Homeobox Gene and MicroRNA Duplication in Chelicerata

Daniel J. Leite

Degree awarded by
Oxford Brookes University

This thesis is submitted in partial fulfillment of the requirements
of the award of Doctor of Philosophy

August 2017

Acknowledgements

So far as people, I thank Alistair McGregor greatly for giving me the opportunity to be emersed in the realm of science. I hugely appreciate the generosity of Nigel Groome, and the studentship scheme that helped fund me through the PhD. My spider buddy Anna Schönauer has always been a great companion, for scientific chats, jokes, and keeping me somewhat sane. The write up time was made bareable with the love and kindness of Eirini Kapsala, thank you. There are many others both in the McGregor lab, from other labs and around Oxford Brookes who have been great people to be around for the PhD. I deeply thank my parents Linda and Brian Leite, who gave me freedom and ecouragement to be myself.

I would also like to mention some music and musicians that became staples in my diet during the PhD. All the bioinformatics was accompanied by Pink Floyd, the power of your music lifted me from my computer. For the thesis write up, my companion was John Martyn, your voice and soul is unmatched, you gave me feelings when I had none. Keeping a groove through it all was Parliament-Funkadelic, if ever I was lethargic you put me back in time.

I have to acknowledge Sci-Hub, which without I would not have been able to research and learn from others. It is appauling that we hide knowledge from one another yet commonly say education should become accessable to all. Especially in science, the freedom of information is essential to drive the progression of knowledge. I hope that the community changes this promptly.

Lastly and most importantly, I would like to give my respect to life, especially the spiders that were tortured and killed by me during the PhD. It is fascinating to be a part of this increadible process of life and to have a mind to study it. Sadly though, I feel our desire to obtain glimpses of understanding outways our desire to make sure it will be around to continue studying in generations to come. Four billion years have generated more than we will ever understand but seemingly not more than we can destroy.

Abstract

Animals have evolved a vast array of morphological phenotypes that has probably been generated by the evolution of core developmental toolkit genes. The expansion of these toolkits via gene duplications is correlated with the evolution of multi-cellularity and morphological complexity in metazoans. Small tandem duplications are a common source of additional genetic material in animals, while larger-scale events are relatively rarer and until recently whole genome duplications (WGDs) had only been described in vertebrates and rotifers. Interestingly though, recent research in chelicerates, a sister-group to the other arthropods, has provided compelling evidence of WGDs in the lineages of Xiphosura (horseshoe crabs) and Arachnopulmonata (spiders, scorpions). To gain further support and insights to these possible WGD events, homeobox genes and microRNAs (miRNAs) were systematically identified and characterised in chelicerates with evidence for WGDs, as well as lineages with no data supporting WGD. This identified pervasive duplication of homeobox genes and miRNAs in arachnopulmonates, which may have been caused by a WGD event that is independent of those in xiphosurans. Analysis of the spider *Parasteatoda tepidariorum* genome identified duplicate Hox, NK, SINE, TALE and LIM clusters and several miRNA clusters, providing extensive synteny support for a WGD in this lineage. Furthermore, the expression patterns of homeobox paralogs have diverged during embryogenesis and may represent cases of sub- and neo- functionalisation. Additionally, the mir-3791 family in *P. tepidariorum*, has expanded due to the proposed WGD and further tandem duplications. This family has orthology to the insect and nematode miRNAs that are involved in the maternal to zygotic transition (MZT). A subset of *Pt-mir-3791* paralogs were also predicted to be involved in the MZT in the spider, possibly indicating a more ancestral role of this miRNA family than previously thought. These results offer additional support for WGDs in chelicerates as well as regulatory divergences between paralogs that may correspond to morphological evolution. These findings provide an excellent platform to further compare vertebrate and invertebrate lineages that have independently experienced WGDs and to understand the role of these events in the evolution of animal diversity.

Table of Contents

ACKNOWLEDGEMENTS **1**

ABSTRACT **3**

TABLE OF CONTENTS **5**

LIST OF FIGURES **8**

LIST OF TABLES **9**

ABBREVIATIONS **10**

PUBLICATIONS **11**

1 INTRODUCTION **13**

1.1 DEVELOPMENT AND THE EVOLUTION OF ANIMAL MORPHOLOGY 13

1.1.1 Genetic Toolkits in Development *13*

1.1.2 Evolution of Developmental Networks *13*

1.2 HOMEOBOX GENES 15

1.2.1 The Homeodomain *15*

1.2.2 Roles in Development *16*

1.3 MICRORNAS 19

1.3.1 Biogenesis and Function *19*

1.3.2 Roles in Development *19*

1.4 GENE DUPLICATION 21

1.4.1 Evolution after Gene Duplication *21*

1.4.2 Expansion of Homeobox containing Genes in Metazoans *21*

1.4.3 Expansion of miRNAs in Metazoans *23*

1.4.4 Gene Duplication in Chelicerata *24*

1.5 CHELICERATES 26

1.5.1 Phylogenetics of Chelicerata *26*

1.5.2 Resources for Studying Gene Duplication in Chelicerates *29*

1.6 AIMS 32

1.6.1 Duplication of Homeobox Genes in Chelicerata *32*

1.6.2 Duplication of miRNAs in Chelicerata *33*

1.6.3 Developmental Roles of Duplicated miRNAs *34*

2 MATERIALS AND METHODS **35**

2.1 BIOINFORMATICS 35

2.1.1 Analysis of RNA Sequencing from P. tepidariorum Embryogenesis *35*

2.1.2 Identification and Characterisation of Homeobox Genes *35*

2.1.3 P. tepidariorum Culture and Small RNA Sequencing *36*

2.1.4 Analysis of Small RNA Sequencing Data and Identification of P. tepidariorum miRNAs 36
2.1.5 Identification of P. tepidariorum miRNA Orthologs in Other Species *37*
2.1.6 Quantification of Reads Associated with miRNAs in P. tepidariorum *38*
2.1.7 Comparisons of Relative Arm Usage *39*
2.1.8 Identification of miRNA Targets During Maternal to Zygotic Transition (MZT) *40*
2.2 MOLECULAR BIOLOGY 40
2.2.1 Cloning of Gene Fragments and Probe Synthesis *40*
2.2.2 In Situ Hybridization (ISH) *41*

3 DUPLICATION OF HOMEOBOX GENES IN SPIDERS AND SCORPIONS 43
3.1 COMPARISON OF HOMEOBOX FAMILIES IN CHELICERATES AND OTHER ARTHROPODS 44
3.2 PERVASIVE DUPLICATION OF HOMEOBOX GENES IN ARACHNOPULMONATES 46
3.2.1 Arachnopulmonates Share Many Duplicated Homeobox Families *46*
3.2.2 Duplicated Hox clusters in Arachnopulmonata *46*
3.2.3 WGDs and Tandem Duplications of Homeobox Genes in P. tepidariorum *50*
3.3 EXPRESSION OF HOMEOBOX GENES IN A *P. TEPIDARIORUM* EMBRYOGENESIS 52
3.4 EXPRESSION DIVERGENCE OF DUPLICATED *P. TEPIDARIORUM* HOMEOBOX GENES IN THE EMBRYONIC TRANSCRIPTOME 52
3.4.1 Divergence of Dispersed Paralogs in a P. tepidariorum Embryogenesis *53*
3.4.2 Divergence of Tandem Paralogs in a P. tepidariorum Embryogenesis *55*
3.5 *IN SITU* HYBRIDISATION (ISH) OF *P. TEPIDARIORUM* HOMEOBOX GENES REVEALS DIVERGENT EXPRESSION PATTERNS 57
3.5.1 Cux Genes *57*
3.5.2 Hmx Genes *60*
3.5.3 Pitx Genes *63*
3.5.4 Zfh Genes *66*
3.6 DISCUSSION 70
3.6.1 Homeobox Repertoires in Chelicerates *70*
3.6.2 Possible Ancient Whole Genome Duplication in the Ancestor of Arachnopulmonates 71
3.6.3 Divergence in the Expression of Homeobox Paralogs *72*
3.6.4 Duplication and Divergence of Homeobox Genes in Arachnopulmonates *75*

4 DUPLICATION OF MIRNAS IN XIPHOSURA AND ARACHNOPULMONATA 75
4.1 MICRORNA PREDICTION AND ANNOTATION IN *P. TEPIDARIORUM* 78
4.2 IDENTIFICATION OF CONSERVED *P. TEPIDARIORUM* MIRNA FAMILIES IN OTHER CHELICERATES 81
4.3 NOVEL *P. TEPIDARIORUM* MIRNAS ARE LARGELY SPECIES-SPECIFIC 82
4.4 INCIDENCE OF MULTI COPY MIRNA FAMILIES IN OTHER CHELICERATES 82
4.5 SYNTENY OF MIRNA DUPLICATIONS 83
4.5.1 Small and Large Scale Duplications of miRNAs in P. tepidariorum *83*
4.5.2 Small and Large Scale Duplications of miRNAs in Other Chelicerates *84*
4.6 DIVERGENCE IN ARM USAGE IN MULTI COPY MIRNAS 88

4.7 Discussion 90
4.7.1 The Evolution of the P. tepidariorum miRNA Repertoire 90
4.7.2 Evidence of Duplication of both Coding and Noncoding Genes in P. tepidariorum 90
4.7.3 Duplication and Divergence of miRNAs in Multiple Chelicerate Lineages 91
4.7.4 Evolution of miRNA Function 92

5 THE PREDICTED ROLE OF MIRNAS IN THE MATERNAL TO ZYGOTIC TRANSITION IN *P. TEPIDARIORUM* 93
5.1 Identification of MZT in *P. tepidariorum* 94
5.1.1 Transitions in the Embryonic mRNA Transcriptome of P. tepidariorum 94
5.1.2 Early Transitions in the Embryonic miRNA Transcriptome of P. tepidariorum 95
5.2 Predicting miRNA Targets at Phase II of the MZT in *P. tepidariorum* 97
5.3 Paralogs of *miR-3791* and *mir-184* are Predicted to Regulate MZT 99
5.4 Predicted Targets that are also Maternally Loaded in Other Species 101
5.4.1 Many Targets are Maternally Supplied in Other Metazoans 101
5.4.2 Predicted Role of mex-3 in Axis Specification of P. tepidariorum 101
5.5 Discussion 103
5.5.1 Timing of MZT Corresponds with Germ Disc Formation and Gastrulation 103
5.5.2 Ecdysozoans Share Common Clustered miRNAs that Regulate the MZT 104
5.5.3 Maternal Supplied mex3 may help to set up Egg Axes in P. tepidariorum 104
5.5.4 Other Mechanisms of Maternal mRNA Clearance 105

6 GENERAL DISCUSSION AND FUTURE DIRECTIONS 107
6.1 Duplication of Homeobox Genes in Chelicerates 107
6.2 Duplication of miRNAs in Chelicerates 110
6.3 MicroRNAs in the MZT of *P. tepidariorum* 111
6.4 Conclusion 114

REFERENCES 115

APPENDIX 129

List of Figures

Figure 1.1: Homeobox genes encode a homeodomain and a variety of other functional domains. ... 17
Figure 1.2: Body plans and known Hox gene expression in chelicerate orders. ... 18
Figure 1.3: An overview of miRNA biogenesis and effector pathways. ... 20
Figure 1.4: Evolution of the hypothetical metazoan "Giga-homeobox cluster". ... 22
Figure 1.5: Phylogeny and Hox repertoire of chelicerates and myriapods. ... 28
Figure 1.6: The development and life cycle of the spider *Parasteatoda tepidariorum*. ... 30
Figure 2.1: MicroRNA prediction and annotation in *P. tepidariorum*. ... 37
Figure 2.2: Detection of *P. tepidariorum* miRNA orthologs in other chelicerate species. ... 39
Figure 2.3: Quantification of reads assigned to miRNAs, CDS and other regions of the *P. tepidariorum* genome. ... 39
Figure 3.1: Comparison of homeobox repertoires in arthropods reveals pervasive duplication in arachnopulmonates. ... 45
Figure 3.2: Duplication of Hox clusters in arachnopulmonates. ... 48
Figure 3.3: Duplicated Hox clusters in *P. tepidariorum*. ... 49
Figure 3.4: Homeobox gene clustering in the *P. tepidariorum* genome. ... 51
Figure 3.5: Expression of homeobox genes in *P. tepidariorum* expressed from S1 to S10. ... 54
Figure 3.6: *Cux* genes in arthropods. ... 58
Figure 3.7: Expression of *Cux* paralogs in *P. tepidariorum*. ... 59
Figure 3.8: *Hmx* genes in arthropods. ... 60
Figure 3.9: Expression of *Hmx* paralogs in *P. tepidariorum*. ... 62
Figure 3.10: *Pitx* genes in arthropods. ... 63
Figure 3.11: Expression of *Pitx* paralogs in *P. tepidariorum*. ... 65
Figure 3.12: *Zfh* genes in arthropods. ... 67
Figure 3.13: Expression of *Zfh* paralogs in *P. tepidariorum*. ... 69
Figure 3.14: Divergence in the expression of *P. tepidariorum* homeobox gene paralogs. ... 74
Figure 4.1: Distribution and assignment of pooled small RNA-Seq from *P. tepidariorum*. ... 79
Figure 4.2: MicroRNA family copy number in *P. tepidariorum* and other ecdysozoans. ... 82
Figure 4.3: Clusters of miRNA paralogs in the *P. tepidariorum* DoveTail genome assembly. ... 85
Figure 4.4: Duplication of miRNA clusters in *P. tepidariorum* and other chelicerates ... 87
Figure 4.5: Relative arm usage changes in miRNAs. ... 89
Figure 5.1: MicroRNA and mRNA expression in *P. tepidariorum* development. ... 95
Figure 5.2: Early stages of miRNA and mRNA expression in *P. tepidariorum* separate from later stages. ... 96
Figure 5.3: *Mir-3791* cluster contains paralogs predicted to regulate MZT. ... 100
Figure 5.4: Regulation of the predicted axis patterning gene *mex3* in *P. tepidariorum*. ... 102

Appendix Figure 3.1: Maximum likelihood tree of the homeodomain from ANTP class homeobox genes. ... 134
Appendix Figure 3.3: Maximum likelihood tree of homeodomains from homeobox genes, excluding ANTP and PRD classes. ... 135
Appendix Figure 3.4: Maximum likelihood tree of *P. tepidariorum* homeobox genes. ... 136
Appendix Figure 3.5: Maximum likelihood tree of *P. phalangioides* homeobox genes. ... 137
Appendix Figure 3.6: Maximum likelihood tree of *C. sculpturatus* homeobox genes. ... 138
Appendix Figure 3.7: Maximum likelihood tree of *I. scapularis* homeobox genes. ... 139
Appendix Figure 4.1: Distribution of number of miRNAs per conserved miRNA family. ... 145
Appendix Figure 4.2: Dispersed paralogs of miRNAs in the *P. tepidariorum* DoveTail genome assembly. ... 146
Appendix Figure 5.1: MEX3 alignment from *P. tepidariorum*, *T. castaneum*, *C. elegans*, *Ciona savignyi* and humans. ... 147
Appendix Figure 5.2: Alignments of the *PCPB1 Homo sapiens* PolyC-binding protein to *P. tepidariorum* proteins (subject). ... 148

List of Tables

Table 2.1: Primers used for cloning DNA fragments for probe synthesis. ... 41
Table 4.1: Mapping of small RNA sequencing to *P. tepidariorum* genome ... 78
Table 4.2: MicroRNA annotation and copy number in *P. tepidariorum* ... 80
Table 5.1: Differentially regulated genes in *P. tepidariorum* development ... 94
Table 5.2: Predicted miRNAs and their mRNA targets in *P. tepidariorum* MZT ... 98

Appendix Table 3.1: *P. tepidariorum* homeobox genes ... 130
Appendix Table 3.2: *P. phalangioides* homeobox genes ... 130
Appendix Table 3.3: *C. sculpturatus* homeobox genes ... 132
Appendix Table 3.4: *I. scapularis* homeobox genes ... 133
Appendix Table 3.5: MicroRNAs and non-Hox genes in *P. tepidariorum* Hox cluster scaffolds ... 140
Appendix Table 4.1: *P. tepidariorum* hairpin miRNA sequences ... 141
Appendix Table 4.2: *P. tepidariorum* mature miRNA sequences ... 143

Abbreviations

AP	**Anterior-posterior**
ATAC-Seq	Assay for transposase-accessible chromatin sequencing
Ch	Chelicerae
ChIP-Seq	Chromatin immunoprecipitation sequencing
DAPI	4′,6-Diamidino-2-phenylindole dihydrochloride
DNA	Deoxyribonucleic acid
eRNAi	Embryonic RNA interference
GRN	Gene regulatory network
ISH	*in situ* hybridisation
L	Walking leg
miRNA	MicroRNA
mRNA	Messenger RNA
MYA	Million years ago
MZT	Maternal to zygotic transition
ncRNA	Non coding RNA
O	Opisthosoma
piRNA	Piwi-interacting RNA
Pp	Pedipalp
pre-miRNA	Precursor microRNA
pri-miRNA	Primary microRNA
pRNAi	Parental RNA interference
RISC	RNA silencing complex
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
rRNA	Ribosomal RNA
S	Stage
tRNA	Transfer RNA
UTR	Untranslated region
WGD	Whole genome duplication

Publications

Schwager EE[†], Sharma PP[†], Clarke T[†], **Leite DJ**[†], Wierschin T[†], Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, Buffry AD, Chao H, Dinh H, Doddapaneni H, Dugan S, Eibner C, Extavour CG, Funch P, Garb J, Gonzalez LB, Gonzalez VL, Griffiths-Jones S, Han Y, Hayashi C, Hilbrant M, Hughes DST, Janssen R, Lee SL, Maeso I, Murali SC, Muzny DM, Nunes da Fonseca R, Paese CLB, Qu J, Ronshaugen M, Schomburg C, Schönauer A, Stollewerk A, Torres-Oliva M, Turetzek N, Vanthournout B, Werren JH, Wolff C, Worley KC, Bucher G, Gibbs RA, Coddington J, Oda H, Stanke M, Ayoub NA, Prpic N-M, Flot J-F, Posnien N, Richards S, McGregor AP. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*. 15(1).

[†] Equal contribution

Leite DJ, McGregor AP. 2016. Arthropod evolution and development: recent insights from chelicerates and myriapods. *Curr Opin Genet Dev*. 3993-100.

Leite DJ, Ninova M, Hilbrant M, Arif S, Griffiths-Jones S, Ronshaugen M, McGregor AP. 2016. Pervasive microRNA Duplication in Chelicerates: Insights from the Embryonic microRNA Repertoire of the Spider *Parasteatoda tepidariorum*. *Genome Biol Evol*. 8(7): 2133-44.

Schönauer A, Paese CL, Hilbrant M, **Leite DJ**, Schwager EE, Feitosa NM, Eibner C, Damen WG, McGregor AP. 2016. The Wnt and Delta-Notch signalling pathways interact to direct pair-rule gene expression via caudal during segment addition in the spider *Parasteatoda tepidariorum*. *Development*. 143(13): 2455-63.

Schwager EE, Schoenauer A, **Leite DJ**, Sharma PP, McGregor AP. 2015. Chelicerata. *In:* Wanninger, A (ed.) *Evolutionary Developmental Biology of Invertebrates.* Spinger.

Hilbrant M, Almudi I, **Leite DJ**, Kuncheria L, Posnien N, Nunes MD, McGregor AP. 2014. Sexual dimorphism and natural variation within and among species in the *Drosophila* retinal mosaic. *BMC Evol Biol.* 14(240): 1-13.

Jacobs LA, Bewicke-Copley F, Poolman MG, Pink RC, Mulcahy LA, Baker I, Beaman EM, Brooks T, Caley DP, Cowling W, Currie JM, Horsburgh J, Kenehan L, Keyes E, **Leite D**, Massa D, McDermott-Rouse A, Samuel P, Wood H, Kadhim M, Carter DR. 2013. Meta-analysis using a novel database, miRStress, reveals miRNAs that are frequently associated with the radiation and hypoxia stress-responses. *PLoS One*. 8(11): e80844.

1 INTRODUCTION

1.1 Development and the Evolution of Animal Morphology

1.1.1 Genetic Toolkits in Development

The process of evolution has given rise to many different forms of life ranging from single celled prokaryotes to complex multicellular organisms. Some lineages of plants, fungi and animals have independently evolved the extraordinary process of development, which sees single cells go through precisely orchestrated proliferation and differentiation to build organisms of hugely varying morphological diversity. Many of the core gene regulatory networks (GRNs) utilised in development are conserved between species (Rokas 2008). The types of genes involved are often transcription factors (TFs) such as homeobox genes and signalling pathways such as Wnt and Delta/Notch (Randazzo *et al.* 1991; Sidow 1992; Rothbächer *et al.* 1995; Abzhanov and Kaufman 1999; Onuma *et al.* 2001; Carroll *et al.* 2005; Schenkelaars *et al.* 2017). These toolkit genes often provide various patterning instructions such as morphogen gradients and oscillations, and cell-cell communications, and can be co-opted in different contexts to regulate the development of a variety of organs, tissues and structures (Carroll *et al.* 2005; Korkut and Budnik 2009; Werner *et al.* 2010; Krol *et al.* 2011; Koshikawa *et al.* 2015). In some cases the network of interactions between toolkits genes have been mapped out into large system wide contexts (Levine and Davidson 2005). These have provided a framework of known interactions between TFs and their *cis*-regulatory binding sites in downstream target genes. However, during development, the transcriptional profile changes in time and space and therefore different subsets of these networks are employed as development progresses (Arif *et al.* 2015; Clark and Akam 2016). This is achieved through modular elements in *cis*-regulatory region that bind different TFs to regulate the same gene at different times and places (Gompel *et al.* 2005; McGregor *et al.* 2007). It is thought that this modularity in part facilitates the evolution of these toolkit GRNs can therefore alter development and consequently the adult phenotype to produce huge variety in morphology (McGregor *et al.* 2007; Kvon *et al.* 2016; Gaiti *et al.* 2017; Halfon 2017). However, note that in the case of developmental systems drift, GRNs can also be rewired differently but still produced the same phenotypic output (Frankel *et al.* 2012; Sommer 2012; Halfon 2017).

1.1.2 Evolution of Developmental Networks

Evolution of developmental programs can be associated with changes in many different elements within GRNs (Halfon 2017).

Orthologous protein coding genes in these networks can have high sequence and functional conservation between species due to negative selection on functional domains (Rokas

and Carroll 2008). For example the functional domains of transcription factors, such as the homeodomain found in the superclass of homeobox genes, can have many DNA binding sites in *cis*-regulatory regions of downstream genes and interacting co-factors (Ortiz-Lombardia *et al.* 2017). Consequently changes in the functional domains of protein coding genes like these will be highly pleiotropic and can drastically affect the function and regulation of many downstream genes, often resulting in deleterious consequences. However changes in the coding regions of genes can also alter GRNs and give rise to phenotypic change (Pick and Heffer 2012). For example in insects the homeobox gene *Ultrabithorax* (*Ubx*) has evolved a repressor domain that inhibits *Distal-less* (*Dll*) in the abdomen and therefore is associated with the lack of limbs on this tagma (Vachon *et al.* 1992; Ronshaugen *et al.* 2002). In contrast, the crustacean *Ubx* does not contain the repressor motif and therefore *Dll* is able to regulate the formation of limbs on the abdomen (Vachon *et al.* 1992; Ronshaugen *et al.* 2002).

Alternatively, *cis*-regulatory regions can evolve so that a gene can be regulated in different cellular contexts, such as the seven stripes of the homeobox gene *even-skipped* or pigmentation genes in *Drosophila* (Gompel *et al.* 2005; Prud'homme *et al.* 2006; Prud'homme *et al.* 2007). The *cis*-regulatory enhancers can be considered as modular input components to the regulation of a gene. However, unlike the logic of triplet codons encoding for amino acids, the syntax rules of enhancer functions are more degenerate and varied (Nirenberg and Leder 1964; McGregor *et al.* 2007; Farley *et al.* 2015; Buffry *et al.* 2016; Farley *et al.* 2016). So that the spacing, direction, affinity and arrangement of transcription factors binding to enhancers can produce subtle changes in transcription rates, levels and spatio-temporal expression (Farley *et al.* 2015; Farley *et al.* 2016). The degenerate and complex nature of enhancer function also facilitates greater DNA substitutions, as mutations do not necessarily have as great an effect on regulation as non-synonymous mutations in coding domains (Prud'homme *et al.* 2007). These characteristics of enhancers mean that there can be relatively few genes in a genome, but each gene can evolve to be precisely regulated in many different ways during development (Prud'homme *et al.* 2006). Evolution of the *cis*-regulatory logic to a gene can affect its expression and therefore alter morphologies. For example, the expression of *Sonic hedgehog* (*Shh*) in developing mouse limbs is dependent on the interaction with a long-range *cis*-regulatory module (Lettice *et al.* 2003; Kvon *et al.* 2016). This module is highly conserved across vertebrates that possess limbs, however, in the snakes a 17 bp deletion in the long-range enhancer causes a loss of the binding site for the transcription factor ETS1 (Kvon *et al.* 2016). Insertion of the snake specific enhancer into the endogenous mouse enhancer causes a loss of limbs, highlighting the evolution of this enhancer and its crucial role in development (Kvon *et al.* 2016). Furthermore, changes in the *cis*-regulatory regions of a key regulator of trichome development, *shavenbaby,* has resulted in different trichome patterning in the larva between species of *Drosophila* (McGregor *et al.* 2007; Frankel *et al.* 2012).

The intergenic regions of genomes also host many transcribed loci that produce functional non-coding RNA (ncRNA) products (Gaiti *et al.* 2017). There are many different classes of ncRNAs and they have diverse roles in regulatory biological process such as modulating chromatin structure (Bhan and Mandal 2015; Bianchi *et al.* 2017; Tang *et al.* 2017), transcription rates (Goodrich and Kugel 2006), synthesis of proteins (Bartel 2004; Eichhorn *et al.* 2014) and alternative splicing (Gonzalez *et al.* 2015). While some ncRNAs are essential components of the cellular processes like rRNAs and tRNAs, many other ncRNAs such as microRNAs, piRNAs, long-ncRNAs and circular RNAs provide fine tuning and buffering regulatory capabilities in GRNs (Alberti and Cochella 2017; Rotini *et al.* 2017). Evolution of ncRNAs can therefore also change GRNS and this can also lead to morphological evolution (Arif *et al.* 2015; Gaiti *et al.* 2017).

Thus existing proteins and non-coding genetic information can evolve to effect developmental outcomes. However, while these toolkit components may evolve to create developmental differences, it is the expansion of toolkit genes that are correlated with the advent of metazoan multi-cellularity and the increase in morphological complexity (Rokas 2008; Erwin 2009; Gaiti *et al.* 2017). These expansions have not come from the relatively rare *de novo* formation of genes, but rather through the process of gene duplication (Ohno 1970; Taylor and Raes 2004; Van de Peer *et al.* 2009; Pascual-Anaya *et al.* 2012; Schmitz and Bornberg-Bauer 2017). These genes include the transcription factor superclass of homeobox genes, which have expanded in metazoans, especially bilaterians, via both tandem and large-scale duplications events (Garcia-Fernandez 2005; Hoegg and Meyer 2005; Holland 2015; Ferrier 2016). Furthermore, the expansion in the number and diversity of non-coding elements, such as microRNAs (miRNAs) also coincides with increases in cellular and phenotypic complexity (Sempere *et al.* 2006; Grimson *et al.* 2008; Peterson *et al.* 2009; Wheeler *et al.* 2009; Tarver *et al.* 2015). Therefore, while changes in the writing and co-option of existing GRNs underlie the evolution of development, the addition of new genetic material through gene duplication of developmental toolkit genes is also important.

1.2 Homeobox Genes

1.2.1 The Homeodomain

Homeobox containing genes encode a large superclass of transcription factors that are characterised by the presence of a homeodomain. The homeodomain sequence is usually 60 amino acids in length and folds to form a structure with three α-helices and an N-terminal domain (Fig. 1.1) (Ortiz-Lombardia *et al.* 2017). It is the third α-helix and N-terminal domain that provides the specificity in the binding of the homeodomain to the major and minor groove of the DNA double helix, respectively (Ortiz-Lombardia *et al.* 2017). Motifs within the third α-helix and N-terminus are therefore highly conserved (Fig. 1.1). This allows for the characterisation of many homeobox genes based solely on the homeobox sequence (Holland *et al.* 2007). However, there are also a

variety of other binding domains found in metazoan homeobox genes, which provide additional identification characteristics and biological functions (Fig. 1.1) (Burglin and Affolter 2016).

1.2.2 Roles in Development

Homeobox genes have many diverse roles in development and often display severe phenotypes when their regulation is disrupted (Zagozewski *et al.* 2014). Their function as transcription factors means that they can activate expression of many downstream genes. For instance, *Pax6/eyeless* (*ey*), a PRD class homeobox gene, is found to regulate eye development throughout metazoans (Kozmik 2008). Removal of *ey* expression can lead to the partial or complete loss of eye development (Clements *et al.* 2009). Furthermore, ectopic expression of *ey* in the imaginal discs of *Drosophila* can lead to the development of structures very similar to the complex compound eye (Halder *et al.* 1995). Eye development in *Drosophila* requires the activation of thousands of genes, many of which are likely to be expressed by the ectopic activation of *ey*. This example highlights the potential for homeobox genes to activate many downstream genes to build large morphological structures and the severe consequences when their expression is manipulated. These are some of the characteristics that have lead people to call them master regulators in developmental networks (Davis and Rebay 2017).

Perhaps the most well-known and studied homeobox containing genes are those that belong to the Hox cluster, due their role in patterning the body plan of bilaterians (Lewis 1978; Kaufman *et al.* 1980; Abzhanov and Kaufman 1999; Ferrier 2016; Hrycaj and Wellik 2016). The positioning of the Hox genes in the cluster corresponds to the spatio-temporal expression along the anterior-posterior axis (Fig. 1.2). The collinearity of Hox expression pattern and arrangement within the cluster is critical for patterning the identity of body segments. For instance the evolution of the Hox expression domains in vertebrates relates to changes seen in the axial identity (Burke *et al.* 1995; Carroll *et al.* 2005). Forelimb formation in mice, chickens and geese is positioned at the anterior region of *Hoxc6* expression, at the transition from cervical to thoracic axial regions. However, in the python the expression of Hoxc6 and Hoxc8 as shifted more anteriorly, which relates to the extended region of thoracic axial identity in these animals.

Furthermore, the positioning and number of breathing organs in chelicerates show diversity based on the anterior domain of *Abd-B* expression. In scorpions, breathing organs called book lungs are positioned in opisthosomal segments three (O3) to six (O6) (Schwager *et al.* 2015), which corresponds with the anterior expression boundary of *Abd-B* the O6 segment (Fig. 1.2) (Sharma *et al.* 2014c; Schwager *et al.* 2015). However, in spiders *Abd-B* is expressed more anteriorly in the opisthosoma and the breathing organs are present in O2 and O3 (Fig. 1.2) (Schwager *et al.* 2015; Schwager *et al.* 2017). Furthermore, the expression of *Abd-B* in Opiliones is found even more anteriorly than in spiders and scorpions, and corresponds with the development of breathing organs only in O2 (Fig. 1.2) (Sharma *et al.* 2012; Schwager *et al.* 2015).

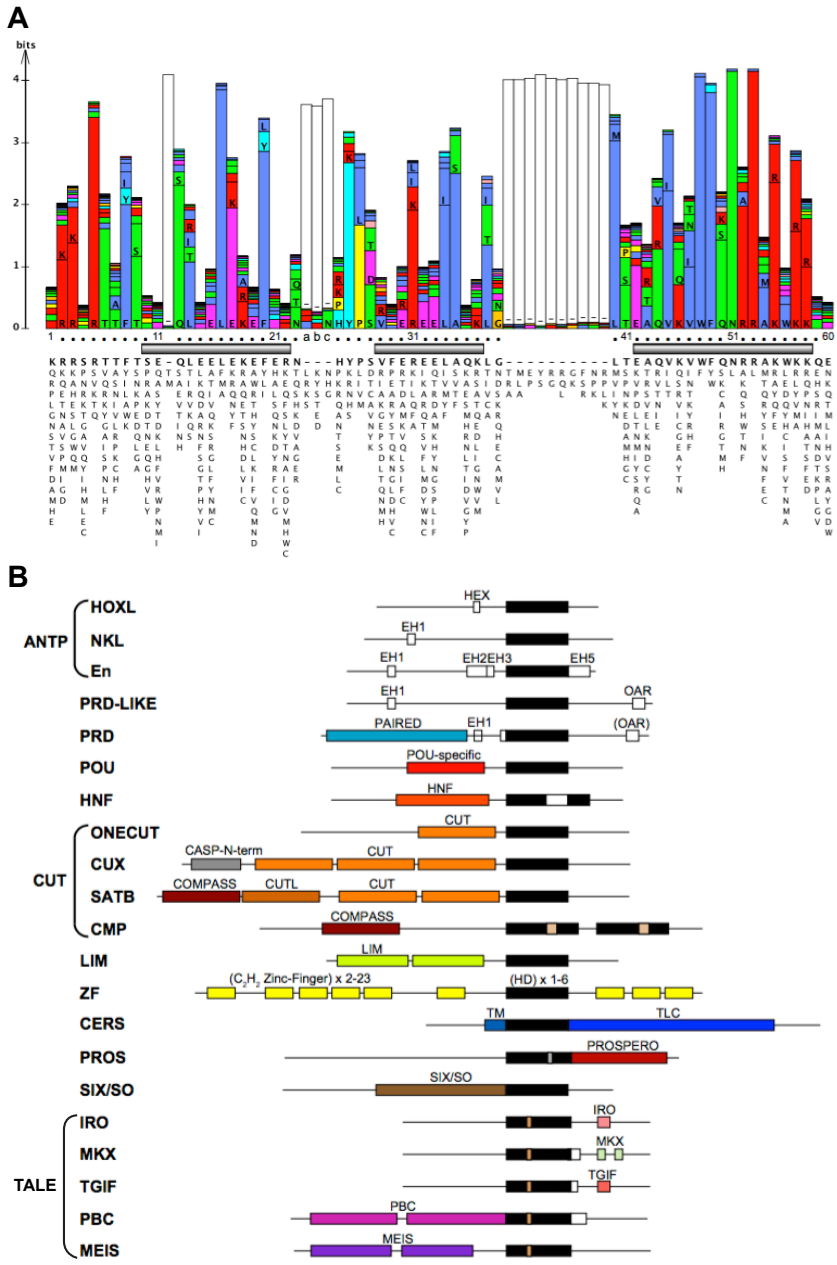


Figure 1.1: Homeobox genes encode a homeodomain and a variety of other functional domains. (A) Homeodomain protein logo generated primarily from *Drosophila* homeodomain proteins using LogoBar (Perez-Bercoff *et al.* 2006; Pérez-Bercoff *et al.* 2010). The higher the bar, the stronger a position is conserved. Letters inside the bars indicate amino acid residues. Open bars indicate gap regions that were introduced to accommodate longer atypical homeodomains. Numbers underneath are based on the standard numbering for homeodomains with 60 residues, and "abc" marks the positions of the three extra residues in loop 1 of TALE homeodomains. The three alpha helices are indicated with shaded boxes. At the bottom, the consensus sequence (most frequent residue) is shown; residues underneath each position are listed in decreasing order of frequency of occurrence. **(B)** Schematic representation of conserved domains and motifs associated with homeodomain proteins in animals. Not all domains and motifs are shown, which is mostly the case for motives conserved only within families. Figures adapted from Burglin and Affolter (2016).

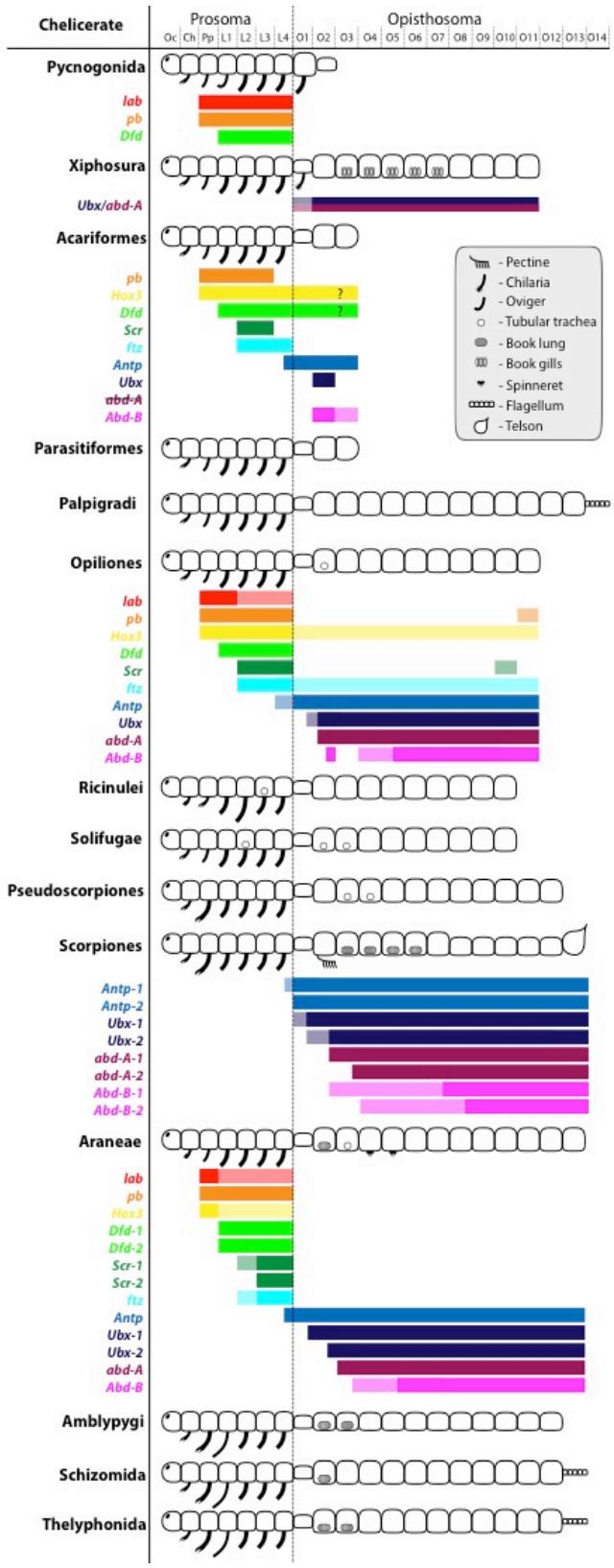


Figure 1.2: Body plans and known Hox gene expression in chelicerate orders. The number of prosomal segments is conserved in Euchelicerata; variation in segment number of Pycnogonida is not shown. The euchelicerate prosoma consists of six appendage-bearing segments: a pair of chelicerae, a pair of pedipalps, and four pairs of walking legs. There is considerable innovation in the function of these limbs, such as the long-range tactile first legs of Amblypygi or the muscular pedipalps of Scorpions. In contrast to the prosoma, opisthosomal segment number is variable both between and within orders. The respiratory organs that are found in chelicerates differ in morphology as well as number and position. For instance, Xiphosura has book gills rather than the book lungs that are found in many other chelicerates. A tracheal respiratory system occurs in several arachnid orders as well as in derived spiders. The position of these is variable and can occur within the prosoma and/or opisthosoma (simplified to show typical positions of spiracles in Acariformes and Parasitiformes). Palpigrade opisthosomal "sacs" are of dubitable homology and are not shown here. Appendages shown that are specific to certain orders include pectins (scorpions), ovigers (pycnogonids), and chilaria (horseshoe crabs). In conjunction with morphological studies, expression of Hox genes in chelicerates has been characterised in Pycnogonida (Jager *et al.* 2006), Xiphosura (Popadić and Nagy 2001), Acariformes (Telford and Thomas 1998; Barnett and Thomas 2013), Opiliones (Sharma *et al.* 2012), Scorpions (Sharma *et al.* 2014c), and Araneae (Damen *et al.* 1998; Damen and Tautz 1999; Schwager *et al.* 2007). The variable anterior expression boundaries of posterior Hox genes are strongly consistent with involvement in patterning opisthosomal segment identity. Note that *abd-A* has been lost in mites. The conserved expression domains of *lab*, p*b*, and *Dfd* in the prosoma of Pycnogonida has contributed significantly to understanding segmental homology of arthropod head segments. In both Araneae and Scorpions, Hox paralogs have been observed to have both spatial and temporal expression differences. *Oc* ocular segment, *Ch* chelicerae, *Pp* pedipalps, *L1-L4* walking legs, *O1-O14* opisthosomal segments. Figure adapted from Schwager *et al.* (2015).

1.3 MicroRNAs

1.3.1 Biogenesis and Function

MicroRNAs (miRNAs) are a class of small noncoding RNAs (~22 nt) that regulate gene expression by repressing translation of targeted mRNAs (Bartel 2004; Kawamata *et al.* 2011; Eichhorn *et al.* 2014). Current understanding of canonical miRNA biogenesis in metazoans is that the miRNA primary transcript folds into a hairpin structure that is cleaved by the endonuclease Drosha forming the so-called precursor miRNA hairpin (pre-miRNA) (Fig. 1.3). Pre-miRNAs are then exported to the cytoplasm, where they are further cleaved by Dicer to produce a double stranded RNA molecule (Fig. 1.3) (Bartel 2004; Chendrimada *et al.* 2005). One of the two RNA strands is subsequently loaded into the RNA Induced Silencing Complex (RISC) (Fig. 1.3) (Gong *et al.* 2014; Guennewig *et al.* 2014). The mature miRNA guides the RISC to specific mRNAs through complementary binding of the miRNA seed region to recognition element in the mRNA 3' UTRs (Fig. 1.3). Once bound, the miRISC represses translation by destabilising the mRNA, repressing binding of translation initiation factors, and can lead to the degradation of transcripts (Fig. 1.3) (Bartel 2004; Chendrimada *et al.* 2005; Juvvuna *et al.* 2012; Fukaya *et al.* 2014).

Normally, one of the two arms of a miRNA hairpin is preferentially incorporated into the RISC. However, some pre-miRNAs can generate two functional products, one from each arm, which normally target different mRNA repertoires (Marco *et al.* 2012b; Gong *et al.* 2014; Guennewig *et al.* 2014). The dominant arm loaded into AGO is termed the miRNA mature strand and the other passenger strand is referred to as the miRNA star (miRNA*). This mode of processing has been found to play a role in cell differentiation (Zhou *et al.* 2010) and disease progression (Pink *et al.* 2015). Moreover, for several miRNAs, it has been observed that differences have evolved between insects with respect to which arm of a conserved pre-miRNA is dominantly loaded into the RISC (Marco *et al.* 2010; Griffiths-Jones *et al.* 2011). This evolutionary change is likely to have altered the target repertoires of these miRNAs between species, because the two mature products from the precursor miRNA have different sequences and are therefore expected bind to distinct recognition sites in different mRNA targets (Marco *et al.* 2012b). Furthermore, the functional use of both miRNA and miRNA* arms is evident due to the conservation of sites in 3' UTRs for both mature products (Okamura *et al.* 2008).

1.3.2 Roles in Development

Comparative studies in animals such as vertebrates (Heimberg *et al.* 2008; Field *et al.* 2014; Fromm *et al.* 2015), ecdysozoans (de Wit *et al.* 2009; Campbell *et al.* 2011; Rota-Stabelli *et al.* 2011; Wiegmann *et al.* 2011), flatworms (Fromm *et al.* 2013) and annelids (Helm *et al.* 2012) have contributed greatly to our knowledge of microRNA function and evolution (Hatfield *et al.* 2005;

Okamura *et al.* 2008; Marco *et al.* 2010; Griffiths-Jones *et al.* 2011; Marco *et al.* 2012a; Chen *et al.* 2014; Marco 2014; Mohammed *et al.* 2014a; Mohammed *et al.* 2014b; Yang *et al.* 2014).

MicroRNAs have diverse roles in development and an individual miRNAs can have multiple mRNA targets (John *et al.* 2004; Ambros 2011; Alberti and Cochella 2017). Knockout screens of miRNAs in *Caenorhabditis elegans* and *Drosophila melanogaster* have identified varying levels of detected phenotypic consequences of knockouts, with greater affects seen in *D. melanogaster* (Miska *et al.* 2007; Alvarez-Saavedra and Horvitz 2010; Chen *et al.* 2014). Furthermore, miRNAs display both fine-tuning peripheral functions as well as important master regulatory functions in development (Rajman and Schratt 2017). Perturbed miRNA function has been shown to disrupt many developmental processes in metazoans such as embryonic survival and viability (Chen *et al.* 2014), stem cells proliferation (Hatfield *et al.* 2005), metamorphosis in insects (Sokol *et al.* 2008; Ambros 2011), organ development (Chen *et al.* 2014) and vertebral number and identity (Wong *et al.* 2015). They are also found to act through feedback and feed forward loops to return regulatory fluctuations of input signals to a steady state output, which helps canalise development (Posadas and Carthew 2014).

MicroRNAs have also been shown to regulate aspects of the maternal to zygotic transition (MZT) in a range of deuterostome and protostome species. The MZT involves the clearance of maternal transcripts and the onset of zygotic transcription. Experiments using vertebrates such as frogs and zebrafish have identified an orthologous mir-427/mir-430 family that target maternal transcripts and are required for the MZT (Giraldez *et al.* 2005; Giraldez *et al.* 2006; Lund *et al.* 2009). Similarly, invertebrates such as insects and *C. elegans* also deploy miRNAs during the MZT (Bushati *et al.* 2008; Alvarez-Saavedra and Horvitz 2010; Zondag *et al.* 2012; Ninova *et al.* 2014, 2016). However, there is a lack in understanding the regulation of the MZT in spiders and whether miRNAs are involved.

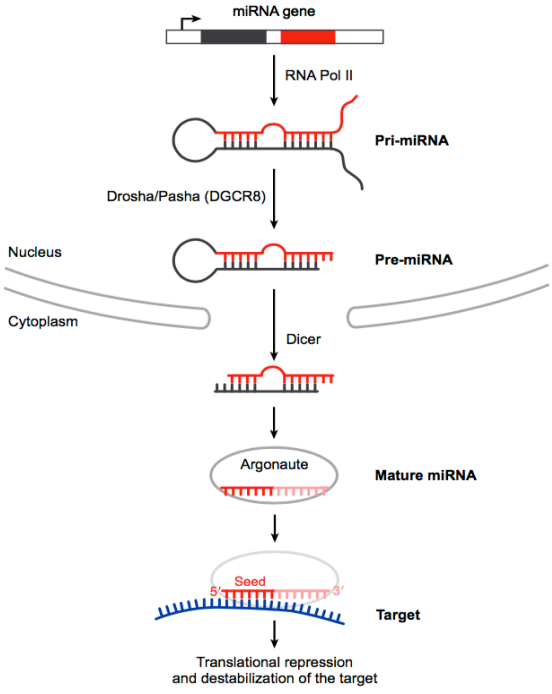


Figure 1.3: An overview of miRNA biogenesis and effector pathways. The transcription of a primary miRNA (pri-miRNA) by RNA polymerase II is followed by its processing by Drosha in complex with Pasha/DGCR8, into a precursor hairpin (pre-miRNA); these steps occur in the nucleus. After nuclear export, processing of the pre-miRNA by Dicer occurs in the cytoplasm. One strand of the mature duplex is preferentially loaded onto an Argonaute family member. The miRNA then guides the Argonaute and associated factors to its targets (blue) specified by base pairing to nucleotides 2-7 at the 5′ end of the miRNA (the seed sequence, in red). Figure taken from (Alberti and Cochella 2017).

1.4 Gene duplication

1.4.1 Evolution after Gene Duplication

The process of gene duplication provides new genetic material that can evolve to rewire gene regulatory networks (Ohno 1970; Taylor and Raes 2004; Huminiecki and Conant 2012). Duplications of regions containing genes, either coding or non-coding, generates further copies called paralogs. Initially when genes duplicate there are dosage effects that can change the stoichiometric relationships between the paralogs and their interacting partners (Innan and Kondrashov 2010). These effects can be disruptive to the function of the regulatory network and therefore paralogs are often lost (Innan and Kondrashov 2010). However, if there are either positively selected dosage effects due to duplication or a lack of dosage effect from duplications, then paralogs maybe retained (Innan and Kondrashov 2010). Once retained the paralogs of a gene at first may have similar or identical coding regions. The presence of two copies of the same gene in the genome can therefore relax selective pressures on sequence evolution of the paralogs. However, depending on the scale of the duplication the *cis*-regulatory controls of a paralog may or may not be copied in the process of duplication. This could have immediate consequences on the spatial and temporal expression of the duplicated paralog (Lynch and Katju 2004). Thereafter, evolution of the coding and *cis*-regulatory regions of paralogs may lead to functional diversion via two routes.

After gene duplication the functional role of the paralogs can subfunctionalise so that they split the function of the ancestral single copy (Ohno 1970; Taylor and Raes 2004; Huminiecki and Conant 2012). This can occur either through changes to *cis*-regulatory and/or coding regions. In the case of *cis*-regulatory evolution the enhancers of the paralogs may diverge so that expression of the paralogs may collectively resemble that of the ancestral single copy expression pattern. Similarly, coding changes can occur where functional domains are differentially lost between paralogs. Subfunctionalisation therefore creates a decoupling in the selective pressures between paralogs and facilitates divergences in the *cis*-regulatory and coding regions of genes. These changes can continue to happen so that neofunctionalisation can occur, providing additional functions to a paralog that were not present in the ancestral single copy (Ohno 1970; Taylor and Raes 2004; Huminiecki and Conant 2012). These evolutionary processes both affect the wiring of gene regulatory networks and developmental outcomes.

1.4.2 Expansion of Homeobox containing Genes in Metazoans

During the evolution of metazoans the expansion of homeobox gene and miRNA repertoires via duplication has been associated with multicellularity and the increase in morphological complexity (Garcia-Fernandez 2005; Hoegg and Meyer 2005; Pascual-Anaya *et al.* 2012; Holland 2015). The initial expansion of homeobox containing genes started prior to the metazoan lineage (Fig. 1.4) (Ferrier 2016). The divergence of these proto-homeobox paralogs early in pre-metazoan and metazoan evolution created several functional classes of homeobox genes with conserved

homeodomains (Fig. 1.4) (Pascual-Anaya *et al.* 2012; Ferrier 2016). It is thought that in pre-bilaterians these homeobox genes formed a large Giga-homeobox cluster, containing several families, including the addition of the metazoan specific ANTP class (Fig. 1.4) (Ferrier 2016). Continued tandem duplications of the different classes generated clusters of homeobox genes such as the ParaHox, SuperHox, SINE/Six, TALE/Irx, PRD/HRO clusters (Fig. 1.4) (Ferrier 2016). These became fragmented in the genome of the bilaterian ancestor, and subsequently have been retained to varying degrees in different lineages during bilaterian evolution (Fig. 1.4) (Ferrier 2016).

Tandem duplication of homeobox genes later in metazoan evolution also corresponds to the evolution of morphological changes. In the arthropod lineage Hox3 has independently duplicated in various lineages of insects (van der Zee *et al.* 2005). For example in *Tribolium*

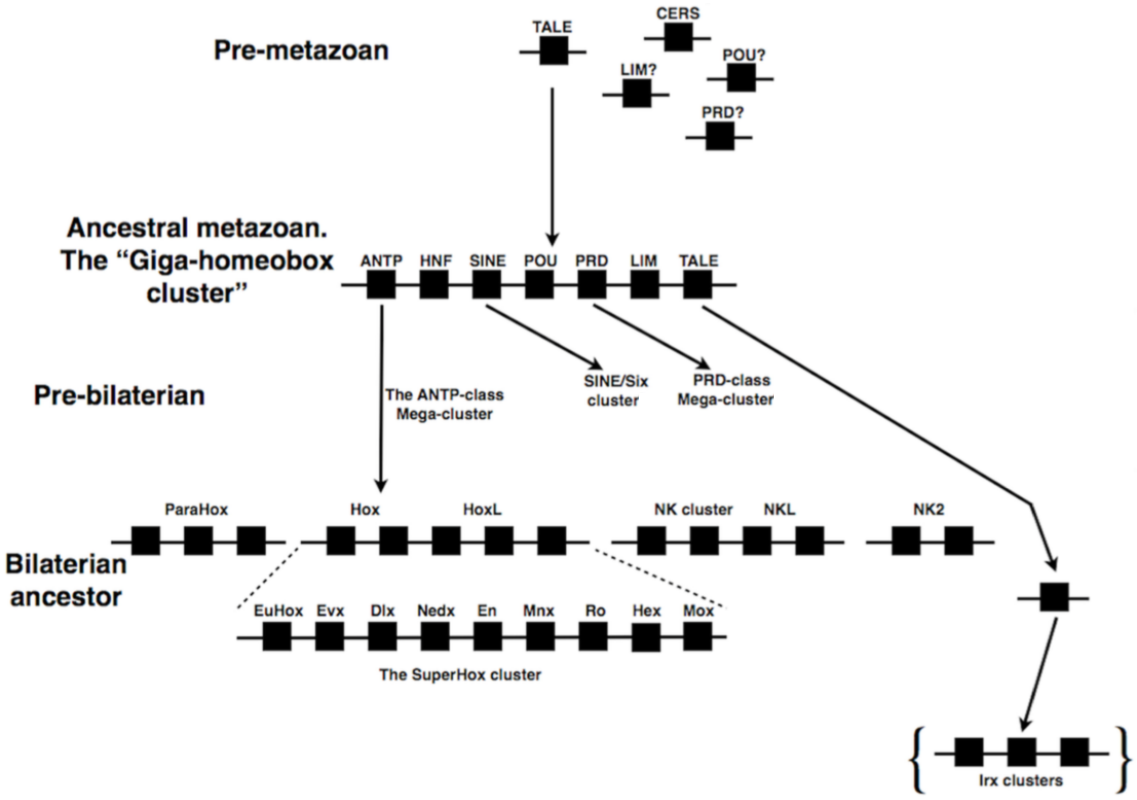


Figure 1.4: Evolution of the hypothetical metazoan "Giga-homeobox cluster". Prior to the origin of animals there were only a small number of homeobox genes, including TALE and CERS-class genes along with several further genes of uncertain affinities (Sebe-Pedros *et al.* 2011). The genomic arrangement of these genes is also unclear. In an ancestral metazoan a clustered array of homeobox genes likely existed, consisting of the precursors of several classes: the "Giga-homeobox cluster." The deduction of the composition of this array is described in the text. Further members of this array may be revealed by analyses of additional metazoan genome sequences. Sub-components of this Giga-cluster include the ANTP-class Mega-cluster (Pollard and Holland 2000), which had dispersed into at least four sub-components distributed on distinct chromosomes by the time of the bilaterian (protostome–deuterostome) ancestor (Hui *et al.* 2012). One of these sub-components was the SuperHox cluster, composed of the true Hox genes (EuHox genes) and several additional ANTP-class genes that were Hox-linked (HoxL). A second sub-component was the "NK cluster" genes with several NK-linked (NKL) genes. Further gene clusters deriving from within the Giga-cluster included the SINE/Six cluster and the PRD-class Mega-cluster. The Iroquois/Irx cluster is different from the other clusters described here because it expanded to a three-gene cluster independently in several distinct lineages (denoted by the brackets), most likely from a single gene state in the bilaterian ancestor. Continuous horizontal lines indicate clustering on the same chromosome. Figure adapted from (Ferrier 2016).

castaneum there are two copies of *Hox3*/*zen*, though in *Drosophila* there are two *zen* paralogs and a third copy called *bicoid* (*bcd*) (Berleth *et al.* 1988; Driever and Nüssiein-Volhard 1988; McGregor 2005). *Bcd* has diverged from the *zen* role in serosa development to acquire the function as a anterior-posterior (AP) axis determinant that is maternal supplied (McGregor 2005). Furthermore, in the ditrysian butterflies *Hox3* has duplicated to form four tandem duplicates of *Hox3*/*Shx* genes (Ferguson *et al.* 2014). These paralogs show divergent expression patterns in serosa development and may have subfunctionalised (Ferguson *et al.* 2014). Therefore, tandem duplications have been important in both expanding developmental toolkit gene repertoires and contribute to evolutionary changes.

While the occurrence of tandem and chromosomal duplications are fairly common, whole genome duplications (WGDs) are observed less frequently. The polyploidisation of genomes is associated with adaption, speciation and diversification of organisms (Van de Peer *et al.* 2009; Huminiecki and Conant 2012). They have been identified in ciliates, fungi, in several animal lineages and in many plant lineages (Van de Peer *et al.* 2009).

In animals, two rounds of WGDs have been identified in the ancestor of vertebrates (Putnam *et al.* 2008). Additional duplications have also occurred within vertebrates leading to teleost fish and another in salmonid fish (Mungpakdee *et al.* 2008; Pascual-Anaya *et al.* 2012; Berthelot *et al.* 2014). A hallmark of these WGD events is the presence of duplicated clusters of genes (Garcia-Fernandez 2005; Pascual-Anaya *et al.* 2012). The evolution of homeobox genes and their arrangement into several clusters prior to the bilaterian ancestor has made them attractive candidates to identify WGD (Pascual-Anaya *et al.* 2012; Ferrier 2016). Differential retention across lineages concomitant with coding and regulatory changes in paralogous Hox genes relate to the evolutionary changes in the vertebrates (Lynch *et al.* 2006). Therefore while tandem duplications may have helped to build gene repertoires, WGDs may have great influence on the evolutionary potential of a lineage.

1.4.3 Expansion of miRNAs in Metazoans

Variation in miRNA function and the expansion of repertoires in genomes has become increasingly connected with the evolution of cellular differentiation, morphological complexity and even multi-cellularity (Heimberg *et al.* 2008; Peterson *et al.* 2009; Wheeler *et al.* 2009; Campo-Paysaa *et al.* 2011; Guerra-Assunção and Enright 2012; Arif *et al.* 2013; Tarver *et al.* 2015). MicroRNAs are found in plants, in single and multicellular algae, sponges and eumetazoans, but are absent from choanoflagellates, placozoans and ctenophorans (Grimson *et al.* 2008; Erwin 2009; Tarver *et al.* 2013; Tarver *et al.* 2015; Gaiti *et al.* 2017). Furthermore, the bilaterian biogenic enzymes that process miRNAs and the structure of their miRNA hairpins differ from non-bilaterians (Bråte *et al.* 2016; Moran *et al.* 2017). These patterns indicate that miRNAs and their biogenic enzymes likely originated in the pre-metazoan ancestor but have subsequently been independently lost in several lineages early in animal evolution (Bråte *et al.* 2016; Moran *et al.* 2017). Thereafter, the number of

miRNA families has increased in bilaterian lineages (Wheeler *et al.* 2009; Tarver *et al.* 2013). Interestingly, the length of the 3' UTRs in animals increases as the number of cell types per species increases (Chen *et al.* 2012). Lengthening of 3' UTRs generates more available miRNA targeting of mRNAs. So not only have miRNA repertoires expanded but also the regions that they target have evolved to accommodate a greater degree of regulatory interactions (Chen *et al.* 2012).

Tandem and dispersed duplications have been important in the expansion of some miRNA, though surprisingly expansions have occurred through the *de novo* formation of miRNAs found in clusters that coincide with the evolution of gnathostomes and eutherians (Hertel *et al.* 2006; Griffiths-Jones *et al.* 2011; Mohammed *et al.* 2014b). The emergence of novel miRNAs is due to their short sequence that can readily evolve in genomes to form hairpin structures (Marco *et al.* 2013). If generated close to other miRNAs, for instance in a cluster, then they may be localised around available miRNA biogenesis machinery and therefore become processed (Marco *et al.* 2013). Duplications and *de novo* formation have generated clusters of miRNAs that regulate similar sets of genes. Some of these clusters have important functions in development. For example the mir-427/mir-430 family in vertebrates, the *mir-309* cluster in insects, and the mir-35-42 family in nematodes are all clusters of miRNAs that regulate MZT (Giraldez *et al.* 2005; Giraldez *et al.* 2006; Bushati *et al.* 2008; Lund *et al.* 2009; Alvarez-Saavedra and Horvitz 2010; Zondag *et al.* 2012; Ninova *et al.* 2014, 2016).

1.4.4 Gene Duplication in Chelicerata

Until recently, vertebrates were the only metazoans where WGD had been identified. However, recent evidence suggests that there have been independent WGDs in several chelicerate lineages.

There is evidence for two WGDs in the common ancestor of horseshoe crabs, exemplified by the presence of four Hox clusters in extant species (Fig. 1.5) (Ohno 1970; Hoegg and Meyer 2005; Nossa *et al.* 2014; Kenny *et al.* 2015a). Additionally, in scorpions, two Hox clusters as well as duplicates of other homeobox genes have been found (Fig. 1.5) (Cao *et al.* 2013; Sharma *et al.* 2014c; Di *et al.* 2015; Qu *et al.* 2015; Sharma *et al.* 2015). In *P. tepidariorum* 28% of annotated transcripts are expressed from duplicated genes (Posnien *et al.* 2014) and duplicates of some Hox genes have been found in this spider and in *Cupiennius salei* (Fig. 1.5) (Schwager *et al.* 2007; Pechmann *et al.* 2015). Recent analysis (including data from this thesis) has also identified two Hox clusters in *P. tepidariorum* and other synteny signatures reminiscent of a WGD in spiders and scorpions (Schwager *et al.* 2017). Moreover, estimates of the timing of duplications in chelicerates suggest that the WGD in spiders and scorpions is independent to the WGDs in the horseshoe crab lineage (Schwager *et al.* 2017). Furthermore, In contrast to the horseshoe crabs, scorpions and spiders, Hox gene duplications have not been found in other chelicerate orders like Acariformes, Parasitiformes or Opiliones, with the exception of duplicate *ftz* genes in the mite *Tetranychus urticae* (Fig. 1.5) (Grbic *et al.* 2011; Sharma *et al.* 2012). Though, *abd-A* appears to have been lost in

T. urticae, suggesting a slightly unusual Hox arrangement in the mite (Grbic *et al.* 2011) (Fig.1.5 and 1.6).

There is also evidence for the sub-functionalisation and neo-functionalisation of Hox paralogs in chelicerates. For example, shifts in the anterior expression boundaries of paralogous Hox genes in scorpions are thought to correlate with the evolution of body plan heteronomy (Sharma *et al.* 2014c). In *P. tepidariorum*, RNAi knockdown of *lab-1,* but not *lab-2,* results in loss of pedipalps and the first walking legs, suggesting divergent functions have evolved between these paralogs (Pechmann *et al.* 2015). In addition, in the horseshoe crabs, *Tachypleus tridentatus* and *Carcinoscorpinus rotundicauda,* the spatial expression of *Dfd* and *Scr* paralogs has evolved within and between species, although the implications of these differences are not yet known (Kenny *et al.* 2015a).

