THE UNIVERSITY
of EDINBURGH

The evolution of Dipteran *Argonaute* genes through duplication, selection and functional specialisation

Samuel H. Lewis

A thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2015

Contents

Lay Summary

All organisms need to regulate their genes correctly and fight off parasites. *Argonaute* genes play a key role in these processes in almost all plants, fungi and animals, but are present in varying numbers in different species. This variation is produced by gene duplication, which can allow new duplicates to alter their function, changing key aspects of the whole organism. Previous work has identified isolated examples of *Argonaute* gene duplication, but most work on *Argonaute* function has been focused on a few model species such as the fruit fly. Little attention has been paid to how frequently or rapidly *Argonaute* genes duplicate, how they evolve after duplication, and the functional diversity that such evolution may produce.

In Chapter 2 I measured how often different *Argonaute* genes duplicate in 86 different fly species, ranging from mosquitoes to hoverflies. I found that the rate of duplication varies between different *Argonaute* genes and different fly species, and that *Argonaute* genes evolve more rapidly at the DNA sequence level after they duplicate, suggesting that *Argonaute* genes may be taking on different functions after duplication. In Chapter 3 I investigated this potential for functional change in three fly species, and found that in each case *Argonaute* duplicates have specialized to a function in the testes. In Chapter 4 I measured how much natural selection is acting on these duplicates, and found that many are evolving very quickly at the sequence level, and appear to be under strong selection for their testis-specific function. I conclude that frequent duplication and rapid evolution are likely to have produced a hitherto unappreciated diversity of *Argonaute* functions.

Abstract

The RNA interference (RNAi) mechanism is a conserved system of nucleic acid manipulation, based on the interaction between small RNA guide molecules and Argonaute effector proteins. RNAi pathways are found in the vast majority of eukaryotes, and have diversified into a broad array of functions including gene regulation, antiviral immunity and transposable element (TE) suppression. Many of these functional innovations coincide with duplication of *Argonaute* genes, suggesting that gene duplication may be a key driving force in the diversification of RNAi. However, few studies have explicitly investigated *Argonaute* evolution after duplication. In this thesis, I focused on the impact of gene duplication on the evolution of *Argonaute* genes.

Argonaute genes in different species exhibit a broad array of functions; however, most of our knowledge of *Argonaute* function in the arthropods is based on studies in *D. melanogaster*. To compare the rate of duplication and its evolutionary effect between different *Argonaute* subclades, I quantified gene turnover rates and evolutionary rate change in *Argonaute* genes from 86 Dipteran species (Chapter 2). I find that duplication rate varies widely between subclades and lineages, and that duplication drives an increase in evolutionary rate, suggesting that functional divergence after *Argonaute* duplication is prevalent throughout the Diptera.

In the *obscura* group of *Drosophila* I identified a series of recent duplications of *Argonaute2* (*Ago2*), which has antiviral and anti-TE functions in *D. melanogaster*. To quantify the extent of functional divergence between these paralogues, I measured the expression of paralogues from three species (*D. subobscura*, *D. obscura* and *D. pseudoobscura*), in different tissues and under viral challenge (Chapter 3). I find that the majority of *Ago2* paralogues have specialised to a derived testis-specific role, potentially to suppress TE activity or meiotic drive. While CRISPR-Cas9 mediated knockout of these genes ultimately proved unsuccessful (Chapter 5), the selective importance of their derived function is suggested by its multiple independent origins.

Functional novelty, as appears to have evolved in the *obscura* group *Ago2* paralogues, is often driven by strong selection. To quantify the evolutionary rate and positive selection on these paralogues, I gathered intraspecies polymorphism data for all paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, combining this with publicly-available population genomic data for *D. pseudoobscura* (Chapter 4). I find that the majority of paralogues in all species have extremely low diversity, indicative of recent selection, and identify recent selective sweeps on three paralogues in *D. pseudoobscura*. This suggests that the majority of *Ago2* paralogues in the *obscura* group are evolving under strong positive selection.

In this thesis I have aimed to quantify the effect of gene duplication on *Argonaute* evolution. I find that *Argonaute* genes duplicate frequently in some lineages, resulting in the evolution of derived functions that may be driven by positive selection. This suggests that functional diversification is prevalent in eukaryotic RNAi, and is likely to coincide with expansion of the *Argonaute* gene family.

Declaration

I declare that I have composed this thesis, under guidance from Dr. Darren Obbard, and with contributions from other authors for individual chapters as stated below. Chapters 2-5 are manuscripts that have been published or are in preparation, therefore I have kept the use of "we". Chapters 1 and 6 were written solely as thesis chapters, and therefore use "I". I confirm that I carried out all experimental work, except where explicitly stated below.

Chapter 2

Dr. Heli Salmela performed protein structural modelling and provided feedback on writing, and some sequence data was gathered by Dr. Obbard. I gathered all other data, performed all other analyses, and wrote the chapter under guidance from Dr. Obbard.

Chapter 5

Drosophila pseudoobscura embryo microinjections were performed by Dr Sang Chan at the Fly Facility, University of Cambridge, UK. I performed all other organismal and molecular work.

14.01.16

Acknowledgements

First and foremost, I would like to thank Darren Obbard for providing an enormous amount of teaching and help with the practical aspects, for always offering advice and guidance, and for being equally generous when buying drinks. Many thanks to Pedro Vale for providing me with gainful employment during the final months and valuable advice throughout, and Tom Little for being a great second supervisor. I'm very grateful to Claire Webster for welcoming me to Edinburgh and giving me a firm grounding in lab technique, the Helens for being so helpful and always making me smile, and Fergal Waldron, Billy Palmer and Nathan Medd for being such great lab mates. Thanks also to Elliott Chapman and Sang Chan for providing advice and practical help with CRISPR, and all in the Evolutionary Genetics and IEEG lab groups for listening to me and providing insightful feedback.

I'd like to thank all of the people in IEB, who have made it such an enjoyable place to do a PhD. Particular thanks to Emma Hodcroft, Reuben Nowell, Hannah Froy, Lucy Carter, Kevin Donnelly, Manon Ragonnet, Darren Parker, Maarit Mäenpää and Elisa Schaum, who often made me think and always made me laugh. I'd also like to thank Manon Ragonnet for her support in the final year, and for always looking out for me. Thanks to everyone in my cohort, particularly Elisa Anastasi, Georgina Brennan, Jess Flood and Richard Allen, without whom the highs would have been less enjoyable and the lows less bearable. I wouldn't have got to or through my PhD without my parents and sister, who have always given me love and support, and provided enthusiasm and perspective in equal measure. My biggest thanks go to Sarah Mattey for her unfailing kindness, support and humour, for patiently reading numerous thesis chapters, and for helping me to remember that there are more important things than flies.

Lastly, I'd like to thank Dr Marks and Matthew Cobb for inspiring me, and Andrea Betancourt and David Finnegan for reading this thesis.

1. General Introduction

RNA interference (RNAi) is an ancient and conserved mechanism of nucleic acid manipulation, in which small RNA molecules guide Argonaute proteins to cleave, inhibit or regulate nucleic acid targets. RNAi is found in the vast majority of eukaryotes, where it carries out a wide array of functions including antiviral defence, gene regulation and the suppression of transposable elements (TEs). This diversity coincides with expansions of the *Argonaute* gene family, which has undergone ancient and recent gene duplications in a number of lineages. Gene duplication has underpinned the evolutionary diversification of many other mechanisms, resulting in the emergence of novel and adaptive traits; however, there are still large gaps in our knowledge of how gene duplicates (paralogues) in the RNAi pathway evolve. This thesis therefore focuses on the impact of gene duplication on the evolution of RNAi.

1.1. The role of gene duplication in the evolution of functional diversity

1.1.1. Evolutionary trajectories of paralogues

Gene duplication is a fundamental force in evolution (Ohno 1970), and occurs at a rate that is orders of magnitude higher than the base mutation rate (Katju and Bergthorsson 2013), with one mutation accumulation experiment in *D. melanogaster* estimating the duplication rate at 1.25 x 10^{-7} per base per generation, and the base mutation rate at 5.49 x 10^{-9} per base per generation (Schrider *et al.* 2013). Gene duplication can occur by four main mechanisms: DNA replication error, non-allelic homologous recombination, transposition and whole-genome duplication (reviewed in Hastings *et al.* 2009). After duplication, the resulting paralogues can evolve in ways traditionally classified as pseudogenization,

conservation, neofunctionalization and subfunctionalization (Figure 1.1).

Firstly, and most commonly (Lynch and Conery 2000; Hughes and Liberles 2007), one paralogue may retain the ancestral function, while the other paralogue accumulates deleterious substitutions which ablate its function (pseudogenization). Secondly, both paralogues may retain the function carried out by the ancestral gene before duplication (conservation). This may occur to increase the expression level of the ancestral gene, as illustrated by the extraordinarily high copy number of ribosomal RNA genes in *Saccharomyces cerevisiae* (~140 paralogues), *Drosophila melanogaster* (~200 paralogues) and humans (~400 paralogues) (reviewed in Eickbush and Eickbush 2007). More rarely (Clark 1994), both paralogues may retain the ancestral function to provide an essential function with a buffer against deleterious mutations (Haldane 1932).

Thirdly, one paralogue may retain the ancestral function while the other paralogue experiences relaxed selective constraints, resulting in the evolution of an entirely new function or pattern of expression (neofunctionalization). This is seen for RNase1 & RNase1B in the douc langur: RNase1 displays optimum activity at pH 7.4, whereas RNase1B has optimum activity at the lower pH of the stomach, enabling the douc langur to exploit the novel niche of folivory (Zhang *et al.* 2002). Lastly, the two paralogues can each specialize to a subset of the original functions (subfunctionalization), either by complementary degenerative mutations in each paralogue (termed "duplication-degeneration-complementation" (DDC); Force *et al.* 1999), or by positive selection driving specialization to complementary functions. This is demonstrated by paralogues of the *engrailed* gene in the ray-finned fishes, which have partitioned the expression patterns of the ancestral gene, with *eng1* expressed only in the pectoral appendage bud, and *eng1b* expressed solely in the spinal chord and hindbrain (Force *et al.* 1999).

Despite their differences, these trajectories are not mutually exclusive: rather, subfunctionalization can lead to neofunctionalization. One consequence of subfunctionalization is to prevent one or both paralogues accumulating deleterious mutations and being lost by pseudogenization, increasing both the retention of paralogues and the frequency of neofunctionalization (Rastogi and Liberles 2005). This process, termed "subneofunctionalization" (He and Zhang 2005), describes the evolutionary trajectories of a large proportion of paralogues in yeast (He and Zhang 2005), and may have driven the evolution of paralogues of the RNA polymerase (Pol) subunit *NRPD2*/*NRPE2* in the genus *Viola*, which display the combination of conservation and positive selection characteristic of the subneofunctionalization process (Marcussen *et al.* 2010).

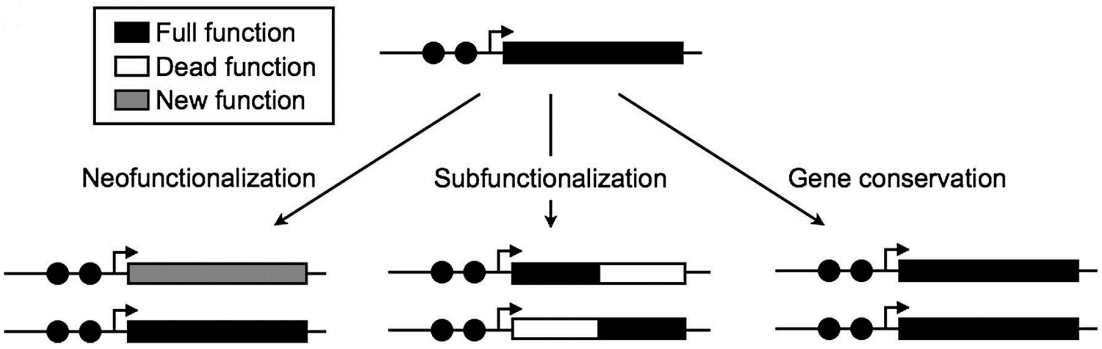


Figure 1.1.: Evolutionary trajectories of paralogues.
After duplication, paralogues may evolve a new function (neofunctionalization), take on a subset of the functions of a multifunctional ancestral gene (subfunctionalization), be conserved to carry out their original function (gene conservation), or be lost through pseudogenization (not shown) (Figure reproduced with permission from Hahn 2009, Fig. 1).

1.1.2. Correlates of gene duplication

There are three main factors that affect the frequency of duplication (duplicability) at a particular locus: essentiality (normally defined by a lethal knockout phenotype), centrality (the number of interactions with other genes), and specificity of expression. There is a negative relationship between essentiality and duplicability, with essential genes duplicating less frequently in *C. elegans* (Woods *et al.* 2013) and humans (Nguyen *et al.* 2008). A similar negative relationship exists between centrality and duplicability, as seen in *D. melanogaster* (Dopman and Hartl 2007) and yeast (Prachumwat and Li 2006). This may be linked to the overrepresentation of ohnologues (duplicates arising from a whole-genome duplication) among copy number variants (CNVs) linked to neurodegenerative disease (McLysaght *et al.* 2014): ohnologues are generally retained to maintain dosage balance with interacting genes, which would be disrupted by a further duplication. In contrast, specificity of expression is positively correlated with duplicability, evidenced by the higher duplication rates of tissue-specific genes compared with ubiquitously-expressed genes in *D. melanogaster* (Dopman and Hartl 2007), and more frequent duplication of genes limited to the cellular periphery of *S. cerevisiae* (Prachumwat and Li 2006).

After duplication, a common fate for recent duplicates is the evolution of a testis-biased or testis-specific pattern of expression. Individual examples of this pattern are provided by the paralogue *Tre2* in hominoids (Paulding *et al.* 2003), the orphan gene *Poldi* in the genus *Mus* (Heinen *et al.* 2009), and the

paralogue *Sdic* in *Drosophila* (Nurminsky *et al.* 1998). Multi-locus analyses have confirmed the prevalence of this pattern in *Drosophila*. In *D. melanogaster*, 50% of paralogues that have retroposed from the X chromosome are testis-specific (Betrán *et al.* 2002), and all five of the orphan genes investigated by Levine *et al.* 2006 were testis-biased or testis-specific. A similar pattern has been found in the *D. yakuba*/*D. erecta* clade, where 8 orphan genes exhibit testis-specific expression (Begun *et al.* 2007). These observations led to the proposal of the "out of the testis" hypothesis (Kaessmann *et al.* 2009; Kaessmann 2010), which proposes that novel functions will evolve predominantly in the testis. Mechanistically, the open chromatin structure and high transcription rate of the autosomes in the testes (Kleene 2001) result in a promiscuous expression environment, which can cause otherwise non-functional duplicates to be expressed. This exposes these duplicates to the strong selective pressures of the testis, which are enriched for rapidly evolving genes in humans (Nielsen 2005) and *Drosophila* (Haerty *et al.* 2007), preventing their loss by pseudogenization and resulting in the evolution of new functions (Kaessmann *et al.* 2009; Kaessmann 2010). More recently, genome-wide studies in *Drosophila* have confirmed the relevance of the out-of-the-testis hypothesis for both young paralogues (Assis and Bachtrog 2013) and young orphan genes (Palmieri *et al.* 2014). Additionally, comparison of the tissue-specificity of young and old paralogues has shown that testis-specificity is not a fixed pattern; instead, paralogues appear to go through a period of testis-specificity immediately after duplication, but then evolve expression in a broader range of tissues and carry out more protein-protein interactions as they age (Assis and Bachtrog 2013).

These studies show that expression patterns can be highly informative regarding the evolution of gene duplicates. Before duplication, expression pattern is itself predictive of the likelihood of duplication, as well as being correlated with other predictors of duplication such as connectivity (Prachumwat and Li 2006; Dopman and Hartl 2007). After duplication, specialisation to different tissues can reflect subfunctionalization or neofunctionalization (Hahn 2009), lack of expression is sometimes underpinned by pseudogenization, and testis-specific expression indicates the potential emergence of novel functions (Kaessmann *et al.* 2009; Kaessmann 2010).

1.1.3. The adaptive significance of gene duplication

Gene duplication is a key process in the emergence of phenotypic innovations. This is exemplified by the three paralogues of the *knickopf* gene in the beetle *Tribolium castaneum*, all of which function in

the formation of chitin, but two of which have specialized to different moults in the development of the cuticle (Chaudhari *et al.* 2014). Additionally, these three paralogues have been conserved since their origin in the ancestor of insects, suggesting that their specific roles may have played an important role in insect evolution (Chaudhari *et al.* 2014). The diversity of leaf shapes seen in the Brassicaceae has also been driven by duplication: a duplication of the *LMI1* homeobox gene early in Brassicaceae evolution produced *RCO1*, the regulatory regions of which evolved to restrict its expression to the base of developing leaves, transforming leaf serrations into deep lobes and resulting in the evolution of more complex leaf morphology (Vlad *et al.* 2014). Other paralogues have contributed to evolutionary change despite a lack of divergence at the sequence level, as demonstrated by amylase, which aids in the digestion of starch in humans, and which increases in copy number in populations with a history of high starch diets (Perry *et al.* 2007).

Due to the selective advantage conferred by some duplications, a number of paralogues evolve under strong selection. The strength of selection can be estimated by computing the ratio of non-synonymous to synonymous polymorphisms (Ka/Ks, or ω) for individual sites or whole genes, with $\omega<1$ indicating purifying selection, $\omega=1$ suggesting neutrality, and $\omega>1$ a sign of positive selection. Using this approach, strong positive selection was found soon after paralogue fixation in the desaturase gene family, members of which play key roles in evolutionary divergence and speciation through their contribution to the formation of cuticular hydrocarbons (Keays *et al.* 2011). Similar evidence of positive selection was found for *Opn4x*, an ancient paralogue of the vertebrate photopigment melanopsin, which also displays some evidence for functional divergence (Dong *et al.* 2012). Moreover, positive selection was identified after each of the duplication events that produced the four paralogues of vitellogenin in *Formica* ants, which have also evolved divergent protein structures and caste-specific expression patterns (Morandin *et al.* 2014).

It is clear from these studies that gene duplication produces many paralogues that evolve adaptively, resulting in functional divergence and phenotypic innovation. However, the resulting functional diversity makes it difficult to measure how frequently adaptive functional change occurs. Previous metrics based on expression patterns (Assis and Bachtrog 2013) or knockout lethality (Woods *et al.* 2013) have provided valuable insights, but are too coarse to capture all functional change, meaning that our knowledge of the functional evolution of paralogues is still in large part informed by individual molecular studies.

1.1.4. Gene duplication as a solution to fitness trade-offs

Many genes evolve under selective constraint imposed by fitness trade-offs. This constraint is often imposed by pleiotropy, when one gene influences numerous traits which exert multiple, and sometime conflicting, selective pressures (reviewed in Stearns 2010). The escape from adaptive conflict model (EAC) decribes how this constraint can be lifted by gene duplication, which allows the resulting paralogues to specialize to one function of the pleiotropic ancestral gene, resulting in a fitness increase due to the increased efficiency with which each paralogue can carry out its function (Hittinger and Carroll 2007). There are two expectations for paralogues evolving through this process, which differentiate it from the other trajectories outlined in Section 1.1.1. Firstly, both paralogues will evolve under positive selection, as both accumulate adaptive substitutions; this is in contrast with neofunctionalization, where only one paralogue experiences positive selection driven by the novel function, while the other is constrained under purifying selection to retain the ancestral function. Secondly, under EAC the efficiency of both ancestral functions will increase; this contrasts with neofunctionalization, which increases the efficiency of only the novel function, and the DDC model of subfunctionalization, which proceeds through a neutral process and involves no increase in efficiency (Section 1.1.1; Hittinger and Carroll 2007; Des Marais and Rausher 2008).

There are a number of examples of paralogues evolving under the EAC model. This model has been invoked for paralogues of the *dihydroflavonol-4-reductase* (*DFR*) gene in morning glory (*Ipomoea*), which underwent positive selection soon after duplication, and only one of which has retained the ancestral function in the anthocyanin pathway, which it carries out with higher efficiency than the multifunctional ancestral gene (Des Marais and Rausher 2008). The EAC model has also been invoked to explain the evolution of the yeast glucosidase enzymes: by inferring and synthesising ancestral proteins sampled throughout glucosidase evolutionary history, Voordeckers *et al.* 2012 have shown that ancestral forms could metabolise a broad range of maltose-like and isomaltose-like sugars, but at a lower rate than more recent paralogues, which have specialized to either maltose or isomaltose.

These examples illustrate how gene duplication can provide a mechanism for the resolution of antagonistic pleiotropy, and underline the significance of this mechanism in facilitating adaptive evolution that would otherwise be constrained by conflicting selection pressures. These studies also suggest that while duplication of multifunctional genes with a high degree of connectivity may be rare (see Section 1.1.2), when such duplications do occur they can often lead to functional divergence driven by positive

selection.

1.2. RNAi-related mechanisms

1.2.1. Basic mechanism

RNAi is an ancient system for processing nucleic acids that is directed by small RNA (sRNA) guides, which bind Argonaute (Ago) proteins and guide them to nucleic acid targets, which Argonaute then manipulates by cleavage, translational inhibition or degradation (reviewed in Ding 2010). There are a number of different pathways, which differ in their targets, sRNAs and Argonautes, but all proceed through the same basic mechanism (Figure 1.2). All RNAi mechanisms start with the production of a double-stranded sRNA from a nucleic acid target, which can be exogenous (e.g. a virus) or endogenous (e.g. a host locus or transposable element) in origin (reviewed in Kim *et al.* 2009). The production of sRNAs can be achieved by transcription of a host locus encoding an sRNA, or the "dicing" of a target nucleic acid by a Dicer protein, which is sometimes first produced from a target RNA by an RNA-dependent RNA polymerase (RdRP; Wassenegger and Krczal 2006). Each of these sRNAs is then processed and loaded into an Argonaute protein, at which point one strand of the sRNA (the passenger strand) is ejected, leaving a single-stranded sRNA bound to the Argonaute. This sRNA then guides the Argonaute to the target nucleic acid, and binds to it through complementary base pairing. Once the sRNA has bound, the Argonaute cleaves the target or inhibits its transcription or translation, thereby preventing the target from acting in the cell.

1.2.2. Argonaute domain architecture and protein structure

All full-length Argonaute proteins have four functional domains (MID, PAZ, PIWI & N) which play distinct functional roles. The MID domain facilitates sRNA loading by interacting with the 5' sRNA end, and can confer sRNA-binding specificity by recognizing particular 5' bases (Ma *et al.* 2005; Frank *et al.* 2010). The PAZ domain also plays a role in loading sRNA guides, binding the 3' end and preventing the sRNA from degradation (Hur *et al.* 2013). In contrast, the N domain ensures that the passenger strand is ejected from the sRNA-Argonaute complex, thereby permitting complementary base-pairing

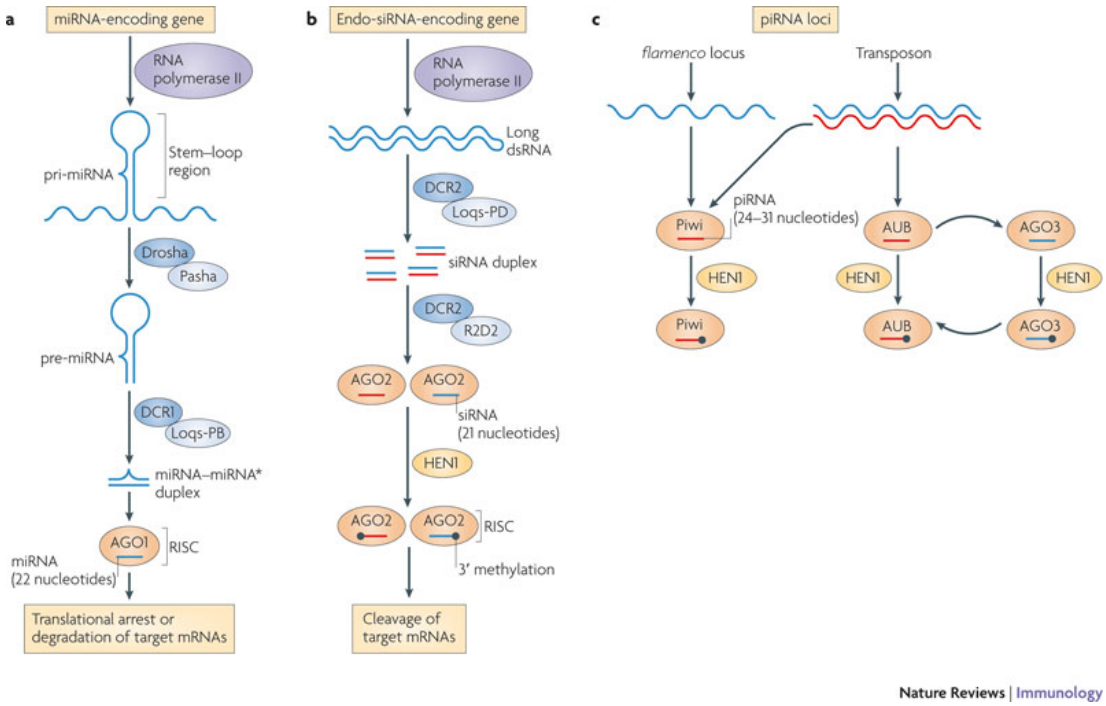


Figure 1.2.: The Metazoan RNAi pathways.
Each RNAi pathway proceeds through the same basic mechanism. First, a target nucleic acid enters the cell and is chopped into small RNAs (sRNAs), each of which is loaded into an Argonaute protein. The sRNA then guides the Argonaute to the target and binds by complementary base pairing, whereupon the Argonaute cleaves the target or prevents its transcription or translation (Figure reproduced with permission from Ding 2010, Fig. 1).

between the sRNA guide and the target (Kwak and Tomari 2012). Finally, the PIWI domain functions in target binding, with the PIWI domain of most Argonautes containing a DDX triad which catalyses target cleavage (reviewed in Swarts *et al.* 2014b). However, some Argonautes have an inactive DDX triad, such as *D. melanogaster* Ago1 and human Ago1, Ago3 & Ago4 (reviewed in Meister 2013), resulting in inhibition of the target through means other than cleavage.

