2797c	2002	5.53E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-279a	18068	4.99E-03	1854	4.95E-03	1.0	3188	3.04E-03	0.6
Pae-miR-279b	30680	8.47E-03	2051	5.48E-03	0.6	4564	4.35E-03	0.5
Pae-miR-279c	847	2.34E-04	53	1.42E-04	0.6	124	1.18E-04	0.5
Pae-miR-279d	27122	7.49E-03	2374	6.34E-03	0.8	4410	4.20E-03	0.6
Pae-miR-281	47901	1.32E-02	1292	3.45E-03	0.3	2067	1.97E-03	0.1
Pae-miR-282	852	2.35E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-283	156	4.31E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-285	2503	6.91E-04	3	8.01E-06	0.0	7	6.67E-06	0.0
Pae-miR-2a	9979	2.75E-03	111	2.96E-04	0.1	424	4.04E-04	0.1
Pae-miR-2b	2183	6.02E-04	37	9.88E-05	0.2	60	5.71E-05	0.1
Pae-miR-2c	10186	2.81E-03	71	1.90E-04	0.1	162	1.54E-04	0.1
Pae-miR-3000	46	1.27E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-307	415	1.15E-04	12	3.20E-05	0.3	15	1.43E-05	0.1
Pae-miR-308	5385	1.49E-03	576	1.54E-03	1.0	1608	1.53E-03	1.0
Pae-miR-31	64433	1.78E-02	959	2.56E-03	0.1	2418	2.30E-03	0.1
Pae-miR-316	130	3.59E-05	8	2.14E-05	0.6	0	0.00E+00	0.0
Pae-miR-317	6220	1.72E-03	512	1.37E-03	0.8	598	5.70E-04	0.3
Pae-miR-3286	323	8.91E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-33	254	7.01E-05	12	3.20E-05	0.5	18	1.71E-05	0.2
Pae-miR-3338	21	5.80E-06	0	0.00E+00	0.0	13	1.24E-05	2.1
Pae-miR-34	13370	3.69E-03	657	1.75E-03	0.5	956	9.11E-04	0.2
Pae-miR-375	1217	3.36E-04	6	1.60E-05	0.0	20	1.90E-05	0.1
Pae-miR-6094	496	1.37E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-6307	624	1.72E-04	59	1.58E-04	0.9	132	1.26E-04	0.7
Pae-miR-7	391	1.08E-04	44	1.18E-04	1.1	43	4.10E-05	0.4
Pae-miR-745	1883	5.20E-04	19	5.07E-05	0.1	40	3.81E-05	0.1
Pae-miR-750	1945	5.37E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-8	634	1.75E-04	78	2.08E-04	1.2	290	2.76E-04	1.6
Pae-miR-87	346	9.55E-05	54	1.44E-04	1.5	63	6.00E-05	0.6
Pae-miR-9	9534	2.63E-03	130	3.47E-04	0.1	217	2.07E-04	0.1
Pae-miR-927	167	4.61E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-929	11	3.04E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-92a	6220	1.72E-03	198	5.29E-04	0.3	426	4.06E-04	0.2
Pae-miR-92b	1640	4.53E-04	0	0.00E+00	0.0	110	1.05E-04	0.2

Pae-miR-932	126	3.48E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-965	352	9.71E-05	32	8.55E-05	0.9	52	4.95E-05	0.5
Pae-miR-970	8319	2.30E-03	194	5.18E-04	0.2	593	5.65E-04	0.2
Pae-miR-971	18	4.97E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-981	18	4.97E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Pae-miR-988	84	2.32E-05	9	2.40E-05	1.0	13	1.24E-05	0.5
Pae-miR-989	538441	1.49E-01	288099	7.69E-01	5.2	873884	8.32E-01	5.6
Pae-miR-993	37297	1.03E-02	4611	1.23E-02	1.2	8301	7.91E-03	0.8
Pae-miR-998	1389	3.83E-04	52	1.39E-04	0.4	127	1.21E-04	0.3
Pae-miR-iab-4	706	1.95E-04	15	4.01E-05	0.2	33	3.14E-05	0.2
Pae-miR-iab-8	30	8.28E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-013	42783	1.18E-02	1714	4.58E-03	0.4	4723	4.50E-03	0.4
Par-017	41987	1.16E-02	1935	5.17E-03	0.4	5586	5.32E-03	0.5
Par-058	2394	6.61E-04	185	4.94E-04	0.7	342	3.26E-04	0.5
Par-080	810	2.24E-04	70	1.87E-04	0.8	171	1.63E-04	0.7
Par-094	100	2.76E-05	0	0.00E+00	0.0	128	1.22E-04	4.4
Par-099	640	1.77E-04	46	1.23E-04	0.7	91	8.67E-05	0.5
Par-104	905	2.50E-04	38	1.01E-04	0.4	73	6.95E-05	0.3
Par-110	663	1.83E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-111	291	8.03E-05	85	2.27E-04	2.8	128	1.22E-04	1.5
Par-117	673	1.86E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-118	82	2.26E-05	19	5.07E-05	2.2	67	6.38E-05	2.8
Par-129	752	2.08E-04	148	3.95E-04	1.9	577	5.50E-04	2.6
Par-145	373	1.03E-04	237	6.33E-04	6.1	0	0.00E+00	0.0
Par-146	333	9.19E-05	11	2.94E-05	0.3	60	5.71E-05	0.6
Par-152	372	1.03E-04	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-160	248	6.84E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-168	365	1.01E-04	0	0.00E+00	0.0	60	5.71E-05	0.6
Par-181	144	3.97E-05	0	0.00E+00	0.0	8	7.62E-06	0.2
Par-189	87	2.40E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-211	114	3.15E-05	11	2.94E-05	0.9	0	0.00E+00	0.0
Par-212	125	3.45E-05	63	1.68E-04	4.9	0	0.00E+00	0.0
Par-220	95	2.62E-05	3	8.01E-06	0.3	10	9.52E-06	0.4
Par-222	81	2.24E-05	57	1.52E-04	6.8	14	1.33E-05	0.6
Par-240	75	2.07E-05	21	5.61E-05	2.7		0.00E+00	0.0
Par-247	44	1.21E-05	125	3.34E-04	27.5	146	1.39E-04	11.5
Par-259	77	2.13E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-262	59	1.63E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-268	47	1.30E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-269	92	2.54E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-280	57	1.57E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-284	54	1.49E-05	15	4.01E-05	2.7	29	2.76E-05	1.9
Par-290	33	9.11E-06	8	2.14E-05	2.3	0	0.00E+00	0.0
Par-305	51	1.41E-05	6	1.60E-05	1.1	0	0.00E+00	0.0
Par-324	50	1.38E-05	0	0.00E+00	0.0	9	8.57E-06	0.6
Par-340	49	1.35E-05	52	1.39E-04	10.3	70	6.67E-05	4.9

Par-341	48	1.32E-05	52	1.39E-04	10.5	70	6.67E-05	5.0
Par-350	21	5.80E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-371	23	6.35E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-377	19	5.24E-06	3	8.01E-06	1.5	0	0.00E+00	0.0
Par-404	19	5.24E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-405	19	5.24E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-425	38	1.05E-05	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-426	36	9.94E-06	13	3.47E-05	3.5	0	0.00E+00	0.0
Par-471	27	7.45E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-523	20	5.52E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-536	10	2.76E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-542	11	3.04E-06	0	0.00E+00	0.0	97	9.24E-05	30.4
Par-543	14	3.86E-06	0	0.00E+00	0.0	5	4.76E-06	1.2
Par-554	14	3.86E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-567	15	4.14E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-602	11	3.04E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-605	7	1.93E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-614	8	2.21E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-634	17	4.69E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-658	6	1.66E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-659	16	4.42E-06	6	1.60E-05	3.6	0	0.00E+00	0.0
Par-700	10	2.76E-06	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-737	3	8.28E-07	0	0.00E+00	0.0	0	0.00E+00	0.0
Par-738	9	2.48E-06	5	1.34E-05	5.4	0	0.00E+00	0.0

Supplement S5.2

Alignment of the newly-identified B-0401/C-0173 miRNA with known precursor sequences for miR-2763 across the Lepidoptera. The *P. aegeria* sequence was predicted from the genome but not identified in the deep sequencing data presented in Chapter 3

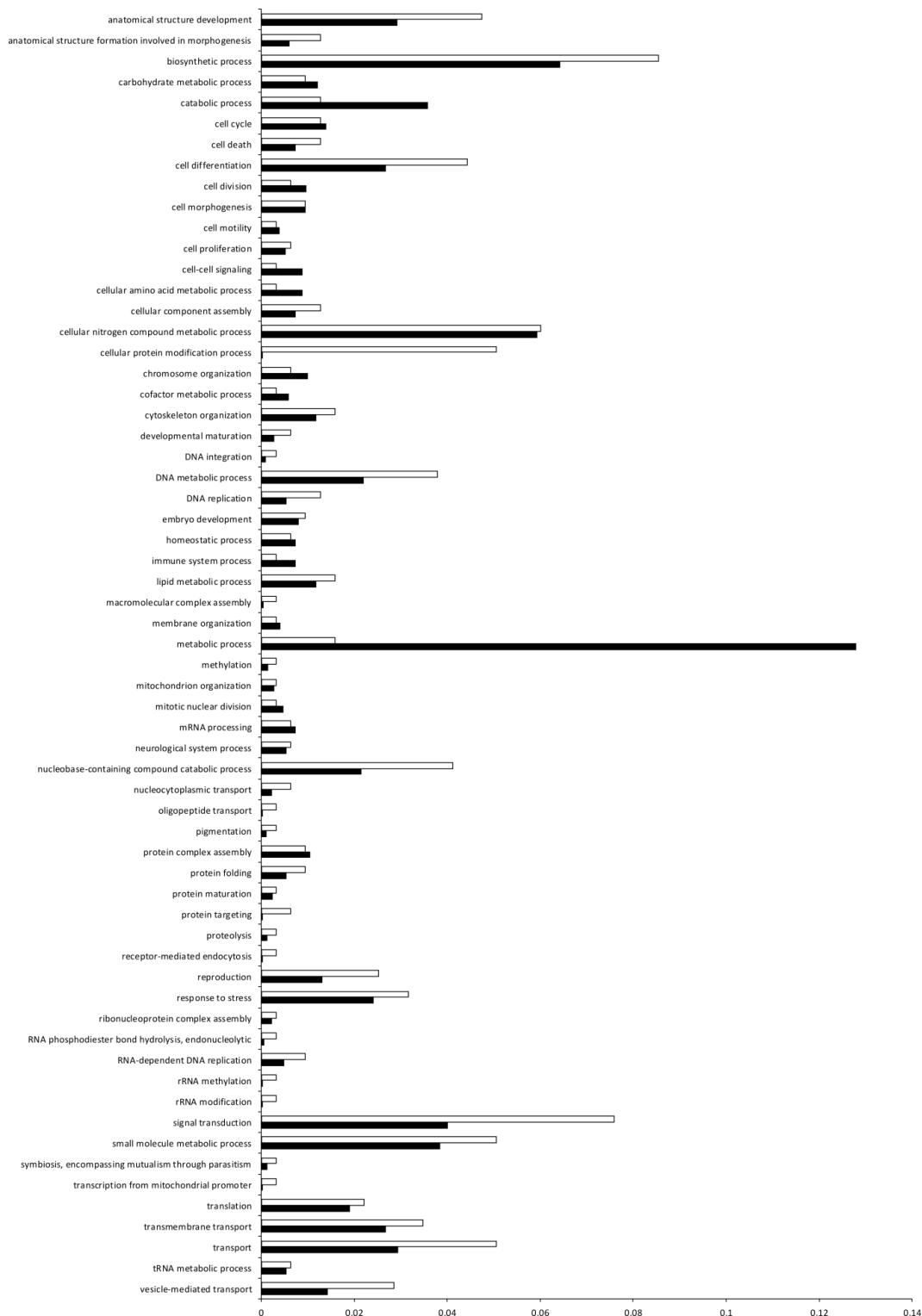
```

B-0401/C-0173      AGCUGUGCUAAUUAACAUAUCCUCACAAUUUUUCCAAAGUAGUGGGCGUAAUUUAUCGUGAG  --CUAGCUAUAGAUAAUUUAUGCUCAUUACUUUGGAUAUGUUGUGGAA  -----CUUACAUC
P. aegeria (prediction)  -----AUUACAUCUACAUAUUUCCAAAGUAGUGGGCGUAAUUUAUCGUGAG  --CUAGCUAUAGAUAAUUUAUGCUCAUUACUUUGGAUAUGUUGUGGAA  -----
D. plexippus          -----UCCAAAGUAGUGGACGUAUUUAUCGUGAG  --CCAACUAUAGAUAAUUUAUGCUCAUUACUUUGGAUAUGUUGUGCG  --
H. melipomene        ---CGUAUAAUUUAUUAUCUACAACGUCACCCAAAGUAGUAGACGUAUUUAUCGUGAG  --CUAUCUAUAGAUAAUUUAUGCUCAUUACUUUGGAUAUGUUGUGGAA  -----
B. mori              -----GGCAUUGUACUCGCAAGUAGUAGAAUUAUUAUCGUGAGAUUUUAUUAUGCUCAUUUAUCUUUGGAUAUGUUGUGGAA  -----
M. sexta             -----UACAAUACACUGGGCAUUAUCUCCAAAGUAGUAGAAUUAUUAUCGUGAGAUUUUAUUAUGCUCAUUUAUCUUUGGAUAUGUUGUGGAA  -----
P. c-album           -----AAUUUAUUCGGGCUACAUAUCCCAAGUAGUAGACGUAUUUAUCGUGGA  --AUGAGCCU--GAUAUUUAUGCUCAUUACUUUGGAUAUGUUGAGCGA  -----

```

Supplement S5.3

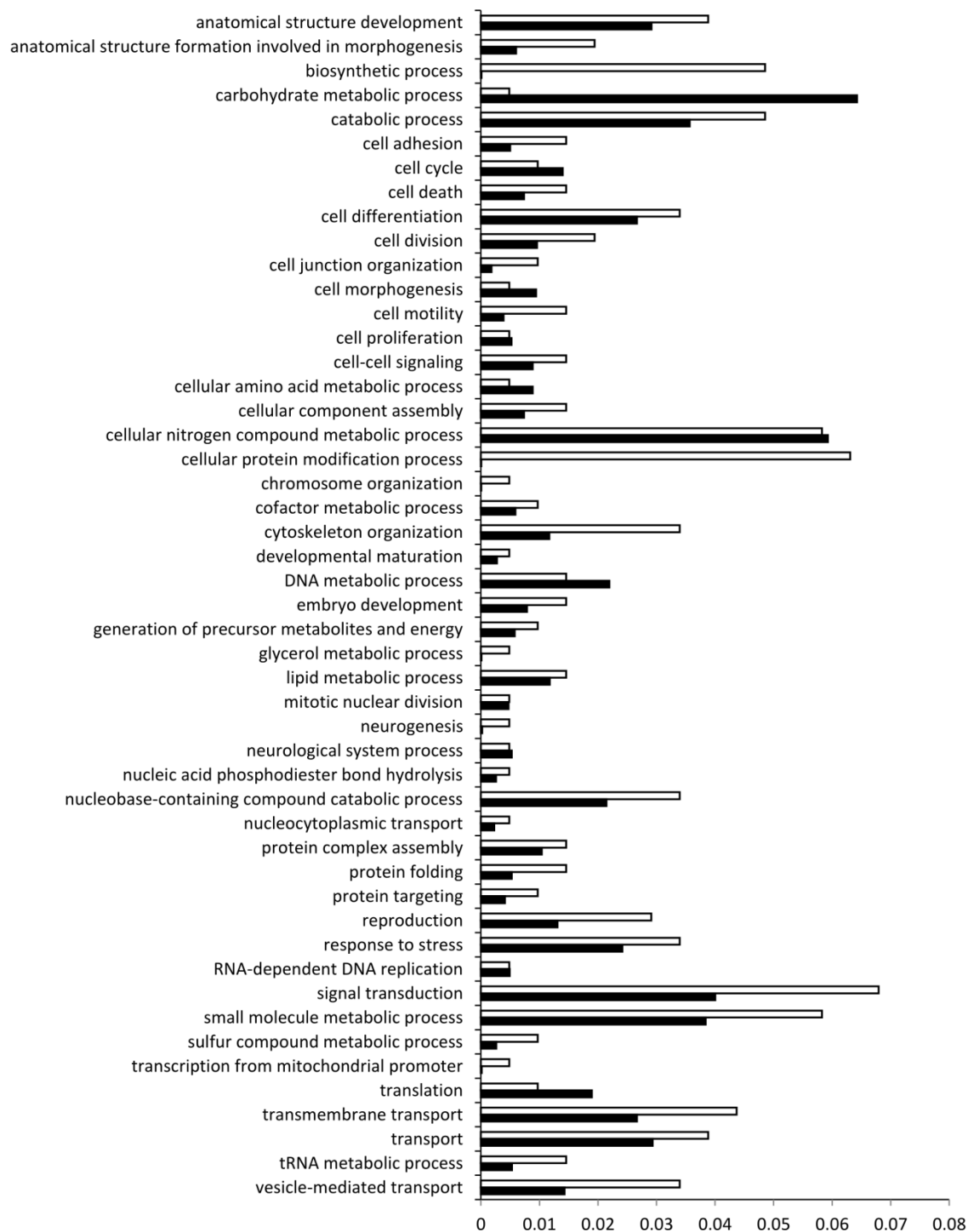
GO term distributions for the target prediction dataset of each 'ovary set' miRNA relative to the combined transcriptome assembly

miR-263a-5p

GO term	Combined		miR-263a-5p	
	Raw	Normalised	Raw	Normalised
anatomical structure development	15	0.029128796	657	0.047468354
anatomical structure formation involved in morphogenesis	4	0.006029705	136	0.012658228
biosynthetic process	27	0.064242962	1449	0.085443038
carbohydrate metabolic process	3	0.012148082	274	0.009493671
catabolic process	4	0.035690534	805	0.012658228
cell cycle	4	0.013965861	315	0.012658228
cell death	4	0.007359787	166	0.012658228
cell differentiation	14	0.026645977	601	0.044303797
cell division	2	0.009576591	216	0.006329114
cell morphogenesis	3	0.009443582	213	0.009493671
cell motility	1	0.003857238	87	0.003164557
cell proliferation	2	0.005231656	118	0.006329114
cell-cell signaling	1	0.008867213	200	0.003164557
cellular amino acid metabolic process	1	0.008867213	200	0.003164557
cellular component assembly	4	0.007359787	166	0.012658228
cellular nitrogen compound metabolic process	19	0.059277322	1337	0.060126582
cellular protein modification process	16	4.43361E-05	1	0.050632911
chromosome organization	2	0.009975615	225	0.006329114
cofactor metabolic process	1	0.005896697	133	0.003164557
cytoskeleton organization	5	0.011704722	264	0.015822785
developmental maturation	2	0.002748836	62	0.006329114
DNA integration	1	0.000931057	21	0.003164557
DNA metabolic process	12	0.021946353	495	0.037974684
DNA replication	4	0.005409	122	0.012658228
embryo development	3	0.00789182	178	0.009493671
homeostatic process	2	0.007315451	165	0.006329114
immune system process	1	0.007359787	166	0.003164557
lipid metabolic process	5	0.011749058	265	0.015822785
macromolecular complex assembly	1	0.000399025	9	0.003164557
membrane organization	1	0.003990246	90	0.003164557
metabolic process	5	0.127865218	2884	0.015822785
methylation	1	0.00146309	33	0.003164557
mitochondrion organization	1	0.002748836	62	0.003164557
mitotic nuclear division	1	0.004743959	107	0.003164557
mRNA processing	2	0.007226779	163	0.006329114
neurological system process	2	0.005275992	119	0.006329114
nucleobase-containing compound catabolic process	13	0.021458657	484	0.041139241
nucleocytoplasmic transport	2	0.002261139	51	0.006329114
oligopeptide transport	1	4.43361E-05	1	0.003164557
pigmentation	1	0.001064066	24	0.003164557
protein complex assembly	3	0.010418976	235	0.009493671
protein folding	3	0.005275992	119	0.009493671
protein maturation	1	0.002305476	52	0.003164557
protein targeting	2	8.86721E-05	2	0.006329114

proteolysis	1	0.001152738	26	0.003164557
receptor-mediated endocytosis	1	0.000133008	3	0.003164557
reproduction	8	0.01307914	295	0.025316456
response to stress	10	0.024163157	545	0.03164557
ribonucleoprotein complex assembly	1	0.002261139	51	0.003164557
RNA phosphodiester bond hydrolysis, endonucleolytic	1	0.000487697	11	0.003164557
RNA-dependent DNA replication	3	0.004921303	111	0.009493671
rRNA methylation	1	0.000177344	4	0.003164557
rRNA modification	1	0.000177344	4	0.003164557
signal transduction	24	0.040035469	903	0.075949367
small molecule metabolic process	16	0.038395034	866	0.050632911
symbiosis, encompassing mutualism through parasitism	1	0.001197074	27	0.003164557
transcription from mitochondrial promoter	1	0.000133008	3	0.003164557
translation	7	0.018975837	428	0.022151899
transmembrane transport	11	0.026645977	601	0.034810127
transport	16	0.029350477	662	0.050632911
tRNA metabolic process	2	0.005320328	120	0.006329114
vesicle-mediated transport	9	0.014276214	322	0.028481013

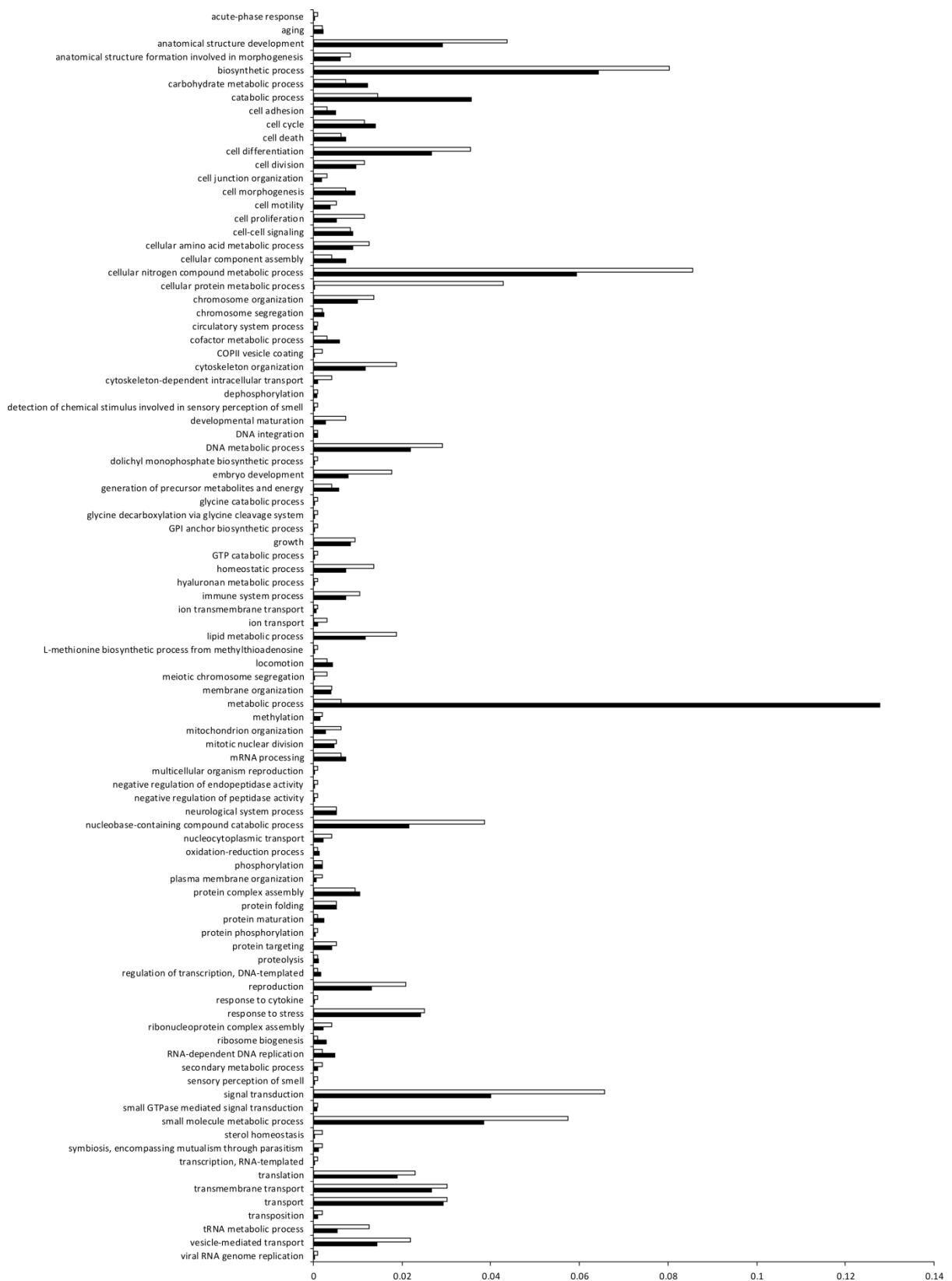
miR-263a-3p



GO term	Combined		miR-263a-3p	
	Raw	Normalised	Raw	Normalised
anatomical structure development	657	0.0291288	8	0.03883495
anatomical structure formation involved in morphogenesis	136	0.0060297	4	0.01941748
biosynthetic process	2	8.867E-05	10	0.04854369
carbohydrate metabolic process	1449	0.064243	1	0.00485437
catabolic process	805	0.0356905	10	0.04854369
cell adhesion	113	0.00501	3	0.01456311
cell cycle	315	0.0139659	2	0.00970874
cell death	166	0.0073598	3	0.01456311
cell differentiation	601	0.026646	7	0.03398058
cell division	216	0.0095766	4	0.01941748
cell junction organization	41	0.0018178	2	0.00970874
cell morphogenesis	213	0.0094436	1	0.00485437
cell motility	87	0.0038572	3	0.01456311
cell proliferation	118	0.0052317	1	0.00485437
cell-cell signaling	200	0.0088672	3	0.01456311
cellular amino acid metabolic process	200	0.0088672	1	0.00485437
cellular component assembly	166	0.0073598	3	0.01456311
cellular nitrogen compound metabolic process	1337	0.0592773	12	0.05825243
cellular protein modification process	1	4.434E-05	13	0.0631068
chromosome organization	1	4.434E-05	1	0.00485437
cofactor metabolic process	133	0.0058967	2	0.00970874
cytoskeleton organization	264	0.0117047	7	0.03398058
developmental maturation	62	0.0027488	1	0.00485437
DNA metabolic process	495	0.0219464	3	0.01456311
embryo development	178	0.0078918	3	0.01456311
generation of precursor metabolites and energy	130	0.0057637	2	0.00970874
glycerol metabolic process	2	8.867E-05	1	0.00485437
lipid metabolic process	265	0.0117491	3	0.01456311
mitotic nuclear division	107	0.004744	1	0.00485437
neurogenesis	5	0.0002217	1	0.00485437
neurological system process	119	0.005276	1	0.00485437
nucleic acid phosphodiester bond hydrolysis	58	0.0025715	1	0.00485437
nucleobase-containing compound catabolic process	484	0.0214587	7	0.03398058
nucleocytoplasmic transport	51	0.0022611	1	0.00485437
protein complex assembly	235	0.010419	3	0.01456311
protein folding	119	0.005276	3	0.01456311
protein targeting	93	0.0041233	2	0.00970874
reproduction	295	0.0130791	6	0.02912621
response to stress	545	0.0241632	7	0.03398058
RNA-dependent DNA replication	111	0.0049213	1	0.00485437
signal transduction	903	0.0400355	14	0.06796117
small molecule metabolic process	866	0.038395	12	0.05825243
sulfur compound metabolic process	59	0.0026158	2	0.00970874
transcription from mitochondrial promoter	3	0.000133	1	0.00485437

translation	428	0.0189758	2	0.00970874
transmembrane transport	601	0.026646	9	0.04368932
transport	662	0.0293505	8	0.03883495
tRNA metabolic process	120	0.0053203	3	0.01456311
vesicle-mediated transport	322	0.0142762	7	0.03398058

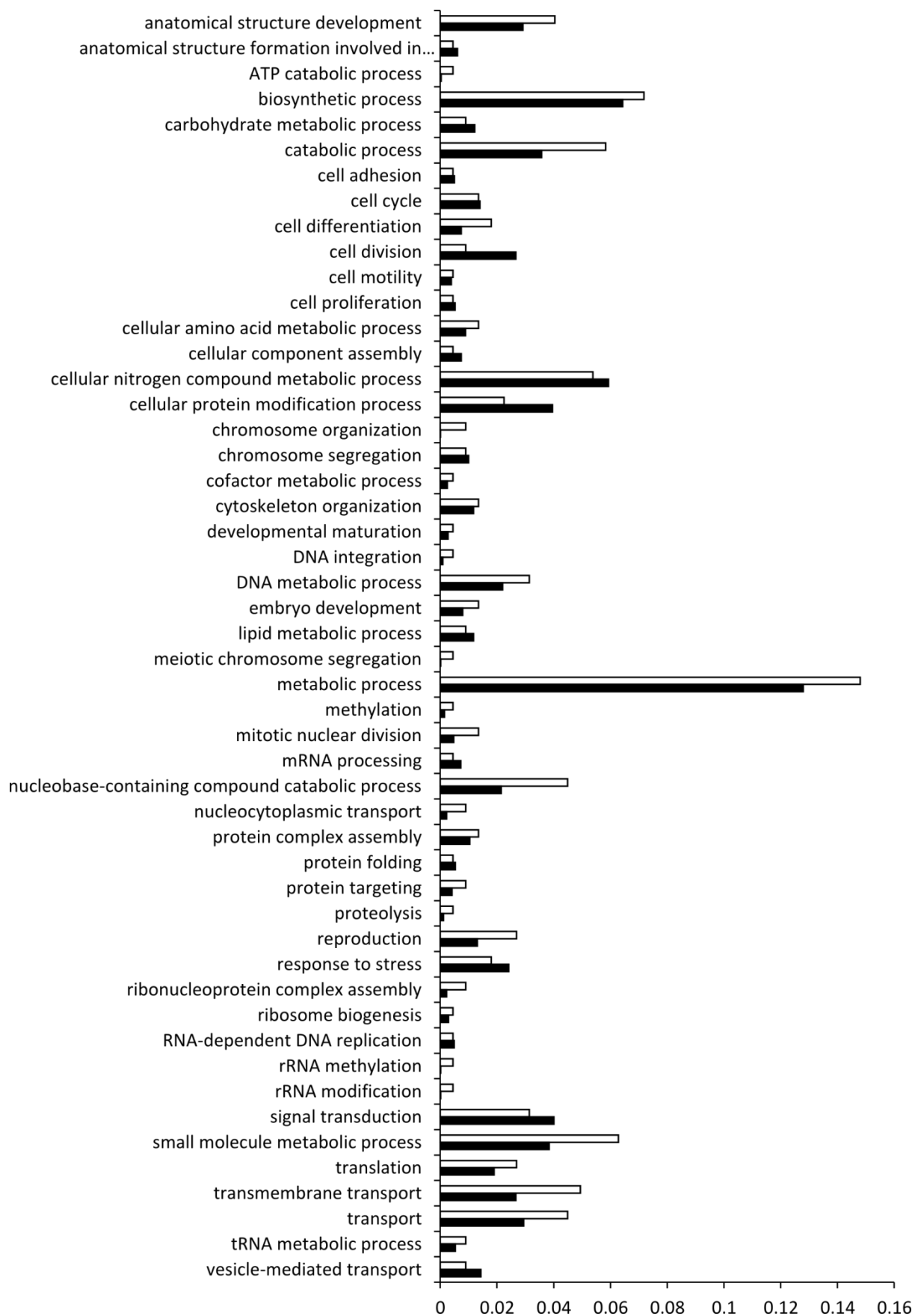
miR-989-5p



GO term	Raw	Combined	Raw	miR-989-5p
		Normalised		Normalised
acute-phase response	1	0.0010428	2	8.86721E-05
aging	2	0.0020855	49	0.002172467
anatomical structure development	42	0.0437956	657	0.029128796
anatomical structure formation involved in morphogenesis	8	0.008342	136 144	0.006029705
biosynthetic process	77	0.080292	9	0.064242962
carbohydrate metabolic process	7	0.0072993	274	0.012148082
catabolic process	14	0.0145985	805	0.035690534
cell adhesion	3	0.0031283	113	0.005009976
cell cycle	11	0.0114703	315	0.013965861
cell death	6	0.0062565	166	0.007359787
cell differentiation	34	0.0354536	601	0.026645977
cell division	11	0.0114703	216	0.009576591
cell junction organization	3	0.0031283	41	0.001817779
cell morphogenesis	7	0.0072993	213	0.009443582
cell motility	5	0.0052138	87	0.003857238
cell proliferation	11	0.0114703	118	0.005231656
cell-cell signaling	8	0.008342	200	0.008867213
cellular amino acid metabolic process	12	0.012513	200	0.008867213
cellular component assembly	4	0.004171	166 133	0.007359787
cellular nitrogen compound metabolic process	82	0.0855057	7	0.059277322
cellular protein modification process	41	0.0427529	892	0.039547772
chromosome organization	13	0.0135558	225	0.009975615
chromosome segregation	2	0.0020855	55	0.002438484
circulatory system process	1	0.0010428	19	0.000842385
cofactor metabolic process	3	0.0031283	133	0.005896697
COPII vesicle coating	2	0.0020855	7	0.000310352
cytoskeleton organization	18	0.0187696	264	0.011704722
cytoskeleton-dependent intracellular transport	4	0.004171	23	0.00101973
dephosphorylation	1	0.0010428	17	0.000753713
detection of chemical stimulus involved in sensory perception of smell	1	0.0010428	3	0.000133008
developmental maturation	7	0.0072993	62	0.002748836
DNA integration	1	0.0010428	1	4.43361E-05
DNA metabolic process	28	0.0291971	495	0.021946353
dolichyl monophosphate biosynthetic process	1	0.0010428	2	8.86721E-05
embryo development	17	0.0177268	178	0.00789182
generation of precursor metabolites and energy	4	0.004171	130	0.005763689
glycine catabolic process	1	0.0010428	5	0.00022168
glycine decarboxylation via glycine cleavage system	1	0.0010428	5	0.00022168
GPI anchor biosynthetic process	1	0.0010428	1	4.43361E-05
growth	9	0.0093848	188	0.008335181
GTP catabolic process	1	0.0010428	7	0.000310352
homeostatic process	13	0.0135558	165	0.007315451
hyaluronan metabolic process	1	0.0010428	2	8.86721E-05

immune system process	10	0.0104275	166	0.007359787
ion transmembrane transport	1	0.0010428	12	0.000532033
ion transport	3	0.0031283	22	0.000975393
lipid metabolic process	18	0.0187696	265	0.011749058
L-methionine biosynthetic process from methylthioadenosine	1	0.0010428	3	0.000133008
locomotion	3	0.0031283	97	0.004300599
meiotic chromosome segregation	3	0.0031283	4	0.000177344
membrane organization	4	0.004171	90	0.003990246
metabolic process	6	0.0062565	288	0.127865218
methylation	2	0.0020855	4	0.00146309
mitochondrion organization	6	0.0062565	33	0.002748836
mitotic nuclear division	6	0.0062565	62	0.002748836
mRNA processing	5	0.0052138	107	0.004743959
multicellular organism reproduction	6	0.0062565	163	0.007226779
negative regulation of endopeptidase activity	1	0.0010428	2	8.86721E-05
negative regulation of peptidase activity	1	0.0010428	3	0.000133008
neurological system process	1	0.0010428	1	4.43361E-05
nucleobase-containing compound catabolic process	5	0.0052138	119	0.005275992
nucleocytoplasmic transport	37	0.0385819	484	0.021458657
oxidation-reduction process	4	0.004171	51	0.002261139
phosphorylation	1	0.0010428	30	0.001330082
plasma membrane organization	2	0.0020855	47	0.002083795
protein complex assembly	2	0.0020855	13	0.000576369
protein folding	9	0.0093848	235	0.010418976
protein maturation	5	0.0052138	119	0.005275992
protein phosphorylation	1	0.0010428	52	0.002305476
protein targeting	1	0.0010428	11	0.000487697
proteolysis	5	0.0052138	93	0.004123254
regulation of transcription, DNA-templated	1	0.0010428	26	0.001152738
reproduction	1	0.0010428	38	0.001684771
response to cytokine	20	0.0208551	295	0.01307914
response to stress	1	0.0010428	2	8.86721E-05
ribonucleoprotein complex assembly	24	0.0250261	545	0.024163157
ribosome biogenesis	4	0.004171	51	0.002261139
RNA-dependent DNA replication	1	0.0010428	65	0.002881844
secondary metabolic process	2	0.0020855	111	0.004921303
sensory perception of smell	2	0.0020855	20	0.000886721
signal transduction	1	0.0010428	3	0.000133008
small GTPase mediated signal transduction	63	0.0656934	903	0.040035469
small molecule metabolic process	1	0.0010428	16	0.000709377
sulfur compound metabolic process	55	0.0573514	866	0.038395034
symbiosis, encompassing mutualism through parasitism	2	0.0020855	59	0.002615828
transcription, RNA-templated	2	0.0020855	27	0.001197074
translation	1	0.0010428	1	4.43361E-05
transmembrane transport	22	0.0229406	428	0.018975837
transport	29	0.0302398	601	0.026645977
	29	0.0302398	662	0.029350477

transposition	2	0.0020855	23	0.00101973
tRNA metabolic process	12	0.012513	120	0.005320328
vesicle-mediated transport	21	0.0218978	322	0.014276214
viral RNA genome replication	1	0.0010428	1	4.43361E-05