Duplications of many other developmental genes have been found in spiders and some of these are correlated with phenotypic evolution. For example, in *P. tepidariorum dachshund-2,* but not *dachshund-1*, is required for the development of the patella leg segment (Turetzek *et al.* 2016b). Therefore the evolution of the patella in chelicerates is associated with gene duplication and neo-functionalisation (Turetzek *et al.* 2016b). Several retinal determination genes (RDGs) are also duplicated in the spiders *P. tepidariorum* and *C. salei* including *sine oculis* (*so*), and *six3* (Samadi *et al.* 2015; Schomburg *et al.* 2015) and paralogs of RDGs are differentially expressed during the development of different spider eyes (Samadi *et al.* 2015; Schomburg *et al.* 2015). For example, in *P. tepidariorum*, *so-1* is expressed in all eyes while *so-2* expression is restricted to anterior lateral eyes (Schomburg *et al.* 2015). Therefore, the neo-functionalisation and/or sub-functionalisation of paralogous RDGs could have contributed to the evolution of the spider visual system (Samadi *et al.* 2015; Schomburg *et al.* 2015).

The repertoires of venom genes have also expanded in spiders and scorpions due to tandem duplications (Cao *et al.* 2013; Clarke *et al.* 2014; Haney *et al.* 2014; Sanggaard *et al.* 2014; Xu *et al.* 2014; Xu *et al.* 2015; Haney *et al.* 2016) and there has been convergent evolution of venom in centipedes and spiders with the neo-functionalisation of an ancestral arthropod hormone to form a class of HAND toxins (Undheim *et al.* 2015). Among araneomorphs, the copy numbers of silk genes have also expanded by tandem duplication, which is perhaps related to innovations in their usage of silk (Clarke *et al.* 2014; Haney *et al.* 2014; Sanggaard *et al.* 2014).

The presence of WGDs in two independent linages of chelicerates, as well as lineages where there is no evidence of WGD, provides an excellent opportunity to study evolution 'before' and 'after' WGD. Comparisons between these organisms will likely provide insight into how gene duplication has contributed to their evolutionary and developmental differences. Furthermore, the duplications in chelicerates provide an independent invertebrate example to compare to WGDs seen in vertebrates.

1.5 Chelicerates

1.5.1 Phylogenetics of Chelicerata

The chelicerates are a sub-phylum of the arthropods. Resolving the phylogenetic relationships among arthropods has improved our understanding of developmental differences that define the evolution of certain lineages, as well as similarities that suggest shared ancestral mechanisms. It is now generally accepted that Chelicerata are the sister branch to the monophyletic Mandibulata (Myriapoda and Pancrustacea) in the arthropod phylogeny (Fig. 1.5) (Rota-Stabelli *et al.* 2011; Borner *et al.* 2014; Chipman *et al.* 2014; Rehm *et al.* 2014; Sharma *et al.* 2014b; Pepato and Klimov 2015). This is consistent with differences in gene content and expression that have been identified among these arthropods. For example, the cuticular proteins of the CPR family (Cuticular Proteins with the Rebers and Riddiford consensus) regulate the development of the arthropod exoskeleton. However, while the RR-2 subfamily of CPR genes is present in all arthropods, the RR-1 subfamily is restricted to the Mandibulata (Fig. 1.5) (Willis 2010; Chipman *et al.* 2014; Ioannidou *et al.* 2014; Kenny *et al.* 2015b). Furthermore, particular components of the juvenile hormone signalling pathway are restricted to hexapods (CYP15A1/CYP15C1) or mandibulates (JHE and CYP306A1), and chelicerates likely use ponasterone A (25-deoxy-20E) rather than 20-hydroxyecdysone (20E) as their main moulting hormone (Fig. 1.5) (Grbic *et al.* 2011; Cao *et al.* 2013; Qu *et al.* 2015). In addition, the expression of *cap-n-collar* differs between chelicerates and mandibulates, and is associated with the evolution of the mandibles in the latter (Sharma *et al.* 2014a).

While the relationship of the Arthropoda subphyla appears to be resolved, reconstructing a robust Chelicerata phylogeny has remained challenging. The sea spiders (Pycnogonida) represent a euchelicerate out-group to Chelicerata (Fig. 1.5) (Sharma *et al.* 2014b). The chelicerate group of extant horseshoe crabs (Xiphosura) forms a sister-group to the arachnids (Fig. 1.5). Within Arachnida molecular evidence supports the Tetrapulmonata clade containing the Pedipalpi and Araneae (Fig. 1.5) (Sharma *et al.* 2014b). The Araneae taxon contains over 45,000 described spider species and comprise of two clades, the Mesothelae and the more abundant Opisthothelae (Garrison *et al.* 2016). The Opisthothelae are divided into the Mygalomorphae and Araneomorphae based on phylogenetic analysis and the characteristic downward facing or opposing orientation of their chelicerae, respectively (Garrison *et al.* 2016).

The Pedipalpi are the sister group to Araneae, and contain several less well-known taxons (Amblypygi, Schizomida, Thelyphonida) (Fig. 1.5) (Schwager *et al.* 2015). The Tetrapulmonata are also thought to group with scorpions to form the Arachnopulmonata clade (Fig. 1.5) (Sharma *et al.* 2014b). However, there is inconsistent phylogenetic support for the Poecilophysidea clade (Solifugae and Acariformes) and the positioning of other non-Arachnopulmonata arachnids (Fig. 1.5) (Sharma *et al.* 2014b; Pepato and Klimov 2015). This is due to the long branch lengths of certain groups such as Acariformes and Parasitiformes (Sharma *et al.* 2014b).

The segmented body plan of chelicerates can be split into the anterior prosoma and posterior opisthosoma regions (Fig. 1.2). The prosoma region bares the six pairs of appendages: the chelicerae, pedipalps, and four pairs of walking legs. Note that in sea spiders and horseshoe crabs there is another appendage on the first opisthosomal segment, which is not present in arachnids (Fig. 1.2) (Schwager *et al.* 2015). There are considerable differences in the morphology between appendages and between species. For example the pedipalps, which are modified legs, have evolved different segment numbers, and in scorpions are pincers, where as in male spiders they are modified to deliver sperm to the female epigyne (Fig. 1.2) (Michalik and Ramírez 2014). Chelicerate opisthosomas also vary in morphology, with different segment numbers and organ diversity. The breathing organs of the Xiphosuran are book gills, which are present in opisthosomal segments three to seven (Fig. 1.2). In contrast, arachnids breath via either tubular trachea and/or book gills. These structures vary in their position among chelicerate orders (Fig. 1.2). Other opisthosomal structures can be found in distinct lineages, such as spinnerets in spiders, the telson sting and pectine sensory structures of scorpions, and the reduced flagellum segments of the opisthosoma in Schizomida and Thelyphonida (Fig. 1.2). Furthermore, spiders and scorpions have both evolved venom to immobilise their prey (Fig. 1.2).

The evolutionary position of chelicerates in arthropods, their phenotypic diversity and various adaptive features make them an attractive system to study the effects of gene duplication. There are already some insights into how gene duplication has contributed to the evolution of segment identity, appendages, and eyes. A growing range of genomic resources in chelicerates provide an opportunity to study gene duplication systematically and the molecular tools in *P. tepidariorum* are ideally for studying particular candidates during development.

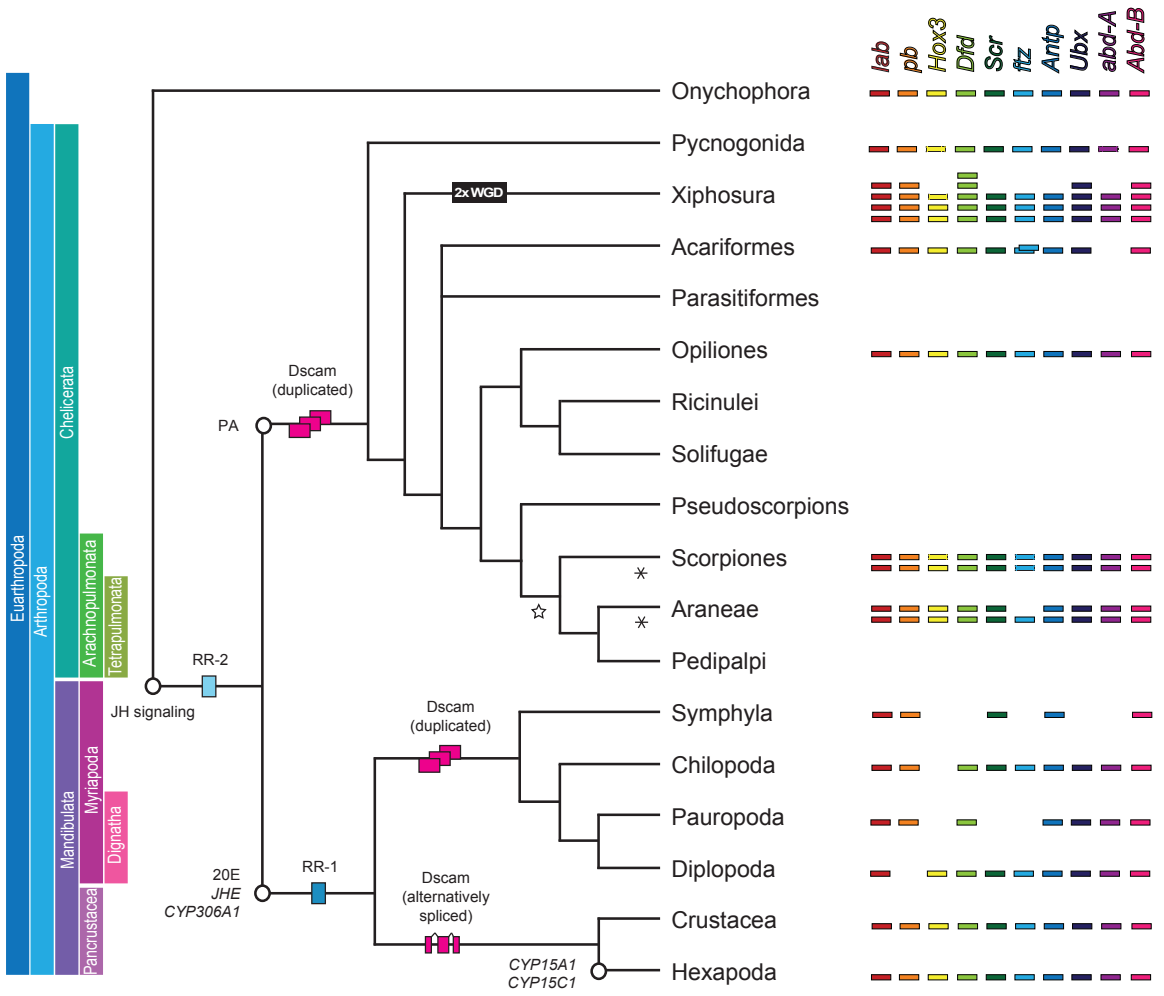


Figure 1.5: Phylogeny and Hox repertoire of chelicerates and myriapods. Juvenile hormone signalling was likely present in the common ancestor of arthropods. Changes in the components of the JH pathway, such as the mandibulate specific CYP306A1 and JHE, have lead to divergence in signalling between chelicerates and myriapods. *Dscam* genes (pink boxes) are highly duplicated in chelicerates and myriapods, but *Dscam* is alternatively spliced in to generate extensive isoform diversity in pancrustaceans. The cuticular proteins RR-2 (light blue box) are found across arthropods, but RR-1 genes (dark blue box) are restricted to mandibulates. The Hox gene repertoires are shown for each clade. Boxes with dotted outlines indicate variation in the presence or absence between species within a clade. Two whole genome duplications (2x WGD) have been proposed in HSCs. It is possible that scorpions and spiders have undergone independent large scale duplications (*asterisk*), or shared duplication in the Arachnopulmonata ancestor (*star*). JH, juvenile hormone; PA, Ponasterone A; 20E, 20-hydroxyecdysone. Phylogeny based on (Schwager *et al.* 2015), and Hox gene repertoires from (Cook *et al.* 2001; Manuel *et al.* 2006; Grbic *et al.* 2011; Chipman *et al.* 2014; Nossa *et al.* 2014; Sharma *et al.* 2014c; Di *et al.* 2015; Kenny *et al.* 2015a; Kenny *et al.* 2015b; Pechmann *et al.* 2015; Sharma *et al.* 2015; Schwager *et al.* 2017). Figure taken from (Leite and McGregor 2016).

1.5.2 Resources for Studying Gene Duplication in Chelicerates

Growing genomic and transcriptomic resources in chelicerates allows for comparisons between several orders and with the extensive data available for insects and other arthropods. For chelicerate lineages with possible WGD, sequenced spider genomes include both Araneomorphae (*Parasteatoda tepidariorum*, *Stegodyphus mimosarum, Nephila clavipes*)(Sanggaard *et al.* 2014; Babb *et al.* 2017; Schwager *et al.* 2017) and Mygalomorphae (*Acanthoscurria genticulata*). There are genomes available for two scorpions (*Mesobuthus martensii* and *Centruroides sculpturatus*)(Cao *et al.* 2013; Schwager *et al.* 2017) and three of the four extent horseshoe crabs (*Limulus polyphemus*, *Carcinoscorpius rotundicauda* and *Tachypleus tridentatus*)(Nossa *et al.* 2014; Kenny *et al.* 2015a). For the chelicerate lineages that do not appear to have undergone a WGD, genomes exist for the orders of Parasitiformes (*Ixodes scapularis*, *Metaseiulus occidentalis*, *Rhipicephalus microplus*, *Varroa destructor*)(Hoy 2009; Lawson *et al.* 2009; Cornman *et al.* 2010) and Acariformes (*Tetranychus urticae*)(Grbic *et al.* 2011). These provide valuable information to analyse coding and non-coding genes and their genomic synteny.

There are also many transcriptomic resources that have been generated in chelicerates. Some of these resources were produced to investigate phylogenetic relationships of ecdysozoans (Borner *et al.* 2014; Sharma *et al.* 2014b; Garrison *et al.* 2016). Transcriptomes have come from a range of spiders including *P. tepidariorum*, *Latrodectus hesperus*, *S. mimosarum*, *A. geniculata*, which have been used to study gene repertoires as well as the evolution of silk and venom genes (Haney *et al.* 2014; Clarke *et al.* 2015; Xu *et al.* 2015; Haney *et al.* 2016; Schwager *et al.* 2017). Transcriptomes have been extensively used in scorpions and horseshoe crabs to study the duplication of homeobox genes (Kenny *et al.* 2015a; Sharma *et al.* 2015). These next generation sequencing resources not only provide estimates for gene repertoires but they also provide a wealth of sequence information that can be used for such things as phylogenetics and predicting functional interactions.

While bioinformatics can provide global analysis of large genomic resources it is also important to study the developmental aspects of gene regulation in embryos. There are several emerging chelicerates models that are used in evolutionary and developmental studies, however the spider *P. tepidariorum* is the most developed model with a well assembled genome (Hilbrant *et al.* 2012; Schwager *et al.* 2017). Well fed *P. tepidariorum* females can lay cocoons of several hundred embryos every week (Hilbrant *et al.* 2012). Embryos within a cocoon tend to develop synchronously making it ideal to separate for staging and fixations (Hilbrant *et al.* 2012). There is a clear and detailed staging description of *P. tepidariorum* development (Mittmann and Wolff 2012).

At stage one (S1), embryos start off with the nucleus at the centre. Initially the nucleus divides synchronously intralecithally in the syncytium (Fig. 1.6)(Mittmann and Wolff 2012). As nuclei continue to divide, they surface evenly across the spherical egg and by S2 are fully cellularised (Mittmann and Wolff 2012; Pechmann 2016). Cell shape rearrangements organise cells

to one hemisphere during the formation of the germ disc at S3 (Fig. 1.6)(Pechmann 2016). A group of mesenchymal cells called the cumulus forms during S4 in the centre of the germ disc and migrates to the periphery of the germ disc at S5, effectively breaking the radial to bilateral symmetry (Fig. 1.6)(Akiyama-Oda and Oda 2003). The region to which the cumulus migrates to opens up to form the dorsal field at S6. The dorsal field continues to contract around the embryo until S7, so that the centre of the embryo becomes the posterior, and the side opposite to where the cumulus migrated to forms the anterior. The break in radial to bilateral symmetry creates a germ band that segments in the anterior via splitting waves and the posterior through oscillation in gene expression (Akiyama-Oda and Oda 2006; Kanayama *et al.* 2011; Schönauer *et al.* 2016). Appendage buds for the chelicerae (Ch), pedipalps (Pp) and four walking limbs (L1-L4) develop during S8 while segmentation is still occurring (Fig. 1.6). At later stages the appendages segment, and the dorsal tissue grows around the yolk, enveloping it during inversion. By S14 a constriction can be seen between the appendage bearing prosomal region and the opisthosomal region, which bares book lungs and trachea in O2 and O3, respectively, and spinnerets in O4 and O5 (Fig. 1.6). The newly hatched spiders are initially motionless and take few days to strengthen the cuticle, develop hairs and become mobile (Fig. 1.6). Maturity to adulthood takes just over a month and individuals are easy to handle for experimentation.

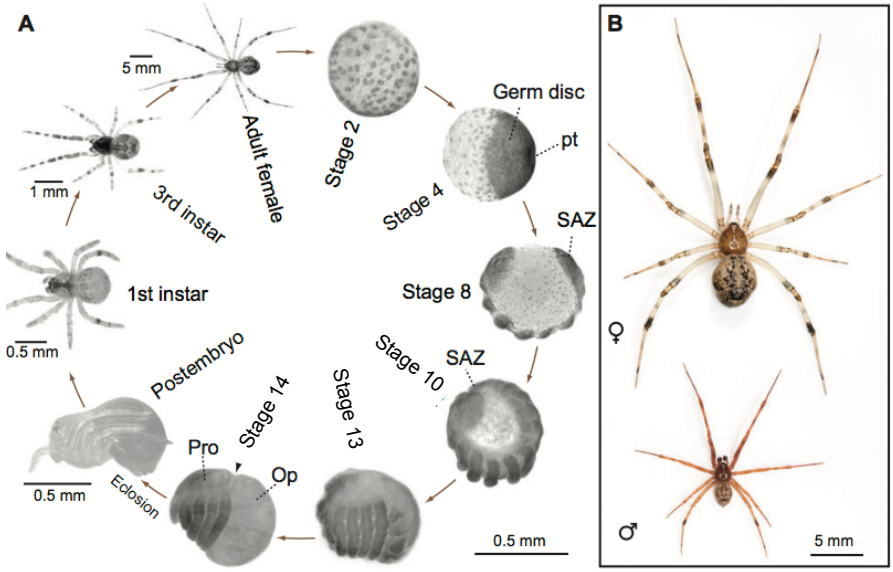


Figure 1.6: The development and life cycle of the spider *Parasteatoda tepidariorum*. (A) Stages of embryonic and post-embryonic development: stage 2, cellularisation is complete, blastoderm formation; stage 4, germ disc with primary thickening (*pt*) in the centre; stage 8, early germ band with segment addition zone (*SAZ*); stage 10, elongated germ band with limbs; stage 13, the end of inversion; stage 14, with distinct prosoma (*Pro*) and opisthosoma (*Op*) with constriction between them (*arrowhead*); postembryo; 1st instar, which exits the cocoon; 3rd instar, a free-foraging instar stage; adult female. Staging after Mittmann and Wolff (2012). Anterior is to the left in all images. Scale bar is given with respect to the stage 2-14 embryos. **(B)** Adult female (*top*) and male (*bottom*). Anterior is to the top. Figure adapted from (Hilbrant *et al.* 2012) and (Schwager *et al.* 2015).

Protocols for fixations and in situ hybridisation (ISH) have worked well to study the expression patterns of many developmental genes in *P. tepidariorum* (Akiyama-Oda and Oda 2003; Hilbrant *et al.* 2012). Furthermore, disruption of gene expression can be achieved with parental and embryonic and embryonic RNAi (pRNAi and eRNAi respectively)(Hilbrant *et al.* 2012). The effects of pRNAi are ideal for looking at systemic defects in development when gene expression is knockdown (Akiyama-Oda and Oda 2006). In contrast, eRNAi at different stages can be used to knockdown genes in different sizes of clonal populations (Kanayama *et al.* 2011; Schönauer *et al.* 2016; Pechmann *et al.* 2017). This has been a useful tool to study both the genetic regulation of axis specification, germ disc formation, cumulus migration, germ band formation, the dynamics of segmentation in the anterior and posterior, and appendage development and their phenotypic differences (Akiyama-Oda and Oda 2006; Kanayama *et al.* 2011; Pechmann *et al.* 2011; Pechmann *et al.* 2015; Pechmann 2016; Schönauer *et al.* 2016; Pechmann *et al.* 2017). Recent work has also been made to establish the over expression of genes using capped mRNA (Schönauer *et al.* 2016; Pechmann *et al.* 2017).

Together the genomic resources available from chelicerates and the tools developed in *P. tepidariorum* provide an excellent platform to study the evolution of development after gene duplication. To further what has already been revealed about gene duplication in chelicerates, I have used some of these genomics resources, as well as producing new sequencing data and analysing currently unpublished transcriptomes. Furthermore, ISH was used to investigate the divergences in the expression of duplicate genes.

1.6 Aims

1.6.1 Duplication of Homeobox Genes in Chelicerata

To further understand the evolution of homeobox genes, complete repertoires from different species provide an important perspective. However, there has only been systematic identification of homeobox genes in scorpion and horseshoe crab species. Therefore in chapter three I have:

- Used transcriptome resources to identify homeobox gene repertoires in two spiders, a scorpion and a tick, and compared these to the homeobox repertoires of other arthropods.
- Investigated the evolution of homeobox clusters using the well-assembled DoveTail *P. tepidariorum* genome.
- Used chelicerate genomes to investigate the scale of duplication by searching for remains of duplicated homeobox clusters.
- Observed divergences in the expression of duplicate homeobox genes in *P. tepidariorum* using an embryonic transcriptome.
- Used sequence alignments and phylogenies to study the divergence in homeodomains.
- Studied the spatial and temporal expression patterns of duplicated genes using ISH in *P. tepidariorum*.

1.6.2 Duplication of miRNAs in Chelicerata

To date, the characterisation of miRNAs in chelicerates has been limited primarily to mites and ticks, representing a rather narrow sampling of this subphylum (Barrero *et al.* 2011; Rota-Stabelli *et al.* 2011; Zhou *et al.* 2013; Liu *et al.* 2014a; Luo *et al.* 2015). There is also no analysis of any non-coding genes in chelicerates and whether they are duplicated. To gain further insights into the evolution of non-coding genes in chelicerates, miRNAs need to be first characterised in other chelicerate lineages. Therefore in chapter four I have:

- Sequenced small RNA from ten embryonic stages of *P. tepidariorum* development and used the data to predict the embryonic miRNA repertoire.
- Annotated and identified the orthology of *P. tepidariorum* miRNAs, and identify paralogs.
- Used the *P. tepidariorum* miRNA sequences to search for orthologs in other spiders and in scorpions, mites, ticks and horseshoe crabs. Compared the gains and losses of miRNAs in across these chelicerates in conjunction with other ecdysozoans.
- Used chelicerate genomes to study the synteny and duplication of miRNA clusters.
- Studied the evolution of miRNA arm switching compared to other arthropods.

1.6.3 Developmental Roles of Duplicated miRNAs

To investigate a role of duplicated miRNAs in *P. tepidariorum*, the most expanded miRNA family identified in Chapter Two was studied. This family had orthology to the mir-3791 family, which is located in a cluster of miRNAs that regulate the maternal to zygotic transition (MZT) in insects. Current understanding of the MZT is limited in *P. tepidariorum* and other chelicerates. Further investigations will provide insight into this crucial phase in development in chelicerates and an evolutionary comparison to what is known in insects. Therefore in chapter five I have:

- Used the embryonic expression profiles of mRNAs and miRNAs to identify the MZT in *P. tepidariorum*.
- Used differential expression; correlation analysis; and miRNA target prediction software to predict miRNAs expressed at the MZT and their mRNA targets.
- Assessed the sequence similarity of the MZT miRNAs and their expression profiles in *P. tepidariorum* development.
- Studied the roles of target mRNA genes in *P. tepidariorum* embryogenesis.
- Compared the possible mechanisms involved in *P. tepidariorum* MZT to what is currently known in other species.

2 MATERIALS AND METHODS

2.1 Bioinformatics

2.1.1 Analysis of RNA Sequencing from *P. tepidariorum* Embryogenesis

To obtain insights into the regulation of mRNAs during *P. tepidariorum* embryogenesis RNA was sequenced from ten developmental stage consisting of stages (S) S1-S4, S5 early and late, S6-S8 and S10 (Mittmann and Wolff 2012). Fastq files of trimmed single-end 100 bp reads from MiSeq Illumina sequencing of two biological replicates for each developmental stages were kindly provided by Hiroki Oda, Osaka, Japan. Read quality was first confirmed using FastQC v0.11.2 (Babraham Bioinformatics 2011). Paired runs for each stage were aligned to the *P. tepidariorum* genome and supplying a reference transcriptome for gene models (Schwager *et al.* 2017) using TopHat v2 (Kim *et al.* 2012). Outputs files were sorted and indexed with Samtools v1.2 (Li *et al.* 2009a) and expression levels were quantified using HTSeq-count (-s no) (Anders *et al.* 2015) and custom Perl scripts. Heatmaps were generated in R v3.2.3 (R Core Team 2015) using the ComplexHeatmap package (Gu *et al.* 2016). To observe read depth, reads were mapped to the *P. tepidariorum* DoveTail genome (Schwager *et al.* 2017) using Bowtie2 v2.2.7 (Langmead and Salzberg 2012), output files were converted to BedGraph files using Bedtools (Quinlan and Hall 2010) and visualised in R v3.2.3 (R Core Team 2015) with the Sushi package (Phanstiel *et al.* 2014).

2.1.2 Identification and Characterisation of Homeobox Genes

Existing protein models from *P. tepidariorum*, *C. sculpturatus* (Schwager *et al.* 2017) and *I. scapularis* (BioProject: PRJNA16232) were obtained for homeobox gene identification. Natascha Turetzek, Göttingen, Germany, kindly provided an assembled nucleotide transcriptome from the spider *Pholcus phalangioides*. The ORFs were first predicted in *P. phalangioides*, using TransDecoder v3.0.0 (Haas *et al.* 2013). Sequence homology and protein domains of predicted ORFs were then analysed respectively with BLASTP v2.2.28+ (-evalue $1e^{-6}$) (Altschul *et al.* 1990), using the UniProt Swiss-Prot database (UniProt 2015), and HMMER v3.1 (Wheeler and Eddy 2013), using the Pfam v30.0 database (Finn *et al.* 2016). The proteins from the four species were then queried using BLASTP v2.2.28+ (Altschul *et al.* 1990) against all homeodomains present in the ten species from HomeoDB (Zhong *et al.* 2008; Zhong and Holland 2011). All hits were scanned for homeodomains using the CDD search tool (Marchler-Bauer *et al.* 2015). Hits that contained at least one homeobox sequence were manually checked for the completeness of the homeobox domain sequences. The homeobox genes were initially annotated according to the best match to the HomeoDB query and classified based on Holland *et al*, 2007. Phylogenetic analysis was then used to better validate these annotations using addition homeodomain sequences from *Strigamia*

maritima (Chipman *et al.* 2014), *Apis mellifera*, *Tribolium castaneum* and *Drosophila melanogaster* (Zhong *et al.* 2008; Zhong and Holland 2011). Alignments of all homeodomains were generated with ClustalW (Larkin *et al.* 2007) and then amino acid substitution models were predicted using the model with the lowest AIC scores in MEGA v7 (Kumar *et al.* 2016). Maximum likelihood phylogenetic analysis was performed on the RAxML BlackBox server (Stamatakis *et al.* 2008) with the LG model and 100 bootstraps and trees were inspected with FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Alignments of each species (*P. tepidariorum*, *P. phalangioides*, *C. sculpturatus*, *I. scapularis*) homeobox gene repertoires were also generated with ClustalW (Larkin *et al.* 2007). Amino acid substitution models were predicted with MEGA v7 (Kumar *et al.* 2016) and maximum likelihood phylogenetic reconstruction was run with PhyML v3.0 (Guindon *et al.* 2010) with 500 bootstraps, maximum likelihood gamma distribution was estimated with the PhyML option –a e. Trees were visualised with FigTree v1.4.2

For alignments of *Cux*, *Hmx*, *Pitx* and *Zfh* homologs, full protein sequences and homeodomains from *Drosophila melanogaster* and *Tribolium castaneum* were aligned with above species and visualised in Geneious v10.2.3 (Kearse *et al.* 2012) (*Dm-Cux*: NP_001138174, *Tc-Cux*: XP_015836586, *Dm-Hmx*: NP_001262674.1, *Tc-Hmx*: XP_967446.1, *Dm-Pitx*: NP_001138130.2, *Tc-Pitx*: CBL87030, *Dm-Zfh2*: NP_524623, *Tc-Zfh*: XP_015834028.1). For these genes individual maximum likelihood trees were made, first predicting amino acid substitution models with MEGA v7 (Kumar *et al.* 2016), then run using PhyML v3.0 (Guindon *et al.* 2010) with 500 bootstraps and visualised with FigTree v1.4.2.

2.1.3 *P. tepidariorum* Culture and Small RNA Sequencing

P. tepidariorum (from a strain collected in Göttingen, Germany) were raised at 25°C and fed on a diet of *Drosophila* and *Gryllodes sigillatus*. Total RNA was extracted from the embryos of ten healthy cocoons, corresponding to stages 1 to 10 of embryogenesis in this spider (Mittmann and Wolff 2012), using QIAzol (Qiagen) following the manufacturer's instructions. Small RNA (~15-30 nt) libraries were prepared with the Illumina TruSeq® Small RNA Sample Prep Kit according to the manufacturer's instructions, and library quality was assessed on the 2200 TapeStation instrument. Size selection between 18-30 bp was performed by excision of RNA from a 6% DNA PAGE gel, 1 mm (Invitrogen). Sequencing was performed on the Illumina HiSeq 2000 platform at the University of Manchester Genomic Technologies facility, and yielded a total of 177,789,607 reads (100 bp). The raw read data has been deposited in the Sequence Read Archive, with accession PRJEB13119.

2.1.4 Analysis of Small RNA Sequencing Data and Identification of *P. tepidariorum* miRNAs

The raw read quality was confirmed using FastQC v0.11.2 (Babraham Bioinformatics 2011), after which adapter sequences were trimmed using the Cutadapt v1.4.1 tool (Martin 2011), and reads

≥16 bp were retained (Fig. 2.1A). Reads were then mapped against the first *P. tepidariorum* genome assembly (GCA_000365465.1) using Bowtie v1.0.0 (Langmead *et al.* 2009). All mapped reads were analysed using miRDeep2 v0.0.5 (Friedländer *et al.* 2008) to predict miRNAs (Fig 2.1A). To preliminarily identify orthologs, the predicted *P. tepidariorum* miRNAs hairpins were then compared against metazoan miRNA precursor sequences from miRBase v21 (Kozomara and Griffiths-Jones 2014) using BLASTN v2.2.28+ (Altschul *et al.* 1990) (Fig 2.1B). Predicted miRNAs that did not have significant sequence similarity to any other miRNA in miRBase were then manually inspected. Those that met the following criteria were discarded from the study: fewer than ten reads mapping to the locus; poorly defined Dicer processing sites, defined as less than 50% of reads for a given mature miRNA having the same 5' end; the predicted mature miRNA sequence had a BLAST hit to more than ten loci in the *P. tepidariorum* genome with two or fewer mismatches. Predicted paralogous miRNAs were aligned using Clustal-Omega (Sievers *et al.* 2011) and then manually inspected in RALEE v0.8 (Griffiths-Jones 2005).

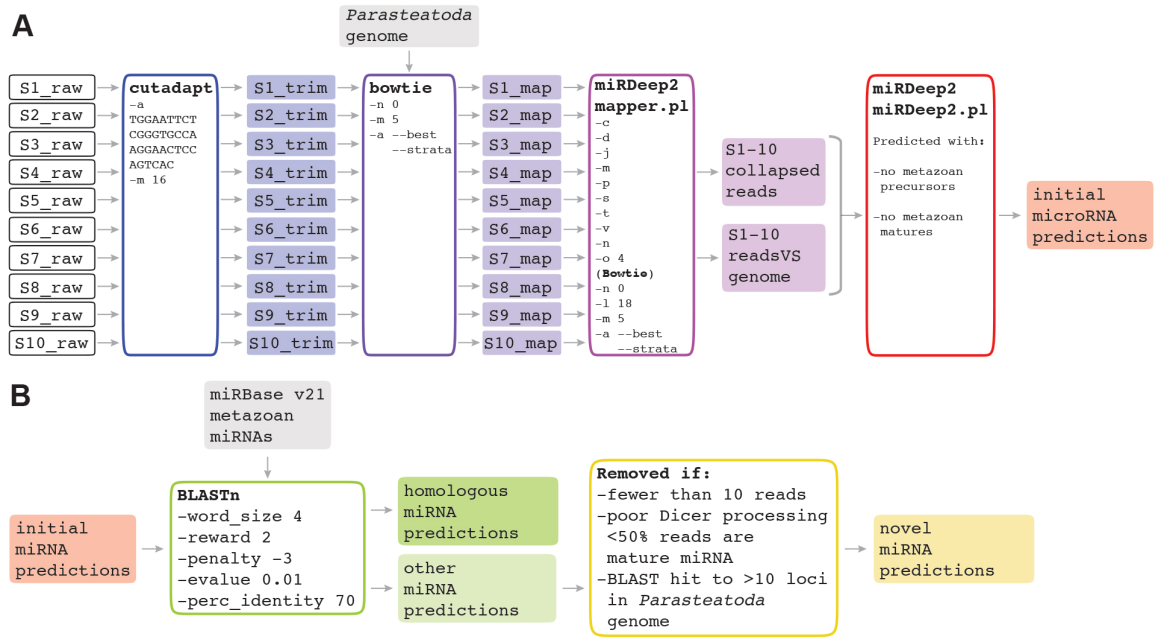


Figure 2.1: MicroRNA prediction and annotation in *P. tepidariorum*. The workflows of the software and their parameters to predict miRNAs in *P. tepidariorum* **(A)** and then annotated them based on sequence similarity to other known metazoan miRNA **(B)**.

2.1.5 Identification of *P. tepidariorum* miRNA Orthologs in Other Species

Orthologs of all *P. tepidariorum* pre-miRNAs were searched for (Fig. 2.2) in the genomes of two other spiders (*Acanthoscurria* geniculata - GCA_000661875.1 (Sanggaard *et al.* 2014), *Stegodyphus mimosarum* - GCA_000611955.2 (Sanggaard *et al.* 2014)), a scorpion (*Centruroides sculpturatus* - GCA_000671375.1), four parasitiforms (*Ixodes scapularis* - GCA_000208615.1 (Lawson *et al.* 2009), *Metaseiulus occidentalis* - GCA_000255335.1 (Hoy 2009), *Rhipicephalus microplus* - GCA_000181235.2, *Varroa destructor* - GCA_000181155.1 (Cornman *et al.* 2010)), an acariform

Tetranychus urticae - GCA_000239435.1 (Grbic *et al.* 2011)) the horseshoe crab (*Limulus polyphemus* - GCA_000517525.1 (Nossa *et al.* 2014)), five mandibulates (*Strigamia maritima* - GCA_000239455.1 (Chipman *et al.* 2014), *Daphnia pulex* - GCA_000187875.1, *Tribolium castaneum* - GCA_000002335.2 (Richards *et al.* 2008), *Apis mellifera* - GCA_000002195.1, *Drosophila melanogaster* - GCA_000001215.4) and a nematode (*Caenorhabditis elegans* - GCA_000002985.3s) using a combination of BLASTN v2.2.28+ (Altschul *et al.* 1990) and INFERNAL v1.1.1 (Nawrocki *et al.* 2009). For each species, the predicted miRNA repertoire was supplemented with known miRNAs from miRBase v21 (Kozomara and Griffiths-Jones 2014) and for *Rhipicephalus microplus* from Barrero *et al.* (2011). For each miRNA, all BLAST and INFERNAL hits were aligned using Clustalw2 (Larkin *et al.* 2007) in RALEE v0.8 (Griffiths-Jones 2005). Putative orthologs were confirmed by manual inspection of the multiple sequence alignment using RALEE v0.8 (Griffiths-Jones 2005), looking for high similarity of one or both mature arms of the hairpin precursor. To determine the structure and minimum free energy of all chelicerate miRNAs detected by BLAST and INFERNAL searches, precursor miRNA sequences were analysed with RNAfold v1.8.4 using default settings (Gruber *et al.* 2008). All predictions retained had greater than -0.2 kcal/mol/nt (Kozomara and Griffiths-Jones 2014). Scaled synteny maps of miRNAs were made in R v3.2.3 (R Core Team 2015) with custom scripts.

2.1.6 Quantification of Reads Associated with miRNAs in *P. tepidariorum*

To identify the number of reads of certain length that map to miRNAs, CDSs, repetitive loci and other genomic regions, reads from each stage were first mapped with Bowtie v1.0.0 (Langmead *et al.* 2009) to the *P. tepidariorum* genome (GCA_000365465.1) with no mismatches (Fig. 2.3). Genomic mapped reads were then mapped against the 148 *P. tepidariorum* miRNAs and unmapped reads from this alignment were mapped against the CDS (Posnien *et al.* 2014). The remaining mapped reads were considered to map to other regions of the genome with no mismatches. Initially unmapped reads were then re-mapped against miRNAs and CDSs in a similar process as before but allowing for one mismatch (Fig 2.3). Expression levels of mature miRNAs were calculated with a custom Perl script that accounted for reads that mapped to multiple miRNA sequences and normalised expression by total miRNA reads per sample. Heatmaps were generated in R v3.2.3 (R Core Team 2015) using the ComplexHeatmap package (Gu *et al.* 2016).

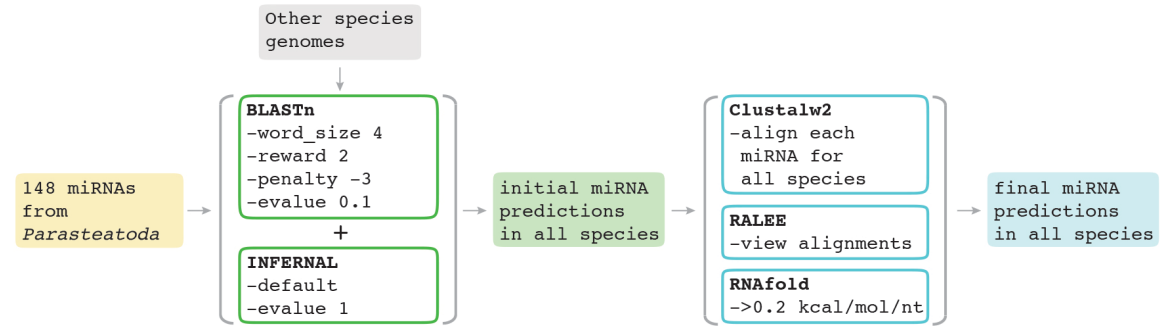


Figure 2.2: Detection of *P. tepidariorum* miRNA orthologs in other chelicerate species.

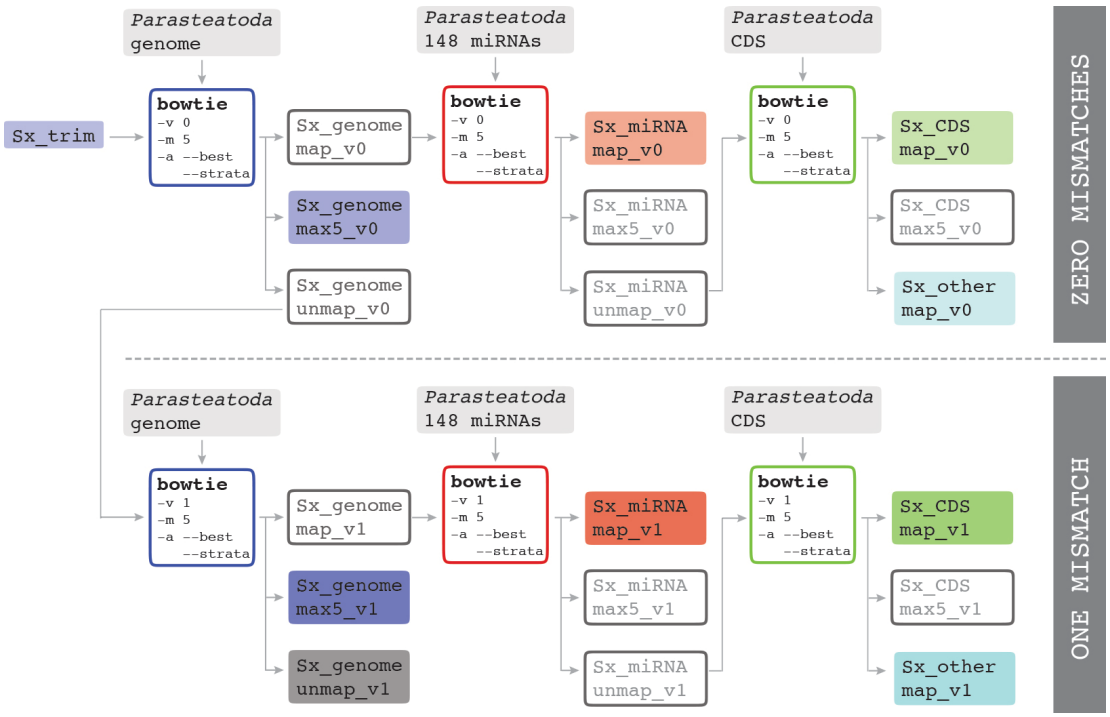


Figure 2.3: Quantification of reads assigned to miRNAs, CDS and other regions of the *P. tepidariorum* genome.

2.1.7 Comparisons of Relative Arm Usage

To compare the relative arm usage of *P. tepidariorum* microRNAs to those of other species a similar approach to Marco *et al.* (2010) was used. The average number of reads per experiment catalogued in miRBase for each microRNA arm for *D. melanogaster* and *T. castaneum* were obtained from miRBase v21 (Kozomara and Griffiths-Jones 2014). For the multi copy microRNA families in *D. melanogaster* and *T. castaneum* the average arm usage across the family was used. Arm usage of *P. tepidariorum* microRNAs with unique 5' and 3' mature sequences was also quantified (App. Table 4.2). *P. tepidariorum* microRNAs with one unique mature sequence (App. Table 4.2) were also analysed, with the caveat that the expression of the other arm was possibly over or under estimated. Note that arm usage of *P. tepidariorum* microRNAs that have non-unique sequences for both mature products could not be reliably assessed because reads could not be unambiguously mapped to one location (App. Table 4.2).

2.1.8 Identification of miRNA Targets During Maternal to Zygotic Transition (MZT)

To identify miRNA repression of maternal mRNAs at the MZT, DESeq2 (Love *et al.* 2014) was used to identify differentially expressed mRNAs between successive stages and perform principal component analysis. Heatmap clustering, using the ComplexHeatmap package (Gu *et al.* 2016), was used to identify clustering of similar stages of both miRNA and mRNA data. Principle component analysis in R v3.2.3 (R Core Team 2015) was used identify transitions in expression profiles of miRNA data. To identify the miRNAs upregulated at the predicted maternal-to-zygotic transition (MZT) the fold change was calculated between S2 and S3. Custom Perl scripts were written to identify negative correlating miRNAs and mRNA. For this, S5e and S5l in the mRNA transcriptome were averaged and paired to S5 from the miRNA sample; and S9 was omitted from the miRNA data because it was absent from the mRNA data set. Target predictions with MFE < -15 kcal/mol were made using miRanda v3.3a (John *et al.* 2004) with the mature miRNAs identified (App. Table 4.2) (Leite *et al.* 2016) and the 3' UTRs from the *P. tepidariorum* transcriptome (Schwager *et al.* 2017). MEME (Bailey *et al.* 2010) was used to identify motifs in the 3' UTRs of target genes predicted at the MTZ with the following parameters; -nmotifs 3 -minw 6 -maxw 20.

2.2 Molecular Biology

2.2.1 Cloning of Gene Fragments and Probe Synthesis

For the synthesis of RNA probes for *in situ* hybridisation (ISH), gene fragments were cloned into TOPO-TA (ThermoFisher Scientific) so that T3 and T7 sites could be used for the transcription of both sense and anti-sense RNA probes. The cDNA libraries for coding genes were generated from QuantiTech (Qiagen) using RNA extracted (Qiazol), following standard protocols, from a complete range *P. tepidariorum* embryo stages (S1-S14). Genomic DNA was obtained with standard salt extraction for cloning of miRNA fragments. Primers (Table 2.1) were used to amplify gene fragments using PCR with the OneTaq® Quick-Load® 2x Master Mix with Standard Buffer (NEB BioLabs) so that the additional adenosine overhangs on PCR fragments would pair with thymine overhangs in the TOPO-TA vector. TOPO-TA cloning followed all standard protocols, except reaction and cell volumes were halved. Cells were plated on 100 µg/mL ampicillin or 50 µg/mL kanamycin agar (Lennox), colonies were checked with M13 primers, and positive colonies were grown overnight at 37˚C in 4 mL of 100 µg/mL ampicillin or 50 µg/mL kanamycin LB broth (Lennox). Plasmids were extracted with the Plasmid DNA Mini Kit (EZNA) and were sequenced (Eurofins Genomics and SourceBioscience). To validate cloning and select best plasmids, sequences were aligned to the *P. tepidariorum* genomes (Schwager *et al.* 2017), and when available, relevant transcriptome (Posnien *et al.* 2014; Schwager *et al.* 2017) or miRNA sequences (Leite *et al.* 2016), using ClustalW (Larkin *et al.* 2007) and/or BLASTN v2.2.28+ (Altschul *et al.* 1990). Anti-sense RNA probes were transcribed with T3 (11031163001 - Roche)/T7 polymerase (10881775001 - Roche), with DIG RNA labelling mix (11277073910 - Roche), from clean PCR fragments generate

using M13 primers and TOPO-TA clones, following standard protocols. Probes were purified using a lithium chloride precipitation, by adding 2x LiCl to transcription reaction and storing at -20ºC for > 2 hours, centrifuged at 14,000 rpm @ 4ºC for 30 mins. The supernatant was carefully removed and the pellet was washed with 70% ethanol and vortexed. Centrifugation at 14,000 rpm @ 4ºC for 15 mins to reform pellet and discard the supernatant, air dry and resuspended in RNAase free water.

2.2.2 *In Situ* Hybridization (ISH)

Colourmetric ISH was performed as previously described (Akiyama-Oda, 2003). Briefly, embryos were dechorionated in bleach, washed in water then fixed in 5.5% formaldehyde:PEMS:heptane. Fixations were stopped, washes with PEMS-tween 20 (0.1%) (Sigma - P9416) and transitioned to 100% methanol and stored at -20oC. Embryos were rehydrated into PBS-t and vitelline membranes removed then fixed for 20 minutes in 5% formaldehyde:PBS-t. After washing embryos were incubated in proteinase K (Roche – 03115828001) for 2 minutes then fixed for 20 minutes in 5% formaldehyde:PBS-t. Embryos were washed in PBS-t then pre-hybridised at 65◦C for an hour and then hybridised with probes overnight at 65◦C. After multiple washes in PBS-t, embryos were blocked (Roche - 1096176) for 30 minutes then anti-DIG Fab-AP (1:2000) (Roche – 11093274910) or anti-Fluorescein Fab-AP (1:2000) (Roche – 11426338910) was added for 3-4 hours. After washes in PBS-t then staining buffer, ISH was developed using NBT/BCIP (Roche – 11681451001) for DIG probes or INT/BCIP (Roche – 11681460001) for Fluorescein probes. Embryos were counterstained with DAPI (Roche – 10236276001) for ~20 mins to visualise nuclei. Images were taken on a Zeiss Axio Zoom V.16.

Table 2.1: Primers used for cloning DNA fragments for probe synthesis.

Gene	Forward primer	Reverse primer	Fragment size (bp)
mex3	TCCAGTATTCCAGGCCAAACT	TCTGACGGGACCAAAATCTGA	633
mir-10	GCCTAAAGTAAGTATAGCTATTAACAGTCT	CATCTAAAGAATACCATATTTTAAACCCGT	2063
Scr-A	ATGAGTTCGTATCAGTTCGTAAA	ATAGTTGATGCCATCTTATGTTCC	806
Scr-B	TTTGTGAACTCATTAACTTCTGCTA	AGTCTGAACTCACCGTGTAAGT	908
Cux1	GTTGGGCCTATCTATCCGCACT	CCACTGCTTGATGGAGGTGGTA	756
Cux2	GGAGCAGAAGGAAGCCTTGAGA	TCCTGTTCGTTACCGCTGTCAT	774
Hmx1	ACCTCCCGCATTTTTACCGGTA	ACTAGGTAGAGGGGCACGTGTA	709
Hmx2	GCACCTTCGATAGCCACAAG	AGCTTCTATGTCAGCTGCCA	703
Pitx1	GACCCATCACAGACTGGACA	CCAGCATATGTCCCAAAGCC	804
Pitx2	ACTTGCCTGACTCAGCATCT	TAAGGACAAGGAGGTGCAGG	629
Zfh1	CCAGAGACTACCCCAACACCTG	TAGGCAATGGCGTAACACCAGA	568
Zfh2	AAGTTCCTTTGAGCATCGCAGC	GTCTCGGTCGAGGAAATGGACA	501

3 Duplication of Homeobox Genes in Spiders and Scorpions

Homeobox genes encode an important group of transcription factors distinguished by a homeodomain and are known to regulate many aspects of development. Gene duplication has played an important role in the expansion of homeobox repertoires in metazoans (Garcia-Fernandez 2005). The duplication of single homeobox genes can lead to evolutionary changes, for example *Hox3* in insects (McGregor 2005; Ferguson *et al.* 2014). Furthermore, large-scale duplication such as WGD can be identified by the presences of multiple Hox clusters, such as those seen in vertebrates (Hoegg and Meyer 2005; Mungpakdee *et al.* 2008; Pascual-Anaya *et al.* 2012). Studying homeobox genes in chelicerates offers continued insights into understanding how gene duplications have contributed to chelicerate homeobox repertoires and what their roles are in the morphological evolution of these animals.

Recent genome and transcriptome sequencing of chelicerate species have identified many duplicated homeobox genes in certain lineages (reviewed in Leite and McGregor 2016). For example, genome sequencing in three horseshoe crab species has identified the presence of many duplicate homeobox genes, including the presence of four Hox clusters, indicating possibly two rounds of WGD in the ancestor of these animals (Nossa *et al.* 2014; Kenny *et al.* 2015a). Additionally, genome sequencing of *Mesobuthus martensii* and systemic surveying of homeobox genes in this scorpion identified pervasive duplication of homeobox genes and two complete Hox clusters (Cao *et al.* 2013; Di *et al.* 2015). Further transcriptome sequencing of many other scorpion species also provide support that this lineage of chelicerates contains duplicated Hox clusters (Sharma *et al.* 2015). Spiders are another lineage of chelicerates where multiple duplicated genes have been reported. These have included duplicates of Hox genes and other homeobox genes (Schwager *et al.* 2007; Pechmann *et al.* 2015; Samadi *et al.* 2015; Schomburg *et al.* 2015; Schwager *et al.* 2017; Turetzek *et al.* 2017). However, there has not yet been a systemic characterisation of all homeobox genes in spiders and a comparison to other chelicerate homeobox repertoires.

Therefore, this chapter aimed to systematically identify homeobox gene repertoires, particularly in spiders and further explore the scorpion repertoires of these genes. This was then used as a platform to assess the evolutionary features of these genes such as lineage specific retention and loss, expression levels, spatial expression differences and synteny relationships.

3.1 Comparison of Homeobox Families in Chelicerates and other Arthropods

To systematically identify homeobox repertoires in chelicerate species the characteristic homeodomain motif was searched for within protein predictions from transcriptomes of the two spiders *P. tepidariorum* (App. Table 3.1 and App. Fig. 3.4) and *Pholcus phalangioides* (App. Table 3.2 and App. Fig. 3.5), a scorpion *Centruroides sculpturatus* (App. Table 3.3 and App. Fig. 3.6) and a tick *Ixodes scapularis* (App. Table 3.4 and App. Fig. 3.7). A total of 80, 78, 82 and 70 homeobox families were identified respectively, indicating a comparable dataset from each species relative to other arthropods, including another scorpion (*Mesobuthus martensii* – 82) (Di *et al.* 2015), a myriapod (*Stigamia maritima* – 83) (Chipman *et al.* 2014) and two insects (*D. melanogaster* - 80 and *T. castaneum* - 80) (Zhong *et al.* 2008; Zhong and Holland 2011). All of the families identified were derived from transcriptomic sequences and predicted protein models, it is therefore possible that pseudogenes may be present but this was not considered in this analysis.

Maximum likelihood trees of the homeodomains from arthropod ANTP (App. Fig. 3.1), PRD (App. Fig. 3.2) and other (App. Fig. 3.3) classes generally formed monophyletic groups of homeobox families, suggesting reasonable robustness in the annotation of chelicerate homeobox genes. Comparisons of families between these repertoires suggest particular patterns of retention and loss of homeobox families in arthropod lineages (Fig. 3.1). Overall, 52 of the 87 homeobox families were present in all species surveyed. The HNF class and Dmbx family, which is present in vertebrates (Zhong *et al.* 2008; Zhong and Holland 2011), are also present in chelicerates, the myriapod (Chipman *et al.* 2014) but are absent from insects (Fig. 3.1). The only family suggested to be absent from chelicerates compared to other arthropods surveyed was Pax2/5/8. In contrast chelicerates contain the Barx family, which is absent from the myriapod and insects surveyed. Furthermore, both scorpions have support for two additional families; the Nk8 family, which is a ANTP class, and groups closest with the Scro family; and an additional SINE class family Six7 that grouped close to Six4/6 and also contain SIX domains (Fig. 3.1, App. Fig. 3.2 and App. Table 3.3). This family may be specific to scorpions or have been present in the arachnopulmonate ancestor and subsequently lost in spiders.

There were also some families that were missing from spiders but present in other chelicerates and arthropods, including Bsx, Hlx, Mslsx and Mkx. The only family retained in the insects, a myriapod and spiders, but not other chelicerates was Nedx (Fig. 3.1). Of the repertoires compared the tick had the fewest families. This is consistent with finding fewer coding genes and miRNA families present in this lineage (Fig. 4.2) (Schwager *et al.* 2015; Leite *et al.* 2016). Finally, specific BLAST searches for PROS class genes, identified a *Pros* gene (MMa30254) in *M. martensii*, which had previously not been identified (Di *et al.* 2015) (Fig. 3.1).

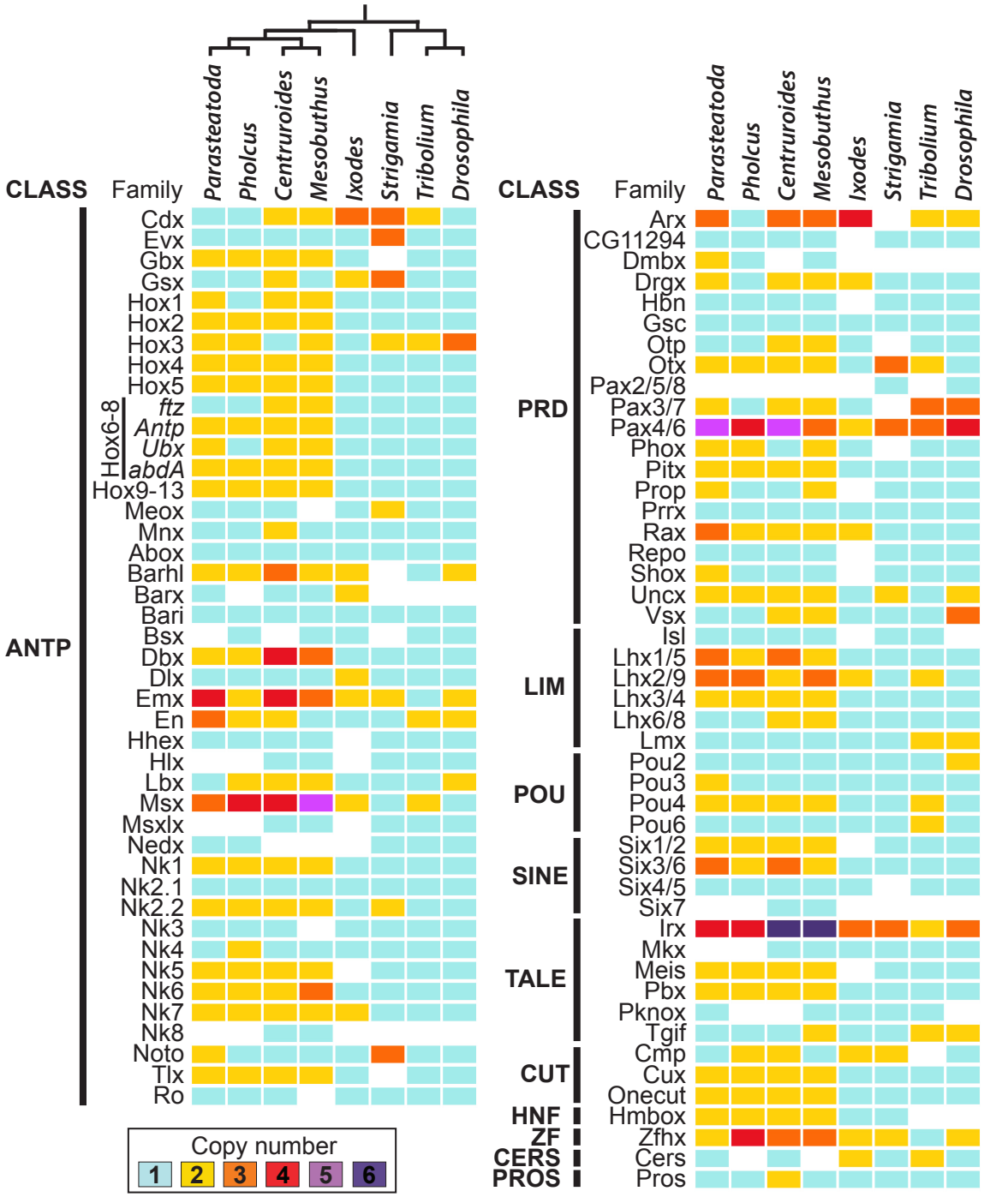


Figure 3.1: Comparison of homeobox repertoires in arthropods reveals pervasive duplication in arachnopulmonates. The copy number of homeobox families is generally greater in arachnopulmonates compared to other arthropods across all classes, except Cers and Pros. Homeobox genes are grouped based on Holland *et al*, (2007) and number of paralogs in each family is colour coded. The Hox6-8 family has been broken down further to show specific copy number of *ftz*, *Antp*, *Ubx*, and *abdA*.

3.2 Pervasive Duplication of Homeobox Genes in Arachnopulmonates

3.2.1 Arachnopulmonates Share Many Duplicated Homeobox Families

Despite the number of homeobox families being fairly similar between arthropod species surveyed, the actual number of genes varied considerably between arachnopulmonate species and other arthropods. The two spiders *P. tepidariorum* and *P. phalangioides* contained 145 and 132 homeobox genes, respectively, and the scorpion *C. sculpturatus* contained 156 (App. Table 3.1-3.3), as does *M. martensii* (Di *et al.* 2015). However the tick (App. Table 3.4), centipede (Chipman *et al.* 2014), beetle and fly contain 96, 113, 105 and 104, respectively (Zhong *et al.* 2008). This indicates that many more of the arachnopulmonate homeobox families are comprised of multiple genes copies compared to other arthropods. Indeed, 56%, 49%, 59%, 57% of homeobox families in *P. tepidariorum*, *P. phalangioides*, *C. sculpturatus* and *M. martensii* were duplicated, compared to 20% in the tick, 16% in the centipede, and 15% in the beetle and fly. In total, there were 34 families that were duplicated in all four arachnopulmonate species (Fig. 3.1), which may indicate that these were commonly duplicated and subsequently maintained in both lineages. Interestingly, 21 of these 34 families were not duplicated in any non-arachnopulmonate species surveyed. Furthermore, there were 38 families that were duplicated in both spiders, whereas 46 families were duplicated in both scorpions (Fig. 3.1). However, due to the assembly quality of the *C. scrulpturatus* genome, it is unclear if scorpions have retained more duplicated families than spiders, since the possible shared duplication event, or whether they have undergone further lineage specific duplications.

Of the families that were duplicated in both insects, with the exception the TALE class Tgif family, all of them were also duplicated in arachnopulmonates, suggesting these may have been duplicated in the ancestor of all arthropods, or were independent duplications in the different lineages. However, such families like Emx, Msx, Arx, Pax4/6 and Irx, which are duplicated in non-arachnopulmonates species, are present in more than two copies in arachnopulmonates (Fig. 3.1), indicating that there has been a further round of duplication specific to arachnopulmonates.