As well as this domain architecture, the 3-D protein structure of Argonautes plays a vital role in their function. All Argonautes have a bilobal structure (Figure 1.3), with the PAZ and N domains on one lobe and the Piwi domain on the other, forming a channel into which the target nucleic acid slots (Song *et al.* 2003; Song *et al.* 2004; Schirle and Macrae 2012). These two lobes are connected by the MID domain, which forms the binding pocket for the sRNA guide (Song *et al.* 2004; Elkayam *et al.* 2012; Schirle and Macrae 2012). This structure displays a remarkable degree of evolutionary conservation, being found in bacteria (Song *et al.* 2004), fungi (Nakanishi *et al.* 2012) and animals (Schirle and Macrae 2012), reinforcing its essential role in Argonaute function (Swarts *et al.* 2014b).

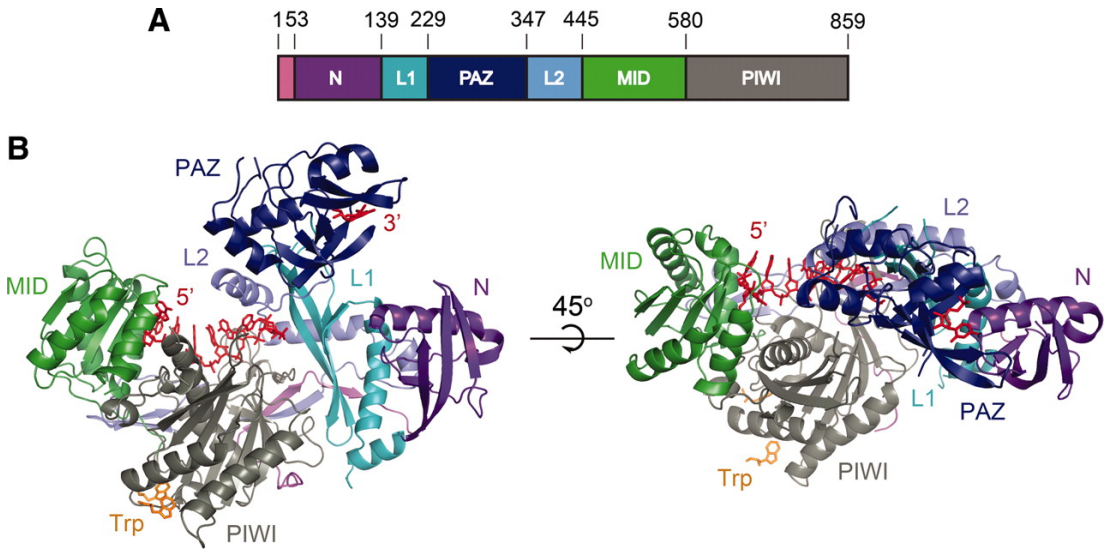


Figure 1.3.: The domain architecture and crystal structure of human Argonaute2.
Human Argonaute2 illustrates the conserved Argonaute protein structure: the PAZ and MID domains form a binding pocket for the sRNA guide, while the N and PIWI domains form a catalytic core for target binding and cleavage (Figure reproduced with permission from Schirle and Macrae 2012, Fig. 1).

1.2.3. Canonical sRNA classes, characteristics and targets

All RNAi pathways are guided by sRNAs, which fall into three main classes: microRNAs (miRNAs), short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). These classes differ in their biogenesis, their length and base frequency biases, the Argonautes to which they bind, and the type of nucleic acid that they target. There are several key differences in these characteristics between plant and animal sRNAs, but here I focus on sRNAs from animals.

In animals, miRNA loci are found in exons, introns and intergenic regions, and are frequently located in clusters which may be transcribed in a single polycistronic unit (reviewed in Axtell *et al.* 2011). Most miRNAs are transcribed from host-encoded loci by RNA polymerase II (Lee *et al.* 2004a), producing a primary miRNA (pri-miRNA) which has a hairpin-loop structure. This pri-miRNA is truncated in the nucleus by Drosha, producing a ~70nt precursor miRNA (Lee *et al.* 2003), which is then cleaved by Dicer in the cytoplasm, producing a mature miRNA (Hutvágner *et al.* 2001; Lee *et al.* 2004b). miRNAs are ~22nt long, have 5' monophosphate and 3' hydroxl termini (Ding 2010), and bind members of the Ago subfamily of Argonautes. Once bound, miRNAs mediate gene regulation by binding to host transcripts, inhibiting their translation or inducing mRNA degradation (reviewed in Huntzinger and Izaurralde 2011). The specificity of target binding is achieved despite incomplete base complementarity between the miRNA and the target mRNA: instead, target specificity is largely determined by the "seed"

region at nucleotide positions 2-7, and to a lesser extent by positions 13-16 (Ha and Kim 2014). This lack of specificity makes it difficult to cluster miRNAs into discrete groups based on target; however, an miRNA family is generally defined as a group of miRNAs that share the same seed region (Ha and Kim 2014). Under this classification scheme, it appears that gene duplication has played a major role in the expansion of some miRNA families, such as the let-7 family that has 14 members in humans (Roush and Slack 2008).

In contrast, siRNAs are produced through the "dicing" of dsRNA by Dicer. This produces 19-23nt sRNAs that have 5' monophosphate and 3' hydroxl termini (Ding 2010), as well as characteristic 2nt 3' overhangs due to the dual RNaseIII domains of Dicer, each of which catalyzes cleavage of one strand of the dsRNA at sites 2nt apart (Zhang *et al.* 2004). All siRNAs bind members of the Ago subfamily of Argonautes, which cleave the target nucleic acid once the siRNA has bound. There are two subclasses of siRNAs, which share the common siRNA characteristics mentioned above, but are derived from different sources. Firstly, virally-derived siRNAs (vsiRNAs) are viral in origin, and are derived either from the genomes of dsRNA viruses (reviewed in Bronkhorst and Rij 2014), or from overlapping viral transcripts (Bronkhorst *et al.* 2012). Secondly, endogenous siRNAs (endo-siRNAs) originate from the host genome, either from host transcripts or TE-derived transcripts in the soma (Czech *et al.* 2008), with the latter having a 5' U bias (Chung *et al.* 2008).

Finally, piRNAs are produced from TEs and some host genes through two pathways (primary and secondary), which have mainly been characterised in *D. melanogaster* (Li *et al.* 2009; Mohn *et al.* 2015; Han *et al.* 2015) but are also conserved in *Mus musculus* (Mohn *et al.* 2015; Han *et al.* 2015). In the primary piRNA pathway (Figure 1.4, left), *D. melanogaster* Piwi or Aub binds to a TE-derived antisense transcript, triggering its cleavage and the formation of a piRNA (Mohn *et al.* 2015; Han *et al.* 2015). This process also triggers the formation of further piRNAs in a phased pattern across the transcript by Zucchini and Armitage, which cleave only at uridine (Mohn *et al.* 2015; Han *et al.* 2015). Primary piRNAs can then feed into the secondary piRNA pathway (Figure 1.4, right), also termed the Ping-Pong amplification loop, where *D. melanogaster* Ago3 and Aub produce further antisense and sense piRNAs respectively from target transcripts (Li *et al.* 2009). Some of these secondary piRNAs can also leave the Ping-Pong cycle and feed back into the primary piRNA pathway, triggering further phased production of piRNAs by Zucchini and Armitage (Mohn *et al.* 2015; Han *et al.* 2015). Through these two processes, the piRNA pool can be both diversified (by the primary pathway) and amplified (by the secondary pathway) (Han *et al.* 2015). piRNAs are primarily ~25-30nt long, are methylated at the 3' end, and have a 5'U bias, and those that are produced by the Ping-Pong pathway also have a bias for A

at the 10th position (reviewed in Luteijn and Ketting 2013). All piRNAs are defined by their binding to Piwi subfamily Argonaute proteins, and direct the silencing of TEs (Kalmykova *et al.* 2005) through the formation of heterochromatin (Sienski *et al.* 2012).

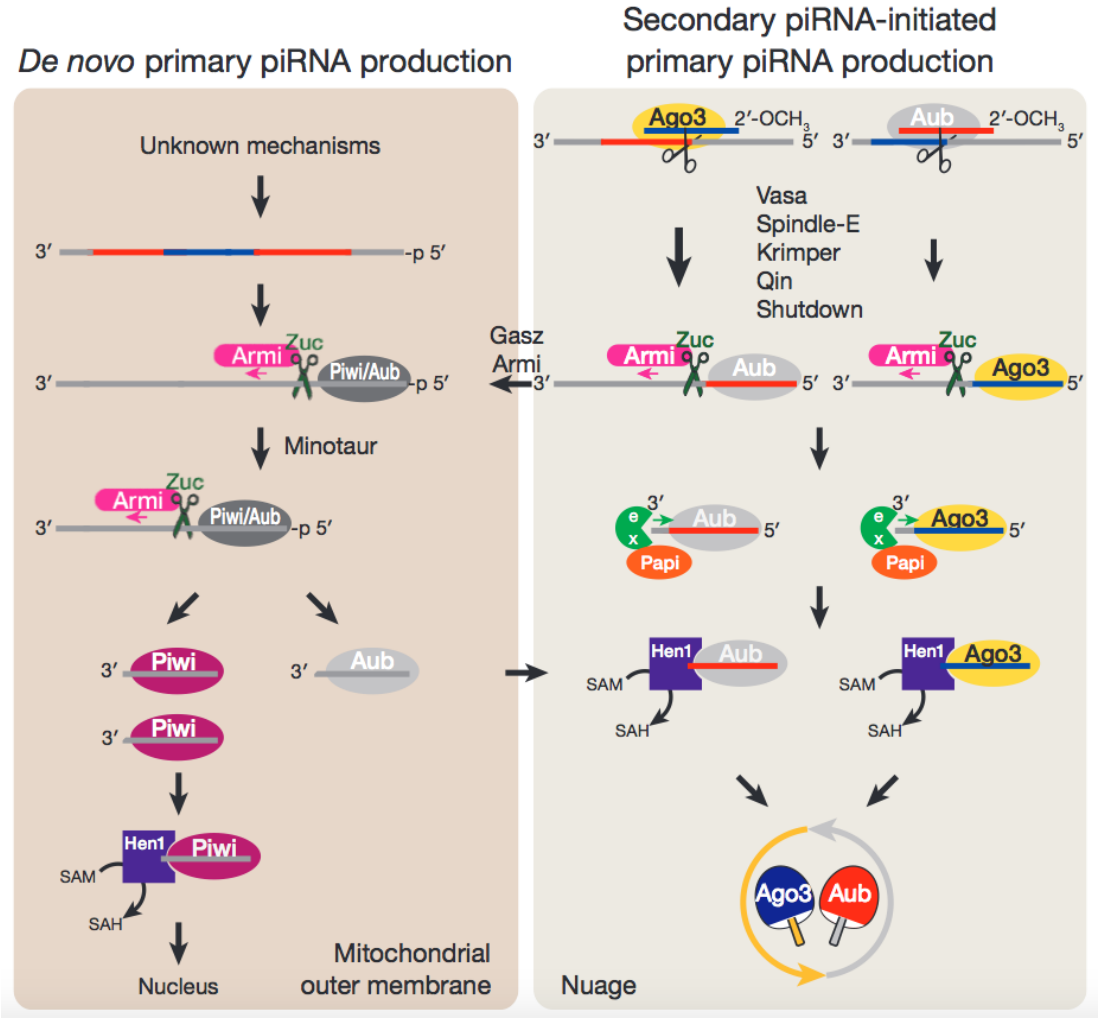


Figure 1.4.: A recently revised model for piRNA biogenesis.
In the primary pathway, binding of an antisense transcript by Piwi or Aub triggers phased piRNA production along the length of the transcript by Zucchini or Armitage. In the secondary pathway, Ago3 and Aub amplify the piRNA signal by producing antisense and sense piRNAs respectively (Figure reproduced with permission from Han *et al.* 2015, Fig. S8).

1.3. Comparative RNAi

1.3.1. Phylogenetic distribution of *Argonaute* genes

The tight functional link between Argonautes and sRNAs is reflected by the ancient origin of *Argonaute* genes, which have been found in prokaryotes and eukaryotes, and are therefore likely to have originated in an early prokaryote (Cerutti and Casas-Mollano 2006; Mukherjee *et al.* 2013; Swarts *et al.* 2014b). Argonautes function in a variety of prokaryotic silencing mechanisms that are analogous to RNAi, but are guided by DNA instead of or in addition to RNA (Olovnikov *et al.* 2013; Swarts *et al.* 2014a; Swarts *et al.* 2015). These mechanisms inhibit the invasion and uptake of foreign DNA, and hence have been speculated to be an ancient defence against invading plasmids and bacteriophages (Olovnikov *et al.* 2013; Swarts *et al.* 2014a; Swarts *et al.* 2015). *Argonaute* genes are widely distributed throughout the prokaryotes; however, prevalent horizontal gene transfer (HGT) precludes estimation of the rates of *Argonaute* duplication and loss (Makarova *et al.* 2009). While the domain architecture and protein structure of some of these Argonautes is similar to eukaryotic Argonautes, others are quite divergent: in particular, a truncated form of Argonaute (short pAgo) appears to have arisen early in bacterial evolution, which lacks the PAZ and MID domains (Makarova *et al.* 2009). Additionally, *pPiwi-RE*, a divergent *Argonaute* which encodes a PIWI domain but no PAZ or MID domain (Burroughs *et al.* 2013), is scattered across the prokaryotic phylogeny (Swarts *et al.* 2014b), although HGT again makes it difficult to trace the origin of this clade. In many species, short pAgo and pPiwi-RE proteins have inactivated PIWI domains, but the genes encoding them are located near predicted nucleases. If the *Argonaute* and the neighbouring nuclease are coexpressed in a single operon, as is the case for many other bacterial genes in close proximity (Aravind 2000), these Argonaute proteins could participate in an RNAi-like process despite their divergent structure (Makarova *et al.* 2009; Burroughs *et al.* 2013).

Argonaute genes are also distributed throughout the eukaryotes, and have undergone numerous ancient duplications early in eukaryotic evolution (Figure 1.5). However, sparse taxon sampling and evolutionary rate heterogeneity make these gene trees difficult to infer, introducing uncertainty around the precise dates of early duplication events. For example, *Piwi* is present in animals and some Protozoa (Swarts *et al.* 2014b), but completely absent from plants (Cerutti and Casas-Mollano 2006; Swarts *et al.* 2014b): this suggests that plants diverged from other eukaryotes first, followed by the duplication that produced *Piwi*, after which Protozoa diverged, some of which lost *Piwi* (Mueller *et al.* 2014). However, the unresolved root of this tree means that the alternative scenario is also possible: a duplication in the

eukaryotic ancestor could have produced *Piwi*, followed by the divergence of Protozoa (some of which lost Piwi), after which animals diverged from plants, all of which lost *Piwi*.

Denser taxon sampling and a shorter timescale mean that the order of events within separate eukaryotic lineages can be inferred with more certainty. In plants, there were at least three duplication events after the divergence of green unicellular algae, establishing four main *Argonaute* families (clades I-IV) before the evolution of multicellularity in plants (Singh *et al.* 2015). In the Metazoa, *Ago* duplicated early to form *Ago1* and *Ago2*, the latter of which appears to have been subsequently lost from deuterostomes and nematodes (Mukherjee *et al.* 2013). There were also duplications in the Piwi subfamily early in Metazoan evolution, either one duplication event in the ancestor of the protostomes and deuterostomes, or two duplication events separately in these two lineages, which produced *Hili* and *Hiwi* (and their homologues) in deuterostomes, and *Ago3* and *Piwi/Aub* (and their homologues) in protostomes (Swarts *et al.* 2014b).

These early duplication events have resulted in most eukaryotic species having at least one *Argonaute*; however, on rare occasions *Argonaute* subclades have been lost. This is demonstrated by multiple independent losses of the Piwi subfamily genes during nematode evolution (Sarkies *et al.* 2015), as well as the loss of other piRNA pathway genes such as *Hen1* (Sarkies *et al.* 2015), which methylates the 3' end of piRNAs (Horwich *et al.* 2007). A more dramatic example of *Argonaute* loss is provided by the budding yeasts, in which *Argonaute* has been lost entirely on several independent occasions (Drinnenberg *et al.* 2009; Drinnenberg *et al.* 2011). This has been linked to symbiosis with the killer virus, which confers a selective advantage on RNAi-deficient species by inhibiting the growth of other fungi that do not carry killer, and the presence of which is broadly correlated with the absence of RNAi across the fungi (Drinnenberg *et al.* 2011).

1.3.2. Phylogenetic distribution of other RNAi genes

In keeping with the wide distribution of *Argonaute* genes, most other RNAi pathway components appear to have an ancient origin; however, these other components appear to have evolved later than the *Argonaute* genes, in the most recent common ancestor (MRCA) of eukaryotes. The *Dicers* are assumed to have arisen early in the course of eukaryotic evolution, as they are found in the majority of eukaryotes but are absent from prokaryotes (Cerutti and Casas-Mollano 2006; Mukherjee *et al.* 2013). However,

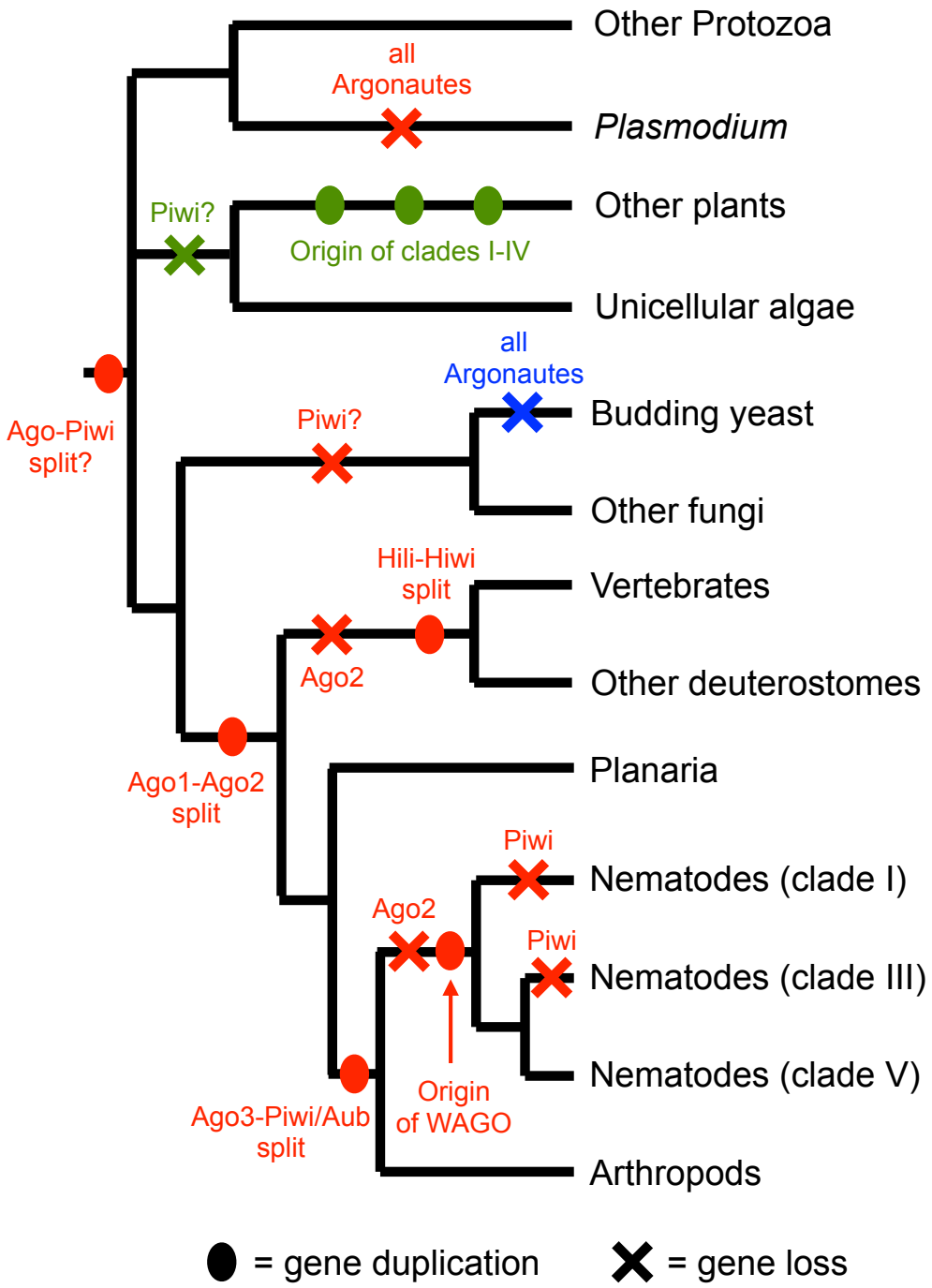


Figure 1.5.: A possible model for early events in eukaryotic Argonaute evolution.
There have been numerous expansions and contractions of the *Argonaute* family in different eukaryotic lineages, which are tentatively placed on this phylogeny; however, due to sparse taxon sampling and evolutionary rate heterogeneity, many of these dates remain uncertain (based on Cerutti and Casas-Mollano 2006 Fig. 2, Mukherjee *et al.* 2013 Fig. S12, Swarts *et al.* 2014b Supp. data 4, Sarkies *et al.* 2015 Fig. 7E, & Singh *et al.* 2015 Fig. 2).

some prokaryotes encode proteins with the RNaseIII (Cerutti and Casas-Mollano 2006) and RNA helicase (Shabalina and Koonin 2008) domains that are also found in the Dicer protein, suggesting that *Dicer* may have arisen by the fusion of existing genes (Shabalina and Koonin 2008). *Dicer* appears to have expanded in plants just before or very soon after the divergence of the mosses, giving rise to four paralogues which underwent lineage-specific loss (Mukherjee *et al.* 2013). *Dicer* also duplicated once early in the evolution of animals, after which both paralogues were retained in arthropods and some basal taxa, whereas one was lost from deuterostomes and nematodes (Mukherjee *et al.* 2013). Further lineage-specific expansions and losses have occurred in isolated animal lineages, including a paralogue of *Dicer1* (involved in gene regulation) in ticks that has evolved a derived antiviral role (Schnettler *et al.* 2014).

Similarly to *Dicer*, *RdRP* is absent from prokaryotes but was present in the MRCA of eukaryotes (Cerutti and Casas-Mollano 2006). *RdRP* appears to have duplicated twice in this ancestor to produce three paralogues: *RDR*α, *RDR*β and *RDR*γ (Zong *et al.* 2009). *RDR*α has been retained in plants, fungi and animals, whereas *RDR*β was lost early in the evolution of plants, and *RDR*γ was lost early in animal evolution (Zong *et al.* 2009). There have also been lineage-specific expansions and losses of each *RdRP*, including the loss of all *RdRPs* in insects and vertebrates (Zong *et al.* 2009), although sparse taxon sampling precludes reliable dating of these events. Less is known about the evolutionary history of other members of the RNAi pathway, although a pan-eukaryotic distribution has been found for *Hen1* (Tkaczuk *et al.* 2006), which methylates piRNA 3' ends (Horwich *et al.* 2007), and *SIDs* (Obbard *et al.* 2009b), which facilitate cellular uptake of dsRNA (Feinberg and Hunter 2003). This suggests that the majority of the RNAi pathway was present in early eukaryotes, and underwent evolutionary diversification or specialization in response to lineage-specific selection pressures, resulting in differential expansion and loss of genes throughout the mechanism (Obbard *et al.* 2009b).

1.3.3. Evolutionarily ancient contrasts in RNAi mechanisms

Despite its conservation across the eukaryotes, the RNAi mechanism has evolved a number of differences between separate eukaryotic kingdoms. This is illustrated by the contrasting modes of action seen in animal and plant miRNAs, despite their shared function in gene regulation. As outlined in Section 1.2.3, miRNA loci in animals are frequently clustered, and can be located within introns or exons, whereas in plants the vast majority of miRNA loci are non-clustered, and are rarely found within introns

or exons (Axtell *et al.* 2011). The cellular location of miRNA production from these loci also differs: in plants, miRNA biogenesis occurs exclusively in the nucleus (reviewed in Voinnet 2009), whereas in animals there are distinct nuclear and cytoplasmic stages (reviewed in Ha and Kim 2014). After biogenesis, plants miRNAs are invariably methylated at the 3' end by Hen1, whereas only a small fraction of animal miRNAs are 3' methylated (Axtell *et al.* 2011). After maturation, miRNAs also bind targets with contrasting degrees of complementarity: in animals, each miRNA binds to its targets with a low degree of sequence complementarity, meaning that each miRNA targets several different transcripts by inducing translational inhibition (Ha and Kim 2014). In contrast, each plant miRNA binds a single host transcript with a high degree of complementarity, inducing mRNA cleavage (Voinnet 2009). These extensive differences have even led some authors to speculate that miRNAs have separate evolutionary origins in plants and animals (Axtell *et al.* 2011).

Differences have also evolved in the RNAi-mediated suppression of TEs. Firstly, there is a stark contrast in sRNA length: in animals, TE silencing in the germline is guided by piRNAs (~25-30nt), whereas in plants TEs are silenced through the action of 21nt & 24nt siRNAs. Secondly, these sRNAs are produced through different mechanisms: in animals, piRNAs are produced in a Dicer-independent manner, whereas in plants siRNA production is Dicer-dependent. Thirdly, these sRNAs bind different Argonautes: piRNAs bind Piwi subfamily Argonautes (Luteijn and Ketting 2013), whereas plant siRNAs bind Ago subfamily Argonautes (reviewed in Parent *et al.* 2012). Despite these differences, an interesting similarity exists in the dual TE suppression mechanisms that have evolved in both animals and plants. In plants, 24nt siRNAs mediate pre-transcriptional TE suppression through the RNA-dependent DNA methylation (RdDM) pathway (Matzke and Mosher 2014), which directs methylation and histone modification at TE loci in a similar manner to the animal piRNA pathway (Sienski *et al.* 2012). In contrast, plant 21nt siRNAs guide post-transcriptional RNA-mediated cleavage of TE transcripts (reviewed in Ito 2013), through a mechanism analogous to the endo-siRNA pathway of animals (Czech *et al.* 2008; Chung *et al.* 2008).

1.3.4. Recent duplications and functional divergence of *Argonaute* genes

Differences in RNAi function have also arisen after recent *Argonaute* expansions. In humans, four Ago subfamily genes were formed from duplications early in the evolution of the vertebrates (Swarts *et al.* 2014b), and have since diverged in the mechanism by which they silence their mRNA targets: human

AGO2 (hAgo2) has retained the catalytic DDX triad in its PIWI domain (Section 1.2.2) and therefore cleaves its targets, whereas hAGO1, hAGO3 and hAGO4 have a non-catalytic PIWI domain, and inhibit target translation (Meister 2013). There are isolated reports of individual miRNAs binding preferentially to individual Argonautes, such as the non-canonical production of the microRNA miR-451, the precursor of which is specifically cleaved by Ago2 (Dueck *et al.* 2012). However, there is conflicting evidence regarding the extent of this specificity, with some studies reporting an assortative mechanism which directs different miRNAs to different Argonautes (Burroughs *et al.* 2011), and others concluding that miRNAs bind randomly to different Argonautes (Wang *et al.* 2012).