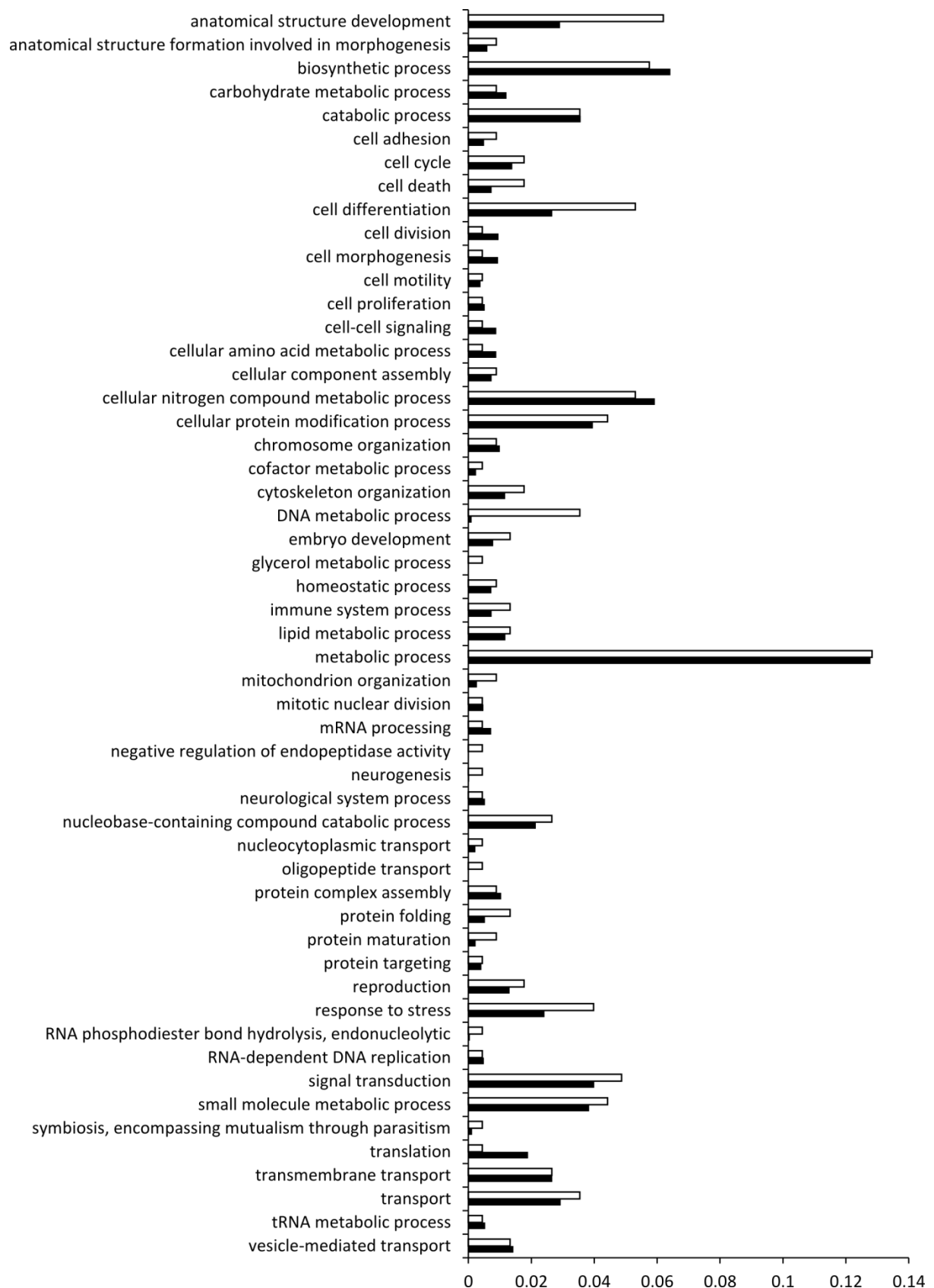
miR-2763-5p

GO term	Combined		miR-2763-5p	
	Raw	Normalised	Raw	Normalised
anatomical structure development	657	0.0291288	9	0.0403587
anatomical structure formation involved in morphogenesis	136	0.0060297	1	0.0044843
ATP catabolic process	6	0.000266	1	0.0044843
biosynthetic process	1449	0.064243	16	0.0717489
carbohydrate metabolic process	274	0.0121481	2	0.0089686
catabolic process	805	0.0356905	13	0.058296
cell adhesion	113	0.00501	1	0.0044843
cell cycle	315	0.0139659	3	0.0134529
cell differentiation	166	0.0073598	4	0.0179372
cell division	601	0.026646	2	0.0089686
cell motility	87	0.0038572	1	0.0044843
cell proliferation	118	0.0052317	1	0.0044843
cellular amino acid metabolic process	200	0.0088672	3	0.0134529
cellular component assembly	166	0.0073598	1	0.0044843
cellular nitrogen compound metabolic process	1337	0.0592773	12	0.0538117
cellular protein modification process	892	0.0395478	5	0.0224215
chromosome organization	1	4.434E-05	2	0.0089686
chromosome segregation	225	0.0099756	2	0.0089686
cofactor metabolic process	55	0.0024385	1	0.0044843
cytoskeleton organization	264	0.0117047	3	0.0134529
developmental maturation	62	0.0027488	1	0.0044843
DNA integration	21	0.0009311	1	0.0044843
DNA metabolic process	495	0.0219464	7	0.0313901
embryo development	178	0.0078918	3	0.0134529
lipid metabolic process	265	0.0117491	2	0.0089686
meiotic chromosome segregation	4	0.0001773	1	0.0044843
metabolic process	2884	0.1278652	33	0.1479821
methylation	33	0.0014631	1	0.0044843
mitotic nuclear division	107	0.004744	3	0.0134529
mRNA processing	163	0.0072268	1	0.0044843
nucleobase-containing compound catabolic process	484	0.0214587	10	0.044843
nucleocytoplasmic transport	51	0.0022611	2	0.0089686
protein complex assembly	235	0.010419	3	0.0134529
protein folding	119	0.005276	1	0.0044843
protein targeting	93	0.0041233	2	0.0089686
proteolysis	26	0.0011527	1	0.0044843
reproduction	295	0.0130791	6	0.0269058
response to stress	545	0.0241632	4	0.0179372
ribonucleoprotein complex assembly	51	0.0022611	2	0.0089686
ribosome biogenesis	65	0.0028818	1	0.0044843
RNA-dependent DNA replication	111	0.0049213	1	0.0044843
rRNA methylation	4	0.0001773	1	0.0044843
rRNA modification	4	0.0001773	1	0.0044843
signal transduction	903	0.0400355	7	0.0313901

small molecule metabolic process	866	0.038395	14	0.0627803
translation	428	0.0189758	6	0.0269058
transmembrane transport	601	0.026646	11	0.0493274
transport	662	0.0293505	10	0.044843
tRNA metabolic process	120	0.0053203	2	0.0089686
vesicle-mediated transport	322	0.0142762	2	0.0089686

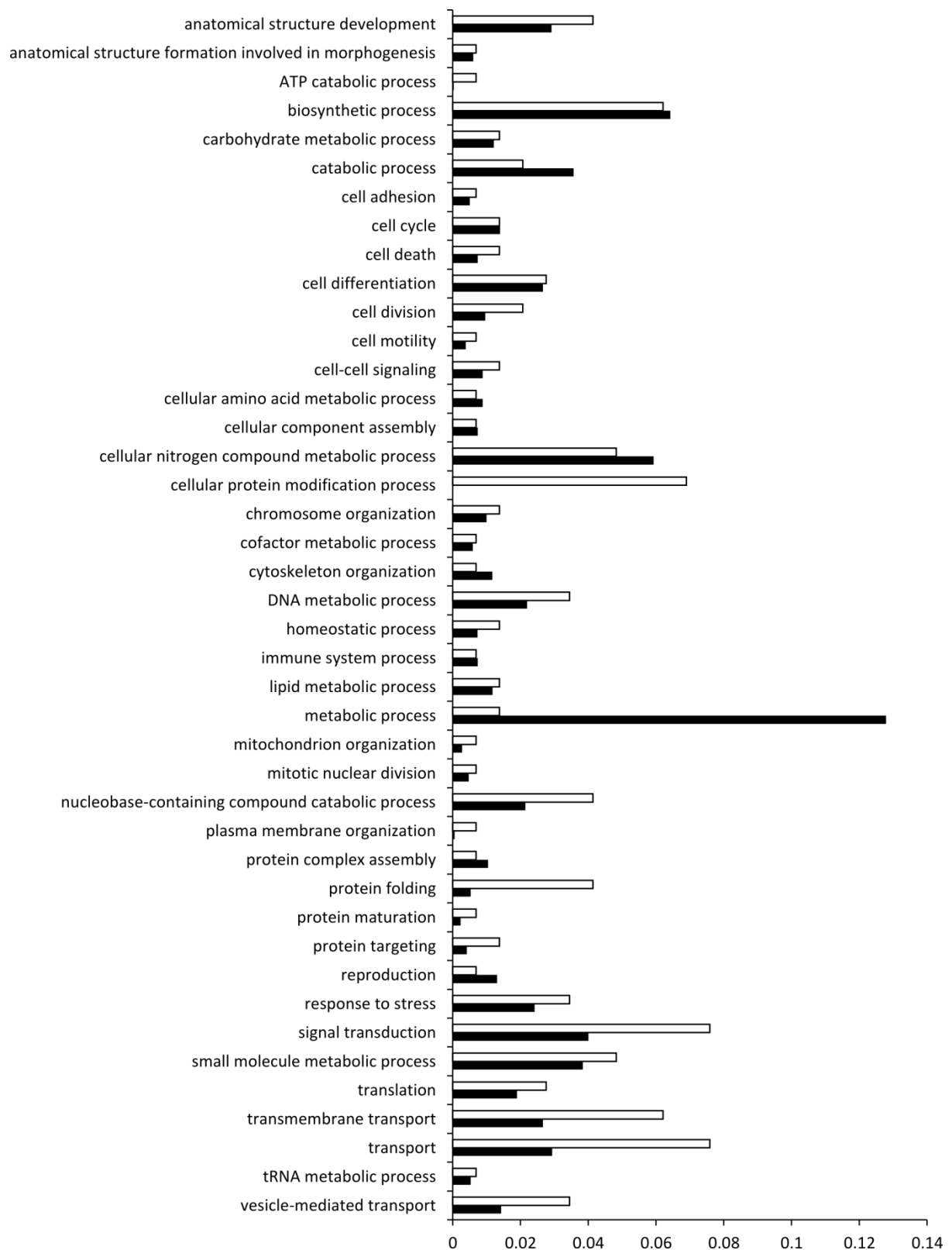
miR-2763-3p

GO term	#contigs
small molecule metabolic process	1
cellular protein modification process	1
signal transduction	1
nucleobase-containing compound catabolic process	1

B-0070/C-0072-3p

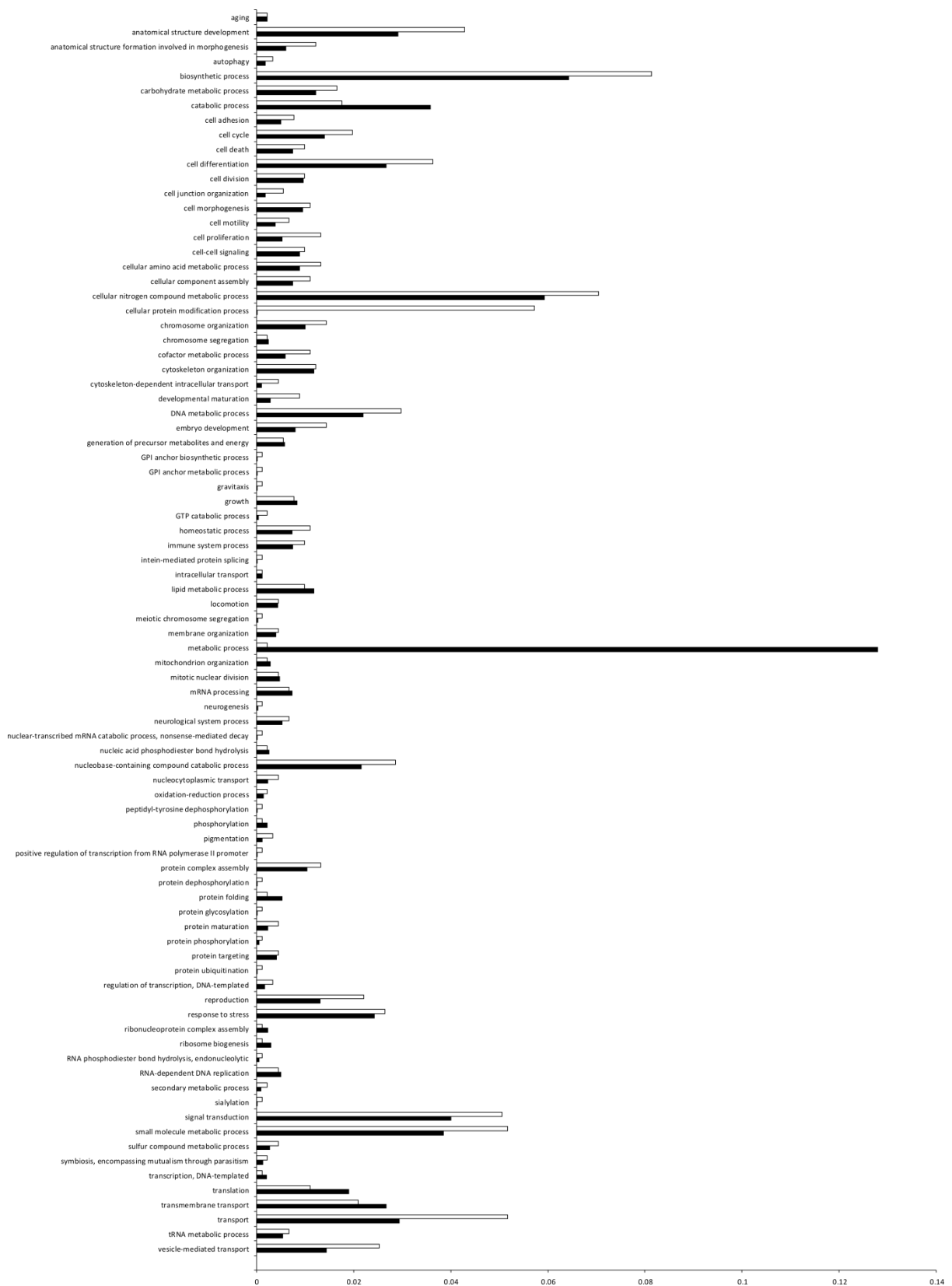
GO term	Combined		B-0070/C-0072-3p	
	Raw	Normalised	Raw	Normalised
anatomical structure development	657	0.029128796	14	0.061946903
anatomical structure formation involved in morphogenesis	136	0.006029705	2	0.008849558
biosynthetic process	1449	0.064242962	13	0.057522124
carbohydrate metabolic process	274	0.012148082	2	0.008849558
catabolic process	805	0.035690534	8	0.03539823
cell adhesion	113	0.005009976	2	0.008849558
cell cycle	315	0.013965861	4	0.017699115
cell death	166	0.007359787	4	0.017699115
cell differentiation	601	0.026645977	12	0.053097345
cell division	216	0.009576591	1	0.004424779
cell morphogenesis	213	0.009443582	1	0.004424779
cell motility	87	0.003857238	1	0.004424779
cell proliferation	118	0.005231656	1	0.004424779
cell-cell signaling	200	0.008867213	1	0.004424779
cellular amino acid metabolic process	200	0.008867213	1	0.004424779
cellular component assembly	166	0.007359787	2	0.008849558
cellular nitrogen compound metabolic process	1337	0.059277322	12	0.053097345
cellular protein modification process	892	0.039547772	10	0.044247788
chromosome organization	225	0.009975615	2	0.008849558
cofactor metabolic process	55	0.002438484	1	0.004424779
cytoskeleton organization	264	0.011704722	4	0.017699115
DNA metabolic process	23	0.00101973	8	0.03539823
embryo development	178	0.00789182	3	0.013274336
glycerol metabolic process	2	8.86721E-05	1	0.004424779
homeostatic process	165	0.007315451	2	0.008849558
immune system process	166	0.007359787	3	0.013274336
lipid metabolic process	265	0.011749058	3	0.013274336
metabolic process	2884	0.127865218	29	0.128318584
mitochondrion organization	62	0.002748836	2	0.008849558
mitotic nuclear division	107	0.004743959	1	0.004424779
mRNA processing	163	0.007226779	1	0.004424779
negative regulation of endopeptidase activity	3	0.000133008	1	0.004424779
neurogenesis	5	0.00022168	1	0.004424779
neurological system process	119	0.005275992	1	0.004424779
nucleobase-containing compound catabolic process	484	0.021458657	6	0.026548673
nucleocytoplasmic transport	51	0.002261139	1	0.004424779
oligopeptide transport	1	4.43361E-05	1	0.004424779
protein complex assembly	235	0.010418976	2	0.008849558
protein folding	119	0.005275992	3	0.013274336
protein maturation	52	0.002305476	2	0.008849558
protein targeting	93	0.004123254	1	0.004424779
reproduction	295	0.01307914	4	0.017699115
response to stress	545	0.024163157	9	0.039823009
RNA phosphodiester bond hydrolysis, endonucleolytic	11	0.000487697	1	0.004424779

RNA-dependent DNA replication	111	0.004921303	1	0.004424779
signal transduction	903	0.040035469	11	0.048672566
small molecule metabolic process	866	0.038395034	10	0.044247788
symbiosis, encompassing mutualism through parasitism	27	0.001197074	1	0.004424779
translation	428	0.018975837	1	0.004424779
transmembrane transport	601	0.026645977	6	0.026548673
transport	662	0.029350477	8	0.03539823
tRNA metabolic process	120	0.005320328	1	0.004424779
vesicle-mediated transport	322	0.014276214	3	0.013274336

B-0070/C-0072-5p

GO term	Combined		B-0070/C-0072-5p	
	Raw	Normalised	Raw	Normalised
anatomical structure development	657	0.029128796	6	0.04137931
anatomical structure formation involved in morphogenesis	136	0.006029705	1	0.00689655
ATP catabolic process	6	0.000266016	1	0.00689655
biosynthetic process	1449	0.064242962	9	0.06206897
carbohydrate metabolic process	274	0.012148082	2	0.0137931
catabolic process	805	0.035690534	3	0.02068966
cell adhesion	113	0.005009976	1	0.00689655
cell cycle	315	0.013965861	2	0.0137931
cell death	166	0.007359787	2	0.0137931
cell differentiation	601	0.026645977	4	0.02758621
cell division	216	0.009576591	3	0.02068966
cell motility	87	0.003857238	1	0.00689655
cell-cell signaling	200	0.008867213	2	0.0137931
cellular amino acid metabolic process	200	0.008867213	1	0.00689655
cellular component assembly	166	0.007359787	1	0.00689655
cellular nitrogen compound metabolic process	1337	0.059277322	7	0.04827586
cellular protein modification process	1	4.43361E-05	10	0.06896552
chromosome organization	225	0.009975615	2	0.0137931
cofactor metabolic process	133	0.005896697	1	0.00689655
cytoskeleton organization	264	0.011704722	1	0.00689655
DNA metabolic process	495	0.021946353	5	0.03448276
homeostatic process	165	0.007315451	2	0.0137931
immune system process	166	0.007359787	1	0.00689655
lipid metabolic process	265	0.011749058	2	0.0137931
metabolic process	2884	0.127865218	2	0.0137931
mitochondrion organization	62	0.002748836	1	0.00689655
mitotic nuclear division	107	0.004743959	1	0.00689655
nucleobase-containing compound catabolic process	484	0.021458657	6	0.04137931
plasma membrane organization	13	0.000576369	1	0.00689655
protein complex assembly	235	0.010418976	1	0.00689655
protein folding	119	0.005275992	6	0.04137931
protein maturation	52	0.002305476	1	0.00689655
protein targeting	93	0.004123254	2	0.0137931
reproduction	295	0.01307914	1	0.00689655
response to stress	545	0.024163157	5	0.03448276
signal transduction	903	0.040035469	11	0.07586207
small molecule metabolic process	866	0.038395034	7	0.04827586
translation	428	0.018975837	4	0.02758621
transmembrane transport	601	0.026645977	9	0.06206897
transport	662	0.029350477	11	0.07586207
tRNA metabolic process	120	0.005320328	1	0.00689655
vesicle-mediated transport	322	0.014276214	5	0.03448276

B-0210/C-0160-3p



GO term	Combined		B-0210-3p	
	Raw	Normalised	Raw	Normalised
aging	49	0.002172467	2	0.002197802
anatomical structure development	657	0.029128796	39	0.042857143
anatomical structure formation involved in morphogenesis	136	0.006029705	11	0.012087912
autophagy	40	0.001773443	3	0.003296703
biosynthetic process	1449	0.064242962	74	0.081318681
carbohydrate metabolic process	274	0.012148082	15	0.016483516
catabolic process	805	0.035690534	16	0.017582418
cell adhesion	113	0.005009976	7	0.007692308
cell cycle	315	0.013965861	18	0.01978022
cell death	166	0.007359787	9	0.00989011
cell differentiation	601	0.026645977	33	0.036263736
cell division	216	0.009576591	9	0.00989011
cell junction organization	41	0.001817779	5	0.005494505
cell morphogenesis	213	0.009443582	10	0.010989011
cell motility	87	0.003857238	6	0.006593407
cell proliferation	118	0.005231656	12	0.013186813
cell-cell signaling	200	0.008867213	9	0.00989011
cellular amino acid metabolic process	200	0.008867213	12	0.013186813
cellular component assembly	166	0.007359787	10	0.010989011
cellular nitrogen compound metabolic process	1337	0.059277322	64	0.07032967
cellular protein modification process	1	4.43361E-05	52	0.057142857
chromosome organization	225	0.009975615	13	0.014285714
chromosome segregation	55	0.002438484	2	0.002197802
cofactor metabolic process	133	0.005896697	10	0.010989011
cytoskeleton organization	264	0.011704722	11	0.012087912
cytoskeleton-dependent intracellular transport	23	0.00101973	4	0.004395604
developmental maturation	62	0.002748836	8	0.008791209
DNA metabolic process	495	0.021946353	27	0.02967033
embryo development	178	0.00789182	13	0.014285714
generation of precursor metabolites and energy	130	0.005763689	5	0.005494505
GPI anchor biosynthetic process	1	4.43361E-05	1	0.001098901
GPI anchor metabolic process	2	8.86721E-05	1	0.001098901
gravitaxis	1	4.43361E-05	1	0.001098901
growth	188	0.008335181	7	0.007692308
GTP catabolic process	7	0.000310352	2	0.002197802
homeostatic process	165	0.007315451	10	0.010989011
immune system process	166	0.007359787	9	0.00989011
intein-mediated protein splicing	3	0.000133008	1	0.001098901
intracellular transport	25	0.001108402	1	0.001098901
lipid metabolic process	265	0.011749058	9	0.00989011
locomotion	97	0.004300599	4	0.004395604
meiotic chromosome segregation	4	0.000177344	1	0.001098901
membrane organization	90	0.003990246	4	0.004395604
metabolic process	2884	0.127865218	2	0.002197802

mitochondrion organization	62	0.002748836	2	0.002197802
mitotic nuclear division	107	0.004743959	4	0.004395604
mRNA processing	163	0.007226779	6	0.006593407
neurogenesis	5	0.00022168	1	0.001098901
neurological system process	119	0.005275992	6	0.006593407
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3	0.000133008	1	0.001098901
nucleic acid phosphodiester bond hydrolysis	58	0.002571492	2	0.002197802
nucleobase-containing compound catabolic process	484	0.021458657	26	0.028571429
nucleocytoplasmic transport	51	0.002261139	4	0.004395604
oxidation-reduction process	30	0.001330082	2	0.002197802
peptidyl-tyrosine dephosphorylation	2	8.86721E-05	1	0.001098901
phosphorylation	47	0.002083795	1	0.001098901
pigmentation	24	0.001064066	3	0.003296703
positive regulation of transcription from RNA polymerase II promoter	2	8.86721E-05	1	0.001098901
protein complex assembly	235	0.010418976	12	0.013186813
protein dephosphorylation	2	8.86721E-05	1	0.001098901
protein folding	119	0.005275992	2	0.002197802
protein glycosylation	3	0.000133008	1	0.001098901
protein maturation	52	0.002305476	4	0.004395604
protein phosphorylation	11	0.000487697	1	0.001098901
protein targeting	93	0.004123254	4	0.004395604
protein ubiquitination	3	0.000133008	1	0.001098901
regulation of transcription, DNA-templated	38	0.001684771	3	0.003296703
reproduction	295	0.01307914	20	0.021978022
response to stress	545	0.024163157	24	0.026373626
ribonucleoprotein complex assembly	51	0.002261139	1	0.001098901
ribosome biogenesis	65	0.002881844	1	0.001098901
RNA phosphodiester bond hydrolysis, endonucleolytic	11	0.000487697	1	0.001098901
RNA-dependent DNA replication	111	0.004921303	4	0.004395604
secondary metabolic process	20	0.000886721	2	0.002197802
sialylation	2	8.86721E-05	1	0.001098901
signal transduction	903	0.040035469	46	0.050549451
small molecule metabolic process	866	0.038395034	47	0.051648352
sulfur compound metabolic process	59	0.002615828	4	0.004395604
symbiosis, encompassing mutualism through parasitism	27	0.001197074	2	0.002197802
transcription, DNA-templated	45	0.001995123	1	0.001098901
translation	428	0.018975837	10	0.010989011
transmembrane transport	601	0.026645977	19	0.020879121
transport	662	0.029350477	47	0.051648352
tRNA metabolic process	120	0.005320328	6	0.006593407
vesicle-mediated transport	322	0.014276214	23	0.025274725

B-0210/C-0160-5p

GO term	#contigs
anatomical structure development	1
biosynthetic process	3
carbohydrate metabolic process	1
cellular nitrogen compound metabolic process	2
cellular protein modification process	3
DNA metabolic process	2
immune system process	1
nucleobase-containing compound catabolic process	2
nucleocytoplasmic transport	1
response to stress	3
RNA-dependent DNA replication	1
signal transduction	1
small GTPase mediated signal transduction	1
small molecule metabolic process	3
translation	1
transmembrane transport	2
transport	2
tRNA metabolic process	2

Par-340/341-5p

GO term	#contigs
biosynthetic process	3
carbohydrate metabolic process	1
cellular nitrogen compound metabolic process	2
cellular protein modification process	1
nucleobase-containing compound catabolic process	1
nucleocytoplasmic transport	1
response to stress	1
RNA-dependent DNA replication	1
signal transduction	2
small GTPase mediated signal transduction	1
small molecule metabolic process	2
tRNA metabolic process	2

Par-340-3p

GO term	#contigs
biosynthetic process	1
cellular nitrogen compound metabolic process	1
nucleocytoplasmic transport	1
small GTPase mediated signal transduction	1
vesicle-mediated transport	1
cell adhesion	1
protein targeting	1

Par-341-3p

GO term	#contigs
cell division	1

Par-247-3p

(No predicted GO terms)

Par-247-5p

GO term	#contigs
anatomical structure development	1
biosynthetic process	2
cell adhesion	1
cell differentiation	1
cellular nitrogen compound metabolic process	1
cellular protein modification process	3
DNA metabolic process	1
neurological system process	1
protein targeting	1
transmembrane transport	1
transport	2
vesicle-mediated transport	1

Chapter 6:

General Discussion

- **Synopsis of research**
- **A new model for metazoan miRNA evolution**
- **Intronic exaptation and miRNA evolution**
- **miRNAs, biological robustness, and the canalization of development**

1. Synopsis of research

This thesis represents an original contribution to our understanding of the role of miRNA innovation in the evolution of development within the Metazoa. At the level of individual genes, I present case studies of two intronic miRNAs. The mammalian Hox cluster miRNA, miR-615, provides a model for the evolution of intronic miRNAs and their regulation. miR-2768, a lepidopteran innovation, represents the first functionally validated study of an intronic miRNA performing a complementary function to its host gene. I also identify a core set of miRNAs which are highly upregulated or uniquely expressed in the ovaries of the Speckled Wood butterfly, and discuss the roles of these miRNAs in regulation of oogenesis and ovarian function. On a phylogenetic level, comparative analyses of miRNA complements in butterflies and moths reveal a highly dynamic miRNA landscape with continual gains and losses of newly emergent miRNA genes. The observation that conserved miRNAs were acquired in a burst at the base of the lepidopteran tree, rather than being homogenously gained over all lineages, is consistent with published work in other clades. These data supports a gene capture model whereby, amidst the constant turnover of young miRNAs at the tips of phylogenies, variable selection pressures influence the rate at which new miRNA genes become fixed in the genome. This model represents a new way of thinking about miRNA evolution and challenges the current dogma that metazoan miRNAs are continuously gained and rarely lost in the course of evolution.

2. A new model for metazoan miRNA evolution

2.1 The miRNA capture model

The work presented in Chapter 3 represents the characterization of whole-body miRNA transcriptomes pooled across all life cycle stages in the Speckled Wood butterfly and the Horse chestnut Leafminer moth. This complements existing literature on miRNA discovery in other lepidopteran species. Combining all available data on lepidopteran miRNAs with BLASTn analyses of genome sequences permitted the scoring of each

miRNA for presence or absence across the Lepidoptera and the reconstruction of the most likely scenario for lepidopteran miRNA evolution.

The most significant finding from this project was that the miRNAs present in modern butterflies and moths did not arise evenly throughout the evolution of the clade. While there is clearly an abundance of young, lineage-specific miRNAs on terminal branches of the tree, the majority of miRNA innovation subsequently retained by selection appears concentrated along the common lepidopteran stem following divergence from Trichoptera. These seemingly conflicting observations led to the proposal of the 'miRNA capture model'. In line with existing models of miRNA turnover at the tips of phylogenies (Lyu et al., 2014; Meunier et al., 2013), this model describes how a constant high rate of *de novo* miRNA precursor generation across a clade is roughly balanced with a roughly equal rate of gene loss that eliminates most precursors shortly after they arise. The critical difference between my model and the current understanding of miRNA evolution is an emphasis that the rate of miRNA fixation ('capture') in the genome is heterogeneous across lineages, with the occasional relaxation of selection pressures along specific lineages leading to the fixation of more new precursors and the observed bursts of miRNA evolution which occur throughout the evolution of the Metazoa.

This work is now being taken forward with Prof. Michael Bonsall with the aim of building a mathematical model that simulates the gain-loss/capture cycle and makes testable predictions for miRNA evolution.

2.2 Applying the capture model beyond miRNA evolution

My proposed 'capture model' is not inherently restricted to describing miRNA evolution. I argue that this model may be applied to describe gene evolution in other situations involving *de novo* gene evolution. The *de novo* acquisition of new genes is a major driver of miRNA diversity but is also known to contribute to the evolution of other classes of genes, including protein coding genes.

The classical concept of descent with modification was traditionally thought to suffice for explaining protein diversification. 'Molecular tinkering' of this nature, where protein-coding genes are descended from existing protein-coding genes through duplication

and/or modification of function (Jacob, 1977), was proposed to explain the majority of useful protein diversity. In contrast, *de novo* innovation of new genes from previously non-coding DNA was long assumed to be unimportant to protein evolution. However, recent advances in DNA sequencing technologies have led to increasingly detailed characterization of animal genomes and the identification of 'orphan genes' which are seemingly unrelated to any other existing genes (Begun, Lindfors, Thompson, & Holloway, 2006; Heinen, Staubach, Häming, & Tautz, 2009). Clearly documented cases of *de novo* protein evolution have also been published (Cai, Zhao, Jiang, & Wang, 2008; Heinen et al., 2009; Knowles & McLysaght, 2009), each identifying a functional ORF in a focal species which is homologous to a non-coding RNA in multiple outgroup species. Such observations have raised doubts on the viewpoint that all protein innovation was concentrated at one point early in evolution and that all subsequent protein evolution was achieved through tinkering with existing material (Tautz & Domazet-Lošo, 2011).

How can the same evolutionary model be used to characterise the evolution of two different kinds of genes? The *de novo* gain of a new protein gives rise to similar selective pressures as the emergence of a new miRNA. Newly emergent proteins, like newly emergent miRNAs, have the potential to disrupt existing genetic regulatory networks. A new miRNA or protein is much less likely to confer a selective benefit than it is to be selectively neutral or disadvantageous. Most newly emergent miRNA and protein coding genes will therefore be quickly lost from the genome due to the lack of selective pressure to retain them, resulting in high rates of turnover. The net rate of novel gene retention in both cases is the result of balancing the rate of gene gain through *de novo* birth and the rate of gene loss. In the event of a relaxation of selection pressures on the genome, there is a higher chance that a newly emergent gene confers a selective benefit to the system, allowing for the fixation of more young genes in the genome and giving rise to increased rates of gene gain.

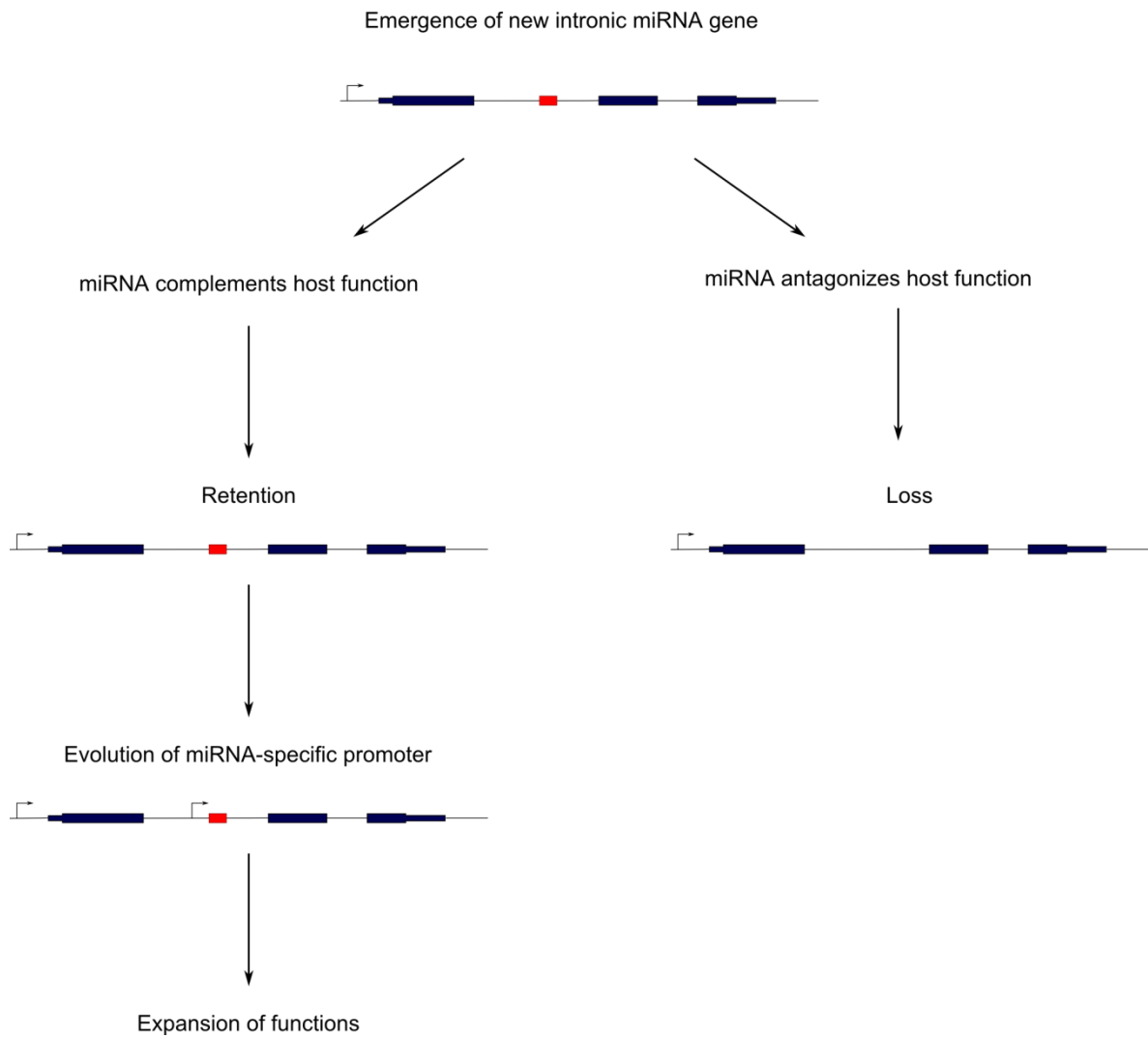
Published data on *de novo* protein acquisition are in line with the predictions of the capture model. A 2011 report on the *de novo* evolution of protein coding genes (Tautz & Domazet-Lošo, 2011) presents striking parallels to my data on miRNA gene evolution within the Lepidoptera (Chapter 3). In this study, the authors analyse the likely time of origin of orphan genes which subsequently give rise to new gene families ('founder genes') and study the emergence rates of these genes across three different lineages leading to *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana*. All lineages show a high rate of recent gene acquisition on terminal branches on lepidopteran and mammalian phylogenies. This is strongly suggestive of a high rate of

founder gene turnover (gain and loss) at the tips of phylogenies and mirrors the pattern of miRNA evolution observed in the Lepidoptera (Chapter 3) and in mammals (Meunier et al., 2013). In mice and *Arabidopsis* the observed distribution of older founder genes is very much in line with the pattern predicted by the capture model, with low rates of gain interspersed with bursts of gene innovation. The fly lineage shows a slightly different pattern with a more gradual increase in the rate of founder gene gain over time following the emergence of the Insecta. This may be due to a more prolonged window for gene capture; it will be of interest to build on the findings of Chapter 3 and study whether the evolution of older hexapod miRNAs follows the same trend.

3. Intronic exaptation and miRNA evolution

In this thesis, I present case studies of two different intronic miRNA genes – the eutherian-specific *mir-615* (Chapter 2) and the lepidopteran-specific *mir-2768* (Chapter 4). Both are hosted within the introns of homeobox genes with known roles in developmental regulation.

Intronic exaptation, the process by which a novel miRNA emerges through the modification of a pre-existing intronic sequence, is thought to be a major source of miRNA novelty. Under this model, the presence of an existing transcriptional unit complete with promoter sequences facilitates the emergence of a new miRNA by obviating the need to evolve separate transcriptional control sequences. This idea that introns could represent a rich source of unstructured primary transcripts capable of gradual modification into a suitable stem-loop miRNA precursor has been discussed by several different authors (Berezikov, 2011; Campo-Paysaa et al., 2011; Meunier et al., 2013) and is general well-accepted. On the contrary, functions of miRNAs arising via this route in relationship to those of their host genes have not been well characterized. I hypothesise that a new miRNA emerging through this route would need to have functions which are not antagonistic to those of its host gene, or experience negative selection (Figure 6.1). miRNAs with functions that complemented host gene function may have undergone positive selection. It may also be possible that a new intronic miRNA does not interact with any pathways involving the host protein, in which case its retention or loss would be governed by the balance between drift and any other selective effects it may have on the genome.

**Figure 6.1**

Model for the fate of emergent intronic miRNAs. Initial transcription of the miRNA is controlled solely by the host promoter and the miRNA is necessarily co-expressed with the host gene. miRNAs with functions that interfere with host gene function are selected against. Intronic miRNAs retained for long periods over evolutionary time are therefore likely to have functions complementing those of their host genes. In time, intronic miRNAs may acquire their own promoter elements and thereby expand their expression and function independently of their host genes.