3.2.2 Duplicated Hox clusters in Arachnopulmonata

The ANTP group is the largest homeobox class and is split into the HOXL and NK subclasses (Holland *et al.* 2007). The HOXL subclass contains the famous anterior-posterior patterning Hox genes, which commonly group together as a cluster within arthropod genomes (Fig. 3.2 and 3.3A). In most extant arthropods there are usually single copies of each Hox gene. However the Hox3 family has independently undergone tandem duplications in insect lineages, such as in higher dipterans to lead to *bicoid* (McGregor 2005), and *Shx* paralogs in lepidopterans (Ferguson *et al.* 2014). Like other arthropods, the *I. scapularis* genome contains single copies of all arthropod Hox genes arrange in a cluster (with no tandem duplications of the Hox3 family) (Fig. 3.2). However, in contrast to the tick and other arthropods, many of the Hox genes are duplicated in

arachnopulmonates species (Fig 3.1 and 3.2). Two copies of Hox2/*labial* (*lab*), Hox4/*Deformed* (*Dfd*), Hox5/*Sex-combs reduced* (*Scr*), two Hox6-8 subfamilies *Antennapedia* (*Antp*) and *abdominal-A* (*abdA*) and Hox9-13/*Abdominal-B* (*AbdB*) are present in all the arachnopulmonate species surveyed (Fig. 3.1). There were also two copies of *Hox3* in both spiders and *M. martensii*, but only one copy was recovered from *C. sculpturatus*. Only one copy of Hox6-8/*Ultrabithorax* (*Ubx*) was identified in *P. phalangioides*, whereas two copies were identified in *P. tepidariorum* and both scorpions. Another Hox6-8 member *fushi tarazu* (*ftz*) was also present in two copies in scorpion genomes, however only a single copy was present in each of the spiders (Fig. 3.1). Previously expression patterns analysis of all *P. tepidariorum* Hox genes found that paralogs have differences in both their temporal and spatial patterns (Schwager *et al.* 2017).

Synteny analysis of these Hox genes identified the presence and arrangement of Hox clusters in chelicerate genomes. In the tick genome there is a single scaffold (NW_002784743.1) containing all ten Hox genes (Fig. 3.2). However, in each of the *P. tepidariorum*, *C. sculpturatus*, and previously described *M. martensii* genomes (Cao *et al.* 2013; Di *et al.* 2015), the paralogous Hox genes can be assembled into two putative clusters (Fig. 3.2). In *C. sculpturatus* the nineteen Hox genes are present on twelve scaffolds (Fig. 3.2), whereas the DoveTail assembly of the *P. tepidariorum* genome contains nineteen Hox genes on just three scaffolds (Fig. 3.3A). One of these scaffolds, #833, contains representatives of all Hox genes (cluster A), except *ftz*, and the other cluster contains all ten Hox, but they are found on two scaffolds breaking the cluster between *Pt-Hox3-B* and *Pt-Dfd-B* (cluster B) (Fig. 3.3A). The transcriptional orientations of *P. tepidariorum* Hox genes are all in the same direction, indicating that there have been no inversions that alter transcription orientations, as seen in some insects (Fig. 3.3A) (Pace *et al.* 2016). The size of these two clusters, at around 3 Mb, is generally larger than that of other arthropods, except acariforms and parasitiforms, and the fragmented clusters in *Drosophila* and *Bombyx mori* (Pace *et al.* 2016).

To investigate if there is further conservation of protein coding genes between the two Hox clusters of *P. tepidariorum*, the putative (non-homeobox) gene models that mapped to the three Hox cluster scaffolds were also analysed for paralogy. In total the three Hox scaffolds contained 98 non-Hox related transcripts (Fig. 3.3A and App. Table 3.5). However, even though five groups of paralogs could be identified (at least 70% similarity and e-value > $9x10^{-35}$), there were no transcripts that had homology and also shared similar position in the different clusters. The only set of paralogs (*e*) that were present in both cluster A and B were found at different positions in each cluster (Fig. 3.3A). The four other groups of paralogs (*a* – *d*) were tandem duplicates all located to cluster A between *Hox3* and *Dfd* (Fig. 3.3A). Despite the lack of retention of non-Hox gene duplicates in *P. tepidariorum* Hox clusters, it is interesting that cluster A contains a series of tandem duplicates in a region where cluster B has either failed to be sufficiently assembled or perhaps has been fragmented in the genome. Potentially the retention of tandem duplicates in this region of cluster A may have been selected for due to a prior loss of this region in cluster B.

However, another possibility is that these tandem duplicated genes have ended up in the Hox cluster A after the WGD and are therefore unique to this cluster.

In addition, within the arthropod Hox clusters are microRNAs, *mir-993*, *mir-10* and *iab-4/8* (Fig. 3.2 and 3.3A). Only *iab-4/8* was retained in both *P. tepidariorum* clusters, where as a single copy of mir-993 is present in cluster A and a single copy of *mir-10* in cluster B. Unlike the *iab-4/8* paralogs and *mir-993b-1*, the *mir-10* microRNA was found within the intron of a gene (aug3.g17638). ISH with a probe designed to target the intronic region of this gene and thus the nascent *mir-10* transcript revealed expression similar to *Scr-B* expression (Fig. 3.3B - M). This suggests this gene containing *mir-10* may be under similar regulatory control as *Scr-B* at this point in development.

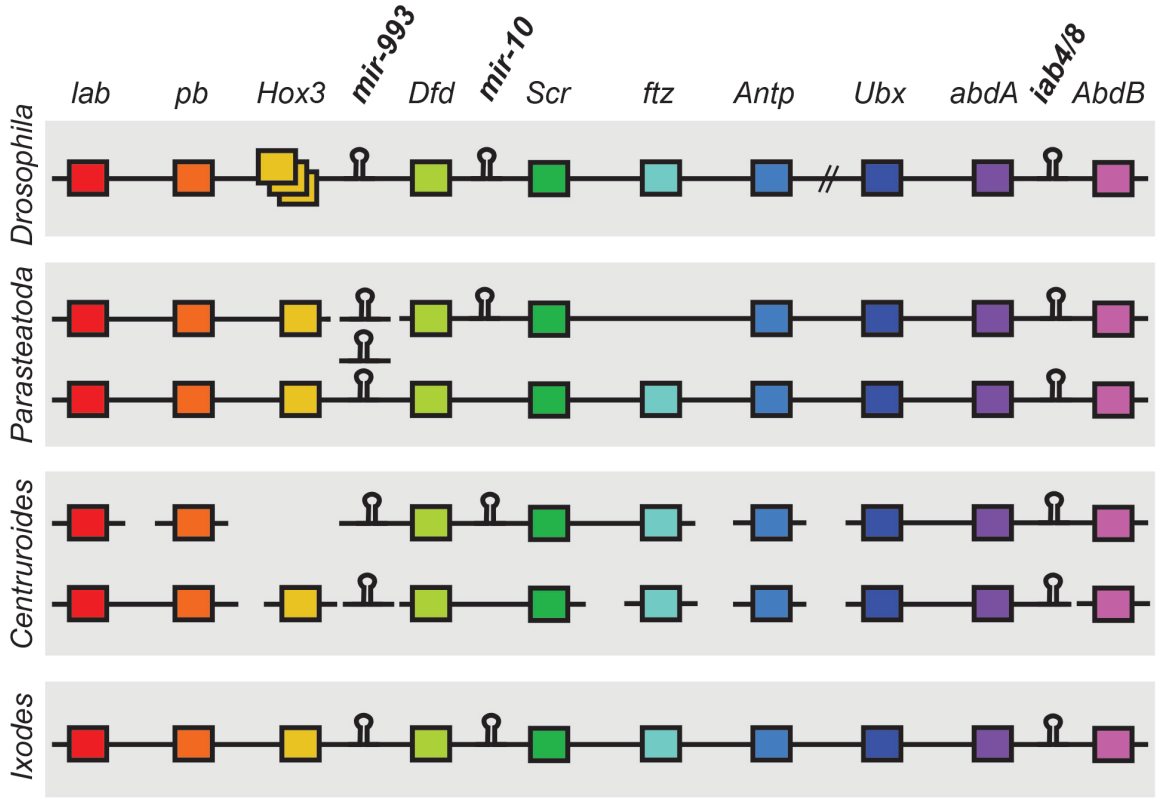


Figure 3.2: Duplication of Hox clusters in arachnopulmonates. In both *Drosophila* and *I. scapularis* genomes there is a single Hox cluster. *I. scapularis* contains single copies of each Hox genes. *Drosophila* has had tandem duplications of *Hox3*. In *P. tepidariorum* and *C. sculpturatus* all Hox genes are duplicated, except *ftz* in *P. tepidariorum* and *Hox3* in *C. sculpturatus*. Black lines represent scaffolds. Hairpin-loops are miRNAs commonly found in the arthropod Hox cluster. The double forward slash in the *Drosophila* cluster marks the break in the cluster.

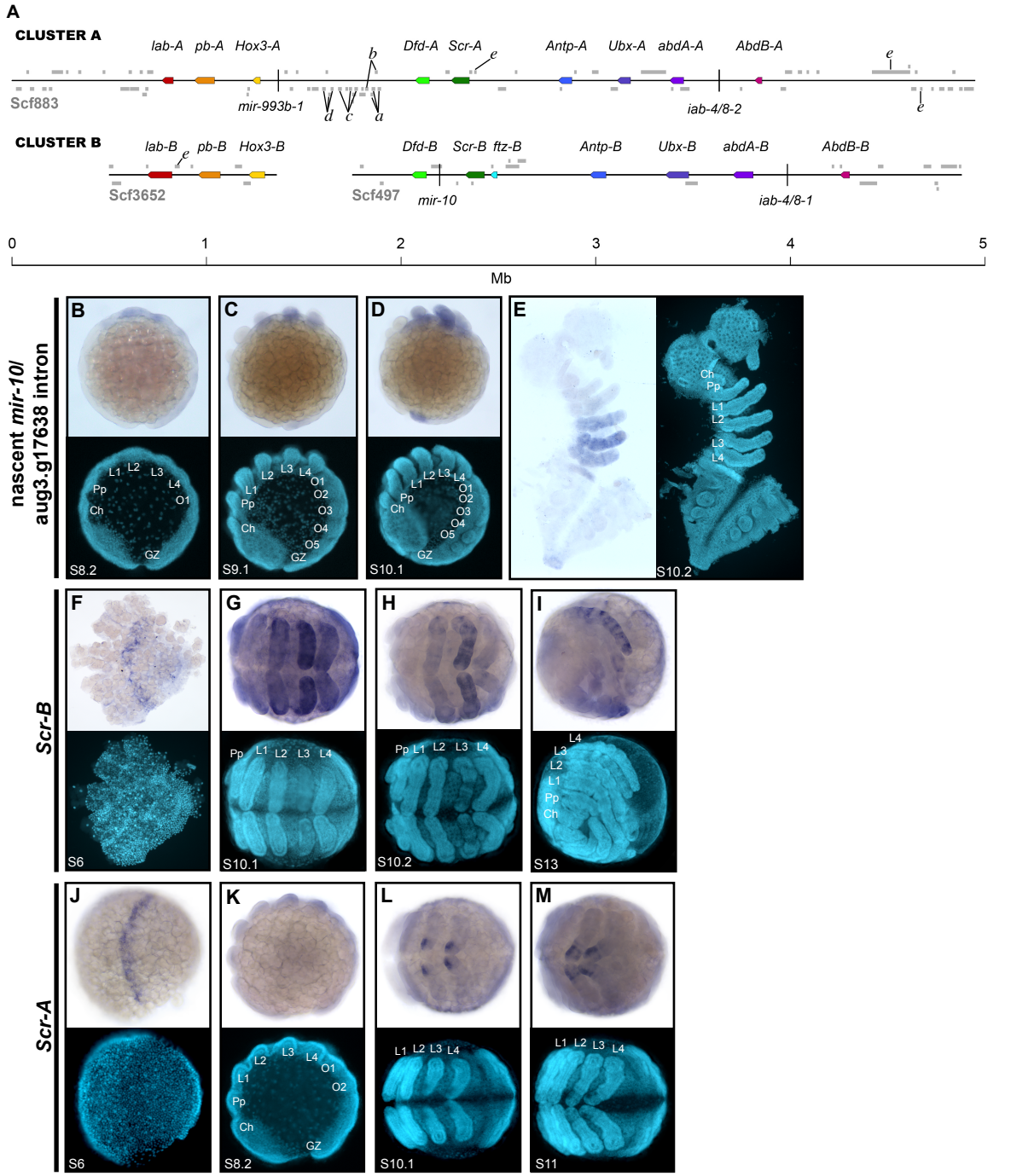


Figure 3.3: Duplicated Hox clusters in *P. tepidariorum*. **(A)** Cluster A is composed of a single scaffold #883, and cluster B is made of two scaffolds #3652 and #497. Hox paralogs have matching colours and arrow indicates the transcription orientation. Non-Hox proteins are shown in grey, with positive strand genes on bottom and negative strand on top. All miRNAs were on the negative strand, shown by perpendicular lines. The single copy of *mir-10*, present in cluster B, appears to be transcribed within an intron of a gene (aug3.g17638). **(B – E)** An ISH against the intron of aug3.g17638, where the *mir-10* sequence is located, produces an expression pattern similar to *Scr-B* **(F – I),** not *Scr-A* **(J – M)**. Bands of expression of both mir-10 **(E)** and Scr-B **(H)** can be seen in L3. **(B-D** and **F-M)** Brightfield images are above and corresponding DAPI stains are below, except for **(E)** which is adjacent. **(B-D** and **G-M)** Wholemount embryos and the S6 flatmount **(F)** are orientated with the anterior to the left. **(E)** Flatmount of embryo orientated with anterior at the top, and the prosomal appendages were purposely removed to ease the flattening of the embryo. Non-Hox paralog groups in **(A)**: ***a*** - aug3.g15609, aug3.g15611, aug3.g15612; ***b*** - aug3.g15610, aug3.g15613; ***c*** - aug3.g15617, aug3.g15620, aug3.g16686; ***d*** - aug3.g16685, aug3.g16684, aug3.g16682; ***e*** - aug3.g15604, aug3.g6142, aug3.g6146, aug3.g27129.

3.2.3 WGDs and Tandem Duplications of Homeobox Genes in *P. tepidariorum*

The presence of duplicated Hox clusters in arachnopulmonates is suggestive of a large-scale duplication event. To identify if there was further support for a large-scale duplication event, the *P. tepidariorum* genome was surveyed for other clusters of homeobox genes and dispersed paralogs (i.e. present on different scaffolds).

Of the 45 families that were duplicated in *P. tepidariorum* there were 40 families that are represented by dispersed paralogs. Some of these dispersed paralogs were present as duplicated clusters in the genome. There were two clusters that contained both *Nk7* and *C15* paralogs, with the same transcriptional orientation conserved between clusters (Fig. 3.4A). Within these two clusters were also other ANTP class genes, but aside from *Msx3* they were all single copy homeobox genes. Another ANTP cluster was also duplicated, with two clusters containing two tandem duplicates of *Emx* paralogs (Fig. 3.4A). Within both clusters the transcriptional direction of *Emx* paralogs was the same. These duplicated ANTP clusters along with the duplication of the Hox clusters provide the compelling evidence for a large-scale duplication. There were some other potential clusters of homeobox genes that were duplicated and retained to some extent. The first of these is the Irx family, where two *mirr* paralogs that were identified on one scaffold and a single copy of *mirr* was present on another scaffold along with *Dmbx2* and *Ap3* (Fig. 3.4C). The second potential duplicated cluster being a scaffold containing two *Lhx1/5* paralogs and another scaffold with a single copy of *Lhx1/5* and one of the *Hgtx* paralogs (Fig. 3.4D). Lastly, there is evidence of a duplicated SINE/Six cluster on scaffolds #121 and #1185. This cluster, found in protostomes and deuterostomes, is usually composed of three genes commonly found in the order *Optix*, *so* and *Six4* (Ferrier 2016). On both of the two scaffolds there are *so* genes, and then one paralog of *Optix* on scaffold #121 and the single copy of *Six4* on scaffold #1185. There are also other paralogs of *Optix* in *P. tepidariorum* but they were dispersed in the genome. Interestingly in insects and myriapods the SINE/Six has degraded and all three genes are dispersed in the genome (Chipman *et al.* 2014; Ferrier 2016). This suggests that the SINE/Six cluster was perhaps present in the arthropod ancestor and then has subsequently been degraded in mandibulates but retained in chelicerates. The clusters of ANTP, PRD, SINE, TALE, and LIM class genes in *P. tepidariorum* suggests that spiders have retained many features of the hypothetical clustering of homeobox genes in the bilaterian ancestor (Ferrier 2016). Furthermore, several of these clusters are duplicated and there are different amounts of gene loss, for example there is relatively low loss in the Hox cluster compared to the NK cluster.

There were in total eight homeobox families with tandemly duplicated paralogs. These were the BarH, Lhx5/9, Pax4/6, Prop and Shox families and the aforementioned mentioned Emx, Irx and Lhx1/5 families (Fig. 3.4). These tandem duplicates were all maintained in the same transcriptional orientation apart from the Pax4/6 cluster. This means that of the retained duplicate homeobox families, 50% were dispersed paralogs, where as only 6% were due to tandem duplications. Additionally, as well as the maintenance of clusters of related ANTP, PRD, TALE,

LIM class genes (Fig. 3.4A-D), there were also eight scaffolds containing at least two homeobox genes from different families (Fig. 3.4E), which may reflect the quality of the DoveTail assembly as some of these scaffolds are larger than the N50 scaffold size (Schwager *et al.* 2017). Overall these synteny relationships imply a greater contribution of WGD than tandem duplication to the expansion of arachnopulmonate homeobox repertoires.

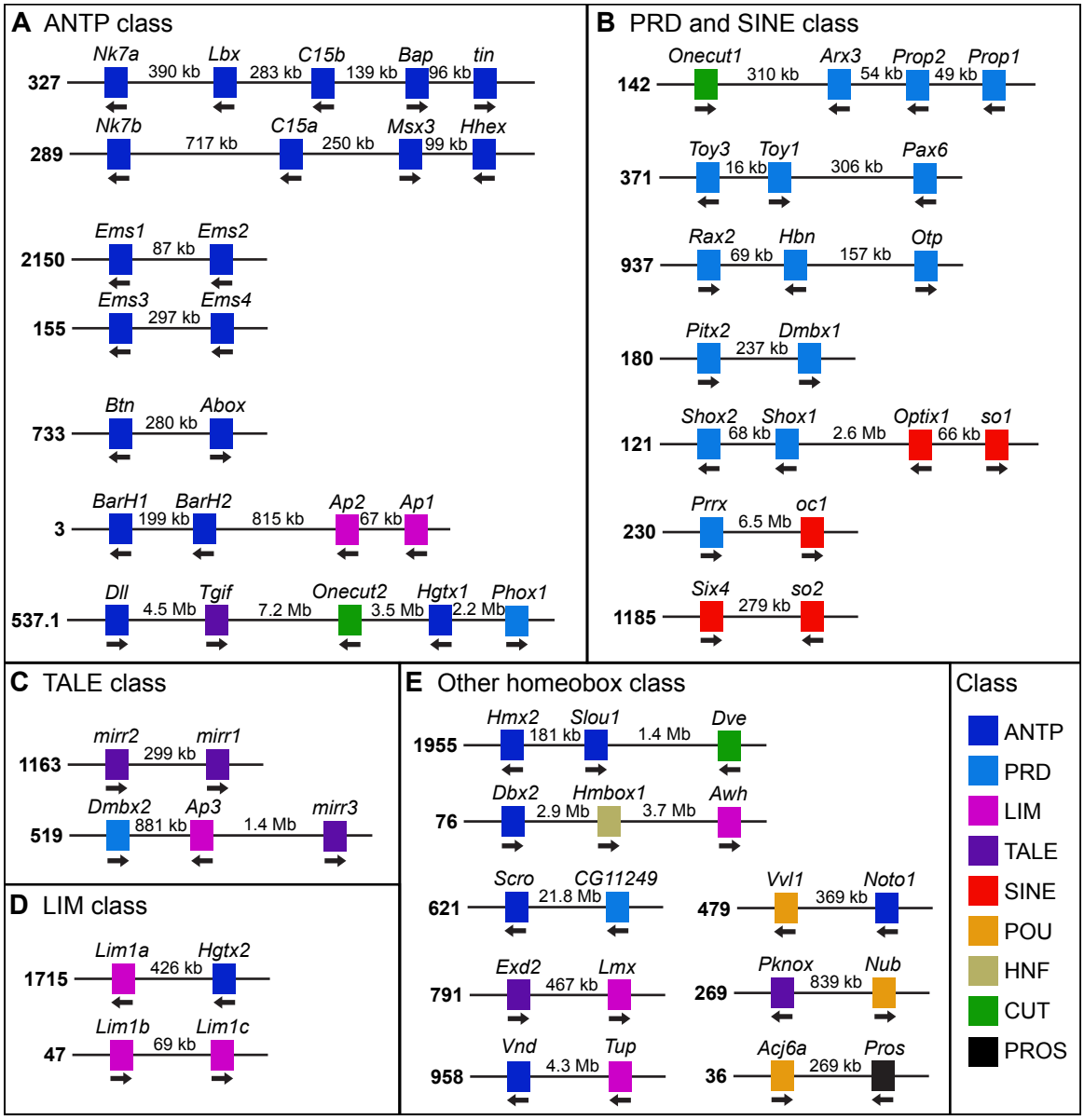


Figure 3.4: Homeobox gene clustering in the *P. tepidariorum* genome. (A) Scaffolds containing at least two ANTP class genes. **(B)** Scaffolds containing PRD and SINE class gene clusters. **(C)** Scaffolds containing the Irx family of the TALE class. **(D)** Scaffolds with Lhx1/5 family of the LIM class. **(E)** Other scaffolds with at least two homeobox genes. All other homeobox genes were located to independent scaffolds. The intergenic distances are indicated in kb or Mb. *P. tepidariorum* DoveTail assembly scaffold numbers are to the left of each cluster. Arrows depict the direction of transcription.

3.3 Expression of Homeobox Genes in a *P. tepidariorum* Embryogenesis

142 out of 145 annotated P. tepidariorum homeobox genes, were expressed in at least one of the ten embryonic stages assayed using RNA sequencing (the exception being one transcript belonging to *Slou2)* (Fig. 3.5). Furthermore, there were no AUGUSTUS gene models that correspond to *Ubx-B* and *abdA-A* (see Materials and Methods). During the course of the time series there is an increase in the average expression of homeobox genes from S1 to S2 (Fig. 3.5A). There is also an increase in the number of genes expressed > 1 log2(RPKM) between the first two stages. This increase is likely the onset of zygotic transcription of homeobox genes (discussed further in Chapter Five). After S2 both the average expression level and number of genes expressed decreases to the lowest levels around early S5 (Fig. 3.5A) after which the number of genes expressed increases and the average expression also increases (Fig. 3.5A).

Interestingly, one homeobox gene that was highly expressed at S1 was *Distal-less* (*Dll*). In *P. tepidariorum Dll* has evolved a novel gap gene like function, with knockdown phenotypes resulting in the loss of the L1 segment. Pechmann *et al.* (2011) noted the early detectable expression of *Dll* at early S5, prior to the development of limb primordia and therefore earlier than expected due to its conserved role in distal appendage formation (Scholtz *et al.* 1998). However, from the RNA-Seq profiling, it appears that *Dll* is expressed from S1, much earlier than previously detected by ISH. Therefore it is possible that *Pt-Dll* is maternally deposited into oocytes, similar to gap genes in other species (Lavore *et al.* 2012).

Other noteworthy genes are *caudal* (*cad*) and *even-skipped* (*eve*), of which earliest onset of expression was previously detected in the caudal lobe around S6 (Schönauer *et al.* 2016). However expression of *Pt-cad* and *Pt-eve* was earlier detected earlier in the transcriptome at S1 and then increasing at S2. This perhaps represents maternal transcripts and then the onset of zygotic transcription of these genes.

3.4 Expression Divergence of Duplicated *P. tepidariorum* Homeobox Genes in the Embryonic Transcriptome

Following gene duplication events, retained paralogs may diverge, leading to sub-functionalisation and neo-functionalisation (Ohno 1970; Lynch and Katju 2004; Amoutzias and Van de Peer 2010; Innan and Kondrashov 2010; McGrath *et al.* 2014). These changes can ultimately rewire gene regulatory networks that affect developmental processes. An important aspect of the evolution of paralogous genes is how they diverge in the timing and location of expression during embryogenesis (Taylor and Raes 2004; Li *et al.* 2009b; Turetzek *et al.* 2016a). During the duplication of genetic material *cis*-regulatory information may or may not be duplicated along with coding regions, which may affect the way paralogs are regulated. In the event of large-scale duplications like WGD, the *cis*-regulatory regions controlling expression of paralogs are likely to

have been duplicated in their entirety, but subsequently accumulate mutations that will change the expression of paralogs. Similarly, coding changes can also change the function of paralogs. To assess the divergence in the expression of *P. tepidariorum* homeobox genes, the expression of paralogs was compared using embryonic transcriptome data set covering the first ten stages of embryogenesis. Furthermore, the expression patterns of a subset of genes representing dispersed duplicates were assayed in *P. tepidariorum* embryos.

3.4.1 Divergence of Dispersed Paralogs in a *P. tepidariorum* Embryogenesis

The RNA-Seq profiling was then analysed to compare the expression of dispersed and tandem paralogs in *P. tepidariorum*. However, it is worth noting that RNA-Seq data generated from whole embryo sequencing does not give sufficient information about the expression of paralogs, so far as spatial patterns and the concentration of transcripts per cell. Therefore inferring the similarities or differences from RNA-Seq as to whether paralogs are divergent or not should be treated with caution.

The spatial and temporal expression of Hox paralogs in *P. tepidariorum* was previously analysed using ISH and showed that Hox genes from both clusters are predominantly expressed in the standard collinear fashion across the AP axis (Schwager *et al.* 2017). Interestingly though, both ISHs and RNA-Seq profiling reveal that one paralog of each Hox gene is always expressed earlier than the other, except for *abdA* paralogs (Schwager *et al.* 2017). Overall the timing of Hox expression in the RNA-Seq data > 1 log2(RPKM) matches well with onset of ISH expression pattern (Fig. 3.5B). However, both *lab-A* and *Dfd-A* were highly expressed in from S1 onwards, indicating earlier expression than detected by ISH (Schwager *et al.* 2017). These results continue to support that *P. tepidariorum* Hox paralogs have diverged in expression characteristics and possibly function roles (Pechmann *et al.* 2015; Schwager *et al.* 2017).

Other dispersed paralogs that were present in clusters were the NK class families Nk7 and C15 (Fig. 3.4A). The *Nk7* paralogs are both expressed at very low levels throughout most of embryogenesis apart from S10 when they both increase in expression (Fig. 3.5B). The *C15* paralogs however show divergence in their temporal and levels of expression, with *C15b* showing increased in expression around S7 to S10, compared to *C15a*, which is barely expressed throughout any of the ten stages (Fig. 3.5B).

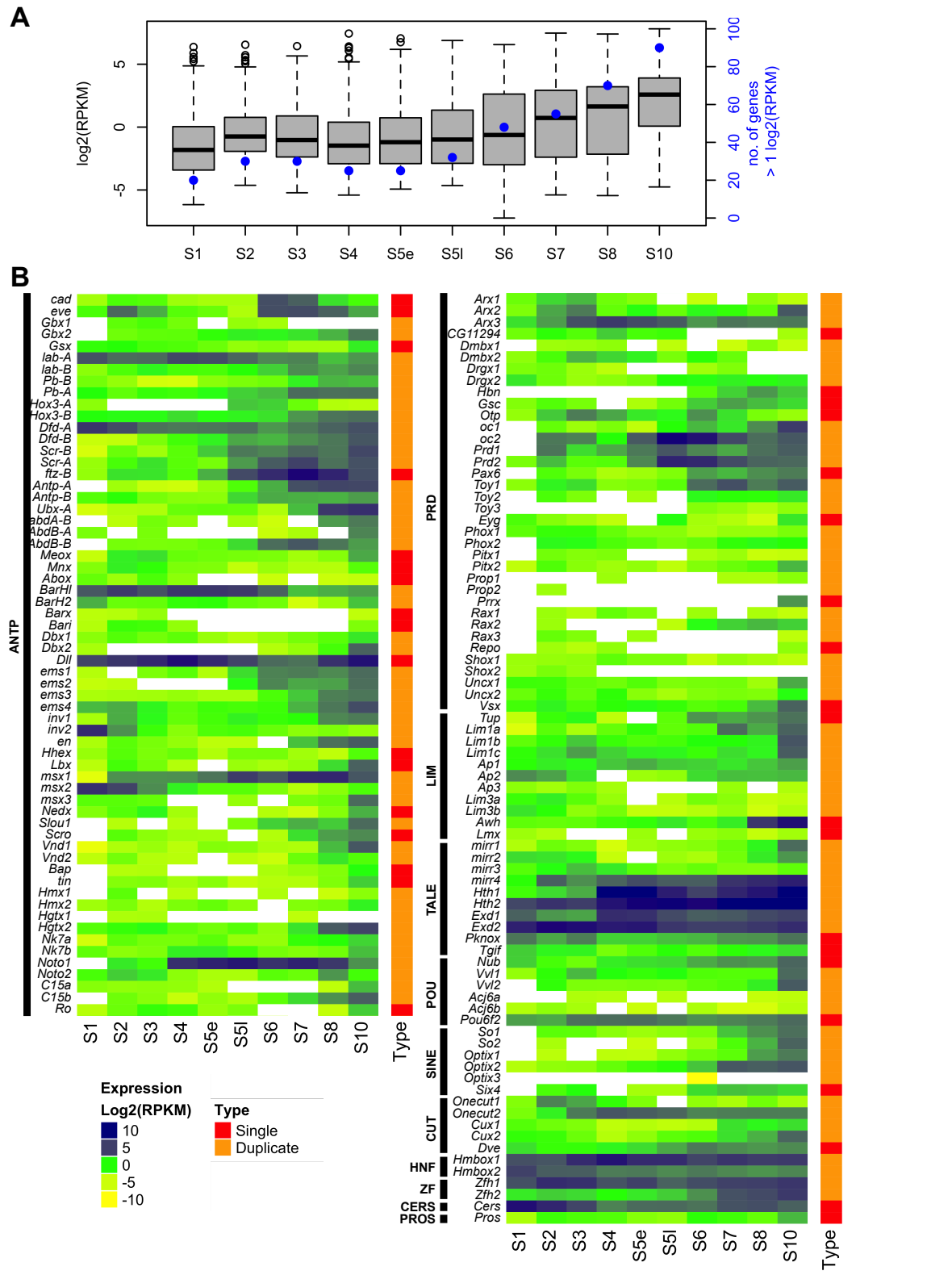


Figure 3.5: Expression of homeobox genes in *P. tepidariorum* expressed from S1 to S10. (A) The expression of all homeobox genes increases from S1 to S2, likely corresponding to onset of zygotic transcription. The numbers of families expressed above 1 log2(RPKM) also increase from S1 to S2. Both the mean expression level and number of families reduces around S4/S5e. After which the mean expression level and number of families continues to increase. **(B)** The transcriptome profile of *P. tepidariorum* AUGUSTUS gene models for Hox genes. Single copy genes per family are in red and duplicates are in orange.

There were also several cases of dispersed (non-clustered) paralogs, which have diverged in the level and timing of their expression (Fig. 3.5B). This included *Hth*, which has one paralog (*Hth2*) expressed throughout all ten stages, where as *Hth1* is only expressed from S4 to S10. Unfortunately, recent analysis of *Hth* in spiders concentrated on the expression of *Hth* in the developing limb appendages, and therefore lacks a comparison to the earlier expression dynamics observed in the *P. tepidariorum* transcriptome (Turetzek *et al.* 2017). Other dispersed paralogs that show aspects of divergence were the en/Inv family. The *en* gene is expressed at S7 in the transcriptome (Fig. 3.5B), which compares well with ISHs that show expression of *en* starts at early S8 in forming segments in *P. tepidariorum* (Schwager 2008). The *Inv1* paralog shows similar expression, however *Inv2* appears to be maternally loaded and down regulated at S2 when zygotic transcription starts (Fig. 3.5B). Therefore, the timing of expression between *en*/*Inv* paralogs suggests that they diverged in function. There were other dispersed duplicates that also showed patterns of divergence between paralogs such as *Gbx*, *Msx*, *Noto, Arx*, *Onecut*, *Hmbox* and *Zfh* (Fig. 3.5B). However, in contrast there were a few examples of dispersed paralogs that maintained expression similarities such as *Pitx*, *Phox*, and *Vvl* (Fig. 3.5B). However, it is possible that expression difference may occur later in development or adult stages and that expression levels do reveal any spatial differences in expression that may have occurred. This suggests that overall there has been evolution of the *cis*-regulation of most dispersed paralogs resulting in divergence in expression levels and transcriptional timing between paralogs.

3.4.2 Divergence of Tandem Paralogs in a *P. tepidariorum* Embryogenesis

There were also a several tandem duplications of homeobox genes in the *P. tepidariorum* genome (Fig. 3.4). Tandem duplicates, similar to dispersed duplicates, exhibited both conserved and divergent expression profiles. The Emx family contains four paralogs, of which pairs of paralogs are found on two different scaffolds (Fig. 3.4A). The pair of paralogs *Emx1* and *Emx2* has similar expression to one another, with expression increasing and maintaining from S6 to S10 (Fig. 3.5B). In contrast the other two paralogs Emx3 and Emx4 are both expressed later from S7/S8 to S10 (Fig. 3.5B). There is some early expression of Emx4, however, overall it appears that *Emx* paralogs on the same scaffold bare greater expression similarity than between paralogs on different scaffolds. Similarly, the Irx family contains four paralogs, with two of these as dispersed duplicates (*mirr3* and *mirr4*) and two tandem duplicates (*mirr1* and *mirr2*) (Fig. 3.4A). The tandem duplicates are both primarily expressed just at S10 (Fig. 3.5C). In contrast the dispersed paralogs *mirr3* is expressed just at S3 and the *mirr4* paralog is expressed from S2 to S10 at fairly consistent levels (Fig. 3.5B). The Lim1/5 family also contained two paralogs on one scaffold and a third paralog on a separate scaffold (Fig. 3.4D). The two Lim1/5 paralogs on the same scaffold had very similar expression, with low levels at S3 but stronger expression at S10 (Fig. 3.5B). In contrast the single Lim1/5 paralog on the other scaffold was not expressed at early stages but instead was expressed from S7 to S10 (Fig. 3.5B). The remaining tandem duplicates, *BarH*, *Prop* and *Shox,* all showed divergent

expression between paralogs (Fig. 3.4A and B, Fig. 3.5B). For example the *BarH1* paralog is strongly expressed from S1 to S6, where as the other paralog appears to be expressed only in S1 and then S10 (Fig. 3.5B). The *Prop* and *Shox* paralogs were also somewhat divergent in expression, however low levels of expression at these stages make it more difficult to compare (Fig. 3.5B). Overall these patterns of conservation and divergence of tandem expression could relate to the time of tandem duplications in respect to a possible WGD. In the situation of the Emx, Irx and Lim1/5 families, it is perhaps most parsimonious to suggest that pairs of duplicates existed prior to the WGD, rather than repeated tandem duplications independently occuring in each lineage. These paralogs may have utilised similar *cis*-regulatory controls, and therefore paralogs would have similar expression patterns. Subsequently, after the WGD the *cis*-regulatory regions of the two clusters diverged, but importantly continued to regulate tandem duplicates on the same scaffold in a similar fashion. In contrast to this the *BarH*, *Prop* and *Shox* tandem duplicates are only present on one scaffold, indicating that they were duplicated in tandem after the WGD event (or that paralogs from the WGD have been completely lost). These genes may have therefore duplicated more recently but diverged more rapidly than tandem duplicates that were also further duplicated in the WGD event.

3.5 *In Situ* Hybridisation (ISH) of *P. tepidariorum* Homeobox Genes Reveals Divergent Expression Patterns

To better characterise the divergence in expression of homeobox genes several candidates were studied using ISH to detect mRNA expression in embryos. All candidates selected were dispersed duplicates. The *Hmx* (ANTP class), *Pitx* (PRD class) and *Zfh* (ZF class) paralogs were selected because of their divergent RNA-Seq profiles, and the *Cux* paralogs (CUT class) for their similar profiles. These genes are also present in other arthropods and some contain annotated functional domains other than the homeodomain, which may provide further information about the evolution and function of these genes.

3.5.1 *Cux* Genes

3.5.1.1 Architecture of Cux proteins in arthropods

The *Cux/Cut* gene is a member of the CUT homeobox family and typically contains one homeodomain and three CUT domains. Additionally there is a C-terminal SNARE-like domain present in non-pancrustacean arthropods, except for *Pt-Cux1* and *Pp-Cux1* (Fig. 3.6A). There are also conserved regions, at positions ~500, ~1550, ~2300 and ~2600 of the alignment, which have no homology to currently annotated protein domains (Fig. 3.6A).

For each arachnopulmonate species two copies of *Cux* were identified, whereas only one copy was found in other arthropods surveyed. All genes contained one complete homeodomain, which had 63% identical residues (Fig. 3.6B). These conserved sites were predominantly concentrated in the first and third α-helices (Fig. 3.6B). Phylogenetic analysis of *Cux* homeodomains, as expected, group chelicerates together, though the spider and scorpion sequences group separately from one another (Fig. 3.6C). An interesting site so far as the duplication in arachnopulmonates is concerned is at position 56, which in *I. scapularis* and the three mandibulates is a leucine (Fig. 3.6B). However, the spiders and scorpions each contain one paralog with a methionine (Fig. 3.6B). The other paralog in both of the spiders is a leucine and the scorpions both have isoleucine. It is likely that leucine was the ancestral amino acid at this position in arthropods. However after the duplication in the arachnopulmonate ancestor, prior to the spider-scorpion split, one of the paralogs evolved to have methionine at position 56. Subsequently, after the spider-scorpion divergence the spiders retained leucine in one paralog, but in scorpions this evolved an isoleucine.

Most *Cux* genes examined here contain three complete CUT domains (Fig. 3.6A). However, both of the scorpion *Cux1* homologs have a partial CUT domain that is closest to the homeodomain but completely lack the other two CUT domains. The other paralog of *C. sculpturatus* also lacks the first CUT domain and only has a partial second domain. The lack of the first two CUT domains in *Pp-Cux1* is likely a result of a partial gene model prediction.

3.5.1.2 *Cux mRNA expression during P. tepidariorum embryogenesis*

In the RNA-Seq profiling both *Cux* paralogs are expressed at low levels early in development (Fig. 3.5B). There is an increase of expression of both paralogs in the final two stages. However, *Cux2* is expressed at higher levels compared to the *Cux1* paralog (Fig. 3.5B). The timing of the onset of expression in the transcriptome matches well with ISH results (Fig. 3.7).

Cux1 was first visible at S10.2 using ISH, and therefore may explain low levels of *Cux1* expression in the RNA-Seq data. The expression of *Cux1* was localised to the very distal tips of the pedipalps and all legs, as well as the very posterior medial region of the SAZ (Fig. 3.7A). In congruence with the transcriptome, *Cux2* expression starts earlier than *Cux1* with signal visible around S9.2 (Fig. 3.7B). There is faint expression in the developing mouth region (Fig. 5.9B) and later expression in the labrum above the stomodeum (Fig. 3.7C). Potential mesodermal expression of *Cux2* in the pedipalps and legs intensifies through development forming bands of stronger expression between limb segment furrows (Fig. 3.7B-D). In the opisthosoma, *Cux2* is first expressed in the tissue between segment furrows and excluding tissue surrounding the ventral sulcus (Fig. 3.7B). By S11 expression in the opisthosoma is fairly ubiquitous but varies in intensity (Fig. 3.7C). However, expression is absent at the book lung opening and as the anterior, posterior and medial spinnerets develop further by S12 (Fig. 3.7D) expression is also absent from the centre of these organs. Only in later stages is expression visible in a region anterior to the anterior furrow (Fig. 3.7C-D).

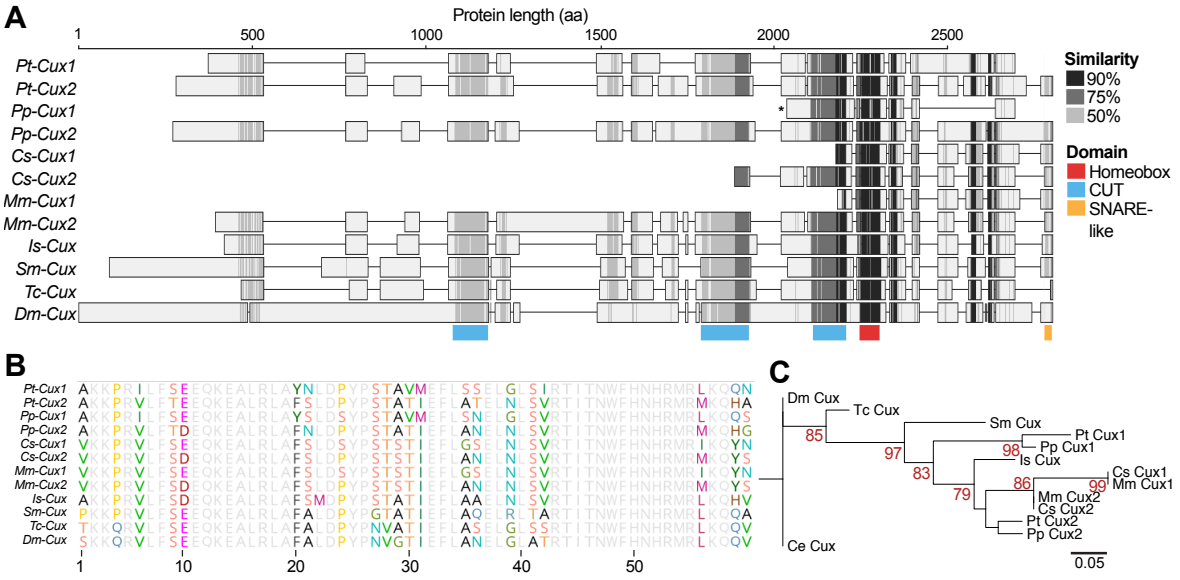


Figure 3.6: *Cux* genes in arthropods. (A) Gene structure of *Cux* genes reveals changes in domain content. **(B)** Homeodomain alignment (<90% conservation highlighted) and their **(C)** phylogenetic reconstruction. Percentage bootstrap scores >70% are shown in red. PhyML run with model JTT; 500 bootstraps; gamma distribution 0.37

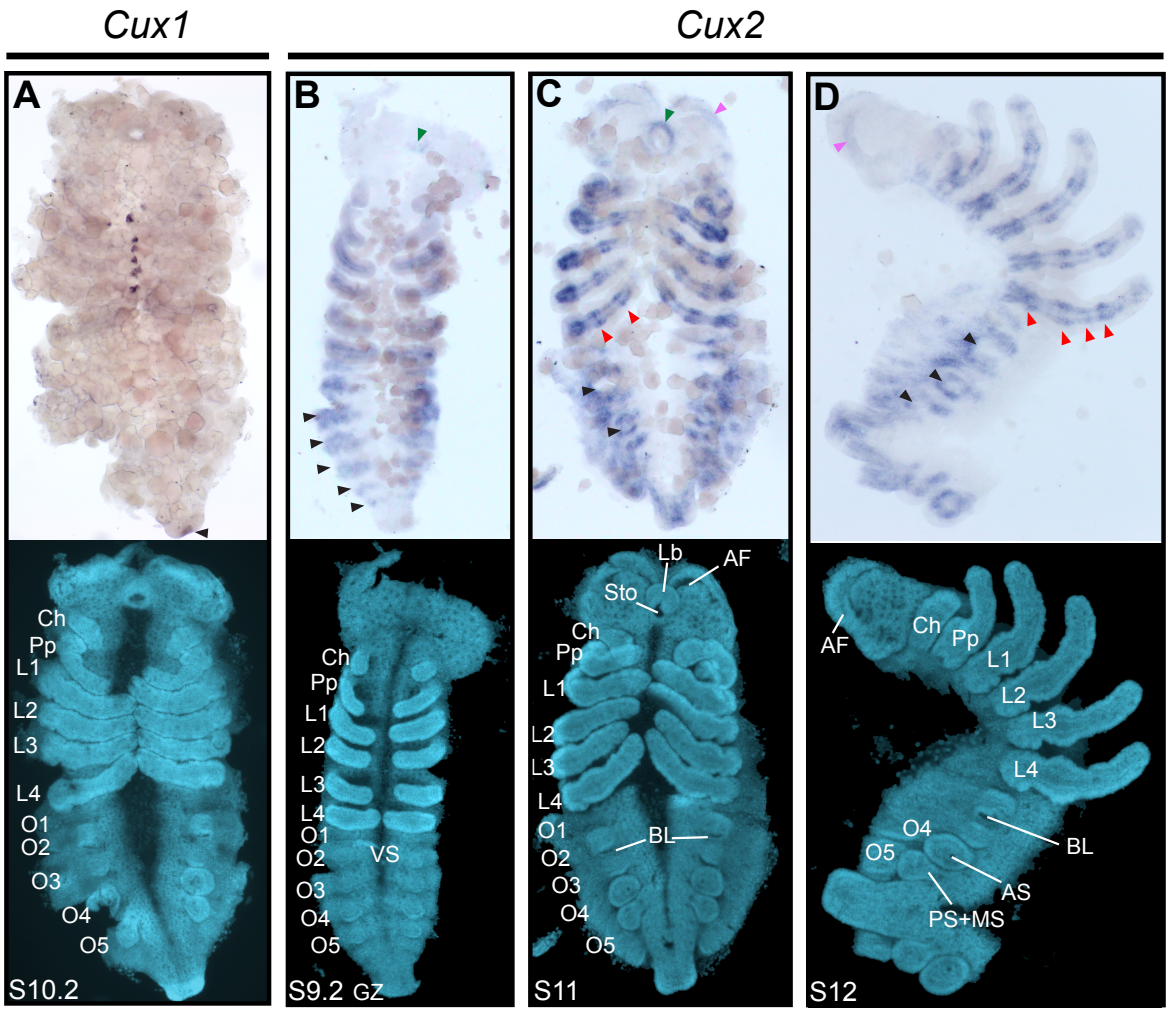


Figure 3.7: Expression of *Cux* paralogs in *P. tepidariorum*. **(A)** *Cux1* expression in the distal region of Pp-L4 and in the very posterior of the opisthosoma (black arrow). **(B – D)** *Cux2* expression in the opisthosoma, which is absent from the ventral sulcus (*VS*), becomes more refined and surrounds the opening of the book lung (*BL*) and anterior (*AS*), posterior (*PS*) and medial spinnerets (*MS*) (black arrows). **(C** and **D)** Possible mesodermal *Cux2* in the limbs form bands of stronger expression (red arrows). **(B)** Expression in the presumptive mouth (green arrow), above the forming stomodeum (*Sto*), becomes more concentrated in the labrum (*Lb*) at later stages **(C)** (green arrow). All embryos are flatmounted with brightfield above and DAPI stainings below. The embryos are orientated with the anterior at the top. In **(D)** the embryo has been dissected in half down the ventral midline so that it could be flatmounted.

3.5.2 *Hmx* Genes

3.5.2.1 Architecture of Hmx *proteins in arthropods*

The *Hmx/Nk5* gene belongs to the large ANTP homeobox family and in spiders and scorpions two copies are present but there is only a single copy in other arthropods surveyed except for *Ixodes* (Fig. 3.8A). *Hmx* contains one homeodomain and flanking it just downstream a conserved AAELEAAN motif (Wang *et al.* 2000), but no other known annotated domain. However, aside from this previously described motif, there are also other conserved regions at positions ~240, ~300, ~560 of the alignment and a region at the C-terminal end of chelicerate specific *Hmx* genes that has no homology to known protein domains (Fig. 3.8A).

Each of the *Hmx* genes surveyed contains a single complete homeodomain except for *Pp-Hmx1*, which may be due to an incomplete gene model (Fig. 3.8B). After excluding *Pp-Hmx1*, the remaining sequences have 82% identical residues, which is reflected in the poor bootstrap support and non-binary grouping of arthropod *Hmx* homeodomains (Fig. 3.8C). The ninth residue contains an amino acid substitution that is specific to chelicerates (T) and differs from mandibulates (S) and *C. elegans* (S) (Zhong *et al.* 2008; Zhong and Holland 2011), however aside from this site there are no other conserved chelicerate specific differences.

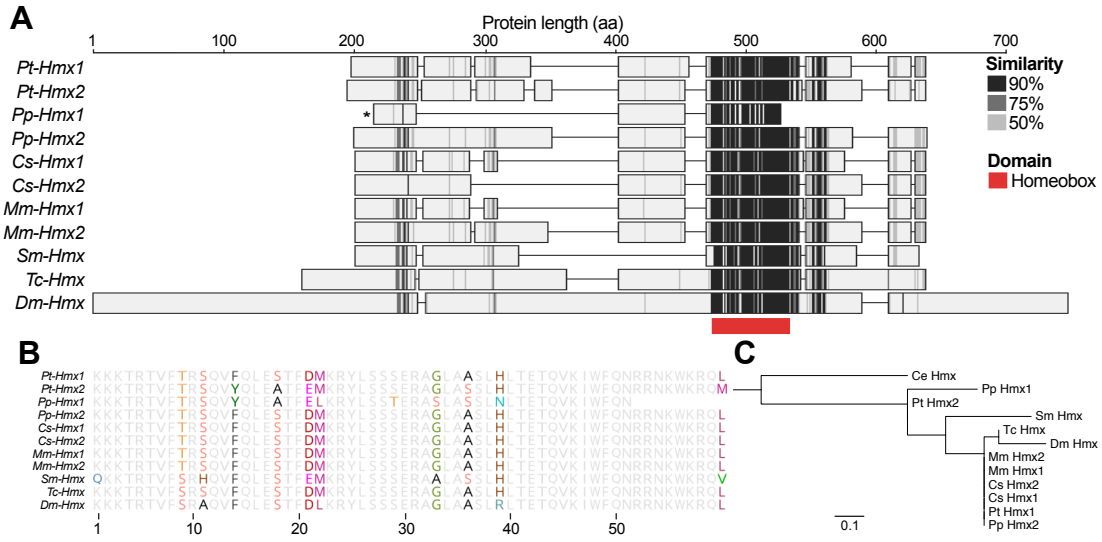


Figure 3.8: *Hmx* genes in arthropods. (A) Gene structure of *Hmx* genes reveals changes in domain content. **(B)** Homeodomain alignment (<90% conservation highlighted) and their **(C)** phylogenetic reconstruction. Percentage bootstrap scores >70% are shown in red. PhyML run with model LG; 500 bootstraps.

3.5.2.2 Hmx *mRNA expression during P. tepidariorum embryogenesis*

In the embryonic expression profiles of *Hmx* genes in *P. tepidariorum* (Fig. 3.5B) both paralogs are expressed by S2, though *Hmx2* is generally more abundantly expressed compared to *Hmx1* throughout the ten stages. However ISH only detected the earliest expression of *Hmx1* from stage eight and of *Hmx2* from stage nine.

Hmx1 expression first appears in the developing prosomal appendage buds (Ch-L4) (Fig. 3.9A). As the limb grows *Hmx1* expression is maintained in the distal region of the appendages (Fig. 3.9B). By stage ten the broad expression domain retracts to form rings of expression approximately as far away from the distal tip of the limb appendage as the most proximal boarder of earlier *Hmx1* expression (Fig. 3.9C).

The ISH of *Hmx2* showed completely non-overlapping expression compared to *Hmx1*. The expression of *Hmx2* first appears at S9 in the pre-chelicerical lobes (Fig. 3.9D). As development continues the expression of Hmx2 is maintain in the region between the lateral and anterior furrow (Fig. 3.9E). By stage thirteen, the medial and lateral sub-division have migrated towards the furrow so that the expression of *Hmx2* appears to derive from beneath the surface of the pre-chelicerical region, where the forebrain develops (Fig. 3.9F). This expression domain of *Hmx2* suggests that it marks early presumptive neural tissue that undergoes invagination through the pre-cheliceral furrows to form some of the brain tissue.

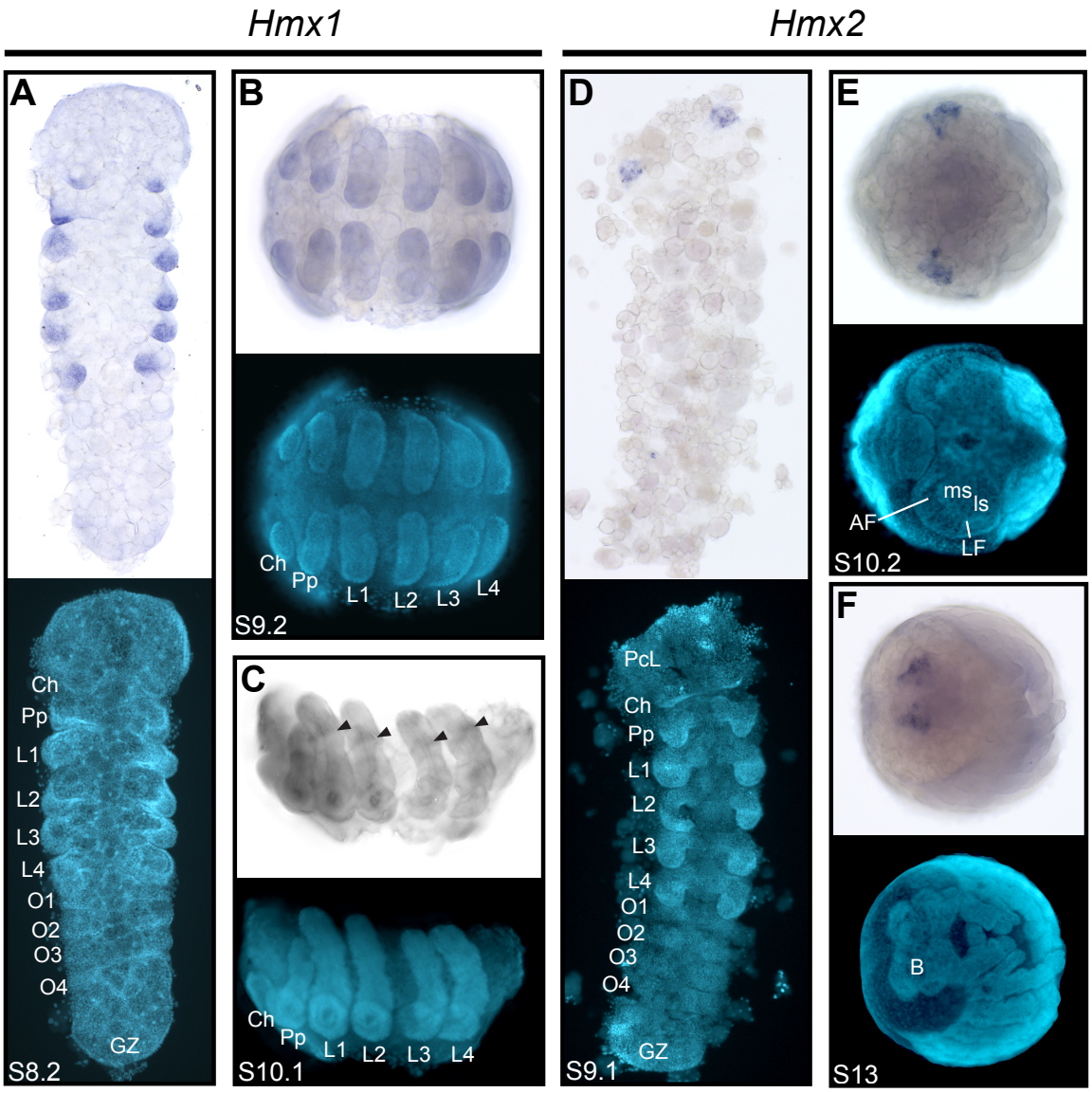


Figure 3.9: Expression of *Hmx* paralogs in *P. tepidariorum*. **(A – C)** Limb specific expression patterns of *Hmx1* form bands in later stage embryos **(C)** (black arrows). **(D)** Expression of *Hmx2* in the pre-cheliceral (*PcL*) region locates to area around the anterior (*AF*) and lateral (*LF*) furrows, where the medial (*ms*) and lateral (*ls*) subdivision migrates **(E)**, as the brain (*B*) develops **(F)**. All brightfield images are above and respective DAPI images below. **(A** and **D)** Flatmounts of embryos are orientated with the anterior at the top. **(B**, **E**, **F)** The wholemounts are orientated with anterior to the left. **(C)** The prosomal appendages were dissected from one side of an embryo and orientated so the anterior is to the left.

3.5.3 *Pitx* Genes

3.5.3.1 Architecture of Pitx *proteins in arthropods*

The *Pitx* gene belongs to the PRD homeobox family and in spiders and scorpions two copies are present but there is only a single copy in other arthropods surveyed. *Pitx* genes commonly contain a single homeodomain and a C-terminal OAR domain (Fig. 3.10A). There are also two regions that have some conservation, one at the N-terminus relating to an eh1 domain (Tolkunova *et al.* 1998; Burglin and Affolter 2016), and another that has no homology to known domains at position ~470 of the alignment.

A single homeodomain is present in all genes, although both *C. sculpturatus* paralogs and the single copy found in *I. scapularis* are missing the C-terminal region of the homeodomain (Fig. 3.10B). Excluding these partial or perhaps truncated sequences, the remaining sequences had 82% identical residues. Despite the grouping together of *C. sculpturatus* paralogs (Fig. 3.10C) it looks from the alignment at positions 11 (R), 34 (L) and 36 (V) (Fig. 3.10C) that *Cs-Pitx2* and *Mm-Pitx2* are homologs. The OAR domain is present in all *Pitx* genes surveyed except for both *C. sculpturatus* paralogs, *Is-Pitx* and *Mm-Pitx2*. The remaining species have 10/15 identical sites in the OAR domain. There is also differential retention of the eh1 domain at the start of *Pitx*. A single parlogs in each spider (*Pt-Pitx1* and *Pp-Pitx2*) and in each scorpion (*Cs-Pitx2* and *Mm-Pitx2*), as well as *Ixodes*, have lost this N-terminal domain.

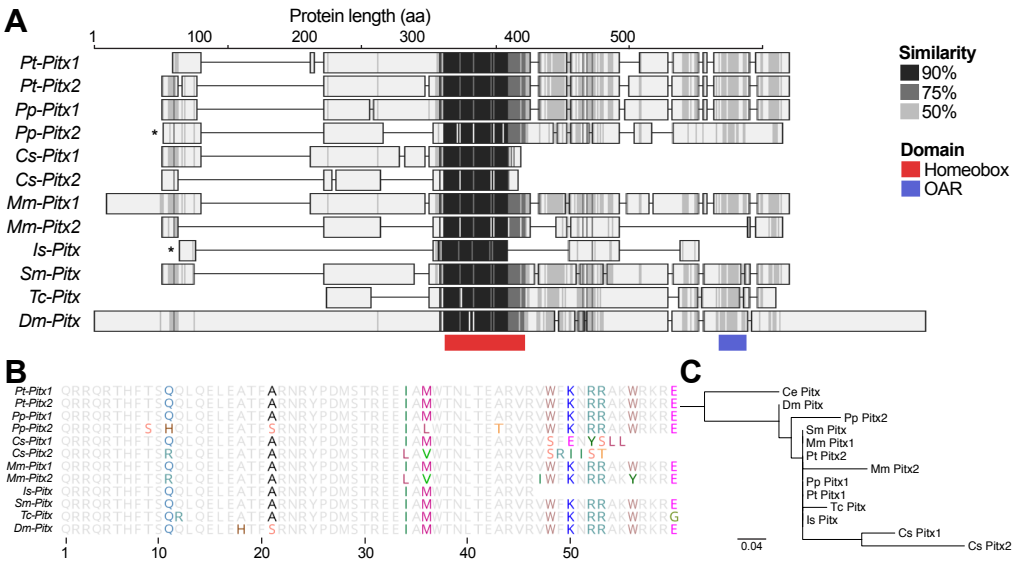


Figure 3.10: *Pitx* genes in arthropods. (A) Gene structure of *Pitx* genes reveals changes in domain content. **(B)** Homeodomain alignment (<90% conservation highlighted) and their **(C)** phylogenetic reconstruction. Percentage bootstrap scores >70% are shown in red. PhyML run with model CpREV; 500 bootstraps.

3.5.3.2 Pitx *mRNA expression during P. tepidariorum embryogenesis*

In the embryonic transcriptome *Pitx1* is expressed at low levels (< -1.9 log2(RPKM)) throughout the ten developmental stages (Fig. 3.5B). *Pitx2* also is lowly expressed (< -0.9 log2(RPKM)) through S1 to S8, but at S10 increases to 1.7 log2(RPKM) (Fig. 3.5B). These expression profiles are also supported by ISH results (Fig. 5.11), which did not detect *Pitx1* until S12 (Fig. 5.11A) and detected *Pitx2* at S10.1 (Fig. 5.11B), which corresponds to the increase in expression in the transcriptome at S10.

Expression of the *Pitx1* paralog was present at S12, with faint expression showing in two regions. There were faint dots of expression to the lateral region to the stomodeum (arrow heads) (Fig. 5.11A). There was also a dot per segment in the ventral tissue close to the margin where the ventral midline opens up as the embryo undergoes inversion (Fig. 5.11A). These dots were in each segment from the first walking limb (L1) until at least opisthosomal segment eight (O8).

The other paralog, *Pitx2*, appeared earlier than *Pitx1*, at around S10.1, with expression in small patches in each segment from L1 to O4 (Fig. 5.11B). The region of expression is somewhat similar to *Pitx1*, however the patches of *Pitx2* are larger than the dots of *Pitx1*. Furthermore, compared to *Pitx1*, the *Pitx2* expression looks to be slightly further from the ventral midline. As development progresses the patches of expression appear up to O6, and the colourmetric signal in S11 embryos intensified compared to S10.1 embryos (prepared and stained in the same ISH experiment) (Fig. 5.11C). By S12 these patches of expression are present to the very posterior opisthosomal segment and also in the coxa (Fig. 5.11D). There is also *Pitx2* expression in the same region as *Pitx1* in small dots close to the ventral midline that by S13 reach O5 (Fig. 5.11E).

Pitx2 also has complex expression in the pre-cheliceral region that starts at S11. Initially the expression is observed in the very anterior of the pre-cheliceral lobe and flanking the sides of the labrum (Fig. 5.11C). As the pre-cheliceral lobes begin to contract together (S12) anterior *Pitx2* expression marks the presumptive edges that meet by S13. Lastly, the initially dispersed *Pitx2* expression around the labrum resolves to form concentrated expression within and just anterior of the labrum (Fig. 5.11C – E).