Numerous duplications of *Argonaute* have occurred in the arthropods. In the tiger shrimp *Penaeus monodon*, a young paralogue of *Ago2* has lost its ancestral role in ubiquitous antiviral defence, and has instead specialized to the germline, where it appears to be functioning in TE suppression (Leebonoi *et al.* 2015). A similar expansion of *Ago2* has occurred in the tick *Ixodes scapularis*, with two of the resulting four paralogues losing their ancestral antiviral activity (Schnettler *et al.* 2014). Interestingly, Dicer1 appears to play a role in antiviral defence in *I. scapularis*, despite the canonical function of arthropod Dicer1 in miRNA-mediated gene regulation; combined with expansions of *Ago2* in other *Ixodes* species (Schnettler *et al.* 2014), this suggests that tick RNAi pathways may be undergoing dynamic evolution.

In mosquitoes, there have been a series of duplications of *Piwi/Aub*, resulting in 6 paralogues in *Culex pipiens* and 7 paralogues in *Aedes aegypti* (Campbell *et al.* 2008). This expansion is accompanied by the recent discovery of piRNAs derived from viruses (vpiRNAs) in the soma of *A. albopictus* (Morazzani *et al.* 2012), and in somatically-derived cell culture of *A. aegypti* (Vodovar *et al.* 2012). The *Piwi/Aub* paralogues in *A. aegpyti* have distinct roles in antiviral defence, with Ago3 and Piwi5 amplifying vpiRNAs through the Ping-Pong mechanism (Miesen *et al.* 2015) while Piwi4 acts as an antiviral effector (Schnettler *et al.* 2013a), and appear to bind TE-derived piRNAs with contrasting origins and sequence characteristics (Miesen *et al.* 2015). The recent origin of these paralogues, combined with their divergent functions, suggest that Piwi/Aub paralogues in mosquitoes have rapidly specialized to distinct roles in novel RNAi pathways.

The most dramatic expansion of *Argonaute* genes currently known is in the nematodes, with *Caenorhabditis elegans* encoding 25 *Argonaute* genes, including 18 members of the highly divergent worm-specific *Agos* (*WAGOs*) (reviewed in Buck and Blaxter 2013). These Argonautes have undergone extensive functional divergence, variously specializing to miRNA (gene regulation), exo-RNAi (antiviral defence) and endo-RNAi (TE suppression) pathways (Yigit *et al.* 2006). Additionally, the divergent WAGOs bind

a novel class of sRNA (Yigit *et al.* 2006), the 22G-RNAs, which are unusual among animals in being produced on an individual basis by an RdRP (rather than in a long dsRNA which is then diced) (Pak and Fire 2007), and which carry out many novel functions such as epigenetic memory formation (reviewed in Buck and Blaxter 2013). Many of the *Argonaute* paralogues in *C. elegans* have homologues in other nematode species, suggesting that *Argonaute* genes have duplicated frequently throughout the evolution of this diverse phylum (Buck and Blaxter 2013). These *Argonaute* paralogues are likely to be intricately involved in the staggering diversity of RNAi mechanisms seen in other nematode clades, several of which have lost *Piwis* and piRNAs independently (Sarkies *et al.* 2015). Instead, TEs are silenced through the evolution of an RdRP-dependent but Dicer-independent mechanism which produces TE-derived 22G-RNAs, or the retention of a potentially ancient RNA-directed DNA methylation mechanism involving both RdRP and Dicer (Sarkies *et al.* 2015).

1.3.5. The evolutionary dynamics of *D. melanogaster Argonaute* genes

Much of our knowledge regarding the functional and mechanistic differences between Argonaute paralogues has come from studies in *D. melanogaster*, which has two genes in the Ago subfamily (*Ago1* & *Ago2*), and three genes in the Piwi subfamily (*Ago3*, *Aubergine* (*Aub*) & *Piwi*). These genes bind different sRNA classes, carry out distinct functions, and evolve under contrasting selection pressures. Ago1 binds miRNAs (Tomari *et al.* 2007) with complementarity to host transcripts, and plays a key role in gene regulation by inhibiting translation and marking transcripts for degradation (reviewed in Eulalio *et al.* 2008). Perhaps unsurprisingly given the conserved nature of this role across the Metazoa, *Ago1* evolves very slowly under strong selective constraint (Obbard *et al.* 2006; Obbard *et al.* 2009b), a pattern reflected in other members of the gene regulatory RNAi pathway, such as *Dicer1* and *R3D1* (Obbard *et al.* 2006).

In contrast, Ago2 binds virally-derived siRNAs (Aliyari *et al.* 2008), and plays a key role in antiviral defence by cleaving viral genomes or transcripts (Rij *et al.* 2006). Additionally, Ago2 binds siRNAs derived from TEs in the soma, and inhibits transposition in the soma (Czech *et al.* 2008; Chung *et al.* 2008). In further contrast to *Ago1*, *Ago2* evolves exceptionally quickly, being among the top 3% of fastest-evolving genes in *D. melanogaster* (Obbard *et al.* 2006), and experiences strong selection (Obbard *et al.* 2006) and repeated selective sweeps (Kolaczkowski *et al.* 2011; Obbard *et al.* 2011). This rapid evolution may be driven by an evolutionary arms race to escape suppressors of RNAi (Obbard *et*

al. 2009b; Kolaczkowski *et al.* 2011), which are encoded both by viruses (reviewed in Bronkhorst and Rij 2014) and transposable elements (Nosaka *et al.* 2012).

As introduced in Section 1.2.3, Piwi subfamily Argonautes interact with piRNAs, which are canonically derived from transposable elements in the germline. Each Argonaute carries out a distinct role in the piRNA pathway: Ago3 and Aub amplify the TE signal by producing antisense and sense piRNAs respectively (reviewed in Luteijn and Ketting 2013), whereas Piwi suppresses TE transposition by directing the formation of heterochromatin (Sienski *et al.* 2012). These functional differences are reflected by contrasting rates of evolution, with *Aub* having a higher dN/dS ratio than *Ago3* and *Piwi* (Obbard *et al.* 2009b), and *Aub* of *Drosophila* displaying evidence for adaptive evolution across the whole gene (Simkin *et al.* 2013) and specifically in its PAZ domain (Kolaczkowski *et al.* 2011).

1.4. Thesis aims

Duplication of *Argonaute* genes has been anecdotally reported in isolated taxa, but a systematic analysis of the rate of *Argonaute* duplication is lacking. Additionally, the frequency of functional change after *Argonaute* duplication remains underexplored, despite the growing body of literature documenting functional divergence between *Argonaute* paralogues. In Chapter 2 I therefore quantify the rate of gene turnover, and analyse the change in evolutionary rate after duplication, in each of the Dipteran *Argonaute* subclades (*Ago1*, *Ago2*, *Ago3* & *Piwi/Aub*). This approach reveals how duplication rate varies between different *Argonaute* subclades and Dipteran lineages, and identifies hotspots of *Argonaute* expansion during Dipteran evolution. By quantifying the direction and extent of evolutionary rate change induced by duplication, this analysis allows inference of how *Argonaute* function changes after duplication, and reveals likely candidates for functional divergence.

A series of duplications of *Ago2* have been noted in *D. pseudoobscura* (Hain *et al.* 2010), but their age and distribution in other related species is unknown. The functions of these paralogues are also uncharacterised, but may be substantially different, given the range of functions carried out by *D. melanogaster* Ago2 and the capacity for paralogous genes to subfunctionalize to complementary roles. In Chapter 3 I carry out a detailed survey of *Ago2* in *D. pseudoobscura* and its relatives in the *obscura* group, in order to reconstruct the evolutionary history of these duplications, and estimate the age of the paralogues in different *obscura* group species. To characterise the functions of these paralogues, I also measure their

expression patterns in 3 *obscura* group species, in different tissues and under viral challenge. When analysed in a phylogenetic framework, these data reveal the speed and extent of *Argonaute* functional change after duplication.

Gene duplication often leads to altered selection pressures on the resulting paralogues, either relaxing constraints because of redundancy, or imposing stronger selection pressures due to the evolution of a new function. When combined with expression data, these data can indicate the strength of selection that may be driving the evolution of a derived function. In Chapter 4 I therefore gather population genetic data for *Ago2* paralogues in 3 species of the *obscura* group, in order to measure their rate of evolution and level of positive selection. When analysed in conjunction with the expression data gathered in Chapter 3, this allows me to infer the evolutionary relevance of any observed differences in expression patterns.

A powerful approach to experimentally explore differences in gene function that may be suggested by contrasting expression patterns is gene knockout by transgenesis. The complexity and cost of transgenic technologies have traditionally limited their use to model species; however, the recent development of CRISPR/Cas9 technology has allowed the production of gene knockouts in non-model species. In Chapter 5, I attempt to use the CRISPR/Cas9 technology to create knockouts of each *Ago2* paralogue in *D. pseudoobscura*, enabling me to test functional hypotheses derived from the expression patterns of paralogues, and quantify the extent of redundancy and divergence between these paralogues.

By quantifying the divergence in expression patterns between *Argonaute* paralogues of varying ages, this work will give a valuable insight into the speed with which functional divergence can occur after gene duplication. Additionally, the measurement of selection pressures acting on very recent paralogues will allow investigation of the relative contributions of positive selection and relaxed selection pressures in the early stages of paralogue evolution. Finally, the comprehensive analysis of Dipteran *Argonaute* expansion and evolution, combined with detailed characterisation of the function and evolutionary dynamics of *Ago2* paralogues in the *obscura* group, will be highly informative regarding the frequency of novel *Argonaute* functions, and the selection pressures under which they evolve.

2. Duplication and diversification of Dipteran *Argonaute* genes

2.1. Introduction

Argonaute genes affect a broad range of processes from development to antiviral immunity, and are found in almost all eukaryotes (Cerutti and Casas-Mollano 2006). They constitute an ancient gene family that was present in the common ancestor of extant prokaryotes and eukaryotes (reviewed in Swarts *et al.* 2014a), and which diverged into Ago and Piwi subfamilies early in eukaryotic evolution (Cerutti and Casas-Mollano 2006; Mukherjee *et al.* 2013). Argonaute proteins are effectors in the RNA interference-related (RNAi) pathways, which can be broadly defined as a system of nucleic acid manipulation by complementary base pairing between small RNA (sRNA) guides and long nucleic acid targets. Each sRNA is loaded into an Argonaute protein, which it guides to a target nucleic acid through complementary base-pairing, resulting in cleavage or translational inhibition of the target (reviewed in Sarkies and Miska 2014). Three broad classes of sRNA can be defined based on their sizes and interactors (reviewed in Kim *et al.* 2009): short interfering RNAs (siRNAs) are ~21-24 nucleotides (nt) long and are produced from viruses, transposable elements (TEs), and some long dsRNA products in the soma; microRNAs (miRNAs) are generally ~22-23nt long and are derived from host-encoded hairpin loops; and Piwi-interacting RNAs (piRNAs) are 24-29nt long, derived largely from intergenic repetitive elements (e.g. TEs) in the germline, and exclusively bind Piwi subfamily Argonaute proteins.

RNAi is well studied in *Arabidopsis thaliana*, where the *Argonaute* gene was first identified (Bohmert *et al.* 1998), and in the nematode *Caenorhabditis elegans*, where the RNAi mechanism was first characterized (Fire *et al.* 1998). Subsequent studies have reported *Argonaute* genes with diverse functions and differences in copy number across different eukaryotic clades (Mukherjee *et al.* 2013), illustrating

that RNAi pathways have a dynamic evolutionary history. For example, in the filamentous fungus *Neurospora crassa* the *Argonaute* homologue *QDE-2* plays a key role in the process of quelling, a form of RNAi-dependent homology-directed gene silencing (reviewed in Billmyre *et al.* 2013), and in budding yeasts *Argonaute* has been completely lost on several independent occasions (Drinnenberg *et al.* 2009; Drinnenberg *et al.* 2011).

Differences in *Argonaute* copy number and function are also found within the animals. For example, humans have 4 *Ago* genes, all of which carry out gene silencing, but only one of which (*Ago2*) has retained its catalytic ability to cleave nucleic acids (reviewed in Meister 2013). In the protostomes, the planarian *Schmidtea mediterranea* has 9 *Piwi* homologues (Palakodeti *et al.* 2008), two of which (*smedwi-2* & *smedwi-3*) play vital roles in regeneration by facilitating the differentiation of pluripotent neoblasts (Reddien *et al.* 2005; Palakodeti *et al.* 2008). In contrast, the *Piwi* genes and their associated piRNAs have been lost independently in several independent lineages of nematodes, with TE suppression carried out instead by DNA methylation mediated by RNA-dependent RNA polymerase and Dicer (Sarkies *et al.* 2015). Interestingly, this loss of *Piwi* has been accompanied by a massive expansion of other *Argonaute* genes in nematodes, with *C. elegans* encoding 25 *Argonautes*, 18 of which fall into the divergent worm-specific *Ago* (*WAGO*) clade: these associate with a novel class of sRNA (22G-RNAs) and carry out derived functions such as epigenetic memory formation (reviewed in Buck and Blaxter 2013).

Recent genome sequences and experimental data from isolated taxa have also revealed numerous arthropods with duplicates of *Argonautes*, some of which have novel and divergent functions. For example, the centipede *Strigamia maritima* has 2 paralogues of *Ago2* and 3 paralogues of *Piwi* (Chipman *et al.* 2014). Duplication of *Ago2* is also seen in the tick *Ixodes scapularis*, which has 3 *Ago2* paralogues, only 2 of which appear to function in antiviral defence (Schnettler *et al.* 2014). Larger expansions are seen in the aphid *Acyrthosiphon pisum*, which has 2 paralogues of *Ago3* and 8 paralogues of *Piwi*, some of which are expressed in the soma (in contrast to *Drosophila melanogaster*, where they are predominantly germline-specific) (Lu *et al.* 2011).

Despite this diversity, much of our functional and mechanistic understanding of arthropod Argonautes comes from studies of *D. melanogaster* (Kataoka *et al.* 2001; Li *et al.* 2002; Pal-Bhadra *et al.* 2004; Vagin *et al.* 2004; Kalmykova *et al.* 2005; Rij *et al.* 2006; Chung *et al.* 2008; Czech *et al.* 2008)), which has five *Argonaute* genes. Ago1 binds miRNAs and regulates gene expression by inhibiting translation of host transcripts (reviewed in Eulalio *et al.* 2008). Ago2 binds siRNAs from two sources. First, virus

specific small interfering RNAs (viRNAs), which are viral in origin and guide Ago2 to cleave viruses or their transcripts, forming an integral part of the antiviral defence mechanism (Li *et al.* 2002; Rij *et al.* 2006). Second, endogenous (endo)-siRNAs, which are derived from TEs, overlapping UTRs and other repetitive sequences in the soma (Chung *et al.* 2008; Czech *et al.* 2008). The remaining three Argonautes are Piwi subfamily members that bind piRNAs in the germline and surrounding tissues: Ago3, Aubergine (Aub) and Piwi (reviewed in Luteijn and Ketting 2013). The piRNAs are differentiated from miRNAs and siRNAs in *D. melanogaster* by their Dicer-independent production and their amplification through the "Ping-Pong" pathway, a positive feedback loop involving Ago3 and Aub (Li *et al.* 2009). In *D. melanogaster*, piRNAs guide Piwi to TEs in euchromatin, where it inhibits transposition (Kalmykova *et al.* 2005) by directing the formation of heterochromatin (Sienski *et al.* 2012).

Comprehensive analysis of *Argonaute* evolution at a eukaryotic, or even metazoan, scale is hindered by limited taxon sampling, wide variation in evolutionary rate, and the presence of ancient and recent duplications and losses (discussed by Philippe *et al.* 2011). The Diptera provide an opportunity to study *Argonaute* evolution in an order that is densely sampled and less divergent, but still shows variation in *Argonaute* copy number and function. Previous reports of *Argonaute* duplication in the Diptera have been limited to isolated taxa, such as the house fly *Musca domestica* (Scott *et al.* 2014), *Drosophila pseudoobscura* (Hain *et al.* 2010), and three mosquito species (Campbell *et al.* 2008). These mosquito duplicates appear to have evolved derived functions: several *Piwi* paralogues in *Aedes aegypti* (Vodovar *et al.* 2012; Schnettler *et al.* 2013a) and *Aedes albopictus* (Morazzani *et al.* 2012) are expressed in the soma, and at least one of the somatically-expressed *Piwi* duplicates in *Aedes albopictus* appears to have functionally diverged to a novel antiviral function (Schnettler *et al.* 2013a).

Gene duplications, such as those that gave rise to the diversity of eukaryotic RNAi pathways, are often associated with changes in selection pressure, which can cause differences in evolutionary rate between the resulting paralogues. Such a difference is seen between *Ago1* and *Ago2* in *D. melanogaster*, with *Ago1* evolving slowly under purifying selection, and *Ago2* evolving rapidly under positive selection (Obbard *et al.* 2006; Obbard *et al.* 2009b; Kolaczkowski *et al.* 2011). Gene duplication can also drive contrasting changes in evolutionary rate in paralogues compared to single-copy orthologues (reviewed in Hahn 2009). If each paralogue specializes to a particular ancestral role, evolutionary rate is expected to increase only in the branches immediately following duplication, before being constrained again by the established functions (e.g. Nielsen *et al.* 2010). Alternatively, if one or both new duplicates evolve a new function under positive selection, all the branches subtending the duplication can evolve more rapidly than the pre-duplication rate (e.g. Morandin *et al.* 2014). These characteristic patterns of selec-

tion therefore enable us to use analyses of evolutionary rate to gain an insight into functional evolution.

Here we take advantage of the diversity available in the sequenced genomes and transcriptomes of Diptera to analyse patterns of *Argonaute* duplication and sequence evolution across 86 species. Contrasting rates of duplication and evolution are commonly associated with differences in function and selection pressure. We find a higher rate of evolution in *Ago2* and *Ago3*, a higher rate of gene turnover in *Piwi/Aub*, and we estimate the date of the duplication that led to the separate *Piwi* and *Aub* subclades. We also find that paralogues of *Ago2*, *Ago3* and *Piwi/Aub* evolve more rapidly after duplication, indicating potential divergence into novel and strongly selected functions.

2.2. Methods

2.2.1. Identification of *Argonaute* homologues

We used tblastx and tblastn (Altschul *et al.* 1997) to identify *Argonaute* homologues in the genomes and transcriptomes of 86 Dipteran species found in Genbank, Flybase, Vectorbase, Diptex, the NCBI Transcriptome Shotgun Assembly or other unpublished transcriptomes (sequence metadata available on request). For each species, we used *Argonautes* from the closest well-annotated relative as queries, or *D. melanogaster* if no homologue from a close relative was available. Where blast returned multiple partial hits, we assigned hits to the correct query sequence by aligning all hits from the target species to all *Argonautes* from the query species, and inferring a neighbour-joining tree. For each query sequence, partial blast hits were then manually curated into complete genes using Geneious v5.6.2 (http://www.geneious.com, Kearse *et al.* 2012). For some species of *Drosophila*, PCR and Sanger sequencing was used as no transcriptomic or genomic data were available, and novel sequences have been submitted to Genbank (accession numbers KR012647-KR012696).

2.2.2. Phylogenetic analysis of Dipteran *Argonautes*

We initially assigned homologues into subclades (*Ago1*, *Ago2*, *Ago3* and *Piwi/Aub*) based on a Bayesian gene tree rooted between the Ago and Piwi subfamilies, with ambiguous alignment positions removed

using Gblocks (Castresana 2000) and with the wasp *Nasonia vitripennis* as the outgroup for each subclade. To minimise the loss of information when removing ambiguous positions, we re-inferred separate Bayesian gene trees for each subclade with no outgroup, using new alignments with ambiguous positions identified by eye and removed. Sequences were aligned using translational MAFFT (Katoh *et al.* 2002) with default parameters. All phylogenies were inferred using the Bayesian approach implemented in MrBayes v3.2.1 (Ronquist and Huelsenbeck 2003) under a nucleotide model, assuming a GTR substitution model with 3 unlinked codon-position classes, gamma-distributed rate variation between sites with no invariant sites, and inferred base frequencies. We ran each analysis for a minimum of 50 million steps, or as long necessary for the tree topologies to reach stationarity (standard deviation of split frequencies between duplicate independent runs <0.01; PSRF 1 and ESS>1000 for all parameters). Samples from the posterior were recorded every 10,000 steps, and a maximum clade credibility tree was inferred using TreeAnnotator (Drummond *et al.* 2012).

2.2.3. Gene turnover rates

To quantify the rate of gene duplication and loss during *Argonaute* evolution, we estimated the rate of gene turnover (λ, the number of gains or losses per million years) for each *Argonaute* subclade using CAFE v3.1 (Han *et al.* 2013a). We also tested whether subclades differed significantly in their rates of gene turnover by using 1000 replicates of CAFE's Monte Carlo resampling procedure. This generates an expected distribution of gene family sizes under a birth-death model, conditioned on the species topology and a set λ value (which we fixed at the value estimated for each subclade), thus providing an estimate of the *p*-value for each of the other subclades. To mitigate potential bias introduced by incomplete genome assemblies, turnover analyses only included species that had at least one gene in each subclade (66 of total 86 species). To assess the potential impact of searching transcriptomes, which will only detect expressed genes (and may therefore lead to erroneous inference of gene loss and falsely inflate the rate of gene turnover), we repeated these analyses with rates of gene gain and loss estimated separately. We find similar results when comparing rates of gene gain and gene turnover, suggesting that missing data is having a negligible effect on our estimates of gene turnover rate.

To provide the independent species-level tree topology for all 66 taxa that is required for this analysis, we manually combined the high-confidence multigene phylogenies presented in Wiegmann *et al.* 2011 and Misof *et al.* 2014. Where these reference trees lacked the relevant taxa (e.g. relationships

below the level of family), we either referred to other published multi-gene phylogenies (Linde *et al.* 2010, Zhang *et al.* 2010 and Dyer *et al.* 2008 for Drosophilidae, *Bactrocera* and *Glossina*, respectively), or inferred a Bayesian phylogeny using the *arginine kinase* gene (*Culicidae*, parameters as above). Conditional on this species topology, we estimated relative branch lengths using BEAST v1.7 (Drummond *et al.* 2012) and a translational MAFFT alignment of the 1:1:1 orthologue *Ago1*, constraining the dates of key nodes to previously inferred dates derived from fossil evidence (as used by Wiegmann *et al.* 2011: root=245mya, Brachycera=200mya, Cyclorrhapha=150mya, Schizophora=70mya). As our primary concern is the difference in relative rates of gene gain and loss for the different subclades, inaccuracies of the absolute timescale will have minimal impact.

2.2.4. Evolutionary rate and positively selected residues

To infer the relative rates of synonymous and non-synonymous substitution (dN/dS=ω) averaged across all sites, we used codeml (PAML, Yang 1997) to fit model M0 (single ω) separately for each subclade (*Ago1*, *Ago2*, *Ago3*, *Piwi/Aub*), conditional on the alignment and tree topology. To test for significant differences between these subclade-specific rates, we fixed ω for each subclade at the value estimated for each of the other subclades, and used Akaike weights to compare the likelihood of these fixed ω values to the likelihood of the ω value estimated from the data for that subclade.

To estimate the change in evolutionary rate after duplication, and to test whether duplicates experienced a transient or sustained change in evolutionary rate, we fitted two variants of the M0 model, each with two separate ω parameters estimated for different branches of the gene tree (illustrated in Figure 2.1). To test for a transient change in evolutionary rate directly after duplication, we fitted a model (which we term "Immediate") that specified one ω for branches immediately after a duplication event, and another ω for all other branches. To test for a sustained change in evolutionary rate following duplication, we fitted a second model (which we term "All descendants") that specified one ω for all branches arising from a duplication event, and another ω for all other branches. For each subclade, Akaike weights were used to estimate the relative support for the M0, "Immediate" and "All descendants" models. To test for positively selected residues in each subclade, we used likelihood ratio tests (LRTs) to compare the fit of two models, each with two site classes. In both models, ω of the first 'constrained' site class was a discretised beta-distribution with eight classes. The models differ in that in the first model (the null "M8a" model) ω of the second 'positively selected' site class is fixed at 1 (neutrality), while in the sec-

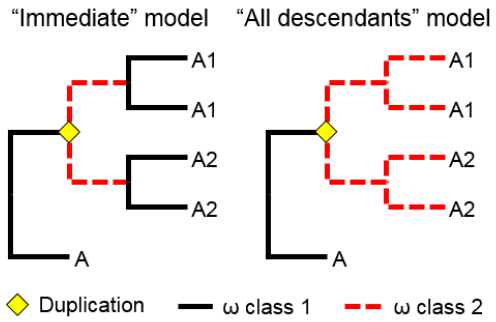


Figure 2.1.: Models fitted to analyse evolution after Argonaute duplication.
"Immediate" models the expectation if selection pressures change only briefly after duplication, whereas "All descendants" models the expectation if paralogues evolve at a consistently different rate.

ond model (the "M8" model) ω of the second site class is constrained to exceed 1. If the LRT indicated a significantly better fit for M8 than M8a given the parameters in the model, individual residues were classed as positively selected if they had a Bayes Empirical Bayes (BEB) posterior probability of >95% that ω>1.

To assess the potential impact of false positives introduced by misalignments (see Jordan and Goldman 2012), we ran all codeml analyses on two alignments for each subclade, the first with no trimming of ambiguous alignment positions (which may represent genuinely rapidly evolving sites), and the second with ambiguous alignment positions identified by eye and removed. All estimates and statistical comparisons of evolutionary rates outlined above were very similar with and without alignment screening: we therefore report results estimated from the untrimmed alignments.

To test for gene conversion between paralogues, we used GARD (Kosakovsky Pond *et al.* 2006) with default settings to identify potential recombination breakpoints in each subclade. This identified no breakpoints in the *Ago2* and *Ago3* subclades, and 1 potential breakpoint each in the *Ago1* and *Piwi/Aub* subclades. Although we could not rule out gene conversion between paralogues in each of these subclades (which can lead to erroneous support for positive selection (Casola and Hahn 2009)), we found very few positively selected sites in both of these subclades, so this effect is likely to have little or no effect on our analyses.