In Chapter 4, I identify and functionally validate a candidate target site for miR-2768-3p within the coding sequence of the *ci* gene. This finding is significant as the *mir-2768* gene is hosted within the intron of *inv*, which is itself involved in repression of *ci*. This presents a rather striking, and perhaps unusual, demonstration of coordinated functions between an intronic miRNA and its host gene. It is unlikely that the majority of intronic miRNAs interact with precisely the same targets as their host genes; amongst the target prediction dataset for miR-615-3p, only PBX3 is known to interact with *Hoxc5* in mice (Ravasi et al., 2010). Future large-scale analysis of intronic miRNAs and their targets would be beneficial to testing the hypothesis that such miRNAs tend to have functions complementing host gene functions, as well as evaluate the scope to which miRNAs and their host genes are involved in similar genetic regulatory pathways.

4. miRNAs, biological robustness, and the canalization of development

Although miRNAs act as post-transcriptional repressors of gene expression, it is critical to note that their primary role is likely to lie in conferring robustness to genetic regulatory networks, rather than functioning as on/off switches for gene expression (Ebert & Sharp, 2012). In Chapter 5, I study the ovarian miRNA transcriptome of *P. aegeria* and annotate an additional 11 miRNAs across two populations which were not found in the pooled miRNA transcriptome presented in Chapter 3. I also identify a group of seven “ovary set” miRNAs which are either ovary-specific or highly overrepresented in ovaries relative to the pooled transcriptome in both populations. This set includes miRNAs such as miR-989 with documented roles in regulation of ovarian function in other insects. Target prediction data against the corresponding ovarian transcriptome reveal large number of targets with roles critical to ovary function for some miRNAs, and very few targets for others. I hypothesise that this difference is related to the type of regulatory interaction employed by each miRNA in the ovary; miRNAs involved in coherent regulation suppress low level transcription of targets which are therefore unlikely to appear in the transcriptome. Those miRNAs with a large number of targets involved in ovary function are more likely to be acting in incoherent loops to buffer expression levels of more highly abundant transcripts.

In Chapter 4, the observation that miR-2768 has a complementary function to that of its host gene product (*Invected*), is also suggestive of a coherent regulatory loop and improved repression of *ci* due to the action of the miRNA in the context of body segmentation and wing compartment patterning. The observation that *ci* is coexpressed with *inv* in the eyespots of some species (Keys, 1999) hints at the possibility that miR-2768 may also be involved in incoherent regulation through modulation rather than complete repression of *ci* transcript expression. Whether or not the same miRNA is capable of exerting incoherent regulation in some contexts and coherent regulation in others is unknown at the moment and worth investigating. It will be of interest to future projects to compare the expression domains of *en/inv*, *mir-2768* and *ci*. This would both promote an understanding of the mode of regulation adopted by miR-2768, as well as identify whether it is capable of independent expression from its host transcript and thereby present a parallel case study to mammalian *mir-615* in the investigation of intronic miRNA regulation and function.

5. Evolutionary conservation of miRNAs and its effect on expression levels and tissue specificity

Contrary to the model presented by Chen & Rajewsky where newly-emergent miRNAs are predicted to have lower expression levels and more stringent tissue specificity than older miRNAs (Berezikov, 2011; K. Chen & Rajewsky, 2007), I found no significant difference in the total expression levels for mature products of conserved and young miRNAs in the pooled whole-body miRNA transcriptomes of the Speckled Wood butterfly (*Pararge aegeria*) and the Horse chestnut Leafminer moth (*Cameraria ohridella*) presented in Chapter 3. Similar observations were made in ovary-specific miRNA transcriptome in the Speckled Wood (Chapter 5). My data indicate that miRNAs exhibit a very wide range of expression levels within each category, with a large degree of overlap in expression between old and young miRNAs. However, a consistent observation across all of my transcriptomes is that conserved miRNAs make up the most highly expressed sequences in each library.

In the ovary, the majority of miRNAs expressed are the products of conserved genes (Chapter 5). The proportion of conserved genes expressed is higher than that in the whole-body pooled transcriptome. This is at odds with the prediction that young

miRNAs have tissue-specific distributions and would be expected to be overrepresented in a tissue-specific miRNA library compared to one from the whole body. It may be possible that individual tissues differ in their capacity to tolerate the expression of new miRNAs. The ovary represents a particularly complicated tissue for studying miRNA expression as its gene expression profile incorporates a transgenerational component; expressed RNAs are required both for ovarian function as well as maternal regulation of early embryogenesis. Additional small RNA sampling across a wider range of tissues is an avenue for further research on tissue specificity in relation to miRNA age.

Concluding Remarks

The results presented in this thesis represent a significant contribution to our understanding of miRNA evolution and its contribution to developmental regulation. A broad perspective on miRNA evolution across the Metazoa is discussed and put forth as a new model with variable miRNA capture rates in response to selective pressures. This integrates and builds upon the current understanding of animal miRNA evolution through continuous miRNA addition to metazoan genomes and high turnover of young miRNAs. Specific case studies performed on the intronic miRNA genes *mir-615* and *mir-2768* shed light on the evolution and transcriptional regulation of such miRNAs as well as their functions relative to their host genes. Finally, I look into the role of miRNAs in the lepidopteran ovary transcriptome, study their expression levels relative to the whole body, and discuss the involvement of this class of small RNAs in conferring robustness to the processes responsible for reproduction and the continuity of future generations.

Appendix A: Scripts

This appendix contains computer scripts used in data processing as described in the Materials and Methods section of each data chapter.

Appendix A1 – fastqtrim.py

```
#!/software/bin/python
# This is a fairly simple python script using Biopython 1.51+
# to read in FASTQ records from stdin, trim to 21 letters,
# and write them to stdout. It does not check the quality
# strings at all, and should therefore work on Sanger,
# Solexa or Illumina 1.3+ FASTQ files equally well.
import sys,string

def reversecomplement(sequence):
    complement = {"A":"T", "T":"A", "C":"G", "G":"C", "N":"N"}

    sequence = sequence[::-1]

    reverse_complement_sequence = ""
    for letter in sequence:
        reverse_complement_sequence += complement[letter]

    return reverse_complement_sequence

def getNameList( fastq, names, sep ) :
    print "Getting names from ",fastq
    f_in = open(fastq,"r")
    i = 4
    j = 0
    for line in f_in :
        if i == 4 and line[0] == "@" :
            id = line.strip()
            id = id.split(sep)[0]
            names.append(id)
            i = 0
            j = j+1
        i = i+1
    f_in.close()
    print "Done, got:",j
    return names

def getRead( f_in, revcmp ) :
    l1 = f_in.readline().strip()
    l2 = f_in.readline().strip()
    l3 = f_in.readline().strip()
    l4 = f_in.readline().strip()

    if revcmp == 1 :
        #print l2
        #print l4
        l2 = reversecomplement(l2)
        l4 = l4[::-1]
        #print "RC"
        #print l2
        #print l4

    if len(l1) < 10 :
        print "No more lines, exiting"
        sys.exit(0)

    return [l1,l2,l3,l4,f_in]

def main( argv ) :
```

```

    if len(argv) < 2 :
        print "Shuffle paired reads, checking for missing reads and re-aligning,
also shuffling "
        print "USAGE: <fastqPart1> <fastqPart2> <shuffle output file> <trim
start pos> <trim end pos> <OPTION: -revcomp>\n"
        sys.exit(0)

    fastq1 = argv[1]
    fastq2 = argv[2]
    out = argv[3]
    p1 = int(argv[4])
    p2 = int(argv[5])

    print "Taking a slice from",p1,"to",p2

    if len(argv) > 6 :
        if argv[6] == "-revcomp" :
            print "Reverse-complementing all reads!!"
            revcmp = 1
        else :
            revcmp = 0
    seperator = "/"
    localLen = 500

    checkMisalign = 0

    if checkMisalign == 1 :
        names1 = []
        names1 = getNameList( fastq1, names1, seperator )
        names2 = []
        names2 = getNameList( fastq2, names2, seperator )
        max1 = len(names1)
        max2 = len(names2)

    f_in1 = open(fastq1, "r")
    f_in2 = open(fastq2, "r")
    f_out = open(out, "w")
    fout1 = fastq1+".trim"
    fout2 = fastq2+".trim"
    f_o1 = open(fout1, "w")
    f_o2 = open(fout2, "w")

    i = 0
    j = 0
    k = 0
    while 1 :

        [a1,a2,a3,a4,f_in1] = getRead(f_in1,revcmp)
        [b1,b2,b3,b4,f_in2] = getRead(f_in2,revcmp)

        if checkMisalign == 1 :

            end_A = i+localLen
            if end_A > max1 :
                end_A = max1

            end_B = j+localLen
            if end_B > max2 :
                end_B = max2

            localA = names1[i:end_A]
            localB = names2[j:end_B]

            nameA = a1.split(seperator)[0]
            if nameA != localA[0] :
                print "ERROR: indices misaligned for A:",nameA,localA[0:3]
                sys.exit(1)
            nameB = b1.split(seperator)[0]

```

```

    if nameB != localB[0] :
        print "ERROR: indices misaligned for B:",nameB,localB[0:3]
        sys.exit(1)

    if nameA != nameB :
        posA = 0
        posB = 0
        for id in localA :
            if id in localB :
                posB = localB.index(id)
                break
            posA = posA + 1

        for ignore in range(posA) :
            [a1,a2,a3,a4,f_in1] = getRead(f_in1,revcmp)
            nameA = a1.split(seperator)[0]

        for ignore in range(posB) :
            [b1,b2,b3,b4,f_in2] = getRead(f_in2,revcmp)
            nameB = b1.split(seperator)[0]

        i = i+posA
        j = j+posB

        if nameA != nameB :
            print "ERROR Attempting to re-align"
            print "ID seperator>"+seperator+"<"
            print nameA,nameB
            sys.exit(1)

    else :
        nameA = a1.split(seperator)[0]
        nameB = b1.split(seperator)[0]

    if k > 50000 :
        print "So far done:",i,j,nameA,nameB
        k = 0

    str1 = a1 + "\n" + a2 + "\n" + a3 + "\n" + a4 + "\n"
    f_out.write(str1)
    str2 = b1 + "\n" + b2 + "\n" + b3 + "\n" + b4 + "\n"
    f_out.write(str2)

    str1 = a1 + "\n" + a2[p1:p2] + "\n" + a3 + "\n" + a4[p1:p2] + "\n"
    f_o1.write(str1)
    str2 = b1 + "\n" + b2[p1:p2] + "\n" + b3 + "\n" + b4[p1:p2] + "\n"
    f_o2.write(str2)

    i = i+1
    j = j+1
    k = k+1

    f_in1.close()
    f_in2.close()
    f_out.close()

#   print record_dict["gi:12345678"]
#
#   for title, seq, qual in FastqGeneralIterator(sys.stdin) :
#       print "@%s\n%s\n+\n%s" % (title, seq[:21], qual[:21])
#   #The print statement will include a trailing newline

main(sys.argv)

```

Appendix A2 – size-select.py

Script provided by Ferdinand Marlétaz

```
#!/usr/bin/env python
import sys

def readfq(fp): # this is a generator function
    last = None # this is a buffer keeping the last unprocessed line
    while True: # mimic closure; is it a bad idea?
        if not last: # the first record or a record following a fastq
            for l in fp: # search for the start of the next record
                if l[0] in '>@': # fasta/q header line
                    last = l[:-1] # save this line
                    break
        if not last: break
        name, seqs, last = last[1:].split()[0], [], None
        for l in fp: # read the sequence
            if l[0] in '@+>':
                last = l[:-1]
                break
            seqs.append(l[:-1])
        if not last or last[0] != '+': # this is a fasta record
            yield name, ''.join(seqs), None # yield a fasta record
            if not last: break
        else: # this is a fastq record
            seq, leng, seqs = ''.join(seqs), 0, []
            for l in fp: # read the quality
                seqs.append(l[:-1])
                leng += len(l) - 1
                if leng >= len(seq): # have read enough quality
                    last = None
                    yield name, seq, ''.join(seqs); # yield a fastq record
                    break
            if last: # reach EOF before reading enough quality
                yield name, seq, None # yield a fasta record instead
                break

minsize=int(sys.argv[2])
maxsize=int(sys.argv[3])

out=open(sys.argv[1].rsplit('.',1)[0]+'ss.fa','w')
ct=0
for header,seq,qual in readfq(open(sys.argv[1])):
    if len(seq)>minsize and len(seq)<maxsize:
        out.write('>'+header+'\n'+seq+'\n')
        ct+=1
    else:
        continue

print ct,'sequences selected for size...'
```


Appendix A3 – contig-stats.pl

```
#!/usr/bin/perl -w
#source http://seqanswers.com/forums/showthread.php?t=2766 author = BENM
#count and maximum length added by Aaron Liston June 8, 2011
use strict;
my ($len,$total,$contigs)=(0,0,1);
my @x;
while(<>){
    if(/^[\>\@]/){
        if($len>0){
            $total+=$len;
            $contigs++;
            push @x,$len;
        }
        $len=0;
    }
    else{
        s/\s//g;
        $len+=length($_);
    }
}
if ($len>0){
    $total+=$len;
    push @x,$len;
}
@x=sort{$b<=>$a} @x;
my $max_value = $x[0];
my ($count,$half)=(0,0);
for (my $j=0;$j<@x;$j++){
    $count+=$x[$j];
    if (($count>=$total/2)&&($half==0)){
        print "sequence #: $contigs\t";
        print "total length: $total\t";
        print "max length: $max_value\t";
        print "N50: $x[$j]\t";
        $half=$x[$j]
    }elseif ($count>=$total*0.9){
        print "N90: $x[$j]\n";
        exit;
    }
}
}
```


Appendix B: List of abbreviations used

This appendix contains non-S.I. abbreviations used in this thesis.

BLAST	Basic Local Alignment Search Tool
BLASTn	Nucleotide BLAST
bp	Base pair
cDNA	Complementary DNA
CDS	Coding sequence
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
dsRNA	Double-stranded RNA
ENCODE	Encyclopedia of DNA Elements
EST	Expressed Sequence Tag
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GEO	Gene Expression Omnibus
LD	Long day
miRNA	microRNA
NCBI	National Centre for Biotechnology Information
nt	Nucleotide
PCR	Polymerase chain reaction
PITA	Probability of Interaction by Target Accessibility
ssRNA	Single-stranded RNA
tBLASTn	Translated nucleotide BLAST (using a protein query)
TSS	Transcription start site
UTR	Untranslated region

Appendix C: Resulting publications

The following is a list of publications associated with the work presented in this thesis.

Accepted publications

Quah, S., Breuker, C. J. B., Holland, P.W.H. (2015). A diversity of conserved and novel ovarian microRNAs in the Speckled Wood (*Pararge aegeria*). *PLOS ONE*, 10(11):e0142243.

Quah, S., Holland, P.W.H (2015). The Hox gene miRNA miR-615: A case study of intronic miRNA evolution. *EvoDevo*, 6(31).

Quah, S., Hui, J. H. L., & Holland, P. W. H. (2015). A Burst of miRNA Innovation in the Early Evolution of Butterflies and Moths. *Molecular Biology and Evolution*, 32(5), 1161-1174.

Kenny, N. J., Quah, S., Holland, P. W. H., Tobe, S. S., & Hui, J. H. L. (2013). How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment? *General and Comparative Endocrinology*, 188, 16–22.

Planned

Quah, S. Bonsall, M., Holland, P. W. H. Mathematical modelling of the miRNA capture model.

A Burst of miRNA Innovation in the Early Evolution of Butterflies and Moths

Shan Quah,¹ Jerome H.L. Hui,² and Peter W.H. Holland^{*,1}

¹Department of Zoology, University of Oxford

²School of Life Sciences and The State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong

*Corresponding author: E-mail: peter.holland@zoo.ox.ac.uk.

Associate editor: Gregory Wray

Abstract

MicroRNAs (miRNAs) are involved in posttranscriptional regulation of gene expression. Because several miRNAs are known to affect the stability or translation of developmental regulatory genes, the origin of novel miRNAs may have contributed to the evolution of developmental processes and morphology. Lepidoptera (butterflies and moths) is a species-rich clade with a well-established phylogeny and abundant genomic resources, thereby representing an ideal system in which to study miRNA evolution. We sequenced small RNA libraries from developmental stages of two divergent lepidopterans, *Cameraria ohridella* (Horse chestnut Leafminer) and *Pararge aegeria* (Speckled Wood butterfly), discovering 90 and 81 conserved miRNAs, respectively, and many species-specific miRNA sequences. Mapping miRNAs onto the lepidopteran phylogeny reveals rapid miRNA turnover and an episode of miRNA fixation early in lepidopteran evolution, implying that miRNA acquisition accompanied the early radiation of the Lepidoptera. One lepidopteran-specific miRNA gene, miR-2768, is located within an intron of the homeobox gene *invected*, involved in insect segmental and wing patterning. We identified *cubitus interruptus* (*ci*) as a likely direct target of miR-2768, and validated this suppression using a luciferase assay system. We propose a model by which miR-2768 modulates expression of *ci* in the segmentation pathway and in patterning of lepidopteran wing primordia.

Key words: microRNA, Lepidoptera, segmentation, wing patterning, Gli, homeobox.

Introduction

During development, the information in an organism's genome is used to coordinate the generation of morphology. Evolutionary modification of this developmental process in different animal lineages has given rise to the vast diversity of body form and function observed across the Metazoa today. Many evolutionary changes to genes have been documented including gene duplications (Brooke et al. 1998; Putnam et al. 2008), coding sequence alterations (Fondon and Garner 2004), and modifications to gene expression (Jeong et al. 2006). Although much focus has been on the evolution of *cis*-regulatory elements that affect gene expression (Putnam et al. 2008; Chan et al. 2010), there are also *trans*-acting factors that should not be overlooked. In particular, there is emerging evidence that evolutionary changes to microRNAs (miRNAs) could have a role to play in the evolution of development (Griffiths-Jones et al. 2011; Hui et al. 2013; Moran et al. 2014; Ninova et al. 2014).

miRNAs are short (~22 nt), noncoding RNAs initially transcribed as longer primary transcripts. Each primary transcript includes one or more imperfect hairpins of approximately 80 nt that are excised in the nucleus to give rise to precursor miRNAs. These are exported to the cytoplasm and undergo further processing prior to being incorporated in the RNA-induced silencing complex (RISC). The mature miRNA then functions to guide RISC to specific messenger RNA transcripts

to bring about downregulation of gene expression, by binding to specific target sequences primarily in the 3'-untranslated regions (3'-UTRs) of protein-coding genes (Kenny et al. 2013). Several miRNAs are known to target genes involved in developmental processes (Yekta et al. 2004; Moran et al. 2014), and a general role in stabilization of developmental pathways in the face of fluctuating environmental conditions has also been proposed (Peterson et al. 2009; Herranz and Cohen 2010).

The miRNA repertoire in a given evolutionary lineage is constantly shifting under the opposing influences of miRNA birth and miRNA death (the gain and loss of miRNA genes). Previous work in vertebrates (Heimberg et al. 2008), mammals (Meunier et al. 2013), drosophilids (Lu et al. 2008; Nozawa et al. 2010; Lyu et al. 2014), and flatworms (Fromm et al. 2013) have suggested that miRNA birth rates can differ substantially in different lineages, while loss is thought to be less frequent than gain. For example, there was a burst of miRNA innovation on the stem lineage leading to the vertebrates following divergence from other chordates, and this has been postulated to be a contributory factor in the evolution of vertebrate complexity (Heimberg et al. 2008). Endoparasitic platyhelminths have undergone significant losses of conserved miRNA families which have not been observed to the same extent in other bilaterian phyla studied. This is proposed to be correlated with the evolution of

© The Author 2015. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

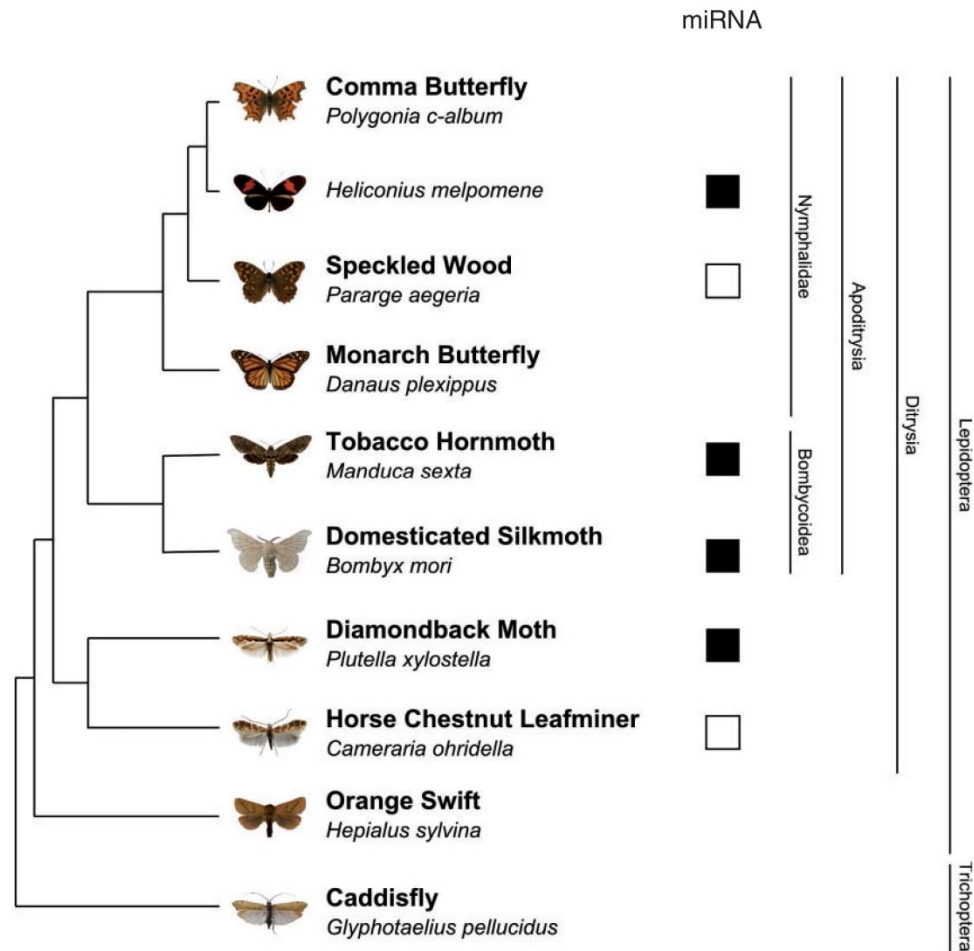


FIG. 1. Phylogeny of the lepidopteran species for which genome sequence data is available. The orange swift genome is of insufficient coverage and was not used for assessment of miRNA distribution. Black squares indicate published miRNA sequencing projects and white squares represent the data presented in this manuscript.

reduced morphological and metabolic complexity (Fromm et al. 2013).

Lepidoptera represents an ideal clade to study the relationship between miRNAs and morphological evolution (fig. 1). This group comprises over 150,000 species of moths and butterflies and is one of the most species-rich insect orders. Lepidopterans share a large number of morphological apomorphies and underwent a major evolutionary radiation during the Cretaceous (Regier et al. 2013; Wahlberg et al. 2013; Kawahara and Breinholt 2014) following their divergence from their sister group Trichoptera (caddisflies). The phylogeny of the Lepidoptera is well-studied, with clearly supported superfamilies identified through both morphological and molecular studies (Regier et al. 2013; Wahlberg et al. 2013). In outline, the vast majority of species diversity is found in the Ditrysia, which makes up 99% of extant lepidopteran species (Wahlberg et al. 2013). Within this clade, the Yponomeutoidea (e.g., *Plutella xylostella*) and Gracillarioidea (e.g., *Cameraria ohridella*) are distant sister groups that

diverged early from the rest of the Ditrysia (Bazin et al. 2013; Regier et al. 2013). Most of the remaining ditrysians form a strongly supported clade—the Apoditrysia (Bazin et al. 2013), which includes the Bombycoidea (including silkmoths and hawkmoths) and the Papilionoidea (butterflies, including the highly diverse nymphalids). One major advantage of using the Lepidoptera for the comparative analysis of miRNA complements is the plethora of genome resources available in the public domain, including sequences from the Domestic Silkmoth *Bombyx mori* (Mita 2004; Xia et al. 2004), the butterflies *Heliconius melpomene* (Heliconius Genome Consortium 2012) and *Danaus plexippus* (Zhan et al. 2011), the Diamondback moth *P. xylostella* (You et al. 2013), and the Tobacco Hornworm hawkmoth *Manduca sexta*. Our group has also recently sequenced and assembled draft genomes for the Horse chestnut Leafminer moth *C. ohridella*, the Speckled Wood butterfly *Pararge aegeria*, the Comma butterfly *Polygonia c-album*, and the Orange Swift moth *Hepialus sylvina* (Ferguson et al. 2014). The same study

also reports a draft genome assembly for the caddisfly *Glyptotendipes pallidus* as a representative of the Trichoptera (fig. 1).

However, while genome sequences can be used to search for known miRNA homologues, sampling of actual small RNAs is required for the identification of novel sequences and hence the determination of the miRNA repertoire in a given species. Prior to a recent report (Etebari et al. 2013) on small RNA sequencing in *P. xylostella* (Yponomeutoidea), miRNA data were limited to *B. mori* and *M. sexta* (Bombycoidea) (Liu et al. 2009, 2010) with limited information from wing discs of *H. melpomene* (Surridge et al. 2011). To enlarge the comparative analysis, ensuring the inclusion of distant lineages, the gracillid *C. ohridella* and the nymphalid butterfly *Pa. aegeria* were selected for small RNA sequencing. Our analyses reveal that a burst of miRNA fixation occurred early in the evolution of the clade. Few miRNA fixations occur on more recent internal branches, even in speciose groups such as the Nymphalidae. We also find high numbers of evolutionarily young, species-specific miRNA sequences, suggesting that high miRNA birth rates are matched by high death rates with few novel genes persisting to integrate stably into developmental programs. We present evidence that one such miRNA has been recruited for a role in segmentation or wing patterning.

Results

Small RNA Sampling in *C. ohridella* and *Pa. aegeria*

Selection of informative species for miRNA sampling to help elucidate the pattern of miRNA evolution across the Lepidoptera was carried out using three essential criteria: phylogenetic placement, accessibility of different life-cycle stages to afford deep sampling of miRNA species, and availability of genomic sequence of sufficient coverage to enable novel miRNA precursor prediction and detection. The phylogenetic criteria focused our attention on basally branching Lepidoptera lineages and on representation of the speciose clade Nymphalidae, particularly because the published nymphalid miRNA data derive from a restricted tissue type (*H. melpomene* wing disc, Surridge et al. 2011).

Species from basally branching lineages with genomic data include the Orange Swift moth *Hep. sylvina* and the Horse-chestnut leafminer *C. ohridella* (fig. 1). To assess suitability for verification of miRNA sequences, we first used a BLASTn search with 20 conserved miRNAs originating in the common hexapod ancestor or earlier (supplementary file S1, Supplementary Material online). This identified only 12 out of 20 query miRNA sequences in the former genome, but 19 out of 20 in the latter. Furthermore, *C. ohridella* represents a basally divergent lineage within the Ditrysia and a distant sister group to *P. xylostella* (Bazin et al. 2013; Regier et al. 2013) and provides an excellent basis for comparative analysis of miRNAs originating at the base of the Ditrysian tree. This invasive species is now extraordinarily common in the UK and most life-cycle stages can be readily collected.

The second species chosen for deep sequencing of small RNAs was the Speckled Wood butterfly (*Pa. aegeria*), a

Table 1. Summary of the Small RNA Sequencing Data Sets from *Cameraria ohridella* and *Pararge aegeria*.

	<i>C. ohridella</i>	<i>Pa. aegeria</i>
Stages sampled	Larvae, pupae, adults	Embryos, larvae, pupae, adults
Raw reads	41,621,320	38,792,711
Processed reads	2,338,028	2,813,742
Total predicted miRNAs	113	139
Conserved miRNAs	87 + 3 ^a	80 + 2 ^a
Lineage-specific miRNAs	23	57

^aIn *C. ohridella*, we identified 87 miRNAs with entries in miRBase 20 as well as three newly reported sequences (Cam-042/Par-110, Cam-110 and Cam-154) shared with other lepidopterans. In *Pa. aegeria*, there were 80 miRNAs with entries in miRBase 20 as well as two newly reported sequences (Cam-042/Par-110 and Par-471). Cam-042 and Par-110 were discovered separately in *C. ohridella* and *Pa. aegeria* and are likely to be homologous miRNAs.

nymphalid. This species also has fairly complete genome coverage (19 out of 20 sequences recovered), and life-cycle stages can be collected from a laboratory-reared population.

Deep Sequencing and miRNA Identification

Small RNA sequencing was undertaken for multiple developmental stages to maximize representation of miRNAs expressed. In *Pa. aegeria* we sampled four time points during embryogenesis (1, 15, 24, and 48 h postoviposition), two larval time points (first and last instar), three pupal time points (0, 3, and 5 days postpupation), and a single adult male and female. *Cameraria ohridella* RNA samples were from larvae, pupae, and adults. In both species, whole bodies, eggs, or pupae were used at each time point for RNA extraction. RNA extracted from individual stages was pooled within each species.

A total of 80,414,031 raw Illumina HiSeq reads were obtained (table 1) and processed according to the schematic in figure 2, designed to ensure removal of spurious reads, degradation products, and other small RNAs. Because small RNA sequencing yields the mature miRNA products (22 nt), not precursor sequences (~80 nt), the miRDeep2 algorithm was used to extract predicted miRNA precursor sequences based on the positions of mapped reads in the genome. All predictions generated by miRDeep2 were first subjected to a free energy filtering step to select stable hairpins. Only genomic sequences fulfilling stringent stability criteria were considered to be candidate miRNA precursors.

Conserved miRNAs in *Cameraria* and *Pararge*

Of the candidate precursors identified, we first identified evolutionarily conserved sequences occurring in more than one species. A total of 87 and 80 “conserved miRNAs” of varying evolutionary ages were identified in *C. ohridella* and *Pa. aegeria*, respectively (supplementary file S2, Supplementary Material online). *Cameraria ohridella* has undergone lineage-specific duplications of several miRNA genes including miR-14, miR-2763, and miR-2796 of which only single copies exist in other Lepidoptera.

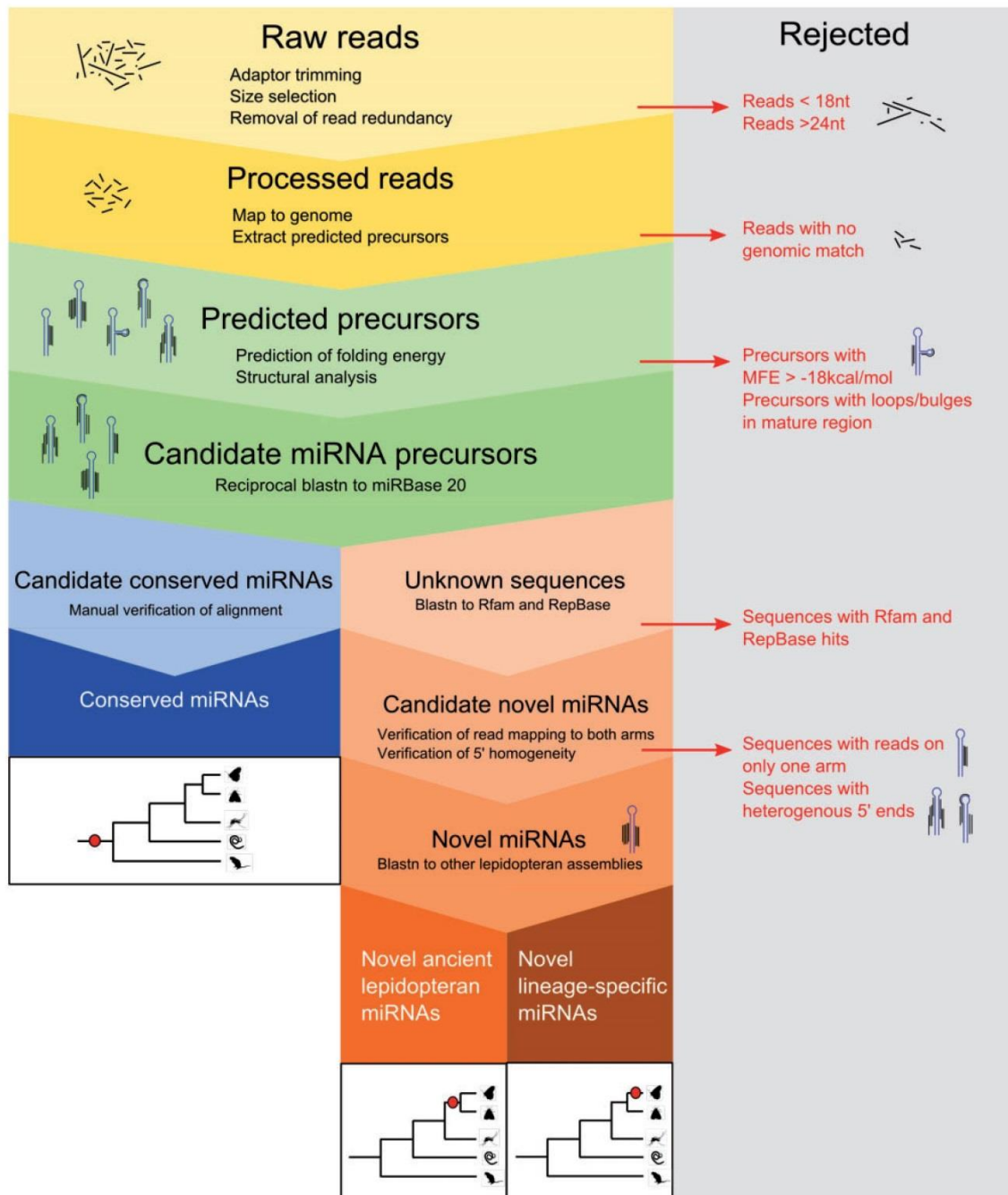


FIG. 2. Data processing pipeline for annotation of miRNAs in *Cameraria ohridella* and *Pararge aegeria* from raw Illumina sequencing output.

Novel miRNA discovery in *Cameraria* and *Pararge*

Following identification of previously documented miRNAs, we filtered the remaining candidates to retain only genuine miRNA genes and classify these into two categories. The first category consists of novel lineage-specific miRNAs found thus far only in the species sampled; the second comprises newly

discovered sequences of older origin which may have been missed in previous miRNA sequencing projects.

To ensure removal of non-miRNA gene sequences from our data set, we subjected all remaining sequences to additional validation (fig. 2). In particular, sequences were checked for 5'-homogeneity in the small RNA reads. Sequences with

highly heterogeneous 5'-ends in their mature reads are unlikely to represent genuine miRNAs as the 5'-end determines the miRNA seed (Kozomara and Griffiths-Jones 2011) and is likely to have effects on target specificity. We discarded these precursors as they could represent degradation products of longer transcripts. As a final conservative criterion, we stipulated that a candidate must have reads mapped to both arms of the hairpin in order to be described as a miRNA.

Comparison to available lepidopteran genome assemblies revealed that our data set includes 23 and 57 lineage-specific miRNAs in *C. ohridella* and *Pa. aegeria*, respectively (supplementary file S2, Supplementary Material online). Only one sequence, "Par471," was identified by BLASTn in several other ditrysian species but has not been reported in miRBase. The recent publication of a *P. xylostella* miRNA sequencing study (Etebari et al. 2013) has identified a predicted precursor (PC-3p-82883_6) which is likely to be Par-471. Conservation of this sequence was verified by alignment and synteny checks as well as structural analysis of predicted precursors in all species where this sequence is found (supplementary file S3, Supplementary Material online).

Phylogenetic Placement of Lepidopteran miRNAs

Using the information revealed by our deep sequencing together with previous studies on lepidopteran miRNAs, we determined the presence or absence of each sequence in each species (fig. 3A) and compared these to a phylogeny of the Lepidoptera (Regier et al. 2013). This revealed the latest possible node of origin for each sequence (fig. 3B). The caddisfly (*G. pellucidus*, Trichoptera) was used as an outgroup to the Lepidoptera. We validated the completeness of this genome by running BLASTn using the same set of 20 conserved miRNAs described previously (supplementary file S1, Supplementary Material online). Nineteen out of 20 of the query sequences were retrieved, suggesting that genome coverage of *G. pellucidus* is adequate and that the likelihood of miRNA genes reported absent as an artefact of genome completeness is low.

As expected, many miRNA precursors had origins prior to the last common ancestor of Lepidoptera and Trichoptera; these are not the focus of this study. A total of 459 sequences were identified which were unique to the Lepidoptera and/or Trichoptera. The majority of these are lineage-specific: 315 from *B. mori*, 10 from *H. melpomene*, 8 from *M. sexta*, 23 from *C. ohridella*, and 57 from *Pa. aegeria*. Of the novel sequences reported in *P. xylostella* (Etebari et al. 2013), we identified 21 sequences with reads mapping to both arms which are likely to be miRNA precursors (supplementary file S4, Supplementary Material online).

A total of 25 potential miRNA precursors were found to be present in Lepidoptera and Trichoptera, and shared by at least two species. Six of these (identified in previous studies) were removed from further analysis as we are not confident that they represent genuine miRNAs under our stringent criteria. These are miR-2739, miR-3389, miR-2733, miR-6495, miR-6496, and miR-6497. The miR-2738 sequence has a predicted secondary structure with an asymmetric bulge in its

mature region; its pattern of read mapping in *B. mori* is suggestive of a degradation product. miR-3389 displays a heterogeneous pattern of read mapping on its 5'-arm, again suggestive of a degradation product, and no reads from its 3'-arm have been reported. miR-2733 maps to several different genomic regions in *B. mori* and may be derived from a repetitive element. The miR-6495, miR-6496, and miR-6497 sequences were found to be derived from ribosomal RNA fragments.

We searched for the remaining 19 sequences in all available lepidopteran and trichopteran genome assemblies using BLASTn. The miR-2779 sequence was removed from further analysis as we are unable to confirm that the predicted precursors for this sequence identified in *B. mori*, *M. sexta*, and *P. xylostella* are homologous to each other by structure or synteny.