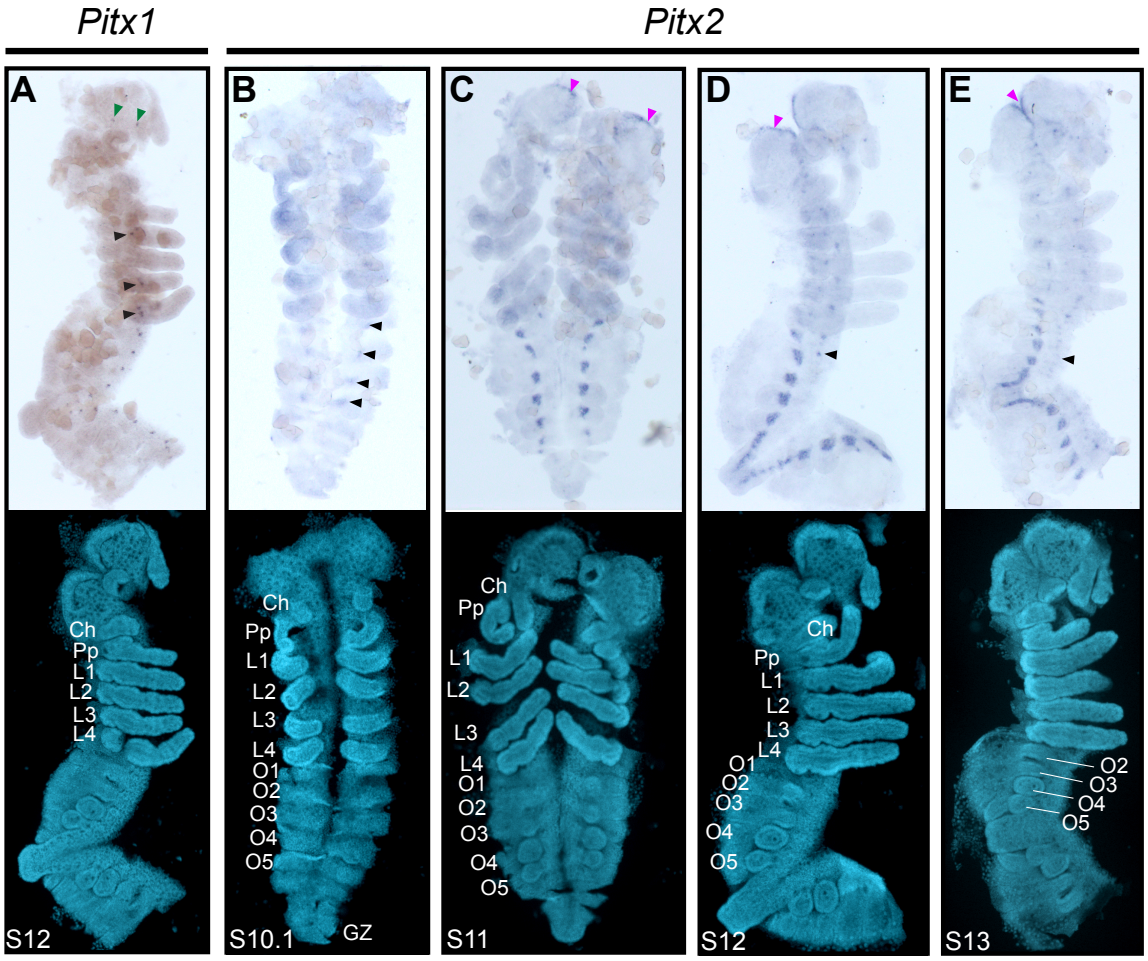


Figure 3.11: Expression of *Pitx* paralogs in *P. tepidariorum*. (A) Small dots of *Pitx1* expression along the ventral midline (black arrows) and by the stomodeum (green arrows). **(B)** Patches of *Pitx2* expression start at S10.1 (black arrows) and **(C)** another row more ventrally by S12 (black arrow). **(C – E)** *Pitx2* expression in the anterior of the pre-cheliceral region (pink arrows) mark regions that join together by S13**.** All brightfield images are above and respective DAPI images below. Embryos are orientated with the anterior at the top. Later stage embryos required the dissection of one side to ease flatmounting.

3.5.4 *Zfh* Genes

3.5.4.1 Architecture of Zfh *proteins in arthropods*

The ZF family of homeobox genes contain a *Zfh* member that has three to four homeodomains as well as multiple zinc-finger domains (Zhong *et al.* 2008; Zhong and Holland 2011). In spiders and scorpions two copies of *Zfh* were found, where as only a single copy of *Zfh* with multiple homeodomains is present in other species surveyed (Fig. 3.12). Aside from the annotated homeodomains and zinc-finger domains there were conserved region that had no homology to known domains at positions ~900, ~1200, ~2100, ~2200, ~2300, ~3100, the regions flanking immediately downstream of each homeodomain a C-terminal end (Fig. 3.12A).

Each *Zfh* gene surveyed had four homeodomains, except for *D. melanogaster* and *P. phalangioides,* which both had three, though the latter is perhaps due to an incomplete protein prediction (Fig. 5.14A). The third *D. melanogaster* homeodomain appears to have been lost and the most C-terminal homeodomain has diverged such that it groups with the third homeodomain of ecdysozoans (Fig. 3.12A and C). The arrangement of *Sm-Zfh2* homeodomains has undergone loss of the conserved 3' flanking regions of homeodomains two to four (Fig. 3.12A). Other differences in gene structure include the presence/absence of zinc-finger domains (Fig. 3.12A). The most N-terminal zinc-fingers (positions ~700-800 of alignment) are not present in *P. phalangioides* (incomplete protein prediction), both *C. sculpturatus* paralogs, *Mm-Zfh2* (incomplete protein prediction) and *Ixodes*, followed by differential retention of an unknown conserved regions at positions ~900 and ~1200 of alignment. Further towards the C-terminus, the zinc-finger domains between positions ~1400-2000, which are well conserved among all species surveyed, have low homology or are absent in *D. melanogaster*.

The homeodomains in *Zfh* have different degrees of divergence between arthropods, with homeodomains one to four sharing respectively 72%, 87%, 88% and 75% (38% including *D. melanogaster*) identical residues. Each of the homeodomains is distinct from each other so that there is strong bootstrap support for grouping each domain separately, except for Dm-Zfh2-D4 (Fig. 3.12C). Within each of these the spiders paralogs are usually grouped together separately from the scorpions, which also group together (Fig. 5.12C).

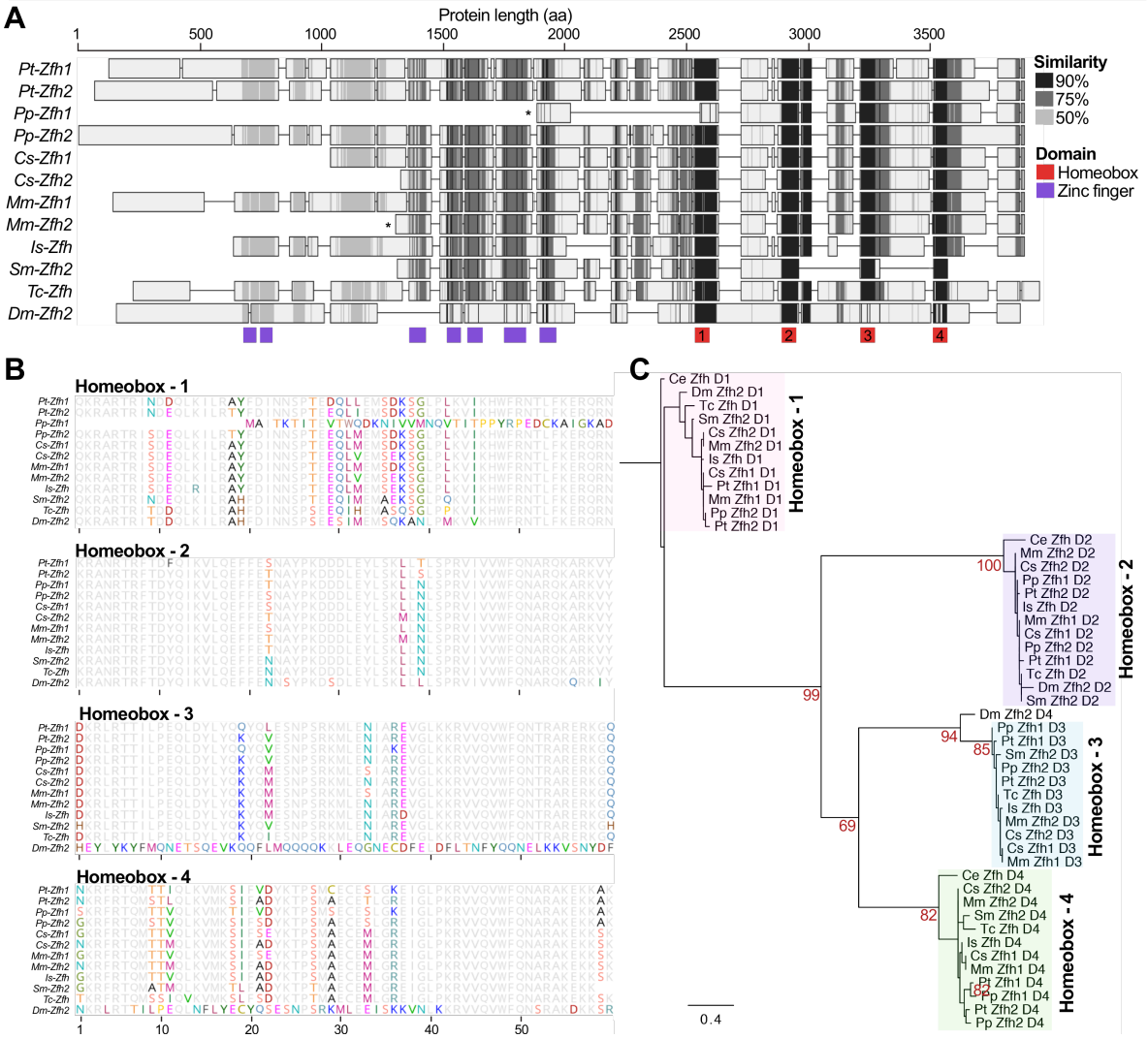


Figure 3.12: *Zfh* genes in arthropods. (**A**) Gene structure of *Zfh* genes reveals changes in domain content. (**B**) Homeodomain alignment (<90% conservation highlighted) and their (**C**) phylogenetic reconstruction. Percentage bootstrap scores >69% are shown in red. PhyML run with model LG; 500 bootstraps; gamma distribution 0.68; invariable sites 0.00.

3.5.4.2 **Zfh** *mRNA expression during P. tepidariorum embryogenesis*

Transcriptomic data provides support that both *Zfh* paralogs are expressed from S1 through to at least S10 in *P. tepidariorum* embryogenesis (Fig. 3.5B). *Zfh1* is supposedly expressed highly and continually, where as *Zfh2* is lowly expressed at first and then expression increases most notably around S7. Despite ISH detection only recovering signal for *Zfh1* at S7, and *Zfh2* at S8.1, this approach did support the RNA-Seq data in that *Zfh1* exhibited continued higher expression than *Zfh2*, and *Zfh2* had gradually increasing expression (Fig. 3.13).

Expression of *Zfh1* at S7 appears to be almost ubiquitous apart from two regions that appear to separate a region anterior of L1, another containing L1-L4 and finally the caudal lobe (Fig. 3.13A). These broad domains of expression resolve in each segment at S8.1 (Fig. 3.13B). In the appendage bearing segments Ch to L4, the expression appears to be absent from the forming buds (Fig. 3.13B and C). By S8.2 the expression in each segment intensifies in the ventral midline and then lateral regions of the pre-cheliceral lobes, with a reduction in expression at the presumptive mouth region (Fig. 3.13C). Furthermore, within the region of previously absent expression in appendage buds, expression reappears within the limb buds (Fig. 3.13D). Generally, the expression domain continues to become more refined by S10.2 to the neural tissue of the ventral midline and pre-cheliceral lobes, as well as being maintained in the proximal segment of Pp to L4 (Fig. 3.13E).

Faint expression of *Zfh2* was first detected at early S8.1 in what appear to be the initial segmental boundaries of the prosoma (Fig. 3.13F). This becomes clearer by late S8.1 and the expression becomes more restricted to the ventral midline, as well as a broad domain covering the pre-cheliceral region (Fig. 3.13G). At later stages (S9.2 to S13) the ventral expression becomes complex and surrounds the coxa, the developing opisthosomal organs and also running along the edges of the ventral sulcus opening (Fig. 3.13H – J). Also the prosomal appendages have strong distal expression and as they elongate, rings of expression form at segmental boundaries. The broad domain of expression in the pre-cheliceral lobes (Fig. 3.13G) later becomes restricted to the labrum and labium by S13 (Fig. 3.13I).

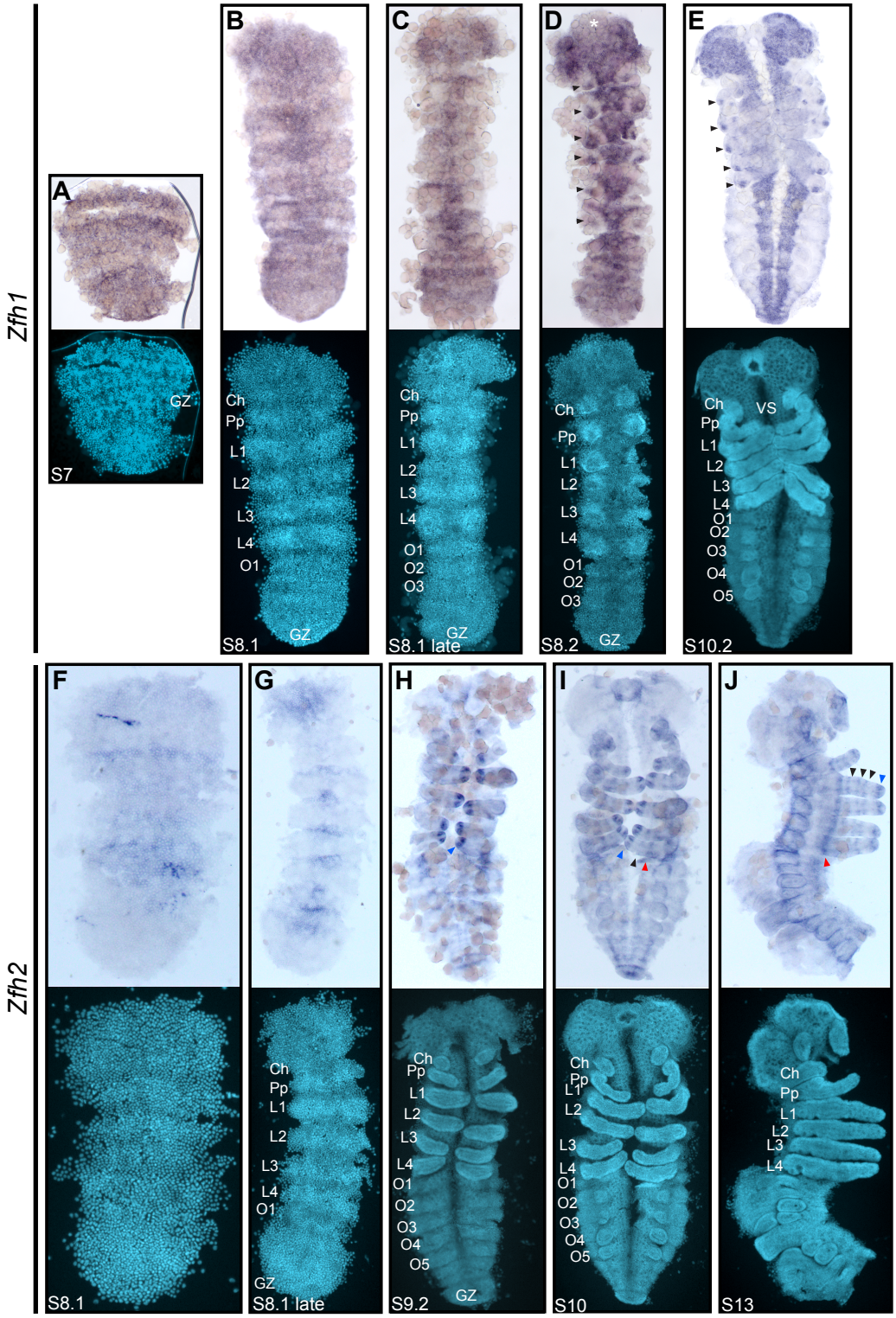


Figure 3.13: Expression of *Zfh* paralogs in *P. tepidariorum*. **(A – E)** Initially broad expression of *Zfh1* becomes restricted to the ventral midline and pre-cheliceral regions, where there are the point-like depressions of the developing neuronal tissue, and is absent from where the ventral sulcus (*VS*) opens. Expression of *Zfh1* is at first absent from the limb buds **(B)** but reappears **(D)** (black arrows) and is maintained in the proximal region of the limbs **(E)** (black arrows). **(F – J)** *Zfh2* expression becomes segmented and is at first restricted to the ventral and pre-checlieral regions. However, later expression also forms in the tips of the limbs **(H)** (blue arrow), and as the limb grows appears to generate bands of expression **(I)** (black arrows) with the first onset of expression marked by the red arrows **(I – J)**. All brightfield images are above and respective DAPI images below. Embryos are orientated with the anterior at the top. Later stage embryos required the dissection of one side to ease flatmounting.

3.6 Discussion

3.6.1 Homeobox Repertoires in Chelicerates

Homeobox genes encode an important group of transcription factors that are well known to be involved in a wide range of developmental programs (Zagozewski *et al.* 2014; Bataille *et al.* 2015; Du and Taylor 2015; Zuniga 2015; Krumlauf 2016). Consequently they have received substantial attention and are often characterised and compared within and between animal genomes. In the arthropod phylum the insects have been sampled the most extensively and robustly. However, other arthropod subphyla such as the basally branching clade of chelicerates have received limited systematic surveys of whole homeobox complements, with only horseshoe crabs and the scorpion *M. martensii* being characterised (Di *et al.* 2015; Kenny *et al.* 2015a). Therefore, in order to expand upon and compare homeobox repertoires in chelicerates, two spiders *P. tepidariorum* and *P. phalangioides*, another scorpion *Centruroides sculpturatus* and lastly the tick *Ixodes scapularis* were surveyed.

Overall there was generally a similar complement of homeobox classes and families present, suggesting that chelicerates share and have retained similar homeobox repertoires as other arthropod subphyla (Fig. 3.1). There were however several families that were specific to scorpions (Nk8 and Six7), and the Nedx family in spiders that was not present in other chelicerates, including horseshoe crabs (Kenny *et al.* 2015a). These particular families may therefore regulate lineage specific features during scorpion and spider development. Furthermore, the Barx family, which was present in chelicerates but not other arthropods, may coordinate specific chelicerate developmental characteristics. Aside from these gross changes in the presence/absence of homeobox families there were also families that contained divergent chelicerate homeodomains such a Cdx, En, Hox3, Hhex, Msxlx and Cers. This implies that there may be functional changes in the homeodomain of these genes. Other functional domains, such as CUT, and zinc finger domains (Fig. 3.6 and 3.12), also displayed similarities and divergences between species. Together, these show that not only homeodomains contribute to changes in protein sequence and therefore predicting functional divergence should not be limited to assessing just the evolution of homeodomains.

The species with perhaps the greatest changes in the complement of homeobox families was the tick *I. scapularis*. In this lineage the fewest families were identified, indicating either loss or incomplete sequence evidence for all families. However, the lineage of parasitiforms, and their related acariforms, has previously been shown to have a greater loss of conserved miRNA families compared to other arachnid lineages (Leite *et al.* 2016). Therefore it is likely that there is actual loss of homeobox genes in *I. scapularis*. Another aspect of mites and ticks that has revealed divergences to other chelicerates is that of long branch lengths, which often disrupts their placement during phylogenetic reconstruction (Sharma *et al.* 2014b). Whether or not there is a connection between gene loss and accelerated evolution of remaining genes is unclear. To uncover these evolutionary

dynamics in chelicerates, further surveying of homeobox repertoires in orders such as Opiliones is needed. Based on their phylogenetic position, Opiliones species are predicted not to have had a WGD, and therefore may provide more comparable genome architecture to mites and ticks (Sharma *et al.* 2014b; Leite and McGregor 2016).

3.6.2 Possible Ancient Whole Genome Duplication in the Ancestor of Arachnopulmonates

Previous work had already identified duplicated homeobox genes in chelicerates (Di *et al.* 2015; Kenny *et al.* 2015a), such as Hox genes in spiders and scorpions (Schwager *et al.* 2007; Sharma *et al.* 2015), as well as other duplicated homeobox genes involved in spider eye development (Samadi *et al.* 2015; Schomburg *et al.* 2015). However, though a scorpion and three horseshoe crabs had been investigated systematically and many duplicated homeobox families were present, there had been no systematic identification of homeobox genes in any spider or tick in order to compare these duplications. The identification of homeobox genes in two spiders, another scorpion and a tick allowed for a comparison of the extent of homeobox duplication within arachnid chelicerates. There were many more duplicated homeobox families in arachnopulmonate species (49-59%) compared to other arthropods surveyed including *I. scapularis* (15-20%) (Fig. 3.1). This proportion of duplicated homeobox families is greater than the levels of duplication identified in either *P. tepidariorum* or *C. sculpturatus* using either the BUSCO (41%) or OMA (20.5%) datasets (Schwager *et al.* 2017). This is perhaps due to the function of homeobox genes as transcription factors, which appear to be enriched in genes retained after WGD events (Maere *et al.* 2005; McGrath *et al.* 2014). Interestingly, 18 homeobox families were represented by two paralogs in all four arachnopulmonates but were only single copy in all other arthropods surveyed. This makes up a considerable function of the 63-78 duplicates identified in *P. tepidariorum* and *C. scuplturatus* compared to mandibulates and ticks using the BUSCO-Ar database (Schwager *et al.* 2017). These common homeobox duplicates may perhaps function in crucial features of arachnopulmonate specific development and have therefore been conserved.

Gene duplications in arachnopulmonates are also not limited to homeobox genes, as many other duplicated developmental genes including Wnts and *frizzled4*, and *dachshund*, as well as venom and silk genes have been identified in the arachnopulmonate linage (Schwager *et al.* 2007; Janssen *et al.* 2010; Haney *et al.* 2014; Clarke *et al.* 2015; Janssen *et al.* 2015; Pechmann *et al.* 2015; Samadi *et al.* 2015; Schomburg *et al.* 2015; Turetzek *et al.* 2016a; Schwager *et al.* 2017; Turetzek *et al.* 2017). Furthermore, non-coding regions of the genome containing miRNAs are also pervasively duplicated in arachnopulmonate genomes (see Chapter Four) (Leite *et al.* 2016). Based on previous understanding of chelicerate phylogenies and recent estimates of the timing of duplication events in chelicerates it is suggested that the WGD in arachnopulmonates is likely to be independent of the two WGDs in the horseshoe crab lineage (Nossa *et al.* 2014; Kenny *et al.* 2015a; Schwager *et al.* 2015; Schwager *et al.* 2017).

Evidence for two WGD in the horseshoe crab lineage derived not only from the copy number of many homeobox genes but importantly from the preservation of four clusters of Hox genes. In vertebrate lineages including teleost and salmonid fish there are multiple Hox clusters present (Hoegg and Meyer 2005; Mungpakdee *et al.* 2008; Pascual-Anaya *et al.* 2012). These have helped to infer that the ancestor of vertebrates underwent two WGDs, followed by a further WGD in teleost fish and another within salmonid fish (Hoegg and Meyer 2005; Mungpakdee *et al.* 2008; Pascual-Anaya *et al.* 2012). Following this logic the presence of intact duplicate Hox clusters in arachnopulmonates implies that the ancestor to this lineage has also undergone a single WGD. In contrast single Hox clusters with single copies of most Hox genes are found in the harvestman *Phalangium opilio*, *I. scapularis* and *T. urticae* (Sharma *et al.* 2012; Pace *et al.* 2016). Further support for a WGD in the arachnopulmonate lineage also exists in the form of duplicated ANTP (NK cluster), SINE, TALE, and LIM class genes (Fig. 3.4). The NK clusters in *P. tepidariorum* contain paralogs of Nk7, Tlx and Emx, though the remaining genes have been reduced to single copies (Fig. 3.4A). However, considering the size of the breaks between the genes in each of the NK clusters, the inferred ancestral order of arachnopulmonate NK clusters (*Nk7*, *Lbx*, *Tlx*, *bap*, *tin*, *Msx*) matches well to the order in the predicted protostome-deuterostome ancestor (Garcia-Fernandez 2005; Ferrier 2016), requiring just an inversion containing *Lbx* and *Tlx* (Fig. 3.4A). Some of the remaining ANTP class genes in *P. tepidariorum* were also clustered, suggestive of remnants of the mega-cluster however these were not duplicated (Fig. 3.4A). Lastly, the presence of the HRO cluster containing *Rax2*, *Hbn and Otp* was also present, provides further evidence along side *S. maritima* that this cluster is found widespread in arthropods and other protostomes (Fig. 3.4B) (Mazza *et al.* 2010; Chipman *et al.* 2014; Ferrier 2016). Despite other *Rax* paralogs in *P. tepidariorum* there are no duplicate HRO clusters (Fig. 3.4B). However, the order of the three genes are different compared other arthropods, suggesting an inversion in the lineage leading to *P. tepidariorum* (Mazza *et al.* 2010). Together these results reveal a possible WGD and subsequently different patterns of gene retention and loss in particular families of homeobox genes in arachnopulmonates. Furthermore, most of the duplicated clusters of homeobox genes that correspond to those hypothesised to exist in the bilaterian ancestor have been retained in the *P. tepidariorum* genome (Ferrier 2016).

3.6.3 Divergence in the Expression of Homeobox Paralogs

An outstanding question about horseshoe crabs and arachnopulmonates is how WGD events have contributed to their evolution and potentially the development of lineage specific features. It has already been shown that the paralogs of *dachshund* in spiders have distinct roles from each other and only one paralog is responsible for the formation of the arachnid specific patella (Turetzek *et al.* 2016a). Furthermore, the arrangement of structures in the opisthosoma of scorpions coincide with the staggered expression of paralogous Hox gene expression (Sharma *et al.* 2014c), suggesting that divergences in Hox paralogs may in part be responsible for innovation of the scorpion body.

Similarly to this the Hox paralogs of spiders also have divergences in their spatial expression (Schwager *et al.* 2007; Schwager *et al.* 2017), as well as expression timing, which can be seen also in developmental transcriptomes (Fig. 3.5B). Functional analysis of the paralogs of the Hox gene *lab* in *P. tepidariorum* identified that only one of the paralogs displays a notable knockdown phenotype of loss of the tritocerebral appendage (Pechmann *et al.* 2015). Other homeobox paralogs also show differential expression within developing eye structures (Samadi *et al.* 2015; Schomburg *et al.* 2015).

To further investigate if other homeobox paralogs follow these patterns of divergence, the coding regions and expression patterns of paralogs were analysed in *P. tepidariorum*. In the homeodomains of each of the paralogs of *Cux*, *Hmx*, *Pitx* and *Zfh* were amino acid changes that may reflect possible functional differences. In addition to the homeodomains there were changes in the other functional regions such as CUT and zinc-finger domains. Future experimentation via ChIP-Seq could help to identify possible divergences in the function of these binding domains.

It was also possible to see that all sets of paralogs have diverged in their expression patterns using ISH. Some, like *Hmx* and *Cux*, showed dramatically different spatial expression patterns, whereas *Pitx* and to some degree *Zfh* paralogs maintained similarities. Therefore, currently there have not been any homeobox paralogs identified in *P. tepidariorum* that display exactly the same expression patterns. It is likely that *cis*-regulatory regions of these genes have evolved and are a primary cause of expression pattern differences.

3.6.3.1 Cux *and* eya *have non-overlapping expression in the heads of spiders and flies*

In comparisons of *P. tepidariorum* mRNA expression patterns to other species it was possible to some extent to also infer which paralog displayed most retained or ancestral characteristics and which had diverged. In *Drosophila*, the single copy of *Cux*/*Cut* is expressed in a non-overlapping pattern in the eye-antennal imaginal disc at mid-late larval stage 2, with *Cut* in the antennal field and *ey* and *eya* in the eye field (Kenyon *et al.* 2003). This bares similarity to the expression of *Cux* and *eya* in *P. tepidariorum*, where *eya* is expressed in possibly non-overlapping regions more anterior to the *Cux* expression seen in the pre-cheliceral furrows and above the stomodeum (Fig. 3.14A) (Schomburg *et al.* 2015).

3.6.3.2 Pt-Hmx1, *not* Pt-Hmx2, *has conserved expression in the brain*

In both *Drosophila* (*Hmx*) and mice (*Hmx2* and *Hmx3*), *Hmx* homologs regulate the development of neural tissue and early expression of *Drosophila Hmx* looks remarkably similar to *Pt-Hmx1* (Wang *et al.* 2000). Therefore it appears that *Pt-Hmx2* has undergone substantial divergence to be expressed in the limbs, in a pattern not reminiscent of neural tissue, and therefore may have potentially undergone neo-functionalisation in *P. tepidariorum* (Fig. 3.14B). Identifying the expression of the ortholog in a chelicerate that has not likely experienced a WGD, such as the harvestman *P. opilio*, would help to clarify if *Pt-Hmx1* show ancestral expression patterns and *Pt-Hmx2* has divergent expression, or whether these paralogs have sub-functionalised. Furthermore,

characterisation of the expression of both paralogs in other spider species would help to identify if the changes between these paralogs are common to spiders or have diverged in the Araneae lineage.

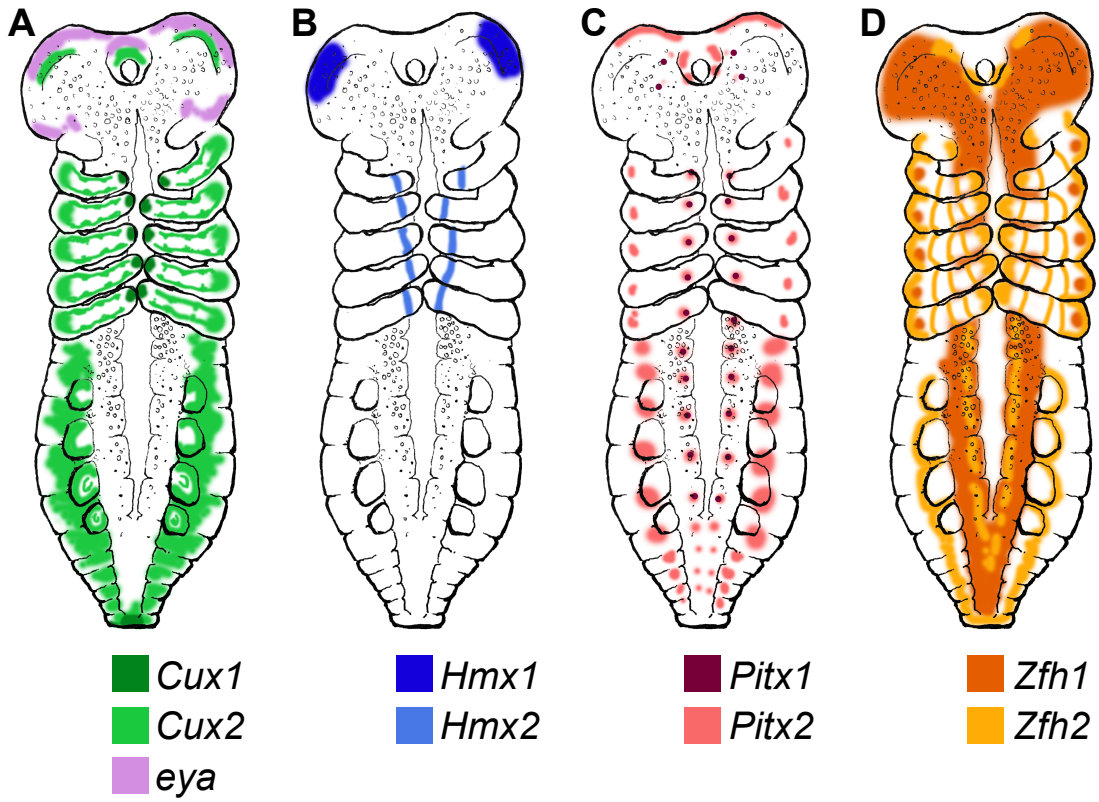


Figure 3.14: Divergence in the expression of *P. tepidariorum* homeobox gene paralogs. A schematic showing the expression patterns of paralogous genes. **(A)** *Cux* paralogs have very divergent expression patterns. The expression of *eya* is also shown, as it has non-overlapping expression with *Cux* also in *Drosophila*. **(B)** *Hmx1* expression in the presumptive brain tissue is perhaps more conserved than the bands of expression of limbs. **(C)** *Pitx* paralogs both have metameric patterns of expression and some overlap between paralogs. **(D)** *Zfh* paralogs exhibit temporal sub-functionalisation, whereby *Pt-Zfh1* matches embryonic expression in developing *Drosophila* larva, and *Pt-Zfh2* resembles the expression of concentric rings in the limbs of developing *Drosophila* pupa.

3.6.3.3 Possible sub-functionalisation of Pitx *paralogs*

Perhaps the most similar expression patterns in *P. tepidariorum* were between *Pitx* paralogs. In *Drosophila*, *Pitx* is expressed in several tissues including a subset of ventral somatic muscles and in neural cells (Vorbrüggen *et al.* 1997). Paralogs in *P. tepidariorum* also show metameric patterning along the ventral, with *Pt-Pitx1* most similar to the *Drosophila* CNS expression and *Pt-Pitx2* having both CNS and muscle tissue like staining (Fig. 3.14C). This suggests that *Pitx* paralogs in *P. tepidariorum* have undergone sub-functionalisation.

3.6.3.4 Pt-Zfh1 *and* Pt-Zfh2 *respectively mimic early and late expression of the* Drosophila *ortholog*

The expression of *Pt-Zfh1* is similar to that of the *Drosophila* ortholog *Zfh2*, with strong expression in the brain and ventral CNS at embryonic stages (Lai *et al.* 1991). Later in *Drosophila* leg imaginal discs *Zfh2* expression goes from an initially broad domain at the centre of the disc, to rings of expression in each segment and expression in the tarsus is necessary for its development (Guarner *et al.* 2014). The characteristics of *Dm-Zfh2* in later development are very reminiscent of firstly the *Pt-Zfh1* expression of the in limb buds, but then *Pt-Zfh2* has rings of expression and is maintained at the distal region of the limbs (Fig. 3.14D). Together it appears that Zfh paralogs have diverged to share early and late roles in with some overlap still in the ventral CNS tissue.

3.6.4 Duplication and Divergence of Homeobox Genes in Arachnopulmonates

The surveying of homeobox genes in chelicerates identified a signal of pervasive gene duplication corresponding to a WGD event that has likely occurred in the ancestor of Arachnopulmonata. This has contributed to the expansion of homeobox genes in these species as well as tandem duplications. Since the WGD the duplicated Hox clusters are still retained in the genomes of spiders and scorpions. Furthermore, other homeobox gene clusters were present in the genome of *P. tepidariorum* such as the duplicates of the NK cluster and the single HRO cluster, revealing that different clusters have had varying levels of retention. Subsequent to the WGD, the CDS regions and spatio-temporal expression dynamics have diverged between paralogs. These differences between paralogs may underlie evolutionary changes that are specific to arachnopulmonates. Sequencing the genomes of more arachnopulmonate and non-arachnopulmonate species will help to clarify the timing of the WGD event. Furthermore, the functional testing of paralogs and their respective orthologs in non-arachnopulmonate chelicerates will provide clues so far as whether paralogs have undergone sub- and/or neo-functionalisation of their roles in development.

4 Duplication of miRNAs in Xiphosura and Arachnopulmonata

The evidence from chelicerate genomes, transcriptomes and gene annotations, such as homeobox genes, currently suggests that the arachnopulmonate and xiphosuran (horseshoe crabs) lineages have independently undergone WGDs (Kenny *et al.* 2015a; Schwager *et al.* 2017). However, the signal of WGD has exclusively been derived from the perspective of coding genes. Further information about the duplication of non-coding genes may help to add compelling evidence for WGD and at least provide a perspective on the evolution of non-coding genes post-duplication in these chelicerate lineages.

MicroRNAs are a well-characterised type of small non-coding RNAs that fine-tune protein synthesis by binding to mRNAs and regulating their translation (Eichhorn *et al.* 2014). Comparative studies in animals such as vertebrates (Heimberg *et al.* 2008; Field *et al.* 2014; Fromm *et al.* 2015), arthropods (Campbell *et al.* 2011; Rota-Stabelli *et al.* 2011; Wiegmann *et al.* 2011), flatworms (Fromm *et al.* 2013) and annelids (Helm *et al.* 2012) have contributed greatly to our knowledge of microRNA biogenesis, function and evolution (Hatfield *et al.* 2005; Okamura *et al.* 2008; Marco *et al.* 2010; Griffiths-Jones *et al.* 2011; Marco *et al.* 2013; Chen *et al.* 2014; Marco 2014; Mohammed *et al.* 2014a; Mohammed *et al.* 2014b; Yang *et al.* 2014). However, in the morphologically diverse group of arthropods the current understanding of microRNA evolution is primarily limited to hexapod lineages. The only chelicerates that have been sampled are mites and ticks (Barrero *et al.* 2011; Rota-Stabelli *et al.* 2011; Zhou *et al.* 2013; Liu *et al.* 2014a; Luo *et al.* 2015), and therefore this represents insufficient taxon sampling to assess microRNA duplication in arachnopulmonates and xiphosurans.

To address the deficiency in chelicerate microRNA repertoires, small RNA sequencing of the first ten stages of *P. tepidariorum* development was used to predict an embryonic microRNA repertoire in this spider (Leite *et al.* 2016). This data set was then used to survey the microRNAs present in other arachnopulmonates and xiphosurans and to identify if there were duplicated microRNAs.

4.1 MicroRNA Prediction and Annotation in *P. tepidariorum*

In total the ten stages of RNA sequencing produced 177,789,607 raw reads (Table 4.1). The distribution of the trimmed read sizes showed two peaks at 22 bp and 29 bp, indicating the presence of a pool of mature miRNAs and presumptive piRNAs, respectively (Fig. 4.1A) (Saito *et al.* 2006). Allowing for up to one-mismatch 64.7% of the trimmed reads mapped to the *P. tepidariorum* genome with 7.3% discarded because they mapped to more than five loci. Reads failing to align with one mismatch, when re-mapped allowing for two mismatches, produced fewer mapped reads to the *P. tepidariorum* genome (<0.5%) compared to the *Tetranychus urticae* genome (<5%), suggesting that either this pool of unmapped reads may be contamination and not just of *P. tepidariorum* origin. However, when using the same unmapped reads that failed to align with one mismatch, to re-mapped with two mismatches against the later available DoveTail assembly of the *P. tepidariorum* genome (Schwager *et al.* 2017), the percentage of mapped reads was considerable (<37%), suggesting that the first assembly was not as complete.

Using only the zero mismatch mapped reads, miRDeep2 analysis initially predicted 278 miRNAs. Of these predicted miRNAs, 130 had either less than 50% of reads that had the same 5' end, low read numbers, or were found repeatedly in the *P. tepidariorum* genome. While these discarded candidates could potentially be functional miRNAs, they were removed to give a conservative final prediction of 148 miRNAs that are expressed during embryogenesis in *P. tepidariorum* (Table 4.2; App. Table 4.1) and that were used for subsequent analysis. Sequences of the hairpins and mature miRNAs for these 148 genes can be found in App. Tables 4.1 and 4.2 respectively.

Table 4.1: Mapping of small RNA sequencing to *P. tepidariorum* genome

Stage	Total reads	Trimmed & >16bp	Mapped to genome (v=0)	Max5 (v=0)	Unmapped (v=0)	microRNA (v=0)	Mapped to genome (v=1)	Max5 (v=1)	Unmapped (v=1)	microRNA (v=1)	% of mapped reads that map to microRNAs
S1	14670183	12533322	5501884	518372	6513066	236684	1671146	227393	4614527	9101	3.4
S2	18181733	16823532	5988725	744152	10090655	315554	2439844	480244	7170567	15666	3.9
S3	17568712	16627206	8265320	864022	7497864	946488	2282025	283322	4932517	56711	9.5
S4	17549176	16419870	8473925	973457	6972488	1065284	2279250	283399	4409839	53508	10.4
S5	20122326	17903718	8933566	847627	8122525	961995	2532699	302259	5287567	52036	8.8
S6	18613685	17009973	9412378	877536	6720059	1239497	2464229	233338	4022492	69513	11.0
S7	17925609	16740756	9727636	1060800	5952320	1599347	2486353	266881	3199086	84175	13.8
S8	21020797	19356951	10465447	1114952	7776552	1569038	2748556	280114	4747882	77950	12.5
S9	15635765	12850020	6620366	742568	5487086	1188753	1927672	270742	3288672	78708	14.8
S10	16501621	12993441	6987947	978240	5027254	1333422	1803126	236352	2987776	81615	16.1
All stages	177789607	159258789	80377194	8721726	70159869	10456062	22634900	2864044	44660925	578983	10.7

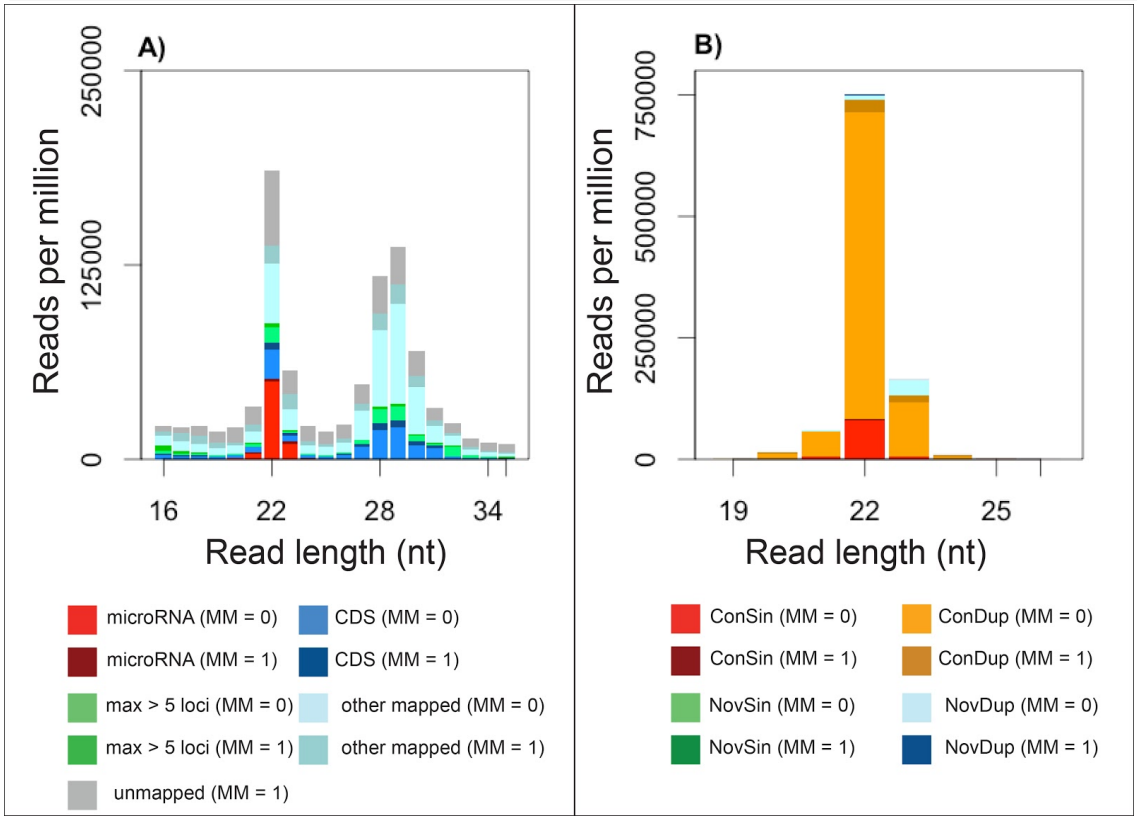


Figure 4.1: Distribution and assignment of pooled small RNA-Seq from *P. tepidariorum*. (A) All the mapped reads and **(B)** miRNAs that are conserved/novels and single copy/multi copy

Orthology of the 148 predicted miRNAs was then inferred using BLAST (Altschul *et al.* 1990) comparisons to metazoan pre-miRNAs from miRBase (Kozomara and Griffiths-Jones 2014) and these genes were grouped into families based on RALEE multiple alignments (Griffiths-Jones 2005). This approach identified 37 conserved families consisting of 101 miRNAs, with similarity to at least one previously annotated metazoan miRNA in miRBase (Table 4.2). The remaining 47 were therefore classified as "novel" *P. tepidariorum* miRNAs despite some having putative orthologs to other arachnopulmonate species surveyed (Fig. 4.2). Note that the possibility cannot be ruled out that the novel miRNA set includes some genes conserved in other species but diverged in sequence past this approach's ability to recognize them as homologs, even using structure-aware RNA homology search tools.

Following from the annotation and alignment process the miRNAs were grouped based on whether they were found as single copy or if the miRNA family had been duplicated and therefore contained multiple miRNAs (multi copy miRNAs). Of the conserved miRNAs, 15 families contained 79 miRNAs, leaving 22 single copy families (Table 4.2). For the novel miRNAs, there were 22 single copy families, and 25 miRNAs belonging to nine families (Table 4.2). The prediction and annotation process therefore suggests that at least 36% of *P. tepidariorum* miRNA families have been duplicated at least once. These duplicated miRNAs were highly abundant in the RNA-Seq data (Fig. 4.1B) with 88.14% (1.12 % per miRNA) and 6.29% (0.25% per miRNA) of the miRNA

reads being assigned to duplicated conserved and duplicated novel miRNA families respectively, and 5.3% (0.24% per miRNA) and 0.24% (0.01% per miRNA) for single copy conserved and single copy novel miRNA families respectively.

Table 4.2: MicroRNA annotation and copy number in *P. tepidariorum*

Single Copy microRNAs		Multi Copy microRNA (copy no. per family)			
Conserved	Novel	Conserved		Novel	
pte-bantam	*pte-mir-11938*	*pte-iab-4*	(2)	*pte-mir-11951*	(2)
pte-let-7	*pte-mir-11939*	*pte-iab-8*	(2)	*pte-mir-11959*	(2)
pte-mir-1	*pte-mir-11940*	*pte-mir-100*	(2)	*pte-mir-11962*	(2)
pte-mir-10	*pte-mir-11941*	*pte-mir-125*	(2)	*pte-mir-11968*	(2)
pte-mir-14	*pte-mir-11943*	*pte-mir-184*	(2)	*pte-mir-11961*	(3)
pte-mir-190	*pte-mir-11944*	*pte-mir-193*	(2)	*pte-mir-11963*	(3)
pte-mir-210	*pte-mir-11945*	*pte-mir-263*	(2)	*pte-mir-11965*	(3)
pte-mir-275	*pte-mir-11946*	*pte-mir-276*	(2)	*pte-mir-11942*	(4)
pte-mir-279	*pte-mir-11948*	*pte-mir-277*	(2)	*pte-mir-11947*	(4)
pte-mir-281	*pte-mir-11949*	*pte-mir-278*	(2)		
pte-mir-305	*pte-mir-11950*	*pte-mir-29*	(2)		
pte-mir-315	*pte-mir-11952*	*pte-mir-375*	(2)		
pte-mir-317	*pte-mir-11953*	*pte-mir-71*	(2)		
pte-mir-34	*pte-mir-11954*	*pte-mir-87*	(3)		
pte-mir-3477	*pte-mir-11955*	*pte-mir-993*	(3)		
pte-mir-3931	*pte-mir-11956*	*pte-mir-92*	(4)		
pte-mir-7	*pte-mir-11957*	*pte-mir-2*	(9)		
pte-mir-745	*pte-mir-11958*	*pte-mir-3791*	(34)		
pte-mir-8	*pte-mir-11960*				
pte-mir-9	*pte-mir-11964*				
pte-mir-96	*pte-mir-11966*				
pte-mir-981	*pte-mir-11967*				

4.2 Identification of Conserved *P. tepidariorum* miRNA Families in Other Chelicerates

All of the 148 predicted *P. tepidariorum* miRNAs were then searched for in the genomes of other chelicerates and also compared to the repertoire of miRNAs present in other selected ecdysozoans (Fig. 4.2). This identified 32 conserved miRNA families from *P. tepidariorum* in two other spiders and a scorpion. The conserved families of mir-14, mir-3477 and mir-3791 were present in *S. mimosarum*, but absent in *A. geniculata* and *C. sculpturatus* (Fig. 4.2). The mir-3791 family appears to be absent from the genomes of all the other chelicerates, except *S. mimosarum* and *L. polyphemus*, but is present in some insects (*T. castaneum* and *A. mellifera*) that were surveyed and also shares sequence similarity with the *mir-35-41* cluster in *C. elegans* (Massirer *et al.* 2012; Fromm *et al.* 2013; McJunkin and Ambros 2014). It is not possible to tell whether the presence of a mir-3791 sequence in the two spiders and *L. polyphemus* but not other chelicerates is due to differential loss or is the result of convergent evolution (Fig. 4.2).

The other arachnids investigated showed variable patterns of retention and loss of the 35 conserved miRNA families found in *P. tepidariorum*. There were nine miRNA families that were present in all of these animals. The lowest number of families in any chelicerate surveyed was 23 in *V. destructor* (Fig. 4.2). Apart from the spiders and scorpion, the mir-193 family is likely to have been lost from the genomes of other chelicerates including *L. polyphemus* (Fig. 4.2). Note that this miRNA is commonly lost in many metazoan lineages (Tarver *et al.* 2013). Other potential losses in chelicerates include mir-14 and mir-3477, which was only present in *P. tepidariorum* and *S. mimosarum*, *L. polyphemus* has also lost mir-981 and mir-279, but has retained the other 31 conserved *P. tepidariorum* miRNA families (Fig. 4.2).

Comparing the conserved miRNA families found in *P. tepidariorum* to other non-chelicerate arthropods reveals that 23 out of 35 are found in the genomes of all the insects, the myriapod and the crustacean in our survey (Fig. 4.2). Interestingly, mir-96 may have been lost in insects but retained in the crustacean, *D. pulex*, as well as most chelicerates, while mir-3931 appears to have evolved in chelicerates as previously reported (Tarver *et al.* 2013), although it may also have been lost in the common ancestors of the other arthropod subphyla (Fig. 4.2).



Figure 4.2: MicroRNA family copy number in *P. tepidariorum* and other ecdysozoans. Presence of *P. tepidariorum* single copy and multi copy miRNA families in other spiders (purple), scorpions (magenta), horseshoe crab (blue) parasitiformes (green), acariformes (dark green), mandibulates (grey) and a nematode (black). The spiders, the scorpion and the horseshoe crab all display large numbers of multi copy miRNA families relative to other species.

4.3 Novel *P. tepidariorum* miRNAs are Largely Species-Specific

Of the 22 novel single copy microRNA families identified in *P. tepidariorum*, two were identified in the genome of *A. geniculata* and three in *S. mimosarum*, while only one (mir-11960) was found in *C. sculpturatus*, and none were found in any other chelicerate (Fig. 4.2). As described above, nine novel multi copy microRNA families were also identified in *P. tepidariorum*. However of these nine families, only mir-11951 was identified in *S. mimosarum* and mir-11942 in *A. geniculata*. Though interestingly, both of these microRNA families also contained duplicates in these other spiders (Fig. 4.2). Taken together this analysis suggests that most of the novel microRNAs identified in *P. tepidariorum* are likely to be young and have recently evolved in the lineage leading to this spider.

4.4 Incidence of Multi Copy miRNA Families in Other Chelicerates

As mentioned above, out of the 35 conserved miRNA families in *P. tepidariorum* 15 families were represented by two or more paralogs. There were also many multi copy families in the other spiders (*S. mimosarum* and *A. geniculata*) and the scorpion (*C. sculpturatus*) with 16, 17 and 16 out of 35 conserved families represented by two or more paralogs, respectively (Fig. 4.2). There was considerable overlap of multi copy families between these species (*S. mimosarum*, *A. geniculata* and *C. sculpturatus*) and *P. tepidariorum* with 12, 11 and 10 families represented by at least two paralogs in the respective species and *P. tepidariorum* (Fig. 4.2). Moreover, eight families are multi copy among all four of these species (Fig. 4.2). In contrast, only up to six (in the parasitiform *M. occidentalis*) multi copy miRNA families were observed in any other chelicerate, or ecdysozoan, surveyed with the notable exception of *L. polyphemus* (Fig. 4.2). In this horseshoe crab, 26 out of the 30 miRNAs found (with respect to the conserved *P. tepidariorum* miRNAs) were represented

by multiple copies, with 21 present as three or more copies. These included mir-8 and mir-10, which were not duplicated in any of the other ecdysozoans that were surveyed (Fig. 4.2).

In the mandibulates, parasitiformes and acariformes analysed, approximately 75-90 % of miRNA are represented by a single miRNA (App. Fig. 4.1). In contrast, only ~40-55 % of miRNAs were present as single copy in the spiders and scorpion, while approximately 35-40 % contain two copies in these species (App. Fig. 4.1). In the horseshoe crab miRNAs were represented by one up to 17 copies; however most contained five copies (App. Fig. 4.1).

4.5 Synteny of miRNA duplications

4.5.1 Small and Large Scale Duplications of miRNAs in *P. tepidariorum*

To further understand the scale of these duplication events the positions of paralogous miRNAs were compared in both the first and second (DoveTail) versions of the *P. tepidariorum* genome as well as other chelicerate genomes. For *P. tepidariorum* the 148 miRNAs identified were located on 122 scaffolds in first genome version. As expected the *P. tepidariorum* miRNAs were located on only 71 scaffolds of the DoveTail genome, indicative of the improved assembly (Fig. 4.3). The genomic arrangement of *P. tepidariorum* miRNA paralogs is suggestive of both large and small scale duplication events.

In the *P. tepidariorum* DoveTail genome there are four scaffolds (#497, #883, #142, #6239) that contain miRNAs found in the arthropod Hox cluster (*mir-993*, *mir-10*, *iab-4/8*) (Fig. 4.3A). Two of these scaffolds #497 and #883 contain Hox genes (discussed in Chapter Three) but scaffolds #142 and #6239 each contain a lone *mir-993* paralog without any other homeobox genes. Scaffolds #497 and #883 both contain paralogs of *iab-4/8*, and scaffold #883 contains *mir-993b-1* and scaffold #497contains the single copy of *mir-10*.

There are several other specific miRNA clusters that are also duplicated in *P. tepidariorum*. The *mir-71*/*mir-2* cluster is an invertebrate specific miRNA cluster that has expanded in arthropods probably due to tandem duplications of *mir-2* (Marco *et al.* 2010; Marco *et al.* 2012a). In the acariform and parasitiform lineages, the copy numbers of *mir-71* and *mir-2* are variable, though usually there is one copy of *mir-71* and two copies of *mir-2* (Fig. 4.2). *P. tepidariorum* contains one cluster containing a single copy of *mir-71* and three copies of *mir-2*, and a second cluster, on a different scaffold, also with one copy of *mir-71*, but with six copies of *mir-2* (Fig. 4.3B). A second miRNA cluster that appeared to be duplicated is the mir-100/let-7/mir-125 cluster. In *P. tepidariorum* there is one complete cluster, and additional copies of *mir-100* and *mir-125* on separate scaffolds but with no copy of *let-7* (Fig. 4.3C). The mir-3791 family, which is represented by 34 paralogs in *P. tepidariorum*, is found clustered in the genome (Fig. 4.3D). Thirteen of the *mir-3791* paralogs are located on one scaffold (786) that is flanked on either side by over 3 Mb of DNA, indicating the probable complete assembly of the cluster. Additionally, there are also two other clusters on scaffolds 73 and 3068 that contain ten and eight *mir-3791* paralogs, respectively. The

remaining three paralogs are dispersed in the genome with one paralog alongside the single copy of *mir-279*.

As well as duplicated miRNA clusters on separate scaffolds there are also miRNA families that had paralogs locating to different scaffolds and so were dispersed duplicates. Eight of these were conserved families (mir-87*, mir-92, mir-184, mir-193, mir-276, mir-277, mir-278, mir-375) and five were novel families (mir-11942*, mir-11947, mir-11963, mir-11965*, mir-11968) (App. Fig. 4.X). The presence of these dispersed duplicates provide further supports the occurrence of large-scale duplications. However, there were also tandem duplicates in *P. tepidariorum* in the above families marked with an asterisk. Further tandem duplications were also observed in six miRNA families (mir-29, mir-263, mir-11951, mir-11959, mir-11961, mir-11962) (App. Fig. 4.X). This evidence shows that both large-scale duplication and small-scale duplications have played a role in the expansion of *P. tepidariorum* miRNAs

4.5.2 Small and Large Scale Duplications of miRNAs in Other Chelicerates

Since the presence of duplicated miRNA clusters and many dispersed paralogs in the *P. tepidariorum* genome is suggestive of a large scale duplication event, it was of interest to identify if other chelicerates shared the duplication of miRNA clusters. Like *P. tepidariorum*, other spiders showed evidence of duplication of the miRNAs commonly found in the arthropod Hox cluster. *S. mimosarum* contains one copy of *mir-10*, two *iab-4/8* paralogs and three *mir-993* paralogs, all of which are located on different scaffolds. *A. geniculata* has a single *mir-10* copy, two *mir-993* paralogs, and three *iab-4/8* paralogs, of which two *iab-4/8* paralogs are on the same scaffold, but all other *A. geniculata* Hox miRNAs are on different scaffolds (Leite *et al.* 2016). Somewhat similar to spiders, scorpions also display duplicated Hox miRNAs (Fig. 3.2). It has already been shown that *M. martensii* has got two copies of *mir-993*, *iab-4/8* and unusually *mir-10* (Pace *et al.* 2016). In *C. scuplturatus* there was only one copy of *mir-10*, but there were two copies of *mir-993* and *iab-4/8*, which due to the quality of the genome can be partly arranged into syntenic clusters with Hox genes (Fig. 3.5). In the *L. polyphemus* there are six *mir-10*, five *mir-993*, and three *iab-4/8* copies, all of which are located on different scaffolds. In contrast to the duplication of Hox miRNAs in spiders, scorpions and horseshoe crabs, the *I. scapularis* genome contain one scaffold with ten Hox genes and one copy of *mir-993*, *mir-10*, and *iab-4/8* (Fig. 3.5).

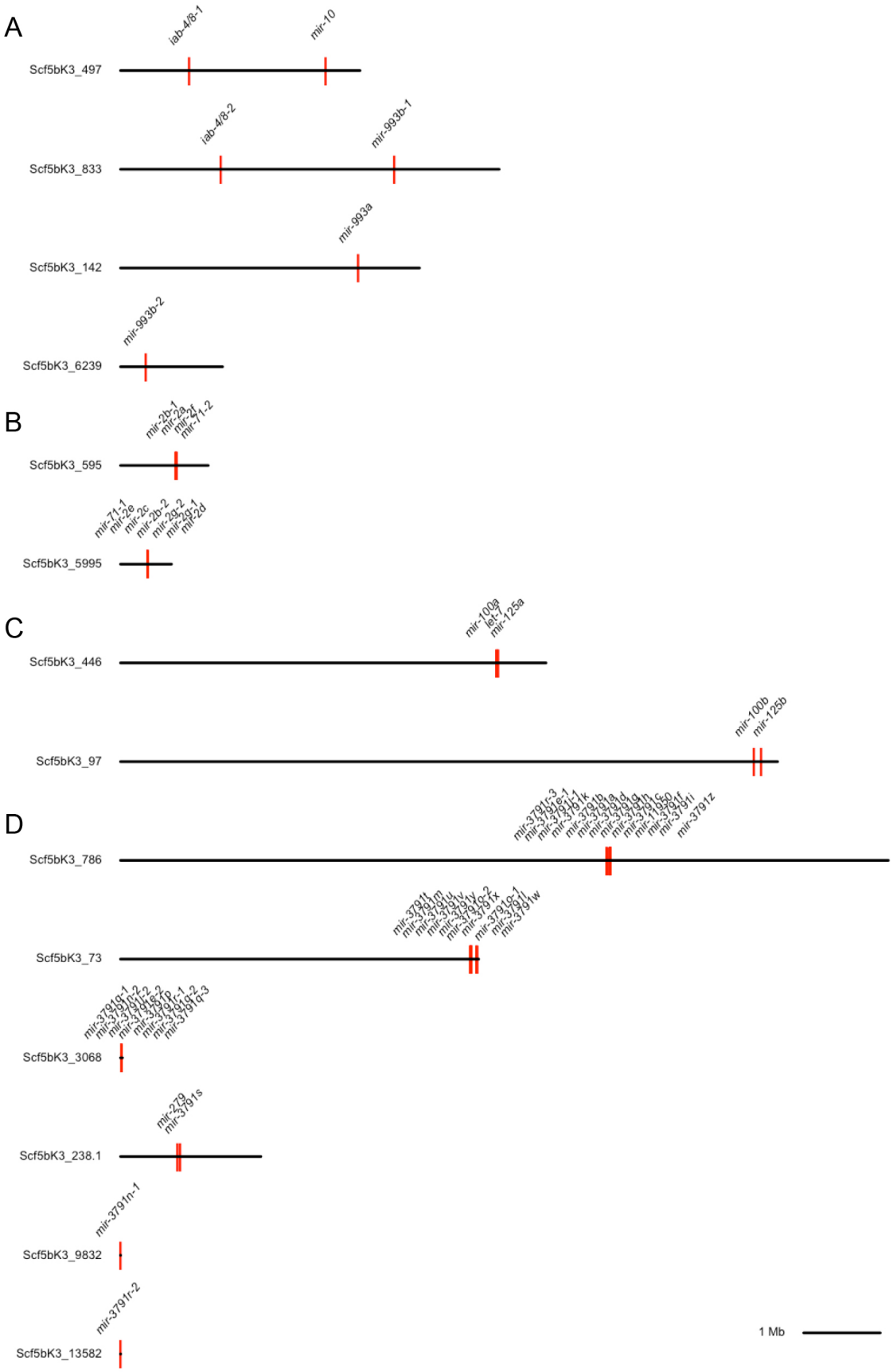


Figure 4.3: Clusters of miRNA paralogs in the *P. tepidariorum* DoveTail genome assembly. **(A)** MicroRNAs that are commonly found in the arthropod Hox cluster. **(B)** Clusters of *mir-2*/*mir-71*. **(C)** Clusters of *mir-100*/*let-7*/*mir-125*. **(D)** Clusters and dispersed paralogs of the mir-3791 family.