2.2.5. Domain mapping and structural modelling

To investigate the distribution of rapidly evolving sites across the domain architecture of each *Argonaute* gene, we inferred the location of each domain in each *Argonaute* gene by searching the Pfam database (Finn *et al.* 2009), and then mapped the mean estimate of ω for each residue across the gene (derived from the BEB posterior distribution under the M8 model in PAML (Yang 1997)). To describe evolutionary rate heterogeneity in the protein structures of each gene, we built structural models based on published X-ray crystallography structures: the *D. melanogaster* Ago1 structure was based on human Ago1 (Faehnle *et al.* 2013), and the structures of *D. melanogaster* Ago2, Ago3 and Piwi were based on human Ago2 (Schirle and Macrae 2012). We used the MODELER software in the Discovery Studio 4.0 Modeling Environment (Accelrys Software Inc., San Diego, 2013) to calculate ten models, and selected the most energetically favourable for each protein. The model quality was assessed with the 3D-profile option, which compares the compatibility of the 3D structure and the sequence. For *D. melanogaster* Ago2, we replaced the inferred PAZ domain structure with the *D. melanogaster* Ago2 PAZ domain structure that has previously been resolved using X-ray crystallography (Song *et al.* 2003). We then mapped ω onto each residue of the structure using PyMol v.1.7.4.1 (Schrödinger, LLC). For both analyses, we used estimates of ω from trimmed alignments to provide a conservative estimate of residue-specific evolutionary rate. Sites that were trimmed out of the alignment were excluded when mapping ω across domains, and were set as ω=0 when mapping ω across structures.

2.2.6. Functional divergence of *Glossina morsitans Argonautes*

To explore the possibility of functional divergence between the *Argonaute* duplicates of the tsetse fly *Glossina morsitans*, we took advantage of transcriptomic data from the carcasses of lactating and non-lactating females (Benoit *et al.* 2014), and from salivary glands of parasitized and unparasitized females (Telleria *et al.* 2014) (accessions SRX287393, SRX287395, SRX342351 & SRX342350 respectively). For each sample, reads were mapped to *G. morsitans* genomic scaffolds using TopHat2 (Kim *et al.* 2013a), processed using picard tools and samtools (Li *et al.* 2009), and read coverage at each gene was estimated using HTSeq (Anders *et al.* 2015). Coverage was converted to reads per kilobase per million reads (RPKM) (Mortazavi *et al.* 2008) to account for variation in transcript length between genes and sequencing effort between samples.

2.3. Results

2.3.1. Duplications of *Ago2*, *Ago3* and *Piwi* occur in different Dipteran lineages

To explore the evolutionary dynamics of *Argonautes* in the Diptera, we quantified the rate of duplication and evolution of *Argonautes* from 86 Dipteran species. We find numerous expansions of *Ago2* and *Piwi/Aub* (including the origin of canonical *Piwi* and *Aub* themselves from their Piwi subfamily ancestor; illustrated in Figures 2.2, 2.4 & 2.6). This is in sharp contrast to *Ago1*, which is present as a single copy orthologue in all Diptera (Figures 2.2 & 2.3), and *Ago3*, which has duplicated only rarely (Figures 2.2 & 2.5).

We also find that the expansions of *Ago2* and *Piwi/Aub* have occurred in different taxa and at different times (Figures 2.2, 2.4 & 2.6). Most duplications of *Ago2* have occurred in the Brachycera, with numerous duplications within the *Glossina* (<84mya), and the *Drosophila obscura* subgroup (<50mya) (Hain et al. 2010). Single duplications of *Ago2* have occurred in the Brachycerans *Drosophila willistoni*, *Scaptodrosophila deflexa*, *Musca domestica* and *Megaselia abdita*, and in the Nematocerans *Belgica antarctica*, *Culex quinquefasciatus* and *Sitodiplosis mosellana* (illustrated in Figure 2.4).

In contrast, most duplications of *Piwi/Aub* have occurred in the Nematocera. Numerous duplications have occurred in the mosquitoes (*Aedes spp.*, *Anopheles spp.* and *Culex quinquefasciatus*) <65mya, and multiple copies are seen in *Lutzomyia longipalpis*, *Sitodiplosis mosellana*, *Chironomus riparius*, *Belgica antarctica* and *Corethrella appendiculata* (shown in Figure 2.6). A duplication at the base of the Brachycera between 182 and 156mya gave rise to the separate *Aub* and *Piwi* subclades (as they occur in *D. melanogaster*, labelled in Figures 2.2 & 2.6). Within these subclades duplications have occurred rarely, only being observed in *Piwi* of the drosophilids *Phortica variegata* and *Scaptodrosophila deflexa*, and in *Aub* of *Teleopsis* species.

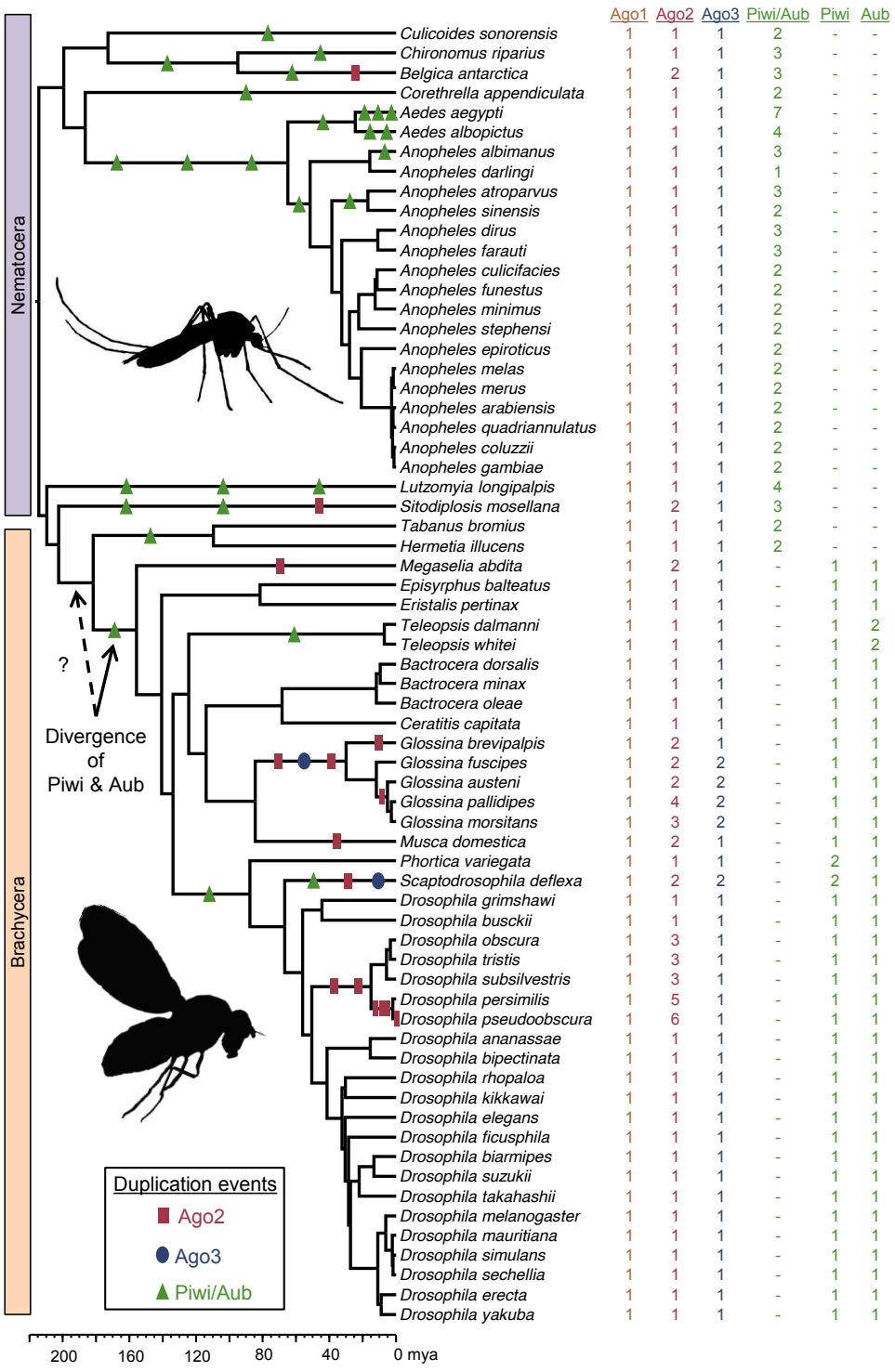


Figure 2.2.: Counts of the number of genes in each *Argonaute* subclade.
Shown are counts for a subsample of 66 Dipteran species with at least one gene in each subclade (out of a total of 86 species). Shapes indicate gene duplication events on branches, inferred by parsimony, and are illustrative only (gene loss is not depicted due to space constraints, thus for some taxa gene counts do not correspond to the number of gene duplications). We estimate the divergence of *Piwi* & *Aub* at 182-156mya (calibrated using dates for key nodes as in Wiegmann *et al.* 2011).

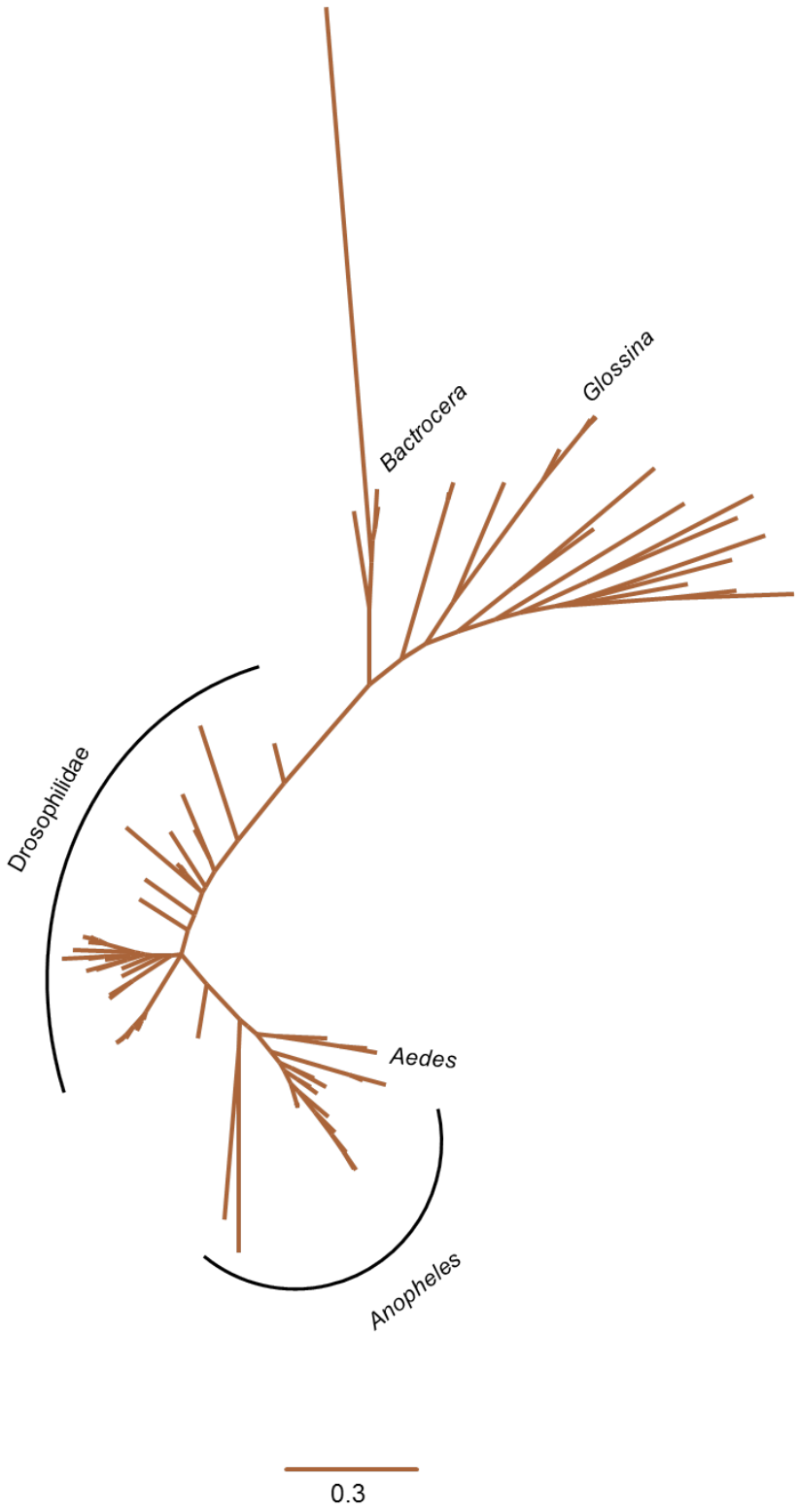


Figure 2.3.: unrooted Bayesian gene tree of Dipteran *Ago1*.
Ago1 has not duplicated and evolves very slowly in the Diptera, resulting in a lack of information that introduces some incongruence between the gene and species trees.

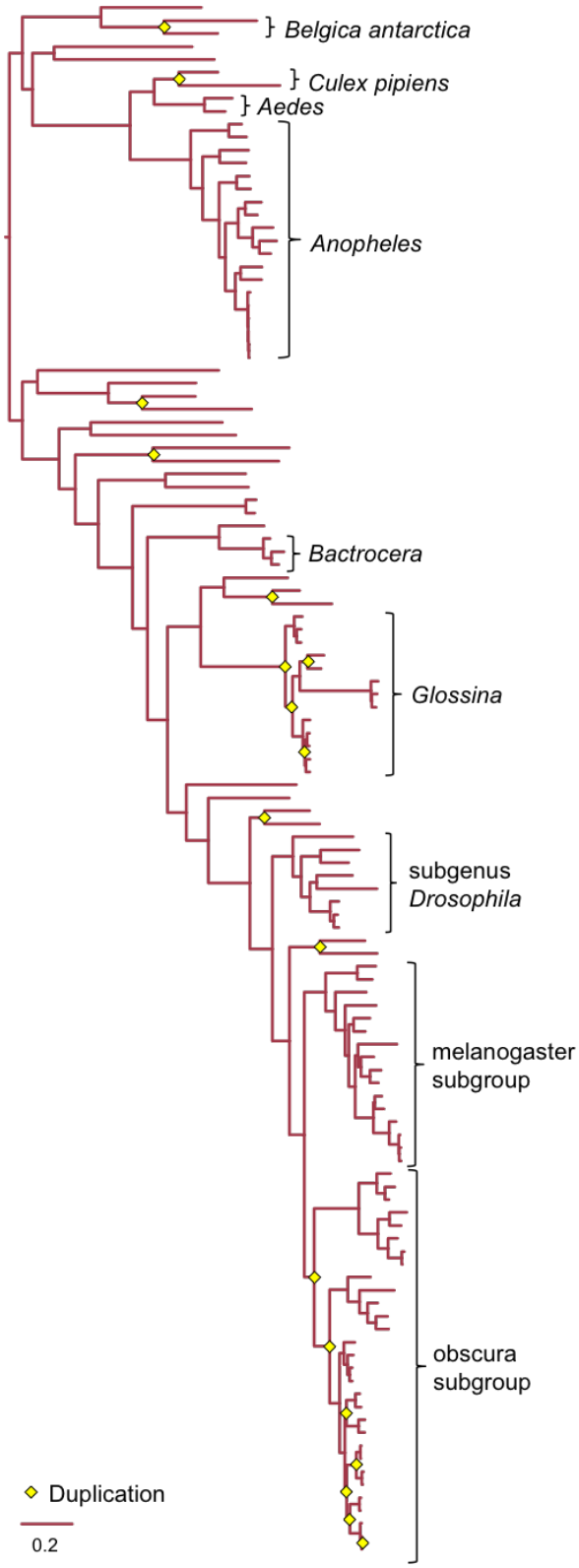


Figure 2.4.: Rooted Bayesian gene tree of Dipteran *Ago2*.
Most duplications of *Ago2* have occurred in the Brachycera, including several in *Glossina* and the *D. obscura* group.

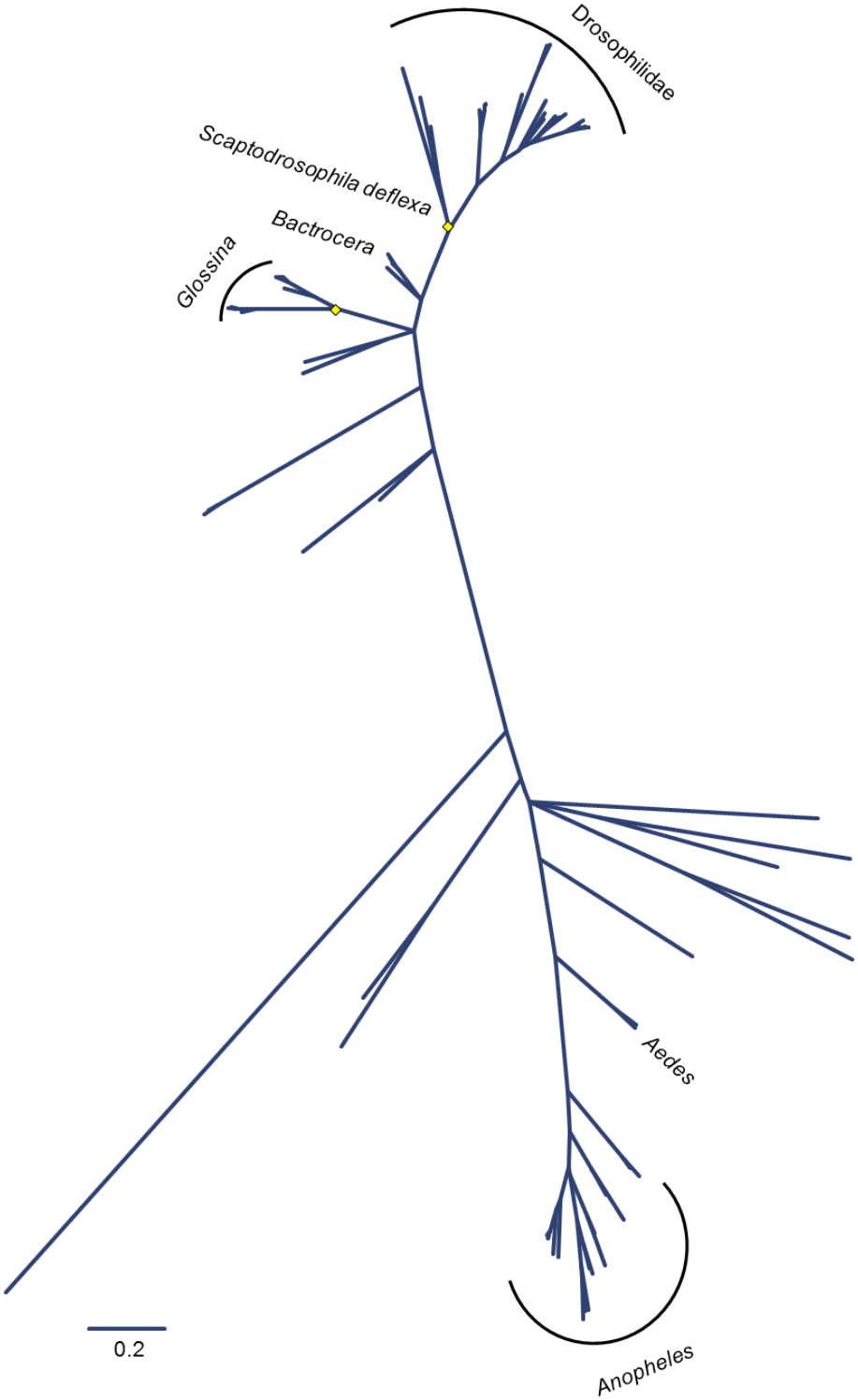


Figure 2.5.: Unrooted Bayesian gene tree of Dipteran *Ago3*.
Ago3 has duplicated twice in the Diptera, once in Scaptodrosophila deflexa and once at the base of the Glossina species.

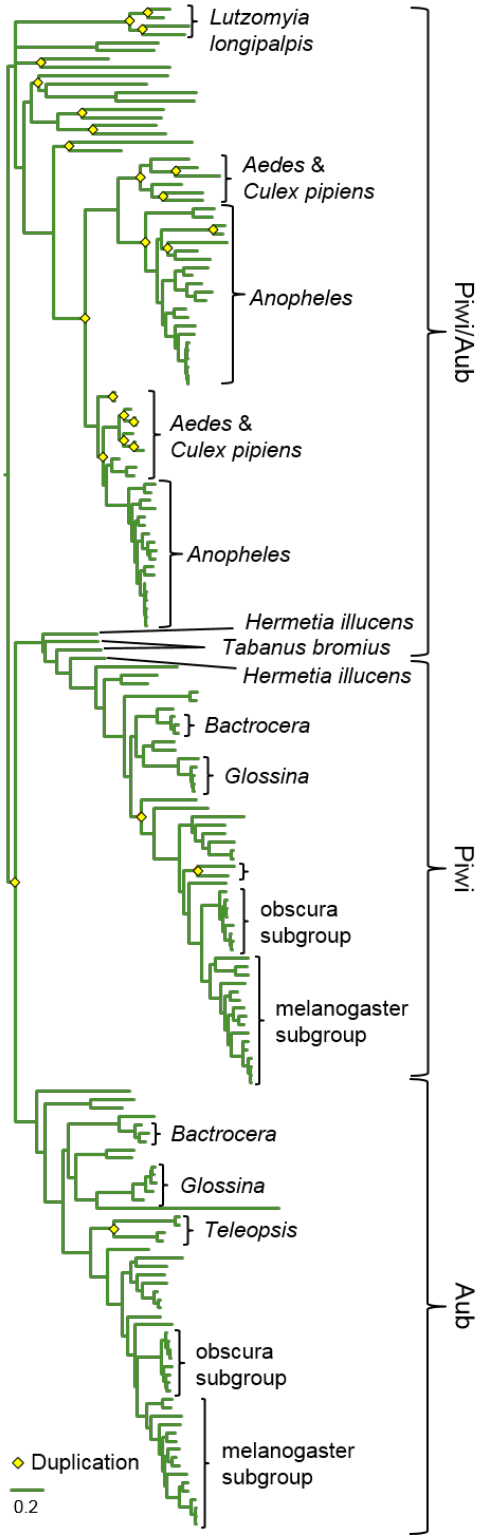


Figure 2.6.: Rooted Bayesian gene tree of Dipteran *Piwi/Aub*.
Most duplications of *Piwi/Aub* have occurred in the Nematocera; however, there was also a duplication early in the Brachycera that led to separate *Piwi* and *Aub* lineages.

2.3.2. *Ago2* and *Piwi* have significantly higher duplication rates than *Ago1* and *Ago3*

To quantify these contrasting patterns of duplication, we used CAFE (Han *et al.* 2013a) to estimate the rate of gene turnover (λ, the number of gains or losses per million years) in each *Argonaute* subclade. We find that gene turnover rate varies considerably among the subclades, with *Ago2* (λ=0.0017), *Ago3* (λ=0.0003) and *Piwi/Aub* (λ=0.0012) having significantly higher gene turnover rates than *Ago1* (λ=1.1516x10^{-10}) (ρ<0.001 based on the expected distribution of gene family sizes under a birth-death model, with λ fixed at the value estimated for *Ago1*). We also find that *Ago2* and *Piwi/Aub* have significantly (ρ<0.001) higher gene turnover rates than *Ago3*, but do not differ significantly from each other (ρ=0.13).

2.3.3. *Argonaute* genes show contrasting rates of evolution before and after duplication

To quantify the rate of protein evolution in each *Argonaute* subclade, and to identify any sites evolving under positive selection, we fitted models using codeml (PAML; Yang 1997). These analyses revealed that *Ago2* has the highest non-synonymous to synonymous substitution ratio (ω=0.14$\pm$0.0015), followed by *Ago3* (ω=0.12$\pm$0.0015), *Piwi/Aub* (ω=0.09$\pm$0.0009), and lastly *Ago1* (ω=0.01$\pm$0.0002). All rates were significantly different from each other (Akaike weight = 1.000 to 3dp in all cases). Scans for positively-selected sites identified five candidate sites in *Ago3* and one in *Piwi/Aub*; however, in both cases the M8 model was not significantly more likely than the null M8a model (for ω estimates and likelihoods under all models see Tables A.1, A.2 & A.3).

To test whether the relative rate of protein evolution changes following duplication, we calculated the likelihood of the data for *Ago2*, *Ago3* and *Piwi/Aub* under two models: the first with a separate evolutionary rate for branches immediately after a duplication event (the "Immediate" model); and the second with a separate rate for all branches subtending a duplication event (the "All descendants" model) (see Figure 2.1). For *Ago2* and *Ago3*, the "All descendants" model had all support (Akaike weight = 1.000 to 3dp for each). For *Piwi/Aub*, however, the "Immediate" model had all support (Akaike weight = 1.000 to 3dp). In each subclade the evolutionary rate increased after duplication, with *Ago2* having the highest rate and *Piwi/Aub* the lowest (Figure 2.7).