We plotted the latest possible origin for each miRNA onto the lepidopteran phylogeny (fig. 3B). Only one sequence (miR-2755) is shared between the Lepidoptera and its sister group Trichoptera. Strikingly, we find that the majority of novel miRNA acquisition is concentrated at the base of the lepidopteran tree, with 11 of these novel sequences likely to have been present in the last common lepidopteran ancestor. Two other sequences (Cam-110 and Cam-154) unite the distantly related sister species *C. ohridella* and *P. xylostella*. There were three further innovations on the Apoditrysian stem. We find very little miRNA acquisition on more recent internodes, with a single gain (miR-6303) in a common ancestor of the Heliconiinae and Nymphalinae. The current miRNA distribution also suggests that several lineage-specific losses may have occurred, though we have less certainty about miRNA loss due to the limitations of genome assemblies.

cubitus interruptus Is a Potential Target for miR-2768

Given the acquisition of new miRNAs early in lepidopteran evolution, we sought to identify target genes. This analysis is hampered by the lack of well-annotated lepidopteran transcriptomes with documented 3'-UTRs, the target of many animal miRNAs. However, applying the PITA target prediction algorithm (Kertész et al. 2007) to annotated coding sequences from the *H. melpomene* transcriptome (Hines et al. 2012) revealed an almost perfectly paired match between miR-2768 and the *cubitus interruptus* (*ci*) transcript. Subsequent analyses of the corresponding region in other lepidopterans reveal that this site is highly conserved between butterflies and moths (fig. 4), with 21 of the 22 bases (including the seed sequence) in miR-2768-3p capable of pairing with the consensus candidate site. This complementarity is greater than that observed between most bilaterian miRNAs and their targets. The miR-2768-3p sequence itself is also completely conserved in all lepidopteran species studied (supplementary file S5, Supplementary Material online) and its precursor gene lies in an intron of the *invected* (*inv*) gene.

To gauge if the candidate site in *ci* is identical between species by chance, or as a by-product of conservation of amino acid sequence, we assessed the overall conservation of the *ci* gene across Lepidoptera. Using genome data for

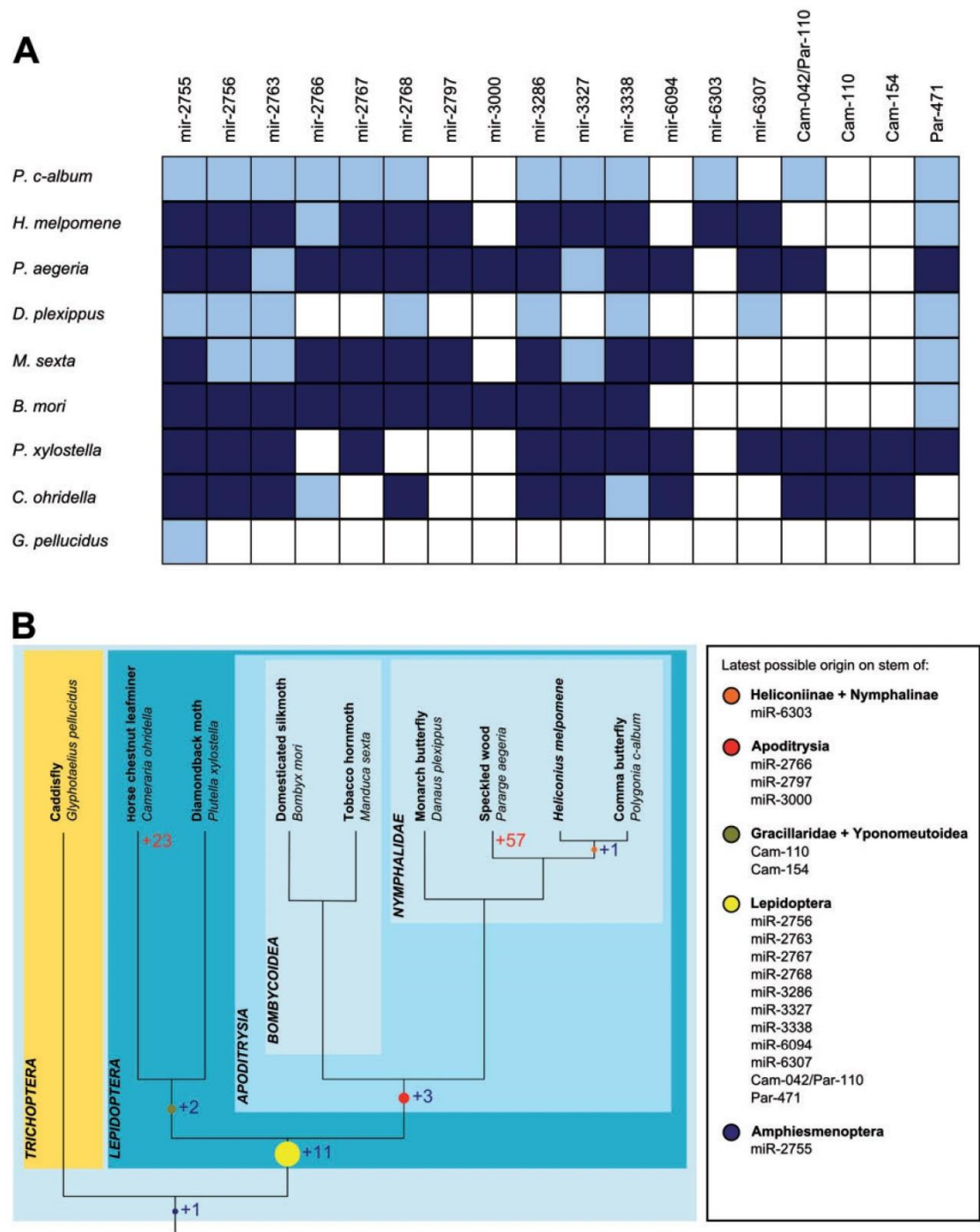


FIG. 3. (A) Presence/absence matrix for each Lepidopteran-specific miRNA family. Light blue squares indicated BLASTn and structural predictions of presence. Dark blue squares indicate evidence from deep sequencing. (B) Phylogenetic distribution of miRNAs in the Lepidoptera and its sister group Trichoptera showing inferred miRNA gains at each node. Lineage-specific gains are only plotted for *Cameraria ohridella* and *Pararge aegeria*.

three lepidopterans (*B. mori*, *D. plexippus*, and *H. melpomene*), we aligned predicted *ci* coding sequences and compared nucleotide and deduced peptide sequences (fig. 4). This revealed that while the peptide sequence of *Ci* is divergent between

different insect orders (supplementary file S5, Supplementary Material online) and only aligns well in the region containing the zinc finger domains, in Lepidoptera it is more conserved across much of its length (fig. 4, supplementary file S5,

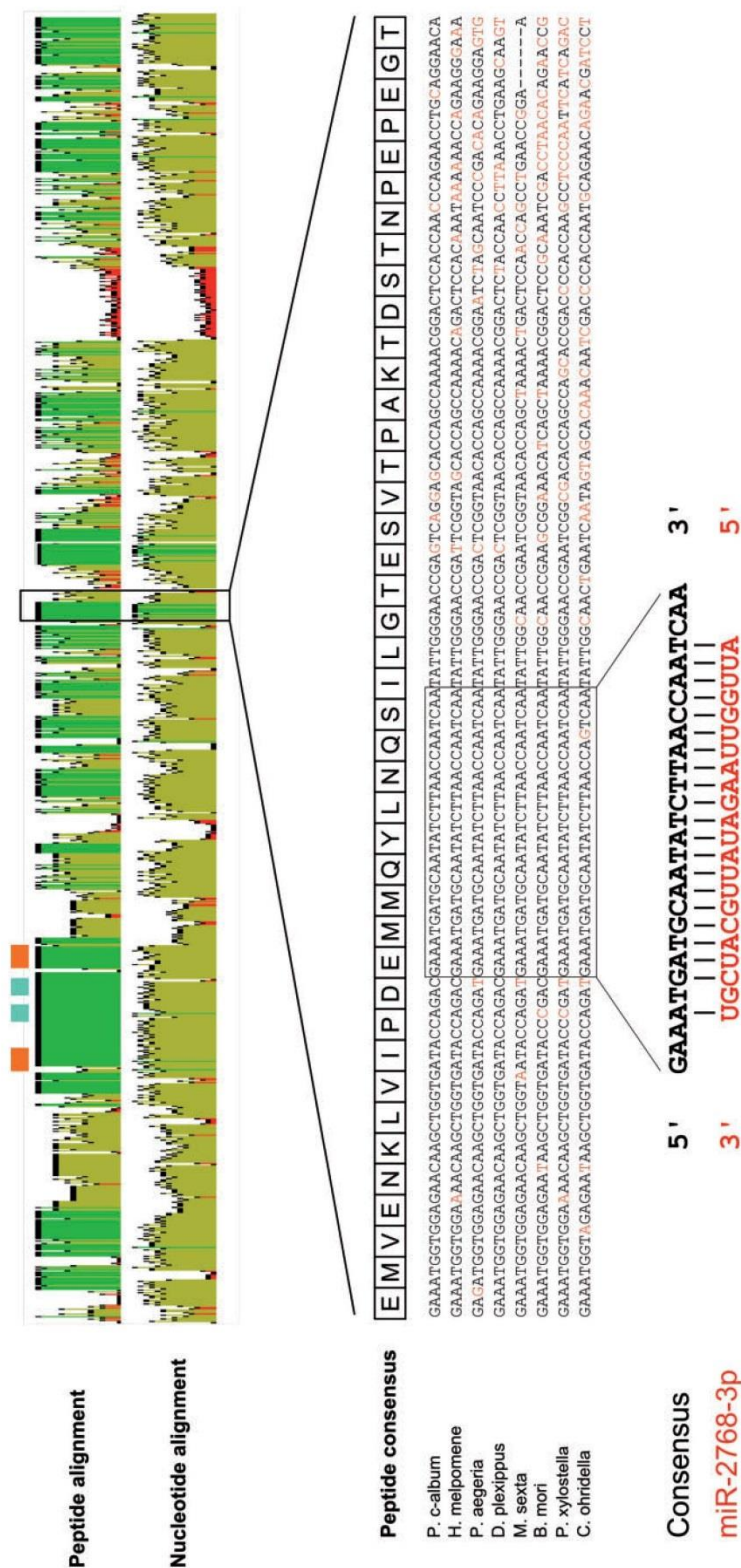


Fig. 4. Conservation of the peptide (top graph) and nucleotide (lower graph) sequence across the length of the *ci* gene in eight lepidopteran species. Positions with 100% mean pairwise identity between all pairs in the column are colored in green. Brown indicates positions with identity under 100% but above 30%. Red indicates less than 30% identity. The orange boxes show the positions of zinc finger C2H2 domains and blue boxes indicate the position of the zinc finger double domains. A closeup is shown of the nucleotide sequence alignment in the region of the putative miR-2768 binding site. Residues differing from the consensus are colored in red.

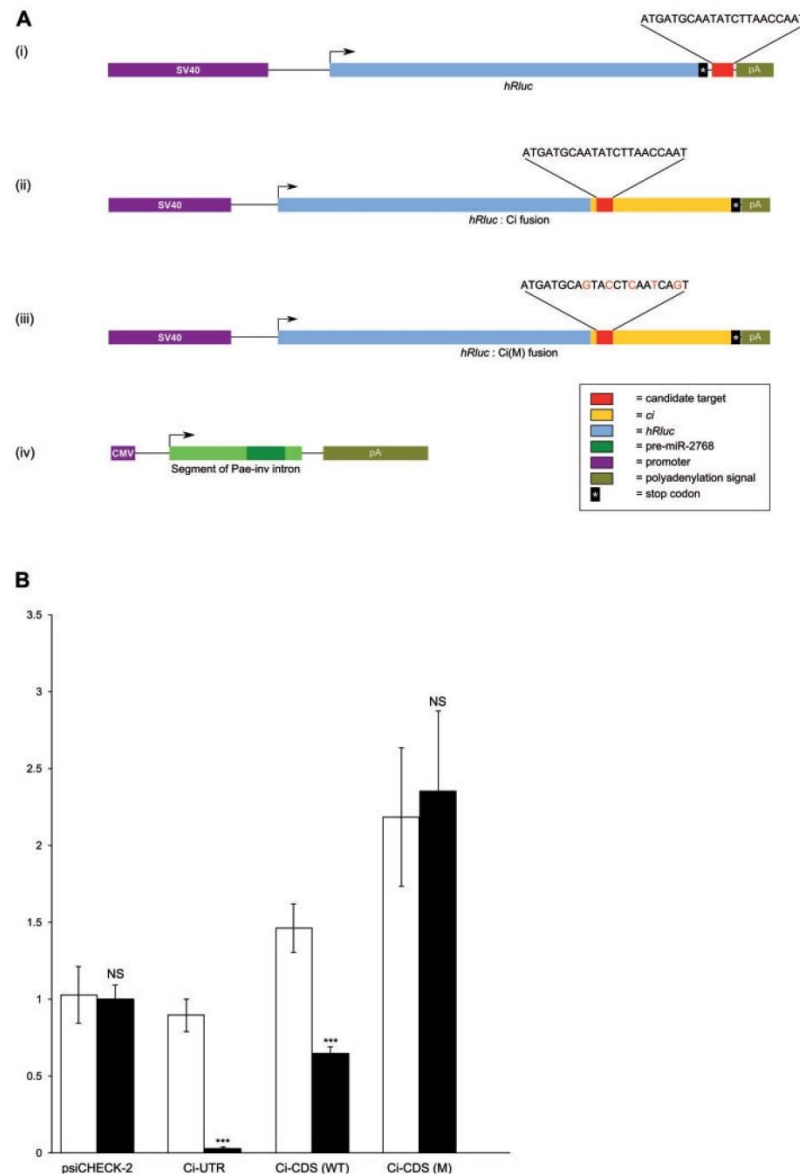


FIG. 5. (A) Constructs used for functional validation of *Pae*-miR-2768 binding to its predicted site in *Pae-ci*. (i) psiCHECK-2: *ci*-UTR, (ii) psiCHECK-2: *ci*-CDS(WT) (iii) psiCHECK-2: *ci*-CDS (M), (iv) pCS2+:2768. (B) Pooled luciferase assay data for the candidate miR-2768 target site in the presence or absence of miR-2768, and for wild-type and mutant *hRluc* fusion constructs in the presence or absence of miR-2768. Error bars represent standard deviation. ****P* < 0.001; NS, not significant.

Supplementary Material online). Consistent with selection acting at the amino acid level, the corresponding nucleotide sequence is weakly conserved with the majority of the coding sequence showing evidence of third codon position wobble. However, there are two clear regions in the predicted *ci* cDNAs where the nucleotide sequence is almost completely conserved, and strikingly one of these contains the putative miR-2768-3p target site. This implies there is a reason other than amino acid constraints that has forced conservation of local sequence, consistent with conservation as a target site for interaction with miR-2768-3p.

In order to functionally test the interaction of miR-2768-3p with its predicted target, we cloned the consensus 22 bp candidate site from lepidopteran *ci* into a luciferase expression vector downstream of the *hRluc* Renilla luciferase stop codon (fig. 5A (i); construct psiCHECK-2: *ci*-UTR). The vector also contains a synthetic firefly luciferase gene (*hLuc+*) as a control for transfection efficiency. We also generated an expression construct for miR-2768 by cloning a 339 bp fragment of the *inv* intron containing the miR-2768 precursor sequence under a constitutive promoter (fig. 5A (iv); construct pCS2+:2768). Following cotransfection, the ratio of

Renilla:firefly luciferase expression is used as a measure of miRNA activity (fig. 5B). We detect a very strong (~37-fold) downregulation of Renilla luciferase expression in the presence of both miR-2768 and its candidate site from *ci*. In the absence of the miRNA, no downregulation is observed. This confirms that the miR-2768 precursor region is capable of being processed into a functional miRNA in cells, and that this downregulates transcripts containing the predicted binding site, such as lepidopteran *ci* gene.

As the identified candidate site lies within the coding sequence, we further tested the ability of miR-2768 to downregulate a target when the site occurs within a coding region. To do this, we modified the *hRluc* gene on the psiCHECK-2 vector to encode a fusion protein between *hRluc* and the *ci* exon containing the target site (fig. 5A (ii); construct psiCHECK-2:ci-CDS). A mutant version of this construct was also generated with five synonymous transition mutations introduced into the candidate target site (fig. 5A (iii); construct ci-CDS(M), [supplementary file S5, Supplementary Material](#) online). These mutations should ensure that the protein sequence is not affected, while simultaneously disrupting miRNA binding. Both constructs were assayed for Renilla:firefly luciferase activity in the presence and absence of miR-2768. We detect downregulation of expression from the wild-type construct in the presence of miR-2768, but there is no downregulation of the mutant fusion construct (fig. 5B). A slight increase in Renilla:firefly luciferase activity was observed in the mutant construct relative to the wild-type irrespective of the presence of miR-2768. Overall, these experiments support our hypothesis that miR-2768 is capable of binding to the target sequence from the *ci* gene and exerting downregulation within coding sequences.

Discussion

miRNAs Discovery in *C. ohridella* and *Pa. aegeria*

Using small RNA sequencing, we identified 113 and 139 miRNA genes in the Horse chestnut Leafminer moth *C. ohridella* and the Speckled Wood butterfly *Pa. aegeria*, respectively. It is likely that these are not fully comprehensive figures, because some conserved ancient miRNAs known from other animals were not detected, perhaps due to highly restricted spatiotemporal expression. Although the number of sequences shared with at least one other species is similar in both organisms (90 in *C. ohridella*; 82 in *Pa. aegeria*), the number of lineage-specific sequences identified in *Pa. aegeria* (57) is more than detected in *C. ohridella* (23) despite equal sequencing depth in both species. We suggest this results from the difference in the range of developmental stages sampled in each species. *Pararge aegeria* was extensively sampled at different time points for each developmental stage, including embryos; in contrast, we lack data on small RNAs expressed during embryonic development in *C. ohridella*. Consistent with this, research in drosophilid fruitflies has shown that many young, fast-evolving miRNAs are highly expressed during embryogenesis and are downregulated during later development (Ninova et al. 2014). It is possible that the early embryo provides a more permissive

environment for gene expression changes; this suggestion is compatible with “phylogenetic hourglass” models that identify early developmental stages as prone to evolutionary change by incorporation of newly evolved genes (Domazet-Lošo and Tautz 2010).

Most of the evolutionarily “older” miRNAs have been reported before. This is not surprising because conserved miRNAs have been the focus of much study. We did identify one additional conserved miRNA, Par-471, from *Pa. aegeria* which we suggest is widely distributed across Lepidoptera, because its precursor sequence could be identified in most lepidopteran genome sequences. This miRNA is also present in a small RNA data set from the Diamondback moth *P. xylostella* (Etebari et al. 2013). Although the precursor has been conserved in sequence and structure across other lepidopterans ([supplementary file S5, Supplementary Material](#) online), it has not been reported in deep sequencing projects in the Apoditrysia. The mature Par-471 miRNA is not highly represented in our pooled *Pa. aegeria* library suggesting it may have a spatiotemporally restricted expression pattern.

Birth and Death of miRNAs in the Lepidoptera

Assigning miRNA genes to nodes on a phylogenetic tree of Lepidoptera yielded two seemingly paradoxical conclusions. First, there is extensive birth of new miRNAs along terminal branches in the phylogeny generating a plethora of young miRNAs. Second, most of the “conserved” miRNAs, shared by more than one species of Lepidoptera, originated early in phylogeny, with little evidence for integration of new miRNAs in the 100 My (Wahlberg et al. 2013) following the initial radiation of the Lepidoptera. How can these two conclusions be reconciled?

The present-day miRNA complement of any given lineage is the net result of the opposing forces of miRNA birth and miRNA death. The birth of new miRNAs may occur through de novo generation of new precursors from existing hairpin-like segments of existing transcripts (Campo-Paysaa et al. 2011) or through the duplication and divergence of existing miRNA genes. miRNAs may also be lost through mutations that disrupt the formation of a stable stem loop structure. The balance between these forces in shaping miRNA evolution throughout the Metazoa has been a subject of debate. Some authors have argued for a continuous addition of miRNAs to metazoan genomes over time with very low rates of secondary loss (Tarver et al. 2013); others have presented evidence for a more fluid model with continuous gains and losses operating on each stem in a clade (Meunier et al. 2013). What is clear, and also supported by our data, is that terminal nodes in a phylogeny tend to have an abundance of young, lineage-specific miRNAs (Lu et al. 2008; Nozawa et al. 2010; Meunier et al. 2013; Lyu et al. 2014). This was reported by Lu et al. (2008) in drosophilids (Lu et al. 2008) and although this study was subsequently criticized for overestimating lineage-specific miRNA gains through sampling of possible degradation products (Berezikov et al. 2010), evidence for a high miRNA birth rate on terminal branches was also later presented for mammals (Meunier et al. 2013) and confirmed in

drosophilids (Nozawa et al. 2010; Lyu et al. 2014). In Lepidoptera, the number of lineage-specific miRNAs identified with available miRNA deep-sequencing data are 23 in *C. ohridella*, 21 in *P. xylostella*, 315 in *B. mori*, 8 in *M. sexta*, 57 in *Pa. aegeria*, and 10 in *H. melpomene*. Although the exact numbers quoted for each lineage vary (probably due to depth of transcriptome sequencing and rigor of miRNA annotation), this is still indicative of the continuous birth of new miRNA sequences on terminal branches within the lepidopteran phylogeny.

It is estimated that a large proportion of new miRNAs (> 94% in drosophilids) will rapidly be lost (Lyu et al. 2014). The quick turnover phase in miRNA evolution is estimated to last for about 30 My in drosophilids, after which the remaining sequences will have been integrated into gene regulatory pathways. It is widely thought that miRNAs incorporated into such pathways are then unlikely to be lost, although there are exceptions. For example, secondary loss of conserved miRNAs have been noted in tapeworms, which may be linked to morphological simplification (Fromm et al. 2013), and loss of conserved miRNAs has also been documented in mammals (Meunier et al. 2013).

It is likely that the constant generation of new miRNA precursors has been going on throughout lepidopteran evolution. Birth rates of novel miRNAs are unlikely to drop substantially given the low requirements for transcripts to adopt a suitable secondary structure and become integrated into the miRNA processing pathway (Campo-Paysaa et al. 2011; Lyu et al. 2014), though it has been postulated that they may sometimes be increased through increased transcriptional activity in the genome (Heimberg et al. 2008). Despite this constant production of new material for selection to act upon, we find that hardly any new miRNAs have been integrated into lepidopteran miRNA repertoires following the establishment of the deeper branches in lepidopteran evolution. The distribution of lepidopteran-specific miRNAs shows the majority of innovation occurring along the Lepidopteran stem following divergence from the Trichoptera. Only a single sequence discovered so far, miR-2755, was acquired along the common lineage leading to the Lepidoptera and Trichoptera, and few or no conserved sequences were acquired in nodes such as Nymphalidae or Bombycoidea. Hence, we suggest that the concentration of new miRNA sequences close to the base of Lepidoptera, around 100 Ma (Wahlberg et al. 2013), was not a burst of accelerated miRNA birth, but rather a decrease in miRNA death rate. We argue that early in lepidopteran evolution, a set of new miRNA genes were incorporated into genetic regulatory pathways. Such pathways might include those governing lepidopteran-specific characters such as the patterning of wing and body scales, color vision, or interaction with anispermis.

The *ci* Gene Is a Likely Target for miR-2768

To understand how novel lepidopteran miRNAs were incorporated into genetic regulatory pathways, it is necessary to identify candidate target genes which these miRNAs regulate. The majority of identified target sites for metazoan miRNAs

lie within 3'-UTRs, possibly due to functional constraints (Grimson et al. 2007; Gu et al. 2009). This phenomenon may be somewhat exaggerated, however, as many miRNA target prediction programs consider only 3'-UTRs in predictions of miRNA binding (Lewis et al. 2005; Wang and El Naqa 2008; Maragkakis et al. 2009). Indeed, miRNA target sites located within coding sequences have been documented in animals (Forman et al. 2008; Tay et al. 2008), including humans (Forman et al. 2008), and have recently been suggested to be a common occurrence in cnidarians (Moran et al. 2014). There is some evidence suggesting that miRNA binding in coding regions may require higher sequence complementarity between the miRNA and its target, and proceed via RISC-mediated cleavage (Duursma et al. 2008).

We found evidence that the transcript from the *ci* gene is a target for downregulation by one of the novel lepidopteran miRNAs, miR-2768, within its coding sequence. All lepidopteran *ci* cDNAs contain a sequence which is an almost perfect complement to miR-2768-3p. This candidate target is strikingly well-conserved at the nucleotide level in all species analyzed, even though the majority of the *ci* coding sequence is not, suggestive of selection for function on its nucleotide sequence rather than only on its peptide sequence. Luciferase assays conducted in cell culture also indicate that this target site is capable of being targeted and downregulated by miR-2768, both when located within the 3'-UTR as well as when present as part of the coding sequence. We suggest this miRNA-target interaction arose early in lepidopteran evolution, since we find that the miR-2768 precursor sequence is present in all lepidopteran genomes examined, but absent in the caddisfly *G. pellucidus*. We also discovered that the miR-2768 gene itself is located in an intron of the homeobox gene *inv*, in the same transcriptional orientation, suggesting a mechanism by which miR-2768 was recruited to regulate *ci*.

We postulate that miR-2768 serves to downregulate the expression of *ci* during butterfly and moth morphogenesis, possibly during segmentation of the embryo or during post-embryonic patterning of the wing primordia, or indeed both processes. *Ci* is the insect homologue of the vertebrate Gli protein family and functions as an effector of Hedgehog (Hh) signaling. The Hh signaling pathway is essential in the establishment of compartment boundaries during anteroposterior segmentation of the insect embryo (Ingham and McMahon 2001), subsequent anteroposterior patterning of the cuticle in *Drosophila* (Sanson et al. 1999), as well as in the wing imaginal disc during late larval and pupal stages in insects (Keys 1999; Ingham and McMahon 2001). Common to these systems is a regulatory circuit (fig. 6) whereby the Hh ligand, expressed in posterior compartment cells under the influence of Engrailed (*En*) and its paralogue *Invected* (*Inv*), signals anteriorly across the compartment boundary to activate *wg* expression. The transcription factor acting downstream of the Hh signal to induce expression of *wg* and other downstream genes is *Ci*. *Wg* signals back across the compartment boundary to maintain expression of *en/inv*. In the *en/inv* domain, *ci* is not expressed due to repression by *En/Inv*. Regulation of *Ci* activity is essential as it mediates many of the downstream actions of Hh. This is achieved both at the transcriptional level through

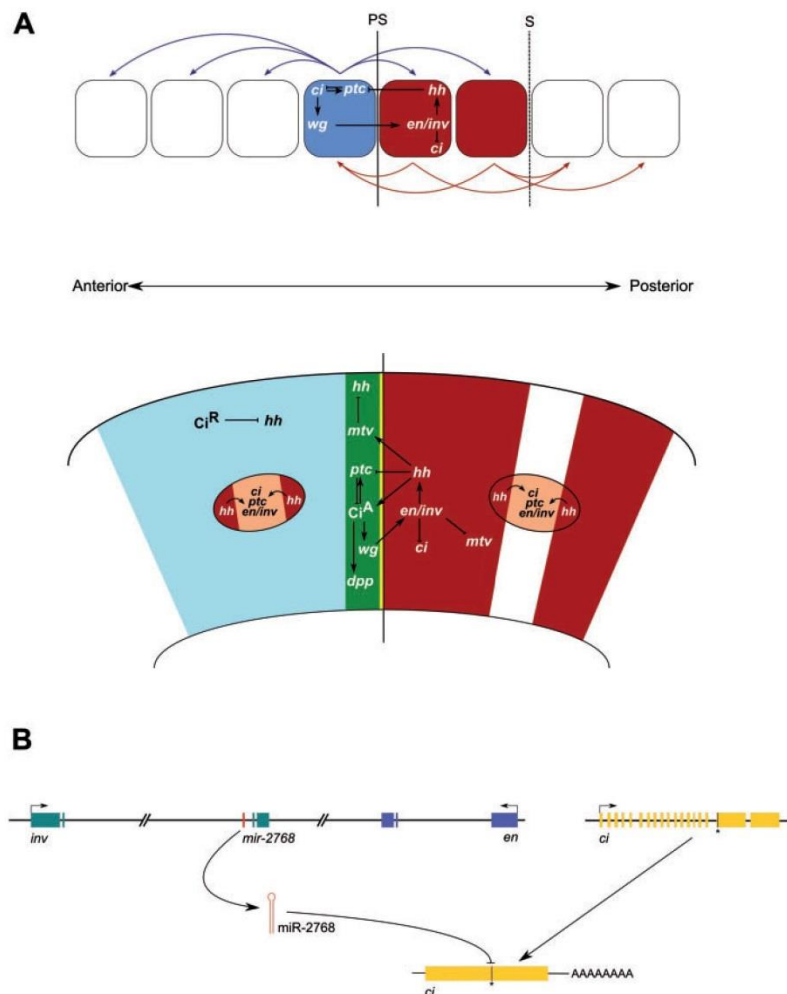


FIG. 6. (A) Involvement of *ci* in the Hedgehog signaling pathway during anteroposterior patterning of the *Drosophila melanogaster* embryo (above) and wing imaginal disc patterning in *Precis coenia* (below). PS, parasegment boundary. S, segment boundary. (B) Model for the contribution of miR-2768 to the repression of *ci*. The asterisk indicates the position of the miR-2768-3p binding site in the coding region of the *ci* messenger RNA.

the actions of En/Inv in repressing *ci* transcription, as well as through posttranslational modifications to the Ci protein which regulate its localization and activity. We propose that in the Lepidoptera, miR-2768 functions to add another level of regulation over the Ci activity at a posttranscriptional level. As miR-2768 is located within the transcription unit of *inv*, it is likely to be coexpressed with *inv* (Rodriguez et al. 2004; Baskerville and Bartel 2005) and may reinforce the repression of *en/inv* on *ci*.

There is another system where Hh signaling has been co-opted for use in patterning where the gene interactions described above have been modified. Many butterflies and moths have large, brightly colored eyespots on their wings. Eyespot morphology and genes involved in determining eyespot color and pattern may differ between species (Brunetti et al. 2001; Oliver et al. 2012), and the regulatory networks behind eyespot focal specification have only been studied in

detail only in Nymphalidae (Oliver et al. 2012). In their report, they propose that a network co-option event occurred in the evolution of eyespot focal patterning and was subsequently modified in different lineages to generate novel patterns. In *Bicyclus anynana* and *Junonia coenia*, focal development first involves an accumulation of *hh* in cells flanking developing foci. Hh signaling subsequently induces expression of *ptc*, *ci*, and *en/inv* in cells which will form the focus (Keys 1999; fig. 6). This process of eyespot specification occurs regardless of whether the focus develops in the anterior or posterior compartment, implying a relaxation of compartmental restrictions. Also relaxed must be the transcriptional repression of *ci* by *en/inv* as both are coexpressed in the same cells. It has been noted that some miRNAs are coexpressed with their target genes, and it has been postulated that their role may be to subtly modulate target mRNA levels. We speculate that perhaps miR-2768, produced from the intron of *inv*, may

subtly regulate the levels of *ci* mRNA in regions of coexpression such as eyespots.

Concluding Remarks

We demonstrate that the majority of miRNA innovation retained by selection within the Lepidoptera arose early in lepidopteran evolution and suggest that novel lepidopteran miRNAs were recruited for regulatory roles in embryonic and wing patterning. We describe a novel gene regulatory network involving *ci*, *inv*, and miR-2768 in butterflies and moths, in which miR-2768, generated from an intron of *inv*, downregulates *ci*. This network may be involved in segmentation and in patterning of wing primordia in Lepidoptera.

Materials and Methods

Sample Collection

Cameraria ohridella larvae and pupae were obtained by dissection from leaf mines on Horse chestnut (*Aesculus hippocastanum*) leaves collected in Oxford University Parks, UK, in July 2013. Adults were collected by incubating leaves in a damp environment in the laboratory at room temperature. Twenty individuals were sampled for each developmental stage. Adults were not sexed.

Pararge aegeria specimens were obtained from a colony at Oxford Brookes University originally founded with specimens from Belgium (Carter et al. 2013). Eggs were collected within 20 min of oviposition and homogenized 1, 15, 24, and 48 h after oviposition. For first instar larvae, a mixture of whole larvae was homogenized for a total mass of 0.10 g. Two final instar larvae (total mass = 0.11 g) were used. Single whole pupae were sampled at 0, 3, and 5 days postpupation. One adult male and one adult female were homogenized separately following removal of the wings and legs.

RNA Extraction and miRNA Sequencing

Small RNA extraction was carried out using the miRVana miRNA Isolation Kit (Life Technologies, AM1560) following the manufacturer's protocol. Tissues were homogenized in miRVana Lysis Buffer using a single bead in the TissueLyser II (Qiagen). All tissues were homogenized fresh.

Quantity and quality of the extracted RNA was assessed using the Experion Automated Electrophoresis System (Bio-Rad) on an Experion HighSens chip. Sample pooling was carried out for each species such that equal amounts of RNA were used from each sampling time point. Library preparation and deep sequencing were carried out at the Beijing Genomics Institute (Hong Kong). The two libraries were barcoded and run on a single lane of the Illumina HiSeq 2000 platform. Raw read and processed data are available on the Gene Expression Omnibus (accession number GSE63644).

Computational Analysis

We developed a conservative pipeline for identification of true miRNA sequences in the small RNA sequencing reads (fig. 2). FastQC (<http://www.bioinformatics.babraham.ac.uk/>

[projects/fastqc](http://www.bioinformatics.babraham.ac.uk/projects/fastqc), last accessed January 23, 2015) was first used to validate read quality. Raw reads were processed by converting FASTQ to FASTA file formats and removing adapter sequences. Reads were collapsed to remove redundancy; collapsed reads less than 18 nt and greater than 24 nt were discarded. Remaining reads were mapped to Bowtie-indexed genome assemblies and those that failed to align were discarded.

The miRDeep2 algorithm (Friedländer et al. 2012) was used to predict precursor hairpins at genomic locations where the reads mapped. Each predicted precursor was then manually inspected for the following criteria before being accepted as a true putative miRNA: stable hairpin structure with less than −18 kcal/mol free energy, at least 18 paired bases on the main stem, and the absence of large internal loops and bulges in mature regions. Free energy of the predicted precursors was calculated using RNAeval.

For conserved miRNA identification, all precursors passing the structural criteria were checked against known precursors in miRBase 20 using the following BLASTn parameters: word_size = 4, reward = 5, penalty = −4, gapopen = 8, and gapextend = 6. BLAST hits were assessed and verified by reciprocal BLASTn. Verified sequences were annotated as conserved miRNAs. Sequences without BLASTn hits to miRBase 20 were subjected to further processing and retained only if they passed the following criteria: homogeneity of 5′-ends in mature products, read mapping to both arms of the hairpin, and no BLASTn hits to RepBase, Rfam, or NCBI bacterial sequences (taxid:2). The positions where each candidate precursor mapped onto the genome were screened to remove duplicate, overlapping predictions.

Phylogenetic Placement of Lepidopteran miRNAs

All documented miRNA precursor sequences were downloaded from miRBase 20 (Kozomara and Griffiths-Jones 2011) and filtered to remove sequences present outside of the Lepidoptera (represented by *B. mori*, *M. sexta*, and *H. melpomene*). Lepidopteran-specific miRBase entries, along with sequences reported here, were then compared by BLASTn against genome assemblies for *H. melpomene* (www.butterflygenome.org, last accessed January 23, 2015), *D. plexippus* (<http://monarchbase.umassmed.edu/>, last accessed January 23, 2015), *M. sexta* (<http://www.hgsc.bcm.edu/arthropods/tobacco-hornworm-genome-project>, last accessed January 23, 2015), *B. mori* (<http://sgp.dna.affrc.go.jp/KAIKObase/> and <http://silkworm.genomics.org.cn/silkdb/>, last accessed January 23, 2015), *P. xylostella* (<http://dbm.dna.affrc.go.jp/px/>, last accessed January 23, 2015), *G. pellucidus*, *Hep. sylvina*, *C. ohridella*, *Pa. aegeria*, and *Polygonia c-album* using the BLASTn parameters described above. Genome assembly data for *G. pellucidus*, *Hep. sylvina*, *C. ohridella*, and *Pa. aegeria* are also publicly available in the Oxford University Research Data archive (doi: 10.5287/bodleian.dury3) and raw sequence data are available in the NCBI BioProject database (accession number PRJNA241175), as reported in Ferguson et al. (2014). Each species was scored for presence or absence of individual miRNA families. Where possible,

syntenic conservation was verified by comparing the positions of the miRNA and its flanking genes between species.

miRNA Target Predictions

Transcriptome data for *H. melpomene* were downloaded from EnsemblMetazoa. The PITA target prediction algorithm (Kertesz et al. 2007) with default parameters was run on all contigs using a cutoff score of -20 . This algorithm computes the difference in free energy gained upon pairing of the miRNA to its target and free energy lost in opening up RNA secondary structure in order to make the site accessible to miRNA binding.

Vector Construction

A synthetic oligonucleotide duplex was constructed consisting of the predicted miR-2768 target site (22 bp) from the *Pa. aegeria ci* gene flanked by restriction sites and inserted into the multiple cloning site of the psiCHECK-2 vector (Promega, C8021) downstream of the *hRluc* stop codon to make the psiCHECK-2:ci-UTR construct. To generate a construct, psiCHECK-2:ci-CDS(WT), with the putative *ci* target site within its coding region, a 369 bp region of the *ci* exon containing the site was cloned and inserted immediately preceding the stop codon of *hRluc*. A mutated version with synonymous transition mutations within the candidate site was generated by polymerase chain reaction (PCR) to give the psiCHECK-2:ci-CDS(M) construct. All constructs were validated by sequencing. For construction of the miR-2768 expression vector, a 339 bp segment containing the miRNA precursor was amplified by PCR from *Pa. aegeria* genomic DNA and inserted into the multiple cloning site of pCS2+.

Luciferase Assay Protocol

HeLa cells (ATCC CCL-2) were cultured in 25-cm² flasks at 37°C, 5% CO₂ in Dulbecco's modified Eagle's medium (Sigma-Aldrich, D5671) supplemented with 10% fetal calf serum, 2 mM L-glutamine, and penicillin/streptomycin. Cells were passaged and transferred to 96-well assay plates (Sigma-Aldrich, CLS3603) 36 h prior to transfection. The time between plating and transfection was optimized experimentally for maximal transfection efficiency. Transfection was carried out using FuGENE HD (Promega, E2311). Each transfection reaction was set up according to the manufacturer's protocol in 100 µl using a total of 2 µg of plasmid DNA and 6 µl of FuGENE HD. Luciferase assays were performed 36 h after transfection with the Dual-Glo Luciferase Assay System (Promega, E2920). Firefly and Renilla luciferase expression was quantified by measuring luminescence with no emission filter on the FLUOstar Omega microplate reader (BMG LABTECH). For each assay, three biological replicates (two for psiCHECK-2:ci-UTR without miR-2768) on separate plates were performed at separate times and relative Renilla and firefly luciferase activities were recorded. Each biological replicate consisted of three or five technical replicates. The mean Renilla:firefly ratio for pCS2+:2768 cotransfected with

blank psiCHECK-2 was used to normalize readings across plates.