There were other specific microRNA clusters that were also duplicated in other spiders, scorpions and the horseshoe crab including, at least two copies of *mir-71* and consistently more than two copies of *mir-2* (Fig. 4.2). In the spiders and the scorpion, these microRNAs were generally found as two separate clusters on different scaffolds (Fig. 4.4A). In the *S. mimosarum* and *C. sculpturatus* genomes there was one scaffold containing one copy of *mir-71* and four copies of *mir-2* (Fig. 4.4A), and a second scaffold with one copy of *mir-71* and three copies of *mir-2* (Fig. 4.4A). In *A. geniculata* there is one complete cluster with one *mir-71* and four *mir-2* copies, all on one scaffold (Fig. 4.4A). There is another scaffold with one copy of *mir-71* and one copy of *mir-2* located ~6 kb from the end of the scaffold. There are also two other copies of *mir-2*, each on ~1 kb scaffolds. The arrangement of *L. polyphemus mir-71* and *mir-2* miRNAs is much more fragmented than other chelicerates described (Fig. 4.4A). In this horseshoe crab eight copies of *mir-71* located on seven scaffolds were identified, of which six also contain at least one copy of *mir-2* (Fig. 4.4A). The seven copies of *mir-2* that are located on six different scaffolds are all found close to the ends of the scaffolds, and may therefore possibly be fragments of larger *mir-71*/*mir-2* clusters (Fig. 4.4A).

A second miRNA cluster that appeared to be duplicated in some species is the *mir-100*/*let-7*/*mir-125* cluster (Fig. 4.4B). However, in all lineages there appeared to be the possible loss or rearrangement of at least one cluster, although it is not possible to exclude that this is an artefact of genome assembly. The arrangement of these miRNAs in *S. mimosarum* is similar to *P. tepidariorum* although the two *mir-100* paralogs are on separate scaffolds (Fig. 4.4B). *A. geniculata* has a complete *mir-100*/*let-7*/*mir-125* cluster and two additional copies of *mir-125* are found on separate scaffolds (Fig. 4.4B). In *C. sculpturatus*, there is a single cluster, which appears to have been rearranged, similar to one of the three *L. polyphemus* clusters (Fig. 4.4B).

Therefore, some Hox cluster miRNAs, the *mir-71*/*mir-2* and *mir-100*/*let-7*/*mir-125* clusters appear to have been duplicated in spiders and scorpions (Arachnopulmonata) as well as *L. polyphemus* but not other chelicerates or mandibulates that were surveyed. Taken together, these results show that both tandem duplications and larger scale duplications of entire clusters, followed by subsequent lineage specific expansion or loss of miRNAs, have contributed to the evolution of chelicerate miRNA repertoires.

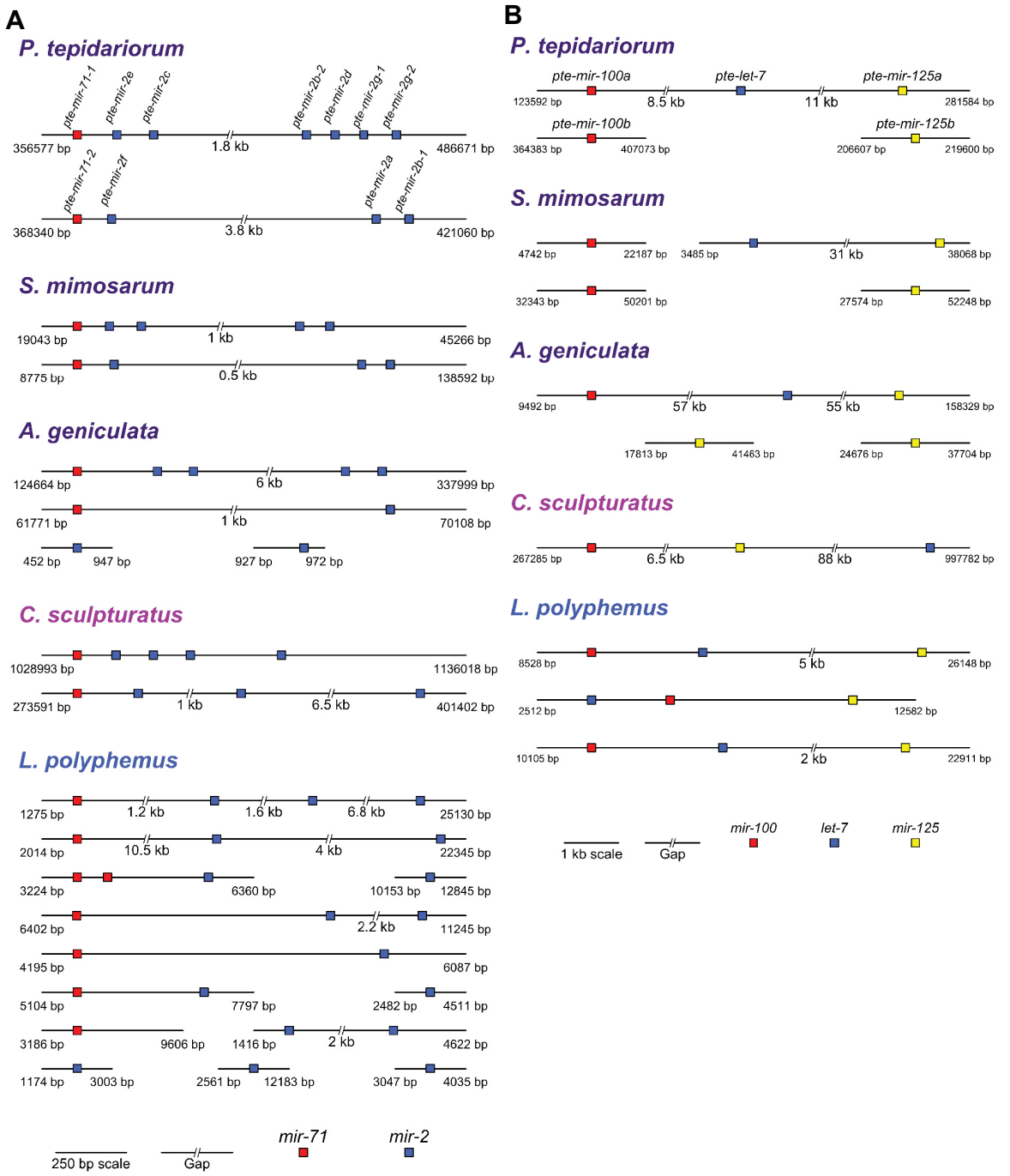


Figure 4.4: Duplication of miRNA clusters in *P. tepidariorum* and other chelicerates. (A) The *mir-71*/*mir-2* cluster is duplicated in spiders (purple), the scorpion (magenta) and the horseshoe crab (blue). Each species displays lineage specific retention, loss and further small segmental duplications. **(B)** The *mir-100*/*let-7*/*mir-125* cluster is duplicated in spiders (purple) and the horseshoe crab (blue). Unlike the *mir-71*/*mir-2* cluster this cluster has lower retention of both clusters and greater rearrangements across species. The position of the left most miRNA and the total scaffold length are indicated on the left and right hand side respectively.

4.6 Divergence in Arm Usage in Multi Copy miRNAs

Duplicate miRNAs have the potential to diversify their functions by a number of evolutionary mechanisms, including arm usage (Okamura *et al.* 2008; de Wit *et al.* 2009; Marco *et al.* 2010; Griffiths-Jones *et al.* 2011). Inspection of miRNA arm usage in *P. tepidariorum* suggests that there is a bias towards the mature sequence deriving from the 3p arm of the hairpin (Fig. 4.5A). The distribution was statistically different from that in both *T. castaneum* (D = 0.2124; *P* = 0.002) and *D. melanogaster* (D = 0.2955; $P = 5.4 \times 10^{-7}$), but was not different between *T. castaneum* and *D. melanogaster* (D = 0.1343; *P* = 0.08) (Kolmogorov-Smirnov test).

The relative arm usage was then compared for individual miRNAs between *P. tepidariorum*, *T. castaneum* and *D. melanogaster* and observed multiple cases of arm switching among these species (Fig. 4.5B-C). Considering first the miRNAs that have unique mature sequences in *P. tepidariorum*, only *mir-275* and *mir-3477* exhibit differences in relative arm usage between *P. tepidariorum* and *T. castaneum,* (Fig. 4.5B). Comparisons between *P. tepidariorum* and *D. melanogaster* shows that *mir-10*, *mir-993a*, *mir-278b*, *mir-281*, also display changes in arm use (Fig. 4.5C).

There were also occurrences of arm switching for miRNAs that contained only one unique mature sequence in *P. tepidariorum*. Between *P. tepidariorum* and *T. castaneum, mir-993b-1*, *mir-2b-1* all exhibit differences in relative arm usage (Fig. 4.5B). Comparing *P. tepidariorum* with *D. melanogaster*, showed that *mir-993b-2*, both copies of *mir-276*, and again *mir-2b-1* exhibited arm switching (Fig. 4.5C).

One of the *P. tepidariorum mir-3791* paralogs were identified to have been subject to relative arm switching, and another paralog that had similar 5p and 3p expression, which is in contrast with the rest of this mostly 3p dominant family in *P. tepidariorum* (Fig. 4.5D). Both of these miRNAs have unique mature sequences in *P. tepidariorum*, though some of the other paralogs of this family do not have unique sequences for both mature products (App. Table 4.2).

It was also possible to compare the arm usage among the paralogs of the nine novel multi copy miRNA families identified in *P. tepidariorum* (Fig. 4.2). The mir-11942 family exhibited differential arm usage between paralogs (Fig. 4.5E). Another family, mir-11961, also showed variation in arm usage between paralogs, however there was less striking differences between paralogs compared to those observed in the mir-11942 family (Fig. 4.5E). Importantly, the paralogs of these two miRNA families all had unique mature sequences in *P. tepidariorum* (App. Table 4.2). The other novel multi copy families showed relatively similar arm usage between paralogs. However, these belonged to miRNA families with one non-unique mature sequence in *P. tepidariorum* in relation to their paralogs (App. Table 4.2).

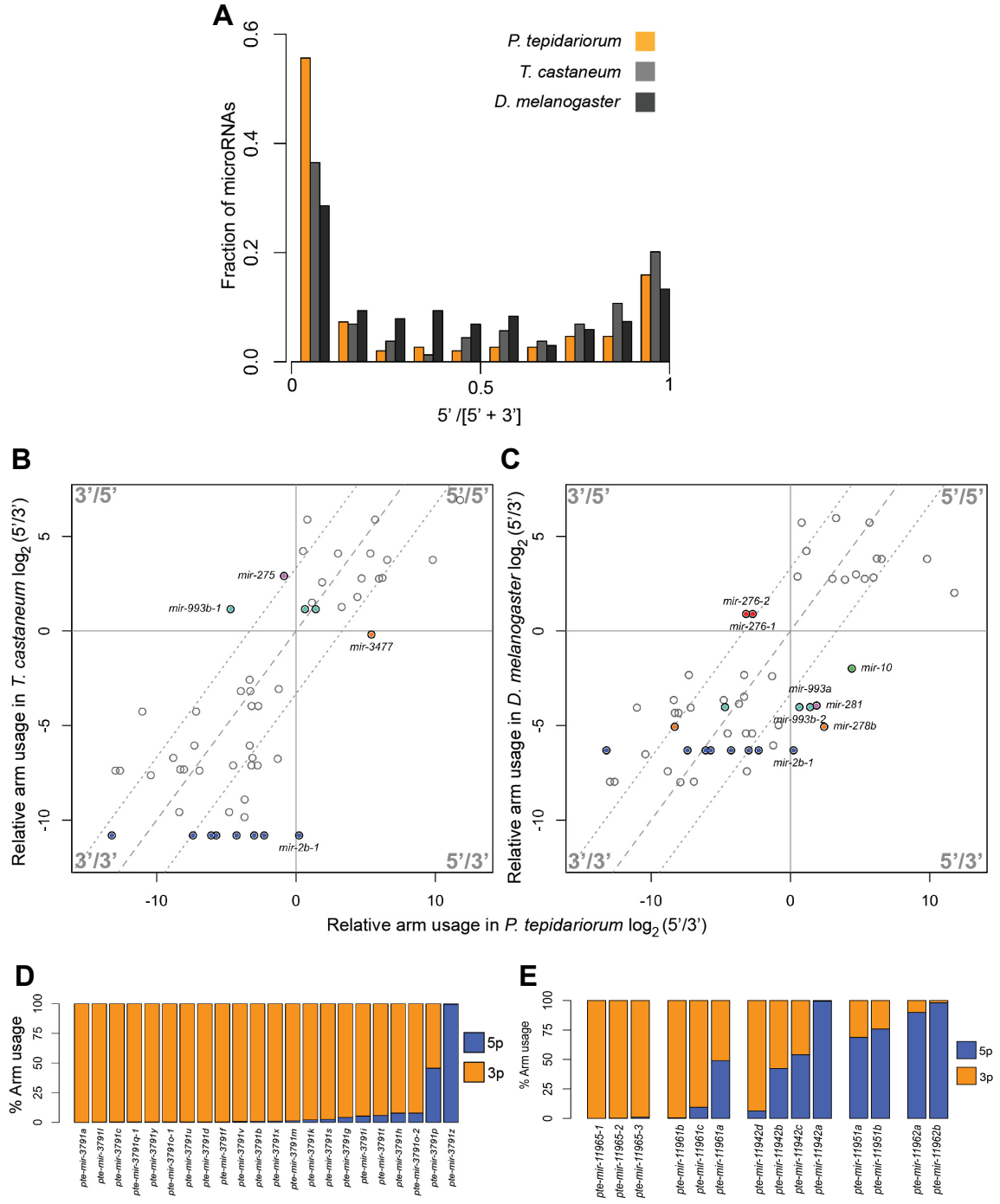


Figure 4.5: Relative arm usage changes in miRNAs. (A) *P. tepidariorum* has bias towards 3p arm usage. The frequency distribution of the ratio of the number of reads assigned to the 5p strand out of total assigned reads to a miRNA show that there is a bias to 3p strand usage in *P. tepidariorum* compared to *T. castaneum* and *D. melanogaster*. **(B – C)** Comparisons of the relative strand usage of *P. tepidariorum* miRNAs to **(B)** *T. castaneum* and **(C)** *D. melanogaster* show miRNAs that have undergone strand switching in the 3'/5' and 5'/3' quadrants. The dashed line indicates the theoretical expectation for conserved arm usage between the two species. Dotted lines limit the boundaries of the dashed line to less than 10-fold differences in arm usage. Note, in **(B)** paralogs of *mir-71* are located in exactly the same place. **(D)** Arm usage changes in the mir-3791 family. Comparisons of the arm usage in the mir-3791 family of *P. tepidariorum* show a general 3p dominance with one paralog having a switched to 5p dominance. **(E)** Arm usage changes in novel *P. tepidariorum* miRNAs. Paralogs in families mir-11942 and mir-11961 show differential arm usage between paralogs, with a switch in the mir-11942 family. Other families have generally similar arm usage between paralogs.

4.7 Discussion

4.7.1 The Evolution of the *P. tepidariorum* miRNA Repertoire

In this chapter the repertoire of miRNAs expressed during embryogenesis in *P. tepidariorum* was predicted and used to identify these genes in other chelicerate genomes to examine patterns of duplication during the evolution of these genes in arthropods and other animals. From the initial 278 miRDeep2 predictions, the analysis focused on a conservative total of 148 miRNAs representing 66 families expressed during *P. tepidariorum* embryogenesis (Fig. 4.2). This number is similar to the complement of 172 miRNAs expressed during *D. melanogaster* embryogenesis (Ninova *et al.* 2014). However, *D. melanogaster* has a total of 256 miRNAs identified across all life stages (Kozomara and Griffiths-Jones, 2014). Therefore, it is likely that further miRNAs expressed later in development and in adults remain to be identified in *P. tepidariorum*. Despite this, more miRNAs were characterised in this spider than previously identified in other arachnids (*R. microplus* [87], *I. scapularis* [49] and *T. urticae* [52]) (Wheeler *et al.* 2009; Barrero *et al.* 2011; Grbic *et al.* 2011) except the tick *Hyalomma anatolicum anatolicum* (Luo *et al.* 2015). Thus *P. tepidariorum* possesses one of the largest chelicerate miRNA repertoires characterised to date, and is the only one where small RNAs have been mapped against the specific corresponding genome sequence.

Approximately half of the miRNA families identified in *P. tepidariorum* by small RNA-Seq were conserved in other metazoans (Fig. 4.2). These included 21 out of 31 families that are common to most bilateral animals, and seven out of the twelve families present in protostomes (Tarver *et al.* 2013). The losses include the mir-33, mir-219, mir-2001 and mir-1993 families, which have been commonly lost in metazoans (Tarver *et al.* 2013). *P. tepidariorum* also contains panarthropod (mir-276 and mir-305), arthropod (iab-4/8 and mir-275) and chelicerate specific (mir-3931) miRNA families (Rota-Stabelli *et al.* 2011; Tarver *et al.* 2013). *P. tepidariorum* also lacks the mir-242 and mir-216 families, which are also lost in other arthropods, nor the mir-31 family, which is not found in Ixodidae, or the mandibulate specific mir-282 or mir-965 families (Rota-Stabelli *et al.* 2011; Tarver *et al.* 2013). It is possible that due to the sampling of embryonic small RNA that some miRNA families have been missed (Tarver *et al.* 2013) as they may be expressed during other life stages. Indeed, it was possible to identify one copy of *mir-133*, two copies of *mir-137* and three copies of *mir-124* (Leite *et al.* 2016) in the genome of *P. tepidariorum* via BLAST searches, further increasing the number of miRNAs that were identified.

4.7.2 Evidence of Duplication of both Coding and Noncoding Genes in *P. tepidariorum*

Over one third of the conserved miRNA families and many of the novel miRNA families are present in multiple copies in *P. tepidariorum* (Table 4.2). This expansion appears to have resulted from both local tandem duplications and larger scale segmental duplications (Fig. 4.3 and 4.4; and

App. Fig. 4.1 and 4.2). Previous reports have found that some protein-coding genes in *P. tepidariorum* have also been duplicated (Posnien *et al.* 2014; Schwager *et al.* 2017). Of the 8917 *P. tepidariorum* transcripts identified as being orthologous to a *Drosophila* gene, approximately 28% are likely to be expressed from duplicated genes (Posnien *et al.* 2014). However, analysis in Schwager *et al.* (2017) revealed lower percentages of genes that were duplicated in arachnopulmonates but only single copy in *T. urticae* and *T. castaneum*. This therefore suggests that there has been greater retention of duplicated miRNAs compared to protein coding genes in *P. tepidariorum*. This difference is similar to previous estimates of miRNAs and coding gene retention following large-scale duplication (Berthelot *et al.* 2014).

4.7.3 Duplication and Divergence of miRNAs in Multiple Chelicerate Lineages

Evidence of gene duplication in chelicerates is not limited to *P. tepidariorum*. Analysis of the transcriptomes of representatives of the Mygalomorphae and Araneomorphae suggests that these spiders are likely to have shared a large-scale duplication event and that retained paralogs may have contributed to the origin and evolution of silk glands (Clarke *et al.* 2015). The Hox genes in the spider *C. salei* and scorpions, including *C. sculpturatus* and *M. martensii*, have also been found to be duplicated, though it is unclear whether these are shared or independent duplication events (Schwager *et al.* 2007; Sharma *et al.* 2014c; Di *et al.* 2015; Sharma *et al.* 2015). However, recent analysis of Hox genes in *P. tepidariorum* and *C. sculpturatus* genomes suggest that arachnopulmonates share a WGD (Schwager *et al.* 2017). In the xiphosuran lineage, there is also evidence of two rounds of WGD, which are thought to be independent of the duplication events that have occurred in spiders or scorpions (Nossa *et al.* 2014; Sharma *et al.* 2014b; Sharma *et al.* 2014c; Kenny *et al.* 2015a; Schwager *et al.* 2015; Schwager *et al.* 2017). The presence of mir-193 in Arachnopulmonata and its absence in all the other chelicerates that were surveyed further supports the hypothesis of independent duplication events.

In addition, this analysis also revealed many duplicate miRNA families in spiders, scorpions and a horseshoe crab (Fig. 4.2 and Fig. 4.4). It was found that *L. polyphemus* had the largest estimated miRNA repertoire among chelicerates. Interestingly, many miRNA families were found to be represented by multiple miRNAs in all three spiders and the scorpion that were surveyed (Fig. 4.2). These findings are consistent with a possible large-scale duplication event in the common ancestor of arachnopulmonates, and the two possibly independent rounds of WGD in horseshoe crabs (Nossa *et al.* 2014; Sharma *et al.* 2014b; Sharma *et al.* 2014c; Kenny *et al.* 2015a; Schwager *et al.* 2015; Schwager *et al.* 2017).

In contrast to the shared duplication of miRNAs among Arachnopulmonata species, there were differences in the retention of paralogs in some families (Fig. 4.2). These patterns of gain and loss may be due to genome assembly quality or a lack of small RNA sequencing from adult stages of *P. tepidariorum*. Alternatively, differential patterns of retention of miRNA paralogs potentially relates to differential retention of duplicate coding genes between these spider species (Clarke *et al.*

2015). It is also possible that the paralogous miRNAs found in some species could have been produced by lineage-specific duplication events. However, further analysis of the genomes and small RNA sequencing of all of these chelicerates is required to investigate these different evolutionary scenarios.

4.7.4 Evolution of miRNA Function

Despite the shared retention of many duplicated genes generated by putative tandem and large-scale duplication events, it is clear from the results that there are also common lineage specific losses (e.g. mir-3791 in *P. tepidariorum*) (Fig. 4.2, 4.3 and 4.4). Changes in miRNA biogenesis may have also contributed to the evolution of their function among chelicerates. The analysis identified that *P. tepidariorum*, like some other ecdysozoans (de Wit *et al.* 2009), has a general 3p arm bias with respect to preferential strand loading into the RISC. This bias is greater than that reported for 3p arm usage compared to *T. castaneum* and *D. melanogaster* (Marco *et al.* 2010) (Fig. 4.5A). This could perhaps be caused by the many paralogs of the mir-3791 and mir-2 families, which were generally 3p dominant, though both of these families do show instances of arm switching (Fig. 4.5B, C and D). However, the miRNA that switched in the mir-2 family did not contain completely unique mature sequences and relative arm switching may have been over or underestimated. Many of the miRNAs that show arm switching, relative to *T. castaneum* and *D. melanogaster*, also belong to multi copy miRNA families. These results suggest that duplication may facilitate functional change between paralogs and provides further evidence that miRNA duplication facilitates changes that can alter strand selection (de Wit *et al.* 2009).

There were also cases of single copy miRNAs undergoing arm switching between species. miR-10-5p is more abundant than miR-10-3p in both *P. tepidariorum* and *T. castaneum,* while miR-10-3p dominates in *D. melanogaster* (Griffiths-Jones *et al.* 2011). In both *P. tepidariorum* and *Drosophila,* mature products from both arms are expressed at detectable levels (Stark *et al.* 2007). The expression of both mature arms may be a feature that contributes to a miRNAs ability to switch strand usage between species.

The pervasive duplication and subsequent divergence in retention and copy number, as well as arm switching, that were identified among chelicerate miRNAs may have led to their sub-functionalisation and neo-functionalisation (Ohno 1970; Taylor and Raes 2004; Li *et al.* 2008; Innan and Kondrashov 2010; Abrouk *et al.* 2012; Huminiecki and Conant 2012; Wang and Adams 2015). These evolutionary differences, therefore, may have contributed to the divergence in the developmental programmes of chelicerates.

5 The Predicted Role of miRNAs in the Maternal to Zygotic Transition in *P. tepidariorum*

During oogenesis maternal instructions are loaded into the oocyte to coordinate the initial phases of development before the zygote becomes transcriptionally active (Tadros and Lipshitz 2009; Lee *et al.* 2014). This regulatory switch is known as the maternal to zygotic transition (MZT) and involves various mechanisms, such as miRNA clearance of maternal mRNAs, as well as changes in the abundance of maternal and zygotic repressors/activators, cell cycle timing and chromatin rearrangements (Tadros and Lipshitz 2009; Giraldez 2010; Lee *et al.* 2014; Liu *et al.* 2014b; Laver *et al.* 2015; Pires *et al.* 2016; Palfy *et al.* 2017)

Deployment of miRNAs to clear maternally supplied mRNAs has been identified in animals and also in plants (Baroux *et al.* 2008; Tadros and Lipshitz 2009). In the animal species studied so far, clusters of duplicate miRNAs have been found to regulate maternal genes. In the vertebrate lineage there appears to be orthologous miRNA families in fish (mir-430) (Giraldez *et al.* 2005; Giraldez *et al.* 2006) and amphibians (mir-427) (Lund *et al.* 2009) that regulate the MZT. Though, it is currently unclear if the orthologous miRNA families in mammals (miR-295 in mice, and miR-302, miR-372, miR-516-520 in humans and primates) that are expressed early in embryogenesis are also involved in the mammalian MZT (Giraldez 2010). Insects have also evolved to use miRNAs derived from a similar cluster to regulate their MZT (Bushati *et al.* 2008; Zondag *et al.* 2012; Ninova *et al.* 2014, 2016). They contain the mir-309~6 cluster which comprises several families (e.g. mir-6, mir-9, mir-279, mir-309, mir3791) that have undergone lineage specific rearrangements and duplications during insect evolution (Ninova *et al.* 2014). The spider *P. tepidariorum* also contains orthologs of *mir-3791* and *mir-279* miRNAs clustered in the genome (Chapter Four) (Leite *et al.* 2016). However, there is no current understanding of the regulation of the MZT in spiders and whether miRNAs are involved.

Therefore, this chapter aims to identify the timing of MZT in *P. tepidariorum* and predict the role of miRNAs in its regulation. Studying the miRNA regulation of the MZT will provide further insights not only into early developmental processes but also into gene duplication and the evolutionary patterns of functionally related clustered genes.

5.1 Identification of MZT in *P. tepidariorum*

5.1.1 Transitions in the Embryonic mRNA Transcriptome of *P. tepidariorum*

The MZT is characterised by the reduction in maternal mRNAs and the increase in transcripts from the zygote, which are commonly broken into an early (phase I) and later phase (phase II) (Tadros and Lipshitz 2009). To identify these changes in *P. tepidariorum* the embryonic expression profiles of both mRNAs and miRNAs were analysed.

The greatest changes in differentially expressed mRNAs that most likely correspond to the MZT were observed between S1-S2 and S3-S4 (Table. 5.1). The transition between S1-S2 contains the most up-regulated genes, possibly indicating the early onset of zygotic transcription at S2 (Table 5.1). This is congruent with findings that inhibition of RNA pol II with alpha-amanitin, arrests embryos as S2 (Pechmann 2016). There are also many down-regulated genes between S1-S2, however, the greatest number of down-regulated genes is observed between S3-S4 (Table 5.1). Of the 528 genes identified as significantly reduced between S3-S4 ($p < 0.05$) only three were not expressed at S1. Therefore the majority of down-regulated genes at S4 are those that likely have been maternally supplied. Furthermore, the dendrogram clustering of mRNAs groups S1-S3 separately from other stages, indicating that these early stages are most similar to one another and that the main regulatory transition occurs between S3 and S4 (Fig. 5.1A). This was additionally tested using principle component (PC) analysis, and showed that the first PC separates S1-S3 samples from S4-S10 (Fig. 5.2A). The second PC separates S1 from S2-S3, and also separates the later stages (S4-S10) in their temporal order (Fig. 5.2A). Thus, the embryonic mRNA expression profile from *P. tepidariorum* suggests that early zygotic genome activation occurs at S2 and specific genes continue to be up-regulated at S4. Furthermore, there is a reduction of maternally loaded mRNAs at S2 but is most severe at S4 (Table 5.1). This suggests that the MZT in *P. tepidariorum*, like other animals, can also be separated into an early and later phases (Tadros and Lipshitz 2009).

Table 5.1: Differentially regulated genes in *P. tepidariorum* development

	Number of Log2FoldChange mRNAs							
	All		***p* < 0.1**		***p* < 0.05**		***p* < 0.01**	
Stages	Up	Down	Up	Down	Up	Down	Up	Down
S1-S2	5402	1585	753	275	580	147	349	42
S2-S3	1440	2291	2	0	0	0	0	0
S3-S4	2414	4104	292	732	187	528	87	291
S4-S5e	1445	2285	5	8	3	2	0	0
S5e-S5l	1617	1973	9	2	7	2	0	0
S5l-S6	2740	1649	24	0	19	0	11	0
S6-S7	1886	2318	6	4	6	2	3	2
S7-S8	1575	1673	16	0	15	0	10	0
S8-S10	2239	1769	121	13	93	10	46	5

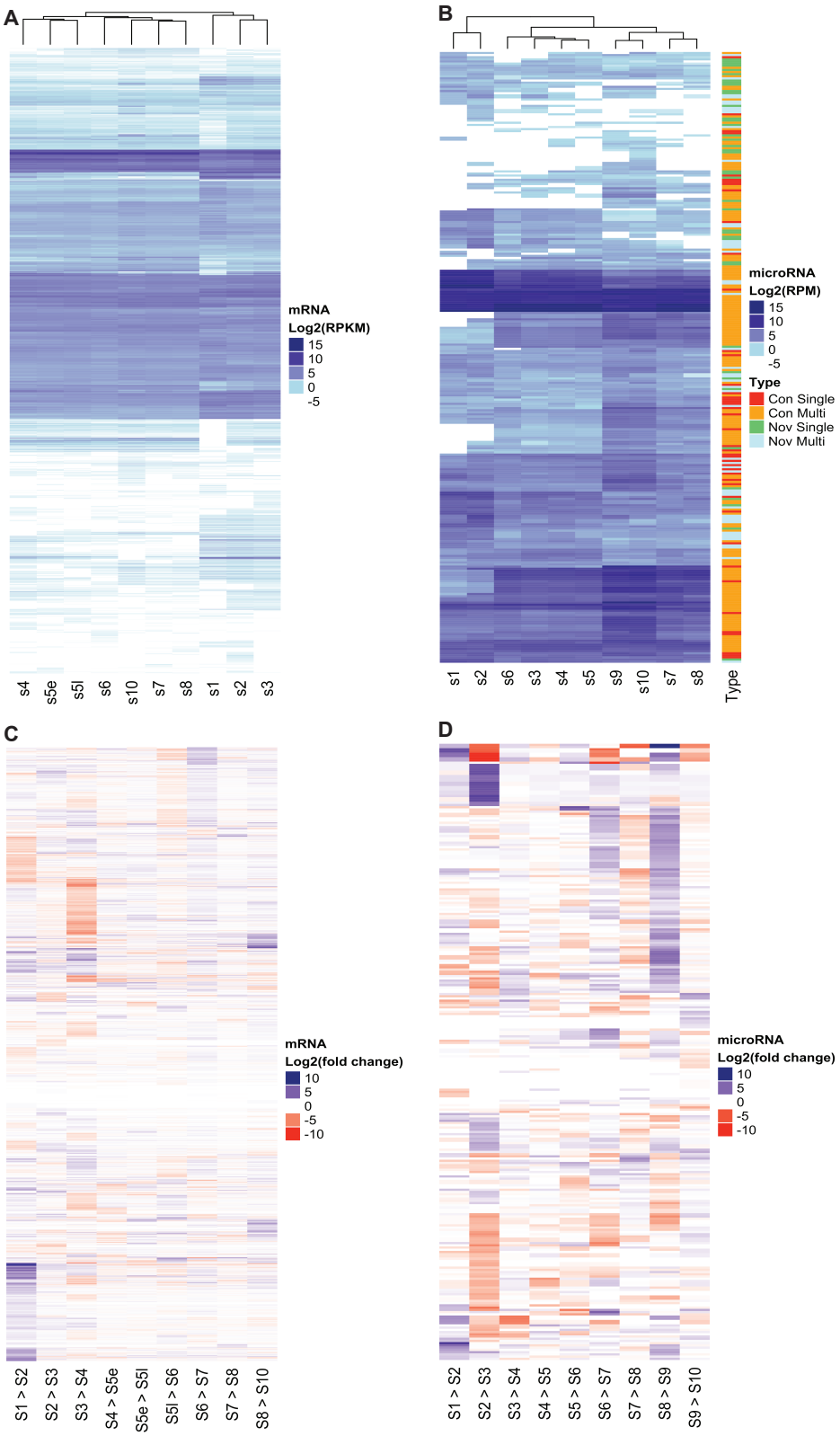


Figure 5.1: MicroRNA and mRNA expression in *P. tepidariorum* development. Heatmaps of mRNAs **(A)** and mature miRNAs **(B)** showing the expression and dendrogram clustering of stages. The miRNA expression clustering groups stages one and two together. The mRNA clustering groups stages one to three together **(B)**. Fold changes from stage to stage for miRNA **(C)** and mRNA **(D)** reveal upregulated miRNAs and down regulated mRNAs between clustered stages. The miRNAs were also annotated in the expression profile **(B)** based on whether they were from conserved/novel families and also whether they were single or multi copy.

5.1.2 Early Transitions in the Embryonic miRNA Transcriptome of *P. tepidariorum*

The miRNA expression profile was then examined for differences in expression that may correspond to the MZT in *P. tepidariorum*. The expression changes between stages were detected with dendrogram clustering (Fig. 5.1B), PC analysis (Fig. 5.2B), and examining changes in abundance of reads assigned to miRNAs from the total small RNA-Seq data (Fig. 5.3D). From the total small RNA sequencing, stages S1 and S2 had similar levels of reads mapping to miRNAs with 3.4 % and 3.9 % respectively (Fig. 5.3D). There appear to be little change between these two stages as they also cluster together in the heatmap dendrogram (Fig. 5.1B) and group close to one another in the first and second PC (Fig. 5.2B). However, the greatest increase in reads that mapped to miRNAs occurred between stages S2 and S3, possibly indicating the onset of zygotic transcription of miRNAs (Fig. 5.3D). At the transition between S2-S3 a particular subset of miRNAs are strongly upregulated, while many other miRNAs are downregulated (Fig. 5.1D). These expression changes between early and later stages are such that the first PC separates them from one another (Fig. 5.2B). Together this suggests that at S3, initiation of zygotic transcription of miRNAs occurs at a similar time as maternal miRNAs are down-regulated.

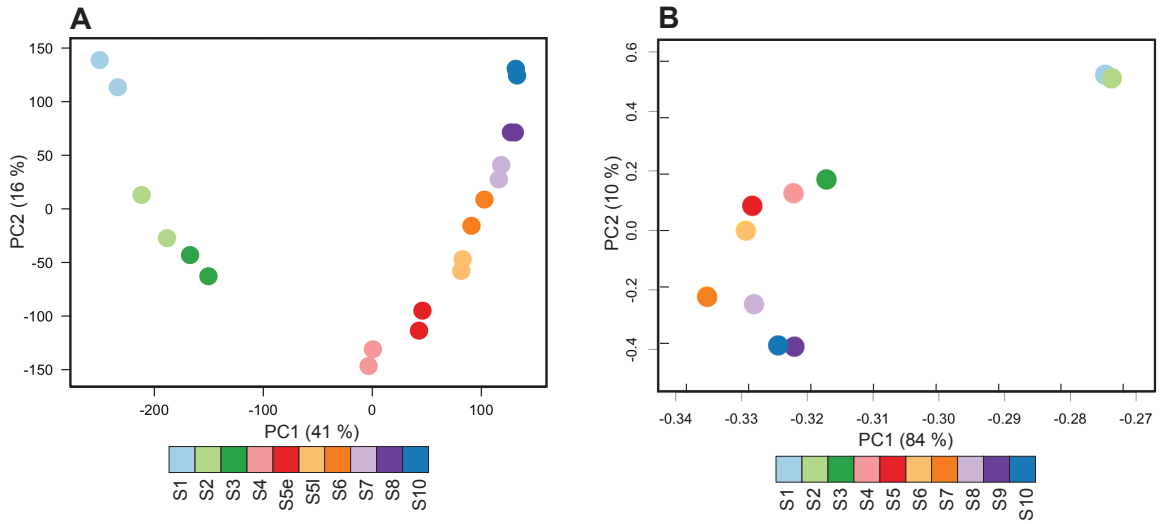


Figure 5.2: Early stages of miRNA and mRNA expression in *P. tepidariorum* separate from later stages. Principal component analysis of mRNAs **(A)** separates the first three stages from later stages by the first principal component. The first stage is also separated from S2 and S3 along the second PC. A somewhat similar result is seen for miRNA data except that the first PC separates only S1 and S2 from later stages **(B)**. For both mRNA and mRNA PCA the second PCs separate later stages by their temporal order.

5.2 Predicting miRNA Targets at Phase II of the MZT in *P. tepidariorum*

The mRNA and miRNA embryonic expression profiles indicate that phase II of the MZT in *P. tepidariorum* involves the greatest up-regulation of miRNAs followed by the greatest decrease in maternal mRNA expression. Therefore, this offers the more obvious window within the MZT to study the role of miRNA regulation of maternally supplied mRNAs.

In order to identify miRNAs and their predicted targets, firstly the miRNAs up-regulated greater than log2 fold at S3 and mRNA down-regulated greater than log2 fold ($p < 0.1$) were used as the differentially expressed gene set representing the phase II transition. Previous miRNA target predictions have utilised correlation analysis to help identify possible targets because negative correlations assume mRNA reduction as targeting miRNA levels increase (Wang and Li 2009). Therefore, the expression levels of phase II up-regulated miRNAs and down-regulated mRNAs were correlated across the embryonic stages. Correlated miRNA-mRNAs with $\leq -0.8 = R$ were retained for target site prediction analysis in the 3' UTRs of mRNAs using miRanda (John *et al.* 2004), which considers both sequence and thermodynamic properties of miRNA-mRNA duplexes. After selection of the differentially expressed genes, assessing the strength of miRNA-mRNA correlations and predicted targeting, eight mature miRNAs were identified, from seven miRNA hairpins that were predicted to target a total of 31 mRNAs (Table 5.2).

Table 5.2: Predicted miRNAs and their mRNA targets in *P. tepidariorum* MZT

***Parasteatoda* AUG3 ID**	**No. CDS isoforms**	**Gene Name**	**Targeting miRNA**	**miRanda Score**	**MFE (kcal/mol)**	**Maternal in species**	**Role**
aug3.g10005	2	*Gbs76A*	miR-184-1-5p	164	-15.78		
aug3.g10040	3	*HMGCR*	miR-184-1-5p	143	-15.45	Porcine	Oocyte competence, cholesterol synthesis (Yuan *et al*, 2011)
aug3.g10585	**1**	***chs***	**miR-3791k-3p**	**147**	**-18.58**	***C. elegans***	**CHS strong downr-reg at 1-cell stage embryo (Stoeckius *et al*, 2014)**
aug3.g12985	1	*MAPKKK7*	miR-3791g-3p	147	-23.32		Other MAPK maternally loaded in other species (Liu *et al*, 2014)
aug3.g1316	1	*Led d 7*	miR-184-1-5p	149	-18.22		
aug3.g13683	1	*CTR1 or SLC31A1*	miR-3791c-5p	145	-17.03		Other SLCs maternally loaded in other species (Liu *et al*, 2014)
aug3.g13758	1	*Lep d 7*	miR-184-1-5p	144	-16.15		
aug3.g13872	1	*MCU*	miR-3791k-3p	140	-21.66	Zebrafish	Bcl-wav and MCU drive gastrula morphogenesis (Prudent *et al*, 2013)
aug3.g15251	2	*DLGAP1*	miR-3791k-3p	141	-17.25	Conserved	Highly conserved maternally supplied (Liu *et al*, 2014)
aug3.g17576	1	*PDIA4*	miR-3791a-5p	145	-17.03	*Haemaphysalis longicornis*	Oocyte development (Liao *et al*, 2008)
aug3.g18528	1	*GLUT1*	miR-184-1-5p	163	-16.55	*Bubalus bubalis*	Cleavage speed (Rajhans *et al*, 2010)
aug3.g23010	1	*BCL2*	miR-3791g-3p	141	-17.74		MEX3 destabilises BCL2 (Donnini et al, 2004); anti-apoptotic factor (Sibl et al, 1997; Johnston et al, 2005)
aug3.g23747	**1**	***CERS-like***	**miR-184-1-5p**	**141**	**-26.60**	***Drosophila***	**Growth and fat body (Bauer *et al*, 2009); Oocyte aging (Kujjo *et al*, 2012)**
aug3.g25381	1	*TSPAN*	miR-3791g-3p	142	-19.56	Zebrafish, conserved	Expressed in animal pole at 1-cell stage (Trikic *et al*, 2011); Highly conserved maternally supplied (Liu *et al*, 2014)
aug3.g2543	1	*n/a*	miR-184-1-5p	148	-15.32		
aug3.g26608	1	*TPI*	miR-3791c-5p	156	-18.11	Conserved	Highly conserved maternally supplied (Liu *et al*, 2014)
			miR-3791c-5p	163	-17.60		
aug3.g26821	2	*CHRNB1*	miR-3791c-5p	147	-15.22		
			miR-3791d-5p	167	-18.05		
			miR-3791f-5p	151	-20.51		
			miR-3791g-5p	167	-20.85		
aug3.g26839	1	*SUCO*	miR-184-1-5p	146	-15.11		Other SUN domain maternally loaded in other species (Liu *et al*, 2014)
aug3.g27626	1	*PCIF1*	miR-184-1-5p	150	-22.59		
aug3.g27831	1	*ORMDL3*	miR-184-1-5p	140	-20.49		
aug3.g2883	3	*CD109*	miR-184-1-5p	150	-18.08		
aug3.g377	4	*ADAMTS*	miR-3791c-5p	154	-16.77	Zebrafish	Maternal and zygotic roles (Brunet *et al*, 2015)
aug3.g4125	1	*GCAT*	miR-3791a-5p	149	-16.74		
aug3.g4152	2	*B4GAT1*	miR-184-1-5p	152	-16.02		
aug3.g5497	1	*RASGRP3*	miR-3791k-3p	149	-18.58		Other RAS family maternally loaded in other species (Liu *et al*, 2014)
aug3.g5633	1	*HSPA5*	miR-184-1-5p	140	-17.15	*Lymnaea stagnalis,* conserved	Highly conserved maternally supplied (Liu *et al*, 2014)
aug3.g6078	3	*LRMP*	miR-184-1-5p	141	-16.70	Zebrafish	Nucleus and centromere attachment (Lindeman and Pelegri, 2012)
aug3.g74	2	*GMPPA*	miR-3791a-5p	165	-15.81		
aug3.g7514	**1**	***mex3***	**miR-184-1-5p**	**140**	**-15.40**	***C. elegans***	**MEX3 destabilises BCL2 (Donnini *et al*, 2004)**
aug3.g8467	1	*GATAD14*	miR-184-1-5p	144	-16.45		
aug3.g9737	1	*irreC*	miR-3791a-5p	148	-16.53		

5.3 Paralogs of *miR-3791* and *mir-184* are Predicted to Regulate MZT

The eight mature miRNAs that were predicted to target maternal mRNAs were mir-184-1-5p and seven mature miRNAs from six paralogs of the mir-3971 family (miR-3791a-5p, miR-3791c-5p, miR-3791d-5p, miR-3791f-5p, miR-3791g-5p, miR-3791g-3p, miR-3791k-3p) (Fig. 5.3A and Table 5.2). Except for three miRNAs (miR-3791d-5p, miR-3791f-5p, miR-3791g-5p), the remainder were predicted to target multiple genes, with miR-184-1-5p targeting the most with 16 targets (Table 5.2). These target predictions are strongly influenced by the seed sequence present in the mature miRNA (John *et al.* 2004; Marco *et al.* 2012b). The two paralogs of *mir-184* both contain the same CUUAUCA seed in the 5p arm (App. Table 4.2). However, the *mir-184-2* paralog does not appear to be strongly up-regulated at S3 (Fig. 5.3B) and therefore would not have been included in the target prediction analysis.

For the *mir-3791* paralogs that had predicted involvement in the MZT, miR-3791f-5p and miR-3791a-5p both had unique seed sequences compared to other *mir-3791* paralogs (Fig. 5.3A and 5.3C). However, the other *mir-3791* paralogs that potentially regulate the MZT contained identical seeds in other *mir-3791* paralogs not predicted to be involved in the MZT. The most widely shared seed (CACCGGG) from miR-3791g-3p and miR-3791k-3p was present in the 3p arms of nineteen other paralogs of the mir-3791 family (mir-3791a/b/c/d/e-1/e-2/f/h/i/j-1/j-2/l/m/n-1/n-2/o-1/o-2/p/s) (Fig. 5.3C). Therefore, the majority of *mir-3791* paralogs may share similar targets for their 3p arm. The CACCGG seed sequence is also shared in the 3p arm of orthologous *mir-3791* miRNAs in *A. mellifera* and *T. castaneum*, as well as the mir-35-42 family in *C. elegans* (Fig. 5.3C). Furthermore, between the mir-3791 and mir-35-42 miRNA families there is also homology in the 5p seed sequence and regions outside of the seed (Fig. 5.3C). It is possible that the mir-3791 family in arthropods is orthologous to the mir-35-42 family in nematodes.

The mir-3791 family is highly expressed throughout *P. tepidariorum* development with ~50% of total miRNA reads per stage mapping to this family (Fig. 5.3D). However, the expression of just the six *mir-3791* paralogs showed a marked increase from S2 to S3, until S8, representing the majority of expression compared to the total mir-3791 family expression (Fig. 5.3D). Previous synteny analysis of the mir-3791 family in Chapter Four showed that the majority of paralogs are found in three clusters in the *P. tepidariorum* DoveTail genome (Fig. 4.3D). Interestingly, the six paralogs from this family that are predicted to have a role in the MZT are all found in one cluster on a single scaffold (Fig. 5.3E). The majority of these are positioned next to each other with only *mir-3791k* located more distally ~40 kb upstream of these. This arrangement and the expression profile of these *mir-3791* paralogs could indicate that they are co-transcribed and/or use the same cis-regulatory elements during the MZT.

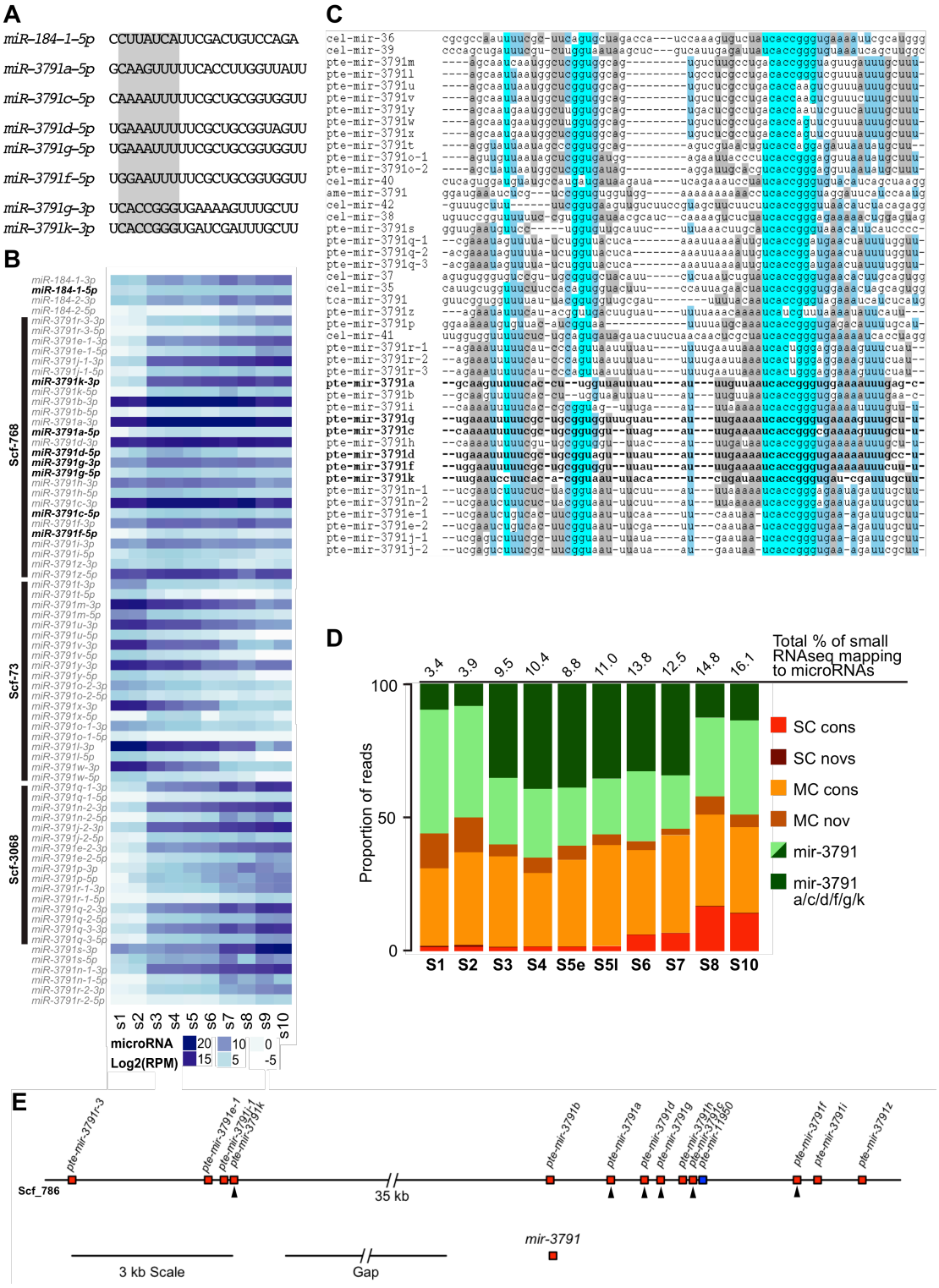


Figure 5.3: *Mir-3791* cluster contains paralogs predicted to regulate MZT. (A) The eight mature miRNAs with the seed sequence highlighted in grey. **(B)** Expression levels of both 5p and 3p arms from the hairpins of all paralogs of the mir-184 and mir-3791 families. The eight mature miRNAs that have predicted MZT roles are in bold. The scaffold IDs are shown for the three large clusters of *mir-3791* paralogs. **(C)** Alignments of all *pte-mir-3791* paralogs show homology to *mir-3791* orthologs from *A. mellifera* and *T. castaneum*, and also the mir-35-42 family from *C. elegans*. **(D)** The mir-3791 family is proportionally highly expressed throughout development. However, the six *mir-3791* paralogs predicted to be involved in the MZT are upregulated at S3 and until S8 constitute the majority of *mir-3791* expression. **(E)** The six *mir-3791* paralogs (arrow heads) all derive from the same cluster found on a single scaffold in the DoveTail assembly of the *P. tepidariorum* genome. SC = single copy; MC = multi copy; cons = conserved miRNA family; novs = novel *P. tepidariorum* miRNA family.

5.4 Predicted Targets that are also Maternally Loaded in Other Species

5.4.1 Many Targets are Maternally Supplied in Other Metazoans

To initially identify possible roles of the 31 targets, gene orthology was based on previous annotations (Schwager *et al.* 2017), except for gene model aug3.g23010, which is a partial gene model and had 40% similarity (p = 0.012) to *BCL2* from *Octopus bimaculoides*. Based on these annotations, there is evidence that 18 genes are maternally loaded into oocytes and/or repressed in the MZT in other species (Table 5.2). Eight of these have previously been reported to be maternally supplied mRNAs that are conserved from a mollusc (*Lymnaea stagnalis*) to humans (Liu *et al.* 2014b), with the remainder involved in the MZT of *Drosophila* (Bauer *et al.* 2009; Kujjo and Perez 2012), zebrafish (Trikic *et al.* 2011; Lindeman and Pelegri 2012; Prudent *et al.* 2013; Brunet *et al.* 2015), the tick *Haemaphysalis longicornis* (Liao *et al.* 2008), *C. elegans* (Donnini *et al.* 2004; Stoeckius *et al.* 2014), *Xenopus* (Sible *et al.* 1997; Johnston *et al.* 2005), the water buffalo *Bubalus bubalis (Rajhans et al. 2010),* and pigs (Yuan *et al.* 2011). Therefore, the majority of genes predicted to be targets of miRNAs at the MZT of *P. tepidariorum* are also maternally supplied in other species.

5.4.2 Predicted Role of *mex-3* in Axis Specification of *P. tepidariorum*

One of the predicted targets of miR-184-1-5p was *muscle excess-3* (*mex-3*), which encodes an RNA-binding protein with two conserved N-terminal KH domains and a C-terminal RING domain (App. Fig. 5.1) (Draper *et al.* 1996). In ascidians, beetles and nematodes, *mex-3* is maternally loaded to the anterior part of the oocytes (Draper *et al.* 1996; Satou 1999; Donnini *et al.* 2004; Schoppmeier *et al.* 2009). The asymmetric expression of *mex-3* has been shown in *T. castaneum* and *C. elegans* to be required for the setup of the AP axis and head development (Draper *et al.* 1996; Schoppmeier *et al.* 2009). In early spider oocytes the expression of *Pt-mex-3* mRNA is not present (Fig. 5.4A). However, as oocytes mature *Pt-mex-3* transcripts are observed in a small patch towards the surface of the egg on one side (Fig. 5.4A). This is likely the location of the nucleus of the oocyte is, as staining at S1 localises to intralecithally dividing nuclei (data not shown).

During the target predict analysis, full length 3' UTRs were used, however it is possible that alternative polyadenylation of transcripts can result in different length 3' UTRs affecting visibility of miRNA targeting (Patraquim *et al.* 2011). Expression coverage of the *Pt-mex-3* locus shows that at S1 exons one and two, which contains the 5' UTR, CDS and proximal 3' UTR, have higher levels of read depth compared to the distal 3' UTR region, which contains the miR-184-1-5p binding site (Fig. 5.4B and 5.4C). It is therefore possible that at S1 not all transcripts possess a full length 3' UTR and therefore cannot be regulated at the predicted miRNA binding site. However, at S2 and S3 the proportion of read depth across all *Pt-mex-3* exons in similar (Fig. 5.4B). This could imply that zygotic transcription activation at S2 produces *Pt-mex-3* transcripts with longer 3' UTRs containing

the predicted miR-184-1-5p binding site. The maternal transcripts with short UTRs may therefore be cleared by other miRNAs that have proximal binding sites or by another mechanism altogether.

In *C. elegans*, the repression of maternal transcripts at the MZT is not thought to be solely due to miRNA repression (Wu *et al.* 2010; Stoeckius *et al.* 2014). RNA binding proteins that recognise polyC motifs in 3' UTRs of mRNAs have also been found to repress maternal transcripts in *C. elegans* (Makeyev and Liebhaber 2002; Stoeckius *et al.* 2014). One gene that was identified to contain a polyC motif in its 3' UTR was *mex-3*. A motif scan of the 31 *P. tepidariorum* genes predicted to be targeted by miRNAs at the MZT also identified two putative polyC motifs in the 3' UTR of *Pt-mex-3* (5.4D). One of these is positioned at the proximal region of the 3' UTR where the read depth is similar to upstream regions, the other at the very end of the 3' UTR (Fig. 5.4B). Expression levels of the two *P. tepidariorum* orthologs of the polyC-binding protein (PCBP) (App. Fig. 5.2) show that one paralog is likely maternally supplied and expression decreases throughout development, the other paralog has a spike in expression around S3 (Fig. 5.4E). Extrapolating from this, maternally supplied *Pt-mex-3* transcripts may be initially regulated by a maternal supplied PCBP paralog where as zygotic *Pt-mex-3* transcripts are perhaps regulated by miRNA targeting as well as a different PCBP paralog that is zygotically expressed.

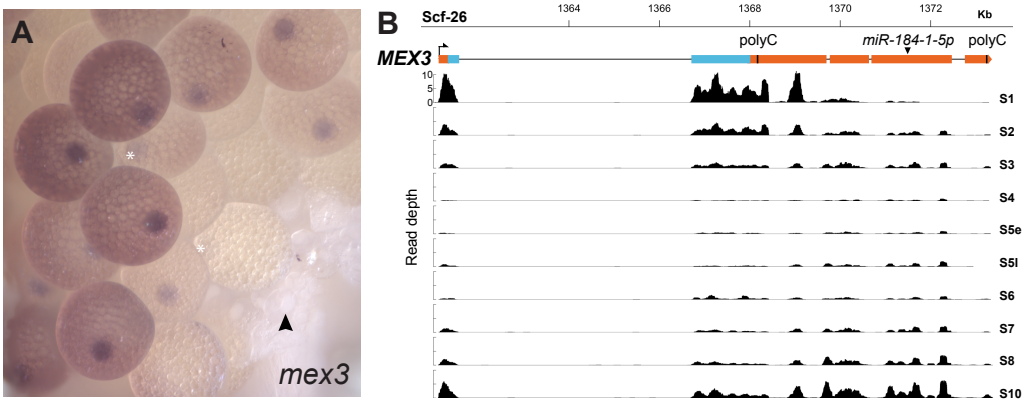


Figure 5.4: Regulation of the predicted axis patterning gene *mex3* in *P. tepidariorum*. (A) Transcripts of *Pt-mex3* mRNA is not visible in immature oocytes (arrow head). As oocytes grow and mature transcripts of *Pt-mex3* are localised to a single small patch towards the surface of the oocyte (white asterisks). **(B)** The *Pt-mex3* locus showing the gene model and the read depth support from exons in the ten RNA-Seq samples. S1 expression appears to have support for a short 3' UTR that would lack the mir-184-1-5p binding site **(C)**. However, at S2, S3 and later expression appears to have coverage throughout all exons, indicating that these transcripts would contain the predicted miR-184-1-5p binding sites. **(D)** There are also putative polyC binding motifs in the *Pt-mex3* 3' UTR. **(E)**The two *P. tepidariorum PCBP* paralogs show one is expressed likely as a maternal factor and the other has a peak in expression at S3.

5.5 Discussion

5.5.1 Timing of MZT Corresponds with Germ Disc Formation and Gastrulation

The maternal to zygotic transition is a feature of early embryogenesis where by initial maternal instructions are cleared and the zygote becomes transcriptionally active (Tadros and Lipshitz 2009; Lee *et al.* 2014). Global analysis of the embryonic miRNAs and mRNAs expression profiles in *P. tepidariorum* identified a molecular signature corresponding to this regulatory transition. In the first stage of *P. tepidariorum* embryogenesis, prior to zygotic transcription, many mRNAs and miRNAs expressed are likely to be primarily maternal supplied. There are reductions in a subset of these transcripts in both S2 and S4, with the greatest number of down regulated genes at S4. The initial reduction of some genes at S2 also coincides with the up-regulation of other mRNAs and therefore marks the onset of zygotic transcription. This associates well with previous findings that inhibition of RNA pol II in early *P. tepidariorum* embryos arrests development at S2, indicating that transcriptional activation is required for embryogenesis to precede (Pechmann 2016). Even though miRNAs also predominantly required RNA pol II for transcription (Lee *et al.* 2004), there appears to be greater fold increases in miRNA expression at S3 compared to S2. This may reflect different regulation of transcription and timing of processing between mRNA and miRNA molecules. However, mis-matches in timing of up-regulation between mRNA and miRNA profiles may also be due to samples deriving from different RNA extractions. Another possibility is that the up-regulation of miRNAs at S3 has a delayed affect on mRNA repression leading to the major down–regulation of transcripts at S4. Therefore, the timing of zygotic genome activation, and the degradation of maternal molecules, occurs in a two phase process in *P. tepidariorum* similar to other animals observed already (Tadros and Lipshitz 2009)

Timing of the miRNA and mRNA transitions corresponds to distinct morphological changes in the organisation of *P. tepidariorum* embryos (Mittmann and Wolff 2012). During S1, nuclei divide intralecithally in a syncytium until there are 16 nuclei, when the process of cellularisation begins (Pechmann 2016). These cells continue to divide and migrate so that by the beginning of S2 there are 64 nuclei at the surface of the embryo (Mittmann and Wolff 2012; Pechmann 2016). mRNA and miRNA repertoires prior to cellularisation and after the cells have reached the embryo surface, indicate the onset of zygotic transcription takes place during this period. Inhibition of RNA pol II with alpha-amanitin, does not appear to stop cellularisation but does prevent germ-disc formation. Therefore the onset of transcription identified here corresponds to previous findings (Pechmann 2016) and occurs when cells have reached the embryo surface.

The transition of from evenly distributed cells at S2 to the formation of the germ disc hemisphere corresponds to the greatest miRNA changes, suggesting they may help to regulate the transition and germ disc formation. After establishment of the germ disc at S3, cells in the central region of the disc aggregate together and form the blastopore (Mittmann and Wolff 2012). The

transition in mRNA profile corresponds then with the internalisation of the cells at the blastopore and the onset of gastrulation at S4 (Mittmann and Wolff 2012). These results suggest that substantial changes in the gene regulatory networks of miRNAs and mRNAs are required for these morphological transitions.

5.5.2 Ecdysozoans Share Common Clustered miRNAs that Regulate the MZT

The expansion of the mir-3791 family and its role predicted role in the MZT might be a common feature of miRNAs that regulate the MZT. In vertebrates and invertebrates, many of the miRNAs involved in maternal transcript clearance are also found clustered together. Homologs *mir-430*/*mir-427* form extensive clusters in both zebrafish and *Xenopus*, respectively (Giraldez *et al.* 2005; Giraldez *et al.* 2006; Lund *et al.* 2009). However, different miRNAs appear to regulate the MZT in insects. In *Drosophila* the *mir-309* cluster (*mir-309*, *mir-3*, *mir-286*, *mir-4*, *mir-5*, and three copies of *mir-6*) is expressed by the zygote to clear maternal transcripts (Bushati *et al.* 2008). This cluster is composed of a number of fast evolving miRNAs, including *mir-3791*, that has also been found to be zygotically transcribed and involved in MZT in both *T. castaneum* and *A. mellifera* (Zondag *et al.* 2012; Ninova *et al.* 2014, 2016). The mir-3791 family present in the *T. castaneum and A. mellifera* clusters is homologous to the *cluster*ed mir-3791 family identified in *P. tepidariorum* and contains the same CACCGGG seed site, which has only one site difference compared to *mir-309* (CACUGGG) in *Drosophila* (Kozomara and Griffiths-Jones 2014; Ninova *et al.* 2016). This CACCGGG seed site is also present in the 3p arm of the *C. elegans mir-35-42* miRNA cluster. Interestingly, this cluster is also involved in maternal mRNA clearance at the MZT by causing deadenylation of targets (Alvarez-Saavedra and Horvitz 2010; Wu *et al.* 2010). Previously it has been suggested that, though divergent, the cluster of miRNAs involved in the fly, beetle and bee MZTs may have derived from a common miRNA orthology within holometabolous insects (Ninova *et al.* 2016). However, these results provide evidence for a more ancient origin for the involvement of this miRNA cluster in the MZT within ecdysozoans. Therefore extensive miRNA duplication and turn over in these families may help in dosage compensation as well as regulating duplicated and newly expressed coding genes that are present in the oocyte (Bartel 2004; Evsikov *et al.* 2006).