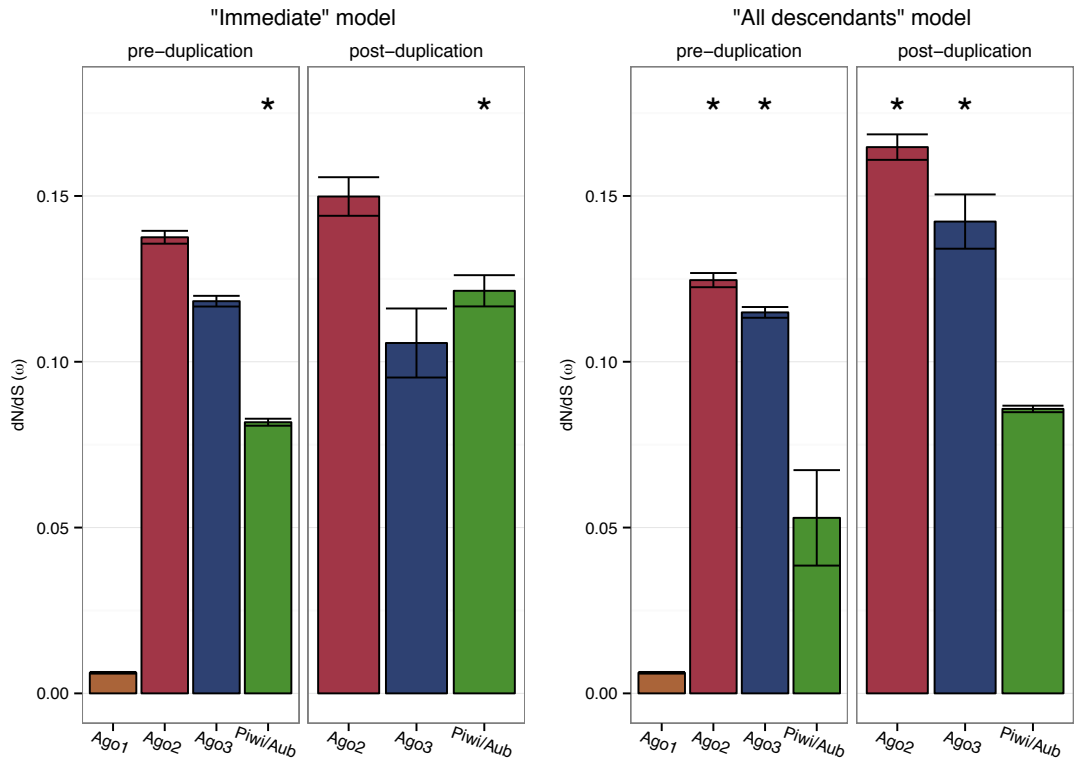


Figure 2.7.: Evolutionary rate estimates before and after duplication, under the "Immediate" and "All descendants" models.
Asterisks indicate the most highly supported model: duplicates of *Piwi/Aub* evolve more quickly immediately after duplication, whereas *Ago2* and *Ago3* paralogues experience a sustained increase in evolutionary rate.

2.3.4. Ago2 displays hotspots of evolution at the RNA binding pocket entrance

To investigate the distribution of rapidly evolving residues, we mapped ω estimates onto the domains and structures of each Argonaute. We find that rapidly evolving residues are spread across all domains of Ago2, Ago3 and Piwi/Aub (Appendix A). We also find that Ago2 has clusters of more rapidly evolving residues at the entrance to the RNA binding pocket, which are not found in Ago1, Ago3 or Piwi/Aub (Figure 2.8). In contrast, the residues that directly contact the sRNA guide are conserved in all Argonautes (Figure 2.8).

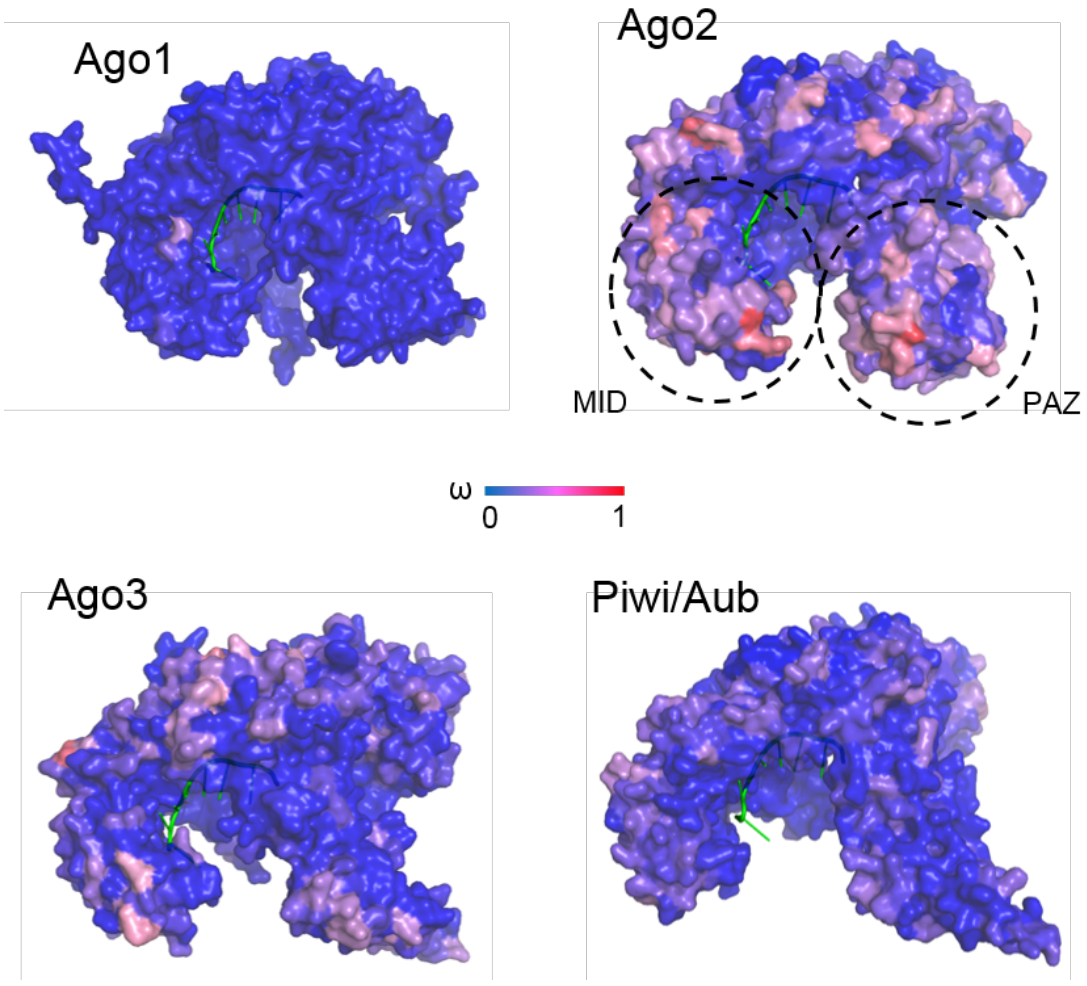


Figure 2.8.: Evolutionary rates mapped onto the protein structures of Dipteran Argonautes.
Residues are coloured in the blue to red spectrum according to their evolutionary rate, with those closer to the red evolving more rapidly. The short RNA in the binding pocket of each protein is coloured bright green. In Ago2, hotspots of evolution are seen at the entrance of the RNA binding pocket; in contrast, evolutionary rate (ω) across the structure of Ago1, Ago3 and Piwi/Aub is uniformly low. The MID and PAZ domains are indicated for Ago2.

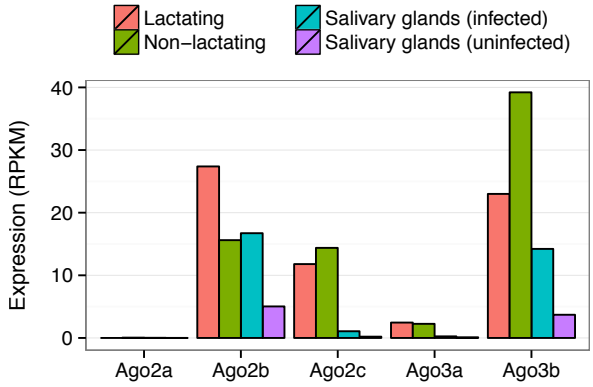


Figure 2.9.: Tissue-specific expression patterns of *Ago2* and *Ago3* in *G. morsitans*.
Paralogues of *Ago2* and *Ago3* show divergent patterns of expression. This is particularly unexpected for *Ago3b*, which is expressed in the salivary glands of *G. morsitans*, despite *D. melanogaster Ago3* being germline-specific.

2.3.5. Differential expression of *Ago2* and *Ago3* paralogues in *G. morsitans*

To test for functional divergence between paralogues of *Ago2* and *Ago3* in *G. morsitans*, we measured the expression of each *Ago2* and *Ago3* paralogue using publicly-available transcriptome data for 4 different tissues. We find differential expression between paralogues of *Ago2*, with *Ago2b* expressed in all tissues, *Ago2c* expressed in lactating and non-lactating females but absent from the salivary glands, and *Ago2a* unexpressed in all tissues (Figure 2.9). We also find differential expression between *Ago3* paralogues: while *Ago3b* is expressed in almost every tissue, *Ago3a* is expressed at only a very low level in lactating and non-lactating females, and is completely absent from the salivary glands.

2.4. Discussion

Our results reveal contrasting patterns of selection and duplication during Dipteran *Argonaute* evolution. The low evolutionary rate and lack of gene turnover in *Ago1* are in agreement with previous studies in *D. melanogaster* (Obbard *et al.* 2006; Obbard *et al.* 2009b), and are consistent with the idea that *Ago1* is carrying out a conserved gene regulatory role in the Diptera as a whole. In contrast, the better fit of the "All descendants" model to duplications in *Ago2* and *Ago3* indicates that paralogues in these subclades have experienced a sustained increase in evolutionary rate, possibly driven by the acquisition of new

functions.

This result is particularly noteworthy in *Ago2*, which is already among the top 3% of the fastest evolving genes in *D. melanogaster* (Obbard *et al.* 2006). Our protein structural modelling suggests that one possible hotspot of adaptive evolution for these paralogues may be the entrance of the RNA binding pocket (Figure 2.8). Relaxation of selection pressures on these residues is unexpected as they form alpha-helices (Panchenko *et al.* 2005; Liu *et al.* 2008), and their rapid evolution may instead be caused by undetected positive selection. The pocket is formed by the PAZ and MID domains, which bind the sRNA guide and form the channel in which the target RNA sits during cleavage (Schirle and Macrae 2012). Given the location of these residues at the mouth of the binding pocket away from the sRNA guide (Figure 2.8), such positive selection could be driving differences in target RNA binding and cleavage. Alternatively, selection could be imposed by viral suppressors of RNAi, which are encoded by numerous viruses to inhibit the antiviral RNAi response, and several of which prevent target cleavage by Ago2 (Wang *et al.* 2006; Mierlo *et al.* 2012; Mierlo *et al.* 2014).

Functional differences between most Dipteran *Argonaute* paralogues have not been characterized experimentally. However, transcriptome data are available for some *G. morsitans* tissues (Benoit *et al.* 2014; Telleria *et al.* 2014), allowing us to explore the possibility of functional divergence in one of the few Dipteran species with expansions of both *Ago2* and *Ago3*. We found differential expression between both sets of paralogues, as well as high expression of *Ago3b* in the salivary glands, which increased upon infection with *T. brucei* (Figure 2.9). Although this observation awaits replication, the canonical germline-specific role of *Ago3* in *D. melanogaster* (Li *et al.* 2009) makes any expression of *Ago3b* in the salivary glands unexpected, and suggests that paralogues of *Ago3* in *G. morsitans* have undergone rapid functional divergence to roles beyond TE suppression. Strikingly, this reflects the general patterns noted for somatically-expressed Piwis across the eukaryotes, which have evolved diverse roles in epigenetic regulation, genome rearrangement and somatic development (reviewed in Ross *et al.* 2014).
The better fit of the "Immediate" model to duplications of *Piwi/Aub* suggests that the evolutionary rate of paralogues in these subclades has been constrained soon after duplication, which may indicate a burst of adaptation to specialize to existing (but distinct) roles. Nevertheless, this is complicated by the age of the *Piwi/Aub* duplicates, the majority of which are very recent and have occurred in terminal lineages. This means that many of the branches immediately after duplication are also terminal branches, and clouds the difference between the "Immediate" and "All descendants" models for *Piwi/Aub*.

An important exception to this is the divergence of separate *Aub* and *Piwi* (*sensu stricto*) lineages, which

resulted from a much older duplication in the Piwi subfamily *Piwi/Aub* subclade. We estimate that this divergence, which happened at the base of the Brachycera, occurred between 182 and 156mya (Figure 2.2). However, the ambiguous identities of the two *Piwi/Aub* paralogues in *Hermetia illucens* and *Tabanus bromius* (Figure 2.6) mean that this duplication could have occurred slightly earlier (200mya). Under either scenario, *Piwi/Aub* paralogues in the vast majority of Nematoceran taxa (including all mosquitoes) are equally homologous to *Aub* and *Piwi*, which in *D. melanogaster* have specialized to distinct roles in the "Ping-Pong" piRNA amplification cycle and TE silencing respectively (reviewed in Luteijn and Ketting 2013). Our finding that this specialization occurred after their divergence from Nematoceran *Piwi/Aub* indicates that piRNA biogenesis and TE silencing may rely on different suites of *Argonaute* genes in the Nematocera. This has recently been confirmed for *A. aegypti* somatic cells, in which Piwi5, Piwi6 and Ago3 carry out "Ping-Pong" amplification of TE-derived piRNAs, while only Piwi5 acts with Ago3 in the "Ping-Pong" amplification of virally-derived piRNAs (Miesen *et al.* 2015). The evolution of this somatic antiviral function, which has also been confirmed in adult mosquitoes (Morazzani *et al.* 2012), demonstrates that the *Piwi/Aub* subclade is not constrained to a germline-specific anti-TE role, but can evolve novel and highly derived functions.

Analysis of gene turnover can be influenced by variability between species in the quality of genome assembly, with poorly assembled genomes potentially leading to genes missing from gene sets (discussed in Han *et al.* 2013b). This can in turn lead to both false inferences of gene loss in the poorly assembled genome, and false inference of gene gain in related species. Additionally, gene turnover analysis can be influenced by recent duplication, which produces very similar paralogues that may be collapsed into one gene during genome assembly, and by gene conversion, which may result in the false inference of gene loss. While we cannot rule out the influence of these last two factors, we minimised the effect of assembly quality by including only species with well assembled genomes (see Section 2.2.3).

In conclusion, we show that Dipteran *Argonaute* subclades differ widely in their rates of gene turnover and protein evolution. This suggests a high degree of evolutionary lability in *Argonaute* function across a wide range of taxa, with some genes maintaining a conserved role while others evolve new functions. This is in agreement with previous experimental studies in isolated taxa and our preliminary analysis of publicly available *G. morsitans* transcriptomic data. We also find that gene turnover rates vary widely between taxa, and that paralogues of *Ago2*, *Ago3* and *Piwi* experience distinct selection pressures after duplication. This points to a possible evolutionary mechanism for the functional overlap frequently observed between different *Argonaute* subclades: different taxa may undergo expansions of different *Argonaute* subclades (as we have shown for *Ago2* in the Brachycera versus *Piwi/Aub* in the Nemato-

cera), which can then specialize to distinct aspects of existing roles, or adapt to fulfil new functions.

3. Phylogenetic and expression analysis of *Argonaute2* paralogues in the *obscura* group

3.1. Introduction

Gene duplication is one of the most prevalent processes by which novel gene functions can evolve (Ohno 1970). While most new paralogues are lost through the unconstrained accumulation of substitutions (Lynch and Conery 2000; Hughes and Liberles 2007), a minority can be retained by selection. These paralogues may keep their ancestral function(s) (e.g. ribosomal RNA genes; Eickbush and Eickbush 2007), specialise to one existing function (e.g. *engrailed*; Force *et al.* 1999), or evolve an entirely new function (e.g. *RNase1B*; Zhang *et al.* 2002). For many paralogues in *Drosophila*, however, this is preceded by a period of testis-biased or testis-specific expression (Betrán *et al.* 2002; Levine *et al.* 2006; Assis and Bachtrog 2013; Palmieri *et al.* 2014). The "out-of-the-testis" hypothesis suggests that this pattern is due to the high transcriptional rate of the testis, which causes otherwise non-functional paralogues to be expressed, exposing them to strong selection against deleterious mutations and increasing their rate of retention (Kaessmann *et al.* 2009; Kaessmann 2010). As they age, these testis-specific paralogues evolve more protein-protein interactions and specialize to other tissues, indicating the evolution of derived functions (Assis and Bachtrog 2013).

Gene duplication has been a key influence in the evolution of the *Argonaute* gene family, which arose in early prokaryotes and is now found in the majority of eukaryotes (Swarts *et al.* 2014b). Duplication early in eukaryotic evolution produced two distinct *Argonaute* subfamilies (Ago & Piwi; Cerutti and Casas-Mollano 2006), and subsequent duplications have produced *Argonaute* genes with derived func-

tions in a diverse range of taxa (detailed in Chapter 2). In *D. melanogaster*, there are two members of the Ago subfamily (*Ago1* and *Ago2*) and three members of the Piwi subfamily (*Ago3*, *Aubergine* (*Aub*) and *Piwi*). All of the Piwi subfamily proteins bind Piwi-interacting RNAs (piRNAs) derived from host gene 3' UTRs (Robine *et al.* 2009) and transposable elements (TEs) (reviewed in Luteijn and Ketting 2013). Those that bind TE-derived piRNAs suppress transposition by either amplifying the piRNA signal through the "Ping-Pong" pathway (Ago3 & Aub; Li *et al.* 2009), or directing heterochromatin formation (Piwi; Sienski *et al.* 2012). In contrast, Ago1 binds miRNAs (Ha and Kim 2014), and regulates gene expression by inhibiting the translation of host transcripts (reviewed in Eulalio *et al.* 2008). Lastly, Ago2 plays key roles in antiviral defence (Rij *et al.* 2006), somatic TE suppression (Czech *et al.* 2008; Chung *et al.* 2008) and dosage compensation (Menon and Meller 2012).

This variety of functions is reflected in the distinct expression patterns that *Argonaute* genes exhibit, some of which appear to be evolutionarily ancient. In both vertebrates and invertebrates, members of the Ago subfamily that function in gene regulation are expressed in the vast majority of developmental stages and tissues (González-González *et al.* 2008; Williams and Rubin 2002; Celniker *et al.* 2009). Within each cell, this *Ago*-mediated gene regulation is localized to P-bodies, which are cytoplasmic foci of mRNA degradation (reviewed in Pfaff and Meister 2013). Similarly, *Agos* involved in antiviral defence are expressed ubiquitously in *D. melanogaster* (Celniker *et al.* 2009), the moth *M. sexta* (Garbutt and Reynolds 2012) and the tick *I. scapularis* (Schnettler *et al.* 2014). In these species expression of antiviral *Agos* is constitutive (Celniker *et al.* 2009; Garbutt and Reynolds 2012; Schnettler *et al.* 2014), a pattern which presumably existed prior to the radiation of the arthropods; in contrast, the expression of antiviral *Agos* in the honey bee *Apis mellifera* is induced by viral challenge (Galbraith *et al.* 2015). In vertebrates, an antiviral role for *Agos* has only been found in embryonic stem cells (Maillard *et al.* 2013) and suckling mice (Li *et al.* 2013). It is therefore an open question whether the ubiquitous antiviral *Ago* function evolved in the early arthropods, or in the vertebrate-invertebrate ancestor, and was subsequently lost in vertebrates (possibly due to interaction with other immune mechanisms).

In contrast, the Piwi subfamily is germline-specific in the vast majority of Metazoa (reviewed in Ross *et al.* 2014), suggesting that it specialized to the germline early in Metazoan evolution. This reflects the key roles of *Piwi* genes in the maintenance of spermatogenesis and fertility in a wide range of taxa (Cox *et al.* 1998; Kuramochi-Miyagawa *et al.* 2004; Houwing *et al.* 2007), which they achieve primarily through the suppression of transposition (Vagin *et al.* 2004; Kalmykova *et al.* 2005; Houwing *et al.* 2007; Grivna *et al.* 2006). However, different Piwi proteins localize to different parts of germline cells, where they carry out distinct functions. In *D. melanogaster*, Ago3 and Aub are exclusively cytoplas-

mic and are found in the nuage, an electron-dense perinuclear organelle (Al-Mukhtar and Webb 1971), where they amplify the piRNA signal in the "Ping-Pong" pathway (Li *et al.* 2009). Contrastingly, Piwi moves between the cytoplasm and the nucleus, being loaded with a piRNA in the cytoplasm and then moving across the nuclear membrane to target TEs for suppression (reviewed in Luteijn and Ketting 2013). Additionally, *Piwi* is expressed in the stem cells of sponges (*Ephyatia fluviatilis*; Funayama *et al.* 2010), jellyfish (*Podocoryne carnea*; Seipel *et al.* 2004), molluscs (*Aplysia californica*; Rajasethupathy *et al.* 2012), arthropods (*D. melanogaster*; Cox *et al.* 1998) and mammals (*Homo sapiens*; Sharma *et al.* 2001), indicating that it functions more broadly in the maintenance of cell pluripotency. This function is not universal, however, as *Piwi* has been lost in flukes and tapeworms, which instead appear to maintain stem cell pluripotency through the action of the flatworm-specific group 4 *Argonaute* genes (Skinner *et al.* 2014).

In contrast to these conserved patterns of expression, other *Argonaute* genes have recently evolved derived patterns of expression, which underlie rapid functional divergence. A major mechanism for this divergence appears to be gene duplication, which has occurred throughout *Argonaute* evolution (Cerutti and Casas-Mollano 2006; Mukherjee *et al.* 2013; Swarts *et al.* 2014b), and is prevalent within the Diptera (Chapter 2). For example, in *Glossina morsitans* a paralogue of *Ago3* (which is ancestrally germline-specific) has evolved expression in the salivary glands (Chapter 2). Similarly, somatic expression of Piwi subfamily genes has been reported in the rhesus macaque *Macaca mulatta* (Yan *et al.* 2011) and the mosquito *Aedes albopictus* (Morazzani *et al.* 2012; Vodovar *et al.* 2012; Schnettler *et al.* 2013a), the latter of which have evolved a novel antiviral function (Morazzani *et al.* 2012; Vodovar *et al.* 2012; Schnettler *et al.* 2013a). More rarely, Ago subfamily genes have specialized to a germline-specific role, as seen for a paralogue of *Ago2* in the tiger shrimp *Penaeus monodon*, which has specialized to suppress TEs (Leebonoi *et al.* 2015).

An expansion of *Argonaute* genes has recently been reported in *D. pseudoobscura* (Hain *et al.* 2010); however, the age of these duplications, their presence or absence in related species, and their functions are yet to be characterized. These paralogues therefore represent an ideal system in which to explore the functional evolution of *Argonaute* genes after duplication. Additionally, the comparison of expression patterns across multiple species provides us with a valuable opportunity to estimate the speed and frequency with which new functions evolve.

3.2. Aims

Duplication of *Ago2* has recently been reported in *D. pseudoobscura*; however, the existence of these paralogues has not been validated, and their age and relationship with *Ago2* paralogues in related species remains unexplored. Additionally, duplication of *Argonaute* has previously led to the evolution of derived functions and expression patterns; however, whether these recent *Ago2* duplicates are functional, and what these functions are, has not been tested. Finally, despite the range of functions carried out by *Argonaute* genes, the speed and frequency with which this functional diversity can evolve is relatively unexplored. To address these questions, our aims were as follows:

1. Validate the existence of the *D. pseudoobscura* paralogues by sequencing, and identify paralogues of *Ago2* in other species of the *obscura* group.

2. Characterise the phylogenetic relationships between *Ago2* paralogues in the *obscura* group, and date the duplication events from which they arose.

3. Quantify the expression levels of *D. pseudoobscura Ago2* paralogues in different tissues and under viral infection, in order to detect tissue-specific or inducible expression patterns.

4. Measure the tissue-specific and antiviral expression patterns of *Ago2* paralogues in other *obscura* group species, and compare these patterns with *D. pseudoobscura* in a phylogenetic framework.

3.3. Methods

3.3.1. Identification of *Ago2* paralogues in the *obscura* group

In order to quantify the extent and distribution of *Ago2* duplication in the *obscura* group, we used tBLASTx to identify *Ago2* paralogues in *D. subsilvestris*, *D. tristis*, *D. lowei*, *D. persimilis*, *D. pseudoobscura*, *D. subobscura* and *D. obscura* (see Chapter 2 for details of sequence data origin). For each species, we used *Ago2* sequences from the closest possible relative as queries, or *D. pseudoobscura* if no closer relative was available. If blast returned partial hits, we aligned all hits from the target species

to all *Argonaute* genes from the query species, and assigned hits to the correct query sequence based on a neighbour-joining tree. For each query sequence, we then manually curated partial blast hits into complete genes using Geneious v5.6.2 (http://www.geneious.com, Kearse *et al.* 2012). For *D. pseudoobscura*, five duplicates of *Ago2* have already been reported (Hain *et al.* 2010); however, because the *D. pseudoobscura* genome has since been reassembled using long read data (English *et al.* 2012), we repeated these BLAST searches. Additionally, we used PCR with degenerate *Ago2* primers to identify and amplify *Ago2* paralogues in *D. azteca* and *D. affinis*.

Duplicates with a high degree of sequence similarity may be collapsed into one locus when assembling a genome from short read data, leading to an underestimate of the size of a gene family. To validate the *Ago2* paralogues identified in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we used PCR to amplify each *Ago2* paralogue from each individual, with manually-designed paralogue-specific primer pairs and a touchdown amplification cycle used to avoid cross-amplification of multiple paralogues (see Appendix B for details of PCR primers and cycle conditions). Unincorporated primers were removed with ExonucleaseI (NEB, Ipswich, MA, USA) and 5' phosphates were removed with Antarctic Phosphatase (NEB, Ipswich, MA, USA), and then the PCR products were sequenced by Edinburgh Genomics (see Appendix B for details of sequencing primers and cycle conditions), using BigDye reagents on a capillary sequencer (Applied Biosystems, Foster City, CA, USA). Finally, we trimmed and assembled Sanger sequence reads using Geneious v.5.6.2 (http://www.geneious.com, Kearse *et al.* 2012).

3.3.2. Locating *D. pseudoobscura Ago2a1 & Ago2a3*

The *obscura* group has undergone a series of chromosomal translocations (Segarra and Aguadé 1992; Schaeffer *et al.* 2008), which may have changed the genomic location of *Ago2* paralogues. The lack of a genome for *D. subobscura* and *D. obscura* precludes the location of their *Ago2* paralogues; however, a high quality genome for *D. pseudoobscura* means that the locations of *Ago2b-Ago2e* were already known in this species. In contrast, *Ago2a1* and *Ago2a3* were located on a 26kb unplaced contig (labelled "Unknown_contig_265" in the *D. pseudoobscura* R3.03 genome), with *Ago2a1* ~3.5kb in from the left end and *Ago2a3* ~2kb away in from the right end. In order to identify the location of *Ago2a1* and *Ago2a3*, we used a combination of BLAST similarity searches and PCR, with reagents and cycling conditions as above.

3.3.3. Domain structures of *Ago2* paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*

The domain architecture of Argonaute proteins is key to their function, with the PAZ domain in particular remaining highly conserved throughout evolution (Swarts *et al.* 2014a); as such, the retention of these domains indicates that an Argonaute is still functional, while their loss may suggest pseudogenization. To infer the location of each domain in each paralogue identified in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we searched the Pfam database (Finn *et al.* 2009).