Supplementary Material

Supplementary files S1–S5 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

The authors thank Casper Breuker, Jean-Michel Carter, and Luca Livraghi for advice and assistance in obtaining material from *Pa. aegeria*; Adrian Smith, Stephen Preston, Xinyang Hong, Anita Schamberger, Rachael Nimmo, and Caroline Bacquet for advice on cell culture; James Tarver, Ferdinand Marletaz, and Jordi Paps for discussions and bioinformatics advice; the Oxford University Museum of Natural History for access to additional lepidopteran specimens for photography. S.Q. acknowledges support from an A*STAR National Science PhD Scholarship and the Elizabeth-Hannah Jenkinson Fund, P.W.H.H. and J.H.L.H. acknowledge support from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)/ERC grant [268513].

References

- Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 11:241–247.
- Bazinnet AL, Cummings MP, Mitter KT, Mitter CW. 2013. Can RNA-seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. *PLoS One* 8:e82615.
- Berezikov E, Liu N, Flynt AS, Hodges E, Rooks M, Hannon GJ, Lai EC. 2010. Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*. *Nat Genet* 42:6–9; author reply 9–10.
- Brooke NM, Garcia-Fernández J, Holland PW. 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392: 920–922.
- Brunetti CR, Selegue JE, Monteiro A, French V, Brakefield PM, Carroll SB. 2001. The generation and diversification of butterfly eyespot color patterns. *Curr Biol* 11:1578–1585.
- Campo-Paysaa F, Sémon M, Cameron RA, Peterson KJ, Schubert M. 2011. microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol Dev* 13:15–27.
- Carter J-M, Baker SC, Pink R, Carter DRF, Collins A, Tomlin J, Gibbs M, Breuker CJ. 2013. Unscrambling butterfly oogenesis. *BMC Genomics* 14:283.
- Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Ptx1 enhancer. *Science* 327:302–305.
- Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468: 815–818.
- Duursma AM, Kedde M, Schrier M, le Sage C, Agami R. 2008. miR-148 targets human DNMT3b protein coding region. *RNA* 14:872–877.
- Etebari K, Hussain M, Asgari S. 2013. Identification of microRNAs from *Plutella xylostella* larvae associated with parasitization by *Diadegma semiclausum*. *Insect Biochem Mol Biol* 43:309–318.
- Ferguson L, Marlétaz F, Carter J-M, Taylor WR, Gibbs M, Breuker CJ, Holland PWH. 2014. Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS Genet* 10:e1004698.

- Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*. 101: 18058–18063.
- Forman JJ, Legesse-Miller A, Collier HA. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A*. 105: 14879–14884.
- Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 40:37–52.
- Fromm B, Worren MM, Hahn C, Hovig E, Bachmann L. 2013. Substantial loss of conserved and gain of novel microRNA families in flatworms. *Mol Biol Evol*. 30:2619–2628.
- Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M. 2011. MicroRNA evolution by arm switching. *EMBO Rep*. 12:172–177.
- Grimson A, Farh KK-H, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*. 27:91–105.
- Gu S, Jin L, Zhang F, Sarnow P, Kay MA. 2009. Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nat Struct Mol Biol*. 16:144–150.
- Heimberg AM, Sempere LF, Moy VN, Donoghue PCJ, Peterson KJ. 2008. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A*. 105:2946–2950.
- Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*. 487:94–98.
- Herranz H, Cohen SM. 2010. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes Dev*. 24: 1339–1344.
- Hines HM, Papa R, Ruiz M, Papanicolaou A, Wang C, Nijhout HF, McMillan WO, Reed RD. 2012. Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation. *BMC Genomics*. 13:288.
- Hui JHL, Marco A, Hunt S, Mellington J, Griffiths-Jones S, Ronshaugen M. 2013. Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. *Nucleic Acids Res*. 41:3352–3361.
- Ingham PW, McMahon AP. 2001. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev*. 15:3059–3087.
- Jeong S, Rokas A, Carroll SB. 2006. Regulation of body pigmentation by the abdominal-B Hox protein and its gain and loss in *Drosophila* evolution. *Cell*. 125:1387–1399.
- Kawahara AY, Breinholt JW. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc Biol Sci*. 281: 20140970.
- Kenny N, Quah S, Holland PWH, Tobe SS, Hui JHL. 2013. How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment? *Gen Comp Endocrinol*. 16–22.
- Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007. The role of site accessibility in microRNA target recognition. *Nat Genet*. 39: 1278–1284.
- Keys DN. 1999. Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science*. 283:532–534.
- Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 39:D152–7.
- Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 120:15–20.
- Liu S, Li D, Li Q, Zhao P, Xiang Z, Xia Q. 2010. MicroRNAs of *Bombyx mori* identified by Solexa sequencing. *BMC Genomics*. 11:148.
- Liu S, Zhang L, Li Q, Zhao P, Duan J, Cheng D, Xiang Z, Xia Q. 2009. MicroRNA expression profiling during the life cycle of the silkworm (*Bombyx mori*). *BMC Genomics*. 10:455.
- Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu C-L. 2008. The birth and death of microRNA genes in *Drosophila*. *Nat Genet*. 40:351–355.
- Lyu Y, Shen Y, Li H, Chen Y, Guo L, Zhao Y, Hungate E, Shi S, Wu C-L, Tang T. 2014. New MicroRNAs in *Drosophila*—birth, death and cycles of adaptive evolution. *PLoS Genet*. 10:e1004096.
- Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, et al. 2009. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*. 37:W273–W276.
- Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Res*. 23:34–45.
- Mita K. 2004. The Genome sequence of silkworm, *Bombyx mori*. *DNA Res*. 11:27–35.
- Moran Y, Fredman D, Praher D, Li XZ, Wee LM, Rentzsch F, Zamore PD, Technau U, Seitz H. 2014. Cnidarian microRNAs frequently regulate targets by cleavage. *Genome Res*. 651–663.
- Ninova M, Ronshaugen M, Griffiths-Jones S. 2014. Fast-evolving microRNAs are highly expressed in the early embryo of *Drosophila virilis*. *RNA*. 20:360–372.
- Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in *Drosophila* species. *Genome Biol Evol*. 2:180–189.
- Oliver JC, Tong X-L, Gall LF, Piel WH, Monteiro A. 2012. A single origin for nymphalid butterfly eyespots followed by widespread loss of associated gene expression. *PLoS Genet*. 8:e1002893.
- Peterson KJ, Dietrich MR, McPeck MA. 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays*. 31:736–747.
- Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature*. 453:1064–1071.
- Regier JC, Mitter C, Zwick A, Bazinet AL, Cummings MP, Kawahara AY, Sohn J-C, Zwick DJ, Cho S, Davis DR, et al. 2013. A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS One*. 8:e58568.
- Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res*. 14:1902–1910.
- Sanson B, Alexandre C, Fascetti N, Vincent J-P. 1999. Engrailed and hedgehog make the range of wingless asymmetric in *Drosophila* embryos. *Cell*. 98:207–216.
- Surridge AK, Lopez-Gomollon S, Moxon S, Maroja LS, Rathjen T, Nadeau NJ, Dalmay T, Jiggins CD. 2011. Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene*. *BMC Genomics*. 12:62.
- Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PCJ, Peterson KJ. 2013. miRNAs: small genes with big potential in metazoan phylogenetics. *Mol Biol Evol*. 30: 2369–2382.
- Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I. 2008. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*. 455:1124–1128.
- Wahlberg N, Wheat CW, Peña C. 2013. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PLoS One*. 8:e80875.
- Wang X, El Naqa IM. 2008. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*. 24:325–332.
- Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, et al. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*. 306:1937–1940.
- Yekta S, Shih I-H, Bartel DP. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*. 304:594–596.
- You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat Genet*. 45: 220–225.
- Zhan S, Merlin C, Boore JL, Reppert SM. 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell*. 147: 1171–1185.

RESEARCH

Open Access



The Hox cluster microRNA miR-615: a case study of intronic microRNA evolution

Shan Quah and Peter W. H. Holland* 

Abstract

Background: Introns represent a potentially rich source of existing transcription for the evolution of novel microRNAs (miRNAs). Within the Hox gene clusters, a miRNA gene, *miR-615*, is located within the intron of the *Hoxc5* gene. This miRNA has a restricted phylogenetic distribution, providing an opportunity to examine the origin and evolution of a new miRNA within the intron of a developmentally-important homeobox gene.

Results: Alignment and structural analyses show that the sequence is highly conserved across eutherian mammals and absent in non-mammalian tetrapods. Marsupials possess a similar sequence which we predict will not be efficiently processed as a miRNA. Our analyses suggest that transcription of *HOXC5* in humans is accompanied by expression of *miR-615* in all cases, but that the miRNA can also be transcribed independently of its host gene through the use of an intragenic promoter. We present scenarios for the evolution of *miR-615* through intronic exaptation, and speculate on the acquisition of independent transcriptional regulation. Target prediction and transcriptomic analyses suggest that the dominant product of *miR-615* is involved in the regulation of growth and a range of developmental processes.

Conclusions: The *miR-615* gene evolved within the intron of *Hoxc5* in the ancestor of placental mammals. Using *miR-615* as a case study, we propose a model by which a functional miRNA can emerge within an intron gradually, by selection on secondary structure followed by evolution of an independent miRNA promoter. The location within a Hox gene intron is of particular interest as the miRNA is specific to placental mammals, is co-expressed with its host gene and may share complementary functions.

Keywords: Homeobox, miRNA, Mammal, Hoxc5, Intron

Background

Intronic miRNAs provide an opportunity to gain insight into de novo miRNA innovation and its contribution to the evolution of development. Pre-existing transcribed RNAs, particularly in the form of introns, are thought to be a major source from which new miRNAs evolve [1, 2]. Here we analyse the origin and evolution of *miR-615*, a mammalian intronic miRNA located within a Hox cluster. We collate, analyse and integrate data on the regulation, expression pattern and likely biological roles of *miR-615*, in concert with comparative analysis, and

propose how the origin of this miRNA contributed to eutherian evolution and development.

Homeobox genes form a large, diverse superclass with a wide range of roles in metazoan developmental biology. Most encode transcription factors which exert regulation over the expression of other genes during embryogenesis, thereby having a direct influence on the resultant adult morphology [3]. Perhaps the best characterised of the homeobox genes are the Hox genes, whose primary roles are in the specification of anteroposterior patterning in the developing bilaterian embryo. Hox genes also have other important roles during animal morphogenesis such as patterning of the limbs [4], embryonic kidney [5, 6] and reproductive tract [7]. Metazoan Hox clusters contain miRNA genes of varying ages and evolutionary origin. The *miR-10* gene is conserved between arthropods

*Correspondence: peter.holland@zoo.ox.ac.uk
Department of Zoology, University of Oxford, South Parks Road,
Oxford OX1 3PS, UK

and mammals and occurs in the same position in insect and vertebrate Hox clusters [8]. The *miR-196* and the bi-directionally transcribed *miR-iab-4* genes are chordate-specific [9] and arthropod-specific [10], respectively. Another Hox cluster miRNA gene, *miR-615*, is less well-characterised and has been reported to be restricted to either mammals or amniotes [11]. A detailed analysis of the phylogenetic distribution of *miR-615* is currently lacking. Unlike *miR-10*, *miR-196* and *miR-iab-4*, which are intergenic, *miR-615* is nested within an intron of the *Hoxc5* gene in mouse and human. This positioning facilitates the evolutionary analysis of *miR-615* because the highly conserved nature of Hox gene exons allows alignment and comparison of intronic regions.

As *miR-615* occurs in the same orientation as *Hoxc5*, it is likely to be co-transcribed with its host gene. Work on other miRNAs supports a general concept of co-expression of intronic miRNAs with their host transcripts [12, 13], as well as revealing complementary functions performed by an intronic miRNA and its host gene product [14], although there is also evidence for some intronic miRNAs also having specific promoters [12, 15, 16].

Using DNA sequence data, derived from the public domain and obtained through our own experiments, we investigate the phylogenetic distribution of *mir-615* across tetrapods. We perform structural predictions and evaluate minimum folding energies to evaluate whether candidate pre-*miR-615* transcripts could be efficiently processed into mature miRNAs, and argue that this most likely occurs only in eutherian mammals. We also integrate RNA-seq data with chromatin modification and expressed sequence tag (EST) studies to argue for the existence of a promoter specific to *mir-615* located within the transcription unit of *Hoxc5*. Finally, we carry out target predictions for the dominant 3p product of *miR-615* and compare these to expression domains to suggest likely in vivo functions of this miRNA, and how these contributed to the evolution and development of mammals.

Methods

Conservation analysis for *mir-615*

Eutherian *Hoxc5* intronic sequences, along with those from *Xenopus tropicalis* and *Anolis carolinensis* were obtained from the UCSC Genome Browser [17] and the Ensembl genome database [18]. The *Sarcophilus harrisii* sequence was obtained from Ensembl while *Macropterus eugenii* sequence data were obtained from GenBank (accession: JN378720.1). Monotreme DNA was obtained from Stephen Donnellan in the South Australian Museum via the Oxford University Museum of Natural History (CITES registration number: GB026). Three species were sampled: Platypus (*Ornithorhynchus anatinus*),

Short-beaked echidna (*Tachyglossus aculeatus*) and Long-beaked echidna (*Zaglossus bruijnii*). Xenarthran samples were provided by Mads Bertelsen at the Copenhagen Zoo; xenarthran genomic DNA was obtained from sloth (*Choeloeopus didactylus*) heart and anteater (*Myrmecophaga tridactyla*) liver by phenol–chloroform extraction. Data for American pika (*Ochotona princeps*), lesser Egyptian jerboa (*Jaculus jaculus*) and common shrew (*Sorex araneus*) were obtained by running BLASTn using the hsa-miR-615 sequence against draft genome assemblies for these species. Alignment was carried out using MAFFT [19].

Primer sequences used for amplification of all *Hoxc5* intronic sequences are: 5'-TTGGACTTAAGCATCAC TTTCCACCG-3' and 5'-CCAGAGTCTGGTAGCGC GTGTAAGTGG-3'. These were designed to bind a region in the flanking coding exons conserved between most tetrapods (Additional file 1: Supplement S1). The annealing temperature was 60 °C for anteater and 55 °C for sloth. PCR products were gel purified using the illustra GFX PCR and DNA Gel Band Purification Kit (GE Healthcare, 28-9034-71). The sloth sequence was cloned into the pGEM-T Easy Vector System (Promega) before sequencing; the anteater amplicon was directly sequenced following gel purification with the same primers used for amplification.

Structural criteria for the annotation of functional miR-615 are: stable hairpin structure with <−18 kcal/mol free energy, at least 18 paired bases on the main stem and the absence of large internal loops and bulges in mature regions. Minimum free energy and predicted secondary structure was calculated for each putative miR-615 sequence using RNAeval.

Bioinformatics and target prediction

Data used to determine expression of *mir-615* and *HOXC5* in human cell lines were obtained from the ENCODE project and visualised on the UCSC Genome Browser (<http://genome.ucsc.edu/ENCODE/>).

A list of publicly released datasets used in bioinformatic analysis of *mir-615* and mRNA expression is provided in Additional file 2: Supplement S2. These data were obtained from the Gene Expression Omnibus [20] and the ENCODE project. In the microarray studies, a gene was determined to be expressed based on the ABS_CALL value (P = present, A = absent, if provided). If this was not provided, a cut-off *p* value of <0.05 was used to determine expression. Raw data for the transcriptome assemblies were quality checked using FastQC, followed by trimming of the first 15 bases using the fastqtrim.py script. Trimmed reads were assembled using Trinity and resultant contig FPKMs determined with RSEM. BLASTn with standard parameters was used to

determine Hox gene expression in each transcriptome, with appropriate sequences from the corresponding species obtained from GenBank. Small RNAseq data were analysed through the miRDeep2 pipeline using the quantifier.pl script to map reads onto known human miRNAs. Handling of raw reads prior to mapping was carried out as described in Quah et al. [14]. Read counts for mouse developmental stages as presented by Chiang et al. [21] were obtained from miRBase [22].

Target prediction for hsa-miR-615 was run using the downloadable executable for the PITA prediction algorithm [23] on all human cDNAs downloaded from Ensembl BioMart (genome assembly: GRCh38.p2). Predicted targets were ranked by score and a cutoff score of -20 was applied. Documented functions for predicted targets were obtained from UniProtKB/Swiss-Prot (<http://www.uniprot.org/> [24]). RPKM values for tissue-specific expression of each gene were obtained from the Human Protein Atlas [25], except for cerebellum where RPKM values were obtained from gene quantifications available through the ENCODE project (accession: ENCFF917ANZ).

mir-615 promoter prediction

Histone methylation, DNase I and CAGE data used to find the position of the *mir-615* promoter were obtained using ENCODE data tracks available through the UCSC Genome Browser (<http://genome.ucsc.edu/ENCODE/>). The EST data track used ('Human ESTs Including Unspliced') was generated at UCSC from publicly available EST sequencing data.

Results

A sequence with high similarity to *mir-615* is present in therian mammals

miR-615 is reported as being restricted to either mammals or amniotes [11] but current literature lacks a detailed analysis of its evolutionary history. We first sought to characterise its phylogenetic distribution to determine its most likely time of origin.

Available entries for *mir-615* on miRBase [22] are limited to two of the four main clades of eutherian mammals (Laurasiatheria and Euarchontoglires, Fig. 1). To determine whether miR-615 is present in other amniotes, we searched for its precursor gene sequence in the genomes of other eutherians, plus a marsupial (*Monodelphis domestica*), a monotreme (*Ornithorhynchus anatinus*) and three non-mammalian vertebrates (chicken *Gallus gallus*, anole lizard *Anolis carolinensis* and zebrafish *Danio rerio*). The non-eutherian species did not produce definitive matches, while a positive control BLASTn run using the same parameters successfully recovered *mir-615* in mouse (*Mus musculus*). Within eutherian mammals,

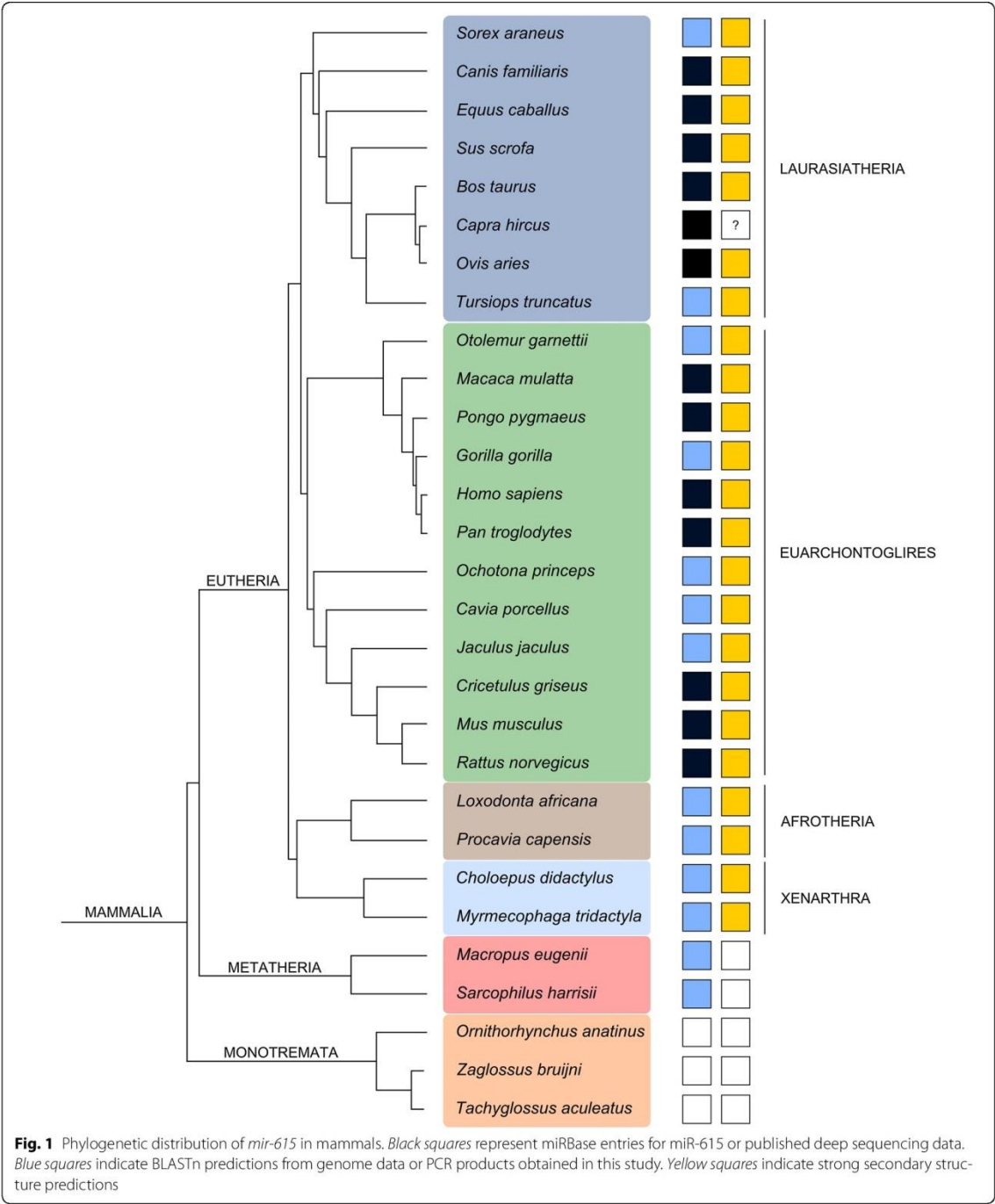
we were able to identify the *mir-615* gene by BLASTn in the elephant (*Loxodonta africana*; Afrotheria), but not in the sloth (*Choloepus hoffmanni*; Xenarthra).

The syntenic position of *mir-615* within a highly conserved Hox gene allows for direct comparisons of the corresponding region to be made between vertebrate species. In human and mouse, *mir-615* is located within the intron of *Hoxc5*. We performed alignments of this intron and its flanking coding exons across the four main eutherian lineages (Laurasiatheria, Euarchontoglires, Afrotheria, and Xenarthra), non-eutherian mammals and other tetrapods to examine sequence conservation across the syntenic position of *mir-615* (Fig. 2, Additional file 3: Supplement S3A). As no sequence data for *Hoxc5* were available for Xenarthra, we sequenced the corresponding region from sloth (*Choloepus didactylus*) and anteater (*Myrmecophaga tridactyla*) using primers designed against highly conserved regions within vertebrate *Hoxc5* coding regions (Additional file 1: Supplement S1). Sequence alignments reveal strong similarity to other eutherian *mir-615* loci (Fig. 2). In marsupials, both the Tasmanian devil (*Sarcophilus harrisii*) and the tamar wallaby (*Macropus eugenii*) possess a sequence resembling eutherian *mir-615* within *Hoxc5*. However, the regions corresponding to the mature miR-615 products differ slightly from the corresponding eutherian versions. We were unable to amplify the *Hoxc5* gene from any monotremes with the same primers. This genomic region is also missing in the platypus (*Ornithorhynchus anatinus*) genome assembly.

These alignment data suggest that the *mir-615* region is strongly conserved at the nucleotide sequence level within the Eutheria, and to a lesser extent in Metatheria, and confirm the lack of an orthologous sequence in monotremes and non-mammalian tetrapods (Fig. 1).

Candidate miR-615 precursors fulfil structural and energetic criteria only in Eutheria

Having established that all therian mammals (eutherians and marsupials) studied possess a sequence similar to the annotated *mir-615* genes of human and mouse (Figs. 1, 2), we analysed these sequences for fulfilment of published structural and energetic criteria used in miRNA annotation. This analysis reveals that all the eutherian *mir-615* orthologues are likely to produce functional pre-miR-615. Analysis of aligned eutherian pre-miR-615 structures using RNAz is strongly supportive of the existence of a conserved fold (Additional file 3: Supplement S3B). However, in the two marsupials analysed minimum folding energy (MFE) for the putative pre-miR-615 homologues (-39.40 kcal/mol in *M. eugenii* and -42.60 kcal/mol in *S. harrisii*) are higher than those observed for all eutherian miR-615 precursors (mean MFE -59.40 kcal/



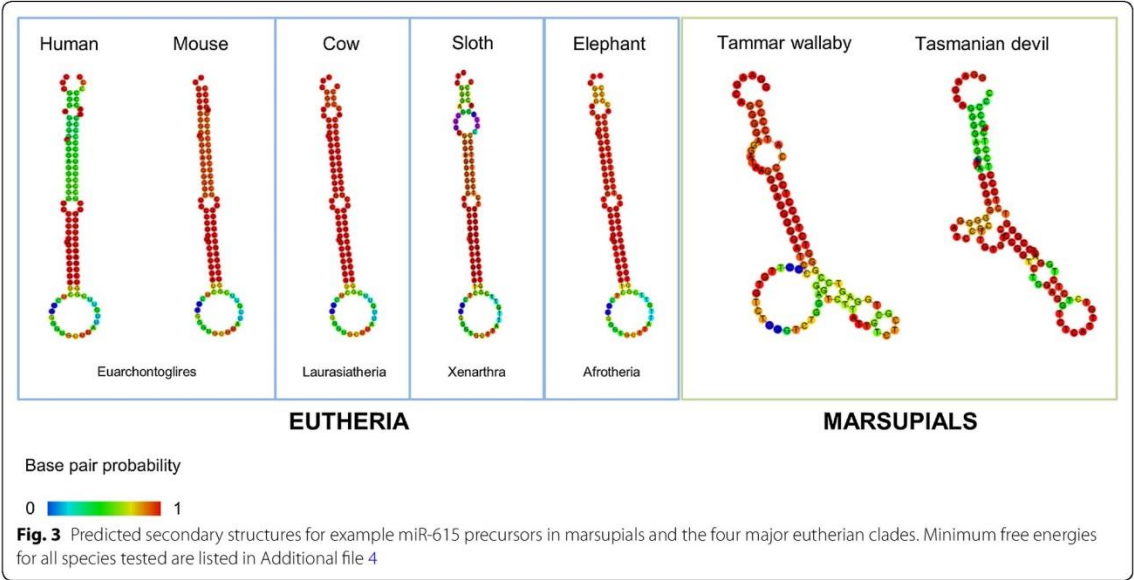
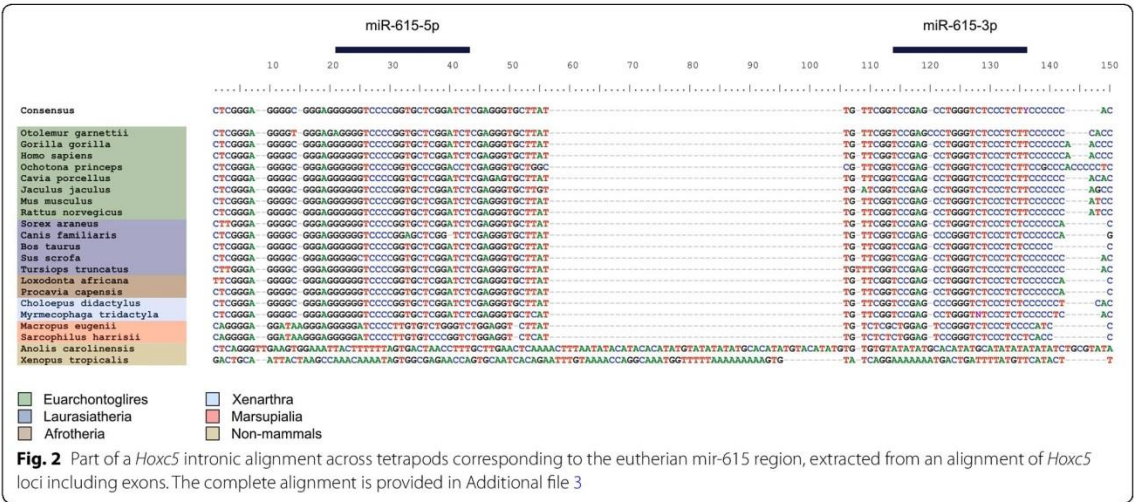
mol). Both marsupial sequences also have large internal loops, bulges and mismatches which are likely to preclude efficient entry into the miRNA pathway (Fig. 3). These predicted structures also differ between *M. eugenii* and *S. harrisii*, suggesting a lack of selection on the secondary structure of this region in marsupials. It is therefore likely that marsupials lack the ability to generate mature miR-615.

miR-615 expression in mammalian cell lines and adult tissues

Using small RNA deep sequencing data from the ENCODE project [26] obtained through the UCSC Genome Browser [17], we analysed miR-615 expression

in 36 human cell lines and tissues (Additional file 5: Supplement S5). miR-615 is expressed in the majority of these datasets. A minority of cell lines lacked expression of miR-615, notably those with immune-related functions (CD14+, CD34+ and CD20+) as well as H1 embryonic stem cells, H1-neurons and HAoEC. There is no evidence for antisense transcription of miR-615 in any of the samples. In all cases where miR-615 is expressed, the miR-615-3p product is dominant.

Since the above browser included only ENCODE data released prior to 2013, we extended these analyses using available raw sequence reads from more recent small RNA deep sequencing projects available through the ENCODE Project Portal using the miRDeep2 quantifier



module. We first confirmed that this approach gives comparable results to the UCSC browser information by analysing the raw sequence reads from the human K562 and GM12878 cell lines (Additional file 6: Supplement S6A). Using other deposited raw sequencing data sets, we then analysed the expression of miR-615 in five organs (cerebellum, cerebral cortex, heart, testis and kidney) in human and mouse [1], as well as ovary in goats [27] and sheep [28]. These data indicate that miR-615 is expressed in kidneys, cerebellum, testis and ovary (Additional file 6: Supplement S6B). Although miR-615 was not detected in whole mouse testis, it is present in the Sertoli cells and spermatocytes (see [1], Additional file 6: Supplement S6B). In the adult goat ovary, expression of both mature products of *mir-615* is increased during pregnancy (see [27], Additional file 6: Supplement S6C). Our analysis of existing small RNA sequencing data in the mouse (obtained from [21]) reveals that miR-615 expression is detected in whole mouse embryos (Additional file 7: Supplement S7). Expression of the miRNA is variable between replicates at each developmental stage, but displays a general trend of increase through embryogenesis and persistence in the newborn.

Expression of miR-615 can occur independently from *Hoxc5*

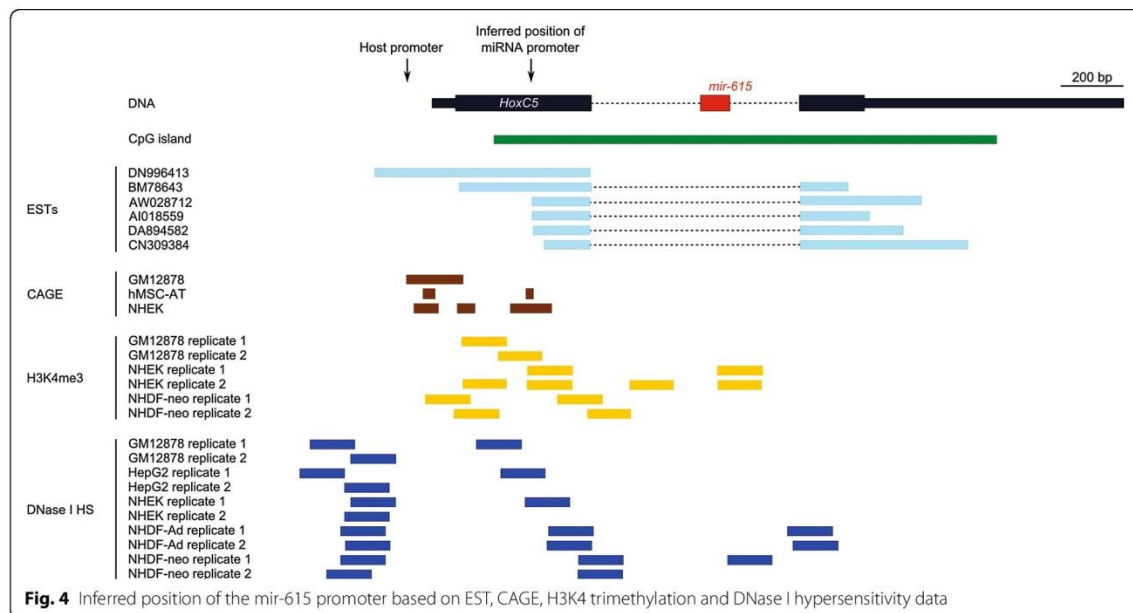
Comparing the expression patterns of *mir-615* and those of *Hoxc5* would help test whether expression of *mir-615* is solely driven by its host transcription unit, or whether there is independent regulation of this intronic miRNA. Out of the 36 cell lines and tissue types used to survey miR-615 expression, 34 also had available RNA-seq data (Additional file 5: Supplement S5), which we analysed to study expression of *HOXC5*. We found evidence for simultaneous expression of both miR-615 and *HOXC5* in several datasets, such as the adipose-derived hMSC-AT mesenchymal stem cell line and NHDF human dermal fibroblasts. Some cell lines, such as H1 embryonic stem cells, express neither the miRNA nor its host gene. However, it is notable that in many of the cell lines where miR-615 is found, including the lymphoblastoid cell line GM12878 and the leukaemia-derived K562 cell line, there is little to no expression of *HOXC5*. To validate the observation that *HOXC5* is not expressed in K562 cells, we constructed a de novo transcriptome assembly for this cell line using raw data from the ENCODE Consortium [29] and again did not find evidence for *HOXC5* expression (Additional file 8: Supplement S8). There are no instances of *HOXC5* expression without concurrent expression of miR-615 in all the cell lines surveyed. Taken together, these data suggest that miR-615 is generated both from transcription of *HOXC5* and from its own promoter.

Analysis of the reported chromatin landscape surrounding *mir-615* in ENCODE human cell lines provides evidence for the existence and location of a miRNA-specific promoter (Fig. 4). Trimethylation of histone H3 on lysine 4 (H3K4me3), a marker of active chromatin associated with transcriptional start sites (TSSs) is distributed throughout the first exon and the intron of *HOXC5* in cell lines that express *mir-615*. Cell lines which express neither *HOXC5* nor *mir-615*, represented by H1-hESC and AG044540, have minimal H3K4 trimethylation at the locus. DNase I hypersensitivity data also provides a measure of chromatin accessibility and is concentrated in two regions, one ~150 bp upstream of the annotated TSS for *HOXC5*, and the other distributed more broadly across the first exon and the 5' end of the intron. A third region of DNase I hypersensitivity is also seen at the 3' end of the intron, but only in NHDF cells.

The strongest indication of an independent promoter for *mir-615* comes from 5' cap analysis of gene expression (CAGE) data, which identify potential initiation sites within the *HOXC5* locus through the detection of 5' capped RNA transcripts. These sites map to both the known *HOXC5* TSS and a region about ~200 bp upstream of the exon–intron boundary, corresponding well with the 5' ends of several expressed sequence tags (ESTs). This position also aligns with the DNase I hypersensitivity sites. Collectively, these analyses of chromatin accessibility and transcription at the *HOXC5* locus, coupled with evidence for expression of *mir-615* in the absence of *HOXC5*, strongly support of the existence of an internal miRNA-specific promoter, in addition to the *HOXC5* gene promoter.

Candidate targets for miR-615-3p

Identifying likely targets for miR-615, integrated with the expression pattern data presented above, would contribute to our understanding of how this miRNA contributes to eutherian biology and evolution. Dominance of the 3p arm is observed in all cases where we have observed expression of *mir-615*, hence we undertook target prediction analysis for miR-615-3p against all known human cDNAs (Additional file 9: Supplement S9). This suggested that potential targets include transcripts encoding proteins with roles in development and function of the nervous system (e.g. KIF1A and PAX6), transcriptional regulation (e.g. YY1AP1), interaction with Hox proteins (PBX3) and cell cycle control (e.g. RSL1D1). For each target above the cut-off, we ascertained expression levels in tissues known to express miR-615 (kidney, testis, ovary and cerebellum). Some targets, such as *AES*, are co-expressed with *mir-615* in all tissues examined. Others are expressed with the miRNA in some tissues but not in others; *PAX6*, for example, is expressed in the cerebellum



but not in the kidney or gonads. GO term distribution was compared between the 72 predicted targets and the input set of human cDNAs (Additional file 10: Supplement S10). GO terms upregulated in the target prediction dataset relative to all human cDNAs include those associated with growth and development, as well as those involved with signalling pathways (e.g. Notch) known to regulate development [30].

Discussion

Origins of eutherian *mir-615*

Sequence alignment and structural prediction suggests that *mir-615* generates functional mature products in all eutherian mammals. A similar sequence exists in the *Hoxc5* intron of marsupials and adopts a very approximate hairpin structure (Figs. 2, 3), but has not been identified as a miRNA gene [31]. This secondary structure, while not a completely random fold, does not conform to the energetically stable stem loop required for efficient entry into the miRNA processing pathway. We speculate that it is perhaps occasionally captured by the miRNA processing machinery, but is unlikely to be processed as efficiently as the eutherian precursor.