5.5.3 Maternal Supplied *mex3* may help to set up Egg Axes in *P. tepidariorum*

The transition from S2 to S3 in *P. tepidariorum* development sees the condensing of cells to form the germ-disc in one hemisphere (Mittmann and Wolff 2012; Pechmann 2016). Activation of zygotic transcription has shown to be required for proper germ-disc formation (Pechmann 2016). However, it is currently unknown in the spider how the transition from spherical to radial symmetry is regulated. One likely potential is that axes could be determined by maternally loaded mRNAs. One of the candidates, mex3, identified in the *P. tepidariorum* MZT also provides axis information and is asymmetrically loaded in other species. At the 1-cell stage of *C. elegans* eggs

mex3 mRNA forms a high-to-low AP concentration so that at the 2-cell stage the anterior blastomere (AB) has a greater level of *mex3* than in P_1 (Draper *et al.* 1996). Subsequently mRNA and protein levels decrease in the AB lineage, but are maintained in the P-granules of the germ-line. Disruption of the distribution of *mex3* in *C. elegans* inhibits its repressive role in determining cell fate (Draper *et al.* 1996). In *T. castaneum*, *mex3* is zygotically expressed in the anterior head region early at the blastoderm stage and is required for head development (Schoppmeier *et al.* 2009). It is therefore suggested that *mex3* possibly represents an ancestral ecdysozoan anterior determinant (Schoppmeier *et al.* 2009). In nematodes, *P. tepidariorum* and the butterfly *Pararge aegeria* (Carter *et al.* 2015) *mex3* is expressed in oocytes, where as in the beetle it is zygotically expressed. It is possible that *mex3* was ancestrally loaded as a maternal component and in the case of *T. castaneum* has become zygotical transcribed. However, based on the transcriptome profiles of *Pt-mex3*, there are likely also zygotic transcripts synthesised prior to its repression and S4 (Fig. 5.4B). To identify how *mex3* functions in *P. tepidariorum* axis formation, further *in situ* hybridisations in later stages will identify if it is asymmetrically distributed. Furthermore, parental RNAi in *P. tepidariorum* would be ideal for knockdown of maternal transcripts to assess their role in axis formation (Akiyama-Oda and Oda 2006).

5.5.4 Other Mechanisms of Maternal mRNA Clearance

From the hundreds of differentially expressed genes at the MZT, there were just a final few tens of candidate genes that were predicted to be targets of miRNAs. While this is likely to be an under estimation of the number of targeting interactions, there are likely other mechanisms that are responsible for RNA clearance. One of the *P. tepidariorum* MZT targets *mex3*, which is a post-transcriptional regulator, is also down regulated in *C. elegans*, though at the oocyte to 1-cell stage (Stoeckius *et al.* 2014). However, Stoeckius *et al*, (2014) identified a polyC motif in the 3' UTR of *mex3* that is recognised by a polyC-binding protein, to be the critical regulator of *mex3* mRNA deadenylation. The presence of both the multiple polyC motifs in *Pt-mex3*, and in other *P. tepidariorum* MZT genes, and two homologs of *PCBPs* that have maternally loaded and dynamic expression at the MZT, indicate that PCBP targeting may contribute to the maternal mRNA clearance alongside miRNAs. This network of miRNAs, *mex3* and *PCBPs* perhaps forms a cascade of negative interactions. Maternally supplied *mex3* may help to promote pluripotency, as well as regulate cell proliferation until stage three, through deadenylation of target transcripts (Jiang *et al.* 2012; Pereira *et al.* 2013; Zhu *et al.* 2015). After which, proliferation reduces (Mittmann and Wolff 2012) as mex3 mRNAs are down regulated, perhaps destabilised by mir-184 and PCBPs, therefore derepressing mex3 targets and allowing cell differentiation as gastrulation occurs (Draper *et al.* 1996).

Furthermore, in the pool of small RNA sequence data there were an uncharacterised pool of reads at 29 bp length, potentially relating to piRNAs (Saito *et al.* 2006). This other class of small non-coding RNAs have been shown to be involved in early development, the specification of cell

fate (Thomson and Lin 2009; Akkouche *et al.* 2017). Future characterisation of these transcripts will help to identify conserved ecdysozoan piRNAs and others that are chelicerate specific.

6 GENERAL DISCUSSION AND FUTURE DIRECTIONS

The expansion of homeobox gene and miRNA repertoires, via duplication events in the genomes of early metazoans, is thought to underlie the evolution of multi-cellularity and further developmental adaptations that generate morphological complexity (Garcia-Fernandez 2005; Heimberg *et al.* 2008; Peterson *et al.* 2009; Wheeler *et al.* 2009; Campo-Paysaa *et al.* 2011; Guerra-Assunção and Enright 2012; Holland 2015; Tarver *et al.* 2015; Hrycaj and Wellik 2016). In the early bilaterian, the homeobox genes have undergone extensive tandem duplications that have formed clusters of homeobox classes (Ferrier 2016). The identification of duplicated homeobox clusters, such as the Hox clusters, in the vertebrate lineage has provided support for WGDs during vertebrate evolution (Hoegg and Meyer 2005). These include two rounds of WGD prior to the gnathostome ancestor, another WGD at the base of teleost fish and again in the salmonid fish (Mungpakdee *et al.* 2008).

These WGDs in vertebrates are hypothesised to contribute to their morphological diversity, but until recently there were no other metazoan lineages to compare this phenomenon to (Van de Peer *et al.* 2009). However, recent genome sequencing of horseshoe crabs, which belong to Chelicerata, the sister subphyla of the other Arthropoda, has identified four Hox clusters, indicating the occurrence of two WGD events in the evolution of these animals (Nossa *et al.* 2014; Kenny *et al.* 2015a). Furthermore, the genomes of spiders and scorpions have also been shown to duplicate Hox clusters (Schwager *et al.* 2017). The prospect of WGDs in a morphological diverse group of animals provides an exciting opportunity to compare the evolution of metazoans after WGD events. However, the systematic characterisation of homeobox gene repertoires had only been carried out in horseshoe crabs and a scorpion (Nossa *et al.* 2014; Di *et al.* 2015; Kenny *et al.* 2015a). Furthermore, the characterisation of miRNAs is limited to chelicerate species not associated with WGDs (Barrero *et al.* 2011; Rota-Stabelli *et al.* 2011; Zhou *et al.* 2013; Liu *et al.* 2014a; Luo *et al.* 2015). Therefore, to gain further perspective on gene duplication in chelicerates I have aimed to expand the characterisation of homeobox genes and miRNAs across arachnid species that have possible experienced WGDs, as well as chelicerates that have not shown any previous evidence of WGDs. Furthermore, I have investigated aspects of regulation and development that may relate to the diversification in function of duplicate genes.

6.1 Duplication of Homeobox Genes in Chelicerates

To characterise the repertoires of homeobox genes in more chelicerates, the transcriptomes of two spiders, a scorpion and a tick were analysed. Detection of homeobox genes via sequence searches for their homeodomains was sufficient to capture repertoires in each species that contained generally similar families to other arthropods (Fig. 3.1). The phylogenetic analysis was able to group the majority of homeobox predictions into previously described homeobox families (App.

Fig. 3.1-3.3). However, the phylogenetic analysis was not used to assess the timing of the duplication event between spiders and scorpions.

Work by Schwager *et al.* (2017) has however studied the timing of duplications and found a signature prior to the spider-scorpion split. However, their attempts to reconcile the phylogenetic relationships of Hox paralogs did not provide consistent support for a WGD in the arachnopulmonate ancestor (Schwager *et al.* 2017). This could be due to the duplication event and divergence between spiders and scorpions occurring at least 430 MYA (Schwager *et al.* 2017). Since then the paralogy relationships that existed in the common ancestor may have degraded so that distinguishing true ohnologs from WGDs has become very difficult (Ohno 1970). Sequencing of genomes and transcriptomes from more arachnid orders may produce further insights into these phylogenetic relationships. Currently there is a lack of information from other Tetrapulmonata groups (Amblypygi, Schizomida and Thelyphonida), which could help confirm if the WGD was common to all Arachnopulmonata lineages or, if however unlikely based on the evidence so far, is independent between spiders and scorpions (Fig. 1.6). Furthermore, sequencing of animals such as pseudoscorpions and solifuges, which are arachnids outside of the arachnopulmonate group, could continue to support that the WGD is restricted to the arachnopulmonates or extends to other arachnids and is even in fact common to events in horseshoe carbs after all (Fig. 1.6).

Nevertheless, the overwhelming proportion of duplicated homeobox genes in several spider and scorpion species indicates that the arachnopulmonate lineage has undergone pervasive duplication (Fig. 3.1). Furthermore, the lack of many duplicated genes in the tick continues to support that the duplication is independent between arachnopulmonates and xiphosurans. Synteny analysis of the homeobox genes identified Hox, NK, PRD, SINE, LIM and TALE clusters in the *P. tepidariorum* genome. These correspond to the clusters proposed to have been dispersed after the fragmentation of the Giga-homeobox cluster in the genome of the bilaterian ancestor (Fig. 3.4)(Ferrier 2016). Of these, the Hox, NK, SINE, LIM and TALE clusters were all duplicated in the *P. tepidariorum* genome (Fig. 3.4). It is parsimonious to suggest a WGD and loss of the duplicate PRD cluster, rather than repeated chromosomal duplications that would duplicate each cluster once, with the exception of the PRD cluster.

The largest and most complete duplicated homeobox cluster identified in *P. tepidariorum* is the Hox cluster (Fig. 3.3A and 3.4). Duplicated Hox clusters were also identified in the scorpion *C. scuplturatus*, but not in the tick (Fig. 3.2). Previous work in the scorpion *M. martensii* also identified duplicated clusters, however other spider genomes (*S. mimosarum* and *A. geniculata*) are currently too fragmented to be able to assemble Hox paralogs into two complete clusters (Di *et al.* 2015; Schwager *et al.* 2017). However, these are likely assembly issues rather than the actual fragmentation of the Hox clusters in these spiders. To gain a better perspective on the evolution of homeobox synteny in chelicerates it will be important to sequence new genomes but with a focus to

improve N50 scaffold sizes. This would improve the power of synteny analysis and could therefore reveal duplicated regions even after chromosomal rearrangements.

RNA-Seq and ISH analysis of duplicate *P. tepidariorum* homeobox genes revealed divergences in their spatio-temporal expression during embryogenesis (Fig. 3.3B-M 3.14). To gain further insights into the divergences of homeobox expression, the RNA-Seq could be extended to contain a complete profile throughout embryogenesis as well as samples at each instar and adult stages (Paps *et al.* 2015). Furthermore, if this were also undertaken in several spider species, as well as non-WGD lineages like Opiliones, it would help to identify the patterns of divergence between the expressions of paralogs. Previous work in molluscs identified that novel homeobox genes were more likely to have their peak in expression either very earl or very late in the life cycle (Paps *et al.* 2015). It would be interesting to identify if the retained duplicates in spiders show a bias to early and late expression (Paps *et al.* 2015).

The spatio-temporal expression differences between paralogs suggest that their *cis*-regulatory regions have diverged (Fig. 3.3B-M 3.14). ATAC-Seq (assay for transposase-accessible chromatin using sequencing) could be used to identify open chromatin states around paralogs that may represent divergent enhancers (Buenrostro *et al.* 2013). If the powerful gene editing tool of CRISPR/cas9 (Doetschman and Georgieva 2017) can be established in *P. tepidariorum*, then these enhancer fragments could be cloned upstream of a reporter GFP and inserted into the *P. tepidariorum* genome at S1 to test if the fragments are sufficient to drive the expression characteristics. Furthermore, in parallel ATAC-Seq in Opiliones could identify enhancers of the single copy orthologs. These regions could be used to replace endogenous enhancers in duplicated orthologs in spider genomes to try and recapitulate the expression of the single copy ortholog in a WGD lineage. Additionally, the enhancer information for clustered homeobox genes could be used for chromosome conformation capture techniques, such as 4C, to assess the chromatin architectural differences between duplicate clusters (Acemel *et al.* 2016). An attractive candidate to study would be the Hox cluster both in WGD lineages and non-WGD lineages of chelicerates. Evolutionary differences in topologically associated domains of duplicate Hox clusters in zebrafish have been linked to changes in the content of neighbouring genes to the Hox clusters (Acemel *et al.* 2016). In comparison to amphioxus these are thought to relate to structural changes to the Hox clusters that represent changes to the animal body plan (Acemel *et al.* 2016). Differences in non-Hox related genes in the Hox clusters of *P. tepidariorum* were observed (Fig. 3.3A). Furthermore, there were both differences in the number of non-Hox related genes between chelicerate species, and also the sizes of the Hox clusters varies between chelicerates (this thesis and Pace *et al.* 2016). These experiments would help to identify the chromatin and *cis*-regulatory changes that have caused the expression pattern divergences between paralogs after a WGD.

There were also changes to the homeodomains that may result in the recognition of different DNA motifs and therefore the regulation of different downstream targets (Ortiz-

Lombardia *et al.* 2017). Chromatin immunoprecipitation sequencing (ChIP-Seq) could be used to identify these targets of endogenous paralogous homeobox genes. Rather than relying on developing antibodies that are specific enough to differentiate between paralogs, CRISPR could be used to tag endogenous genes with a V5 epitope that can be used to capturing in ChIP-Seq (Mazzoni *et al.* 2011; Shlyueva *et al.* 2016). Using this method, ChIP-Seq could again also be carried out in the harvestman to identify the targets of orthologous homeobox genes, and how the duplicated orthologs in *P. tepidariorum* have diverged from the single copy in the harvestman. Situations where the profiles of binding sites in the harvestman are shared between the duplicate orthologs in *P. tepidariorum*, may represent a scenarios of sub-functionalisation. Neo-functionalisation may occur where one paralog in *P. tepidariorum* has a binding site(s) in a target gene that neither the other paralog nor the harvestman ortholog share. However, it is possible that these harvestman profiles have evolved and do not represent true ancestral binding patterns. Therefore, increasing sampling of species from different taxa could more robustly predict true neo-functionalisations relevant to specific lineages.

6.2 Duplication of miRNAs in Chelicerates

The sequencing of small RNA and mapping to the *P. tepidariorum* genome successfully identified conserved metazoan miRNA families but also families specific to spider and *P. tepidariorum* lineages (Leite *et al.* 2016). Many of the conserved families identified were duplicated in *P. tepidariorum* as well as other spiders, a scorpion and the horseshoe crab *L. polyphemus* (Fig. 4.2). There were generally twice the number of paralogs per family in *L. polyphemus* compared to arachnopulmonates, which fits with two WGDs in the Xiphosura lineage and only one WGD proposed in the Arachnopulmonata lineage (Nossa *et al.* 2014; Kenny *et al.* 2015a; Schwager *et al.* 2017).

The synteny analysis of miRNAs in the genomes of chelicerates identified two different miRNA clusters that were common to spiders, a scorpion and *L. polyphemus*. The two clusters (*mir-71*/*mir-2* and *mir-100*/*let-7*/*mir-125*) were also duplicated in these species, providing further support that a large-scale duplication has occurred in the evolution of these chelicerates (Fig. 4.3 and 4.4). The *mir-100*/*let-7*/*mir-125* cluster consists of the ancient eumetazoan *mir-100* and *mir-125* miRNAs, and *let-7* that was acquired in the bilaterian ancestor (Wheeler *et al.* 2009; Tarver *et al.* 2013). Roles for miRNAs of this cluster have been identified in developmental timing through the regulation of ecdysteroid signalling (Reinhart *et al.* 2000; Belles 2017). Ecdysone and juvenile hormone (JH) are also important regulators of developmental morphology (Mirth and Shingleton 2014; Mendes and Mirth 2016). Altering the levels of 20E and JH affects the expression levels of miRNAs in the *mir-100*/*let-7*/*mir-125* cluster (Rubio *et al.* 2012). However, components of JH and ecdysone signalling, which are present in mandibulates, are not found in chelicerates (Fig. 1.6) (Leite and McGregor 2016). Therefore, if the components of the signalling pathways differ, the

regulatory role of the chelicerate *mir-100/let-7/mir-125* cluster may also differ. Furthermore, it will be interesting to identify if the duplication and retention patterns of *mir-100/let-7/mir-125* clusters has had an affect on developmental timing and morphology in chelicerates that have undergone WGDs.

The approach to detect orthologs of *P. tepidariorum* miRNAs in other chelicerates knowingly failed to capture miRNA families that would be novel to other species lineages. Therefore, additional small RNA-Seq in other species that have sequenced genomes, such as *S. mimosarum*, *C. sculpturatus* and *L. polyphemus* is needed to provide more reliable annotations of conserved miRNAs as well as identifying novel miRNA families in these lineages. This data would help to uncover the *de novo* miRNA formations that have been retained within a lineage and others that are younger newly evolved miRNAs that may have more transient adaptive roles (Lyu *et al.* 2014).

An aspect of miRNA processing that can evolve is the whether the 5p or 3p stand is dominantly loaded into the RISC. Loading of alternate strands, either between species or paralogous miRNAs is evidence that these miRNAs have arm-switched during evolution (Griffiths-Jones *et al.* 2011). Several cases of arm switching between species and paralogs were detected in *P. tepidariorum* (Fig. 4.5). However, there was no functional analysis to determine if this might affect miRNA targeting in the spider. To address the arm switching between species, a miRNA that arm-switches, such as *mir-10*, could be cloned from the spider and used to replace the endogenous *mir-10* ortholog in *Drosophila*. This would help reveal if the spider *mir-10* sequence was sufficient to instruct strand selection, or if *P. tepidariorum* miRNA biogenic machinery is required for spider specific processing.

6.3 MicroRNAs in the MZT of *P. tepidariorum*

The maternal to zygotic transition (MZT) is an important phase in early development that requires the clearing of maternal transcripts (Tadros and Lipshitz 2009). MicroRNAs have been found in both vertebrate and invertebrates to be one of the components that clear maternal transcripts during the MZT (Giraldez *et al.* 2005; Giraldez *et al.* 2006; Bushati *et al.* 2008; Lund *et al.* 2009; Tadros and Lipshitz 2009; Giraldez 2010). Interestingly the miRNAs that have been found in these animals to regulate the MZT all belong to miRNA clusters. Similarly in *P. tepidariorum* the majority of miRNAs predicted to target maternal transcripts came from the mir-3791 family. This family is the largest in *P. tepidariorum* with 34 copies, with the majority of paralogs forming three clusters in the spider genome (Fig. 5.3). The addition of miRNAs in the genome via duplications and *de novo* formations are enriched around clusters of miRNAs (Marco *et al.* 2013; Lyu *et al.* 2014). The extensive tandem duplications of the mir-3791 family may represent an example of this phenomenon in *P. tepidariorum*.

In insects and nematodes the mir-3791 family (which is orthologous to the mir-35-42 family in nematodes) and neighbouring clustered miRNAs are also involved in the MZT (Bushati *et al.* 2008; Zondag *et al.* 2012; Ninova *et al.* 2014). However, in *P. tepidariorum* not all of the *mir-3791* clusters appear to regulate the MZT (Fig. 5.3). The paralogs of *mir-3791* that are predicted to regulate the MZT all derive from one particular cluster (Fig. 5.3). Other *mir-3791* clusters in *P. tepidariorum* displayed increases in expression later in embryogenesis rather than early expression related to the MZT. Therefore it is possible that the other *mir-3791* clusters in *P. tepidariorum* could have diverged from their regulatory role in the MZT. The duplication of the mir-3791 family and its predicted role in the MZT provides an avenue to study the role of miRNAs in an important development process in multicellular organisms.

A crucial aspect to deciphering the regulation of maternal mRNAs by the *mir-3791* candidates is to identify their co-localisation. Double ISH detection of the *mir-3791* paralogs expressed early development and maternal expressed target genes such as *Pt-mex-3* will help to ascertain this information. The *Pt-mex-3* ISH identified a possible asymmetry early in the oocyte prior to fertilisation. The signal of *Pt-mex-3* appeared to localise to the nucleus, which was positioned at the periphery of the oocyte. However, prior reports of initial nucleus position and divisions in *P. tepidariorum* during S1 describe them happening intralecithally at the centre of the embryo, not at the periphery (Mittmann and Wolff 2012; Pechmann 2016). In *Drosophila* the anterior-posterior axis is predominantly set up by maternal factors, where as asymmetric positioning of the nucleus determines the dorsal-ventral axis (van Eeden and St Johnston 1999; Grimm and Wieschaus 2010). Therefore, the localisation of *Pt-mex-3* and the nucleus to the periphery of oocytes prior to fertilisation may be predetermining the axes of *P. tepidariorum* embryos. To further characterise these investigations, ISHs in later stages need to be done to determine if the asymmetry is continued throughout development. If the miRNA and mRNA expression patterns overlap then it is possible for the miRNA to repress the transcript.

To validate the miRNA targets both cell culture and *in vivo* reported assays can be used. Luciferase assays can validate targeting in cell cultures using the endogenous 3' UTR of *mex-3* cloned downstream of the reporter *Renilla reniformis* luciferase (Hui *et al.* 2013). This is constitutively expressed along with a control firefly luciferase that is use for normalisation. Co-transfection of this plasmid with another plasmid expressing the miRNA hairpin sequence into cells should result in the reduction of *renilla* relative to firefly; compared to transfecting cells with just the *mex-3* 3' UTR reporter plasmid. The predicted target site could then be removed from the 3' UTR to inhibit miRNA repression, therefore increasing the *renilla* to firefly ratio.

Furthermore, to test these interactions *in vivo* two capped mRNA constructs (Kanayama *et al.* 2011; Schönauer *et al.* 2016; Pechmann *et al.* 2017), one GFP with an endogenous *mex-3* 3' UTR and the other RFP with the miRNA target site removed from the *mex-3* 3' UTR, could be co-injected into S1 embryos. If the 3' UTRs were bound by the endogenous miRNAs at the target site

then the GFP signal would be expected to reduce at S2 onwards, relative to the RFP. To ensure that the different fluorochromes would not affect the interpretation of results, either two different fluorochromes with similar stability properties could be used or the fluorochromes could be swapped between the constructs.

Other insights will also come from disrupting the GRNs in development. Parental RNAi (pRNAi) against *mex-3* would help to identify if it has a role in axis development. Furthermore, pRNAi against the *PCBP* paralogs that are predicted to regulate and repress *mex-3* may result in the continued expression of *mex-3*, which may affect developmental timing. Injecting of *mir-3791* constructs, to over-express mature *mir-3791* miRNAs in the early embryo may disrupt the role of maternally loaded genes and therefore arrest development. It may also clear maternal repressors of zygotic genome activation and therefore initiate transcription earlier than in wild type embryos.

6.4 Conclusion

There is now compelling evidence from both coding and non-coding genes that the lineages of Xiphosura and Arachnopulmonata, in Chelicerata have independently undergone WGDs. The retention of duplicate clusters of genes, such as the Hox clusters, represents hallmarks of WGDs similar to that seen independently in vertebrates. Expression patterns of paralogous homeobox genes have diverged and may relate to sub-functionalisation and/or neo-functionalisation. The duplication of certain miRNAs may have evolved functions in the MZT and general aspects of developmental timing. Therefore, the WGDs in chelicerates provides a invertebrate comparison to the vertebrate WGDs. Studies in these organisms will shed light on the role that gene duplication has on the diversification of GRNs in development and the evolution of morphological diversity.

References

Abrouk M, Zhang R, Murat F, Li A, Pont C, Mao L, Salse J. 2012. Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. *Plant Cell*. 24(5): 1779-92

Abzhanov A, Kaufman TC. 1999. Homeotic genes and the arthropod head: Expression patterns of the *labial*, *proboscipedia*, and *Deformed* genes in crustaceans and insect. *PNAS*. 96: 10224-10229

Acemel RD, Tena JJ, Irastorza-Azcarate I, Marletaz F, Gomez-Marin C, de la Calle-Mustienes E, Bertrand S, Diaz SG, Aldea D, Aury JM, et al. 2016. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat Genet*. 48(3): 336-41

Akiyama-Oda Y, Oda H. 2003. Early patterning of the spider embryo: a cluster of mesenchymal cells at the cumulus produces *Dpp* signals received by germ disc epithelial cells. *Development*. 130(9): 1735-1747

Akiyama-Oda Y, Oda H. 2006. Axis specification in the spider embryo: *dpp* is required for radial-to-axial symmetry transformation and *sog* for ventral patterning. *Development*. 133: 2347-2357

Akkouche A, Mugat B, Barckmann B, Varela-Chavez C, Li B, Raffel R, Pelisson A, Chambeyron S. 2017. Piwi Is required during *Drosophila* embryogenesis to license dual-strand piRNA clusters for transposon repression in adult ovaries. *Mol Cell*. 66(3): 411-419 e4

Alberti C, Cochella L. 2017. A framework for understanding the roles of miRNAs in animal development. *Development*. 144(14): 2548-2559

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215: 403-410

Alvarez-Saavedra E, Horvitz HR. 2010. Many families of *C. elegans* microRNAs are not essential for development or viability. *Curr Biol*. 20(4): 367-73

Ambros V. 2011. MicroRNAs and developmental timing. *Curr Opin Genet Dev*. 21(4): 511-517

Amoutzias G, Van de Peer Y 2010. Single-gene and whole-genome duplications and the evolution of protein-protein interaction networks. *In:* G, C.-A. (ed.) *Evolutionary genomics and systems biology*. UK: Wiley-Blackwell.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31(2): 166-169

Arif S, Kittelmann S, McGregor AP. 2015. From *shavenbaby* to the naked valley: trichome formation as a model for evolutionary developmental biology. *Evol Dev*. 17(1): 120-6

Arif S, Murat S, Almudi I, Nunes MD, Bortolamiol-Becet D, McGregor NS, Currie JM, Hughes H, Ronshaugen M, Sucena É, et al. 2013. Evolution of *mir-92a* underlies natural morphological variation in *Drosophila melanogaster*. *Current Biology*. 23(6): 523-8

Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, Kuntner M, Higgins L, Hayashi CY, Agnarsson I, et al. 2017. The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. *Nat Genet*. 49(6): 895-903

Babraham Bioinformatics. 2011. *FASTQC: A Quality Control Tool for High Throughput Sequence Data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bailey TL, Bodén M, Whitington T, Machanick P. 2010. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*. 11(179)

Barnett AA, Thomas RH. 2013. Posterior Hox gene reduction in an arthropod: *Ultrabithorax* and *Abdominal-B* are expressed in a single segment in the mite *Archegozetes longisetosus*. *Evodevo*. 4(23): 1-12

Baroux C, Autran D, Gillmor CS, Grimanelli D, Grossniklaus U. 2008. The maternal to zygotic transition in animals and plants. *Cold Spring Harb Symp Quant Biol*. LXXIII

Barrero RA, Keeble-Gagnere G, Zhang B, Moolhuijzen P, Ikeo K, Tateno Y, Gojobori T, Guerrero FD, Lew-Tabor A, Bellgard M. 2011. Evolutionary conserved microRNAs are ubiquitously expressed compared to tick-specific miRNAs in the cattle tick *Rhipicephalus* (*Boophilus*) *microplus*. *BMC Genomics*. 12: 328

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 116(2): 281-297

Bataille L, Frendo JL, Vincent A. 2015. Hox control of *Drosophila* larval anatomy; The alary and thoracic alary-related muscles. *Mech Dev*. 138 Pt 2: 170-6

Bauer R, Voelzmann A, Breiden B, Schepers U, Farwanah H, Hahn I, Eckardt F, Sandhoff K, Hoch M. 2009. *Schlank*, a member of the ceramide synthase family controls growth and body fat in *Drosophila*. *EMBO J*. 28(23): 3706-16

Belles X. 2017. MicroRNAs and the evolution of insect metamorphosis. *Annu Rev Entomol*. 62: 111-125

Berleth T, Burri M, Thoma G, Bopp D, Richstein S, Frigerio G, Noll M, Nüssiein-Volhard C. 1988. The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J*. 7(6): 1749-1756

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. 5: 3657

Bhan A, Mandal SS. 2015. LncRNA *HOTAIR*: A master regulator of chromatin dynamics and cancer. *Biochim Biophys Acta*. 1856(1): 151-64

Bianchi M, Renzini A, Adamo S, Moresi V. 2017. Coordinated actions of microRNAs with other epigenetic factors regulate skeletal muscle development and adaptation. *Int J Mol Sci*. 18(4)

Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. 2014. A transcriptome approach to ecdysozoan phylogeny. *Mol Phylogenet Evol*. 80: 79-87

Bråte J, Neumann RS, Fromm B, Haraldsen AAB, Grini PE, Shalchian-Tabrizi K. 2016. Pre-metazoan origin of animal miRNAs. *bioRxiv*. DOI: 10.1101/076190

Brunet FG, Fraser FW, Binder MJ, Smith AD, Kintakas C, Dancevic CM, Ward AC, McCulloch DR. 2015. The evolutionary conservation of the A Disintegrin-like and Metalloproteinase domain with Thrombospondin-1 motif metzincins across vertebrate species and their expression in teleost zebrafish. *BMC Evol Biol*. 15: 22

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 10(12): 1213-8

Buffry AD, Mendes CC, McGregor AP. 2016. The functionality and evolution of eukaryotic transcriptional enhancers. *Adv Genet*. 96: 143-206

Burglin TR, Affolter M. 2016. Homeodomain proteins: an update. *Chromosoma*. 125(3): 497-521

Burke AC, Nelson CE, Morgan BA, Tabin CJ. 1995. Hox genes and the evolution of vertebrate axial morphology. *Development*. 121: 333-346

Bushati N, Stark A, Brennecke J, Cohen SM. 2008. Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in *Drosophila*. *Curr Biol*. 18(7): 501-6

Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc Natl Acad Sci U S A*. 108(38): 15920-15924

Campo-Paysaa F, Semon M, Cameron RA, Peterson KJ, Schubert M. 2011. microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol Dev*. 13(1): 15-27

Cao Z, Yu Y, Wu Y, Hao P, Di Z, He Y, Chen Z, Yang W, Shen Z, He X, et al. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. *Nat Commun*. 4: 2602

Carroll SB, Grenier JK, Weatherbee SD 2005. *From DNA to Diversity,* Oxford, Blackwell Publishing.

Carter JM, Gibbs M, Breuker CJ. 2015. Divergent RNA localisation patterns of maternal genes regulating embryonic patterning in the butterfly *Pararge aegeria*. *PLoS One*. 10(12): e0144471

Chen CY, Chen ST, Juan HF, Huang HC. 2012. Lengthening of 3'UTR increases with morphological complexity in animal evolution. *Bioinformatics*. 28(24): 3178-3781

Chen YW, Song S, Weng R, Verma P, Kugler JM, Buescher M, Rouam S, Cohen SM. 2014. Systematic study of *Drosophila* microRNA functions using a collection of targeted knockout mutations. *Dev Cell*. 31(6): 784-800

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R. 2005. *TRBP* recruits the Dicer complex to *Ago2* for microRNA processing and gene silencing. *Nature*. 436(7051): 740-744

Chipman AD, Ferrier DE, Brena C, Qu J, Hughes DS, Schroder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol*. 12(11): e1002005

Clark E, Akam M. 2016. *Odd-paired* controls frequency doubling in *Drosophila* segmentation by altering the pair-rule gene regulatory network. *Elife*. 5

Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. 2015. Spider transcriptomes identify ancient large-scale gene duplication event potentially important in silk gland evolution. *Genome Biol Evol*. 7(7): 1856-70

Clarke TH, Garb JE, Hayashi CY, Haney RA, Lancaster AK, Corbett S, Ayoub NA. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the sil gland gene toolkit. *BMC Genomics*. 365(15): 1-17

Clements J, Hens K, Merugu S, Dichtl B, de Couet HG, Callaerts P. 2009. Mutational analysis of the *eyeless* gene and phenotypic rescue reveal that an intact Eyeless protein is necessary for normal eye and brain development in *Drosophila*. *Dev Biol*. 334(2): 503-12

Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M. 2001. Hox genes and the phylogeny of the arthropods. *Curr Biol*. 11(10): 759 - 763

Cornman SR, Schatz MC, Johnston SJ, Chen YP, Pettis J. 2010. Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee *Apis mellifera*. *BMC Genomics*. 11: 602

Damen WG, Hausdorf M, Seyfarth EA, Tautz D. 1998. A conserved mode of head segmentation in arthropods revealed by the expression pattern of Hox genes in a spider. *Proc Natl Acad Sci U S A*. 95(18): 10665–10670

Damen WG, Tautz D. 1999. *Abdominal-B* expression in a spider suggests a general role for *Abdominal-B* in specifying the genital structure. *J Exp Zool B Mol Dev Evol*. 258: 85-91

Davis TL, Rebay I. 2017. Master regulators in development: Views from the *Drosophila* retinal determination and mammalian pluripotency gene networks. *Dev Biol*. 421(2): 93-107

de Wit E, Linsen SE, Cuppen E, Berezikov E. 2009. Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res*. 19(11): 2064-74

Di Z, Yu Y, Wu Y, Hao P, He Y, Zhao H, Li Y, Zhao G, Li X, Li W, et al. 2015. Genome-wide analysis of homeobox genes from *Mesobuthus martensii* reveals Hox gene duplication in scorpions. *Insect Biochem Mol Biol*. 61: 25-33

Doetschman T, Georgieva T. 2017. Gene editing with CRISPR/Cas9 RNA-directed nuclease. *Circ Res*. 120(5): 876-894

Donnini M, Lapucci A, Papucci L, Witort E, Jacquier A, Brewer G, Nicolin A, Capaccioli S, Schiavone N. 2004. Identification of *TINO*: a new evolutionarily conserved *BCL-2* AU-rich element RNA-binding protein. *J Biol Chem*. 279(19): 20154-66

Draper BW, Mello CC, Bowerman B, Hardin J, Priess JR. 1996. *MEX-3* is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos. *Cell*. 87: 205-216

Driever W, Nüssiein-Volhard. 1988. A gradient of bicoid protein in *Drosophila* embryos. *Cell*. 54: 83-93

Du H, Taylor HS. 2015. The role of Hox genes in female reproductive tract development, adult function, and fertility. *Cold Spring Harb Perspect Med*. 6(1): a023002

Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, Hsu SH, Ghoshal K, Villén J, Bartel DP. 2014. mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell*. 56(1): 104-115

Erwin DH. 2009. Early origin of the bilaterian developmental toolkit. *Philos Trans R Soc Lond B Biol Sci*. 364(1527): 2253-61

Evsikov AV, Graber JH, Brockman JM, Hampl A, Holbrook AE, Singh P, Eppig JJ, Solter D, Knowles BB. 2006. Cracking the egg: molecular dynamics and evolutionary aspects of the transition from the fully grown oocyte to embryo. *Genes Dev*. 20(19): 2713-27

Farley EK, Olson KM, Levine MS. 2015. Regulatory principles governing tissue specificity of developmental enhancers. *Cold Spring Harb Symp Quant Biol*. 80: 27-32

Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc Natl Acad Sci U S A*. 113(23): 6508-13

Ferguson L, Marletaz F, Carter JM, Taylor WR, Gibbs M, Breuker CJ, Holland PW. 2014. Ancient expansion of the Hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS Genet*. 10(10): e1004698

Ferrier DEK. 2016. Evolution of homeobox gene clusters in animals: The Giga-cluster and primary vs. secondary clustering. *Frontiers in Ecology and Evolution*. 4

Field DJ, Gauthier JA, King BL, Pisani D, Lyson TR, Peterson KJ. 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. *Evol Dev*. 16(4): 189-96

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 44(D1): D279-85

Frankel N, Wang S, Stern DL. 2012. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *PNAS*. 109(51): 20975-20979

Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering miRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*. 26: 407-415

Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, et al. 2015. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet*. 49: 213-42

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(12): 2619-2628

Fukaya T, Iwakawa HO, Tomari Y. 2014. MicroRNAs block assembly of eIF4F translation initiation complex in *Drosophila*. *Mol Cell*. 56(1): 67-78

Gaiti F, Calcino AD, Tanurdzic M, Degnan BM. 2017. Origin and evolution of the metazoan non-coding regulatory genome. *Dev Biol*. 427(2): 193-202

Garcia-Fernandez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet*. 6(12): 881-92

Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, Hedin M, Kocot KM, Ledford JM, Bond JE. 2016. Spider phylogenomics: untangling the spider tree of life. *PeerJ*. 4: e1719

Giraldez AJ. 2010. microRNAs, the cell's Nepenthe: clearing the past during the maternal-to-zygotic transition and cellular reprogramming. *Curr Opin Genet Dev*. 20(4): 369-75

Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, Schier AF. 2005. MicroRNAs regulate brain morphogenesis in zebrafish. *Science*. 308(5723): 833-8

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. 2006. Zebrafish *miR-430* promotes deadenylation and clearance of maternal mRNAs. *Science*. 312(5770): 75-9

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature*. 433: 481-487

Gong J, Wu Y, Zhang X, Liao Y, Sibanda VL, Liu W, Guo AY. 2014. Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing. *RNA Biology*. 11(11): 1375-1385

Gonzalez I, Munita R, Agirre E, Dittmer TA, Gysling K, Misteli T, Luco RF. 2015. A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. *Nat Struct Mol Biol*. 22(5): 370-6

Goodrich JA, Kugel JF. 2006. Non-coding-RNA regulators of RNA polymerase II transcription. *Nature*. 7: 612-616

Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, Grbic V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature*. 479(7374): 487-92

Griffiths-Jones S. 2005. RALEE--RNA ALignment editor in Emacs. *Bioinformatics*. 211: 257-259

Griffiths-Jones S, Hui JH, Marco A, Ronshaugen M. 2011. MicroRNA evolution by arm switching. *EMBO Rep*. 12(2): 172-7

Grimm O, Wieschaus E. 2010. The *Bicoid* gradient is shaped independently of nuclei. *Development*. 137(17): 2857-62

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*. 455(7217): 1193-7

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. *Nucleic Acids Res*. 36: W70–W74

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data *Bioinformatics*. 32(18): 2847-2849

Guarner A, Manjon C, Edwards K, Steller H, Suzanne M, Sanchez-Herrero E. 2014. The *zinc finger homeodomain-2* gene of *Drosophila* controls *Notch* targets and regulates apoptosis in the tarsal segments. *Dev Biol*. 385(2): 350-65

Guennewig B, Roos M, Dogar AM, Gebert LF, Zagalak JA, Vongrad V, Metzner KJ, Hall J. 2014. Synthetic pre-miRNAs reveal dual-strand activity of *miR-34*a on *TNF-α*. *RNA*. 20(1): 61-75

Guerra-Assunção JA, Enright AJ. 2012. Large-scale analysis of microRNA evolution. *BMC Genomics*. 13(218): 1-12

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3): 307-21

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8(8): 1494-512

Halder G, Callaerts P, Gehring WJ. 1995. Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science*. 267: 1788-1792

Halfon MS. 2017. Perspectives on gene regulatory network evolution. *Trends Genet*. 33(7): 436-447

Haney RA, Ayoub NA, Clarke TH, Hayashi CY, Garb JE. 2014. Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. *BMC Genomics*. 366(15): 1-18

Haney RA, Clarke TH, Gadgil R, Fitzpatrick R, Hayashi CY, Ayoub NA, Garb JE. 2016. Effects of gene duplication, positive selection and shifts in gene expression on the evolution of the venom gland transcriptome in widow spiders. *Genome Biol Evol*.

Hatfield SD, Shcherbata HR, Fischer KA, Nakahara K, Carthew RW, Ruohola-Baker H. 2005. Stem cell division is regulated by the microRNA pathway. *Nature*. 435(7044): 974-8

Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ. 2008. MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA*. 105(2946-2950)

Helm C, Bernhart SH, Siederdissen CHz, Nickel B, Bleidorn C. 2012. Deep sequencing of small RNAs confirms an annelid affinity of Myzostomida. *Molecular Phylogenetics and Evolution*. 64(1): 198-203

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, Students of Bioinformatics Computer Labs a. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics*. 7: 25

Hilbrant M, Damen WG, McGregor AP. 2012. Evolutionary crossroads in developmental biology: the spider *Parasteatoda tepidariorum*. *Development*. 139(15): 2655-62

Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet*. 21(8): 421-4

Holland PW. 2015. Did homeobox gene duplications contribute to the Cambrian explosion? *Zoological Lett*. 1: 1

Holland PW, Booth HA, Bruford EA. 2007. Classification and nomenclature of all human homeobox genes. *BMC Biol*. 5: 47

Hoy MA. 2009. The predatory mite *Metaseiulus occidentalis*: mitey small and mitey large genomes. *BioEssays: news and reviews in molecular, cellular and developmental biology*. 31(5): 583-590

Hrycaj SM, Wellik DM. 2016. Hox genes and evolution. *F1000Res*. 5

Hui JH, 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 Res*. 41(5): 3352-61

Hui JH, McDougall C, Monteiro AS, Holland PW, Arendt D, Balavoine G, Ferrier DE. 2012. Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. *Mol Biol Evol*. 29(1): 157-65

Huminiecki L, Conant GC. 2012. Polyploidy and the evolution of complex traits. *Int J Evol Biol*. 2012: 292068

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 11(2): 97-108

Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ. 2014. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models. *Insect Biochem Mol Biol*. 52: 51-9

Jager M, Murienne J, Clabaut C, Deutsch J, Le Guyader H, Manuel M. 2006. Homology of arthropod anterior appendages revealed by Hox gene expression in a sea spider. *Nature*. 441: 506-508

Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, Hopfen C, Colbourne JK, Budd GE, Brown SJ, et al. 2010. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. *BMC Evol Biol*. 10(372): 1-21

Janssen R, Schönauer A, Weber M, Turetzek N, Hogvall M, Goss GE, Patel NH, McGregor AP, Hilbrant M. 2015. The evolution and expression of panarthropod *frizzled* genes. *Frontiers in Ecology and Evolution*. 3

Jiang H, Zhang X, Luo J, Dong C, Xue J, Wei W, Chen J, Zhou J, Gao Y, Yang C. 2012. Knockdown of *hMex-3A* by small RNA interference suppresses cell proliferation and migration in human gastric cancer cells. *Mol Med Rep*. 6(3): 575-80

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human MicroRNA targets. *PLoS Biol*. 2(11): e363

Johnston J, Chan R, Calderon-Segura M, McFarlane S, Browder LW. 2005. The roles of *Bcl-xL* in modulating apoptosis during development of *Xenopus laevis*. *BMC Dev Biol*. 5: 20

Juvvuna PK, Khandelia P, Lee LM, Makeyev EV. 2012. Argonaute identity defines the length of mature mammalian miRNAs. *Nucleic Acids Res*. 40(14): 6808-6820

Kanayama M, Akiyama-Oda Y, Nishimura O, Tarui H, Agata K, Oda H. 2011. Travelling and splitting of a wave of hedgehog expression involved in spider-head segmentation. *Nat Commun*. 2(500)

Kaufman TC, Lewis R, Wakimoto B. 1980. Cytogenetic analysis of chromosome 3 in *Drosophila melanogaster*: The homeotic gene complex in polytene chromosome interval 84A-B. *Genetics*. 94: 115-133

Kawamata T, Yoda M, Tomari Y. 2011. Multilayer checkpoints for miRNA authenticity during RISC assembly. *EMBO reports*. 12(9): 944-949

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28(12): 1647-9

Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PW, Chu KH, et al. 2015a. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity (Edinb)*.

Kenny NJ, Shen X, Chan TT, Wong NW, Chan TF, Chu KH, Lam HM, Hui JH. 2015b. Genome of the rusty millipede, *Trigoniulus corallinus*, illuminates diplopod, myriapod, and arthropod evolution. *Genome Biol Evol*. 7(5): 1280-95

Kenyon KL, Ranade SS, Curtiss J, Mlodzik M, Pignoni F. 2003. Coordinating proliferation and tissue specification to promote regional identity in the *Drosophila* head. *Developmental Cell*. 5(3): 403-414

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2012. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14(4): R36

Korkut C, Budnik V. 2009. WNTs tune up the neuromuscular junction. *Nat Rev Neurosci*. 10(9): 627-634

Koshikawa S, Giorgianni MW, Vaccaro K, Kassner VA, Yoder JH, Werner T, Carroll SB. 2015. Gain of *cis*-regulatory activities underlies novel domains of *wingless* gene expression in *Drosophila*. *Proc Natl Acad Sci U S A*. 112(24): 7524-9

Kozmik Z. 2008. The role of *Pax* genes in eye evolution. *Brain Res Bull*. 75(2-4): 335-9

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence miRNAs using deep sequencing data. *Nucleic Acids Res*. 42: D68-73

Krol AJ, Roellig D, Dequeant ML, Tassy O, Glynn E, Hattem G, Mushegian A, Oates AC, Pourquie O. 2011. Evolutionary plasticity of segmentation clock networks. *Development*. 138(13): 2783-92

Krumlauf R. 2016. Hox genes and the hindbrain: A study in segments. *Curr Top Dev Biol*. 116: 581-96

Kujjo LL, Perez GI. 2012. Ceramide and mitochondrial function in aging oocytes: joggling a new hypothesis and old players. *Reproduction*. 143(1): 1-10

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 33(7): 1870-4

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissieres V, Pickle CS, Plajzer-Frick I, Lee EA, et al. 2016. Progressive loss of function in a limb enhancer during snake evolution. *Cell*. 167(3): 633-642 e11

Lai Z, Fortini ME, Rubin GM. 1991. The embryonic expression patterns of *zfh-1* and *zfh-2*, two *Drosophila* genes encoding novel zinc-finger homeodomain proteins *Mechanisms of Development*. 34: 123-134

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4): 357-9

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10(R25)

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. ClustalW and ClustalX version 2. *Bioinformatics*. 23(21): 2947-2948

Laver JD, Marsolais AJ, Smibert CA, Lipshitz HD. 2015. Regulation and function of maternal gene products during the maternal-to-zygotic transition in *Drosophila*. *Curr Top Dev Biol*. 113: 43-84

Lavore A, Pagola L, Esponda-Behrens N, Rivera-Pomar R. 2012. The gap gene giant of *Rhodnius prolixus* is maternally expressed and required for proper head and abdomen formation. *Dev Biol*. 361(1): 147-55

Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV. 2009. VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res*. 37: D583-587

Lee MT, Bonneau AR, Giraldez AJ. 2014. Zygotic genome activation during the maternal-to-zygotic transition. *Annu Rev Cell Dev Biol*. 30: 581-613

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*. 23: 4051-4060

Leite DJ, McGregor AP. 2016. Arthropod evolution and development: recent insights from chelicerates and myriapods. *Curr Opin Genet Dev*. 39: 93-100

Leite DJ, Ninova M, Hilbrant M, Arif S, Griffiths-Jones S, Ronshaugen M, McGregor AP. 2016. Pervasive microRNA duplication in chelicerates: Insights from the embryonic microRNA repertoire of the spider *Parasteatoda tepidariorum*. *Genome Biol Evol*. 8(7): 2133-44

Lettice LA, Heaney SJH, Purdie LA, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*. 12(14): 1725-1735

Levine MS, Davidson EH. 2005. Gene regulatory networks for development. *PNAS*. 102(14): 4936-4942

Lewis EB. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature*. 276: 565-570

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25(16): 2078-9

Li J, Musso G, Zhang Z. 2008. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome Biol*. 9(8): R132

Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. 2009b. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics*. 10 Suppl 6: S8

Liao M, Boldbaatar D, Gong H, Huang P, Umemiya R, Harnnoi T, Zhou J, Tanaka T, Suzuki H, Xuan X, et al. 2008. Functional analysis of protein disulfide isomerases in blood feeding, viability and oocyte development in *Haemaphysalis longicornis* ticks. *Insect Biochem Mol Biol*. 38(3): 285-95

Lindeman RE, Pelegri F. 2012. Localized products of futile cycle/lrmp promote centrosome-nucleus attachment in the zebrafish zygote. *Curr Biol*. 22(10): 843-851

Liu B, Dou W, Ding TB, Zhong R, Liao CY, Xia WK, Wang JJ. 2014a. An analysis of the small RNA transcriptome of four developmental stages of the citrus red mite (*Panonychus citri*). *Insect Mol Biol*. 23(2): 216-29

Liu MM, Davey JW, Jackson DJ, Blaxter ML, Davison A. 2014b. A conserved set of maternal genes? Insights from a molluscan transcriptome. *Int J Dev Biol*. 58(6-8): 501-11

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15(12): 550

Lund E, Liu M, Hartley RS, Sheets MD, Dahlberg JE. 2009. Deadenylation of maternal mRNAs mediated by *miR-427* in *Xenopus laevis* embryos. *RNA*. 15(12): 2351-63

Luo J, Liu GY, Chen Z, Ren QY, Yin H, Luo JX, Wang H. 2015. Identification and characterization of microRNAs by deep-sequencing in *Hyalomma anatolicum anatolicum* (Acari: Ixodidae) ticks. *Gene*. 564(2): 125-33

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet*. 20(11): 544-9

Lynch VJ, Roth JJ, Wagner GP. 2006. Adaptive evolution of Hox-gene homeodomains after cluster duplications. *BMC Evol Biol*. 6: 86

Lyu Y, Shen Y, Li H, Chen Y, Guo L, Zhao Y, Hungate E, Shi S, Wu CI, Tang T. 2014. New microRNAs in *Drosophila*--birth, death and cycles of adaptive evolution. *PLoS Genet*. 10(1): e1004096

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *PNAS*. 102(15): 5454 –5459

Makeyev AV, Liebhaber SA. 2002. The poly(C)-binding proteins: A multiplicity of functions and a search for mechanisms. *RNA*. 8: 265-278

Manuel M, Jager M, Murienne J, Clabaut C, Le Guyader H. 2006. Hox genes in sea spiders (Pycnogonida) and the homology of arthropod head segments. *Dev Genes Evol*. 216: 481-491

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res*. 43(Database issue): D222-6

Marco A. 2014. Sex-biased expression of microRNAs in *Drosophila melanogaster*. *Open Biol*. 4: 140024

Marco A, Hooks K, Griffiths-Jones S. 2012a. Evolution and function of the extended miR-2 microRNA family. *RNA Biology*. 9(3): 242-248

Marco A, Hui JH, Ronshaugen M, Griffiths-Jones S. 2010. Functional shifts in insect miRNA evolution. *Genome Biol Evol*. 2: 686-696

Marco A, MacPherson JI, Ronshaugen M, Griffiths-Jones S. 2012b. MicroRNAs from the same precursor have different targeting properties. *Silence*. 3(8)

Marco A, Ninova M, Ronshaugen M, Griffiths-Jones S. 2013. Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Res*. 41(16): 7745-52

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17: 10-12

Massirer KB, Perez SG, Mondol V, Pasquinelli AE. 2012. The miR-35-41 family of miRNAs regulates RNAi sensitivity in *Caenorhabditis elegans*. *PLoS Genetics*. 8(3): e1002536

Mazza ME, Pang K, Reitzel AM, Martindale MQ, Finnerty JR. 2010. A conserved cluster of three PRD-class homeobox genes (*homeobrain*, *rx* and *orthopedia*) in the Cnidaria and Protostomia. *Evodevo*. 1(3)

Mazzoni EO, Mahony S, Iacovino M, Morrison CA, Mountoufaris G, Closser M, Whyte WA, Young RA, Kyba M, Gifford DK, et al. 2011. Embryonic stem cell-based mapping of developmental transcriptional programs. *Nature Methods* 8(12): 1056-1060

McGrath CL, Gout JF, Johri P, Doak TG, Lynch M. 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res*. 24(10): 1665-75

McGregor AP. 2005. How to get ahead: the origin, evolution and function of *bicoid*. *Bioessays*. 27(9): 904-13

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature*. 448(7153): 587-90

McJunkin K, Ambros V. 2014. The embryonic mir-35 family of miRNAs promotes multiple aspects of fecundity in *Caenorhabditis elegans*. *G3*. 4: 1747-1754

Mendes CC, Mirth CK. 2016. Stage-specific plasticity in ovary size is regulated by insulin/Insulin-like growth factor and ecdysone signaling in *Drosophila*. *Genetics*. 202(2): 703-19

Michalik P, Ramírez MJ. 2014. Evolutionary morphology of the male reproductive system, spermatozoa and seminal fluid of spiders (Araneae, Arachnida) – Current knowledge and future directions. *Arthropod Structure & Development*. 43(4): 291-322

Mirth CK, Shingleton AW. 2014. The roles of juvenile hormone, insulin/target of rapamycin, and ecdysone signaling in regulating body size in *Drosophila*. *Commun Integr Biol*. 7(5)

Miska EA, Alvarez-Saavedra E, Abbott AL, Lau NC, Hellman AB, McGonagle SM, Bartel DP, Ambros VR, Horvitz HR. 2007. Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genet*. 3(12): e215

Mittmann B, Wolff C. 2012. Embryonic development and staging of the cobweb spider *Parasteatoda tepidariorum* C. L. Koch, 1841 (syn.: *Achaearanea tepidariorum*; Araneomorphae; Theridiidae). *Dev Genes Evol*. 222(4): 189-216

Mohammed J, Bortolamiol-Becet D, Flynt AS, Gronau I, Siepel A, Lai EC. 2014a. Adaptive evolution of testis-specific, recently evolved, clustered miRNAs in *Drosophila*. *RNA*. 20(8): 1195-209

Mohammed J, Siepel A, Lai EC. 2014b. Diverse modes of evolutionary emergence and flux of conserved microRNA clusters. *RNA*. 20(12): 1850-63

Moran Y, Agron M, Praher D, Technau U. 2017. The evolutionary origin of plant and animal microRNAs. *Nature Ecology & Evolution*. 1(3): 0027

Mungpakdee S, Seo HC, Angotzi AR, Dong X, Akalin A, Chourrout D. 2008. Differential evolution of the 13 Atlantic salmon Hox clusters. *Mol Biol Evol*. 25(7): 1333-43

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 25: 1335–1337

Ninova M, Ronshaugen M, Griffiths-Jones S. 2014. Fast-evolving microRNAs are highly expressed in the early embryo of *Drosophila virilis*. *RNA*. 20(3): 360-72

Ninova M, Ronshaugen M, Griffiths-Jones S. 2016. MicroRNA evolution, expression, and function during short germband development in *Tribolium castaneum*. *Genome Res*. 26(1): 85-96

Nirenberg M, Leder P. 1964. RNA codewords and protein synthesis: The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science*. 145: 1399-1407

Nossa CW, Havlak P, Yue JX, Vincent KY, Brockmann J, Putman NH. 2014. Joint assembly and genetic mapping of the Atlantic horshoe crab genome reveals ancient whole genome duplication. *Giga Science*. 9(3): 1-21

Ohno S 1970. *Evolution by gene duplication,* Berlin, Springer-Verlag.

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC. 2008. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat Struct Mol Biol*. 15(4): 354-63

Onuma Y, Takahashi S, Asashima M, Kurata S, Gehring WJ. 2001. Conservation of *Pax 6* function and upstream activation by *Notch* signaling in eye development of frogs and flies. *PNAS*. 99(4): 2020-2025

Ortiz-Lombardia M, Foos N, Maurel-Zaffran C, Saurin AJ, Graba Y. 2017. Hox functional diversity: Novel insights from flexible motif folding and plastic protein interaction. *Bioessays*. 39(4)

Pace RM, Grbic M, Nagy LM. 2016. Composition and genomic organization of arthropod Hox clusters. *Evodevo*. 7: 11

Palfy M, Joseph SR, Vastenhouw NL. 2017. The timing of zygotic genome activation. *Curr Opin Genet Dev*. 43: 53-60

Paps J, Xu F, Zhang G, Holland PW. 2015. Reinforcing the egg-timer: recruitment of novel lophotrochozoa homeobox genes to early and late development in the pacific oyster. *Genome Biol Evol*. 7(3): 677-88

Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernàndez J. 2012. Evolution of Hox gene clusters in deuterostomes. *BMC Dev Biol*. 13(26): 1-14

Patraquim P, Warnefors M, Alonso CR. 2011. Evolution of Hox post-transcriptional regulation by alternative polyadenylation and microRNA modulation within 12 *Drosophila* genomes. *Mol Biol Evol*. 28(9): 2453-60

Pechmann M. 2016. Formation of the germ-disc in spider embryos by a condensation-like mechanism. *Front Zool*. 13: 35

Pechmann M, Benton MA, Kenny NJ, Posnien N, Roth S. 2017. A novel role for *Ets4* in axis specification and cell migration in the spider *Parasteatoda tepidariorum*. *Elife*. 6(e27590)

Pechmann M, Khadjeh S, Turetzek N, McGregor AP, Damen WG, Prpic NM. 2011. Novel function of *Distal-less* as a gap gene during spider segmentation. *PLoS Genet*. 7(10): e1002342

Pechmann M, Schwager EE, Turetzek N, Prpic NM. 2015. Regressive evolution of the arthropod tritocerebral segment linked to functional divergence of the Hox gene *labial*. *Proc Biol Sci*. 282(1814)

Pepato AR, Klimov PB. 2015. Origin and higher-level diversification of acariform mites - evidence from nuclear ribosomal genes, extensive taxon sampling, and secondary structure alignment. *BMC Evol Biol*. 15: 178

Pereira B, Le Borgne M, Chartier NT, Billaud M, Almeida R. 2013. MEX-3 proteins: recent insights on novel post-transcriptional regulators. *Trends Biochem Sci*. 38(10): 477-9

Pérez-Bercoff A, Burglin TR, . 2010. LogoBar - Visualizing protein sequence logos with gaps. In: Fung GPC (ed) *Sequence and Genome Analysis: Methods and Applications II*. iConcept Press Ltd., Hong Kong, pp 57-70.

Perez-Bercoff A, Koch J, Burglin TR. 2006. LogoBar: bar graph visualization of protein logos with gaps. *Bioinformatics*. 22(1): 112-4

Peterson KJ, Dietrich MR, McPeek MA. 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *BioEssays: news and reviews in molecular, cellular and developmental biology*. 31: 736-747

Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics*. 30(19): 2808-10

Pick L, Heffer A. 2012. Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. *Ann N Y Acad Sci*. 1256: 15-32

Pink RC, Samuel P, Massa D, Caley DP, Brooks SA, Carter DR. 2015. The passenger strand, miR-21-3p, plays a role in mediating cisplatin resistance in ovarian cancer cells. *Gynecol Oncol*. 137(1): 143-151

Pires CV, Freitas FC, Cristino AS, Dearden PK, Simoes ZL. 2016. Transcriptome analysis of honeybee (*Apis Mellifera*) haploid and diploid embryos reveals early zygotic transcription during cleavage. *PLoS One*. 11(1): e0146447

Pollard S, Holland PW. 2000. Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr Biol*. 10(17): 1059-1062

Popadić A, Nagy L. 2001. Conservation and variation in *Ubx* expression among chelicerates. *Evol Dev*. 3(6): 391-396

Posadas DM, Carthew RW. 2014. MicroRNAs and their roles in developmental canalization. *Curr Opin Genet Dev*. 27: 1-6

Posnien N, Zeng V, Schwager EE, Pechmann M, Hilbrant M, Keefe JD, Damen WG, Prpic NM, McGregor AP, Extavour CG. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. *PLoS One*. 9(8): e104885

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *PNAS*. 104: 8605– 8612

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. 2006. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature*. 440(7087): 1050-3

Prudent J, Popgeorgiev N, Bonneau B, Thibaut J, Gadet R, Lopez J, Gonzalo P, Rimokh R, Manon S, Houart C, et al. 2013. Bcl-wav and the mitochondrial calcium uniporter drive gastrula morphogenesis in zebrafish. *Nat Commun*. 4: 2330

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature*. 453(7198): 1064-71

Qu Z, Kenny NJ, Lam HM, Chan TF, Chu KH, Bendena WG, Tobe SS, Hui JH. 2015. How did arthropod sesquiterpenoids and ecdysteroids arise? Comparison of hormonal pathway genes in noninsect arthropod genomes. *Genome Biol Evol*. 7(7): 1951-9

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26(6): 841-2

R Core Team 2015. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rajhans R, Kumar GS, Dubey PK, Sharma GT. 2010. Effect of timing of development on total cell number and expression profile of *HSP-70.1* and *GLUT-1* in buffalo (*Bubalus bubalis*) oocytes and preimplantation embryos produced in vitro. *Cell Biol Int*. 34(5): 463-8

Rajman M, Schratt G. 2017. MicroRNAs in neural development: from master regulators to fine-tuners. *Development*. 144(13): 2310-2322

Randazzo FM, Cribbs DL, Kaufman TC. 1991. Rescue and regulation of *proboscipedia*: a homeotic gene of the Antennapedia Complex. *Development*. 113: 257-271

Rehm P, Meusemann K, Borner J, Misof B, Burmester T. 2014. Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. *Mol Phylogenet Evol*. 77: 25-33

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*. 403: 901-905

Richards S, Gibbs RA, Weinstock GM, Brown SJ. 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature*. 452(7190): 949-955

Rokas A. 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet*. 42: 235-51

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*. 25(9): 1943-53

Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. *Nature*. 415: 914-917

Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc Biol Sci*. 278(1703): 298-306

Rothbächer U, Laurent MN, Blitz IL, Watabe T, Marsh JL, Cho KWY. 1995. Functional conservation of the *Wnt* signaling pathway revealed by ectopic expression of *Drosophila dishevelled* in *Xenopus*. *Dev Biol*. 170: 717-721

Rotini A, Martinez-Sarra E, Pozzo E, Sampaolesi M. 2017. Interactions between microRNAs and long non-coding RNAs in cardiac development and repair. *Pharmacol Res*.