3.3.4. Phylogenetic analysis of *Ago2* genes in *Drosophila*

To characterise the evolutionary relationships between the *Ago2* paralogues identified in Section 3.3.1 and other *Drosophila* species, we aligned sequences using translational MAFFT (Katoh *et al.* 2002) with default parameters, and inferred a gene tree using the Bayesian approach implemented in BEAST v1.8.1 (Drummond *et al.* 2012) under a nucleotide model. We assumed a HKY substitution model with two unlinked codon-position classes (1st+2nd & 3rd), variation between sites modelled by a gamma distribution with four categories, and base frequencies estimated from the data. We used the default priors for all parameters, except tree shape (for which we specified a birth-death speciation model) and the *Drosophila-Sophophora* split. To estimate a timescale for the tree, we specified a normal distribution for the date of this node using values based on Obbard *et al.* 2012, with a mean value of 32mya, standard deviation of 7mya, and lower and upper bounds of 15mya and 50mya respectively. We ran the analysis for 50 million steps, recording samples from the posterior every 1,000 steps, and inferred a maximum clade credibility tree with TreeAnnotator v1.8.1 (Drummond *et al.* 2012).

3.3.5. Tissue-specific expression patterns of *Ago2* paralogues

To assess whether *Ago2* paralogues are expressed, and to explore the possibility that they have specialized to different tissues, we analysed the spatial expression pattern of each *Ago2* paralogue in *D. subobscura*, *D. obscura* and *D. pseudoobscura*. For each species, we extracted RNA from the head, testis/ovaries and carcass of 48-96hr virgin adults, with males and females extracted separately. Each sample consisted of 8-15 individuals in *D. subobscura*, 10 individuals in *D. obscura* and 15 individuals

in *D. pseudoobscura*. RNA was extracted from each tissue using TRIzol reagent (Ambion, Carlsbad, CA, USA) and a chloroform/isopropanol extraction, and treated twice with TURBO DNase (Ambion, Carlsbad, CA, USA), before being reverse-transcribed using M-MLV reverse transcriptase (Promega, Madison, WI, USA) primed with random hexamers. We then quantified the expression of *Ago2* paralogues in these samples with qPCR, using Fast Sybr Green (Applied Biosystems, Foster City, CA, USA) and custom-designed paralogue-specific qPCR primer pairs, all of which displayed single melt curve peaks and operated at 95-105% efficiency as quantified by serial dilution (see Appendix B.1 for a representative example). Due to their high level of sequence similarity (99.9% identity), no primer pair could distinguish between *Ago2a1* and *Ago2a3*, so these two genes are presented together as "*Ago2a*". All qPCR reactions for each sample were run in duplicate, and scaled to the internal reference gene Ribosomal Protein L32 (RpL32, see Appendix B for qPCR primers and cycling conditions). To capture any genetic variation and allow generalization across genotypes, we carried out five replicates per species, each in a different wild-type background: DPG1, DPG2, DPG4, DPG6 and DPG12 for *D.subobscura*; DPG1, DPG2, DPG3, DA45 and DA46 for *D. obscura*; and MV2, MV11, MV15, MV25 and MV32 for *D. pseudoobscura*. *D. obscura* DA45 and DA46 were collected from Moor Lane (Derbyshire, UK) by Ben Longdon in August 2008, and the origins of all other backgrounds are described in Section 4.3.2.

To provide an informal comparison with the expression pattern of *Ago2* before duplication (an "ancestral" expression pattern), we used the BPKM (bases per kilobase of gene model per million mapped bases) values for *Ago2* calculated from RNA-seq data from the body, head, ovary and testis of *D. melanogaster* by the FlyAtlas experiment (Brown *et al.* 2014). For comparability with the qPCR data detailed above, we scaled each BPKM value to the value for *RpL32* in each tissue. Due to the design of the FlyAtlas experiment, the body and head data are derived from pooled samples of equal numbers of males and females.

3.3.6. Expression of *Ago2* paralogues in *D. pseudoobscura* embryos

To investigate whether transcripts of any *Ago2* paralogues in *D. pseudoobscura* are being contributed to embryos by sperm, and to rule out early developmental expression, we collected embryos within a 30 minute window after being laid, extracted RNA and synthesised cDNA as detailed in section 3.3.5. We then used qPCR to measure the expression of each *Ago2* paralogue, with reagents and primers as detailed in section 3.3.5, in two separate genetic backgrounds (MV8 and MV10). To compare this with

the estimated ancestral expression pattern of *Ago2* before duplication, we used the BPKM (bases per kilobase of gene model per million mapped bases) values for *Ago2* calculated from RNA-seq data from embryos of *D. melanogaster* by the FlyAtlas experiment (Brown *et al.* 2014), scaled to the BPKM value for *RpL32* in embryos.

3.3.7. Expression of *Ago2* paralogues on viral challenge

To test whether any *Ago2* paralogues have evolved inducible expression under viral infection, we exposed 48-96hr old virgin males and females of *D. subobscura*, *D. obscura* and *D. pseudoobscura* to Drosophila C virus (DCV). We infected individuals by puncturing the cuticle at the top of the pleural suture, using a pin contaminated with DCV at a dose of ~4x10^7 tissue culture ID^{50} per ml (1 TCID^{50} = the dose needed to kill 50% of inoculated tissue culture cells). To estimate the ancestral antiviral expression pattern of *Ago2*, we exposed virgin males and females of *D. melanogaster* to DCV using the same method. All flies were incubated at 18C on a 12L:12D light cycle, with *D. melanogaster* kept on Lewis medium and *D. subobscura*, *D. obscura* and *D. pseudoobscura* kept on banana medium (see Appendix B for recipes). We sampled 4-7 individuals per species at 0, 8, 16, 24, 48 and 72 hours post infection. Three replicates were carried out per species, each in a different genetic background: FR32, FR35 and FR39 for *D. melanogaster* (isofemale lines established and collected from Montpellier, France by Penny Haddrill in August 2010); DPG1, DPG2 and DPG12 for *D. subobscura*; DPG1, DPG3 and DA45 for *D. obscura*; and MV2, MV8 and MV10 for *D. pseudoobscura*. At each time point we extracted RNA, synthesised cDNA and used qPCR to quantify the expression of each *Ago2* paralogue, with two technical replicates per sample, and with reagents, primers and cycling conditions as detailed in section 3.3.5.

3.3.8. Expression of other *Argonaute* gene family members in *D. pseudoobscura* tissues and embryos

The loss of a function in one gene can impose selection to retain duplicates of another gene, which may then evolve to compensate for this loss; alternatively, the evolution of a new function in one gene can relax selection to maintain this function in a gene already carrying out this role, leading to its loss. To assess the possible interaction between duplications of *Ago2* and expression patterns of the rest of the *Argonaute* gene family, we designed qPCR primers for *D. pseudoobscura Ago1*, *Ago3*, *Aub* and

Piwi. We then used qPCR to measure the expression of these genes in the six tissues from five different genetic backgrounds detailed in section 3.3.5, and the embryos from two genetic backgrounds detailed in section 3.3.6, with reagents and cycling conditions as detailed in section 3.3.5. To provide an informal comparison with the "ancestral" expression patterns of the other members of the *Argonaute* gene family (*Ago3*, *Aub* and *Piwi*) when *Ago2* is present as a single copy, we used the BPKM (bases per kilobase of gene model per million mapped bases) values for each gene calculated from RNA-seq data from the body, head, ovary and testis of *D. melanogaster* by Brown *et al.* 2014, scaled to the BPKM value for *RpL32* in that tissue.

3.4. Results

3.4.1. *Ago2* duplicates frequently in the *obscura* group, and moves frequently in *D. pseudoobscura*

To survey the extent of *Ago2* duplication in the *obscura* group, we identified *Ago2* paralogues in 9 *obscura* group species using blast. In total, we find 30 full-length *Ago2* paralogues, indicating frequent duplication. We find two paralogues in *D. subobscura* and three in *D. obscura*, all of which have intact PAZ and PIWI domains. We also find six full-length *Ago2* paralogues in *D. pseudoobscura*, five of which (*Ago2a3-Ago2e*) were reported by Hain *et al.* 2010, and one of which (*Ago2a1*) has not been previously identified. We also identify two truncated pseudogenes in *D. pseudoobscura*, one closely related to *Ago2a* (*Ago2aψ)* and another very similar to *Ago2b* (*Ago2bψ*). All full-length paralogues are 2-3kb long, each has the PAZ and Piwi domains characteristic of *Ago2*, and each shows no copy-number polymorphism in our sample. The length of the 5' end varies widely between paralogues; however, this may be driven by high repeat content, which has previously been reported for the 5' end of *D. melanogaster Ago2* (Meyer *et al.* 2006).

We used PCR to locate the previously-unplaced contig carrying *Ago2a1* and *Ago2a3*, placing it in reverse-orientation on chromosome XL-group1a, with the right end at position 3,463,701, and the left end predicted to be at position 3,489,689. This region contains a single annotation in the *D. pseudoobscura* R3.03 genome, a 2,485bp pseudogene (GA30567) that contains a 1125bp region with high (99%) sequence similarity to the 3' end of *Ago2a3*. Combined with the previously-established locations of

Ago2b-e, this reveals that *Ago2* paralogues are distributed across the *D. pseudoobscura* genome: *Ago2b*, *Ago2bψ* and *Ago2c* are on chromosome 2 (Muller element E); *Ago2e* is on chromosome 4 (Muller B); *Ago2a1*, *Ago2a3* and *Ago2aψ* are on the left arm of the X chromosome (Muller A); and *Ago2d* is on the right arm of the X chromosome (Muller A/D). Synteny comparison with *D. melanogaster Ago2* shows that the ancestral locus is *Ago2d*. Combined with our phylogenetic analysis (Figure 3.2), this allows us to infer the order and timing of movement of *D. pseudoobscura Ago2* paralogues around the genome (Figure 3.1). Firstly, the *Ago2a1-e* ancestor duplicated ~21mya to form *Ago2a1-d* and *Ago2e*, the latter of which moved onto chromosome 2L. Next, the 3L arm fused with the X chromosome, moving *Ago2a1-d* onto the X: this happened after the divergence of the *obscura* group into Palearctic (e.g. *D. subobscura*) and Nearctic (e.g. *D. pseudoobscura*) clades (Segarra and Aguadé 1992), which we estimate occurred 7-13mya. *Ago2a1-d* then duplicated ~4.5mya, forming *Ago2c-d* and *Ago2a1-b*, the latter of which moved onto chromosome 2. After this, *Ago2a1-b* duplicated ~3.5mya, producing *Ago2b* and *Ago2a1-3*, the latter of which moved onto the left arm of the X chromosome. This was followed by a duplication of *Ago2c-d* ~1.7mya, forming *Ago2d* and *Ago2c*, the latter of which moved onto chromosome 2. Finally, *Ago2a1-3* duplicated ~10kya, producing *Ago2a1* and *Ago2a3* in tandem.

3.4.2. The age of *Ago2* paralogues varies widely

To characterize the relationships between *Ago2* paralogues in the *obscura* group, and estimate the date of the duplication events that produced them, we carried out a phylogenetic analysis. This demonstrates that there are three main *Ago2* clades in the *obscura* group, which were produced from relatively old duplications. The *Ago2e* subclade is the oldest, and diverged from other *Ago2* paralogues ~21mya. The *Ago2a* and *Ago2f* subclades are more recent, and were produced by a gene duplication event ~14mya. There has been a more recent duplication of *Ago2a* on the *D. affinis-D. azteca* lineage (~5mya); however, low support for this node means that these paralogues could also nest within the recent expansion seen in *D. pseudoobscura* and *D. persimilis*, with one paralogue sister to the *Ago2a1-Ago2b* subclade and the other sister to the *Ago2c-Ago2d* subclade.

Our phylogenetic analysis also reveals the loss of some *Ago2* paralogues in some species. Specifically, *Ago2e* has been lost in *D. subobscura*, *Ago2f* has been lost in *D. pseudoobscura*, *D. persimilis* and *D. azteca*, and both have been lost in *D. lowei*. While some of these losses may be the result of incomplete genome assemblies or unexpressed genes in transcriptome surveys, we validated the losses in *D. pseudoobscura* and *D. subobscura* by extensive PCR. Combined with the existence of two pseudogenes in *D. pseudoobscura* (Section 3.4.1), this demonstrates rapid turnover of *Ago2* in the *obscura* group.

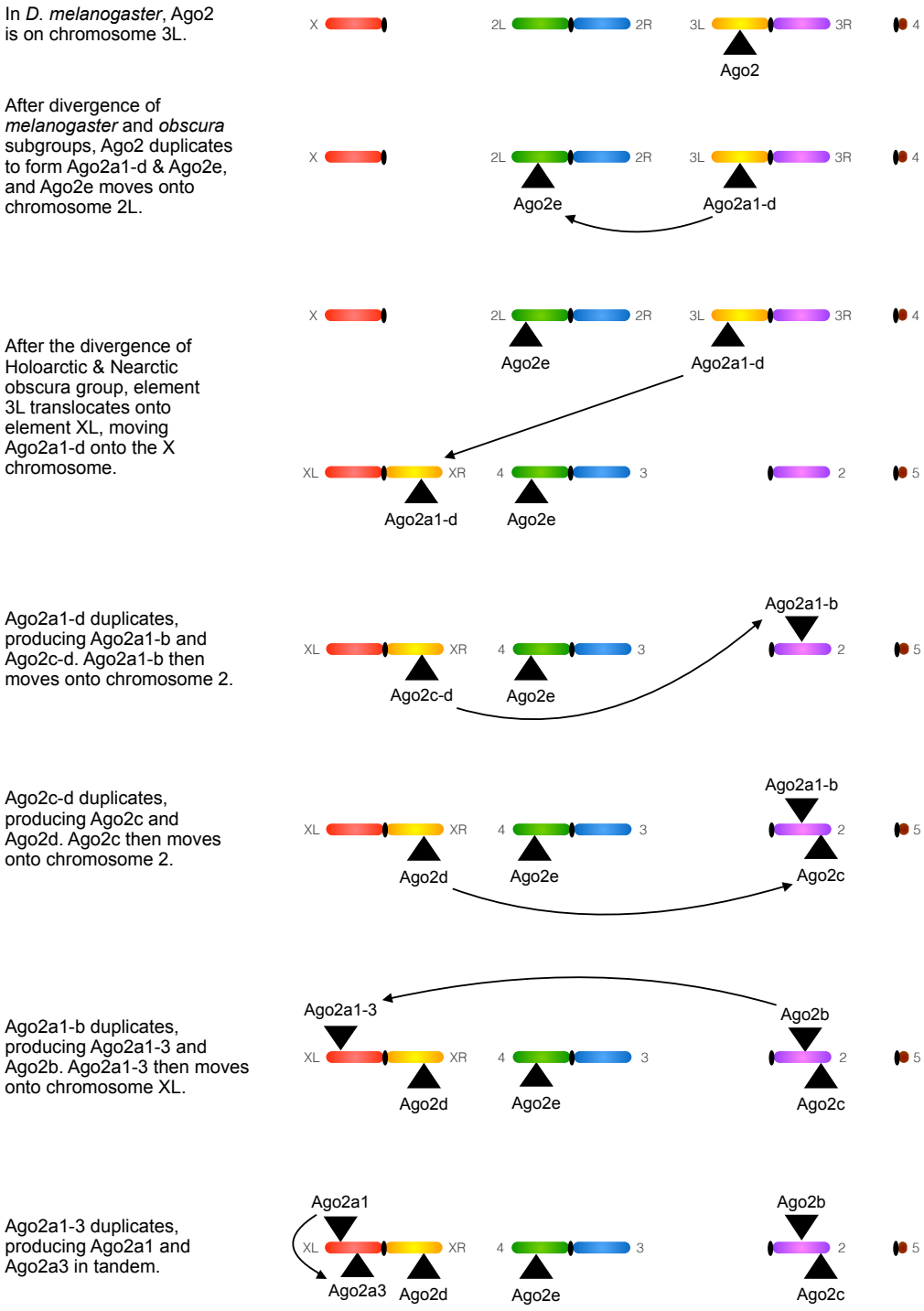


Figure 3.1.: The course of duplications and translocations of *Ago2* paralogues in *D. pseudoobscura*. A complex series of duplications and translocations has produced six *Ago2* paralogues in *D. pseudoobscura*, located on four different chromosome arms (chromosome arms adapted from Schaeffer *et al.* 2008, Fig. 1).

Figure 3.2.: A time-scaled Bayesian gene tree of *Ago2* genes in *Drosophila*. Duplication events are marked by yellow diamonds, and nodes with less than 100% posterior support are labelled. *Ago2* has duplicated frequently in the *obscura* group, and also in *D. willistoni* and *S. deflexa*. After duplication, *Ago2* paralogues have specialized to the testis three times independently, indicating an adaptive basis for testis-specificity.

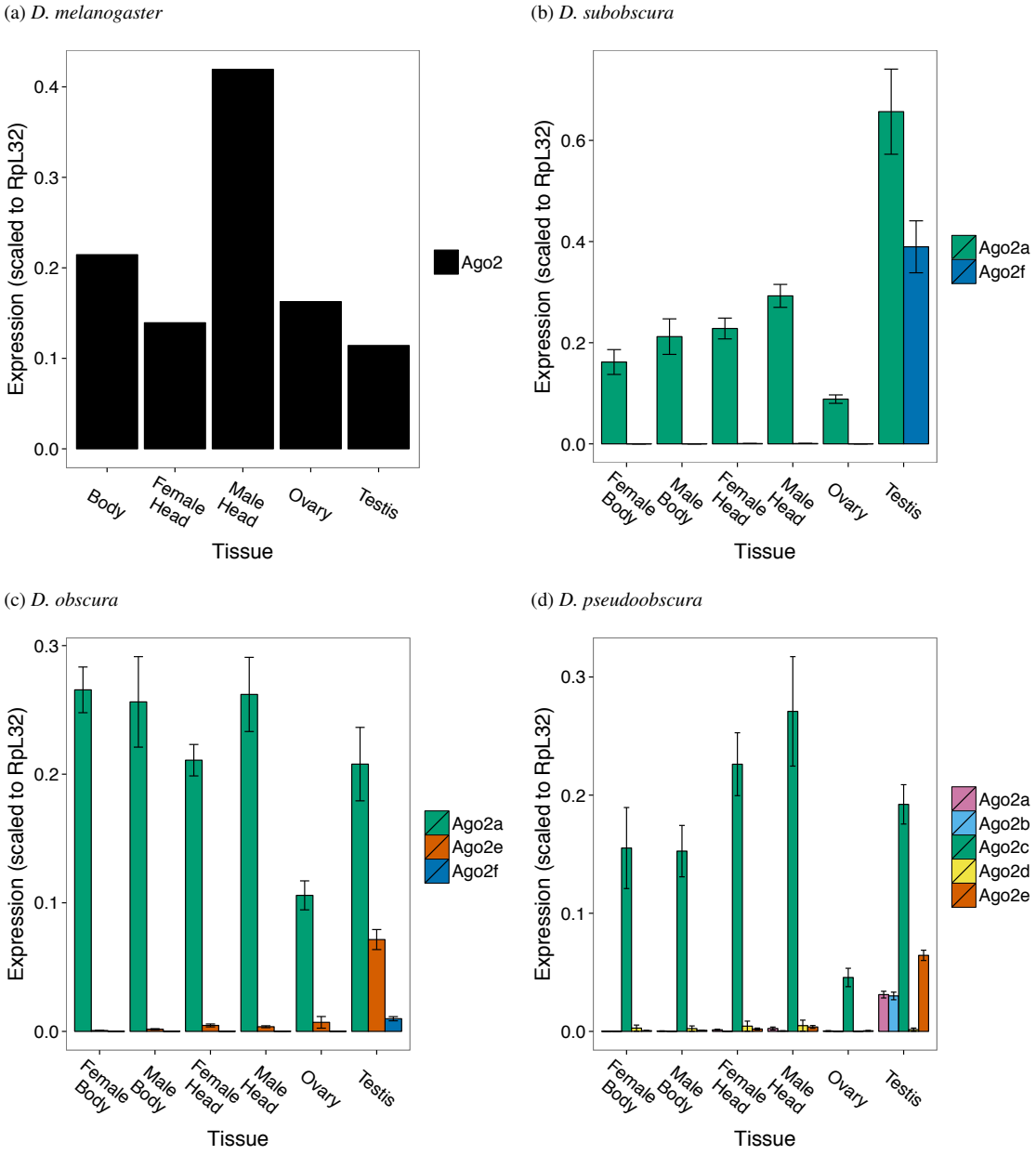


Figure 3.3.: Tissue-specific expression patterns of *Ago2* paralogues.
In each *obscura* group species, one paralogue has retained the ancestral ubiquitous expression pattern, while the others have specialized to the testis (with the exception of *D. pseudoobscura Ago2d*). Comparisons across tissues should be made with caution, given the possibility of variation in expression level of the reference gene. Error bars indicate 1 standard error estimated from 2 technical replicates of each of 5 biological replicates, each from a different genetic background. No error bars are available for the *D. melanogaster* expression levels, as these were estimated from a single RNA-seq experiment (Brown *et al.* 2014).

3.4.3. The majority of *Ago2* paralogues have specialized to the testis

Given that *Ago2* duplicates in the *obscura* group have been retained for many millions of generations, we tested for potential functional differences by quantifying the tissue-specific expression patterns of paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*. In each species, the *Ago2* paralogues exhibit striking differences in their tissue-specific patterns of expression (Figure 3.3). An approximation of the pre-duplication expression pattern of *Ago2* can be obtained from *D. melanogaster*, in which the single copy of *Ago2* is expressed in all adult tissues (Figure 3.3a) and in the embryo (Figure 3.4). In *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we find that one paralogue has a similar ubiquitous expression pattern in adult tissues (Figure 3.3b-d) and in the embryo (Figure 3.4). Interestingly, the ubiquitously expressed paralogue in *D. subobscura* and *D. obscura* is the ancestral gene (*Ago2a* in both cases), but in *D. pseudoobscura* another paralogue (*Ago2c*) has evolved the ubiquitous expression pattern, and the ancestral gene (*Ago2d*) is not expressed at a detectable level in any tissue. In contrast, every other paralogue in each *obscura* group species is expressed only in the testis (Figure 3.3b-d). However, none of these testis-specific paralogues is detectable in embryos (Figure 3.4), indicating that they are not loaded into embryos by sperm.

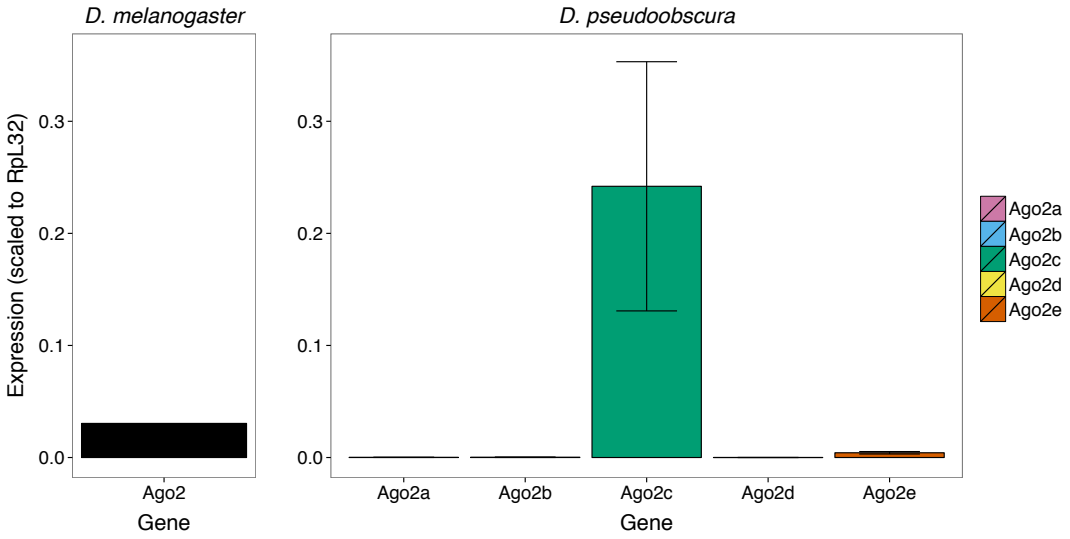


Figure 3.4.: Expression of *Ago2* in embryos of *D. melanogaster* and *D. pseudoobscura*.
In *D. pseudoobscura*, only *Ago2c* has retained the ancestral pattern of expression in embryos, with each testis-specific paralogue being unexpressed. For *D. pseudoobscura*, error bars indicate 1 standard error estimated from 2 technical replicates of each of 2 biological replicates, each from a different genetic background.

3.4.4. Expression of testis-specific *Ago2* paralogues is not induced under viral infection

As the majority of *Ago2* paralogues appear to have specialized to the testis, we tested whether ubiquitous expression of the testis-specific paralogues was triggered under viral challenge. We also tested whether viral challenge caused upregulation of ubiquitously expressed *Ago2* paralogues, or any expression of *D. pseudoobscura Ago2d*. The ancestral expression pattern of *Ago2* can again be estimated from *D. melanogaster*, in which *Ago2* is expressed in both females and males throughout the timecourse of viral infection (Figure 3.5a). A similar pattern is seen for the ubiquitously expressed paralogues of *D. subobscura* (*Ago2a*, Figure 3.5b), *D. obscura* (*Ago2a*, Figure 3.5c) and *D. pseudoobscura* (*Ago2c*, Figure 3.5d), which maintain a high level of expression throughout the timecourse. In contrast, the testis-specific paralogues in each species show no induction and are expressed at a very low rate at all time points, with the differences between them reflective of their different expression levels in the testis under resting conditions (Figure 3.3).

3.4.5. Other *Argonaute* gene family members in *D. pseudoobscura* tissues

To investigate the correlation between the duplication and specialization of *Ago2* and the expression of other members of the *Argonaute* gene family, we measured the expression levels of *Ago1*, *Ago3*, *Aub* and *Piwi* in *D. pseudoobscura* using qPCR. As in sections 3.4.3 and 3.4.4, we first estimated the pre-duplication expression patterns of these genes in *D. melanogaster* using previously published RNA-seq data (Brown *et al.* 2014). When comparing *Argonaute* genes between different tissues, we find that *Ago1* expression is broadly similar in *D. pseudoobscura* and *D. melanogaster*. In contrast, *Piwi* is expressed more highly in the ovary of *D. melanogaster* than *D. pseudoobscura*, and *Ago3*, *Aub* and *Piwi* are expressed more highly in the testis of *D. pseudoobscura* than *D. melanogaster*.