The presence of functional *mir-615* in eutherians and its absence from marsupials could be explained by either the acquisition of *mir-615* on the lineage leading to the Eutheria following its divergence from marsupials, or through acquisition of the miRNA in a common therian ancestor followed by its lineage-specific loss in

marsupials. While an approximation of the ancestral therian condition might be inferred through examination of the corresponding region in monotremes, we were unable to recover any convincing hits to *Hoxc5* by BLASTn against the *O. anatinus* (platypus) genome assembly, or amplify its intron through PCR in three monotreme species (platypus *O. anatinus*, short-beaked echidna *Tachyglossus aculeatus*, long-beaked echidna *Zaglossus bruijnii*). It is plausible that monotremes have undergone lineage-specific loss of part of the *Hoxc* cluster; posterior *Hoxc* genes (*Hoxc10*–*Hoxc13*) are present in the platypus draft genome while *Hoxc4*, *Hoxc5*, *Hoxc6*, *Hoxc8* and *Hoxc9* are absent [32]. A hypothetical monotreme homolog of *mir-615* could therefore either have been lost with the *Hoxc5* gene, and we did not detect in any other genomic region. Given that that none of the 17 eutherian species studied across the four major eutherian lineages have lost this miRNA or significantly altered either of its mature products, we argue that *mir-615* is more likely to be a eutherian novelty rather than the alternative scenario where it arose on the common mammalian lineage and was lost in monotremes and marsupials.

mir-615 as a model for understanding intronic miRNA evolution

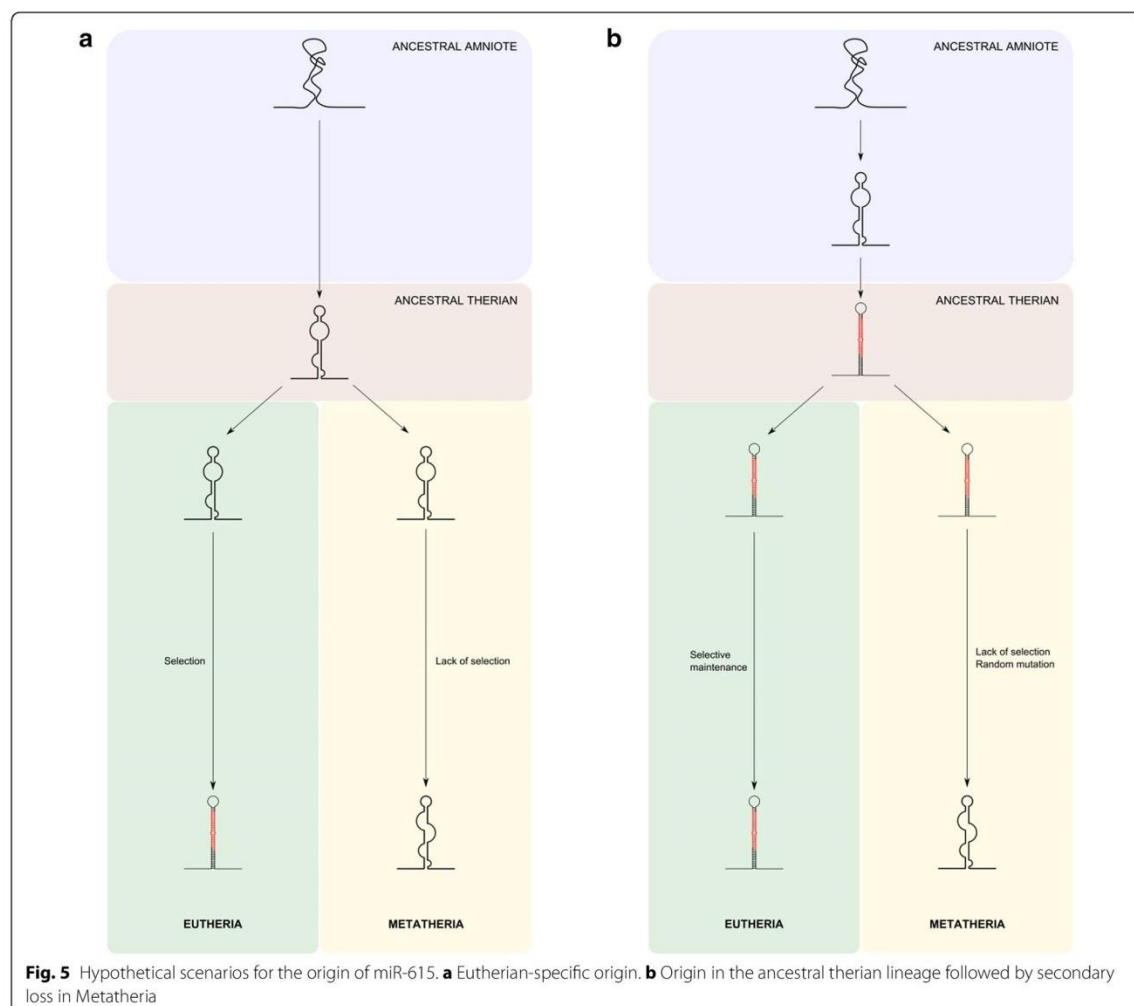
Intronic exaptation has been proposed as a major contributor to miRNA innovation [1, 2]. Under this model, an intronic origin for a novel miRNA eliminates the need to evolve a separate promoter driving miRNA

transcription, at least initially. This reduces the requirement for novel miRNA emergence to the acquisition of a suitable secondary structure by the existing transcript.

A possible model for the origin of miR-615, involving eutherian-specific origin via intronic exaptation, is shown in Fig. 5a. In this scenario, the ancestral sequence was not processed efficiently as a miRNA or stably integrated into any genetic regulatory pathways. Following the divergence of the Eutheria from the Metatheria, different evolutionary forces shaped the evolution of the intron in each lineage. Modification of the eutherian sequence, perhaps initially through drift and subsequently through positive selection on adaptive miRNA-target interactions, resulted in the emergence of a precise stem-loop structure which could efficiently enter the miRNA

processing pathway [33]. The lack of either the stabilising mutations, or the lack of suitable selection pressure after the emergence of such a structure, would explain the absence of functional miR-615 in marsupials. The observed sequence-level similarity between the corresponding *Hoxc5* intronic regions in marsupials and eutherians, with a precise stem-loop structure present in eutherians but not marsupials, presents a strong case for the evolution of miR-615 through intronic exaptation.

Although we favour the above model as the most likely mode of origin of miR-615, we note that the predicted secondary structure of the homologous region in the *Hoxc5* intron of modern marsupials resembles an approximate hairpin more than a completely random fold (Fig. 3). This means we should at least consider the



alternative possibility of secondary loss or degradation in the Metatheria (Fig. 5b). Terminal branches of phylogenies tend to be rich in young miRNAs, most of which undergo rapid turnover and will eventually be lost [1, 14, 34]. It is therefore not implausible that miR-615, newly emergent in the ancestral therian lineage, was captured by selection and retained in only in the eutherian lineage but not in Metatheria. Under this model, a young *mir-615* gene would not have been stably integrated into any genetic regulatory networks at the time of the eutherian-metatherian split. Positive selection on any beneficial miRNA-target interactions and drift represent two counteracting forces governing the fate of new miRNA genes in each lineage. In the Eutheria, the effect of positive selection in maintaining *mir-615* in the genome outweighed the degradation of the secondary structure through drift, while the opposite would have been true for the Metatheria. Such a model would push back the origin of *mir-615* to the common therian lineage.

Despite the finding that *Hoxc5* mRNA and *miR-615* are produced together, we also find evidence for regulation of *mir-615* independently from its host gene in modern eutherians. We thus argue there are two modes of miR-615 production. RNA-seq data indicate that mature products of *mir-615* are detected in many human cell lines and tissues with no *HOXC5* expression. This independent transcription of *mir-615* most likely occurs through the use of separate transcriptional control sequences which do not result in production of a functional host transcript (Fig. 4). It has been reported that most miRNA promoters are located within 500 bp upstream of the pre-miRNA sequence [35], suggesting the putative miR-615 specific promoter lies close to the start of the intron or in the first coding exon of *HOXC5*. The region upstream of *mir-615*, within the limits of the large CpG island spanning most of the *HOXC5* locus (Fig. 4), undergoes DNA methylation in pancreatic cancer cell lines and is reported as including a promoter for *mir-615* [36] although no precise promoter elements were identified. Histone methylation, DNase I hypersensitivity and CAGE tag data, combined with analysis of ESTs, points to a probable transcription initiation site within the first exon, approximately 550 bp upstream of *mir-615* (Fig. 4). As the majority of the *HOXC5* locus including the candidate miRNA promoter lies within the CpG island, it is likely that transcription is driven over a broad range of TSSs rather than a precise initiation site [37, 38].

We hypothesise that the acquisition of the CpG island in eutherian *Hoxc5* could have facilitated the evolution of independent promoter sequences for *mir-615*. The existence of TSSs located within exons has been noted [39], and is hypothesised as a significant contributor to truncated non-coding mRNAs [40]. Internal initiation leading

to the production of such transcripts may have enabled *mir-615* to be expressed independently of its host gene. While this would have occurred initially by chance, it would afford subsequent selection the opportunity to increase expression levels, perhaps in a tissue-specific manner. CpG island-associated promoters are known to evolve rapidly [37], which would promote the evolution of *mir-615* regulation. Interestingly, the CpG island covering the putative TSS for *mir-615* along with much of the *Hoxc5* locus in eutherians seems to be absent in marsupials. In the Tasmanian devil (*Sarcophilus harrisii*), the corresponding region is not annotated as a CpG island, and the GC content of the *Hoxc5* intron is lower in marsupials than in eutherians (52 % in Tasmanian devil and tammar wallaby compared to 60 % in human, 60 % in mouse and 62 % in cow). It is more likely that the CpG island was gained in eutherians, rather than present in a common ancestor and lost in marsupials [41].

Putative functions for miR-615

Several studies have explored the roles of miR-615 in the context of human pathologies [36, 42–45]. The miRNA is known to be highly expressed in the prostate cancer cell line LNCaP relative to normal prostate epithelial cells [44], as well as in cirrhotic and cancerous liver tissues and cell lines but not healthy liver [43]. miR-615-3p is also one of five Hox cluster miRNAs significantly upregulated in the prefrontal cortex in patients with Huntington's Disease [45]. It is likely that in these contexts, miR-615-5p is acting as a tumour suppressor [36, 43] through interfering with the *insulin-like growth factor 2 (IGF2)* transcript. However, the involvement of miR-615 in non-pathologic biological processes is currently unknown.

The conservation of *mir-615* across the Eutheria is indicative of biological function. Our analysis of data from Chiang et al. indicates that whole-body expression of miR-615 increases throughout mouse embryonic development and persists in the newborn [21]. Although expression levels across replicates in this study were somewhat variable, the greatest increase in expression occurs between E9.5 and E12.5. miR-615 is also amongst the 10 miRNAs which are most upregulated upon differentiation of human embryonic stem cells [46]. Known tissues in which miR-615 expression is conserved in more than one eutherian species include the cerebellum, kidney and gonads. While the observation that miR-615 is strongly upregulated in goat ovaries during pregnancy [27] is striking, these data come from a single biological replicate and should be interpreted with caution.

We also find that GO terms associated with regulation of growth and development are overrepresented by in a high stringency target prediction dataset for miR-615-3p. This observation, coupled with the high

expression of miR-615-3p in the mouse embryo, suggests a potential role for miR-615-3p in the regulation of growth and development during embryogenesis. Although the expression of *Hoxc5* during axial patterning is likely to result in production of miR-615-3p and may account for some of the observed expression, axial Hox gene expression is initiated prior to E9.5 [47] and is unlikely to explain the upregulation of the miRNA in between E9.5 and E12.5. We hypothesize that the observed increase in expression results from either upregulation through the host gene promoter elsewhere in the developing embryo, or through the miRNA-specific promoter. The observation that ovaries from newborn mice express miR-615-3p at a relatively high level [48] suggests that some of this may result from ovarian expression of miR-615. Other major developmental events occurring between E9.5 and E12.5 include organogenesis of the kidney [49, 50], sex-specific differentiation of the gonads [51] and outgrowth of the limb primordia [52]. A study on the embryonic gonads of the sheep (*Ovis aries*) lends some weight to the hypothesis that miR-615 may be involved in gonadal differentiation [28], since the miRNA is identified as expressed at significantly higher levels (fold change ≥ 2) in ovaries at early gestation (42 days) than in mid-gestation. This difference is not observed in the foetal testis. However, caution should be applied in interpretation of data across different species. Another possibility is that miR-615 is involved in the modulation of embryonic growth; studies on human cancer cell lines have implicated this miRNA in the regulation of cell growth, proliferation and migration [36, 43].

Conclusions

The Hox cluster miRNA *mir-615* represents an interesting case of acquisition of a novel miRNA by intronic exaptation and an unusual example of a new functional gene arising within a conserved Hox gene cluster. We demonstrate using sequence analysis and structural prediction that miR-615 is a eutherian-specific miRNA, although a similar sequence exists in extant marsupials that may have a low probability of miRNA processing. Using a data mining approach combining epigenetic studies with RNAseq and CAGE tag data, we show that miR-615 can be expressed under the control of a Hox gene promoter, as well as independently from its host gene transcript. Based on these observations, we propose a model for intronic miRNA evolution through selection on intronic secondary structure followed by evolution of an independent miRNA promoter. Target prediction data support the involvement of miR-615 in growth and developmental pathways.

Additional files

Additional file 1: Supplement S1. Position of forward and reverse primers used to amplify the *Hoxc5* intronic region in *Xenarthra*. Primers were designed against highly conserved regions within the *Hoxc5* coding sequence. Corresponding sequences for each vertebrate species are shown.

Additional file 2: Supplement S2. Datasets used in analysis of *mir-615* and *Hoxc5* expression.

Additional file 3: Supplement S3. (A) Alignment of tetrapod *Hoxc5* loci including exonic regions. (B) RNAz analysis for eutherian pre-miR-615.

Additional file 4: Supplement S5. Predicted secondary structures and minimum free energies for the region corresponding to *mir-615* across therian mammals.

Additional file 5: Supplement S5. Cell lines available through the ENCODE Consortium (2003–2012) for which small RNA sequencing data are available.

Additional file 6: Supplement S6. (A) Read counts for miR-615 mature products in the GM12878 and K562 cell lines, as generated by the miRD-eep2 quantifier.pl script. Raw small RNA sequencing data were obtained from the ENCODE Consortium. (B) Data from Zhang et al. (2013) showing normalized tissue-specific read counts for the dominant product of miR-615 (miR-615-3p). Sertoli cells, spermatocytes and spermatids were only sampled in mouse [27]. (C) Read counts for both mature products of miR-615 in pooled ovarian sample from six pregnant and six non-pregnant goats. The dataset used for this analysis was obtained from Zhang et al. (2013).

Additional file 7: Supplement S7. miR-615-3p read counts in mouse small RNA libraries as reported in the dataset of Chiang et al. (2010), refer to [21]. Data were obtained through miRBase.

Additional file 8: Supplement S8. RPKM values for *HOXC* genes in a K562 transcriptome assembly. Raw reads were obtained through the ENCODE Consortium.

Additional file 9: Supplement 9. PITA prediction targets for miR-615-3p.

Additional file 10: Supplement 10. GO terms identified in the miR-615-3p target prediction dataset. The normalised frequency of each term is plotted relative to its frequency in a set of all human cDNAs. The dotted line indicates equal frequency of expression in both datasets.

Abbreviations

BLAST: basic local alignment search tool; BLASTn: nucleotide BLAST; Bp: base pair; cDNA: complementary DNA; CITES: convention on international trade in endangered species of wild fauna and flora; ENCODE: encyclopedia of DNA elements; EST: expressed sequence tag; FPKM: fragments per kilobase of transcript per million mapped reads; GEO: gene expression omnibus; miRNA: microRNA; nt: Nucleotide; PCR: polymerase chain reaction; PITA: probability of interaction by target accessibility; ssRNA: single-stranded RNA; TSS: transcription start site; UTR: untranslated region.

Authors' contributions

SQ performed experimental and computational analysis, prepared the figures, and drafted the manuscript. SQ and PWHH conceived and designed the study. Both authors read and approved the final manuscript.

Acknowledgements

We thank Stephen Donnellan and Mads Bertelsen for provision of tissue samples; Shankar Srinivas, Antonio Miranda, Jerome Hui, Jordi Paps, Ferdinand Marlétaz, Nathalie Feiner, Ignacio Maeso, Thomas Dunwell and Sebastian Shimeld for discussions. SQ acknowledges support from the A*STAR National Science Scholarship (PhD). PWHH acknowledges support from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)/ERC grant [268513].

Competing interests

The authors declare that they have no competing interests.

Received: 31 August 2015 Accepted: 25 September 2015

Published online: 07 October 2015

References

- Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H, Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. Birth and expression evolution of mammalian microRNA genes. *Genome Res*. 2013;23:34–45.
- Campo-Paysaa F, Sémon M, Cameron RA, Peterson KJ, Schubert M. microRNA complements in deuterostomes origin and evolution of microRNAs. *Evol Dev*. 2011;13:15–27.
- Holland PWH. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol*. 2012;2:31–45.
- Tabin CJ. Why we have (only) five fingers per hand. Hox genes and the evolution of paired limbs. *Development*. 1992;116:289–96.
- Taguchi A, Kaku Y, Ohmori T, Sharmin S, Ogawa M, Sasaki H, Nishinakamura R. Redefining the in vivo origin of metanephric nephron progenitors enables generation of complex kidney structures from pluripotent stem cells. *Cell Stem Cell*. 2014;14:53–67.
- Patterson LT, Pembaur M, Potter SS. Hoxa11 and Hoxd11 regulate branching morphogenesis of the ureteric bud in the developing kidney. *Development*. 2001;128:153–61.
- Taylor HS, Vanden Heuvel GB, Igarashi P. A conserved Hox axis in the mouse and human female reproductive system. Late establishment and persistent adult expression of the Hoxa cluster genes. *Biol Reprod*. 1997;57:1338–45.
- Tanzer A, Amemiya CT, Kim C-B, Stadler PF. Evolution of microRNAs located within Hox gene clusters. *J Exp Zool B Mol Dev Evol*. 2005;304:75–85.
- Candiani S. Focus on miRNAs evolution: a perspective from amphioxus. *Brief Funct Genomics*. 2012;11:107–17.
- Hui JHL, Marco A, Hunt S, Melling J, Griffiths-Jones S, Ronshaugen M. Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. *Nucleic Acids Res*. 2013;41:3352–61.
- Woltering JM, Durston AJ. MiR-10 Represses HoxB1a and HoxB3a in Zebrafish. *PLoS One*. 2008;3:e1396.
- Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*. 2005;11:241–7.
- Agranat-Tamir L, Shomron N, Sperling J, Sperling R. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. *Nucleic Acids Res*. 2014;42:4640–51.
- Quah S, Hui JHL, Holland PWH. A burst of miRNA innovation in the Early Evolution of Butterflies and Moths. *Mol Biol Evol*. 2015;32:1161–74.
- Liang Y, Ridzon D, Wong L, Chen C. Characterization of microRNA expression profiles in normal human tissues. *BMC Genom*. 2007;8:166.
- Mi S, Lu J, Sun M, Li Z, Zhang H, Neilly MB, Wang Y, Qian Z, Jin J, Zhang Y, Bohlander SK, Le Beau MM, Larson RA, Golub TR, Rowley JD, Chen J. MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia. *Proc Natl Acad Sci USA*. 2007;104:19971–6.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler AD. The Human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.
- Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, et al. Ensembl 2012. *Nucleic Acids Res*. 2012;42(D1):D749–55.
- Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059–66.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva ANCBIGEO. Archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;40(D1):D84–90.
- Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, Blelloch R, Schroth GP, Nusbaum C, Bartel DP. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*. 2010;24:992–1009.
- Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2011;39:D152–7.
- Kertész M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nat Genet*. 2007;39:1278–84.
- The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2014;43:D204–12.
- Uhlen M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szgyarto CA-K, Odeberg J, Djureinovic D, Tanakan JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, et al. Tissue-based map of the human proteome. *Science*. 2015;347:1260419.
- Katalin F-T, Soirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, Gingeras TR. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*. 2009;457:1028–32.
- Zhang X-D, Zhang Y-H, Ling Y-H, Liu Y, Cao H-G, Yin Z-J, Ding J-P, Zhang X-R. Characterization and differential expression of microRNAs in the ovaries of pregnant and non-pregnant goats (*Capra hircus*). *BMC Genom*. 2013;14:157.
- Torley KJ, da Silveira JC, Smith P, Anthony RV, Veeramachaneni DNR, Winger QA, Bouma GJ. Expression of miRNAs in ovine fetal gonads: potential role in gonadal differentiation. *Reprod Biol Endocrinol*. 2011;9:2.
- Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Fritze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shores N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
- Artavanis-Tsakonas S. Notch signaling. Cell fate control and signal integration in development. *Science*. 1999;284:770–6.
- Yu H, Lindsay J, Feng Z-P, Frankenberg S, Hu Y, Carone D, Shaw G, Pask AJ, O'Neill R, Papenfuss AT, Renfree MB. Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC Genom*. 2012;13:251.
- Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang S-P, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, López-Otin C, Ordóñez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*. 2008;453:175–83.
- Axtell MJ, Westholm JO, Lai EC. Vive la différence. biogenesis and evolution of microRNAs in plants and animals. *Genome Biol*. 2011;12:221.
- Lyu Y, Shen Y, Li H, Chen Y, Guo L, Zhao Y, Hungate E, Shi S, Wu C-I, Tang T. New microRNAs in *Drosophila*—birth, death and cycles of adaptive evolution. *PLoS Genet*. 2014;10:e1004096.
- Zhou X, Ruan J, Wang G, Zhang W. Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol*. 2007;3(3):e37.
- Gao W, Gu Y, Li Z, Cai H, Peng Q, Tu M, Kondo Y, Shinjo K, Zhu Y, Zhang J, Sekido Y, Han B, Qian Z, Miao Y. miR-615-5p is epigenetically inactivated and functions as a tumor suppressor in pancreatic ductal adenocarcinoma. *Oncogene*. 34(13):1629–40.
- Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*. 2006;38:626–35.
- Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. *Genome Res*. 2008;18:1–12.

39. Brodsky AS, Meyer CA, Swinburne IA, Hall G, Keenan BJ, Liu XS, Fox EA, Silver PA. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* 2005;6:R64.
40. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet.* 2007;8:424–36.
41. Suzuki S, Shaw G, Kaneko-Ishino T, Ishino F, Renfree MB. The evolution of mammalian genomic imprinting was accompanied by the acquisition of novel CpG islands. *Genome Biol Evol.* 2011;3:1276–83.
42. Backes C, Meese E, Lenhof HP, Keller A. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.* 2010;38:4476–86.
43. El Tayebi HM, Hosny KA, Esmat G, Breuhahn K, Abdelaziz AI. miR-615-5p is restrictedly expressed in cirrhotic and cancerous liver tissues and its over-expression alleviates the tumorigenic effects in hepatocellular carcinoma. *FEBS Lett.* 2012;586:3309–16.
44. Hult T, Sibbritt T, Wiklund ED, Bert S, Strbenac D, Statham AL, Robinson MD, Clark SJ. Discovery pipeline for epigenetically deregulated miRNAs in cancer: integration of primary miRNA transcription. *BMC Genom.* 2011;12:54.
45. Hoss AG, Kartha VK, Dong X, Latourelle JC, Dumitriu A, Hadzi TC, MacDonald ME, Gusella JF, Akbarian S, Chen J-F, Weng Z, Myers RH. MicroRNAs located in the Hox gene clusters are implicated in Huntington's disease pathogenesis. *PLoS Genet.* 2014;10:e1004188.
46. Bar M, Wyman SK, Fritz BR, Qi J, Garg KS, Parkin RK, Kroh EM, Bendoraite A, Mitchell PS, Nelson AM, Ruzzo WL, Ware C, Radich JP, Gentleman R, Ruohola-Baker H, Tewari M. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells.* 2008;26:2496–505.
47. Deschamps J, van Nes J. Developmental regulation of the Hox genes during axial morphogenesis in the mouse. *Development.* 2005;132:2931–42.
48. Ahn HW, Morin RD, Zhao H, Harris RA, Coarfa C, Chen ZJ, Milosavljevic A, Marra MA, Rajkovic A. MicroRNA transcriptome in the newborn mouse ovaries determined by massive parallel sequencing. *Mol Hum Reprod.* 2010;16:463–71.
49. Vainio S, Lin Y. Coordinating early kidney development: lessons from gene targeting. *Nat Rev Genet.* 2002;3:533–43.
50. Wellik DM. Hox genes and kidney development. *Pediatr Nephrol.* 2011;26:1559–65.
51. Ross AJ, Capel B. Signaling at the crossroads of gonad development. *Trends Endocrinol Metab.* 2005;16(1):19–25.
52. Wanek N, Muneoka K, Holler-Dinsmore G, Burton R, Bryant SV. A staging system for mouse limb development. *J Exp Zool.* 1989;249:41–9.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



RESEARCH ARTICLE

A Diversity of Conserved and Novel Ovarian MicroRNAs in the Speckled Wood (*Pararge aegeria*)

Shan Quah¹, Casper J. Breuker², Peter W. H. Holland^{1*}

1 Department of Zoology, University of Oxford, South Parks Road, Oxford, United Kingdom, **2** Evolutionary Developmental Biology Research Group, Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Gypsy Lane, Headington, Oxford, United Kingdom

* peter.holland@zoo.ox.ac.uk



OPEN ACCESS

Citation: Quah S, Breuker CJ, Holland PWH (2015) A Diversity of Conserved and Novel Ovarian MicroRNAs in the Speckled Wood (*Pararge aegeria*). PLoS ONE 10(11): e0142243. doi:10.1371/journal.pone.0142243

Editor: Andreas Hejnol, Sars International Centre for Marine Molecular Biology, NORWAY

Received: August 19, 2015

Accepted: October 18, 2015

Published: November 10, 2015

Copyright: © 2015 Quah et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data files for miRNA and total RNA sequencing, plus the combined ovarian transcriptome assembly, are available from the Gene Expression Omnibus database (accession number GSE72090).

Funding: This study was funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant 268513 to PWHH and an A* STAR National Science PhD Scholarship to SQ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

microRNAs (miRNAs) are important regulators of animal development and other processes, and impart robustness to living systems through post-transcriptional regulation of specific mRNA transcripts. It is postulated that newly emergent miRNAs are generally expressed at low levels and with spatiotemporally restricted expression domains, thus minimising effects of spurious targeting on animal transcriptomes. Here we present ovarian miRNA transcriptome data for two geographically distinct populations of the Speckled Wood butterfly (*Pararge aegeria*). A total of 74 miRNAs were identified, including 11 newly discovered and evolutionarily-young miRNAs, bringing the total of miRNA genes known from *P. aegeria* up to 150. We find a positive correlation between miRNA age and expression level. A common set of 55 miRNAs are expressed in both populations. From this set, we identify seven that are consistently either ovary-specific or highly upregulated in ovaries relative to other tissues. This 'ovary set' includes miRNAs with known contributions to ovarian function in other insect species with similar ovaries and mode of oogenesis, including miR-989 and miR-2763, plus new candidates for ovarian function. We also note that conserved miRNAs are overrepresented in the ovary relative to the whole body.

Introduction

MicroRNAs (miRNAs) are short regulatory RNAs with important roles in post-transcriptional regulation [1]. They are of significance at all stages during an animal's life cycle from embryonic development [2–5] to post-embryonic development [6,7] and reproduction [5,8]. Understanding the evolution of miRNAs is important to understanding the evolution of growth, reproduction and development.

It has been postulated that newly emergent animal miRNAs are generally expressed at low levels and with spatiotemporally restricted expression domains [9,10]. For example, in a study of miRNA expression in human and chimpanzee brains, more recently acquired miRNAs show a tendency for weaker expression than older miRNAs [10,11]. In mammals, a

Competing Interests: The authors have declared that no competing interests exist.

comprehensive study across multiple tissues demonstrates a positive correlation between miRNA age and expression levels [12].

The Speckled Wood butterfly (*Pararge aegeria*) is an ideal insect species with which we can address similar questions concerning miRNA evolution due to the availability of genome [13], transcriptome [14] and miRNA sequencing [15] data. Here we focus on miRNAs expressed during oogenesis, which are likely to include regulators of egg production and ovary function. Oogenesis in insects must generate a viable egg capable of initiating and sustaining embryonic development after fertilisation [16]. In addition to sustaining the energetic requirements for embryogenesis [17], the egg must also be pre-patterned with a polarity, which in insects occurs largely in the form of localised maternal determinants that drive zygotic spatiotemporal gene expression patterns essential for establishment of body axes [17]. Much is known about the identity and function of maternally provisioned mRNA molecules in insect eggs, but far less about other factors that influence the activity of such mRNAs, such as miRNAs. Furthermore, most of these studies have been performed in drosophilids [18], mosquitoes [19], cockroaches [20] and beetles [21]. Far less is known about the molecular control of oogenesis in butterflies and moths, despite Lepidoptera being one of the largest insect orders. Characterisation of the butterfly miRNA transcriptome would complement the current understanding of both genetic control involved in ovarian function [14] as well as the conservation of such mechanisms across insects.

Materials and Methods

Sample collection and sequencing

We sequenced the ovarian small RNA and mRNA transcriptomes of two females from each of two *P. aegeria* populations kept as lab stocks: (1) Zonza on Corsica, France, and (2) St Hubert, from the south of Belgium. The St Hubert population is the same population for which phenotypic data has previously been collected on maternal effects and egg production [22] as well as ovarian transcriptomic data [14] and for which a draft genome has been characterised [13]. The Zonza stock was newly established in 2014; the St Hubert stock has been in the lab since 2007 (see Carter et al. [14] for details regarding rearing *P. aegeria* in general and for St Hubert specifically).

Fully eclosed females were allowed to mate, and were placed individually in netted cages (0.5 m³) along with a potted *Poa trivialis* grass plant for oviposition and an artificial flower containing a 10% honey solution. Females (kept at 21±0.3°C, LD 16:8) were sacrificed 6 days after mating, when they reached the peak of egg production [22]. Ovaries were dissected and cleared of fat and surrounding tissue (cf. Carter et al. [14]). Total RNA was extracted using TRIzol (Life Technologies). Total RNA samples and analysed by Illumina HiSeq 2000 sequencing at the Beijing Genomics Institute. Each sample was used to generate both a mRNA transcriptome library as well as a small RNA library. The mRNA transcriptome sequencing used paired-end reads of 2 x 100 nucleotides from a 160 bp insert library.

miRNA identification

Small RNA data processing was carried out as previously described [15], except that instead of annotating known miRNAs by comparison to miRBase, precursors were initially scanned by BLASTn against the set of previously reported *P. aegeria* miRNAs [15]. The dataset used in the previous study [15] was generated by pooling RNA in equal amounts from various life-cycle stages of the Speckled Wood butterfly, thereby eliminating potential bias from any one tissue. Sequences without hits against this dataset were subject to additional processing to remove hits against repetitive elements and non-coding RNAs, as well as precursors with reads mapping

only to one arm and those with <95% of reads on either arm with the same 5' starting position. Only sequences with >50 dominant arm reads in Zonza or >18 dominant arms in St Hubert were retained for annotation. The introduction of a minimal read count is because in this study we are focussing on miRNAs with putative ovarian function, whereas in Quah et al. [15] we were searching for all miRNAs; the minimal read counts chosen for the two samples are equivalent after compensating for relative miRNA sequencing depth. All shortlisted candidates were then compared against miRBase 20 to identify known miRNAs which were not have been found in the original *P. aegeria* miRNA survey [15]. The same cutoff values described above were used to determine if a known miRNA was expressed in both ovarian samples.

Read counts for each mature product, as generated by miRDeep2 [23], were cross-checked by running the quantifier.pl script in the miRDeep2 package, using known *P. aegeria* miRNA precursors for read mapping. Both approaches generated similar read count data. Raw read counts for the Zonza and St Hubert samples, as well as for the original pooled dataset, were normalised against the total number of miRNA reads in each sample.

Transcriptome assembly

To assemble a reference ovarian mRNA transcriptome, raw RNAseq reads were processed at source for removal of low quality reads and adaptor trimming. Read quality was then validated by checking FASTQ data using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and the first 16 bases trimmed off all reads. Reads from Zonza and St Hubert were pooled and assembled using the Trinity assembler (version: trinityrnaseq_r20140717) [24] to generate a combined *P. aegeria* ovarian transcriptome assembly containing 54,861 contigs. Contig FPKM (Fragments Per Kilobase of exon per Million of fragments mapped) values for each sample were determined using RSEM [25]. In order to retain only biologically interpretable contigs, we employed a multi-step approach to annotation. Contigs were first screened by BLASTn against the nr database on NCBI to reduce the number of misassembled contigs in the dataset. Remaining contigs were then compared by BLASTn against the *P. aegeria* ovarian transcriptome reported by Carter et al. [14]. A referenced list of proteins with known roles in ovarian and egg function [14] was also used to screen the combined transcriptome assembly by tBLASTn. All BLAST searches were subject to an E-value cutoff of 1e-05. Only contigs with hits to one or more of these datasets were retained, leaving 17,636 annotated contigs (S1 Table) which were further characterised using Blast2GO [26]. This involved screening contigs by BLASTx followed by mapping and annotation. InterProScan and ANNEX were used to augment annotation of GO terms. GO terms were condensed using the generic GO Slim subset.

Raw data for both miRNA and total RNA sequencing are available via the Gene Expression Omnibus (accession: GSE72090) along with the combined ovarian transcriptome assembly.

Statistical analyses

Total read counts (S2 Table) for miRNAs in each of the datasets were used for statistical analysis. miRNAs were designated 'lineage-specific' if found only in *P. aegeria* and 'conserved' if present in any other species. A two-tailed Mann-Whitney U test was conducted to determine if expression levels differed significantly between conserved and lineage-specific miRNAs in any of the datasets. We also conducted a Spearman's rho analysis to examine correlation between miRNA age and expression level.

miRNA target prediction

miRNA target prediction was carried out using the PITA algorithm [27] with default parameters. Only predictions with a score of -10 or below were accepted as predicted hits.

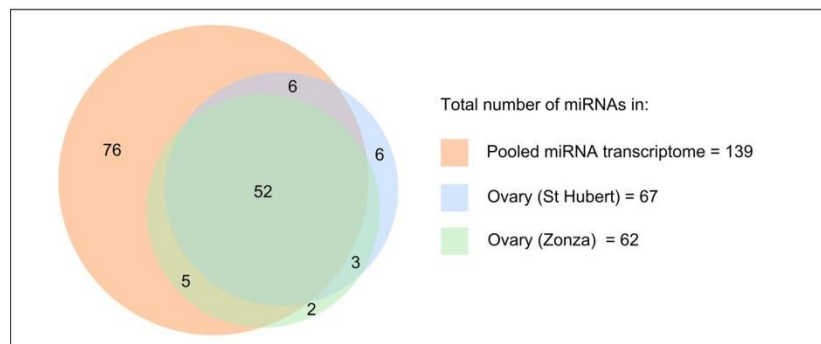


Fig 1. Distribution of miRNAs in each small RNA library.

doi:10.1371/journal.pone.0142243.g001

Results

Small RNA sequencing and data processing

We constructed and sequenced small RNA libraries from ovaries of the Speckled Wood butterfly *P. aegeria* to identify miRNAs expressed in the lepidopteran ovary. To allow for possible differences dependent on geography, climate or diet, we analysed females from two distinct populations from different parts of Europe: St Hubert in Belgium and Zonza in Corsica. Total RNA was extracted from ovaries and used to construct small RNA libraries for each sample. A total of 46,481,949 Illumina HiSeq reads were obtained (22,604, 869 for Corsica and 23,877,080 for Belgium). After filtering reads to remove possible degradation products and non-miRNAs, these were compared to a published dataset of *P. aegeria* miRNAs [15].

Of the 139 previously known miRNAs in *P. aegeria* [15] we obtained reads for 80 and 76 miRNAs in the ovary-specific samples for St Hubert and Zonza respectively (S2 Table). Some of these were expressed at low levels and should not be considered as candidates for ovarian function; using read-count cut-off thresholds, we reduced this number to 58 and 57 in St Hubert and Zonza respectively. Of these, 52 known miRNAs are expressed above cutoff threshold in both populations (Fig 1, S2 Table). In addition, a total of 11 miRNAs that had not been previously described in *P. aegeria* were found; 2 were present in the Zonza sample only, 6 in the St Hubert sample only, and 3 in both (B-0070/C-0072, B-0210/C-0160 and B-0401/C-0173; Table 1). The total number of miRNAs identified as expressed in *P. aegeria* ovary is therefore 74 (63 previously described miRNAs and 11 newly discovered), with a total of 55 miRNAs expressed in the ovary in both populations.

Effect of miRNA age on expression level

We categorised the miRNAs present in the updated *P. aegeria* dataset by age, defining ‘conserved’ miRNAs as those present in *P. aegeria* and at least one other species and ‘lineage-specific’ miRNAs as those not found (or predicted in genome sequence) outside of the Speckled Wood. To test if the 11 newly discovered miRNAs are encoded by lineage-specific genes, or are evolutionarily older, each was investigated by using BLASTn against all known miRNA precursors on miRBase and their inferred pre-miRNA sequences were compared by BLASTn against the available genomes of other Lepidoptera (*Heliconius melpomene*, *Danaus plexippus*, *Manduca sexta*, *Bombyx mori*, *Plutella xylostella* and *Cameraria ohridella*) as well as that of the caddisfly (*Glyptotendipes pellucidus*). One sequence, B-0401/C-0173, was orthologous to miR-2763 (S1 Fig), which likely originated in the common lepidopteran ancestor [15], but was not

Table 1. miRNAs identified in this study which were not found in a whole body pooled transcriptome reported earlier [15].

[illegible]

doi:10.1371/journal.pone.0142243.t001

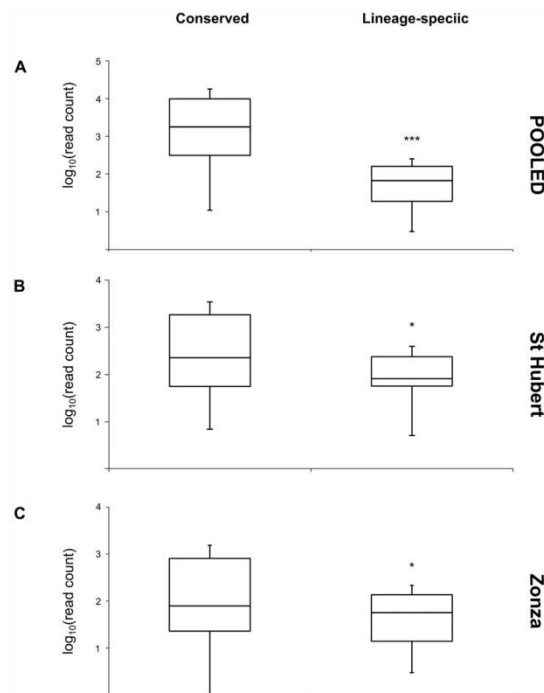


Fig 2. Comparison of miRNA expression levels. Distribution of expression levels for conserved and lineage-specific miRNAs in (A) pooled miRNA transcriptome, (B) St Hubert and (C) Zonza.

doi:10.1371/journal.pone.0142243.g002

previously identified in a whole body small RNA library for *P. aegeria* pooled across multiple developmental stages including adult females [15]; the other 10 sequences are thus far specific to *P. aegeria* and inferred to have arisen more recently. Combining these data with the 139 known *P. aegeria* miRNAs and their times of origin gives a total of 83 conserved and 67 lineage-specific miRNAs in this species.