Rubio M, de Horna A, Belles X. 2012. MicroRNAs in metamorphic and non-metamorphic transitions in hemimetabolan insect metamorphosis. *BMC Genomics*. 13(386): 1-10

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev*. 20(16): 2214-2222

Samadi L, Schmid A, Eriksson BJ. 2015. Differential expression of retinal determination genes in the principal and secondary eyes of *Cupiennius salei* Keyserling (1877). *Evodevo*. 6: 16

Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, Jiang X, Cheng L, Fan D, Feng Y, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. *Nat Commun*. 5: 3765

Satou Y. 1999. *posterior end mark 3* (*pem-3*), an ascidian maternally expressed gene with localized mRNA encodes a protein with *Caenorhabditis elegans* MEX-3-like KH domains. *Dev Biol*. 212: 337-350

Schenkelaars Q, Pratlong M, Kodjabachian L, Fierro-Constain L, Vacelet J, Le Bivic A, Renard E, Borchiellini C. 2017. Animal multicellularity and polarity without Wnt signaling. *Sci Rep*. 7(1): 15383

Schmitz JF, Bornberg-Bauer E. 2017. Fact or fiction: updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. *F1000Res*. 6: 57

Scholtz G, Mittmann B, Gerberding M. 1998. The pattern of *Distal-less* expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of Mandibulata. *Int J Dev Biol*. 42: 801-810

Schomburg C, Turetzek N, Schacht MI, Schneider J, Kirfel P, Prpic NM, Posnien N. 2015. Molecular characterization and embryonic origin of the eyes in the common house spider *Parasteatoda tepidariorum*. *Evodevo*. 6: 15

Schönauer A, Paese CL, Hilbrant M, Leite DJ, Schwager EE, Feitosa NM, Eibner C, Damen WG, McGregor AP. 2016. The Wnt and Delta-Notch signalling pathways interact to direct pair-rule gene expression via *caudal* during segment addition in the spider *Parasteatoda tepidariorum*. *Development*. 143(13): 2455-63

Schoppmeier M, Fischer S, Schmitt-Engel C, Lohr U, Klingler M. 2009. An ancient anterior patterning system promotes *caudal* repression and head formation in ecdysozoa. *Curr Biol*. 19(21): 1811-5

Schwager EE. 2008. *Segmentation of the spider Achaearanea tepidariorum investigated by gene expression and functional analysis of the gap gene hunchback*. PhD, Universität zu Köln

Schwager EE, Schönauer A, Leite DJ, Sharma PP, McGregor AP 2015. Chelicerata. *In:* WANNINGER, A. (ed.) *Evolutionary Developmental Biology of Invertebrates*. Spinger.

Schwager EE, Schoppmeier M, Pechmann M, Damen WG. 2007. Duplicated Hox genes in the spider *Cupiennius salei*. *Front Zool*. 4: 10

Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, et al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*. 15(1)

Sebe-Pedros A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Mol Biol Evol*. 28(3): 1241-1254

Sempere LF, Cole CN, McPeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol*. 306(6): 575-88

Sharma PP, Gupta T, Schwager EE, Wheeler WC, Extavour CG. 2014a. Subdivision of arthropod cap-n-collar expression domains is restricted to Mandibulata. *Evodevo*. 5(1): 3

Sharma PP, Kaluziak ST, Perez-Porro AR, Gonzalez VL, Hormiga G, Wheeler WC, Giribet G. 2014b. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol Biol Evol*. 31(11): 2963-84

Sharma PP, Santiago MA, Gonzalez-Santillan E, Monod L, Wheeler WC. 2015. Evidence of duplicated Hox genes in the most recent common ancestor of extant scorpions. *Evol Dev*. 17(6): 347-55

Sharma PP, Schwager EE, Extavour CG, Giribet G. 2012. Hox gene expression in the harvestman *Phalangium opilio* reveals divergent patterning of the chelicerate opisthosoma. *Evol Dev*. 14(5): 450-63

Sharma PP, Schwager EE, Extavour CG, Wheeler WC. 2014c. Hox gene duplications correlate with posterior heteronomy in scorpions. *Proc Biol Sci*. 281(1792)

Shlyueva D, Meireles-Filho AC, Pagani M, Stark A. 2016. Genome-Wide Ultrabithorax Binding Analysis Reveals Highly Targeted Genomic Loci at Developmental Regulators and a Potential Connection to Polycomb-Mediated Regulation. *PLoS One*. 11(8): e0161997

Sible JC, Anderson JA, Lewellyn AL. 1997. Zygotic transcription is required to block a maternal program of apoptosis in *Xenopus* embryos. *Dev Biol*. 189: 335-346

Sidow A. 1992. Diversification of the *Wnt* gene family on the ancestral lineage of vertebrates. *PNAS*. 89: 5098-5102

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 7: 539

Sokol NS, Xu P, Jan YN, Ambros V. 2008. *Drosophila let-7* microRNA is required for remodeling of the neuromusculature during metamorphosis. *Genes Dev*. 22(12): 1591-6

Sommer RJ. 2012. Evolution of regulatory networks: nematode vulva induction as an example of developmental systems drift. *Adv Exp Med Biol*. 751: 79-91

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol*. 57(5): 758-71

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res*. 17(12): 1865-79

Stoeckius M, Grun D, Kirchner M, Ayoub S, Torti F, Piano F, Herzog M, Selbach M, Rajewsky N. 2014. Global characterization of the oocyte-to-embryo transition in *Caenorhabditis elegans* uncovers a novel mRNA clearance mechanism. *EMBO J*. 33(16): 1751-66

Tadros W, Lipshitz HD. 2009. The maternal-to-zygotic transition: a play in two acts. *Development*. 136(18): 3033-42

Tang Y, Wang J, Lian Y, Fan C, Zhang P, Wu Y, Li X, Xiong F, Li X, Li G, et al. 2017. Linking long non-coding RNAs and SWI/SNF complexes to chromatin remodeling in cancer. *Mol Cancer*. 16(1): 42

Tarver JE, Cormier A, Pinzon N, Taylor RS, Carre W, Strittmatter M, Seitz H, Coelho SM, Cock JM. 2015. microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga *Ectocarpus*. *Nucleic Acids Res*. 43(13): 6384-98

Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PC, Peterson KJ. 2013. miRNAs: small genes with big potential in metazoan phylogenetics. *Mol Biol Evol*. 30(11): 2369-82

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*. 38: 615-43

Telford MJ, Thomas RH. 1998. Expression of homeobox genes shows chelicerate arthropods retain their deutocerebral segment. *Proc Natl Acad Sci U S A*. 95: 10671–10675

Thomson T, Lin H. 2009. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu Rev Cell Dev Biol*. 25: 355-76

Tolkunova EN, Fujioka M, Kobayashi M, Deka D, Jaynes JB. 1998. Two distinct types of repression domain in engrailed:One interacts with the *Groucho* corepressor and is preferentially active on integrated target genes. *Molecular and Cellular Biology*. 18(5): 2804-2814

Trikic MZ, Monk P, Roehl H, Partridge LJ. 2011. Regulation of zebrafish hatching by tetraspanin *cd63*. *Plos One*. 6(5)

Turetzek N, Khadjeh S, Schomburg C, Prpic NM. 2017. Rapid diversification of *homothorax* expression patterns after gene duplication in spiders. *BMC Evol Biol*. 17(1): 168

Turetzek N, Pechmann M, Schomburg C, Schneider J, Prpic NM. 2016a. Neofunctionalisation of a duplicate *dachshund* gene underlies the evolution of a novel leg segment in arachnids *Molecular Biology and Evolution*. 33(1): 109-121

Turetzek N, Pechmann M, Schomburg C, Schneider J, Prpic NM. 2016b. Neofunctionalization of a Duplicate dachshund Gene Underlies the Evolution of a Novel Leg Segment in Arachnids. *Mol Biol Evol*. 33(1): 109-21

Undheim EA, Grimm LL, Low CF, Morgenstern D, Herzig V, Zobel-Thropp P, Pineda SS, Habib R, Dziemborowicz S, Fry BG, et al. 2015. Weaponization of a hormone: Convergent recruitment of hyperglycemic hormone into the venom of arthropod predators. *Structure*. 23(7): 1283-92

UniProt C. 2015. UniProt: a hub for protein information. *Nucleic Acids Res*. 43(Database issue): D204-12

Vachon G, Cohen B, Pfeifle C, McGuffin ME, Botas J, Cohen SM. 1992. Homeotic genes of the Bithorax complex repress limb development in the abdomen of the *Drosophila* embryo through the target gene *Distal-less*. *Cell*. 71(3): 437-450

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet*. 10(10): 725-32

van der Zee M, Berns N, Roth S. 2005. Distinct functions of the *Tribolium zerknullt* genes in serosa specification and dorsal closure. *Curr Biol*. 15(7): 624-36

van Eeden F, St Johnston D. 1999. The polarisation of the anterior-posterior and dorsal-ventral axes during *Drosophila* oogenesis. *Curr Opin Genet Dev*. 9: 396-404

Vorbrüggen G, Constien R, Zilian O, Wimmer EA, Dowe G, Taubert H, Noll M, Jäckle H. 1997. Embryonic expression and characterization of a *Ptx1* homolog in *Drosophila*. *Mechanisms of Development*. 68(1-2): 139-147

Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in *Arabidopsis* and Brassica. *Genome Biol Evol*. 7(3): 646-55

Wang W, Lo P, Fransch M, Lufkin T. 2000. *Hmx*: an evolutionary conserved homeobox gene family expressed in the developing nervous system in mice and *Drosophila*. *Mechanisms of Development*. 99: 123-137

Wang YP, Li KB. 2009. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*. 10: 218

Werner T, Koshikawa S, Williams TM, Carroll SB. 2010. Generation of a novel wing colour pattern by the *Wingless* morphogen. *Nature*. 464(7292): 1143-8

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ. 2009. The deep evolution of metazoan microRNAs. *Evol Dev*. 11(1): 50-68

Wheeler TJ, Eddy SR. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*. 29(19): 2487-9

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, et al. 2011. Episodic radiations in the fly tree of life. *Proc Natl Acad Sci U S A*. 108(14): 5690-5695

Willis JH. 2010. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochem Mol Biol*. 40(3): 189-204

Wong SF, Agarwal V, Mansfield JH, Denans N, Schwartz MG, Prosser HM, Pourquie O, Bartel DP, Tabin CJ, McGlinn E. 2015. Independent regulation of vertebral number and vertebral identity by microRNA-196 paralogs. *Proc Natl Acad Sci U S A*. 112(35): E4884-93

Wu E, Thivierge C, Flamand M, Mathonnet G, Vashisht AA, Wohlschlegel J, Fabian MR, Sonenberg N, Duchaine TF. 2010. Pervasive and cooperative deadenylation of 3'UTRs by embryonic microRNA families. *Mol Cell*. 40(4): 558-70

Xu X, Duan Z, Di Z, He Y, Li J, Li Z, Xie C, Zeng X, Cao Z, Wu Y, et al. 2014. Proteomic analysis of the venom from the scorpion Mesobuthus martensii. *J Proteomics*. 106: 162-80

Xu X, Wang H, Zhang F, Hu Z, Liang S, Liu Z. 2015. A comparative analysis of the venom gland transcriptomes of the fishing spiders *Dolomedes mizhoanus* and *Dolomedes sulfurous*. *PLoS One*. 10(10): e0139908

Yang JS, Smibert P, Westholm JO, Jee D, Maurin T, Lai EC. 2014. Intertwined pathways for Argonaute-mediated microRNA biogenesis in *Drosophila*. *Nucleic Acids Res*. 42(3): 1987-2002

Yuan Y, Ida JM, Paczkowski M, Krisher RL. 2011. Identification of developmental competence-related genes in mature porcine oocytes. *Mol Reprod Dev*. 78(8): 565-75

Zagozewski JL, Zhang Q, Pinto VI, Wigle JT, Eisenstat DD. 2014. The role of homeobox genes in retinal development and disease. *Dev Biol*. 393(2): 195-208

Zhong Y-F, Butts T, Holland PWH. 2008. HomeoDB: a database of homeobox gene diversity. *Evol Dev*. 10(5): 516-518

Zhong Y-F, Holland PWH. 2011. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev*. 13(6): 567-568

Zhou H, Huang X, Cui H, Luo X, Tang Y, Chen S, Wu L, Shen N. 2010. miR-155 and its star-form partner miR-155* cooperatively regulate type I interferon production by human plasmacytoid dendritic cells. *Blood*. 116(26): 5885-5894

Zhou J, Zhou Y, Cao J, Zhang H, Yu Y. 2013. Distinctive microRNA profiles in the salivary glands of *Haemaphysalis longicornis* related to tick blood-feeding. *Exp Appl Acarol*. 59(3): 339-49

Zhu SJ, Hallows SE, Currie KW, Xu C, Pearson BJ. 2015. A *mex3* homolog is required for differentiation during planarian stem cell lineage development. *Elife*. 4

Zondag L, Dearden PK, Wilson MJ. 2012. Deep sequencing and expression of microRNAs from early honeybee (*Apis mellifera*) embryos reveals a role in regulating early embryonic patterning. *BMC Evol Biol*. 12(211): 1-13

Zuniga A. 2015. Next generation limb development and evolution: old questions, new perspectives. *Development*. 142(22): 3810-20

Appendix

Appendix Table 3.1: *P. tepidariorum* homeobox genes

Classes	Gene Family	Gene Names	Gene Model	Other Features
ANTP	Cdx	*Cad*	aug3.g12643	
	Evx	*Eve*	aug3.g17585	
	Gbx	*Gbx1*	aug3.g14024	
	Gbx	*Gbx2*	aug3.g5351	
	Gsx	*Gsx*	aug3.g7582	
	Hox1	*lab-A*	aug3.g7954	
	Hox1	*lab-B*	aug3.g27130	
	Hox2	*Pb-A*	aug3.g7953	
	Hox2	*Pb-B*	aug3.g27127	
	Hox3	*Hox3-B*	aug3.g27124	
	Hox3	*Hox3-A*	aug3.g16493	
	Hox4	*Dfd-A*	aug3.g15608	
	Hox4	*Dfd-B*	aug3.g17637	
	Hox5	*Scr-A*	aug3.g15606	
	Hox5	*Scr-B*	aug3.g17640	
	Hox6-8	*ftz-B*	aug3.g1787	
	Hox6-8	*Antp-A*	aug3.g15600	
	Hox6-8	*Antp-B*	aug3.g1791	
	Hox6-8	*Ubx-A*	aug3.g15595	
	Hox6-8	*Ubx-B*	n/a	Scf5bK3_497:1506647..1506826(+)
	Hox6-8	*abdA-A*	n/a	Scf5bK3_833:1556584..1556763(+)
	Hox6-8	*abdA-B*	aug3.g25358	
	Hox9-13	*AbdB-A*	aug3.g9048	
	Hox9-13	*AbdB-B*	aug3.g25355	
	Meox	*Btn*	aug3.g6223	
	Mnx	*Exex*	aug3.g1062	
	Abox	*Abox*	aug3.g6219	
	Barhl	*BarHl*	aug3.g8250	
	Barhl	*BarH2*	aug3.g15951	
	Barx	*Barx*	aug3.g18073	
	Bari	*Bari*	aug3.g8698	
	Dbx	*Dbx1*	aug3.g19384	
	Dbx	*Dbx2*	aug3.g26237	
	Dlx	*Dll*	aug3.g15531	
	Emx	*Ems1*	aug3.g2295	
	Emx	*Ems2*	aug3.g2297	
	Emx	*Ems3*	aug3.g26210	
	Emx	*Ems4*	aug3.g26211	
	En	*Inv1*	aug3.g15983	Engrail
	En	*Inv2*	aug3.g15723	
	En	*En*	aug3.g7393	
	Hhex	*Hhex*	aug3.g4164	
	Lbx	*Lbx*	aug3.g16867	
	Msx	*Msx1*	aug3.g15189	
	Msx	*Msx2*	aug3.g4533	
	Msx	*Msx3*	aug3.g4165	
	Nedx	*Nedx*	aug3.g7174	
	Nk1	*Slou1*	aug3.g12478	
	Nk1	*Slou2*+	aug3.g23341	
	Nk2.1	*Scro*	aug3.g26115	
	Nk2.2	*Vnd1*	aug3.g1145	
	Nk2.2	*Vnd2*	aug3.g25082	
	Nk3	*Bap*	aug3.g9800	
	Nk4	*tin*	aug3.g10147	
	Nk5	*Hmx1*	aug3.g24988	
	Nk5	*Hmx2*	aug3.g26282	
	Nk6	*Hgtx1*	aug3.g222	
	Nk6	*Hgtx2*	aug3.g3597	
	Nk7	*Nk7a*	aug3.g11040	
	Nk7	*Nk7b*	aug3.g6892	
	Noto	*Noto1*	aug3.g15243	
	Noto	*Noto2*	aug3.g3361	
	Tlx	*C15a*	aug3.g4167	
	Tlx	*C15b*	aug3.g9802	
	Ro	*Ro*	aug3.g1423	
PRD	Arx	*Arx1*	aug3.g1240	
	Arx	*Arx2*	aug3.g2184	OAR
	Arx	*Arx3*^	aug3.g4151	
	CG11294	*CG11294*	aug3.g6013	
	Dmbx	*Dmbx1*	aug3.g13826	
	Dmbx	*Dmbx2*	aug3.g4686	
	Drgx	*Drgx1*	aug3.g21034	
	Drgx	*Drgx2*	aug3.g4287	
	Hbn	*Hbn*	aug3.g9068	
	Gsc	*Gsc*	aug3.g14050	
	Otp	*Otp*	aug3.g16467	
	Otx	*oc1*	aug3.g25936	OTX
	Otx	*oc2*	aug3.g9172	
	Pax3/7	*Prd1*	aug3.g11337	HTH
	Pax3/7	*Prd2*	aug3.g18397	PAX; HTH
	Pax4/6	*Pax6*	aug3.g11688	PAX; HTH
	Pax4/6	*Toy1**	aug3.g19881	
	Pax4/6	*Toy2**+	aug3.g22155	
	Pax4/6	*Toy3*	aug3.g24194	
	Pax4/6	*Eyg*	aug3.g25686	HTH
	Phox	*Phox1*	aug3.g26020	
	Phox	*Phox2*	aug3.g7566	
	Pitx	*Pitx1*	aug3.g12968	OAR
	Pitx	*Pitx2*	aug3.g13822	OAR
	Prop	*Prop1*	aug3.g10202	
	Prop	*Prop2*	aug3.g20598	
	Prrx	*Prrx*	aug3.g14368	
	Rax	*Rax1*	aug3.g17145	OAR
	Rax	*Rax2*	aug3.g9070	OAR
	Rax	*Rax3*^	aug3.g26283	RVE
	Repo	*Repo*	aug3.g12810	
	Shox	*Shox1*	aug3.g7763	
	Shox	*Shox2**	aug3.g25875	OAR
	Uncx	*Uncx1*	aug3.g2943	
	Uncx	*Uncx2*	aug3.g773	
	Vsx	*Vsx*	aug3.g27186	OAR
LIM	Isl	*Tup*	aug3.g3165	2x LIM
	Lhx1/5	*Lim1a*	aug3.g10485	LIM
	Lhx1/5	*Lim1b*	aug3.g16539	2x LIM
	Lhx1/5	*Lim1c**	aug3.g12957	
	Lhx2/9	*Ap1*	aug3.g2651	2x LIM
	Lhx2/9	*Ap2*	aug3.g2649	2x LIM
	Lhx2/9	*Ap3*	aug3.g6992	
	Lhx3/4	*Lim3a*	aug3.g13622	2x LIM
	Lhx3/4	*Lim3b*	aug3.g18628	2x LIM
	Lhx6/8	*Awh*	aug3.g23781	2x LIM
	Lmx	*Lmx*	aug3.g27743	LIM
POU	Pou2	*Nub*	aug3.g13364	POU
	Pou3	*Vvl1*	aug3.g15237	POU
	Pou3	*Vvl2*	aug3.g15946	POU
	Pou4	*Acj6a*	aug3.g15285	POU
	Pou4	*Acj6b*	aug3.g8303	POU; RVE
	Pou6	*Pou6f2*	aug3.g19365	POU
SINE	Six1/2	*So1*	aug3.g26047	SIX
	Six1/2	*So2*	aug3.g325	SIX
	Six3/6	*Optix1*	aug3.g26046	SIX
	Six3/6	*Optix2*	aug3.g3570	SIX
	Six3/6	*Optix3*^	aug3.g14592	SIX
	Six4/5	*Six4*	aug3.g328	SIX
HNF	Hmbox	*Hmbox1*	aug3.g25741	
	Hmbox	*Hmbox2*	aug3.g16446	HNF
TALE	Irx	*mirr1*	aug3.g12625	
	Irx	*mirr2*	aug3.g13696	IRO
	Irx	*mirr3*	aug3.g4636	
	Irx	*mirr4*	aug3.g8295	
	Meis	*Hth1*	aug3.g13867	
	Meis	*Hth2*	aug3.g3111	MEIS/PHNOX
	Pbx	*Exd1*	aug3.g20631	PBC
	Pbx	*Exd2*	aug3.g27748	PBC
	Pknox	*Pknox*	aug3.g6392	MEIS/PHNOX
	Tgif	*Tgif*	aug3.g11579	
CUT	Onecut	*Onecut1*	aug3.g4142	
	Onecut	*Onecut2*	aug3.g26177	CUT
	Cux	*Cux1*	aug3.g13427	3x CUT
	Cux	*Cux2*	aug3.g3872	3x CUT
	Cmp	*Dve*	aug3.g8928	2x HOX; SATB; ULD
ZFH	Zfhx	*Zfh1*	aug3.g11268	4x HOX; 8x Zinc finger; Atrophin; DUF
	Zfhx	*Zfh2*	aug3.g25763	4x HOX; 10x Zinc finger
CERS	Cers	*Cers*	aug3.g5792	TLC
PROS	Prox	*Pros*	aug3.g15298	HPD

* = Partial proteins
\+ = Partial protein with complete homeobox domain
^ = Possible other homeobox gene family

Appendix Table 3.2: *P. phalangioides* homeobox genes

Classes	Gene Family	Gene Names	Gene Model	Other Features	Classes	Gene Family	Gene Names	Gene Model	Other Features
ANTP	Cdx	*Cad**	c103077_g1_i1		**PRD**	Arx	*Arx*	c107787_g1_i1	OAR
	Evx	*Eve**	c100889_g1_i1			CG11294	*CG11294**	c153673_g1_i1	
	Gbx	*Gbx1*	c90896_g1_i1			Dmbx	*Dmbx**	c74916_g1_i1	
	Gbx	*Gbx2*	c90951_g1_i2			Drgx	*Drgx*	c56103_g2_i1	
	Gsx	*Gsx*	c12892_g1_i1			Hbn	*Hbn*	c106595_g1_i1	
	Hox1	*Lab1*	c103960_g6_i1			Gsc	*Gsc*	c132968_g1_i1	
	Hox2	*Pb1*	c104593_g1_i1			Otp	*Otp*	c103695_g1_i1	OAR
	Hox2	*Pb2*	c104684_g3_i1			Otx	*Oc1*	c112655_g2_i1	
	Hox3	*Hox3a*	c101164_g7_i1			Otx	*Oc2*	c110982_g3_i3	
	Hox3	*Hox3b**	c107945_g3_i3			Pax3/7	*Prd*	c113119_g1_i1	PAX/HTH
	Hox4	*Dfd1*	c95663_g1_i1			Pax4/6	*pax6l**	c88295_g2_i1	PAX/HTH
	Hox4	*Dfd2*	c96808_g3_i3			Pax4/6	*Toy1*	c112264_g2_i1	PAX/HTH
	Hox5	*Scr1*	c101195_g2_i1			Pax4/6	*Toy2*	c100812_g8_i1	PAX/HTH
	Hox5	*Scr2*	c109150_g1_i2			Pax4/6	*Eyg*	c105947_g1_i2	PAX/HTH
	Hox6-8	*Ftz*	c92385_g1_i1			Phox	*Phox1*	c107284_g2_i3	
	Hox6-8	*Antp1*	c105467_g2_i4			Phox	*Phox2*	c91992_g1_i1	
	Hox6-8	*Antp2*	c132392_g1_i1			Pitx	*Pitx1*	c100659_g3_i1	OAR
	Hox6-8	*Ubx*	c103414_g2_i9			Pitx	*Pitx2**	c77939_g1_i1	
	Hox6-8	*abd-A1*	c95590_g1_i1			Prop	*Prop*	c72989_g1_i1	
	Hox6-8	*abd-A2*	c111834_g6_i1			Prrx	*Prrx*	c98453_g1_i1	
	Hox9-13	*Abd-B1*	c97115_g1_i1			Rax	*Rax1*	c102000_g1_i1	OAR
	Hox6-8	*abd-B2*	c104286_g1_i3			Rax	*Rax2*	c110545_g4_i3	OAR
	Meox	*Btn*	c106591_g2_i1			Repo	*Repo*	c102959_g1_i1	
	Exex	*Exex*	c46999_g1_i1			Shox	*Shox*	c107809_g1_i1	
	Abox	*Abox*	c101304_g1_i1			Uncx	*Uncx1*	c64396_g1_i1	
	Barhl	*BarH1*	c103653_g1_i1			Uncx	*Uncx2*	c110656_g1_i4	
	Barhl	*BarH2**	c114352_g1_i1			Vsx	*Vsx*	c99447_g1_i4	OAR
	Bari	*Bari*	c98969_g1_i1		**LIM**	Isl	*Tup*	c110195_g2_i1	2xLIM
	Bsx	*Bsh*	c96123_g1_i1			Lhx1/5	*Lim1a*	c104870_g3_i1	2xLIM
	Dbx	*Dbx1*	c97963_g2_i1			Lhx1/5	*Lim1b*	c109007_g9_i1	2xLIM
	Dbx	*Dbx2*	c99617_g1_i1			Lhx2/9	*Ap1*	c96490_g1_i1	2xLIM
	Dll	*Dll*	c47313_g1_i1			Lhx2/9	*Ap2*	c100046_g1_i1	2xLIM
	Emx	*Ems1*	c100271_g1_i1			Lhx2/9	*Ap3*	c110513_g3_i1	2xLIM
	Emx	*Ems2*	c107662_g2_i7			Lhx3/4	*Lim3a**	c113413_g1_i1	
	Emx	*Ems3*	c85063_g1_i1			Lhx3/4	*Lim3b*	c107542_g2_i1	2xLIM
	en	*en*	c96625_g1_i1	Engrailed		Lhx6/8	*Awh2*	c93387_g1_i3	LIM
	Hhex	*Hhex*	c108606_g1_i1			Lmx	*Lmx*	c82652_g1_i1	2xLIM
	Lbx	*Lbx1*	c107876_g1_i1		**POU**	Pou2	*Nub*	c95823_g2_i3	POU
	Lbx	*Lbx2*	c76425_g1_i2			Pou3	*Vvl*	c98062_g1_i2	POU
	Msx	*Msx1*	c97092_g1_i2			Pou4	*Acj6a**	c30371_g1_i1	POU
	Msx	*Msx2*	c61338_g1_i1			Pou4	*Acj6b**	c76069_g1_i1	POU
	Msx	*Msx3*	c112434_g3_i1			Pou6	*Pou6f**	c99759_g2_i1	POU
	Msx	*Msx4*	c108649_g1_i1		**SINE**	Six1/2	*So1*	c77895_g1_i1	SIX
	Nedx	*Nedx*	c108987_g4_i1			Six1/2	*So2*	c109946_g4_i2	SIX
	Nk1	*Slou1*	c83150_g1_i1			Six3/6	*Optix1*	c104016_g8_i1	SIX
	Nk1	*Slou2*	c95467_g1_i2			Six3/6	*Optix2*	c89951_g1_i1	SIX
	Nk2.1	*Scro*	c101836_g1_i1			Six4/5	*Six4*	c98510_g6_i1	SIX
	Nk2.2	*Vnd1**	c204117_g1_i1		**HNF**	Hmbox	*Hmbox1**	c105474_g4_i1	HNF
	Nk2.2	*Vnd2*	c92423_g1_i1			Hmbox	*Hmbox2*	c90767_g1_i2	HNF
	Nk3	*Bap*	c43919_g1_i1		**TALE**	Irx	*Mirr1*	c61073_g1_i1	
	Nk4	*Tin1*	c98245_g1_i2			Irx	*Mirr2*	c36528_g1_i1	
	Nk4	*Tin2*	c95270_g1_i1			Irx	*Mirr3*	c108290_g3_i1	PDZ
	Nk5	*Hmx1**	c217500_g1_i1			Irx	*Mirr4*	c104725_g2_i1	IRO
	Nk5	*Hmx2*	c57611_g1_i1			Meis	*Hth1*	c81766_g1_i1	MEIS/PHNOX
	Nk6	*Hgtx1*	c97811_g1_i1			Meis	*Hth2*	c103774_g1_i2	MEIS/PHNOX
	Nk6	*Hgtx2**	c83744_g1_i1			Pbx	*Exd1*	c112509_g5_i1	PBC
	Nk7	*Nk7a*	c99350_g1_i2			Pbx	*Exd2*	c112509_g4_i1	
	Nk7	*Nk7b*	c109765_g1_i1			Tgif	*Tgif*	c112305_g4_i1	
	Noto	*Noto*	c94032_g1_i1		**CUT**	Onecut	*Onecut1**	c108240_g1_i3	CUT
	Tlx	*C15a*	c103762_g2_i1			Onecut	*Onecut2*	c102359_g1_i5	CUT
	Tlx	*C15b*	c99315_g1_i1			Cux	*Cux1**	c109052_g1_i1	CUT
	Ro	*Ro*	c95638_g2_i1			Cux	*Cux2*	c108008_g2_i2	3xCUT
						Cmp	*Dve1*	c111460_g3_i1	SATB/ULD;CUTL
						Cmp	*Dve2**	c100802_g2_i1	2xHOX
					ZFH	Zfhx	*zfhl1*	c98720_g1_i1	7xZnf
						Zfhx	*Zfhl2*	c106281_g4_i2	4xZnf
						Zfhx	*Zfh1**	c108341_g1_i1	3xHOX;AD;Lsm12
						Zfhx	*Zfh2*	c109439_g1_i2	4xHOX;7xZnf
					PROS	Prox	*Pros*	c98709_g1_i4	HPD
					Unassigned	Rhox	Rhox-like	c107986_g1_i1	

* = Partial proteins
\+ = Partial protein with complete homeobox domain
^ = Possible other homeobox gene family

Appendix Table 3.3: *C. sculpturatus* homeobox genes

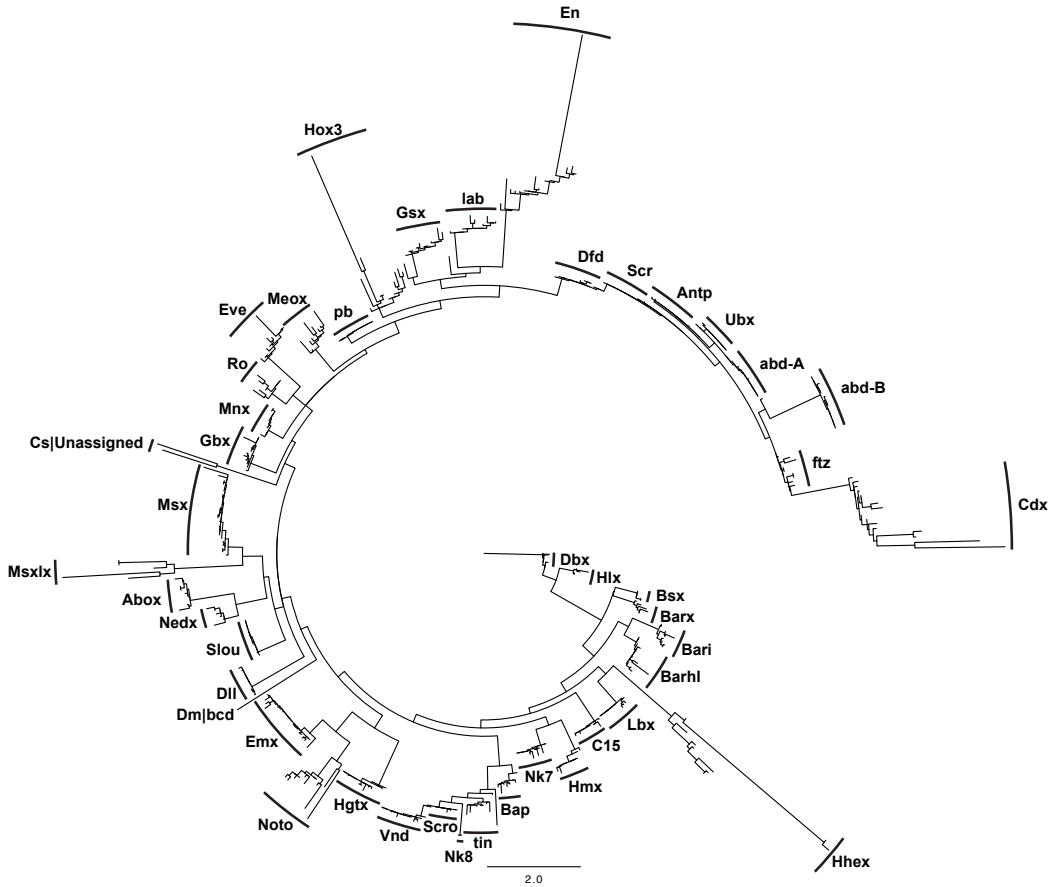
Classes	Gene Family	Gene Names	Gene Model	Other Features	Classes	Gene Family	Gene Names	Gene Model	Other Features
ANTP	Cdx	*Cad1*	CSCU017598		**PRD**	Arx	*Arx1*	CSCU016934	OAR
	Cdx	*Cad2*	CSCU012585			Arx	*Arx2*	CSCU014058	
	Evx	*Eve*	CSCU015763			Arx	*Arx3*	CSCU004778	
	Gbx	*Gbx1**+	CSCU000715			CG11294	*CG11294*	CSCU030184	
	Gbx	*Gbx2**+	CSCU028479			Drgx	*Drgx1*	CSCU000609	
	Gsx	*Gsx*	CSCU010281			Drgx	*Drgx2**+	CSCU025176	
	Unassigned	*Gsx-like*	CSCU008184			Hbn	*Hbn*	CSCU015099	
	Hox1	*Lab-B*	CSCU029776			Gsc	*Gsc*	CSCU004322	
	Hox1	*Lab-A*	CSCU023119			Otp	*Otp1*	CSCU015097	
	Hox2	*Pb-B*	CSCU016897			Otp	*Otp2*	CSCU025909	
	Hox2	*Pb-A*	CSCU010695			Otx	*Oc1**	CSCU005066	OTX
	Hox3	*Hox3-A*	CSCU005339			Otx	*Oc2*	CSCU008374	
	Hox4	*Dfd-A*	CSCU010691			Pax3/7	*Prd1*	CSCU023561	HTH
	Hox4	*Dfd-B*	CSCU005342			Pax3/7	*Prd2**+	CSCU025885	
	Hox5	*Scr-B*	CSCU005346			Pax4/6	*Pax6*	CSCU009125	HTH
	Hox5	*Scr-A**+	CSCU010690			Pax4/6	*Toy1**	CSCU004691	
	Hox6-8	*Ftz-B*	CSCU005348			Pax4/6	*Toy2*	CSCU004685	
	Hox6-8	*Ftz-A*	CSCU010689			Pax4/6	*Toy3*	CSCU004692	
	Hox6-8	*Antp-B*	CSCU005351			Pax4/6	*Eyg*	CSCU010913	HTH
	Hox6-8	*Antp-A*	CSCU010688			Phox	*Phox*	CSCU015671	
	Hox6-8	*Ubx-A*	CSCU030008			Pitx	*Pitx1*	CSCU020592	
	Hox6-8	*Ubx-B*	CSCU020689			Pitx	*Pitx2*	CSCU007998	
	Hox6-8	*abdA-B*	CSCU004419			Prop	*Prop2**+	CSCU002348	
	Hox6-8	*abdA-A*	CSCU020691			Prrx	*Prrx**+	CSCU006574	
	Hox9-13	*AbdB-B*	CSCU004421			Rax	*Rax1*	CSCU022535	OAR
	Hox9-13	*AbdB-A*	CSCU028980			Rax	*Rax2*	CSCU015098	
	Meox	*Btn*	CSCU005132			Repo	*Repo*	CSCU018539	
	Exex	*Exex1*	CSCU004049			Shox	*Shox*	CSCU025915	
	Exex	*Exex2**	CSCU030254			Uncx	*Uncx1*	CSCU024301	
	Abox	*Abox*	CSCU005126			Uncx	*Uncx2*	CSCU012570	
	Barhl	*BarH1**	CSCU016161			Vsx	*Vsx1*	CSCU022240	
	Barhl	*BarH2**+	CSCU016162			Vsx	*Vsx2*	CSCU013732	
	Barhl	*BarH3*	CSCU014407		**LIM**	Isl	*Tup*	CSCU000936	LIM
	Barx	*Barx**+	CSCU024305			Lhx1/5	*Lim1a*	CSCU000898	
	Bari	*Bari*	CSCU018911			Lhx1/5	*Lim1b*	CSCU001114	
	Dbx	*Dbx1*	CSCU006505			Lhx1/5	*Lim1c*	CSCU001113	
	Dbx	*Dbx2*	CSCU021006			Lhx2/9	*Ap1*	CSCU023766	LIM
	Unassigned	*Dbx-like*	CSCU010692			Lhx2/9	*Ap2*	CSCU022642	2x HOX; 2x LIM
	Hhex	*Hhexl**+	CSCU000795			Lhx3/4	*Lim3b*	CSCU027347	
	Dlx	*Dll**+	CSCU004046			Lhx3/4	*Lim3a*	CSCU002617	2x LIM
	Emx	*Ems1*	CSCU001125	156 bp insertion		Lhx6/8	*Awh1**+	CSCU022913	LIM
	Emx	*Ems2*	CSCU022596			Lhx6/8	*Awh2**+	CSCU003258	2x LIM
	Emx	*Ems3*	CSCU001126			Lmx	*Lmx*	CSCU001249	LIM
	Emx	*Ems4**+	CSCU018964		**POU**	Pou2	*Nub*	CSCU014674	POU
	En	*En1**+	CSCU022508	Engrailed		Pou3	*Vvl*	CSCU012403	POU
	En	*En2*	CSCU008610	Engrailed		Pou4	*Acj6a*	CSCU022210	POU
	Hhex	*Hhex**+	CSCU009121			Pou4	*Acj6b**+	CSCU012583	POU
	Hlx	*Hlx*	CSCU015122			Pou6	*Pou6f**+	CSCU021346	POU
	Lbx	*Lbx1*	CSCU006313		**SINE**	Six1/2	*So1*	CSCU000535	SIX
	Lbx	*Lbx2*	CSCU026705			Six1/2	*So2*	CSCU019875	SIX
	Msx	*Msx1*	CSCU016297			Six3/6	*Optix1*	CSCU000537	SIX
	Msx	*Msx2*	CSCU006456			Six3/6	*Optix2*	CSCU019874	SIX
	Msx	*Msx3*	CSCU008264			Six3/6	*Optix3*	CSCU007271	SIX
	Msx	*Msx4*	CSCU008265			Six4/5	*Six4*	CSCU000534	SIX
	Msx	*Msx5*	CSCU009122			Six7	*Six7*	CSCU019878	SIX
	Msxlx	*Msxlx*	CSCU021922		**TALE**	Irx	*Mirr1*	CSCU010447	IRO
	Nk1	*Slou1*	CSCU010836			Irx	*Mirr2*	CSCU010449	
	Nk1	*Slou2*	CSCU013524			Irx	*Mirr3**+	CSCU003601	
	Nk2.1	*Scro*	CSCU012467			Irx	*Mirr5**+	CSCU003602	
	Nk2.2	*Vnd1*	CSCU021921			Irx	*Mirr4**+	CSCU003603	IRO
	Nk2.2	*Vnd2*	CSCU004754			Irx	*Mirr6*	CSCU023360	
	Nk3	*Bap*	CSCU006454			Mkx	*Mkx**+	CSCU014654	
	Nk4	*Tin**+	CSCU006455			Meis	*Hth1**+	CSCU021453	
	Nk5	*Hmx1*	CSCU010835			Meis	*Hth2*	CSCU012140	MEIS/PHNOX
	Nk5	*Hmx2*	CSCU009972			Pbx	*Exd1**+	CSCU029859	
	Nk6	*Hgtx1*	CSCU029876			Pbx	*Exd2*	CSCU001243	PBC
	Nk6	*Hgtx2*	CSCU019444			Tgif	*Tgif*	CSCU001725	
	Nk6	*Hgtx3*	CSCU013459		**CUT**	Onecut	*Onecut1*	CSCU007543	
	Nk7	*NK7a*	CSCU010496			Onecut	*Onecut2*	CSCU014259	
	Nk7	*NK7b*	CSCU006312			Cux	*Cux1*	CSCU010245	
	Nk8	*Nk8*	CSCU007954			Cux	*Cux2*	CSCU002100	CUT
	Noto	*Noto*	CSCU013845			Cmp	*Dve1*	CSCU021829	SATB; ULD
	Tlx	*C15a*	CSCU030322			Cmp	*Dve2*	CSCU001092	2x HOX
	Tlx	*C15b*	CSCU006452		**HNF**	Hmbox	*Hmboxl1*	CSCU010952	
	Ro	*Ro*	CSCU002540			Hmbox	*Hmboxl2*	CSCU025618	
					ZFH	Zfhx	*Zfhl**+	CSCU018249	4x Znf
						Zfhx	*Zfh1*	CSCU021262	4x HOX; 7x Znf
						Zfhx	*Zfh2*	CSCU021822	4x HOX; 9x Znf
					CERS	Cers	*Cers**+	CSCU020408	TLC
					PROS	Prox	*Pros1*	CSCU024206	HPD
						Prox	*Pros2*	CSCU024205	no HPD

* = Partial proteins
\+ = Partial protein with complete homeobox domain
^ = Possible other homeobox gene family

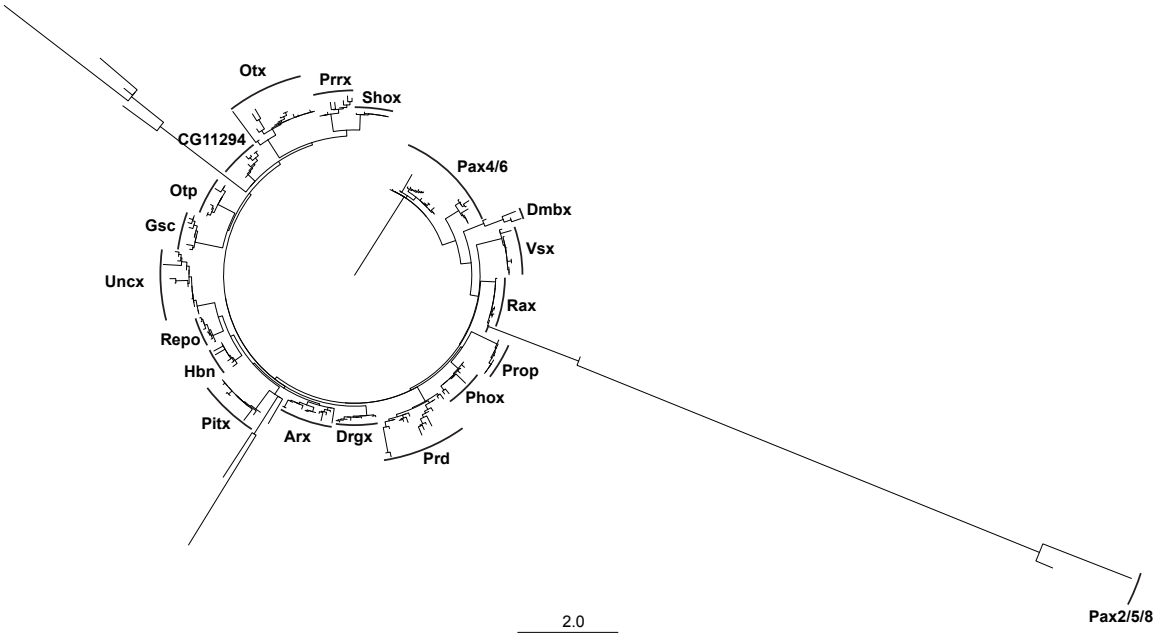
Appendix Table 3.4: *I. scapularis* homeobox genes

Classes	Gene Family	Gene Names	Gene Model	Other Features	Classes	Gene Family	Gene Names	Gene Model	Other Features
ANTP	Cdx	Cad1*+	EEC12818		**PRD**	Arx	Arx1	EEC00161	
	Cdx	Cad2	EEC11508			Arx	Arx2	EEC06233	OAR
	Cdx	Cad3	EEC11509			Arx	Arx3	EEC06234	
	Evx	Eve	EEC01090			Arx	Arx4	EEC11510	
	Gbx	Gbx	EEC06304			Drgx	Drgx1*	EEC02545	
	Gsx	Gsx1*+	EEC05273			Drgx	Drgx2	EEC15360	OAR
	Gsx	Gsx2	EEC13233			Gsc	Gsc	EEC11687	
	Hox1	Lab	EEC15651			Otp	Otp	EEC19355	
	Hox2	Pb	EEC15648			Otx	Oc*	EEC11447	
	Hox3	Hox3	EEC15645			Pax3/7	Prd*+	EEC19941	PAX/HTH
	Hox4	Dfd	EEC15643			Pax4/6	Toy*	EEC07338	PAX/HTH
	Hox5	Scr	EEC15638			Pax4/6	pax6l*+	EEC13842	PAX/HTH
	Hox6-8	Ftz	EEC15635			Phox	Phox	EEC15081	
	Hox6-8	Antp	EEC15630			Pitx	Pitx*	EEC07037	
	Hox6-8	Ubx	EEC15627			Prrx	Prrx	EEC18457	
	Hox6-8	abdA	EEC15624			Rax	Rax1	EEC19769	
	Hox9-13	AbdB	EEC15619			Rax	Rax2	EEC03745	HR1
	Meox	Btn*	EEC19260			Uncx	Uncx	EEC00159	
	Mnx	Exex*+	EEC04902			Vsx	Vsx	EEC11824	
	Abox	Abox*	EEC17078		**LIM**	Lhx1/5	Lim1	EEC17795	
	Barhl	BarH1	EEC09053			Lhx3/4	Lim3	EEC17797	LIM
	Barhl	BarH2	EEC09052			Lhx2/9	Ap1	EEC07790	
	Bsh	Bsh1*+	EEC07361			Lhx2/9	Ap2*+	EEC05711	LIM
	Bsh	Bsh2*+	EEC13007			Lhx6/8	Awh	EEC14711	
	Bari	Bari*+	EEC05088			Lmx	Lmx*+	EEC20369	LIM
	Bsx	Bsh	EEC08153		**POU**	Pou2	Nub	EEC17317	POU
	Dbx	Dbx	EEC11236			Pou3	Vvl	EEC13416	POU
	Dlx	Dll1	EEC12941			Pou4	Acj6	EEC09206	POU
	Dlx	Dll2*	EEC12942			Pou6	Pou6f	EEC05119	POU
	Emx	Ems1	EEC17392		**SINE**	Six3/6	Optix	EEC10081	SIX
	Emx	Ems2	EEC17393			Six4/5	Six4	EEC18860	SIX
	En	En	EEC17545	Engrailed	**TALE**	Irx	Mirr1*+	EEC00346	
	Lbx	Lbx	EEC14640			Irx	Mirr2	EEC00349	IRO
	Msx	Msx1	EEC15666			Irx	Mirr3*	EEC16401	
	Msx	Msx2	EEC08486			Mkx	Mkx	EEC02673	
	Nk1	Slou*+	EEC05904			Pbx	Exd	EEC18913	PBC
	Nk2	Scro	EEC12967			Pknox	Pknox*	EEC13638	MEIS/PHNOX
	Nk22	Vnd	EEC17629			Tgif	Tgif*+	EEC05053	
	Nk3	Bap*+	EEC18891		**CUT**	Onecut	Onecut	EEC01475	
	Nk4	Tin*	EEC15395			Cux	Cut	EEC16775	3xCUT
	Nk6	Hgtx3	EEC19838			Cmp	Dve1	EEC00261	CUTL
	Nk7	Nk7a	EEC04090			Cmp	Dve2*+	EEC00262	2xHOX;SATB/ULD
	Nk7	NK7b	EEC15322		**ZFH**	Zfhx	Zfhl	EEC01527	3xZnf
	Noto	Noto	EEC08664			Zfhx	Zfhx	EEC06824	4xHOX
	Tlx	C15	EEC02597		**HNF**	Hmbox	Hmbox1	EEC05740	2xHOX;HNF
	Ro	Ro	EEC09616			Hmbox	Hmbox2	EEC09248	3xHOX
					CERS	Cers	Cers1	EEC05014	TLC
						Cers	Cers2	EEC05423	TLC
					PROS	Prox	Pros	EEC12842	HPD
					Unassigned	unassigned	vox^	EEC02280	

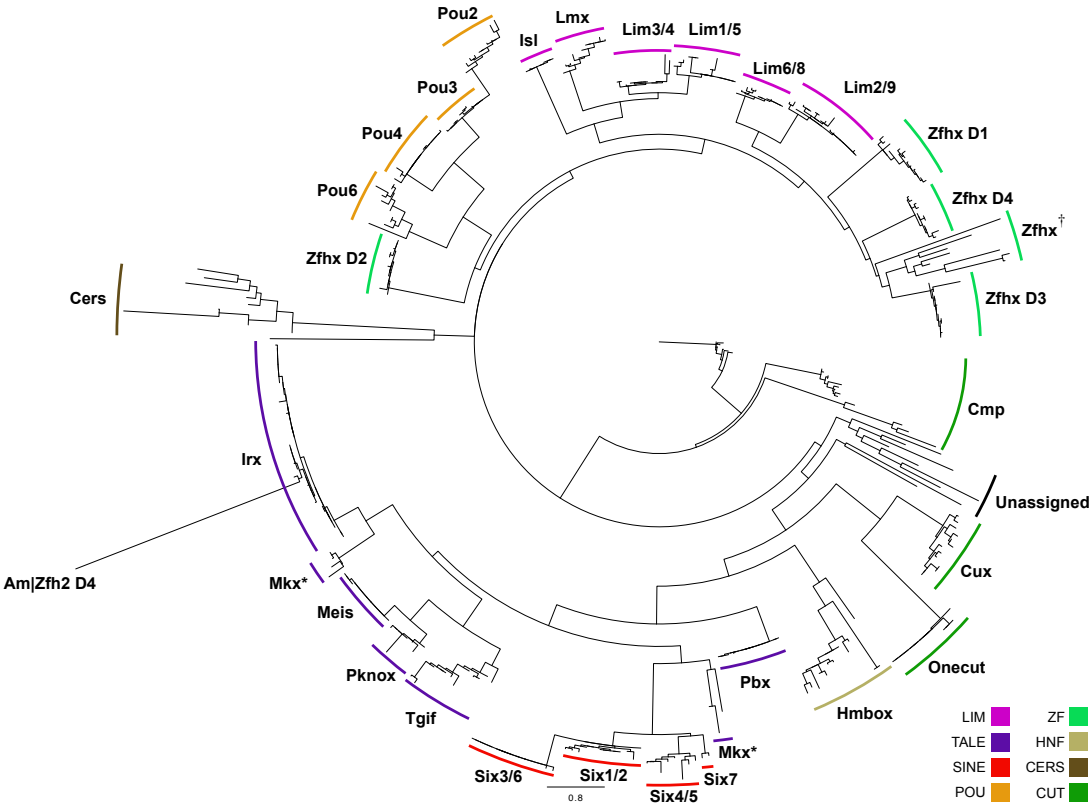
+ = Partial protein with complete homeobox domain
^ = Possible other homeobox gene family



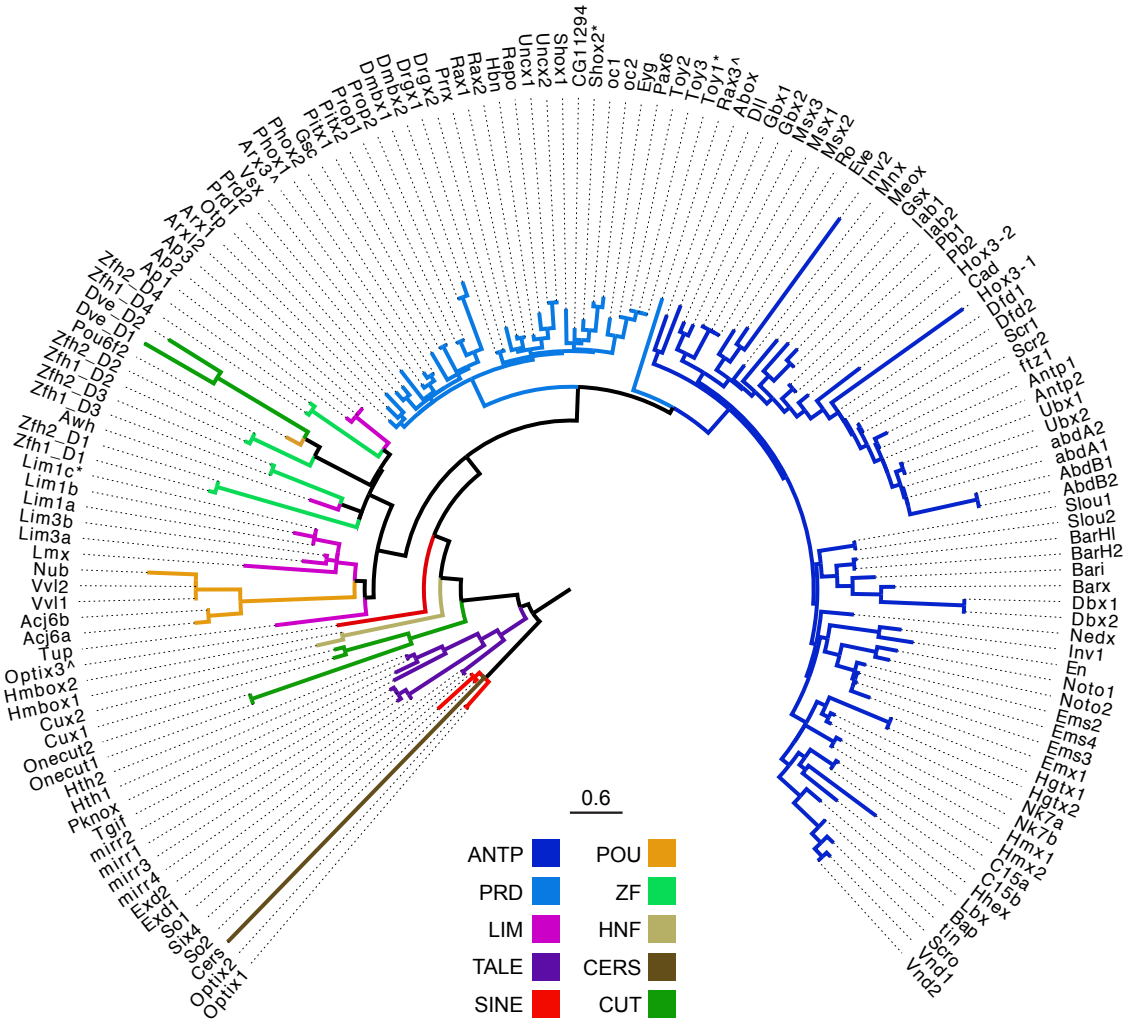
Appendix Figure 3.1: Maximum likelihood tree of the homeodomain from ANTP class homeobox genes. RAxML with model LG; 100 bootstraps; CAT model of heterogeneity. Species in the tree: *P. tepidariorum*, *P. phalangioides*, *C. sculpturatus*, *M. mimosarum*, *I. scapularis*, *S. maritima*, *A. mellifera*, *T. castaneum*, *D. melanogaster*.



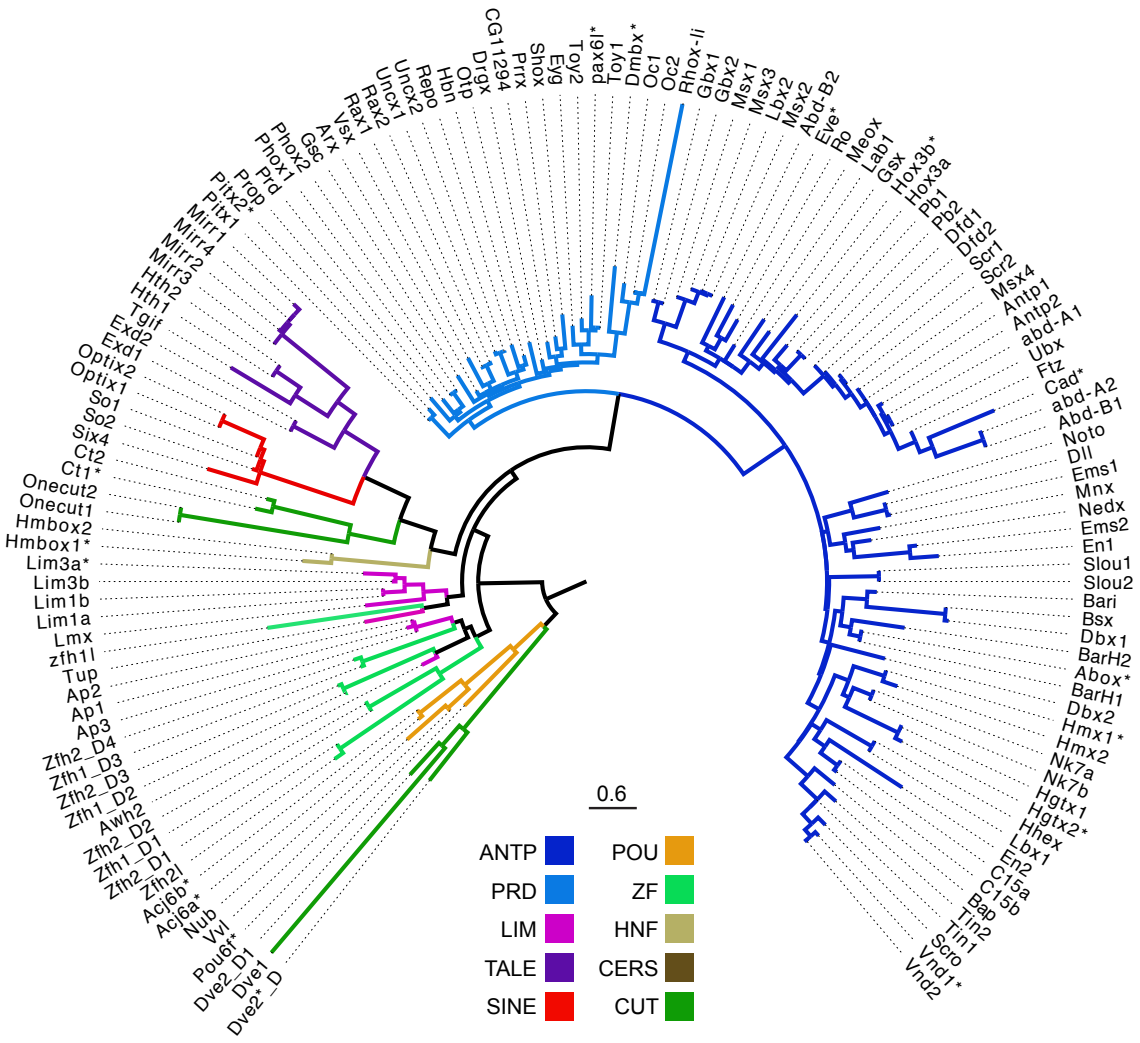
Appendix Figure 3.2: Maximum likelihood tree of the homeodomain from PRD class homeobox genes. RAxML with model LG; 100 bootstraps; CAT model of heterogeneity. Species in the tree: *P. tepidariorum*, *P. phalangioides*, *C. sculpturatus*, *M. mimosarum*, *I. scapularis*, *S. maritima*, *A. mellifera*, *T. castaneum*, *D. melanogaster*.



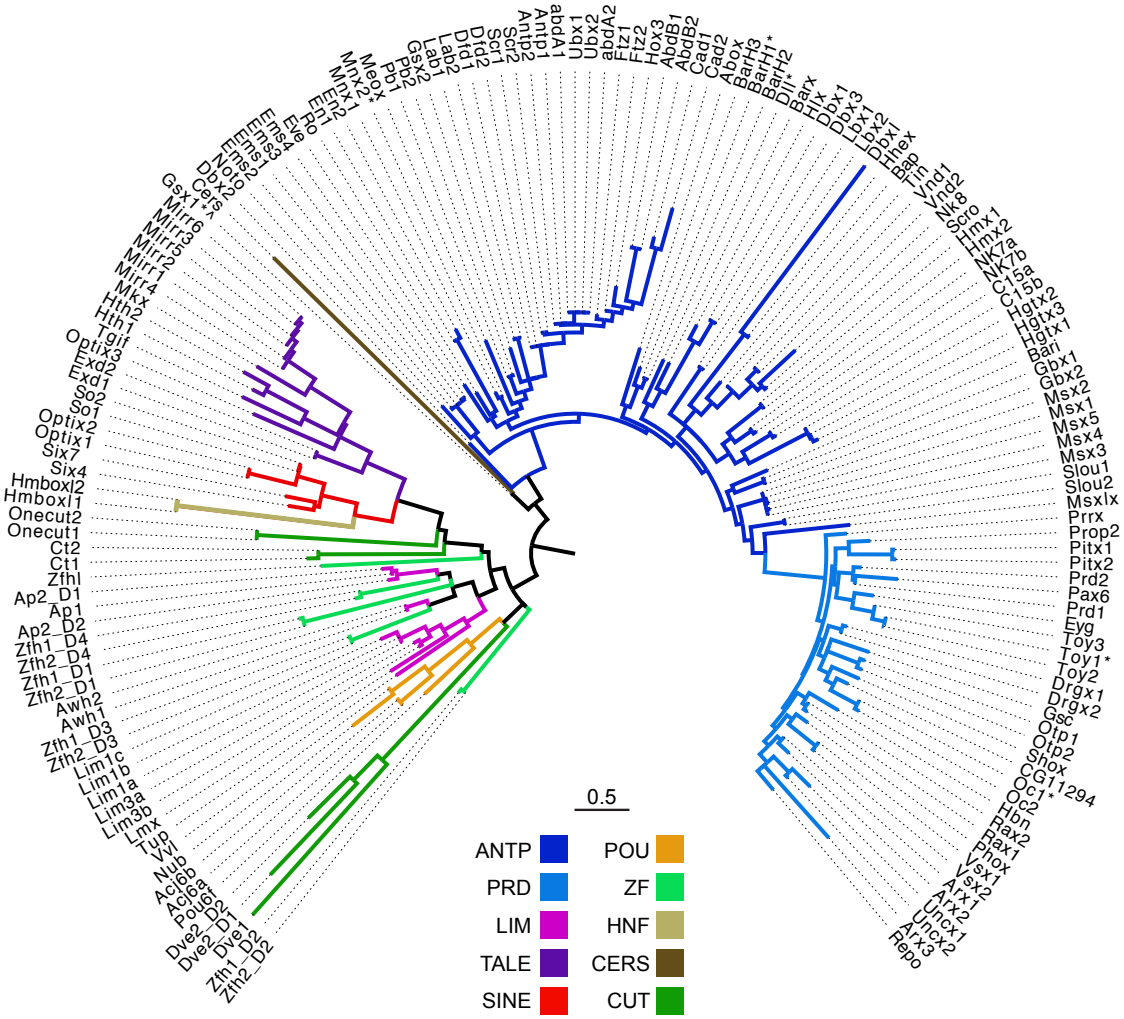
Appendix Figure 3.3: Maximum likelihood tree of homeodomains from homeobox genes, excluding ANTP and PRD classes. RAxML with model LG; 100 bootstraps; CAT model of heterogeneity. Species in the tree: *P. tepidariorum*, *P. phalangioides*, *C. scuplturatus*, *M. mimosarum*, *I. scapularis*, *S. maritima*, *A. mellifera*, *T. castaneum*, *D. melanogaster*.