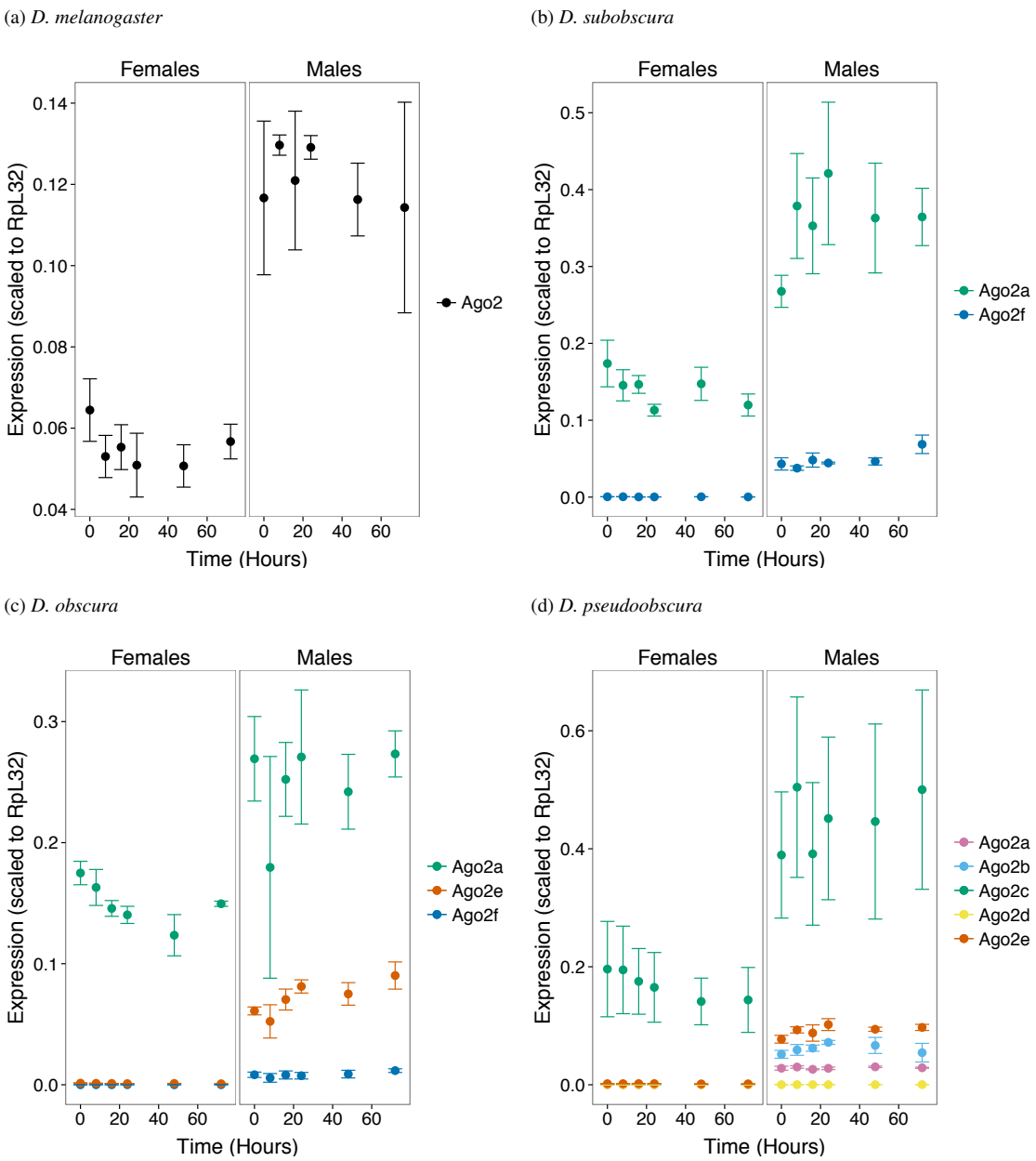


Figure 3.5.: Expression patterns of *Ago2* paralogues under infection with Drosophila C Virus. In each *obscura* group species, only one *Ago2* paralogue has retained the ancestral pattern (illustrated by *D. melanogaster*) of sustained expression under viral challenge. The testis-specific paralogues are not induced on viral infection, and appear to have lost their ancestral role in ubiquitous antiviral defence. Error bars indicate 1 standard error estimated from 2 technical replicates of each of 5 biological replicates, each from a different genetic background.

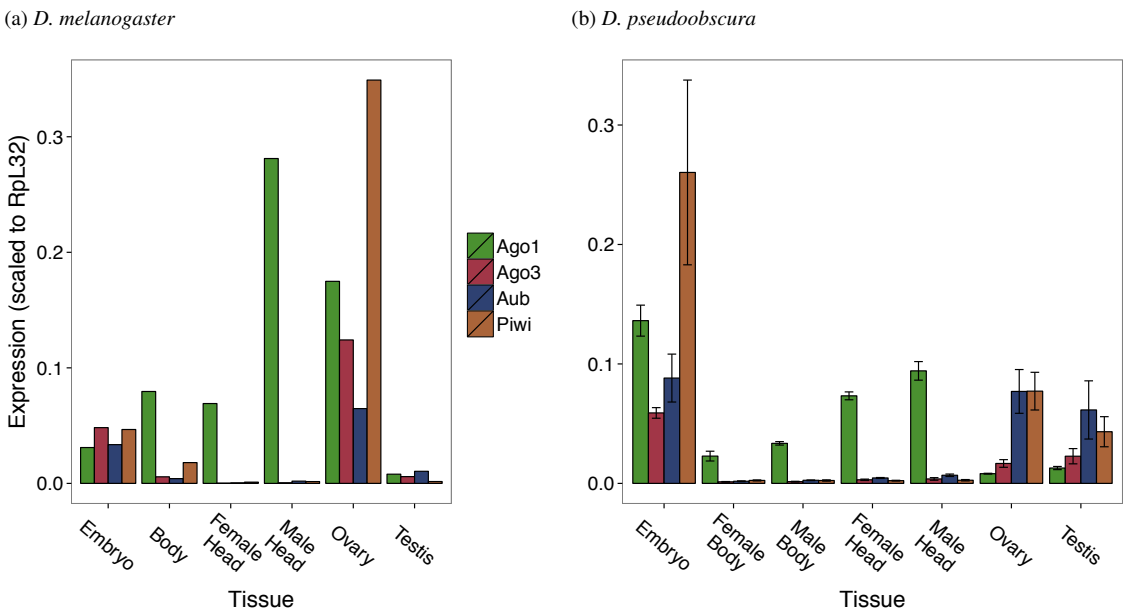


Figure 3.6.: Expression patterns of the other members of the *Argonaute* gene family. *Ago1* is similarly expressed in both species, but *Ago3*, *Aub* and *Piwi* appear to be upregulated in *D. pseudoobscura* testis. For *D. pseudoobscura*, error bars indicate 1 standard error estimated from 2 technical replicates of each of 5 biological replicates, each from a different genetic background.

3.4.6. Specialization to the testis has occurred several times independently and been retained for millions of generations

To estimate the frequency and timing of specialization to the testis, we inferred the position of specialization events on the gene tree of *Ago2* paralogues using the parsimony principle. We find that there have been at least three independent specializations of *Ago2* paralogues to the testis (Figure 3.2). The first occurred after the divergence of the *Ago2e* subclade (12-22mya), and the second after the divergence of *D. affinis* in the *Ago2f* subclade (10-14mya). The third occurred at the base of the *D. pseudoobscura* *Ago2a-Ago2b* subclade (3-4mya), after the divergence of *Ago2a-Ago2b* and *Ago2c-Ago2d*, but before the divergence of *Ago2a* and *Ago2b*.

3.5. Discussion

3.5.1. Functional implications of divergent expression patterns

We show that *Ago2* paralogues in the *Drosophila obscura* group have diverged in their patterns of expression, and are therefore likely to have diverged functionally. The ubiquitous expression of *Ago2* in *D. melanogaster* (Figure 3.3a) provides an estimate of the expression pattern of the ancestral *Ago2* in the *obscura* group before duplication. In each species investigated, one paralogue has retained this expression pattern: *D. subobscura Ago2a*, *D. obscura Ago2a* and *D. pseudoobscura Ago2c* (Figures 3.3b-d), which are the ancestral loci in *D. subobscura* and *D. obscura*, but a recent duplicate in *D. pseudoobscura*. This suggests that these paralogues have retained a role in antiviral defence, and also potentially in dosage compensation (Menon and Meller 2012) and somatic TE suppression (Czech *et al.* 2008; Chung *et al.* 2008), as seen for *D. melanogaster Ago2*.

In contrast, all other paralogues in all three species have lost this ubiquitous expression pattern and specialized to the testis (apart from the unexpressed *D. pseudoobscura Ago2d*). Constitutive ubiquitous expression is not a prerequisite for antiviral function, as many genes that are integral in antiviral defence are induced on viral infection (e.g. members of the Toll (Zambon *et al.* 2005) and Jak-STAT (Dostert *et al.* 2005) signalling pathways in *D. melanogaster*). However, the lack of upregulation that we observe in the testis-specific *Ago2* paralogues (Figures 3.5b-d) shows that they are not induced on viral infection, and are therefore unlikely to carry out a ubiquitous antiviral role. Although data on *Ago2* expression patterns in other taxa is limited, this expression pattern is highly divergent from that of *Ago2* in *Anopheles gambiae* (Keene *et al.* 2004), *Manduca sexta* (Garbutt and Reynolds 2012) and *Apis mellifera* (Galbraith *et al.* 2015), all of which display ubiquitous *Ago2* expression (although in *A. mellifera* this expression is induced on viral challenge; Galbraith *et al.* 2015).

3.5.2. Adaptive basis of testis-specific expression

While specialization to the testis appears to be the dominant fate for *Ago2* paralogues, it does not of itself suggest that such specialization is adaptive, as the out-of-the-testis hypothesis (outlined in section 3.1) suggests that testis-specificity is a dominant fate for recent duplicates of all kinds (Kaessmann *et*

al. 2009; Kaessmann 2010). However, two lines of evidence point to an adaptive basis for the testis-specificity observed for these *Ago2* paralogues. Firstly, the out-of-the-testis hypothesis predicts that paralogues will only retain this specificity for a short time, before their expression patterns broadens to other tissues and they evolve new functions (Kaessmann 2010), as is observed for the majority of paralogues in *D. melanogaster* (Assis and Bachtrog 2013). Contrary to this prediction, the testis-specific *Ago2e* and *Ago2f* have retained their testis-specificity for 12-22 and 10-14 million years respectively (Figure 3.2). Secondly, our phylogenetic analysis reveals that testis-specificity has evolved at least three times independently (Figure 3.2), suggesting a constant selection pressure acting on new *Ago2* paralogues to specialize to the testis.

3.5.3. Possible testis-specific functions

This selection pressure could be imposed by four possible functions: dosage compensation, testis-specific antiviral defence, TE suppression, or the suppression of meiotic drive. Firstly, these paralogues could have a function in dosage compensation, in which *D. melanogaster Ago2* plays an integral role by directing the male-specific lethal (MSL) complex to X-linked genes (Menon and Meller 2012). However, given the absence of the MSL from *Drosophila* testes (Conrad and Akhtar 2012), dosage compensation is unlikely to be the driver of testis-specificity. Secondly, these paralogues could have retained their antiviral function, but specialized to the testis to combat vertically-transmitted viruses such as sigma virus (L'Héritier and Teissier 1937). However, paternally-transmitted viruses are very rare, and all reported to date also pass through the female germline (Longdon and Jiggins 2012). A gene functioning in defence against vertically-transmitted viruses is therefore expected to be expressed in the female germline, from which these paralogues are absent, making an antiviral function less likely.

A third possible function of these testis-specific paralogues is the suppression of TEs and endogenous retroviruses (ERVs), which *D. melanogaster Ago2* carries out in the soma and germline in conjunction with endogenous siRNAs (endo-siRNAs) (Czech *et al.* 2008; Chung *et al.* 2008; Nayak *et al.* 2010). Not only is transposition expected to increase in the germline (Charlesworth and Langley 1989), some TEs transpose preferentially in testes rather than the germline generally. For example, the *Penelope* element is much more active in the testes of *D. virilis* compared to the ovaries (Rozhkov *et al.* 2010), and the *copia* element displays increased expression and higher insertion and excision rates in the testes of males from the 2b line of *D. melanogaster* (Pasyukova *et al.* 1997; Morozova *et al.* 2009). In addition,

several copies of *copia* have been reported in the genome of *D. pseudoobscura* (Biémont and Cizeron 1999), suggesting that testis-specific transposition can occur in this species as well. TEs are canonically suppressed by Piwi subfamily *Argonaute* genes (*Ago3*, *Aub* & *Piwi*), all of which are expressed in *D. pseudoobscura* at comparable or higher levels than *D. melanogaster* (Figure 3.6a), suggesting that any role carried out by these testis-specific *Ago2* paralogues would be happening in addition to the conventional piRNA pathway. It has previously been observed that endo-siRNAs (but not piRNAs) target TEs that are invading (Rozhkov *et al.* 2010) or have recently integrated (Tam *et al.* 2008); *Ago2* paralogues may therefore have been retained and specialized to the testis in *obscura* group species to prevent the integration of new TEs and ERVs.

Lastly, testis-specific *Ago2* paralogues could suppress meiotic drive, where a genetic element propogates itself to the next generation at a greater rate than expected under standard Mendelian segregation, often imposing a fitness cost on the host (Jaenike 2001). Several examples of RNAi-mediated suppression of meiotic drive have been described previously. In *D. simulans*, the *disorter on the X* (*dox*) driver (Tao *et al.* 2007b) is suppressed by *not much yang*, a locus composed of inverted repeats homologous to *dox*, which may be processed into sRNAs and guide Argonaute-mediated suppression of *dox* (Tao *et al.* 2007a). In *D. melanogaster*, the *Stellate* element is suppressed by piRNAs encoded by the Y-linked *Suppressor of Stellate* (*Su*(*St*)) locus (Palumbo *et al.* 1994), which the Argonaute genes Ago3 and Aub bind preferentially in the testis (Aravin *et al.* 2001; Nishida *et al.* 2007; Nagao *et al.* 2010). Additionally, mutation of *Aub* or *Piwi* in *D. melanogaster* results in enhanced propagation of the canonical driver element *Segregation Distorter* (Gell and Reenan 2013).

Meiotic drive systems have frequently been reported in the *obscura* group, for example in *D. obscura* (Gershenson 1928; Sturtevant and Dobzhansky 1936), *D. pseudoobscura* (Sturtevant and Dobzhansky 1936) and other obscura group species (*D. affinis*, *D. athabasca*, *D. azteca* (Sturtevant and Dobzhansky 1936) & *D. persimilis* (Wu and Beckenbach 1983)). Additionally, Y-linked resistance to a sex-ratio distorting drive element has recently been reported in *D. affinis*, although the mechanism of this resistance is not clear (Unckless *et al.* 2015). Moreover, it has been suggested that species with low levels of recombination, such as *D. pseudoobscura* (McGaugh *et al.* 2012), evolve drive systems more readily because of a higher degree of linkage between driver and responder loci (Jaenike 2001). This increased level of meiotic drive may therefore impose selection for the evolution of novel suppression mechanisms, leading to the repeated specialization of *Ago2* paralogues to the testis.

3.5.4. Conclusion

We find that *Ago2* paralogues in the *obscura* group have repeatedly specialized to a testis-specific role. This derived expression pattern appears to be driven by selection, as evidenced by the retention of testis-specificity over millions of generations, and the evolution of this trait on several independent occasions. This as-yet uncharacterised testis-specific function could involve antiviral defence, dosage compensation, or the suppression of TEs or meiotic drive. Given the pleiotropic nature of *Ago2* before duplication, different *Ago2* paralogues could also have specialized to the testis to carry out distinct functions.

4. Population genetics of *Argonaute2* paralogues in the *obscura* group

4.1. Introduction

Gene duplication is an important force in evolution, which can relax selective constraints and lead to functional divergence and phenotypic novelty (Ohno 1970). Most paralogues accrue mutations at the neutral rate and are eventually lost by pseudogenization (Lynch and Conery 2000; Hughes and Liberles 2007); however, a minority are retained by selection, either to conserve the same function (e.g. Gibbons *et al.* 2015) or diversify and adopt new functions (e.g. Chaudhari *et al.* 2014). Moreover, different selection pressures can act on paralogues produced by one duplication event, leading to contrasting patterns of evolution in the same gene family. Such contrasting patterns are seen in the *Argonaute* gene family, and are driven by the broad array of functions carried out by different *Argonaute* genes (Meister 2013).

These diverse functions all proceed through the RNAi mechanism, which is a conserved system of nucleic acid manipulation directed by short guide RNAs (Ding 2010). The role of Argonaute proteins as effectors in this pathway relies on the action of three conserved domains: the PAZ and MID domains facilitate binding of the guide RNA, and the PIWI domain contains an RNase H-like site which cleaves the target RNA (Schirle and Macrae 2012). The essential function of the PAZ and MID domains is highlighted by their retention in *Argonaute* genes across the tree of life (Swarts *et al.* 2014a), and suggests that these domains will be conserved and evolve comparatively slowly. In contrast, the RNase H-like site in the PIWI domain of some *Argonaute* genes has been lost, predominantly following *Argonaute* duplication. For example, humans have four *Ago* genes, only one of which (*Ago2*) has retained a catalytically-active PIWI domain that can cleave a target RNA, with the other three (*Ago1*, *Ago3* & *Ago4*) carrying out a role in miRNA-mediated gene regulation without cleaving their targets (reviewed

in Meister 2013). This indicates that the PIWI domain is not essential for all *Argonaute* functions, and suggests that this domain may evolve more rapidly than PAZ or MID. However, analyses of selection have found that adaptive substitutions are biased away from the functional domains of *D. melanogaster Ago2* (Obbard *et al.* 2006), while a study of multiple species of *Drosophila* found adaptive substitutions in both the PAZ and PIWI domains (Kolaczkowski *et al.* 2011). While these contrasting results may reflect species-specific differences, they may also be driven by the methodological differences between the two studies, making it difficult to draw strong conclusions regarding the evolutionary rate and selection pressure on the domains and structure of Ago2.

In contrast, clear differences in evolutionary rate and selection pressure are seen between the three main Argonaute functions. Firstly, Argonautes function in gene regulation through interaction with host-encoded microRNAs (miRNAs) (reviewed in Eulalio *et al.* 2008). This function is generally associated with a low rate of evolution for *Argonaute1* (*Ago1*), as seen in the mosquito *Aedes aegypti* (Bernhardt *et al.* 2012), *D. melanogaster* (Obbard *et al.* 2006; Obbard *et al.* 2009b), the *Drosophila* more broadly (Kolaczkowski *et al.* 2011), and the Diptera as a whole (Chapter 2). While there are isolated reports of rapid evolution of *Ago1*, such as at the base of the *Drosophila* (Kolaczkowski *et al.* 2011) and after duplication in aphids (Jaubert-Possamai *et al.* 2010), in most organisms *Ago1* evolves very slowly under strong selective constraint.

Argonautes also function in the suppression of transposable elements (TEs), either binding and cleaving their transcripts or guiding DNA methylation and histone modification to reduce their expression (Luteijn and Ketting 2013). This function is associated with rapid evolution and strong selection of anti-TE *Argonaute* genes in teleost fish (Yi *et al.* 2014), as well as *D. melanogaster* (Obbard *et al.* 2009b), the *D. melanogaster* subgroup (Simkin *et al.* 2013) and the *Drosophila* as a whole (Kolaczkowski *et al.* 2011). There are at least two explanations for this rapid evolution: an increase in TE number, or interaction with suppressors of the RNAi mechanism. Firstly, an increase in the diversity of TEs may impose directional selection on anti-TE *Argonaute* genes, to diversify to a wider range of transposition mechanisms. This idea is supported by the association between an increase in the number of TE families in teleosts, and the rapid evolution of the teleost anti-TE RNAi pathway (Yi *et al.* 2014). However, it is contradicted by evidence from *D. melanogaster*, in which there is a negative correlation between TE content and the evolutionary rate of *Aub* and *Ago3*, as well as a positive relationship between TE content and codon usage bias of these genes, indicating that TEs may also constrain evolutionary rate by imposing purifying selection (Castillo *et al.* 2011). Secondly, the high evolutionary rate of anti-TE *Argonaute* genes could be imposed by suppressors of the RNAi mechanism. While there is only a single

report of a TE-encoded suppressor to date (Nosaka *et al.* 2012), if more widespread their interaction with Argonaute proteins could impose strong selection through a host-parasite arms race (Dawkins and Krebs 1979).

Lastly, *Argonaute* genes function in antiviral defence in plants (reviewed in Parent *et al.* 2012), fungi (reviewed in Chang *et al.* 2012), invertebrates (Li *et al.* 2002; Rij *et al.* 2006) and mammals (Li *et al.* 2013; Maillard *et al.* 2013). This function has been associated with the exceptionally rapid evolution of *Ago2* in *Drosophila*. In *D. melanogaster*, *Ago2* is in the top 3% of the fastest-evolving genes (Obbard *et al.* 2006), displays signals of strong positive selection (Obbard *et al.* 2006), and has undergone recent selective sweeps (Obbard *et al.* 2011), which are also evident in *D. simulans* and *D. yakuba* (Obbard *et al.* 2011). Positive selection on *Ago2* is also evident across the *Drosophila* more broadly, with adaptive substitutions in more than 60% of lineages and strong evidence for selective sweeps (Kolaczkowski *et al.* 2011). It has been proposed that this rapid evolution is driven by a host-parasite arms race with viruses (Obbard *et al.* 2009b), which encode viral suppressors of RNAi (VSRs) that inhibit the antiviral RNAi response (reviewed in Bronkhorst and Rij 2014). Strong selection on VSRs is consistent with their host-specificity (Mierlo *et al.* 2014) and numerous independent origins, which have resulted in a range of suppression mechanisms (Li and Ding 2006). Multiple viruses encode VSRs that inhibit RNAi by binding or cleaving Ago2, such as cricket paralysis virus (Nayak *et al.* 2010) and Nora virus (Mierlo *et al.* 2012). While the exact molecular details of this suppression remain unclear, VSRs could impose strong selection pressure on *Ago2*, driving frequent fixation of variants which prevent inhibition by the VSR.

In the *obscura* group, there have been numerous duplications of *Ago2* (Chapter 2). We have previously shown that the resulting paralogues have specialized to the testis on multiple independent occasions, providing a strong indication of functional divergence driven by positive selection (Figure 4.1, Chapter 3). As detailed above, *Argonaute* genes with different functions often evolve at different rates and under contrasting selection regimes, which can be characterised using sequence data. The rate of evolution can be estimated by calculating the proportion of nonsynonymous (dN) and synonymous (dS) differences between a focal gene and the same gene from an outgroup, giving the dN/dS ratio (ω). Under neutrality, ω is expected to be approximately 1; alternatively, ω<1 indicates slow evolution and purifying selection, whereas ω>1 suggests rapid evolution and positive selection. To test explicitly for selection, differences between species can be combined with data on polymorphisms within a species (e.g. the McDonald-Kreitman test; McDonald and Kreitman 1991), or the proportion of polymorphisms within a species can be compared between multiple loci (e.g. the Hudson-Kreitman-Aguadé test; Hudson *et al.* 1987). We

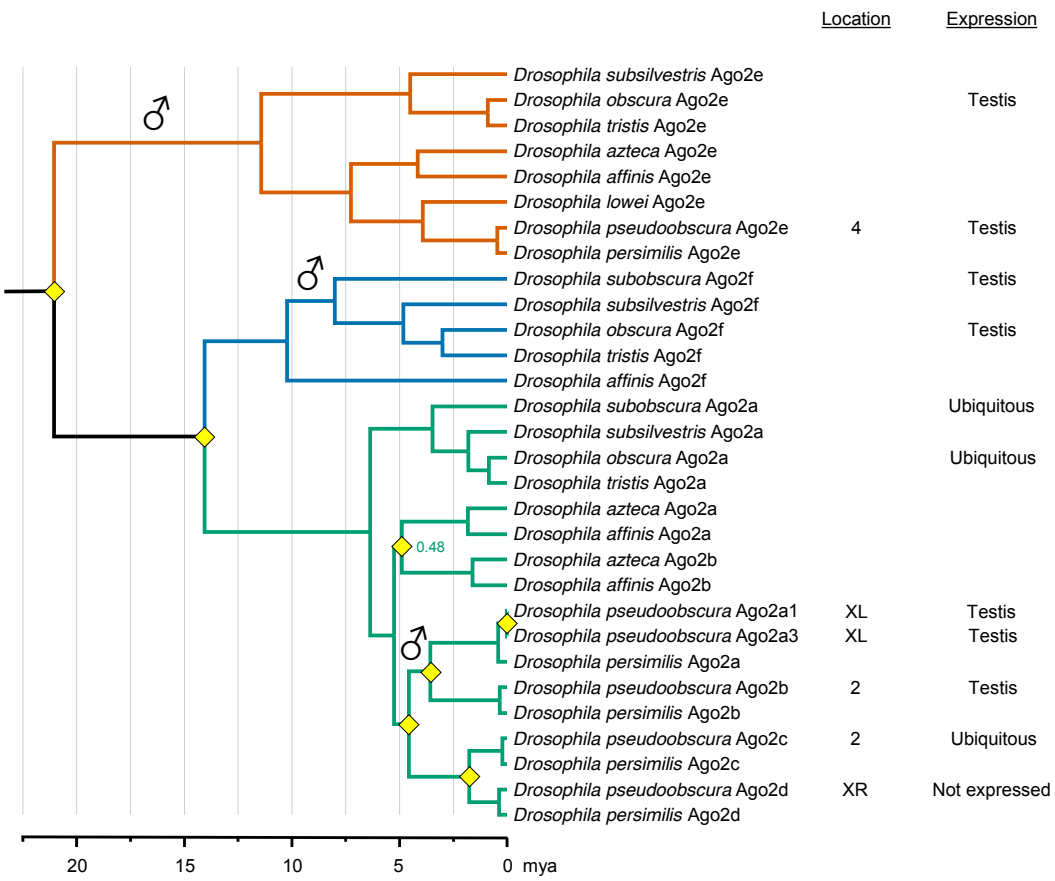


Figure 4.1.: A gene tree of the *Ago2* paralogues in the *obscura* group.
There have been numerous duplications of *Ago2* in the *obscura* group, producing paralogues that have moved across the genome in *D. pseudoobscura*, and specialized to the testis on multiple independent occasions in *D. subobscura*, *D. obscura* and *D. pseudoobscura*. Differences between clades in the dates of some speciation events are likely the result of evolutionary rate differences, and our use of a single node to date the tree (see Chapter 3).

sought to use these techniques to characterize the patterns of evolution and selection acting on *Ago2* paralogues in the *obscura* group, in order to explore the population genetic causes and consequences of their functional divergence.