The total number of reads mapped to each miRNA precursor was taken as a measure of expression level and compared between conserved and lineage-specific miRNAs in St Hubert, Zonza, and the pooled miRNA transcriptome reported in Quah *et al* [15]. There is a large spread of expression levels detected for both conserved and lineage-specific miRNAs in each of the three datasets. A significant difference in median expression levels was observed between conserved and lineage-specific miRNAs (two-tailed Mann-Whitney U test; $U(148) = 802.5$ ($P < 0.001$; pooled sample), $U(148) = 2134$ ($P < 0.05$; St Hubert), $U(148) = 1798$ ($P < 0.05$; Zonza)) in each of the transcriptomes (Fig 2). In addition, the miRNAs with the highest expression levels in each dataset are all members of deeply conserved miRNA gene families, and there is a statistically significant correlation between miRNA age and expression level (Correlation analyses, Spearman's $\rho = 0.582$ ($P < 0.001$; pooled sample), 0.221 ($P < 0.05$; St Hubert), and 0.320 ($P < 0.001$; Zonza)).

Older miRNAs are enriched in the ovary

We studied relative proportions of conserved and lineage-specific miRNAs expressed in the ovary and compared these to the proportions observed in the pooled transcriptome (Fig 3).

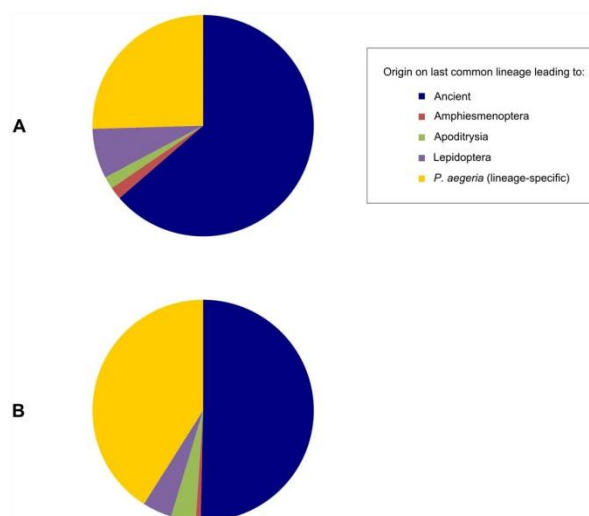


Fig 3. Proportion of miRNAs by age in *P. aegeria* whole body and ovarian transcriptomes. Pie chart showing the proportions of conserved miRNAs originating on each lineage which are (A) present in both St Hubert and Zonza ovarian transcriptomes and (B) present in the original pooled transcriptome reported in Chapter 3. 'Ancient' miRNAs are defined as those originating prior to the emergence of the Amphiesmenoptera (clade consisting of Lepidoptera and Trichoptera).

doi:10.1371/journal.pone.0142243.g003

Ovarian expression of each miRNA was defined by its detection above cut-off levels in both populations. There is a higher proportion of conserved miRNAs (64%) expressed in the ovary compared to that in the pooled miRNA transcriptome (47%, chi-square $P < 0.05$).

Ovarian-specific and ovarian-upregulated miRNAs

The most abundant miRNAs identified in the ovary are (in decreasing order of abundance) miR-989, miR-263a and miR-1. This observation is consistent between St Hubert and Zonza. Other consistently highly-expressed ovarian miRNAs include miR-993 and miR-184 (Table 2). In the pooled whole-body miRNA transcriptome, miR-1 is the most abundantly expressed sequence, followed by miR-989 and let-7; miR-263a and miR-993 are relatively highly expressed in the pooled transcriptome.

Table 2. The ten most abundant miRNAs ranked by expression in the whole body pooled transcriptome and each of the St Hubert and Zonza ovarian transcriptomes.

Rank	Whole body (pooled)	Ovary (St Hubert)	Ovary (Zonza)
1	Pae-miR-1	Pae-miR-989	Pae-miR-989
2	Pae-miR-989	Pae-miR-263a	Pae-miR-263a
3	Pae-let-7	Pae-miR-1	Pae-miR-1
4	Pae-miR-184	Pae-miR-184	Pae-miR-2755
5	Pae-miR-2766	Pae-miR-993	Pae-miR-184
6	Pae-miR-10a	Pae-miR-2755	Pae-miR-993
7	Pae-miR-31	Pae-miR-279d	Pae-let-7
8	Pae-miR-2755	Pae-miR-279b	Par-017
9	Pae-miR-263a	Par-017	Par-013
10	Pae-miR-281	Pae-let-7	Pae-miR-279b

doi:10.1371/journal.pone.0142243.t002

Table 3. The 'ovary set' miRNAs which show greater than 4-fold abundance in both St Hubert (Belgium) and Zonza (Corsica) ovarian transcriptomes relative to the whole body transcriptome. Fold change is calculated from normalized total read. An accurate fold change in abundance cannot be given for the last three miRNAs on the table, as these were detected only in ovarian transcriptomes (both populations) and not at all in the pooled transcriptome.

miRNA	Ovary (St Hubert)				Ovary (Zonza)			
	Total	5'	3'	Fold change	Total	5'	3'	Fold change
Par-247	125	120	5	27.5	146	142	4	11.5
Par-341	52	51	1	10.5	70	69	1	5.0
Par-340	52	51	1	10.3	70	69	1	4.9
Pae-miR-989	288099	32	288067	5.2	873884	88	873796	5.6
Pae-miR-263a	24821	24819	2	4.5	63593	63591	2	4.1
B-0070/C-0072	291	173	118	-	574	130	444	-
B-0210/C-0160	97	9	88	-	262	19	243	-
B-0401/C-0173 (Pae-miR-2763)	27	6	21	-	137	8	129	-

doi:10.1371/journal.pone.0142243.t003

The above analyses reveal which miRNA genes are highly expressed in ovary, but this list includes miRNAs expressed in a range of tissues, as well as ovarian-upregulated sequences. To gain insight into the role of miRNAs in ovarian function, we wished to identify the latter category of miRNAs. Mature and dominant arm read counts for all miRNAs were normalised against the total number of miRNA reads per sample, and then compared with normalised read counts from the pooled miRNA dataset for *P. aegeria* from Quah et al. [15]. miRNAs with dominant arm read counts below a cutoff value in the Belgian and Corsican samples were deemed to not be expressed and excluded from analysis. These analyses identified five miRNAs with > 4-fold upregulation in both libraries (Table 3): Pae-miR-989, Pae-miR-263a, Par-247, Par-340 and Par-341. An additional 13 miRNAs showed up-regulation in just one ovary sample, or were only detected in either ovarian dataset but not in the pooled miRNA transcriptome (Fig 4, S2 Table). As Par-340 and Par-341 share 3' arm reads, we group them together (Par-340/341) for the purpose of analysis. These four 'ovarian-upregulated' miRNA sequences can be combined with the three newly identified 'ovary-specific' miRNAs common to both samples (described above), to form a category of seven 'ovary set miRNAs' for further analysis.

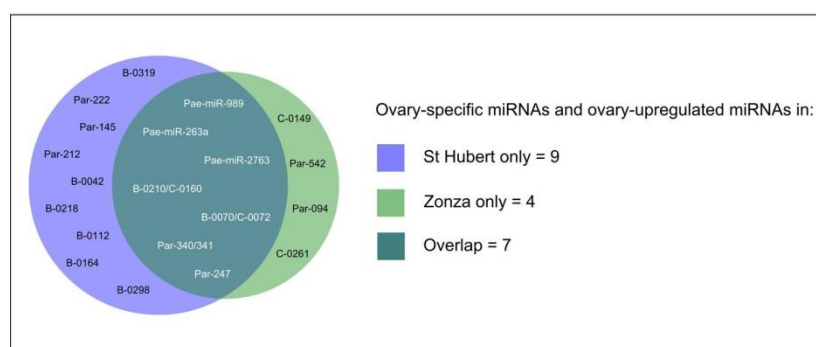


Fig 4. miRNAs with higher abundances in the ovarian datasets relative to the pooled miRNA transcriptome.

doi:10.1371/journal.pone.0142243.g004

Table 4. Trinity assembly statistics for the separate St Hubert and Zonza mRNA transcriptomes as well as a transcriptome combining reads from both populations. Number of contigs is prior to filtering by BLAST to remove potentially misassembled contigs.

	Combined assembly	St Hubert	Zonza
Number of contigs	54861	40915	45206
Total length	64060799	47311956	51104606
Maximum contig length	16200	16174	14446
N50	2236	2295	2130
N90	439	423	426

doi:10.1371/journal.pone.0142243.t004

Identification of putative ovarian miRNA targets

To identify possible mRNAs which may be regulated by the ovary-set miRNAs, we ran target predictions for these miRNAs against ovarian transcriptomes constructed from the same *P. aegeria* total RNA extraction. The Illumina HiSeq 2000 platform was used to generate 26,332,892 read pairs for Zonza and 26,466,283 read pairs for St Hubert.

Reads from both samples were pooled to construct a combined transcriptome assembly (Table 4, S1 Table). The PITA target prediction algorithm [27] was used to identify potential targets in the transcriptome for both 5' and 3' mature products for each of the 7 'ovary-set' miRNAs.

We then sought to identify specific miRNA-mRNA target pairs of interest within the dataset. Predicted target contigs corresponding to genes involved in ovarian and egg function [14] are listed in Fig 5. These contigs cover a variety of functions including regulation of egg

Function	miRNA	Targets
Egg polarity and embryo patterning	B-0210/C-0160-3p	<i>bun, Ets97D, exu, msn, spoonbill</i>
	B-0210_C-0160-3p	<i>Rab11</i>
	Pae-miR-989-3p	<i>msn, spoonbill</i>
	Pae-miR-263a-3p	<i>exu</i>
	Pae-miR-989-5p	<i>bun, Rab11, Pl4Kllalpha, gcl</i>
Morphogenesis	Pae-miR-989-3p	<i>inx3</i>
	Pae-miR-989-5p	<i>blistered</i>
	Pae-miR-263a-5p	<i>stmA</i>
Follicle cell function	B-0210/C-0160-3p	<i>Km, Tie, sdt</i>
	Pae-miR-989-3p	<i>crb, sdt</i>
	Pae-miR-989-5p	<i>mbc, tai</i>
Border cell migration	B-0070/C-0072-5p	<i>Nrg</i>
	B-0401/C-0173-3p	<i>Nrg</i>
	Pae-miR-989-5p	<i>Nrg</i>
Immune defence	Pae-miR-989-3p	<i>hemolin</i>
Stem cell regulation	B-0210/C-0160-3p	<i>Mad</i>
	Pae-miR-989-5p	<i>Mad, osa, RanBPM</i>
	Pae-miR-989-3p	<i>Mad</i>
	Pae-miR-263a-3p	<i>Mad</i>
	Pae-miR-263a-5p	<i>Mad</i>
Oocyte determination	B-0210/C-0160-3p	<i>egl, Dhc64C</i>
	Pae-miR-989-5p	<i>Dhc64C</i>
	Pae-miR-989-3p	<i>egl, Dhc64C</i>
Serosa	Pae-miR-989-5p	<i>ShxD</i>
	Pae-miR-989-3p	<i>ShxD</i>
	B-0210/C-0160-3p	<i>wdb</i>
Lipid storage	Pae-miR-263a-3p	<i>Akt1</i>
	Par-247-5p	<i>Akt1</i>
	Pae-miR-263a-5p	<i>Akt1</i>
Hormone response	B-0210/C-0160-3p	<i>E75</i>
	Pae-miR-989-5p	<i>E75</i>

Fig 5. Predicted targets of 'ovary set' miRNAs with reported roles in ovarian and egg function.

doi:10.1371/journal.pone.0142243.g005

chamber development, oocyte provisioning and defence against pathogens and the external environment. In many cases, multiple ‘ovary set’ miRNAs are predicted to target the same transcript, an example of which is *neuroglian* (*nrg*) involved in regulation of border cell migration in flies [28].

Discussion

Diversity of miRNAs in *Pararge aegeria*

In a previous study, 139 miRNA genes were identified in the Speckled Wood butterfly, *P. aegeria* [15]. Through deep sequencing of small RNAs in adult ovaries from two geographically-distinct populations, we have now increased this number to 150 expressed miRNA genes. Of the 11 new miRNAs identified in the ovarian small RNA transcriptome, six are uniquely expressed in the St Hubert sample, two are unique to Zonza, and three were found in both populations.

A common set of 55 miRNAs are expressed in the ovaries of both St Hubert and Zonza. The most highly expressed of these include miRNAs also known to be highly expressed in other insects, such as miR-989 and miR-1. miR-1 is highly conserved across bilaterians [29] and has documented expression in muscle in vertebrates [30,31] and insects [32]. Its expression is also known to be highly abundant in the ovaries of the cockroach *Blattella germanica* [33]. miR-989 is known to be expressed in the ovary in mosquitoes [19] and flies [34].

Twenty-one of the 150 miRNA genes in the *P. aegeria* miRNA transcriptome are upregulated at least 4-fold in, or are specific to, the ovary in at least one sample. Seven of these are either upregulated or ovary-specific in both populations. We refer to these seven genes as the ‘ovary set’ miRNAs.

miRNA age, ovary expression and expression level

New miRNAs tend to be expressed at low levels and in restricted spatiotemporal domains [9,10]. These characteristics would avoid severe fitness penalties due to the minimal requirements for miRNA-target binding and the resultant large number of potential targets in the transcriptome. Over time, deleterious target sites would be purged from the genome and advantageous target sites preserved, enabling the miRNA to increase its expression level and relax its tissue specificity.

Here we utilise our characterisation of the butterfly ovarian miRNA transcriptome, combined with our previous analysis of a pooled whole body and developmental miRNA transcriptome, to examine expression levels of conserved and lineage-specific miRNAs, as well as their relative contribution to the ovarian miRNA transcriptome. We consider separately the two ways in which overall miRNA expression might be increased—breadth of expression domains and level of transcription. Based on the model described above, we would predict that conserved miRNAs exhibit overall higher expression levels in the pooled transcriptome due to a wider range of expression domains. If evolutionarily-younger miRNAs were more tissue-specific miRNAs, these might be predicted to be under-represented in the pooled transcriptome, and reads might even be missed due to dilution effects.

We observed a larger proportion of conserved miRNAs expressed in the ovary compared to the pooled whole-body transcriptome, somewhat at odds with the predictions made by the model. One possible explanation for this is that the whole-body transcriptome includes RNAs from embryonic stages. It has been suggested that the early embryo represents a permissive environment for newly acquired miRNAs to be expressed and exert their effects [35]. Increased representation of young miRNAs in a library comprised of pooled developmental stages may result from such miRNAs being expressed during embryogenesis. An alternative explanation is

that the capacity of a particular tissue to tolerate experimentation with novel miRNA expression may vary between tissues; tissues with highly complex transcriptomes such as the primate brain, which provides strong evidence for the above model [11], may differ in their robustness and resultant response to new miRNAs compared to those with simpler transcriptomes. Thus, it is possible that while some tissues express a large number of lineage-specific (young) miRNAs, new miRNAs arising under restricted transcriptional control in less permissive tissues are more likely to be eliminated by selection.

Conserved miRNAs are also predicted to display higher expression levels than lineage-specific miRNAs. A seminal study in mammals [12] reported a striking correlation between expression levels of miRNA families and age. In agreement with this report, deeply conserved miRNAs have the highest expression levels in all our datasets and a significant positive correlation exists between miRNA age and expression level. We also observe that both conserved and young miRNAs possess a wide range of expression levels, and that these ranges overlap substantially with each other. The observation that older lepidopteran miRNAs tend to exhibit higher expression levels is in line with the model of Chen & Rajewsky, and also replicates findings observed in mammals [12]. Our finding that some evolutionarily young miRNAs can also be expressed at high levels may also be reconciled with predictions of the model. Given the constant high turnover of young miRNAs in metazoan genomes [12,15,36], deep sequencing of a particular tissue presents a snapshot of the transcriptome at a given point in time. It is unlikely that there is any mechanistic bias behind miRNA birth to predispose newly-evolved miRNAs to emerge under the control of tissue-specific promoters with low expression levels, but it may be the case that such miRNAs are more likely to subsequently avoid elimination through natural selection.

Potential conserved ovarian functions for insect miRNAs

A combined transcriptome assembly for both populations was constructed and used for miRNA target prediction on mature 5' and 3' products of all seven 'ovary set' miRNAs. One of the newly identified miRNAs within the 'ovary set' is the Speckled Wood homologue of miR-2763, previously described in the moths *B. mori* and *Manduca sexta* [37]. Three potential targets were identified for Bmo-miR-2763 in the *B. mori* genome [38]. The first encodes the diapause hormone receptor (*DHR*), expressed in the silkworm ovary [39], where it is involved in transduction of the diapause hormone signal and exerts regulation over embryonic diapause. The second predicted target, *GRF*, encodes an insect protein related to vertebrate germ cell nuclear factor (GCNF). [40]. GRF may function in ecdysteroid signalling in moths [40]. The third putative target, *GATA β* , is implicated in choriogenesis [38]. Although these target predictions were not replicated in *P. aegeria*, this may be due to partial mRNA predictions in the ovary transcriptome data, so some conservation of function is still possible.

Another miRNA with possible conserved ovarian functions across a wide variety of insect species is miR-989: a likely insect-specific miRNA [41] and the most abundant miRNA in the butterfly ovary. We found expression of this mRNA to be upregulated > 5-fold in both butterfly ovary samples relative to the pooled sample, and similarly it is highly expressed in ovaries of several fly species [42]; it is also upregulated in mosquito ovaries in response to blood feeding [19]. The pupal phase, during which oogenesis occurs, coincides with increased expression miR-989 in *B. mori* [38], suggestive of a similar upregulation of this miRNA during egg production in the Lepidoptera. In *Drosophila melanogaster* [34], it was demonstrated that a loss of miR-989 function impairs border cell migration in the egg chamber, leading to sterility.

miR-263a is the third 'ovary set' miRNA found in *P. aegeria* which has been characterised in a variety of other insect species. It has been hypothesised that miR-263a is involved in trans-

generational immune priming in the greater wax moth *Galleria mellonella* [43]. It is strongly upregulated in *G. mellonella* pupae, and is also induced in the larval midgut, the rest of the body, and the eggs of females fed as larvae on diets contaminated with pathogenic bacteria. Our finding that miR-263a is highly expressed in the ovaries of *P. aegeria* lends support to the idea that a role in defence could be conserved across Lepidoptera, but more direct evidence is required to test this.

Candidate targets and their involvement in butterfly ovarian biology

Further insights into the role of each miRNA on ovarian biology can be made by bioinformatic analysis of a curated list of genes implicated in egg and ovary function [14]. In each case, we suggest that the miRNA of interest may be as fine-tuners of gene activity rather than on/off switches for gene expression. This would grant robustness to the system and buffer it against developmental perturbation [44,45].

Hormone signalling

The nuclear receptor *E75* is an early response gene in the ecdysone signalling pathway and is involved in the ovarian response to hormone signalling in silkworms [46] and flies [47]. We identified the *E75* mRNA splice variant *E75B* as a potential target for two of our 'ovary set' miRNAs (Fig 5). Both the predicted miR-989 and B-0210/C-0160 target sites are located at the start codon of this transcript.

Follicle cell function

Follicle cells perform several key functions during insect oogenesis including egg provisioning [48], secretion of maternal ecdysteroids [49], and production of the chorion and eggshell. Follicle cells in flies also provide spatial patterning information to the oocyte [50].

In *Drosophila*, a subset of follicle cells ('border cells') contribute to both anterior patterning as well as the formation of the micropyle through which sperm enter the egg at fertilization [51]. Several 'ovary set' miRNAs, including miR-989 as mentioned above, are predicted to target genes involved in follicle cell function and border cell migration. These include *stardust*, *crumbs* [52], *neuroglian* [28] and *taiman* (*tai*) [53].

Serosa and immune defence

We predict that miR-989 may target the mRNA of *ShxD*, a homeobox gene expressed maternally in ovarioles and zygotically in the serosa [13]. The *Shx* genes are implicated in specifying the location of the extra-embryonic serosal tissue domain in developing oocytes, as well as regulating its subsequent development. As noted above, miR-263a may also be involved in immune defence, through different targets.

Supporting Information

S1 Fig. Alignment of miR-2763 sequences across Lepidoptera.
(DOCX)

S1 Table. Annotation of the combined mRNA transcriptome assembly.
(XLSX)

S2 Table. Read count comparisons for miRNA transcriptomes.
(DOCX)

Acknowledgments

We thank Adam Hargreaves, Jerome Hui and Melanie Gibbs for advice, and Luca Livraghi for assistance in maintaining the butterfly stocks. We also thank two reviewers for very constructive suggestions.

Author Contributions

Conceived and designed the experiments: PWHH SQ. Performed the experiments: SQ CJB. Analyzed the data: SQ CJB PWHH. Contributed reagents/materials/analysis tools: SQ CJB. Wrote the paper: SQ CJB PWHH.

References

- Behura SK (2007). Insect microRNAs: Structure, function and evolution. *Insect Biochem Mol Biol* 37 (1):3–9. doi: [10.1016/j.ibmb.2006.10.006](https://doi.org/10.1016/j.ibmb.2006.10.006) PMID: [17175441](https://pubmed.ncbi.nlm.nih.gov/17175441/)
- Yekta S, Shih I-H, Bartel DP. MicroRNA-directed cleavage of HOXB8 mRNA (2004). *Science* 23; 304 (5670):594–6. doi: [10.1126/science.1097434](https://doi.org/10.1126/science.1097434) PMID: [15105502](https://pubmed.ncbi.nlm.nih.gov/15105502/)
- Pasquinelli AE, Hunter S, Bracht J. (2005). MicroRNAs: a developing story. *Curr Opin Genet Dev.* 15 (2):200–5. doi: [10.1016/j.gde.2005.01.002](https://doi.org/10.1016/j.gde.2005.01.002) PMID: [15797203](https://pubmed.ncbi.nlm.nih.gov/15797203/)
- Jones CI, Newbury SF (2010). Functions of microRNAs in *Drosophila* development. *Biochem Soc Trans.* 38(4):1137–43. doi: [10.1042/BST0381137](https://doi.org/10.1042/BST0381137) PMID: [20659018](https://pubmed.ncbi.nlm.nih.gov/20659018/)
- Lasko P (2011). Posttranscriptional regulation in *Drosophila* oocytes and early embryos. *Wiley Interdisciplinary Reviews: RNA.* 2(3):408–16. doi: [10.1002/wrna.70](https://doi.org/10.1002/wrna.70) PMID: [21957026](https://pubmed.ncbi.nlm.nih.gov/21957026/)
- Kenny NJ, Quah S, Holland PWH, Tobe SS, Hui JHL (2013). How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment? *Gen Comp Endocrinol.* 188:16–22. doi: [10.1016/j.ygcen.2013.02.013](https://doi.org/10.1016/j.ygcen.2013.02.013) PMID: [23480873](https://pubmed.ncbi.nlm.nih.gov/23480873/)
- Lozano J, Montañez R, Belles X (2015). miR-2 family regulates insect metamorphosis by controlling the juvenile hormone signaling pathway. *Proc Nat Acad Sci.* 112(12):3740–3745. doi: [10.1073/pnas.1418522112](https://doi.org/10.1073/pnas.1418522112) PMID: [25775510](https://pubmed.ncbi.nlm.nih.gov/25775510/)
- McBride D, Carré W, Sontakke SD, Hogg CO, Law A, Donadeu FX, et al (2012). Identification of miRNAs associated with the follicular-luteal transition in the ruminant ovary. *Reproduction.* 144(2):221–33. doi: [10.1530/REP-12-0025](https://doi.org/10.1530/REP-12-0025) PMID: [22653318](https://pubmed.ncbi.nlm.nih.gov/22653318/)
- Chen K, Rajewsky N (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet.* 8(2):93–103. doi: [10.1038/nrg1990](https://doi.org/10.1038/nrg1990) PMID: [17230196](https://pubmed.ncbi.nlm.nih.gov/17230196/)
- Berezikov E (2011). Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet.* 12 (12):846–60. doi: [10.1038/nrg3079](https://doi.org/10.1038/nrg3079) PMID: [22094948](https://pubmed.ncbi.nlm.nih.gov/22094948/)
- Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, et al (2006). Diversity of microRNAs in human and chimpanzee brain. *Nat Genet.* 38(12):1375–7. doi: [10.1038/ng1914](https://doi.org/10.1038/ng1914) PMID: [17072315](https://pubmed.ncbi.nlm.nih.gov/17072315/)
- Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al (2013). Birth and expression evolution of mammalian microRNA genes. *Genome Res.* 23(1):34–45. doi: [10.1101/gr.140269.112](https://doi.org/10.1101/gr.140269.112) PMID: [23034410](https://pubmed.ncbi.nlm.nih.gov/23034410/)
- Ferguson L, Marlétaz F, Carter J-M, Taylor WR, Gibbs M, Breuker CJ, et al (2014). Ancient expansion of the Hox cluster in Lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS Genet.* 10(10):e1004698. doi: [10.1371/journal.pgen.1004698](https://doi.org/10.1371/journal.pgen.1004698) PMID: [25340822](https://pubmed.ncbi.nlm.nih.gov/25340822/)
- Carter J-M, Baker SC, Pink R, Carter DRF, Collins A, Tomlin J, et al (2013). Unscrambling butterfly oogenesis. *BMC Genomics.* 14:283. doi: [10.1186/1471-2164-14-283](https://doi.org/10.1186/1471-2164-14-283) PMID: [23622113](https://pubmed.ncbi.nlm.nih.gov/23622113/)
- Quah S, Hui JHL, Holland PWH (2015). A Burst of miRNA Innovation in the Early Evolution of Butterflies and Moths. *Mol Biol Evol.* 32(5):1161–1174. doi: [10.1093/molbev/msv004](https://doi.org/10.1093/molbev/msv004) PMID: [25576364](https://pubmed.ncbi.nlm.nih.gov/25576364/)
- Bastock R, Johnston D St. *Drosophila* oogenesis. *Curr Biol.*2008. 18(23):R1082–7. doi: [10.1016/j.cub.2008.09.011](https://doi.org/10.1016/j.cub.2008.09.011)
- Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. *Development.* 2009; 136 (18):3033–42. doi: [10.1242/dev.033183](https://doi.org/10.1242/dev.033183) PMID: [19700615](https://pubmed.ncbi.nlm.nih.gov/19700615/)
- Barckmann B, Simonelig M. Control of maternal mRNA stability in germ cells and early embryos. *Biochim Biophys Acta.* 2013; 1829(6–7):714–24. doi: [10.1016/j.bbarm.2012.12.011](https://doi.org/10.1016/j.bbarm.2012.12.011) PMID: [23298642](https://pubmed.ncbi.nlm.nih.gov/23298642/)

19. Mead EA, Tu Z. Cloning, characterization, and expression of microRNAs from the Asian malaria mosquito, *Anopheles stephensi*. BMC Genomics. 2008; 9:244. doi: [10.1186/1471-2164-9-244](https://doi.org/10.1186/1471-2164-9-244) PMID: [18500992](https://pubmed.ncbi.nlm.nih.gov/18500992/)
20. Chen W, Jiang GF, Sun SH, Lu Y, Ma F, Li B. Identification of differentially expressed genes in American cockroach ovaries and testes by suppression subtractive hybridization and the prediction of its miRNAs. Mol Genet Genomics. 2013; 288(11):627–38. doi: [10.1007/s00438-013-0777-1](https://doi.org/10.1007/s00438-013-0777-1) PMID: [23996145](https://pubmed.ncbi.nlm.nih.gov/23996145/)
21. Preuss KM, Lopez J a, Colbourne JK, Wade MJ. Identification of maternally-loaded RNA transcripts in unfertilized eggs of *Tribolium castaneum*. BMC Genomics. 2012; 13(1):671. doi: [10.1186/1471-2164-13-671](https://doi.org/10.1186/1471-2164-13-671)
22. Gibbs M, Breuker CJ, Van Dyck H (2013). Flight during oviposition reduces maternal egg provisioning and influences offspring development in *Pararge aegeria* (L.). Physiol Entomol. 35(1):29–39. doi: [10.1111/j.1365-3032.2009.00706.x](https://doi.org/10.1111/j.1365-3032.2009.00706.x)
23. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 40(1):37–52. doi: [10.1093/nar/gkr688](https://doi.org/10.1093/nar/gkr688) PMID: [21911355](https://pubmed.ncbi.nlm.nih.gov/21911355/)
24. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–52. doi: [10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883) PMID: [21572440](https://pubmed.ncbi.nlm.nih.gov/21572440/)
25. Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12:323. doi: [10.1186/1471-2105-12-323](https://doi.org/10.1186/1471-2105-12-323) PMID: [21816040](https://pubmed.ncbi.nlm.nih.gov/21816040/)
26. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21(18):3674–6. doi: [10.1093/bioinformatics/bti610](https://doi.org/10.1093/bioinformatics/bti610) PMID: [16081474](https://pubmed.ncbi.nlm.nih.gov/16081474/)
27. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007). The role of site accessibility in microRNA target recognition. Nat Genet. 39(10):1278–84. doi: [10.1038/ng2135](https://doi.org/10.1038/ng2135) PMID: [17893677](https://pubmed.ncbi.nlm.nih.gov/17893677/)
28. Wei J, Hortsch M, Goode S (2004). Neuroglial stabilizes epithelial structure during *Drosophila* oogenesis. Dev Dyn. 230(4):800–8. doi: [10.1002/dvdy.20108](https://doi.org/10.1002/dvdy.20108) PMID: [15254915](https://pubmed.ncbi.nlm.nih.gov/15254915/)
29. Wheeler BM, Heimberg AM, Moy VN, Sperling E a, Holstein TW, Heber S, et al (2009). The deep evolution of metazoan microRNAs. Evol Dev. 11(1):50–68. doi: [10.1111/j.1525-142X.2008.00302.x](https://doi.org/10.1111/j.1525-142X.2008.00302.x) PMID: [19196333](https://pubmed.ncbi.nlm.nih.gov/19196333/)
30. Chen J-F, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, et al (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nat Genet. 38(2):228–33. doi: [10.1038/ng1725](https://doi.org/10.1038/ng1725) PMID: [16380711](https://pubmed.ncbi.nlm.nih.gov/16380711/)
31. Kloosterman WP, Wienholds E, de Bruijn E, Kauppinen S, Plasterk RHA (2006). In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. Nat Methods. 3(1):27–9. doi: [10.1038/nmeth843](https://doi.org/10.1038/nmeth843) PMID: [16369549](https://pubmed.ncbi.nlm.nih.gov/16369549/)
32. Sokol NS (2005). Mesodermally expressed *Drosophila* microRNA-1 is regulated by Twist and is required in muscles during larval growth. Genes Dev. 19(19):2343–54. doi: [10.1101/gad.1356105](https://doi.org/10.1101/gad.1356105) PMID: [16166373](https://pubmed.ncbi.nlm.nih.gov/16166373/)
33. Cristino AS, Tanaka ED, Rubio M, Piulachs M-D, Belles X (2011). Deep Sequencing of Organ- and Stage-Specific microRNAs in the Evolutionarily Basal Insect *Blattella germanica* (L.) (Dictyoptera, Blattellidae). PLoS One. 28; 6(4):e19350. doi: [10.1371/journal.pone.0019350](https://doi.org/10.1371/journal.pone.0019350) PMID: [21552535](https://pubmed.ncbi.nlm.nih.gov/21552535/)
34. Kugler JM, Chen YW, Weng R, Cohen SM (2013). miR-989 Is required for border cell migration in the *Drosophila* Ovary. PLoS One. 8(7): e67075. doi: [10.1371/journal.pone.0067075](https://doi.org/10.1371/journal.pone.0067075) PMID: [23843983](https://pubmed.ncbi.nlm.nih.gov/23843983/)
35. Ninova M, Ronshaugen M, Griffiths-Jones S (2014). Fast-evolving microRNAs are highly expressed in the early embryo of *Drosophila virilis*. RNA. 20:360–372. doi: [10.1261/ma.041657.113](https://doi.org/10.1261/ma.041657.113) PMID: [24448446](https://pubmed.ncbi.nlm.nih.gov/24448446/)
36. Lyu Y, Shen Y, Li H, Chen Y, Guo L, Zhao Y, et al (2014). New microRNAs in *Drosophila*—birth, death and cycles of adaptive evolution. PLoS Genet. 10(1):e1004096. doi: [10.1371/journal.pgen.1004096](https://doi.org/10.1371/journal.pgen.1004096) PMID: [24465220](https://pubmed.ncbi.nlm.nih.gov/24465220/)
37. Zhang X, Zheng Y, Jagadeeswaran G, Ren R, Sunkar R, Jiang H (2012). Identification and developmental profiling of conserved and novel microRNAs in *Manduca sexta*. Insect Biochem Mol Biol. Elsevier Ltd; 42(6):381–95.
38. Jagadeeswaran G, Zheng Y, Sumathipala N, Jiang H, Arrese EL, Soulages JL, et al (2010). Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development. BMC Genomics. 11:52. doi: [10.1186/1471-2164-11-52](https://doi.org/10.1186/1471-2164-11-52) PMID: [20089182](https://pubmed.ncbi.nlm.nih.gov/20089182/)

39. Homma T, Watanabe K, Tsurumaru S, Kataoka H, Imai K, Kamba M, et al (2006). G protein-coupled receptor for diapause hormone, an inducer of Bombyx embryonic diapause. *Biochem Biophys Res Commun.* 344(1):386–93. doi: [10.1016/j.bbrc.2006.03.085](https://doi.org/10.1016/j.bbrc.2006.03.085) PMID: [16600181](https://pubmed.ncbi.nlm.nih.gov/16600181/)
40. Charles JP, Shinoda T, Chinzei Y (1999). Characterization and DNA-binding properties of GRF, a novel monomeric binding orphan receptor related to GCNF and FTZ-F1. *Eur J Biochem.* 266(1):181–90. doi: [10.1046/j.1432-1327.1999.00842.x](https://doi.org/10.1046/j.1432-1327.1999.00842.x) PMID: [10542063](https://pubmed.ncbi.nlm.nih.gov/10542063/)
41. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39:D152–7. doi: [10.1093/nar/gkq1027](https://doi.org/10.1093/nar/gkq1027) PMID: [21037258](https://pubmed.ncbi.nlm.nih.gov/21037258/)
42. Lucas KJ, Myles KM, Raikhel AS (2013). Small RNAs: A new frontier in mosquito biology. *Trends Parasitol.* 29(6):295–303. doi: [10.1016/j.pt.2013.04.003](https://doi.org/10.1016/j.pt.2013.04.003) PMID: [23680188](https://pubmed.ncbi.nlm.nih.gov/23680188/)
43. Mukherjee K, Vilcinskas A (2014). Development and immunity-related microRNAs of the lepidopteran model host *Galleria mellonella*. *BMC Genomics.* 15(1):705. doi: [10.1186/1471-2164-15-705](https://doi.org/10.1186/1471-2164-15-705)
44. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell.* 123(6):1133–46. doi: [10.1016/j.cell.2005.11.023](https://doi.org/10.1016/j.cell.2005.11.023) PMID: [16337999](https://pubmed.ncbi.nlm.nih.gov/16337999/)
45. Ebert MS, Sharp PA (2012). Roles for MicroRNAs in conferring robustness to biological processes. *Cell.* 149(3):505–24. doi: [10.1016/j.cell.2012.04.005](https://doi.org/10.1016/j.cell.2012.04.005)
46. Papanonis A, Swevers L, Iatrou K (2015). Chorion Genes: A landscape of their Evolution, Structure, and regulation. *Annu Rev Entomol.* doi: [10.1146/annurev-ento-010814-020810](https://doi.org/10.1146/annurev-ento-010814-020810)
47. Ables ET, Bois KE, Garcia CA, Drummond-Barbosa D (2015). Ecdysone response gene E78 controls ovarian germline stem cell niche formation and follicle survival in *Drosophila*. *Dev Biol.* 400(1):33–42. doi: [10.1016/j.ydbio.2015.01.013](https://doi.org/10.1016/j.ydbio.2015.01.013) PMID: [25624267](https://pubmed.ncbi.nlm.nih.gov/25624267/)
48. Swevers L, Iatrou K (2003). The ecdysone regulatory cascade and ovarian development in lepidopteran insects: Insights from the silkworm paradigm. *Insect Biochem Molec.* 33(12):1285–97. doi: [10.1016/j.ibmb.2003.06.012](https://doi.org/10.1016/j.ibmb.2003.06.012)
49. Kadono-okuda K, Amornsak W, Yamashita O (1994). Controlled Ecdysteroid Accumulation in Eggs of the Silkworm, *Bombyx mori*, by an Imidazole Compound (KK-42), and Embryogenesis in These Eggs. *Arch Insect Biochem.* 135(1 994):121–35. doi: [10.1002/arch.940250205](https://doi.org/10.1002/arch.940250205)
50. Saunders C, Cohen RS (1999). The role of oocyte transcription, the 5'UTR, and translation repression and derepression in *Drosophila* gurken mRNA and protein localization. *Mol Cell.* 3(1):43–54. doi: [10.1016/S1097-2765\(00\)80173-2](https://doi.org/10.1016/S1097-2765(00)80173-2) PMID: [10024878](https://pubmed.ncbi.nlm.nih.gov/10024878/)
51. Montell DJ (2003). Border-cell migration: the race is on. *Nat Rev Mol Cell Biol.* 4(1):13–24. doi: [10.1038/nrm1006](https://doi.org/10.1038/nrm1006) PMID: [12511865](https://pubmed.ncbi.nlm.nih.gov/12511865/)
52. Pinheiro EM, Montell DJ (2004). Requirement for Par-6 and Bazooka in *Drosophila* border cell migration. *Development.* 131(21):5243–51. doi: [10.1242/dev.01412](https://doi.org/10.1242/dev.01412) PMID: [15456726](https://pubmed.ncbi.nlm.nih.gov/15456726/)
53. McDonald JA, Pinheiro EM, Montell DJ (2003). PVF1, a PDGF/VEGF homolog, is sufficient to guide border cells and interacts genetically with Taiman. *Development.* 130(15):3469–78. doi: [10.1242/dev.00574](https://doi.org/10.1242/dev.00574) PMID: [12810594](https://pubmed.ncbi.nlm.nih.gov/12810594/)

Bibliography

- Ables, E. T., Bois, K. E., Garcia, C. A., & Drummond-Barbosa, D. (2015). Ecdysone response gene E78 controls ovarian germline stem cell niche formation and follicle survival in *Drosophila*. *Developmental Biology*, 400(1), 33–42.
- Agranat-Tamir, L., Shomron, N., Sperling, J., & Sperling, R. (2014). Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. *Nucleic Acids Research*, 42, 4640–4651.
- Ahn, H. W., Morin, R. D., Zhao, H., Harris, R. A., Coarfa, C., Chen, Z. J., ... Rajkovic, A. (2010). MicroRNA transcriptome in the newborn mouse ovaries determined by massive parallel sequencing. *Molecular Human Reproduction*, 16(7), 463–471.
- Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., & Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Current Biology*, 13(10), 807–818.
- Artavanis-Tsakonas, S. (1999). Notch Signaling: Cell Fate Control and Signal Integration in Development. *Science*, 284(5415), 770–776.
- Axtell, M. J., Westholm, J. O., & Lai, E. C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology*, 12(4), 221.
- Backes, C., Meese, E., Lenhof, H. P., & Keller, A. (2010). A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Research*, 38(13), 4476–4486.
- Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K., ... Tewari, M. (2008). MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells*, 26(10), 2496–2505.
- Barckmann, B., & Simonelig, M. (2013). Control of maternal mRNA stability in germ cells and early embryos. *Biochimica et Biophysica Acta*, 1829(6-7), 714–24.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, 41(D1): D991-D995.
- Baskerville, S., & Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA (New York, N.Y.)*, 11(3), 241–7.
- Bazinet, A. L., Cummings, M. P., Mitter, K. T., & Mitter, C. W. (2013). Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. *PLoS One*, 8(12), e82615.
- Begun, D. J., Lindfors, H. A., Thompson, M. E., Holloway, A. K., (2006). Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics*, 172(3), 1675-1681.
- Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, 12(12), 846–860.
- Berezikov, E., Cuppen, E., & Plasterk, R. H. (2006). Approaches to microRNA discovery. *Nature Genetics*, 38 Suppl(May), S2–7.