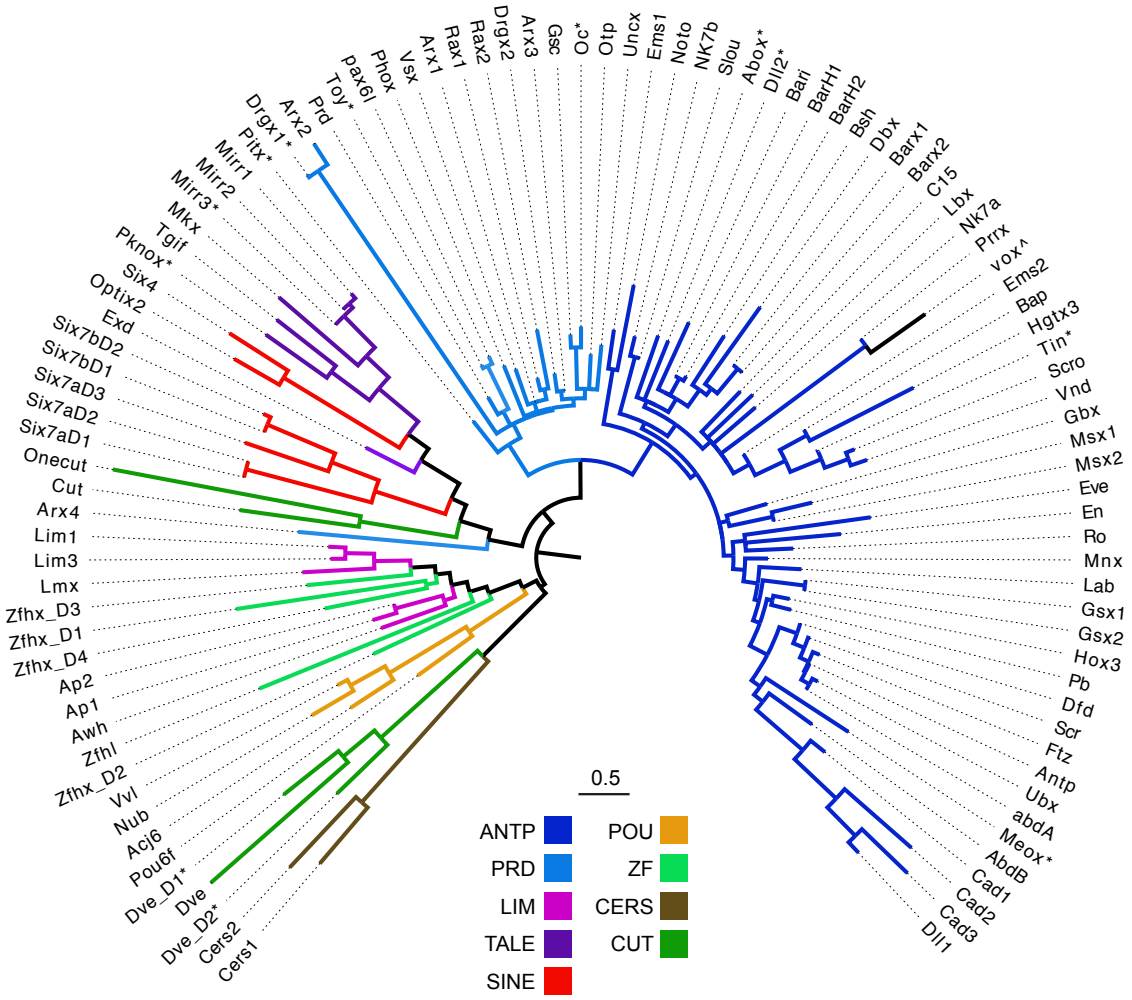
Appendix Figure 3.4: Maximum likelihood tree of *P. tepidariorum* homeobox genes. Asterisks indicate incomplete homeodomain sequence due to partial protein prediction. Carets indicate possible other homeodomain class. PhyML run with model LG; 500 bootstraps; gamma distribution estimated; with tree topology, branch length and rate parameters optimised.



Appendix Figure 3.5: Maximum likelihood tree of *P. phalangioides* homeobox genes. Asterisks indicate incomplete homeodomain sequence due to partial protein prediction. Carets indicate possible other homeodomain class. PhyML run with model LG; 500 bootstraps; gamma distribution 1.16; with tree topology, branch length and rate parameters optimised.



Appendix Figure 3.6: Maximum likelihood tree of *C. sculpturatus* homeobox genes. Asterisks indicate incomplete homeodomain sequence due to partial protein prediction. Carets indicate possible other homeodomain class. PhyML run with model LG; 500 bootstraps; gamma distribution 1.22; with tree topology, branch length and rate parameters optimised.



Appendix Figure 3.7: Maximum likelihood tree of *I. scapularis* homeobox genes. Asterisks indicate incomplete homeodomain sequence due to partial protein prediction. Carets indicate possible other homeodomain class. PhyML run with model LG; 500 bootstraps; gamma distribution 0.81; with tree topology, branch length and rate parameters optimised.

Appendix Table 3.5: MicroRNAs and non-Hox genes in *P. tepidariorum* Hox cluster scaffolds

CLUSTER A					
Scaffold	**AUG3 ID**	**Type**	**Coordinates**		**Strand**
Scf5bK3_833	aug3.g17092	Protein	12404	50863	+
Scf5bK3_833	aug3.g23293	Protein	69343	75684	+
Scf5bK3_833	aug3.g17627	Protein	106347	110612	-
Scf5bK3_833	aug3.g17626	Protein	118772	133977	+
Scf5bK3_833	aug3.g17625	Protein	162806	178288	+
Scf5bK3_833	aug3.g6147	Protein	218413	218847	+
Scf5bK3_833	aug3.g6146	Protein	275778	276122	+
Scf5bK3_833	aug3.g6145	Protein	301707	302738	-
Scf5bK3_833	aug3.g6144	Protein	307126	322226	+
Scf5bK3_833	aug3.g6143	Protein	337603	337836	-
Scf5bK3_833	aug3.g6142	Protein	338753	523542	-
Scf5bK3_833	aug3.g15160	Protein	806677	810378	-
Scf5bK3_833	aug3.g15159	Protein	859738	860286	-
Scf5bK3_833	aug3.g15158	Protein	895962	905313	-
Scf5bK3_833	aug3.g9050	Protein	946875	971066	-
Scf5bK3_833	aug3.g9049	Protein	978297	994389	+
Scf5bK3_833	aug3.g9047	Protein	1203192	1203605	+
Scf5bK3_833	*iab-4-2*	miRNA	1315270	1315330	-
Scf5bK3_833	aug3.g15592	Protein	1473721	1510079	+
Scf5bK3_833	aug3.g15593	Protein	1556746	1556958	+
Scf5bK3_833	aug3.g15594	Protein	1594525	1711382	-
Scf5bK3_833	aug3.g15596	Protein	1841263	1866959	+
Scf5bK3_833	aug3.g15597	Protein	1941513	1941968	-
Scf5bK3_833	aug3.g15598	Protein	1951980	1981315	-
Scf5bK3_833	aug3.g15599	Protein	2053402	2054890	-
Scf5bK3_833	aug3.g15601	Protein	2127888	2128124	+
Scf5bK3_833	aug3.g15602	Protein	2358645	2359130	-
Scf5bK3_833	aug3.g15603	Protein	2416290	2446242	+
Scf5bK3_833	aug3.g15604	Protein	2567754	2568197	-
Scf5bK3_833	aug3.g15605	Protein	2589922	2593240	-
Scf5bK3_833	aug3.g15607	Protein	2705969	2715250	-
Scf5bK3_833	aug3.g15609	Protein	3058522	3064693	+
Scf5bK3_833	aug3.g15610	Protein	3075444	3078817	-
Scf5bK3_833	aug3.g15611	Protein	3088456	3095404	+
Scf5bK3_833	aug3.g15612	Protein	3098432	3100932	+
Scf5bK3_833	aug3.g15613	Protein	3123444	3129408	+
Scf5bK3_833	aug3.g15614	Protein	3138062	3142524	+
Scf5bK3_833	aug3.g15615	Protein	3147681	3148019	+
Scf5bK3_833	aug3.g15616	Protein	3152276	3158397	+
Scf5bK3_833	aug3.g15617	Protein	3177377	3182828	+
Scf5bK3_833	aug3.g15618	Protein	3186877	3195781	+
Scf5bK3_833	aug3.g15619	Protein	3198540	3200902	+
Scf5bK3_833	aug3.g15620	Protein	3207654	3211392	+
Scf5bK3_833	aug3.g15621	Protein	3228174	3229451	+
Scf5bK3_833	aug3.g16686	Protein	3260059	3268113	+
Scf5bK3_833	aug3.g16685	Protein	3301933	3305549	+
Scf5bK3_833	aug3.g16684	Protein	3313854	3318699	+
Scf5bK3_833	aug3.g16683	Protein	3331226	3331591	-
Scf5bK3_833	aug3.g16682	Protein	3342414	3348804	+
Scf5bK3_833	aug3.g16681	Protein	3379940	3386593	+
Scf5bK3_833	aug3.g16680	Protein	3395972	3402247	+
Scf5bK3_833	aug3.g16679	Protein	3422017	3426263	+
Scf5bK3_833	aug3.g16490	Protein	3483712	3483771	-
Scf5bK3_833	aug3.g16491	Protein	3510303	3515677	+
Scf5bK3_833	aug3.g16492	Protein	3539019	3543371	-
Scf5bK3_833	*mir-993b-1*	miRNA	3580234	3580294	-
Scf5bK3_833	aug3.g17571	Protein	3732819	3733390	+
Scf5bK3_833	aug3.g17572	Protein	3737301	3737669	+
Scf5bK3_833	aug3.g7955	Protein	4228876	4248959	-
Scf5bK3_833	aug3.g7956	Protein	4255264	4259480	+
Scf5bK3_833	aug3.g7957	Protein	4263037	4274104	+
Scf5bK3_833	aug3.g21962	Protein	4284106	4287839	-
Scf5bK3_833	aug3.g26787	Protein	4302769	4337764	+
Scf5bK3_833	aug3.g26786	Protein	4339721	4344763	-
Scf5bK3_833	aug3.g26785	Protein	4353244	4386496	+
Scf5bK3_833	aug3.g26784	Protein	4476118	4478640	-
Scf5bK3_833	aug3.g26783	Protein	4600082	4603716	-
Scf5bK3_833	aug3.g26782	Protein	4665319	4666197	-
Scf5bK3_833	aug3.g26781	Protein	4693318	4739518	-
Scf5bK3_833	aug3.g26780	Protein	4748530	4754495	+
Scf5bK3_833	aug3.g26779	Protein	4777972	4789197	+
Scf5bK3_833	aug3.g26778	Protein	4842073	4851719	+
Scf5bK3_833	aug3.g26777	Protein	4868645	4870330	-
Scf5bK3_833	aug3.g21104	Protein	4905707	4923417	+

CLUSTER B					
Scaffold	**AUG3 ID**	**Type**	**Coordinates**		**Strand**
Scf5bK3_3652	aug3.g27132	Protein	4364	14419	-
Scf5bK3_3652	aug3.g27131	Protein	18899	54219	+
Scf5bK3_3652	aug3.g24186	Protein	194154	194351	-
Scf5bK3_3652	aug3.g27129	Protein	342794	356164	-
Scf5bK3_3652	aug3.g27128	Protein	425500	425817	-
Scf5bK3_3652	aug3.g27126	Protein	652138	677132	-
Scf5bK3_3652	aug3.g27125	Protein	695489	721552	+
Scf5bK3_497	aug3.g17307	Protein	11	122833	+
Scf5bK3_497	aug3.g23141	Protein	6902	7108	-
Scf5bK3_497	aug3.g17308	Protein	29298	99590	-
Scf5bK3_497	aug3.g17306	Protein	126667	132933	+
Scf5bK3_497	aug3.g17305	Protein	162001	189110	-
Scf5bK3_497	aug3.g25353	Protein	407286	422740	-
Scf5bK3_497	aug3.g25354	Protein	439278	517702	+
Scf5bK3_497	aug3.g25356	Protein	619530	620693	-
Scf5bK3_497	aug3.g25357	Protein	661526	661888	-
Scf5bK3_497	*iab-4-1*	miRNA	896391	896451	-
Scf5bK3_497	aug3.g25359	Protein	1362095	1413941	+
Scf5bK3_497	aug3.g1790	Protein	2245197	2280004	-
Scf5bK3_497	aug3.g1788	Protein	2279886	2337770	-
Scf5bK3_497	aug3.g1789	Protein	2323755	2324231	-
Scf5bK3_497	aug3.g1786	Protein	2412067	2412285	-
Scf5bK3_497	aug3.g24519	Protein	2514302	2514370	+
Scf5bK3_497	aug3.g17639	Protein	2592144	2592368	+
Scf5bK3_497	*mir-10*	miRNA	2682407	2682467	-
Scf5bK3_497	aug3.g17638	Protein	2676136	2720228	-
Scf5bK3_497	aug3.g14843	Protein	2975718	2983495	+
Scf5bK3_497	aug3.g14844	Protein	3078534	3086147	-

Appendix Table 4.1: *P. tepidariorum* hairpin miRNA sequences

miRNA	Sequence
pte-bantam	ctggttttcccaatgatcatccagaggaatatatgaaatctgagatcattgtgaaagctgatt
pte-iab-4-1	tcgtatactgaatgtatcctgataataattgactctccggtataccttcagtatacgtaac
pte-iab-4-2	tcgtatactgaatgtatcctgagtttaaattttccggtataccttcagtatacgtagc
pte-iab-8-1	tacgtatactgaaggtataccggagagtcaattattatcaggatacattcagtatacgaac
pte-iab-8-2	tacgtatactgaaggtataccggaaaatttaaactcaggatacattcagtatacgaac
pte-let-7	tgaggtagtaggttgtatagtttagaactacacctttaatgggcgaactatacagcttgctaactttcc
pte-mir-1	ccgtgcttccttacttcccatatcttcttaaagagtcgtatggaatgtaaagaagtatggag
pte-mir-10	taccctgtagatccgaatttgtggcaaaaagcataacaaattcggttctagagaggttc
pte-mir-100a	tacccgtagatccgaacttgtgttcacttgaactccacaagtttgattttacgggttga
pte-mir-100b	aacccgtagatccgaacttgtgtcctttgtagacctcccacaagctcgcctctacgggtct
pte-mir-125a	tccctgagaccctaacttgcgatggcccgggtctcccaggttagattctcaggtttt
pte-mir-125b	tccctgagaccctaacttgtgatggattcattgcttcgcacaagttagtgtctcgggcc
pte-mir-14	ggtagagatctgaaggcttacgttcattattcttcgtcagtctttgaatctctccccg
pte-mir-184-1	ccttatcattcgactgtccagactgtatttgcgaactggacggagaactgataagggc
pte-mir-184-2	ccttatcattctcgtgtccaggctgccggttgatcgagatctggacggagaactgataagggc
pte-mir-190	agatatgtttgatattcttggttgtgtcaatttgttcaaccaagaatcagacatatatgg
pte-mir-193a	cgggatttagctggactgatggactgacttcagaaattgtctactggcctgttaagtcccgag
pte-mir-193b	tatgggtcttgacagaacagacgattttataatctactggcctgctaagtcccaca
pte-mir-210	agctgctggccacgtcactagactcgattttagatcttgtgcgtgtgacagcggctat
pte-mir-263a	cttggcactgtaagaattcacagatggatatcaaatcgtggattctttggtgctaaaga
pte-mir-263b	aatggcactggaagaattcacggataatgtttgactccgtggtttctctggtgctataag
pte-mir-275	acgtgctgcatcaggtgcttgtgactgcgccctttcagataagtcaggtacctgaagtagcgcgcgga
pte-mir-276-1	agcgaggtatggagttcttactgtgaaagcagattggtaggaacttcataccatgctcg
pte-mir-276-2	agcgaggtatgaagttcttactgtttcatacagtaggtaggaacttcataccatgctcg
pte-mir-277a	cgttccagaagtgcgtttacattagtgtcttaagaaatctgtaaatgcattttctggaatgta
pte-mir-277b	cgttccagaggtgcattatacatagacatttgtaaatgcaccctctagaaccat
pte-mir-278a	ccggatgaaaattccatccgaccgtatagttgacaggccggtgggattgtcgtccgtct
pte-mir-278b	ccggaagagaattcccaccttaaccgttctatagcaaaaggtcggtgggattttcgtccgtct
pte-mir-279	ggtgagtgggattttagtcctgtatctaaaacactatcatgactagatccacactcatcca
pte-mir-281	aagagagctatccgtggacagtgatagttgatgataactgtcatggaattgctctctttta
pte-mir-29a	accgagttcggatggtggatagatgctattagaaagtctagcaccatttgaattcagttc
pte-mir-29b	ctgatctcagatggtggatagatggatcattgggaaatgtctagcaccatttgaaatcagcgc
pte-mir-2a	ctctcaaggtggctgttatatgtgtgttgaaataatatttcatatcacagctgctttgatgagtg
pte-mir-2b-1	gttgtcaaagtggctgtgacatgttagagatttatgcttcatatcacagccagctttgatgagcg
pte-mir-2b-2	ctcacaaagtgggtgtgtatatgttagtagaatttcatatcacagccagctttgatgagcg
pte-mir-2c	tcatcaaattggcagtgatctgttacatttgaatcatatcacagccatcttgatgatc
pte-mir-2d	gggtcaaactgatgctgtgatgtgtaaattcttcatcatatcacagccagctttgatgagc
pte-mir-2e	acgtcgaagtggctgtgaaatgtattcactttaccatatcacagctatctttgacgtgt
pte-mir-2f	cgtcgaatgtgctctgtgaaatgtgttgtttctccatatcatatcacagccacatttgatgcgc
pte-mir-2g-1	aaactgatgctgtgatgtgtaaattcaattatcatatcacagccagctttgatgag
pte-mir-2g-2	ttgcaaactgatactgtgatgtgtaaattcaaatatcatatcacagccagctttgatgag
pte-mir-305	tttgtacttcattaggtgctctggtgtaaacgagattccaggcatctaagggagtgcaaatg
pte-mir-315	ttttgattgttgctcagaaggcgtgcagacgtcaggctttcgagcaatgatcaaacgt
pte-mir-317	cgagagcctccctgcgttcacaagtgccgtttgctttctgtgaacacagctggtggtatctcagt
pte-mir-34	tggcagtgtggttagctggttgtgttttttctttcaacaaccactatcttctctgccct
pte-mir-3477	taatctcattcggtatggactgatgaatgtttataaaacagtcgtaccatatgagatatc
pte-mir-375-1	gacttgagctgctcgcacaaagctccattaaaatgttttgttcgttcggctcgagtta
pte-mir-375-2	actcgaactcttcgagcaaagctttattttctgttttgttcgttcggctcgagtta
pte-mir-3791a	gcaagtttttcaccttggttatttatatttgttaatcaccgggtggaaaatttgagc
pte-mir-3791b	gcaagtttttcaccttggtaatttatatttgttaatcaccgggtggaaaatttgaac
pte-mir-3791c	caaaatttttcgctgcggtggtttagatttgaaaatcaccgggcgaaaagtttgctt
pte-mir-3791d	tgaaatttttcgctgcggtagtttatatttgaaaatcaccgggtgaaaaatttgcct
pte-mir-3791e-1	tcgaatctgtcacttcggtaatttcaattcaataatcaccgggtgaaagatttgctt
pte-mir-3791e-2	tcgaatctgtcacttcggtaatttcgattcaataatcaccgggtgaaagatttgctt
pte-mir-3791f	tggaatttttcgctgcggtggtttatatttgaaaatcaccgggtgaaaaatttcttt
pte-mir-3791g	tgaaatttttcgctgcggtggtttgtatatttgaaaatcaccgggtgaaaagtttgctt
pte-mir-3791h	caaaatttttcgttgcggtggtttacatttgataatcaccgggtgaaaaatttgctt
pte-mir-3791i	caaaatttttcaccgcggtagtttgaatttaaaaatcaccgggtgaaaattttgttt
pte-mir-3791j-1	tcgagtctttcgcttcggtaatttataatgaataatcaccgggtgaaagattcgctt
pte-mir-3791j-2	tcgagtctttcgcttcggtaatttataatgaataatcaccgggtgaaagattcgctt
pte-mir-3791k	ttgaatccttcacacggtaatttacatctgataatcaccgggtgatcgatttgctt
pte-mir-3791l	agcaattaatggctcggtggcagtgcctcgcctgacaccgggtcgtttatttgcttt
pte-mir-3791m	agcaatcaatggctcggtggcagtgtcttgcctgacaccgggtagttgatttgcttt
pte-mir-3791n-1	tcgaatctttctctacggtaatttcttatttaaaaatcaccgggagaatagatttgctt
pte-mir-3791n-2	tcgaatctttctctacggtaatttcttatttaaaaatcaccgggagaatagatttgctt
pte-mir-3791o-1	agttgttaatagctcggtgatggagaattacccttcaccgggaggttaatatgcttt
pte-mir-3791o-2	agctattaatggctcggtgatagaggattgcacgtcaccgggaggttaatatgcttt
pte-mir-3791p	ggaaaaatgtgttacatcggtaatttttatgcaaatcaccgggtgagacattttgcat
pte-mir-3791q-1	cgaaatagttttatctggttactcaaaattaaaattgtcaccggatgaactattttggtt
pte-mir-3791q-2	acgaaatagttttatctggttactcaaaattaaaattgtcaccggatgaactattttggtt
pte-mir-3791q-3	acgaaatagttttatctggttactcaaaattaaaattgtcaccggatgaactattttggtt
pte-mir-3791r-1	agaaatttttcatcccagttaattttatttttgaattaaatcactgggaggaaagtttctat
pte-mir-3791r-2	agaaatttttcatcccagttaattttatttttgaattaaatcactgggaggaaagtttctat
pte-mir-3791r-3	agaaatttttcatcccagttaattttatttttgaattaaatcactgggaggaaagtttctat
pte-mir-3791s	ggttgaatgttcctggtgttgcattctttaaacttgcatcaccgggtaaacattcatcccc
pte-mir-3791t	aggtattaatagcccggtggcagagtcgtaactgtcaccaggagattaatatgctt
pte-mir-3791u	agcaattaatggctcggtggcagtgtctcgcctgacaccaagtcgtttatttgcttt
pte-mir-3791v	agcaattaatggctcggtggcagtgtctcgcctgacaccaagtcgtttctttgcttt
pte-mir-3791w	agcaatgaatggcttggtggcagtgtctcgcctgacaccagttcgtttatttgcttt
pte-mir-3791x	agcaatgaatggctcggtggcagtgtctcgcctgacaccagttcgtttatttgcttt
pte-mir-3791y	agcaatgaatggcttggtggcagtgactcgcctgacaccaattcgttcatttgcttt
pte-mir-3791z	agaatatttcatacggttgacttgtatttttaaacaaaatcatcgtttaaaatattcatt
pte-mir-3931	atttgtaccgattcaatgatagctttataaaacgaaactactttgagtcggtacgaatca
pte-mir-7	tggaagactagtgattttattgtttgatattcaatttaaacaatagatcactgtctccttt
pte-mir-71-1	tgaaagacatgggtagtgagatgtcttaattccccatctcactaccttgtctttgacg
pte-mir-71-2	tgaaagacatgggtagtgagatgtgtgtaaatagtttcatctcactaccttgtctttgacg
pte-mir-745	agtccatcatcgggttgctagctttttttataattgccagctgccctgtgaagggctgt
pte-mir-8	catcttaccaggcagtattagagtcaatttgaaaatgttcttctaatactgtcaggtaaagatgtc

Appendix Table 4.1: *P.tepidariorum* hairpin miRNA sequences …continued

MicroRNA	Sequence
pte-mir-87a	ccacctgaaaatttgtctcaacctcgtcattcaaaataggtgagcaaagtttcaggtgcct
pte-mir-87b	cagcccggagccttgtctcaacatgtgatgtagtattcagttgagcaaagtttcaggcgaat
pte-mir-87c	gaacctgaaattttgtctcagccggatattttaaactgttgagcaaagtttcaggtgtat
pte-mir-9	tctttggttatctagctgtatgagtgagttttacgtcataaagctaggttaccaaagtta
pte-mir-92a	aaggtcggggtgattggcagtattgtatatgatgcaactaatattgcactcgtcccggcctaa
pte-mir-92b-1	aggtcgggatgactggcagtatttgttctattcaacgaatattgcactcgtcccggccta
pte-mir-92b-2	aggtcgggatgactggcagtattgttcttttcaactaatattgcactcgtcccggccta
pte-mir-92c	agatcaggatgagtggcaatattgctatatgatgccacgaatattgcacttgtcccggtctga
pte-mir-96	tttggcactgtcacatttttgtgtgtctataattctacaaaaattgtgatagtgtcaagc
pte-mir-981	cgggattcacgacgaccgagcactctaaaaaagatgttcgttgtcgttgaaacctgaa
pte-mir-993a	gggacctgtagaatcgggctactgtgtgtttaacaattcagaagctcgtctctacaggaacgt
pte-mir-993b-1	ctaccctgtagatccgggcttttgtagactgcttagcattcagaagctcgtctctacaggtatct
pte-mir-993b-2	ctaccctgtagatccgggcttttgtagtttgtgaatagcattcagaagctcgtatttacaggtatct
pte-mir-11938	tgtctaagaagtttaagtgtccagtgcataatgaaccaggcatctttacttctcagacata
pte-mir-11939	tgttgcttcatatgatctaaagtacttaaaagtcttaagtactagattatgaagcagcaga
pte-mir-11940	tccgtcactgttactactgctaacctaacgttattagaaagtaatagtggcaagt
pte-mir-11941	tgtttaactgcgagtcctgtttatgtttttgaaatttaaacattcttggcagttgagcatt
pte-mir-11942a	cgttaagactgtcagtgtttggctgtgttatattttctgccaaatactgacagtcttaccaa
pte-mir-11942b	tgctaagactgtcagtgtttggtgaattttatttcaaacactggcagtcttgatgg
pte-mir-11942c	tgttaagactgtcagtgtttggtttctatcaattctcaaacactggcagtcttaacaaa
pte-mir-11942d	tgttaagactgtcagtgtttgggtttaaataagtttcaaacactgacagtcttgacaga
pte-mir-11943	ttgcaaagtgaatcgtcagctttaatctcaagtttgttatggtgtacttcaagatgct
pte-mir-11944	cctcgaactctacatggcgccaatactttcagatattggcgccatgtagtgttcgcggac
pte-mir-11945	agcaacatattcactggacggctgatccattcctcaaccatccagtagatatgttgctcgc
pte-mir-11946	tgccacactgaagattttactgactaatttgctgcgtgaactcattcagtgtggcatg
pte-mir-11947a-1	atgtgttcatagtctgttagttagcgtctgcgctaaattaacagactatgaacacattt
pte-mir-11947a-2	atgtgttcatagtctgttagttagcgtctccgctaaattaacagactatgaacacattt
pte-mir-11947b-1	acattcctagcatgttggttagcgtcttagctaaatcagcatggtaggaatgtattt
pte-mir-11947b-2	acattcctagcatgttggttagcgtcatagctaaatcagcatggtaggaatgtattt
pte-mir-11948	tatacgacttttcaacaacttggttgataacgcaaaagttgtcgaacaagtcgtatagt
pte-mir-11949	cgatcttactggaacacaagttttatattgataaacttgtgtcgcagtaagatact
pte-mir-11950	aggcaaagctcctcttcagtaaattatcatgattttaccgggaggtacttggcttaa
pte-mir-11951a	aaagctcttgggaaattatgttctttatttgcgtaacatgttttctccagagcttatc
pte-mir-11951b	aaagctcttgggaaacaatgttgtaagagttaacaaacattttatctcgagggcttatg
pte-mir-11952	gcaatatcgggtttgaacgctgcttgctagataagtgtgtaacgttaacaccgatattgccc
pte-mir-11953	ttctgttcagtgttagtctgagatgttagttttagctttcactgaatggcagc
pte-mir-11954	tgatgtactgagactccgattccaacagctgtttacgaattccagaaccagagtcccagtacaccaaccca
pte-mir-11955	ctgtacgagaggaccttctgaacagagactgttcagaaggtcctctcgtacag
pte-mir-11956	agctcattgagctaaacagaaacagttcataagaatcttcttatgaactgtttctgtttagcttaatgagctct
pte-mir-11957	atgtctcaatggcatagtggttaggacgagcgttaaaccactatgccattgagacattt
pte-mir-11958	gatggttgactccggttatccgagatcccagagaaaaccggagtcatccctcttt
pte-mir-11959-1	gagggtaaactaagcatatccgaaaatcttcggataagctcggttgatcctattt
pte-mir-11959-2	gagggtaaactaagcatatccgaaaatcttcggataagctcggttgatcctattt
pte-mir-11960	atggtaaattgttcactcagatagcaatctgcgtgaacaatttcccatgc
pte-mir-11961a	ggtcagtatctgctgtgatacgggtggcgcagttattcgtatcacaatagtgctagacatt
pte-mir-11961b	gggcggtttctgctgtgaaacgacgattgcagttttccgtttcacagtagaactagctatt
pte-mir-11961c	ggtcggtttctgctgtgagacggcggtcgctgtcgtccgtatcacagcagaatttgacattt
pte-mir-11962a	catatttcagaaatggcatcaagtaattttcacaactgatgtcatttttgaagtatgca
pte-mir-11962b	catatttcaaaaatggcatcacgtcttgctaatatctgatgccatttttgagctatgta
pte-mir-11963a	gtgttttcatggtacaacggttaagcgtgctgctcaatcgctgtgccatgaaaacattt
pte-mir-11963b-1	atgtgttcatggcacaacggttagcgcgtcggttgaaaccgttgtaccatgaaaacattt
pte-mir-11963b-2	atgtgttcatggcacaacggttagcacgtcggctataaaccgttgtaccatgaaaacattt
pte-mir-11964	agataaatgtatatctggtcttgtgaaattaataatcatgactagatattccatttatct
pte-mir-11965-1	gataaaggaatttttgttcatatgtaaatatgaacacaaattccttcatctc
pte-mir-11965-2	gatgaaggaatttgtgttcatagttacatatgaacacaaattccttcatctc
pte-mir-11965-3	agatgaaggaatttgggttcatatgtaaatatgatatgaacacaaattccttcatctc
pte-mir-11966	cccaacctaatatttatttctgagttcaaaaattgatctcagaaataaatattgggttgggt
pte-mir-11967	agatgctgcgctatgagttattttgttttaaataataactcaaagtcgtagcttcggt
pte-mir-11968-1	agggacattgactagcaagctgttgttatgttttatacccagctttctggtaatagtctccct
pte-mir-11968-2	agggacattgactagcaagctgttgttatattttatgcccagctttctggtaatagtctccct

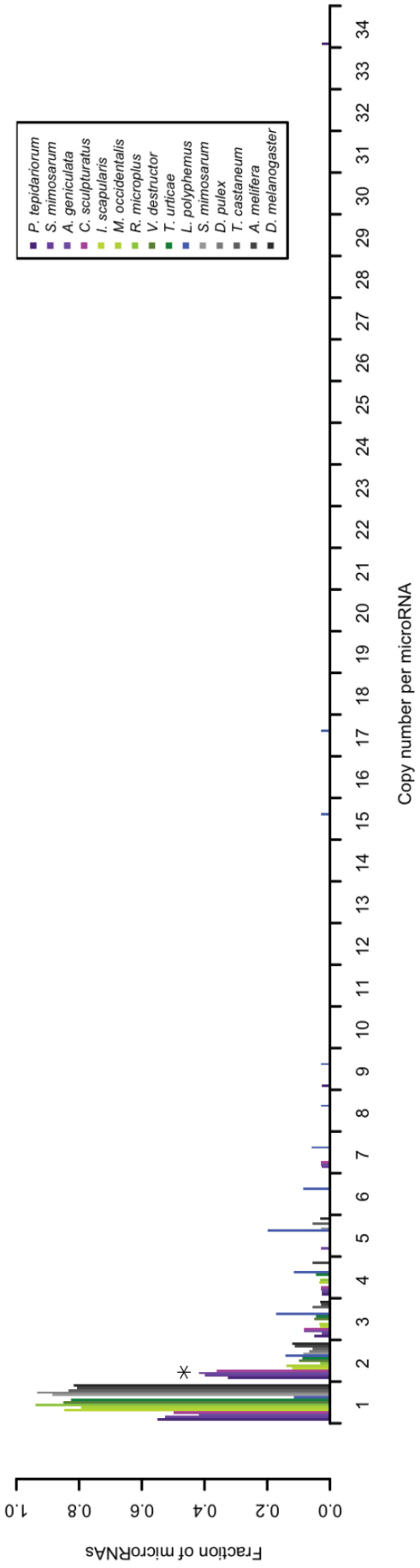
Appendix Table 4.2: *P. tepidariorum* mature miRNA sequences

MicroRNA	5' sequence	3' sequence
bantam	ctggttttcccaatgatcatcca	tgagatcattgtgaaagctgatt
iab-4-1	tcgtatactgaatgtatcctga	cggtataccttcagtatacgtaac
iab-4-2	tcgtatactgaatgtatcctga	cggtataccttcagtatacgtagc
iab-8-1	tacgtatactgaaggtataccgga	aggatacattcagtatacgaac
iab-8-2	tacgtatactgaaggtataccgga	aggatacattcagtatacgaac
let-7	tgaggtagtaggttgtatagtt	ctatacagcttgctaactttcc
mir-1	ccgtgcttccttacttcccata	tggaatgtaaagaagtatggag
mir-10	taccctgtagatccgaatttgt	caaattcggttctagagaggttc
mir-100a	tacccgtagatccgaacttgtg	caagtttgattttacgggttga
mir-100b	aacccgtagatccgaacttgtg	acaagctcgcctctacgggtct
mir-125a	tccctgagaccctaacttgcga	ccaggttagattctcaggtttt
mir-125b	tccctgagaccctaacttgtg	caagttagtgtctcgggcc
mir-184-1	ccttatcattcgactgtccaga	tggacggagaactgataagggc
mir-184-2	ccttatcattctcgtgtccagg	tggacggagaactgataagggc
mir-190	agatatgtttgatattcttggttg	accaagaatcagacatatatgg
mir-193a	cgggatttagctggactgatgga	tactggcctgttaagtcccgag
mir-193b	tatgggtcttgacagaacagac	tactggcctgctaagtcccaca
mir-210	agctgctggccacgtcactaga	ttgtgcgtgtgacagcggctat
mir-263a	cttggcactgtaagaattcacaga	gtggattctttggtgctaaaga
mir-263b	aatggcactggaagaattcacgg	gtggtttctctggtgctataag
mir-275	acgtgctgcatcaggtgcttgtgact	tcaggtacctgaagtagcgcgcgga
mir-276-1	agcgaggtatggagttcttact	taggaacttcataccatgctcg
mir-276-2	agcgaggtatgaagttcttact	taggaacttcataccatgctcg
mir-277a	cgttccagaagtgcgtttaca	taaatgcattttctggaatgta
mir-277b	cgttccagaggtgcattatac	taaatgcaccctctagaaccat
mir-278a	ccggatgaaaattccatccgacc	ccggtgggattgtcgtccgtct
mir-278b	ccggaagagaattcccaccttaacc	tcggtgggattttcgtccgtct
mir-3791a	gcaagtttttcaccttggttatt	tcaccgggtggaaaatttgagc
mir-279	ggtgagtgggattttagtcctg	tgactagatccacactcatcca
mir-3791b	gcaagtttttcaccttggtaatt	tcaccgggtggaaaatttgaac
mir-281	aagagagctatccgtggacagt	tgtcatggaattgctctcttta
mir-29a	accgagttcggatggtggataga	tagcaccatttgaattcagttc
mir-29b	ctgatctcagatggtggataga	tagcaccatttgaaatcagcgc
mir-2a	ctctcaaggtggctgttatatg	tatcacagctgctttgatgagtg
mir-2b-1	gttgtcaaagtggctgtgacatg	tatcacagccagctttgatgagcg
mir-2b-2	ctcacaaagtgggtgtgtatatg	tatcacagccagctttgatgagcg
mir-2c	tcatcaaattggcagtgatctg	tatcacagccatcttgatgatc
mir-2d	gggtcaaactgatgctgtgatgtg	tatcacagccagctttgatgagc
mir-2e	acgtcgaagtggctgtgaaatg	tatcacagctatctttgacgtgt
mir-2f	cgtcgaatgtgctctgtgaaatg	tatcacagccacatttgatgcgc
mir-2g-1	aaactgatgctgtgatgtg	tatcacagccagctttgatgag
mir-2g-2	ttgcaaactgatactgtgatgtg	tatcacagccagctttgatgag
mir-11938	tgtctaagaagtttaagtgtcc	gcatctttacttctcagacata
mir-305	tttgtacttcattaggtgctc	aggcatctaagggagtgcaaatg
mir-315	ttttgattgttgctcagaaggcg	ctttcgagcaatgatcaaacgt
mir-317	cgagagcctccctgcgttcaca	tgaacacagctggtggtatctcagt
mir-34	tggcagtgtggttagctggttgt	aaccactatcttctctgccct
mir-3791c	caaaatttttcgctgcggtggtt	tcaccgggcgaaaagtttgctt
mir-3791o-1	agttgttaatagctcggtgat	tcaccgggaggttaatatgcttt
mir-3791q-1	cgaaatagttttatctggttact	caccggatgaactattttggtt
mir-3791t	aggtattaatagcccggtggc	tcaccaggagattaatatgctt
mir-3791n-1	tcgaatctttctctacggtaatt	tcaccgggagaatagatttgctt
mir-3791n-2	tcgaatctttctctacggtaatt	tcaccgggagaatagatttgctt
mir-3791k	ttgaatccttcacacggtaatt	tcaccgggtgatcgatttgctt
mir-3791d	tgaaatttttcgctgcggtagtt	tcaccgggtgaaaaatttgcct
mir-3791j-1	tcgagtctttcgcttcggtaatt	tcaccgggtgaaagattcgctt
mir-3791j-2	tcgagtctttcgcttcggtaatt	tcaccgggtgaaagattcgctt
mir-3791e-1	tcgaatctgtcacttcggtaatt	tcaccgggtgaaagatttgctt
mir-3791f	tggaatttttcgctgcggtggtt	tcaccgggtgaaaaatttcttt
mir-3791g	tgaaatttttcgctgcggtggtt	tcaccgggtgaaaagtttgctt
mir-3791h	caaaatttttcgttgcggtggtt	tcaccgggtgaaaaatttgctt
mir-3791i	caaaatttttcaccgcggtagtt	tcaccgggtgaaaattttgttt
mir-3791s	ggttgaatgttcctggtgttg	tcaccgggtaaacattcatcccc
mir-3791m	agcaatcaatggctcggtggc	acaccgggtagttgatttgcttt
mir-3791z	agaatatttcatacggttgactt	atcatcgtttaaaatattcatt
mir-3791u	agcaattaatggctcggtggc	acaccaagtcgtttatttgcttt
mir-3791e-2	tcgaatctgtcacttcggtaatt	tcaccgggtgaaagatttgctt
mir-3791p	ggaaaaatgtgttacatcggtaa	tcaccgggtgagacattttgcat
mir-3791w	agcaatgaatggcttggtggc	acaccagttcgtttatttgcttt
mir-3791l	agcaattaatggctcggtggc	acaccgggtcgtttatttgcttt
mir-3791x	agcaatgaatggctcggtggca	acaccagttcgtttatttgcttt
mir-3791y	agcaatgaatggcttggtggc	acaccaattcgttcatttgcttt
mir-3791v	agcaattaatggctcggtggc	acaccaagtcgtttctttgcttt
mir-3791r-1	agaaatttttcatcccagttaatt	tcactgggaggaaagtttctat
mir-3791r-2	agaaatttttcatcccagttaatt	tcactgggaggaaagtttctat
mir-3791r-3	agaaatttttcatcccagttaatt	tcactgggaggaaagtttctat
mir-3791q-2	acgaaatagttttatctggttac	caccggatgaactattttggtt
mir-3791o-2	agctattaatggctcggtgat	tcaccgggaggttaatatgcttt
mir-3791q-3	acgaaatagttttatctggttac	caccggatgaactattttggtt
mir-375-1	gacttgagctgctcgcacaaagc	tttgttcgttcggctcgagtta
mir-375-2	actcgaactcttcgagcaaagc	tttgttcgttcggctcgagtta
mir-3931	atttgtaccgattcaatgatagct	tactttgagtcggtacgaatca
mir-7	tggaagactagtgattttattgtt	caatagatcactgtctccttt
mir-71-1	tgaaagacatgggtagtgagatg	tctcactaccttgtctttgacg
mir-71-2	tgaaagacatgggtagtgagatg	tctcactaccttgtctttgacg
mir-745	agtccatcatcgggttgctagc	cagctgccctgtgaagggctgt
mir-8	catcttaccaggcagtattaga	taatactgtcaggtaaagatgtc
mir-87a	ccacctgaaaatttgtctcaacct	Gtgagcaaagtttcaggtgcct

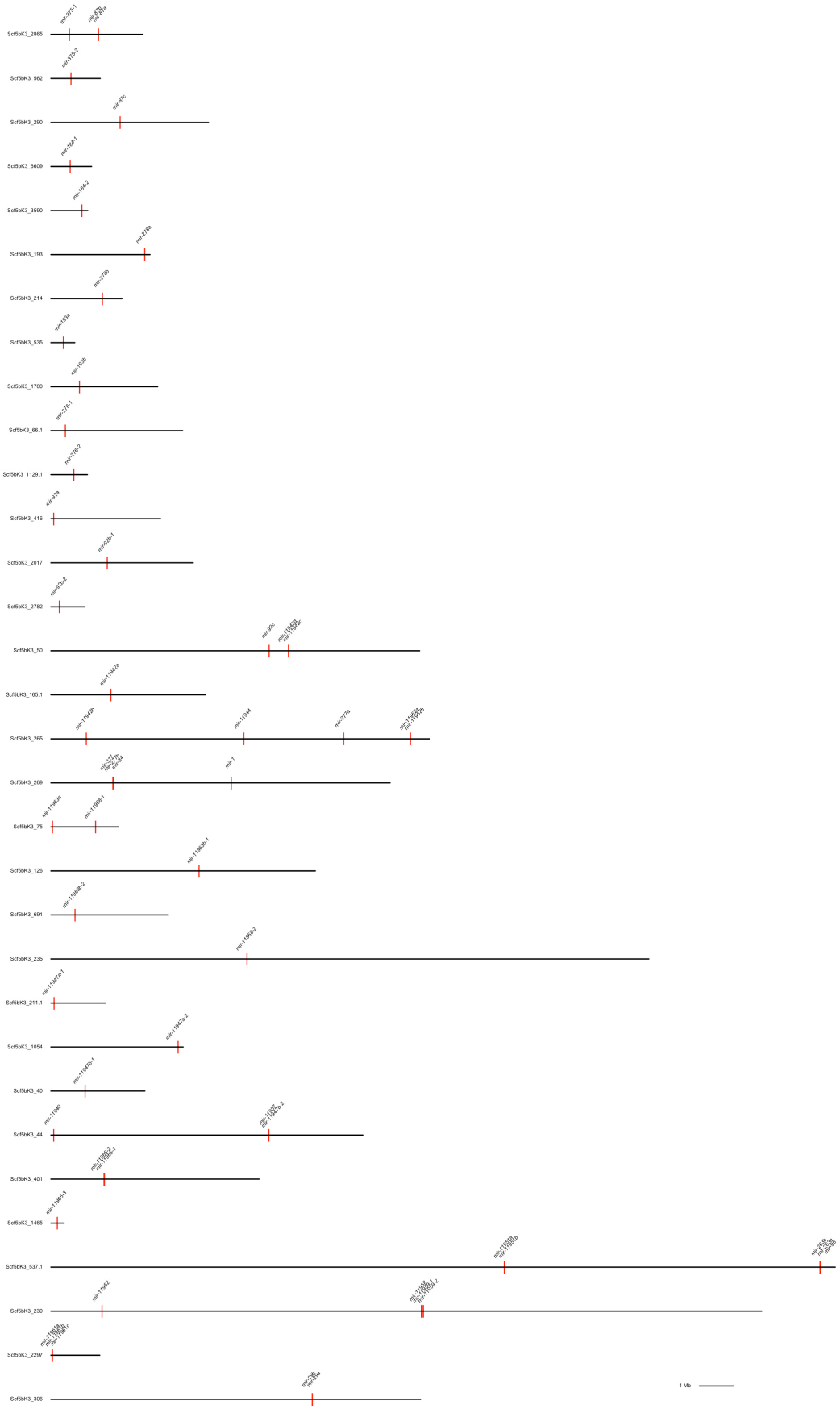
Appendix Table 3.2: *P. tepidariorum* mature miRNA sequences

…continued

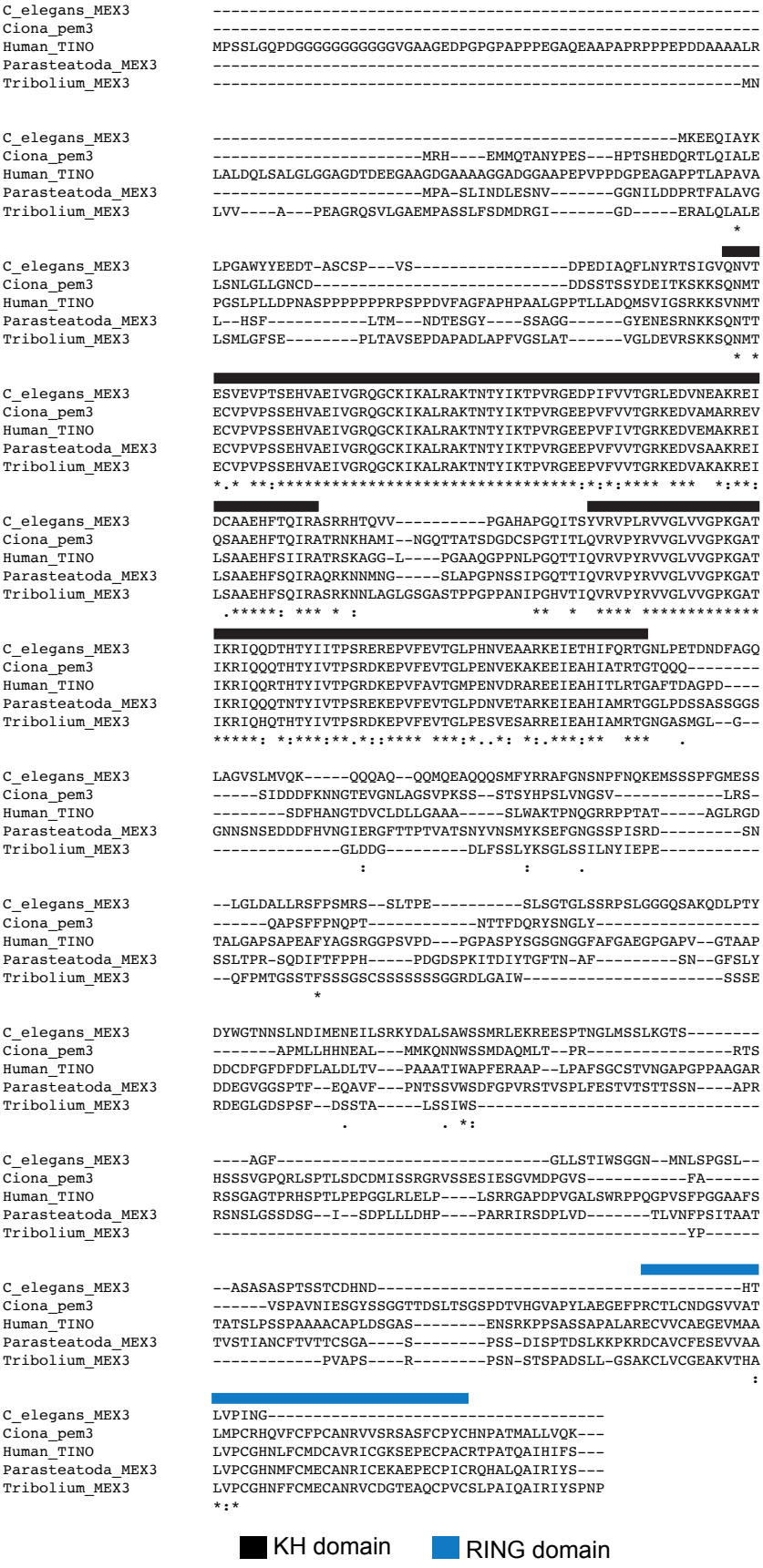
MicroRNA	5′ sequence	3′ sequence
mir-87b	cagcccggagccttgtctcaac	ttgagcaaagtttcaggcgaat
mir-87c	gaacctgaaattttgtctcagcc	ttgagcaaagtttcaggtgtat
mir-9	tctttggttatctagctgtatga	taaagctaggttaccaaagtta
mir-92a	aggtcggggtgattggcagtatt	tattgcactcgtcccggcctaa
mir-92b-1	aggtcgggatgactggcagtatt	tattgcactcgtcccggccta
mir-92c	agatcaggatgagtggcaatatt	tattgcacttgtcccggtctga
mir-92b-2	aggtcgggatgactggcagtatt	tattgcactcgtcccggccta
mir-96	tttggcactgtcacatttttgt	aaaaattgtgatagtgtcaagc
mir-981	cgggattcacgacgaccgagc	ttcgttgtcgttgaaacctgaa
mir-993a	gggacctgtagaatcgggctact	gaagctcgtctctacaggaacgt
mir-993b-1	ctaccctgtagatccgggcttttg	gaagctcgtctctacaggtatct
mir-993b-2	ctaccctgtagatccgggcttttg	gaagctcgtatttacaggtatct
mir-11939	tgttgcttcatatgatctaaagtact	tactagattatgaagcagcaga
mir-11940	tccgtcactgttactactgct	tagaaagtaatagtggcaagt
mir-11941	tgtttaactgcgagtcctgttt	acattcttggcagttgagcatt
mir-11942a	cgttaagactgtcagtgtttggc	caaatactgacagtcttaccaa
mir-11942b	tgctaagactgtcagtgtttgg	aaacactggcagtcttgatgg
mir-11942c	tgttaagactgtcagtgtttggt	aaacactggcagtcttaacaaa
mir-11942d	tgttaagactgtcagtgtttggg	aaacactgacagtcttgacaga
mir-11943	ttgcaaagtgaatcgtcagctt	ttatggtgtacttcaagatgct
mir-14	ggtagagatctgaaggcttac	tcagtcttttgaatctctccccg
mir-11944	cctcgaactctacatggcgcca	ggcgccatgtagtgttcgcggac
mir-11945	agcaacatattcactggacggc	catccagtagatatgttgctcgc
mir-11946	tgccacactgaagattttactg	cgtgaactcattcagtgtggcatg
mir-11947a-1	atgtgttcatagtctgttagtta	ttaacagactatgaacacattt
mir-11947a-2	atgtgttcatagtctgttagtta	ttaacagactatgaacacattt
mir-11948	tatacgacttttcaacaactt	gttgtcgaacaagtcgtatagt
mir-11949	cgatcttactggaacacaagtt	acttgtgtcgcagtaagatact
mir-11950	aggcaaagctcctcttcagtaaa	taccgggaggtacttggcttaa
mir-11951a	aaagctcttgggaaattatgtt	catgttttctccagagcttatc
mir-11951b	aaagctcttgggaaacaatgtt	cattttatctcgagggcttatg
mir-11952	gcaatatcgggtttgaacgctgc	aacgttaacaccgatattgccc
mir-3477	taatctcattcggtatggactga	cagtcgtaccatatgagatatc
mir-11953	ttctgttcagtgttagtctgaga	ttagctttcactgaatggcagc
mir-11954	tgatgtactgagactccgattc	agagtcccagtacaccaaccca
mir-11955	ctgtacgagaggaccttctgaa	ttcagaaggtcctctcgtacag
mir-11956	agctcattgagctaaacagaaa	tctgtttagcttaatgagctct
mir-11947b-1	acattcctagcatgttggtta	tcagcatggtaggaatgtattt
mir-11947b-2	acattcctagcatgttggtta	tcagcatggtaggaatgtattt
mir-11957	atgtctcaatggcatagtggtt	ccactatgccattgagacattt
mir-11958	gatggttgactccggttatc	aaaaccggagtcatccctcttt
mir-11959-1	gagggtaaactaagcatatc	ataagctcggttgatcctattt
mir-11959-2	gagggtaaactaagcatatc	ataagctcggttgatcctattt
mir-11960	atggtaaattgttcactcagat	ctgcgtgaacaatttcccatgc
mir-11961a	ggtcagtatctgctgtgatacg	tatcacaatagtgctagacatt
mir-11961b	gggcggtttctgctgtgaaacg	tttcacagtagaactagctatt
mir-11961c	ggtcggtttctgctgtgagacg	tatcacagcagaatttgacattt
mir-11962a	catatttcagaaatggcatcaa	gatgtcatttttgaagtatgca
mir-11962b	catatttcaaaaatggcatcac	gatgccatttttgagctatgta
mir-11965-3	agatgaaggaatttgggttc	tgaacacaaattccttcatctc
mir-11963a	gtgttttcatggtacaacggtta	tcgctgtgccatgaaaacattt
mir-11963b-1	atgtgttcatggcacaacggtt	ccgttgtaccatgaaaacattt
mir-11963b-2	atgtgttcatggcacaacggtt	ccgttgtaccatgaaaacattt
mir-11964	agataaatgtatatctggtct	tgactagatattccatttatct
mir-11965-1	gataaaggaatttttgttcata	tgaacacaaattccttcatctc
mir-11965-2	gatgaaggaatttgtgttcat	tgaacacaaattccttcatctc
mir-11966	cccaacctaatatttatttctg	agaaataaatattgggttgggt
mir-11967	agatgctgcgctatgagttatt	taactcaaagtcgtagcttcggt
mir-11968-1	agggacattgactagcaagctg	gctttctggtaatagtctccct
mir-11968-2	agggacattgactagcaagctg	gctttctggtaatagtctccct



Appendix Figure 4.1: Distribution of number of miRNAs per conserved miRNA family. Spiders and scorpions display many duplicated miRNA families relative to other species (*). *L. polyphemus* has families with from one to 17 copies.



Appendix Figure 4.2: Dispersed paralogs of miRNAs in the *P. tepidariorum* DoveTail genome assembly.



Appendix Figure 5.1: MEX3 alignment from *P. tepidariorum*, *T. castaneum*, *C. elegans*, *Ciona savignyi* and humans.

```
>aug3.g25290.t1
          Length = 361

 Score =  324 bits (831), Expect = e-109,   Method: Compositional matrix adjust.
 Identities = 185/365 (50%), Positives = 243/365 (66%), Gaps = 50/365 (13%)

Query: 9   GLNVTLTIRLLMHGKEVGSIIGKKGESVKRIREESGARINISEGNCPERIITLTGPTNAI 68
           G +VTLTIRL+M GKEVGSIIGKKG+++K+ REESGA+INIS+G+CPERI+T+TG T+AI
Sbjct: 11  GPSVTLTIRLIMQGKEVGSIIGKKGDNIKKFREESGAKINISDGSCPERIVTVTGCTDAI 70

Query: 69  FKAFAMIIDKLEEDINSSMTNSTAASRPPVTLRLVVPATQCGSLIGKGGCKIKEIRESTG 128
            KAF++I  K EED+ ++ T      +PPVTLRL+VPA+QCGSLIGKGG KIKEIRE TG
Sbjct: 71  LKAFSLIARKFEEDLQNT-TAGINVPKPPVTLRLIVPASQCGSLIGKGGSKIKEIREVTG 129

Query: 129 AQVQVAGDMLPNSTERAITIAGVPQSVTECVKQICLVMLETLSQSPQGRVMTIPYQPMPA 188
           A +QVA +MLPNSTERA+T++G  +++++C+ QIC VM+E+    P+G   TIPY+P PA
Sbjct: 130 ASIQVASEMLPNSTERAVTLSGTAEAISKCIYQICCVMMES---PPKG--ATIPYRPKPA 184

Query: 189 SSPVICAGGQDRCSDAVGYPHATHDLEGPPLDAYSIQGQHTISPLD----------LAKL 238
             PVI AGGQ                      AY++QGQ+ I   D          L KL
Sbjct: 185 MPPVIFAGGQ----------------------AYTVQGQYAIPHPDCQGVELMNPQLTKL 222

Query: 239 NQVARQQSHFAMMHG-GT----------GFAGIDSSSPEVKGYWASLDAS-TQTTHELTI 286
           +Q+A Q +     H  GT            A + +++    G  A+  AS T TT E+ I
Sbjct: 223 HQLALQHAPLLPGHSVGTLSPQGSLCIPALATLAATNSLRPGTAATALASVTTTTTEMNI 282

Query: 287 PNNLIGCIIGRQGANINEIRQMSGAQIKIANPVEGSSGRQVTITGSAASISLAQYLINAR 346
           PN+LIGC+IG+ GA INEIRQ+SGA IKI+N  EGS  R V I+G+  +I+LAQYLIN
Sbjct: 283 PNDLIGCVIGKGGAKINEIRQLSGATIKISNSEEGSKDRTVAISGTPEAINLAQYLINTS 342

Query: 347 LSSEK 351
           + S +
Sbjct: 343 IHSAQ 347



>aug3.g16335.t1
          Length = 449

 Score =  319 bits (817), Expect = e-106,   Method: Compositional matrix adjust.
 Identities = 180/366 (49%), Positives = 242/366 (66%), Gaps = 55/366 (15%)

Query: 5   VTESGLN--VTLTIRLLMHGKEVGSIIGKKGESVKRIREESGARINISEGNCPERIITLT 62
           VT +G+   VTLT+RL+M GKEVGSIIGKKG+++K+ REESGA+INIS+G+ PERI+T+T
Sbjct: 23  VTTNGIAPAVTLTVRLIMQGKEVGSIIGKKGDNIKKFREESGAKINISDGSSPERIVTVT 82

Query: 63  GPTNAIFKAFAMIIDKLEEDINSSMTNSTAASRPPVTLRLVVPATQCGSLIGKGGCKIKE 122
           G T +I KAF+++  K E+D+ +S+T  T   +PPVTLRL+VPA+QCGSLIGKGG KIK+
Sbjct: 83  GSTESILKAFSLVAQKFEDDLQASVTG-TNVPKPPVTLRLIVPASQCGSLIGKGGSKIKD 141

Query: 123 IRESTGAQVQVAGDMLPNSTERAITIAGVPQSVTECVKQICLVMLETLSQSPQGRVMTIP 182
           IRE+TGA +QVA DMLPNSTERA+TI+G  +++T+C+ QIC VM+E+    P+G   TIP
Sbjct: 142 IREATGASIQVASDMLPNSTERAVTISGTAEALTKCIYQICCVMMES---PPKG--ATIP 196

Query: 183 YQPMPASSPVICAGGQDRCSDAVGYPHATHDLEGPPLDAYSIQGQHTISPLDLAKLNQVA 242
           Y+P  A  P++ +G Q                      AY++ GQ+ +   DL KL Q+A
Sbjct: 197 YRPKQAVLPLLFSGAQ----------------------AYAMPGQYALPHPDLNKLQQLA 234

Query: 243 RQQSHFAMMHGGTG--FAGIDSSSPEVKGYWASLDA-------------STQTTHELTIP 287
             Q      H  T     G  S++P+    WASL A             S  +T E+TI
Sbjct: 235 TLQ------HASTAPLLPG-HSATPQA---WASLAAANSVRNLSRAASISPTSTAEMTIS 284

Query: 288 NNLIGCIIGRQGANINEIRQMSGAQIKIANPVEGSSGRQVTITGSAASISLAQYLINARL 347
           N LIGCIIG+ G  INEIRQ+SGA +K++N  EGS  R VTITG+ A+++ AQYLINA +
Sbjct: 285 NELIGCIIGKGGTKINEIRQISGAIVKVSNTEEGSKDRTVTITGTPAAVNCAQYLINANI 344

Query: 348 SSEKGM 353
             K +
Sbjct: 345 ERHKTL 350
```

Appendix Figure 5.2: Alignments of the *PCPB1 Homo sapiens* PolyC-binding protein to *P. tepidariorum* proteins (subject). BLASTp identified two candidate homologs of the human PolyC-binding protein.