- Berezikov, E., Liu, N., Flynt, A. S., Hodges, E., Rooks, M., Hannon, G. J., & Lai, E. C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*. *Nature Genetics*, 42(1), 6–9; author reply 9–10.
- Berezikov, E., Robine, N., Samsonova, A., Westholm, J. O., Naqvi, A., Hung, J. H., ... Lai, E. C. (2011). Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence. *Genome Research*, 21, 203–215.
- Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., & Plasterk, R. H. A. (2006). Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*, 38(12), 1375–1377.
- Béthune, J., Artus-Revel, C. G., & Filipowicz, W. (2012). Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. *EMBO Reports*, 13(8), 716–23.
- Bishop, C. V., Satterwhite, S., Xu, L., Hennebold, J. D., & Stouffer, R. L. (2012). Microarray analysis of the primate luteal transcriptome during chorionic gonadotrophin administration simulating early pregnancy. *Molecular Human Reproduction*, 18(4), 216–227.
- Bohnsack, M. T. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA (New York, N.Y.)*, 10(2), 185–191.
- Braun, L., Cannella, D., Ortet, P., Barakat, M., Sautel, C. F., Kieffer, S., ... Hakimi, M. A. (2010). A complex small RNA repertoire is generated by a plant/fungal-like machinery and effected by a metazoan-like Argonaute in the single-cell human parasite *Toxoplasma gondii*. *PLoS Pathogens*, 6, 1–20.
- Brodersen, P., & Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. *Nature Reviews. Molecular Cell Biology*, 10(2), 141–8.
- Brodsky, A. S., Meyer, C. A., Swinburne, I. A., Hall, G., Keenan, B. J., Liu, X. S., ... Silver, P. A. (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biology*, 6(8), R64.
- Brooke, N. M., Garcia-Fernández, J., & Holland, P. W. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679), 920–2.
- Brunetti, C. R., Selegue, J. E., Monteiro, A., French, V., Brakefield, P. M., & Carroll, S. B. (2001). The generation and diversification of butterfly eyespot color patterns. *Current Biology*, 11(20), 1578–1585.
- Bushati, N., Stark, A., Brennecke, J., & Cohen, S. M. (2008). Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in *Drosophila*. *Current Biology*, 18(7), 501–6.
- Cai, J. J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein coding gene in *Saccharomyces cerevisiae*. *Genetics*, 179, 487–496.
- Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, 10(12), 1957–66.
- Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J., & Schubert, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evolution & Development*, 13(1), 15–27.

- Candiani, S. (2012). Focus on miRNAs evolution: a perspective from amphioxus. *Briefings in Functional Genomics*, 11(2), 107–117.
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., ... Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6), 626–635.
- Carter, J.-M., Baker, S. C., Pink, R., Carter, D. R. F., Collins, A., Tomlin, J., ... Breuker, C. J. (2013). Unscrambling butterfly oogenesis. *BMC Genomics*, 14, 283.
- Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4), 642–55.
- Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., ... Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Ptx1 enhancer. *Science (New York, N.Y.)*, 327(5963), 302–5.
- Charles, J. P., Shinoda, T., & Chinzei, Y. (1999). Characterization and DNA-binding properties of GRF, a novel monomeric binding orphan receptor related to GCNF and FTZ-F1. *European Journal of Biochemistry*, 266(1), 181–190.
- Chen, J.-F., Mandel, E. M., Thomson, J. M., Wu, Q., Callis, T. E., Hammond, S. M., ... Wang, D.-Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nature Genetics*, 38(2), 228–233.
- Chen, K., & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews. Genetics*, 8(2), 93–103.
- Chen, W., Jiang, G. F., Sun, S. H., Lu, Y., Ma, F., & Li, B. (2013). Identification of differentially expressed genes in American cockroach ovaries and testes by suppression subtractive hybridization and the prediction of its miRNAs. *Molecular Genetics and Genomics*, 288(11), 627–638.
- Chen, X. S., Collins, L. J., Biggs, P. J., & Penny, D. (2009). High throughput genome-wide survey of small RNAs from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*. *Genome Biology and Evolution*, 1, 165–175.
- Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., ... Bartel, D. P. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development*, 24(10), 992–1009.
- Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S., ... Arendt, D. (2010). Ancient animal microRNAs and the evolution of tissue identity. *Nature*, 463(7284), 1084–8.
- Chung, A. C., & Cooney, A. J. (2001). Germ cell nuclear factor. *The International Journal of Biochemistry & Cell Biology*, 33(12), 1141–1146.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), 3674–3676.
- Cristino, A. S., Tanaka, E. D., Rubio, M., Piulachs, M.-D., & Belles, X. (2011). Deep Sequencing of Organ- and Stage-Specific microRNAs in the Evolutionarily Basal Insect *Blattella germanica* (L.) (Dictyoptera, Blattellidae). *PLoS ONE*, 6(4), e19350.

- Cullen, B. R. (2004). Transcription and processing of human microRNA precursors. *Molecular Cell*, 16(6), 861–5.
- Daniel, C., Behm, M., & Öhman, M. (2015). The role of Alu elements in the cis-regulation of RNA processing. *Cellular and Molecular Life Sciences*. 1990
- Deschamps, J., & van Nes, J. (2005). Developmental regulation of the Hox genes during axial morphogenesis in the mouse. *Development (Cambridge, England)*, 132(13), 2931–2942.
- Djuranovic, S., Nahvi, A., & Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science (New York, N.Y.)*, 336(6078), 237–40.
- Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325), 815–8.
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74
- Duursma, A. M., Kedde, M., Schrier, M., le Sage, C., & Agami, R. (2008). miR-148 targets human DNMT3b protein coding region. *RNA (New York, N.Y.)*, 14(5), 872–7.
- Ebert, M. S., & Sharp, P. A. (2012). Roles for MicroRNAs in conferring robustness to biological processes. *Cell*, 149, 515–524.
- Edgecombe, G. D., Giribet, G., Dunn, C. W., Hejnol, A., Kristensen, R. M., Neves, R. C., ... Sørensen, M. V. (2011). Higher-level metazoan relationships: recent progress and remaining questions. *Organisms Diversity & Evolution*, 11(2), 151–172.
- El Tayebi, H. M., Hosny, K. A., Esmat, G., Breuhahn, K., & Abdelaziz, A. I. (2012). miR-615-5p is restrictedly expressed in cirrhotic and cancerous liver tissues and its overexpression alleviates the tumorigenic effects in hepatocellular carcinoma. *FEBS Letters*, 586(19), 3309–3316.
- Etebari, K., Hussain, M., & Asgari, S. (2013). Identification of microRNAs from *Plutella xylostella* larvae associated with parasitization by *Diadegma semiclausum*. *Insect Biochemistry and Molecular Biology*, 43(4), 309–18.
- Eulalio, A., Huntzinger, E., & Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature Structural & Molecular Biology*, 15(4), 346–53.
- Fabian, M. R., Cieplak, M. K., Frank, F., Morita, M., Green, J., Srikumar, T., ... Sonenberg, N. (2011). miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT. *Nature Structural & Molecular Biology*, 18(11), 1211–7.
- Ferguson, L., Marlétaz, F., Carter, J.-M., Taylor, W. R., Gibbs, M., Breuker, C. J., & Holland, P. W. H. (2014). Ancient expansion of the Hox cluster in *Lepidoptera* generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS Genetics*, 10(10), e1004698.
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., ... Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Research*, 40(D1), D84–D90
- Fondon, J. W., & Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52), 18058–63.

- Forman, J. J., Legesse-Miller, A., & Collier, H. A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), 14879–84.
- Förstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., ... Zamore, P. D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biology*, 3(7), e236.
- Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1), 37–52.
- Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., & Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. *Genome Research*, 18(1), 1–12.
- Fromm, B., Worren, M. M., Hahn, C., Hovig, E., & Bachmann, L. (2013). Substantial loss of conserved and gain of novel microRNA families in flatworms. *Molecular Biology and Evolution*, 30(12), 2619–28.
- Gao, W., Gu, Y., Li, Z., Cai, H., Peng, Q., Tu, M., ... Miao, Y. (2014). miR-615-5p is epigenetically inactivated and functions as a tumor suppressor in pancreatic ductal adenocarcinoma. *Oncogene*, 34, 1629–1640.
- Gazave, E., Lapébie, P., Ereskovsky, A. V., Vacelet, J., Renard, E., Cárdenas, P., & Borchellini, C. (2011). No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. *Hydrobiologia*, 687(1), 3–10.
- Gibbs, M., Breuker, C. J., & Van Dyck, H. (2010). Flight during oviposition reduces maternal egg provisioning and influences offspring development in *Pararge aegeria* (L.). *Physiological Entomology*, 35(1), 29–39.
- Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., ... Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science (New York, N.Y.)*, 312(5770), 75–9.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–652.
- Griffiths-Jones, S., Hui, J. H. L., Marco, A., & Ronshaugen, M. (2011). MicroRNA evolution by arm switching. *EMBO Reports*, 12(2), 172–7.
- Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1), 91–105.
- Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., ... Bartel, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217), 1193–7.
- Gu, S., Jin, L., Zhang, F., Sarnow, P., & Kay, M. A. (2009). Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nature Structural & Molecular Biology*, 16(2), 144–50.
- Guerra-Assunção, J. A., & Enright, A. J. (2012). Large-scale analysis of microRNA evolution. *BMC Genomics*, 13(1), 218.

- Gustavson, E., Goldsborough, A. S., Ali, Z., & Kornberg, T. B. (1996). The *Drosophila* engrailed and invected genes: partners in regulation, expression and function. *Genetics*, 142(3), 893–906.
- Han, J., Lee, Y., Yeom, K.-H., Nam, J.-W., Heo, I., Rhee, J.-K., ... Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5), 887–901.
- Heinen, T. J. A. J., Staubach, F., Häming, D., Tautz, D. (2009) Emergence of a new gene from an intergenic region. *Current Biology*, 19(18), 1527-1531.
- Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. J., & Peterson, K. J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8), 2946–50. 5
- Hilgers, V., Bushati, N., & Cohen, S. M. (2010). *Drosophila* microRNAs 263a/b confer robustness during development by protecting nascent sense organs from apoptosis. *PLoS Biology*, 8(6).
- Hinas, A., Reimegård, J., Wagner, E. G. H., Nellen, W., Ambros, V. R., & Söderbom, F. (2007). The small RNA repertoire of *Dictyostelium discoideum* and its regulation by components of the RNAi pathway. *Nucleic Acids Research*, 35, 6714–6726.
- Hines, H. M., Papa, R., Ruiz, M., Papanicolaou, A., Wang, C., Nijhout, H. F., ... Reed, R. D. (2012). Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation. *BMC Genomics*, 13, 288.
- Holland, P. W. H. (2012). Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(1), 31-45
- Homma, T., Watanabe, K., Tsurumaru, S., Kataoka, H., Imai, K., Kamba, M., ... Yaginuma, T. (2006). G protein-coupled receptor for diapause hormone, an inducer of *Bombyx* embryonic diapause. *Biochemical and Biophysical Research Communications*, 344(1), 386–393.
- Huang, A., He, L., & Wang, G. (2011). Identification and characterization of microRNAs from *Phaeodactylum tricornutum* by high-throughput sequencing and bioinformatics analysis. *BMC Genomics*, 12, 337.
- Huang, P. J., Lin, W. C., Chen, S. C., Lin, Y. H., Sun, C. H., Lyu, P. C., & Tang, P. (2012). Identification of putative miRNAs from the deep-branching unicellular flagellates. *Genomics*, 99, 101–107.
- Hui, J. H. L., Marco, A., Hunt, S., Melling, J., Griffiths-Jones, S., & Ronshaugen, M. (2013). Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. *Nucleic Acids Research*, 41(5), 3352–61.
- Hulf, T., Sibbritt, T., Wiklund, E. D., Bert, S., Strbenac, D., Statham, A. L., ... Clark, S. J. (2011). Discovery pipeline for epigenetically deregulated miRNAs in cancer: integration of primary miRNA transcription. *BMC Genomics*, 12, 54.
- Ingham, P. W., & McMahon, A. P. (2001). Hedgehog signaling in animal development: paradigms and principles. *Genes & Development*, 15(23), 3059–87.
- Irie, N., & Kuratani, S. (2011). Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature Communications*, 2, 248.
- Jacob, F. (1977) Evolution and tinkering. *Science*. 196(4295), 1161-1166.

- Jagadeeswaran, G., Zheng, Y., Sumathipala, N., Jiang, H., Arrese, E. L., Soulages, J. L., ... Sunkar, R. (2010). Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development. *BMC Genomics*, 11, 52.
- Jeong, S., Rokas, A., & Carroll, S. B. (2006). Regulation of body pigmentation by the Abdominal-B Hox protein and its gain and loss in *Drosophila* evolution. *Cell*, 125(7), 1387–99.
- Jonas, S., & Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, 16(7), 421–433.
- Jones, C. I., & Newbury, S. F. (2010). Functions of microRNAs in *Drosophila* development. *Biochemical Society Transactions*, 38(4), 1137–1143.
- Kadono-okuda, K., Amornsak, W., & Yamashita, O. (1994). Controlled Ecdysteroid Accumulation in Eggs of the Silkworm, *Bombyx mori*, by an Imidazole Compound (KK-42), and Embryogenesis in These Eggs, *Archives of Insect Biochemistry and Physiology*, 135(1994), 121–135.
- Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., ... Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468, 811–814.
- Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., ... Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes & Development*, 19(4), 489–501.
- Kang, W., & Friedländer, M. R. (2015). Computational Prediction of miRNA Genes from Small RNA Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3, 7.
- Katalin, F.-T., Soirova, V., Sachidanandam, R., Assaf, G., Hannon, G. J., Kapranov, P., ... Gingeras, T. R. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, 457(7232), 1028–1032.
- Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066.
- Kawahara, A. Y., & Breinholt, J. W. (2014). Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings. Biological Sciences / The Royal Society*, 281(1788), 20140970.
- Kenny, N. J., Quah, S., Holland, P. W. H., Tobe, S. S., & Hui, J. H. L. (2013). How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment? *General and Comparative Endocrinology*, 188, 16–22.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6), 996–1006.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10), 1278–84.
- Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., & Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Development*, 15(20), 2654–9.

- Keys, D. N. (1999). Recruitment of a hedgehog Regulatory Circuit in Butterfly Eyespot Evolution. *Science*, 283(5401), 532–534.
- Kitano, H. (2007). Biological robustness in complex host-pathogen systems. *Progress in Drug Research*, 64, 239-263.
- Kloosterman, W. P., Wienholds, E., de Bruijn, E., Kauppinen, S., & Plasterk, R. H. A. (2006). In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nature Methods*, 3(1), 27–29.
- Knowles, D.G. & McLysaght, A. (2009). Recent de novo origin of human protein coding genes. *Genome Research*, 19, 1752-1759.
- Kogo, H., Kowa-Sugiyama, H., Yamada, K., Bolor, H., Tsutsumi, M., Ohye, T., ... Kurahashi, H. (2010). Screening of genes involved in chromosome segregation during meiosis I: toward the identification of genes responsible for infertility in humans. *Journal of Human Genetics*, 55(5), 293–299.
- Kok, K. H., Ng, M.-H. J., Ching, Y.-P., & 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. *The Journal of Biological Chemistry*, 282(24), 17649–57.
- Kosik, K. S. (2009). MicroRNAs tell an evo-devo story. *Nature Reviews. Neuroscience*, 10(10), 754–9.
- Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(Database issue), D152–7.
- Kugler, J. M., Chen, Y. W., Weng, R., & Cohen, S. M. (2013). miR-989 Is Required for Border Cell Migration in the Drosophila Ovary. *PLoS ONE*, 8(7). e67075
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543), 853–8.
- Lasko, P. (2011). Posttranscriptional regulation in Drosophila oocytes and early embryos. *Wiley Interdisciplinary Reviews: RNA*. 2(3), 408-416
- Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294(5543), 858–62.
- Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294(5543), 862–4.
- Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), 843–854.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., ... Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing, 425, 415-419
- Lee, Y., Jeon, K., Lee, J., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization, *The EMBO Journal*, 21(17), 4663–4670.
- Levin, M., Hashimshony, T., Wagner, F., & Yanai, I. (2012). Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. *Developmental Cell*, 22, 1101–1108.

- Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1), 15–20.
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
- Liang, Y., Ridzon, D., Wong, L., & Chen, C. (2007). Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics*, 8, 166.
- Lin, W. C., Huang, K. Y., Chen, S. C., Huang, T. Y., Chen, S. J., Huang, P. J., & Tang, P. (2009). Malate dehydrogenase is negatively regulated by miR-1 in *Trichomonas vaginalis*. *Parasitology Research*, 105, 1683–1689.
- Liu, N., Okamura, K., Tyler, D. M., Phillips, M. D., Chung, W.-J., & Lai, E. C. (2008). The evolution and functional diversification of animal microRNA genes. *Cell Research*, 18(10), 985–96.
- Liu, S., Li, D., Li, Q., Zhao, P., Xiang, Z., & Xia, Q. (2010). MicroRNAs of *Bombyx mori* identified by Solexa sequencing. *BMC Genomics*, 11, 148.
- Liu, S., Zhang, L., Li, Q., Zhao, P., Duan, J., Cheng, D., ... Xia, Q. (2009). MicroRNA expression profiling during the life cycle of the silkworm (*Bombyx mori*). *BMC Genomics*, 10, 455.
- Lozano, J., Montañez, R., & Belles, X. (2015). MiR-2 family regulates insect metamorphosis by controlling the juvenile hormone signaling pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 112(12), 2–7.
- Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., ... Wu, C.-I. (2008). The birth and death of microRNA genes in *Drosophila*. *Nature Genetics*, 40(3), 351–5.
- Lucas, K. J., Myles, K. M., & Raikhel, A. S. (2013). Small RNAs: A new frontier in mosquito biology. *Trends in Parasitology*, 6, 295–303.
- Lytle, J. R., Yario, T. A., & Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23), 9667–72.
- Lyu, Y., Shen, Y., Li, H., Chen, Y., Guo, L., Zhao, Y., ... Tang, T. (2014). New microRNAs in *Drosophila*--birth, death and cycles of adaptive evolution. *PLoS Genetics*, 10(1), e1004096.
- Ma, J., Flemr, M., Stein, P., Berninger, P., Malik, R., Zavolan, M., ... Schultz, R. M. (2010). MicroRNA Activity Is Suppressed in Mouse Oocytes. *Current Biology*, 20, 265–270.
- Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., ... Hatzigeorgiou, A. G. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research*, 37(Web Server issue), W273–6.
- Marco, A., Hooks, K., & Griffiths-Jones, S. (2012). Evolution and function of the extended miR-2 microRNA family. *RNA Biology*, 9(3), 242–248.
- Marco, A., Ninova, M., Ronshaugen, M., & Griffiths-Jones, S. (2013). Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Research*, 41, 7745–7752.
- Maxwell, E. K., Ryan, J. F., Schnitzler, C. E., Browne, W. E., & Baxeavanis, A. D. (2012). MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore *Mnemiopsis leidyi*. *BMC Genomics*, 13(1), 714.

- McBride, D., Carré, W., Sontakke, S. D., Hogg, C. O., Law, A., Donadeu, F. X., & Clinton, M. (2012). Identification of miRNAs associated with the follicular-luteal transition in the ruminant ovary. *Reproduction (Cambridge, England)*, 144(2), 221–33.
- McDonald, J. A., Pinheiro, E. M., & Montell, D. J. (2003). PVF1, a PDGF/VEGF homolog, is sufficient to guide border cells and interacts genetically with Taiman. *Development (Cambridge, England)*, 130(15), 3469–3478.
- Mead, E. A., & Tu, Z. (2008). Cloning, characterization, and expression of microRNAs from the Asian malaria mosquito, *Anopheles stephensi*. *BMC Genomics*, 9, 244.
- Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature Reviews. Genetics*, 14(7), 447–59.
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1), 31–46.
- Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., ... Kaessmann, H. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome Research*, 23(1), 34–45.
- Mi, S., Lu, J., Sun, M., Li, Z., Zhang, H., Neilly, M. B., ... Chen, J. (2007). MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 19971–19976.
- Mita, K. (2004). The Genome Sequence of Silkworm, *Bombyx mori*. *DNA Research*, 11(1), 27–35.
- Miyoshi, K., Miyoshi, T., & Siomi, H. (2010). Many ways to generate microRNA-like small RNAs: Non-canonical pathways for microRNA production. *Molecular Genetics and Genomics*, 284, 95–103.
- Montell, D. J. (2003). Border-cell migration: the race is on. *Nature Reviews. Molecular Cell Biology*, 4(1), 13–24.
- Monteys, A. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y., & Davidson, B. L. (2010). Structure and activity of putative intronic miRNA promoters. *RNA (New York, N.Y.)*, 16, 495–505.
- Moran, Y., Fredman, D., Praher, D., Li, X. Z., Wee, L. M., Rentzsch, F., ... Seitz, H. (2014). Cnidarian microRNAs frequently regulate targets by cleavage. *Genome Research*, 24(4), 651–63.
- Moran, Y., Praher, D., Fredman, D., & Technau, U. (2013). The evolution of MicroRNA pathway protein components in Cnidaria. *Molecular Biology and Evolution*, 30, 2541–2552.
- Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., & Proudfoot, N. J. (2008). Primary microRNA transcripts are processed co-transcriptionally. *Nature Structural & Molecular Biology*, 15, 902–909.
- Motzny, C. K., & Holmgren, R. (1995). The *Drosophila cubitus interruptus* protein and its role in the wingless and hedgehog signal transduction pathways. *Mechanisms of Development*, 52(1), 137–150.
- Mukherjee, K., & Vilcinskas, A. (2014). Development and immunity-related microRNAs of the lepidopteran model host *Galleria mellonella*. *BMC Genomics*, 15(1), 705.

- Nie, Q., Fang, M., Jia, X., Zhang, W., Zhou, X., He, X., & Zhang, X. (2011). Analysis of muscle and ovary transcriptome of *Sus scrofa*: Assembly, annotation and marker discovery. *DNA Research*, 18(5), 343–351.
- Ninova, M., Ronshaugen, M., & Griffiths-Jones, S. (2014a). Conserved Temporal Patterns of MicroRNA Expression in *Drosophila* Support a Developmental Hourglass Model. *Genome Biology and Evolution*, 6, 2459–2467.
- Ninova, M., Ronshaugen, M., & Griffiths-Jones, S. (2014b). Fast-evolving microRNAs are highly expressed in the early embryo of *Drosophila virilis*. *RNA (New York, N.Y.)*, 20(3), 360–72.
- Nozawa, M., Miura, S., & Nei, M. (2010). Origins and evolution of microRNA genes in *Drosophila* species. *Genome Biology and Evolution*, 2, 180–9.
- Obernosterer, G., Leuschner, P. J. F., Alenius, M., & Martinez, J. (2006). Post-transcriptional regulation of microRNA expression. *RNA (New York, N.Y.)*, 12(7), 1161–1167.
- Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., & Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, 130(1), 89–100.
- Oliver, J. C., Tong, X.-L., Gall, L. F., Piel, W. H., & Monteiro, A. (2012). A single origin for nymphalid butterfly eyespots followed by widespread loss of associated gene expression. *PLoS Genetics*, 8(8), e1002893.
- Panfilio, K. A. (2008). Extraembryonic development in insects and the acrobatics of blastokinesis. *Developmental Biology*, 313(2), 471–491.
- Papantonis, A., Swevers, L., & Iatrou, K. (2015). Chorion Genes: A Landscape of Their Evolution, Structure, and Regulation. *Annual Review of Entomology*, 60, 177–194.
- Paps, J., Xu, F., Zhang, G., & Holland, P. W. H. (2015). Reinforcing the egg-timer: Recruitment of novel Lophotrochozoa homeobox genes to early and late development in the Pacific oyster. *Genome Biology and Evolution*, 7(3), 677–688.
- Pasquinelli, A. E., Hunter, S., & Bracht, J. (2005). MicroRNAs: a developing story. *Current Opinion in Genetics & Development*, 15(2), 200–5.
- Patterson, L. T., Pembaur, M., & Potter, S. S. (2001). Hoxa11 and Hoxd11 regulate branching morphogenesis of the ureteric bud in the developing kidney. *Development (Cambridge, England)*, 128(11), 2153–61.
- Pavlopoulos, A., Kontarakis, Z., Liubicich, D. M., Serano, J. M., Akam, M., Patel, N. H., & Averof, M. (2009). Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33), 13897–902.
- Peterson, K. J., Dietrich, M. R., & McPeck, M. A. (2009). MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 31(7), 736–47. doi:10.1002/bies.200900033
- Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchellini, C., Boury-Esnault, N., ... Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, 19(8), 706–12. doi:10.1016/j.cub.2009.02.052
- Pigliucci, M. (2008). Is evolvability evolvable? *Nature Reviews. Genetics*, 9, 75–82.

- Pillai, R. S., Bhattacharyya, S. N., & Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? *Trends in Cell Biology*, 17(3), 118–26.
- Pinheiro, E. M., & Montell, D. J. (2004). Requirement for Par-6 and Bazooka in *Drosophila* border cell migration. *Development (Cambridge, England)*, 131(21), 5243–5251.
- Preuss, K. M., Lopez, J. a, Colbourne, J. K., & Wade, M. J. (2012). Identification of maternally-loaded RNA transcripts in unfertilized eggs of *Tribolium castaneum*. *BMC Genomics*, 13(1), 671.
- Prochnik, S. E., Rokhsar, D. S., & Aboobaker, A. A. (2007). Evidence for a microRNA expansion in the bilaterian ancestor. *Development Genes and Evolution*, 217(1), 73–7.
- Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., ... Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198), 1064–71.
- Quah, S., Hui, J. H. L., & Holland, P. W. H. (2015). A Burst of miRNA Innovation in the Early Evolution of Butterflies and Moths. *Molecular Biology and Evolution*, 32(5), 1161–1174.
- Ramaswamy, S. B., Shu, S., Park, Y. I., & Zeng, F. (1997). Dynamics of juvenile hormone-mediated gonadotropism in the Lepidoptera. *Archives of Insect Biochemistry and Physiology*, 35, 539–558.
- Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., ... Hayashizaki, Y. (2010). An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell*, 140(5), 744–752. doi:10.1016/j.cell.2010.01.044
- Regier, J. C., Mitter, C., Zwick, A., Bazinet, A. L., Cummings, M. P., Kawahara, A. Y., ... Mitter, K. T. (2013). A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PloS One*, 8(3), e58568.
- Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., ... Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403(6772), 901–6.
- Robinson, J. M., Sperling, E. A., Bergum, B., Adamski, M., Nichols, S. A., Adamska, M., & Peterson, K. J. (2013). The Identification of MicroRNAs in Calcisponges: Independent Evolution of MicroRNAs in Basal Metazoans. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 320(2), 84–93.
- Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Research*, 14(10A), 1902–10.
- Ross, A. J., & Capel, B. (2005). Signaling at the crossroads of gonad development. *Trends in Endocrinology and Metabolism*. 16(1), 19–25.
- Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10), 1117–24.
- Ruby, J. G., Jan, C. H., & Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149), 83–6.
- Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., & Lai, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Research*, 17(12), 1850–64.

- Saini, H. K., Griffiths-Jones, S., & Enright, A. J. (2007). Genomic analysis of human microRNA transcripts. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45), 17719–24.
- Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews. Genetics*, 8(6), 424–436.
- Sanson, B., Alexandre, C., Fascetti, N., & Vincent, J.-P. (1999). Engrailed and Hedgehog Make the Range of Wingless Asymmetric in *Drosophila* Embryos. *Cell*, 98(2), 207–216.
- Saunders, C., & Cohen, R. S. (1999). The role of oocyte transcription, the 5'UTR, and translation repression and derepression in *Drosophila* gurken mRNA and protein localization. *Molecular Cell*, 3(1), 43–54.
- Schmiedel, J. M., Klemm, S. L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D. S., & van Oudenaarden, A. (2015). MicroRNA control of protein expression noise. *Science*, 348, 128–132.
- Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., & Ast, G. (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biology*, 8(6), R127.
- Shabalina, S. A., & Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. *Trends in Ecology & Evolution*, 23(10), 578–87.
- Shkumatava, A., Stark, A., Sive, H., & Bartel, D. P. (2009). Coherent but overlapping expression of microRNAs and their targets during vertebrate development. *Genes and Development*, 23, 466–481.
- Shomron, N., & Levy, C. (2009). MicroRNA-biogenesis and Pre-mRNA splicing crosstalk. *Journal of Biomedicine & Biotechnology*, 2009, 594678.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539.
- Smalheiser, N. R., & Torvik, V. I. (2005). Mammalian microRNAs derived from genomic repeats. *Trends in Genetics*, 21(6), 322–6.
- Sokol, N. S. (2005). Mesodermally expressed *Drosophila* microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes & Development*, 19(19), 2343–2354.
- Sperling, E. A., Peterson, K. J., & Pisani, D. (2009). Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Molecular Biology and Evolution*, 26(10), 2261–74.
- Stark, A., Brennecke, J., Bushati, N., Russell, R. B., & Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6), 1133–46.
- Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., & Blelloch, R. (2010). MicroRNA Function Is Globally Suppressed in Mouse Oocytes and Early Embryos. *Current Biology*, 20, 271–277.
- Surridge, A. K., Lopez-Gomollon, S., Moxon, S., Maroja, L. S., Rathjen, T., Nadeau, N. J., ... Jiggins, C. D. (2011). Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene*. *BMC Genomics*, 12(1), 62.

- Suzuki, H. I., Yamagata, K., Sugimoto, K., Iwamoto, T., Kato, S., & Miyazono, K. (2009). Modulation of microRNA processing by p53. *Nature*, 460(7254), 529–533.
- Suzuki, S., Shaw, G., Kaneko-Ishino, T., Ishino, F., & Renfree, M. B. (2011). The evolution of mammalian genomic imprinting was accompanied by the acquisition of novel CpG islands. *Genome Biology and Evolution*, 3(1), 1276–1283.
- Swevers, L., & Iatrou, K. (2003). The ecdysone regulatory cascade and ovarian development in lepidopteran insects: Insights from the silkworm paradigm. *Insect Biochemistry and Molecular Biology*, 33(12), 1285–1297.
- Swevers, L., Ito, K., & Iatrou, K. (2002). The BmE75 nuclear receptors function as dominant repressors of the nuclear receptor BmHR3A. *Journal of Biological Chemistry*, 277(44), 41637–41644.
- Tabin, C. J. (1992). Why we have (only) five fingers per hand: hox genes and the evolution of paired limbs. *Development (Cambridge, England)*, 116(2), 289–96.
- Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development (Cambridge, England)*, 136(18), 3033–3042.
- Taguchi, A., Kaku, Y., Ohmori, T., Sharmin, S., Ogawa, M., Sasaki, H., & Nishinakamura, R. (2014). Redefining the in vivo origin of metanephric nephron progenitors enables generation of complex kidney structures from pluripotent stem cells. *Cell Stem Cell*, 14(1), 53–67.
- Tang, F., Kaneda, M., O'Carroll, D., Hajkova, P., Barton, S. C., Sun, Y. A., ... Surani, M. A. (2007). Maternal microRNAs are essential for mouse zygotic development. *Genes and Development*, 21, 644–648.
- Tanzer, A., Amemiya, C. T., Kim, C.-B., & Stadler, P. F. (2005). Evolution of microRNAs located within Hox gene clusters. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, 304(1), 75–85.
- Tarver, J. E., Donoghue, P. C. J., & Peterson, K. J. (2012). Do miRNAs have a deep evolutionary history? *BioEssays*, 34(10), 857–866.
- Tarver, J. E., Sperling, E. A., Nailor, A., Heimberg, A. M., Robinson, J. M., King, B. L., ... Peterson, K. J. (2013). miRNAs: small genes with big potential in metazoan phylogenetics. *Molecular Biology and Evolution*, 30(11), 2369–82.
- Tarver, J. E., Cormier, A., Pinzón, N., Taylor, R. S., Carré, W., Strittmatter, M., Seitz, H., Coelho, S. M., & Cock, J. M. (2015). microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga *Ectocarpus*. *Nucleic Acids Research*, 43(13), 6384–6398.
- Tautz, D. & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12, 692–702.
- Tay, Y., Zhang, J., Thomson, A. M., Lim, B., & Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216), 1124–8.
- Taylor, H. S., Vanden Heuvel, G. B., & Igarashi, P. (1997). A conserved Hox axis in the mouse and human female reproductive system: late establishment and persistent adult expression of the Hoxa cluster genes. *Biology of Reproduction*, 57(6), 1338–45.
- Telfer, W. H. (2009). Egg formation in Lepidoptera. *Journal of Insect Science*, 9, 1–21.

- The Heliconius Genome Consortium. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405), 94–8.
- The UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1), D204–D212.
- Tomari, Y., & Zamore, P. D. (2005). MicroRNA biogenesis: Drosha can't cut it without a partner. *Current Biology*, 15(2), R61–4.
- Torley, K. J., da Silveira, J. C., Smith, P., Anthony, R. V., Veeramachaneni, D. N. R., Winger, Q. A., & Bouma, G. J. (2011). Expression of miRNAs in ovine fetal gonads: potential role in gonadal differentiation. *Reproductive Biology and Endocrinology*, 9, 2.
- Tsang, J., Zhu, J., & van Oudenaarden, A. (2007). MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals. *Molecular Cell*, 26, 753–767.
- Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220), 1260419–1260419.
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220), 1260419
- Vainio, S., & Lin, Y. (2002). Coordinating early kidney development: lessons from gene targeting. *Nature Reviews. Genetics*, 3(7), 533–543.
- Wahlberg, N., Wheat, C. W., & Peña, C. (2013). Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PloS One*, 8(11), e80875.
- Wanek, N., Muneoka, K., Holler-Dinsmore, G., Burton, R., & Bryant, S. V. (1989). A staging system for mouse limb development. *The Journal of Experimental Zoology*, 249(1), 41–49.
- Wang, X., & El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics (Oxford, England)*, 24(3), 325–32.
- Warren, W. C., Hillier, L. W., Marshall Graves, J. A., Birney, E., Ponting, C. P., Grützner, F., ... Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192), 175–183.
- Wei, J., Hortsch, M., & Goode, S. (2004). Neuroglial stabilizes epithelial structure during *Drosophila* oogenesis. *Developmental Dynamics*, 230(4), 800–808.
- Wellik, D. M. (2011). Hox genes and kidney development. *Pediatric Nephrology (Berlin, Germany)*, 26(9), 1559–65.
- Westholm, J. O., & Lai, E. C. (2011). Mirtrons: MicroRNA biogenesis via splicing. *Biochimie*, 93(11), 1897–1904.
- Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. a, Holstein, T. W., Heber, S., & Peterson, K. J. (2009). The deep evolution of metazoan microRNAs. *Evolution & Development*, 11(1), 50–68.
- White, K. P., Hurban, P., Watanabe, T., & Hogness, D. S. (1997). Coordination of *Drosophila* metamorphosis by two ecdysone-induced nuclear receptors. *Science (New York, N.Y.)*, 276(5309), 114–117.

- Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5), 855–62.
- Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews. Genetics*, 13(1), 59–69.
- Woltering, J. M., & Durston, A. J. (2008). MiR-10 Represses HoxB1a and HoxB3a in Zebrafish. *PLoS ONE*, 3(1).
- Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., ... Yang, H. (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science (New York, N.Y.)*, 306(5703), 1937–40.
- Xu, F., Stouffer, R. L., Müller, J., Hennebold, J. D., Wright, J. W., Bahar, A., ... Lindenthal, B. (2011). Dynamics of the transcriptome in the primate ovulatory follicle. *Molecular Human Reproduction*, 17(3), 152–165.
- Yekta, S., Shih, I.-H., & Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science (New York, N.Y.)*, 304(5670), 594–6.
- Yekta, S., Tabin, C. J., & Bartel, D. P. (2008). MicroRNAs in the Hox network: an apparent link to posterior prevalence. *Nature Reviews Genetics*, 9, 789–796.
- You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., ... Wang, J. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics*, 45(2), 220–5.
- Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The monarch butterfly genome yields insights into long-distance migration. *Cell*, 147(5), 1171–85.
- Zhang, X., Zheng, Y., Jagadeeswaran, G., Ren, R., Sunkar, R., & Jiang, H. (2012). Identification and developmental profiling of conserved and novel microRNAs in *Manduca sexta*. *Insect Biochemistry and Molecular Biology*, 42(6), 381–395.
- Zhang, X.-D., Zhang, Y.-H., Ling, Y.-H., Liu, Y., Cao, H.-G., Yin, Z.-J., ... Zhang, X.-R. (2013). Characterization and differential expression of microRNAs in the ovaries of pregnant and non-pregnant goats (*Capra hircus*). *BMC Genomics*, 14(1), 157.
- Zhou, X., Ruan, J., Wang, G., & Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS Computational Biology*, 3(3), 0412–0423.
- Zovoilis, A., Pantazi, A., Smorag, L., Opitz, L., Riester, G. S., Wolf, M., ... Engel, W. (2010). Embryonic stem cell-related miRNAs are involved in differentiation of pluripotent cells originating from the germ line. *Molecular Human Reproduction*, 16(11), 793–803.