4.2. Aims

Gene duplication can drive functional diversification, which is often reflected in contrasting rates of evolution and levels of selection, as has been shown previously for different *Argonaute* genes with contrasting functions. Additionally, we have previously identified hotspots of rapid evolution on the structure of Dipteran *Ago2* (Chapter 2), and discovered that *Ago2* paralogues in the *obscura* group

have specialized to a derived, testis-specific role (Chapter 3). To characterise the population genetic consequences of duplication, and test for the possible role of positive selection in driving this novel function, our aims were as follows:

1. Quantify the level of positive selection acting on each *Ago2* paralogue.

2. Test for a difference in evolutionary rate between ubiquitously-expressed and testis-specific *Ago2* paralogues.

3. Quantify the level of selective constraint on different domains and structural components of each *Ago2* paralogue.

4.3. Methods

4.3.1. Testing for evolutionary rate differences between ubiquitously expressed and testis-specific *Ago2* paralogues

In Chapter 3 we identified several independent origins of testis-specificity after duplication of *Ago2* in the *obscura* group. To compare the evolutionary rates of ubiquitously expressed and testis-specific *Ago2* paralogues, we used codeml (PAML, Yang 1997) to fit a number of variants of the M0 model to the set of 120 Dipteran *Ago2* sequences detailed in section 2.2.1. PAML allows the evolutionary rate (dN/dS or ω) to be estimated separately for different clades, and also provides a maximum likelihood estimate for these separate rates given the data, enabling the testing of hypotheses regarding the effect of a particular event (e.g. specialization to the testis) on the rate of evolution. We fitted two models to test for evolutionary rate change after the evolution of testis-specificity. Firstly, we fitted a model specifying one ω for the *Ago2* paralogues that were verified as testis-specific by qPCR (as detailed in Section 3.4.3), and another ω for the rest of the tree: this is a more conservative model, but potentially has less power to detect evolutionary rate change associated with testis-specificity. Secondly, we fitted a model specifying one ω for the *Ago2f* and *Ago2e* subclades and the *D. pseudoobscura Ago2a-Ago2b* subclade, and another ω for the rest of the tree: this model should have more power to detect rate change, but assumes that the paralogues that we found to be testis-specific (Chapter 3) are accurate

representatives of the other members of the subclades. We also fitted two models to account for rate variation between the *obscura* group *Ago2* subclades. The first model specified a separate ω for the *obscura Ago2a* subclade, the *Ago2e* subclade, the *Ago2f* subclade and the rest of the tree. The second model was the same as the previous model, but with an extra ω specified for the *D. pseudoobscura-D. persimilis Ago2a-Ago2b* subclade (which is testis-specific, in contrast with the rest of the *obscura* group *Ago2a* subclade). We used Akaike weights to assess which model provided the best fit to the data given the number of parameters.

4.3.2. Testing for contrasting patterns of evolution between *Ago2* paralogues

Haplotype sequencing

To characterize the selection pressures acting on *Ago2* paralogues in the *obscura* group, we sequenced the *Ago2* paralogues from six males and six females of each species, each from a different isofemale line. We collected all *D. subobscura* lines and *D. obscura* lines from Blackford Hill (Edinburgh, UK) in September 2010, whereas all *Drosophila pseudoobscura* lines were collected from Mesa Verde National Park (Mesa Verde, CO, USA) by Steve Schaeffer in July 2005 (see Appendix C for further details).

We extracted genomic DNA from each individual using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. We then used PCR to amplify each *Ago2* paralogue from each individual, with manually-designed paralogue-specific primer pairs and a touchdown amplification cycle used to avoid cross-amplification of multiple paralogues (see Appendix B for details of PCR primers and cycle conditions). Unincorporated primers were removed with ExonucleaseI (NEB, Ipswich, MA, USA) and 5' phosphates were removed with Antarctic Phosphatase (NEB, Ipswich, MA, USA), and then the PCR products were sequenced by Edinburgh Genomics (see Appendix B for details of sequencing primers and cycle conditions), using BigDye reagents on a capillary sequencer (Applied Biosystems, Foster City, CA, USA). We trimmed and assembled Sanger sequence reads using Geneious v.5.6.2 (http://www.geneious.com, Kearse *et al.* 2012), and identified polymorphic sites by eye. After sequencing *Ago2a* (annotated as a single gene in the *D. pseudoobscura* genome), we discovered 2 very recent *Ago2a* paralogues (*Ago2a1* & *Ago2a3*, described in Chapter 3), both of which had been cross-amplified. For each *D. pseudoobscura* individual we resequenced *Ago2a3* using primers

binding to its neighbouring locus *GA22965*, and used this sequence to resolve polymorphic sites in the *Ago2a1/Ago2a3* composite sequence, thereby gaining both sequences for each individual.

The presence of polymorphisms in males revealed that all paralogues in *D. subobscura* and *D. obscura* are autosomal (although their exact locations could not be found due to an absence of genomic data). In contrast, genomic and PCR investigations in *D. pseudoobscura* revealed that *Ago2a1*, *Ago2a3* and *Ago2d* are X-linked, whereas *Ago2b*, *Ago2c* and *Ago2e* are located on autosomes (Figure 4.1). For each Ago2 paralogue, we inferred haplotypes from these sequence data using PHASE (Stephens *et al.* 2001), apart from the X-linked paralogues in *D. pseudoobscura* males, for which phase was obtained directly from the sequence data.

Quantifying patterns of evolution in Ago2 paralogues

To quantify differences between paralogues in their population genetic characteristics, we aligned haplotypes using translational MAFFT (Katoh *et al.* 2002), and used DnaSP v.5.10.01 (Librado and Rozas 2009) to calculate the following summary statistics for each *Ago2* paralogue: π (with Jukes-Cantor correction as described in Lynch and Crease 1990) at nonsynonymous (π_a) and synonymous (π_s) sites, Tajima's D (Tajima 1989), the Codon Bias Index (CBI) (Morton 1993) and the effective number of codons (ENC, where 20 indicates the maximum codon usage bias of one codon used for each amino acid, and 61 indicates the minimum bias of each codon used equally) (Wright 1990). To compare the ENC for each gene with the genome as a whole, we used codonW v1.4.2 (Peden 1995) to calculate the ENC for the longest ORF from each gene or transcript in the genomes of *D. subobscura*, *D. obscura* and *D. pseudoobscura* (ORF sets are detailed in Section 4.3.3). In each species, we then compared the ENC values of each *Ago2* paralogue with this genome-wide ENC distribution. To analyse the high codon usage bias of the *Ago2e* subclade, we used the seqinR package (Charif and Lobry 2007) in R 3.2.2 to calculate the GC content of the introns and exons of *Ago2e* in *D. pseudoobscura* and *D. obscura* and the exons of *Ago2e* in *D. subsilvestris*, *D. tristis*, *D. azteca*, *D. affinis*, *D. lowei* and *D. persimilis*.

4.3.3. Quantifying the level of positive selection on each *Ago2* paralogue

McDonald-Kreitman tests

To test for the presence and quantify the strength of positive selection, we performed McDonald-Kreitman (MK) tests (McDonald and Kreitman 1991) on each paralogue in each species using DnaSP v.5.10.01 (Librado and Rozas 2009). The MK test is based on the comparison of the numbers of differences between species at nonsynonymous (Dn) and synonymous (Ds) sites, and polymorphisms within a species at nonsynonymous (Pn) and synonymous (Ps) sites. If all mutations are either neutral or strongly deleterious, the Dn/Ds ratio should be approximately equal to the Pn/Ps ratio; however, if there is positive selection, an excess of nonsynonymous differences is expected (McDonald and Kreitman 1991). Additionally, these numbers can be used to estimate the proportion of nonsynonymous substitutions that are adaptive (α), which can be scaled to the number of synonymous and nonsynonymous sites ($\omega(\alpha)$) (Eyre-Walker 2006). In each test, we assessed significance using a Fisher's exact test. Singleton sites were excluded from the analysis to alleviate the influence of segregating slightly deleterious variants, which inflate the number of nonsynonymous polymorphisms, resulting in downwardly biased estimates when testing for selection (Charlesworth and Eyre-Walker 2008).

To calculate Dn and Ds, a suitably divergent outgroup is needed: if the outgroup is too similar, Dn will be low and the test will be underpowered; whereas if the outgroup is too divergent and synonymous sites have become saturated, Ds will be artificially lowered, inflating the Dn/Ds ratio and introducing false positives. For each paralogue, we therefore chose an outgroup with divergence at synonymous sites (Ks) of 0.1-0.2 where possible. Additionally, the prevalence of duplications and losses of *Ago2* paralogues in the *obscura* group (Figure 4.1, Chapter 3) meant that for some tests, a suitably divergent extant outgroup sequence did not exist. In these cases, we reconstructed hypothetical ancestral sequences using the M0 model in PAML (Yang 1997). Specifically, for each *D. obscura* test we used the corresponding *D. subsilvestris* paralogue as an outgroup, and for each *D. subobscura* test we used the corresponding *D. subobscura-D. obscura* ancestor. For each paralogue within the *D. pseudoobscura Ago2a1-Ago2d* clade we used the ancestor of this clade, and for *D. pseudoobscura Ago2e* we used the *D. pseudoobscura-D. affinis Ago2e* ancestor. To assess the effect of these outgroup choices on our results, we repeated each test with another outgroup. For *D. subobscura* we used *D. subsilvestris*, for *D. obscura* we used *D. tristis*, and for *D. pseudoobscura* we used either *D. lowei* (*Ago2e*), or a different paralogue from *D. pseudobscura* (*Ago2a1-Ago2d*). We find no effect of outgroup choice on the significance of any tests,

and only marginal differences in estimates of α and $\omega(\alpha)$ (Appendix C).

Testing for reduced diversity

Recent selection can also be inferred from a reduction in diversity at a particular locus compared with the genome as a whole. To compare the diversity of each *D. pseudoobscura Ago2* paralogue with the genome as a whole, we inferred the distribution of genome-wide synonymous site diversity. To infer this distribution, we used genomic data for 12 lines generated by McGaugh *et al.* 2012. We mapped short reads to the longest ORF for each gene in the R3.2 gene set using Bowtie2 v2.1.0 (Langmead *et al.* 2009), and estimated synonymous site diversity (θ_W based on fourfold synonymous sites) at each ORF using PoPoolation (Kofler *et al.* 2011). We then plotted the distribution of synonymous site diversity, limited to genes in the size range of 0.75kb - 3kb long for comparability with the *Ago2* paralogues (Chapter 3), and compared the fourfold synonymous site diversity levels of each *D. pseudoobscura Ago2* paralogue with this distribution. Due to the locations of the *D. pseudoobscura* paralogues on autosomes (*Ago2b*, *Ago2c* & *Ago2e*) and the X chromosome (*Ago2a1*, *Ago2a3* & *Ago2d*), and the different population genetic expectations for autosomal and X-linked genes (Vicoso and Charlesworth 2006), we examined separate distributions for autosomal and X-linked genes. To provide an additional test for reduced diversity at *D. pseudoobscura Ago2* paralogues, we performed maximum-likelihood Hudson-Kreitman-Aguadé tests (Wright and Charlesworth 2004), using divergence from *D. affinis* and intraspecies polymorphism data for 84 *D. pseudoobscura* loci generated by Haddrill *et al.* 2010. We performed 63 tests to encompass all one, two, three, four, five and six-way combinations of the paralogues, and calculated Akaike weights from the resulting likelihood estimates to provide an estimate of the level of support for each combination.

To infer a genome-wide distribution of synonymous site diversity for *D. obscura* and *D. subobscura*, for which genomic data are unavailable, we used previously-generated pooled transcriptome data. The *D. subobscura* pool consisted of 338 males (described in Mierlo *et al.* 2014: 38 flies collected July 2011 Sussex 51.100N, 0.164E; 60 flies July 2011 Edinburgh 55.928N, 3.170W; 180 flies August 2011 Perthshire 56.316N, 3.790W; 60 flies October 2011 Edinburgh 55.928N, 3.170W), and the *D. obscura* pool consisted of 222 males (115 flies collected July 2011 Sussex 51.100N, 0.164E; 55 flies August 2011 Perthshire 56.316N, 3.790W; 52 October 2011 Edinburgh 55.928N, 3.170W). To generate a *de novo* transcriptome for each species, we assembled short reads with Trinity r20140717 (Grabherr *et al.* 2011). For each species, we then mapped short reads to the longest ORF for each transcript, estimated

synonymous site diversity at each locus, and plotted the distribution of diversity for genes 0.75kb - 3kb long (as described above for *D. pseudoobscura*). The presence of heterozygous sites in males (identified by Sanger sequencing as detailed in Chapter 3) confirmed that all *Ago2* paralogues in *D. subobscura* and *D. obscura* are autosomal: we therefore compared the synonymous site diversity for these paralogues with the autosomal distribution, and do not show the distributions for X-linked genes.

Selective sweeps

Diversity at a focal locus can be reduced by selection on that locus, but also by selection on other linked loci, depending on their proximity and the rate of recombination (McVean and Charlesworth 2000). Where population genomic data permitted (*D. pseudoobscura*), we therefore investigated whether the diversity of each paralogue was a result of their genomic location, by comparing diversity at each paralogue to diversity in their neighbouring regions. We obtained sequence data for the 50kb either side of each of these paralogues from the 11 whole genomes detailed in McGaugh *et al.* 2012. Due to the very high similarity of these *Ago2* paralogues, they cannot be accurately assembled from short read data. For each genome, we therefore replaced the poorly-assembled region corresponding to the paralogue with one of our own Sanger-sequenced haplotypes, making a set of 11 ca. 102kb sequences for each paralogue. We aligned these sequences using PRANK (Löytynoja and Goldman 2005) with default settings, and calculated Watterson's θ at all sites in a sliding window across each alignment, with a window size of 5kb and a step of 1kb. For *Ago2a1* and *Ago2a3*, which are located in tandem, we analysed the same genomic region. Diversity may also be reduced at these *Ago2* paralogues compared to their neighbouring regions by demographic differences between our sample and the populations sequenced by McGaugh *et al.* 2012, leading to false signatures of selective sweeps. To test the robustness of our results to demographic differences, we repeated these analyses of diversity on a dataset in which our Sanger sequenced haplotypes were removed, leaving only polymorphism data from the populations sequenced by McGaugh *et al.* 2012.

To test explicitly for selective sweeps at each region, we used Sweepfinder (Nielsen *et al.* 2005) to calculate the likelihood of a sweep on each *Ago2* paralogue. Selective sweeps occur when a beneficial variant is driven rapidly to fixation, eliminating standing genetic variation at linked sites, thus changing the allele frequency spectrum in the region around the sweep (Smith and Haigh 1974). Sweepfinder locates these regions by calculating the marginal likelihood of the allele frequency spectrum for each site under models with and without selection, and multiplying the marginal likelihoods for each site in

a given window to gain a maximum composite likelihood under each model for that region (Nielsen *et al.* 2005). It then divides the maximum composite likelihoods with and without a sweep, to obtain a composite likelihood ratio for a sweep in that window (Nielsen *et al.* 2005). Sweepfinder can therefore produce a maximum composite likelihood surface for a chromosomal region, which gives a visual representation of the likelihood of a sweep at any location. Additionally, Sweepfinder ignores haplotype information, therefore the potential disruption of such haplotypes by our combining of Sanger-derived and whole-genome data will not affect the inference of a sweep. In these analyses we specified a gridsize of 20,000, a folded frequency spectrum for all sites, and included invariant sites. Due to their location in tandem, it is impossible to unambiguously attribute a sweep to either *Ago2a1* or *Ago2a3*; for these paralogues, we therefore analysed the composite *Ago2a1/Ago2a3* sequence.

To infer the significance of any observed peaks in the composite likelihood ratio, we used ms (Hudson 2002) to generate 1000 samples of 11 sequences under a coalescent model. We generated separate samples for each region, specifying the same number of polymorphic sites as observed in each region, the sequence length equal to the alignment length, and the effective population size at 10^6 (based on a previous estimate for *D. melanogaster* by Li and Stephan 2006). We specified the recombination rate at 5cM/Mb, a conservative value based on previous estimates for *D. pseudoobscura* (McGaugh *et al.* 2012), which will lead to larger segregating linkage groups and therefore a more stringent significance threshold.

To test for partial or soft selective sweeps at *Ago2e*, we calculated the $n\mathrm{S}_L$ statistic (Ferrer-Admetlla *et al.* 2014) for each SNP in our *Ago2e* haplotype data. The $n\mathrm{S}_L$ statistic detects an increase in haplotype homozygosity caused by a soft or partial sweep, by quantifying the number of segregating sites in the same haplotype as a focal SNP (Ferrer-Admetlla *et al.* 2014). To test for significance of $n\mathrm{S}_L$ at each SNP, we simulated 1,000 sets of 24 haplotypes with the same length and number of segregating sites as the *Ago2e* haplotypes, calculated $n\mathrm{S}_L$ for each simulated haplotype set, and plotted the distribution of these values. To estimate a p-value for each SNP, we then calculated the percentile of the distribution into which the $n\mathrm{S}_L$ for each SNP in the observed data fell.

4.3.4. Investigating variation in selective constraint across the protein structure of Ago2 paralogues

To test for differences in selective constraint between the PAZ domain, PIWI domain and the rest of the gene, we first identified the location of these regions in each *Ago2* paralogue by searching the Pfam database (Finn *et al.* 2009). We then plotted polymorphic sites along the length of each *Ago2* paralogue, and calculated the proportion of sites in each region that were polymorphic (in order to correct for differences in length between the different regions). To investigate variation in selective constraint across the protein structure of Ago2 paralogues, we mapped polymorphic sites identified in each paralogue onto the structure of *D. melanogaster* Ago2. For each paralogue, we identified residues with one or two polymorphic sites, and used translational MAFFT (Katoh *et al.* 2002) to identify the corresponding residues in the *D. melanogaster* Ago2 sequence. We then mapped these polymorphic residues onto the *D. melanogaster* Ago2 3D protein structure described in Section 2.2.5, and visualised polymorphic residues using PyMol v.1.7.4.1 (Schrödinger, LLC).

4.4. Results

4.4.1. Testis-specificity is generally associated with an increase in evolutionary rate

To test for differences in evolutionary rate between testis-specific and ubiquitously expressed *Ago2* paralogues (see Section 3), we fitted models to a set of Dipteran *Ago2* sequences using codeml (PAML, Yang 1997). We find that most support (Akaike weight = 0.99) falls behind a model specifying a different ω for each *obscura* group *Ago2* subclade, and another separate ω for the *D. pseudoobscura-D. persimilis Ago2a-Ago2b* subclade. Under this model, the *D. pseudoobscura-D. persimilis Ago2a-Ago2b* subclade has the highest rate of protein evolution (ω=0.35$\pm$0.058), followed by the *Ago2f* subclade (ω=0.22$\pm$0.016), the *Ago2a* subclade (ω=0.20$\pm$0.015), the *Ago2e* subclade (ω=0.18$\pm$0.011), and finally the other Dipteran *Ago2* sequences (ω=0.13$\pm$0.002). This shows that the evolution of testis-specificity is generally accompanied by an increase in evolutionary rate.

4.4.2. Contrasting sequence characteristics of *Ago2* paralogues in the *obscura* group

We find several contrasts between the population genetic characteristics of the different paralogues (Table 4.1). In *D. subobscura* and *D. obscura*, *Ago2f* has the highest diversity of all paralogues at both synonymous and nonsynonymous sites, as well as the highest π_a/π_s ratio. In *D. pseudoobscura*, *Ago2e* has higher synonymous site diversity than all members of the *Ago2ad* subclade, but lower diversity at nonsynonymous sites, and therefore the lowest π_a/π_s ratio.

Additionally, there is evidence for codon usage bias, which differs in strength between clades. The highest levels are seen in the *Ago2e* clade, which has a low ENC in *D. obscura* (40.36) and *D. pseudoobscura* (34.24), and a high CBI in both species (0.51 & 0.73 respectively). *Ago2e* also displays low ENC compared with the genome as a whole, falling into the 8th percentile for *D. obscura*, and the 1st percentile for *D. pseudoobscura* (Figure 4.2). The *Ago2f* clade also shows codon usage bias but at a weaker level, with *D. subobscura* and *D. obscura Ago2f* having a lower ENC (45.63 & 48.39 respectively) and a higher CBI (0.41 & 0.33 respectively) than any member of the *Ago2a* clade. In the context of the whole genome, however, *Ago2f* does not have unusually high levels of codon usage bias in *D. subobscura* (31st percentile) or *D. obscura* (49th percentile).

Table 4.1.: Genetic diversity and codon usage summary statistics.
All values are displayed to 2dp

Paralogue	π_a **(%)**	π_s **(%)**	π_a/π_s	**ENC**	**CBI**	**Tajima's D**	**p value (Taj. D)**
D. subobscura Ago2ad	0.04	0.66	0.06	56.56	0.22	0.09	>0.10
D. subobscura Ago2f	0.21	1.31	0.16	45.63	0.41	-1.34	>0.10
D. obscura Ago2ad	0.07	0.39	0.17	56.22	0.23	-0.49	>0.10
D. obscura Ago2f	0.62	1.21	0.51	48.39	0.33	-0.62	>0.10
D. obscura Ago2e	0.02	0.43	0.05	40.36	0.51	-0.90	>0.10
D. pseudoobscura Ago2a1	0.17	0.23	0.73	54.08	0.25	-1.05	>0.10
D. pseudoobscura Ago2a3	0.21	0.61	0.34	54.158	0.24	-0.39	>0.10
D. pseudoobscura Ago2b	0.13	0.15	0.87	55.96	0.23	-2.14	<0.05
D. pseudoobscura Ago2c	0.18	0.86	0.21	55.24	0.23	-0.41	>0.10
D. pseudoobscura Ago2d	0.21	0.28	0.74	56.79	0.20	-0.76	>0.10
D. pseudoobscura Ago2e	0.07	2.33	0.03	34.24	0.73	-1.51	>0.10

Table 4.2.: McDonald-Kreitman test results.
Pn & Ps are the number of within-species polymorphisms after singletons have been removed. All values are displayed to 2dp, except $\omega(\alpha)$ which is displayed to 4dp

Paralogue	**Pn**	**Ps**	**Outgroup**	**Dn**	**Ds**	α	$\omega(\alpha)$	**p value**
D. subobscura Ago2a	2	9	*D. subobscura* / *D. obscura Ago2a* ancestor	26	72	0.39	0.0016	0.73
D. subobscura Ago2f	5	13	*D. subobscura* / *D. obscura Ago2f* ancestor	129	150	0.55	0.0011	0.15
D. obscura Ago2a	3	5	*D. subsilvestris Ago2a*	39	65	0.00	0.00	1.00
D. obscura Ago2f	15	12	*D. subsilvestris Ago2f*	92	106	-0.44	-0.0013	0.42
D. obscura Ago2e	1	5	*D. subsilvestris Ago2e*	87	145	0.67	0.0015	0.42
D. pseudoobscura Ago2a1	5	1	*D. pseudoobscura Ago2a1-d* ancestor	74	66	-3.46	-0.0157	0.22
D. pseudoobscura Ago2a3	5	4	*D. pseudoobscura Ago2a1-d* ancestor	72	71	-0.23	-0.0010	1.00
D. pseudoobscura Ago2b	5	1	*D. pseudoobscura Ago2a1-d* ancestor	78	54	-2.46	-0.0136	0.40
D. pseudoobscura Ago2c	8	8	*D. pseudoobscura Ago2a1-d* ancestor	47	78	-0.66	-0.0025	0.42
D. pseudoobscura Ago2d	5	3	*D. pseudoobscura Ago2a1-d* ancestor	85	72	-0.41	-0.0017	0.73
D. pseudoobscura Ago2e	0	17	*D. pseudoobscura Ago2e* & *D. affinis Ago2e* ancestor	77	120	1.00	0.0027	0.00

4.4.3. Positive selection on *D. pseudoobscura Ago2e*

To identify and measure positive selection on the *Ago2* paralogues in the *D. obscura* group, we initially performed McDonald-Kreitman (MK) tests (Table 4.2). We identified strong positive selection on *D. pseudoobscura Ago2e* (Fisher's exact test, p=0.0004), with 100% of substitutions estimated as adaptive (α=1.00). Although this estimate of α may have been biased by the use of a highly divergent outgroup (*D. affinis Ago2e*, Ks=0.30), this would not affect the validity of the test itself, which is instead driven by the extreme skew in pN/pS (0/17). MK tests for all other *Ago2* paralogues were non-significant (Fisher's exact test, p>0.1), implying that there is no positive selection on these loci. Alternatively, this could reflect a lack of power, caused by the low diversity at all of these loci (fewer than 10 polymorphic sites in most cases).

4.4.4. Extremely low diversity in the *Ago2a* & *Ago2e* subclades, and selective sweeps at *D. pseudoobscura Ago2a1/3*, *Ago2b* & *Ago2c*

As a complementary approach to identifying selection, we compared the synonymous site diversity at each *Ago2* paralogue in *D. pseudoobscura* with the distribution of genome-wide synonymous site diversity. We find that all paralogues have unusually low diversity: *Ago2a1*, *Ago2b* and *Ago2c* fall into the 1st percentile, *Ago2a3* and *Ago2d* into the 2nd percentile and *Ago2e* into the 8th percentile (Figure 4.3). ML-HKA tests confirm that the diversity of *Ago2a1-d* is significantly lower than the *D. pseudoobscura* genome as a whole (Akaike weight = 0.99).

We find similar results when comparing the diversity of *D. subobscura* and *D. obscura Ago2* paralogues to transcriptome-derived diversity distributions, suggesting that this low diversity is not limited to *D. pseudoobscura*. In *D. obscura*, *Ago2a* and *Ago2e* fall into the 2nd and 4th percentile respectively, whereas *Ago2f* falls into the 19th percentile (Figure 4.3). In *D. subobscura*, *Ago2a* falls into the 7th percentile, whereas *Ago2f* falls into the 16th percentile (Figure 4.3).

As an additional and more rigorous test for selection on the *Ago2* paralogues in *D. pseudoobscura*, we analysed the neighbouring region around each paralogue for the presence of selective sweeps, allowing us to estimate the strength, significance and location of a sweep for each paralogue. We find strong evidence for recent selective sweeps at or very close to *Ago2a1/3*, *Ago2b* and *Ago2c*, which display

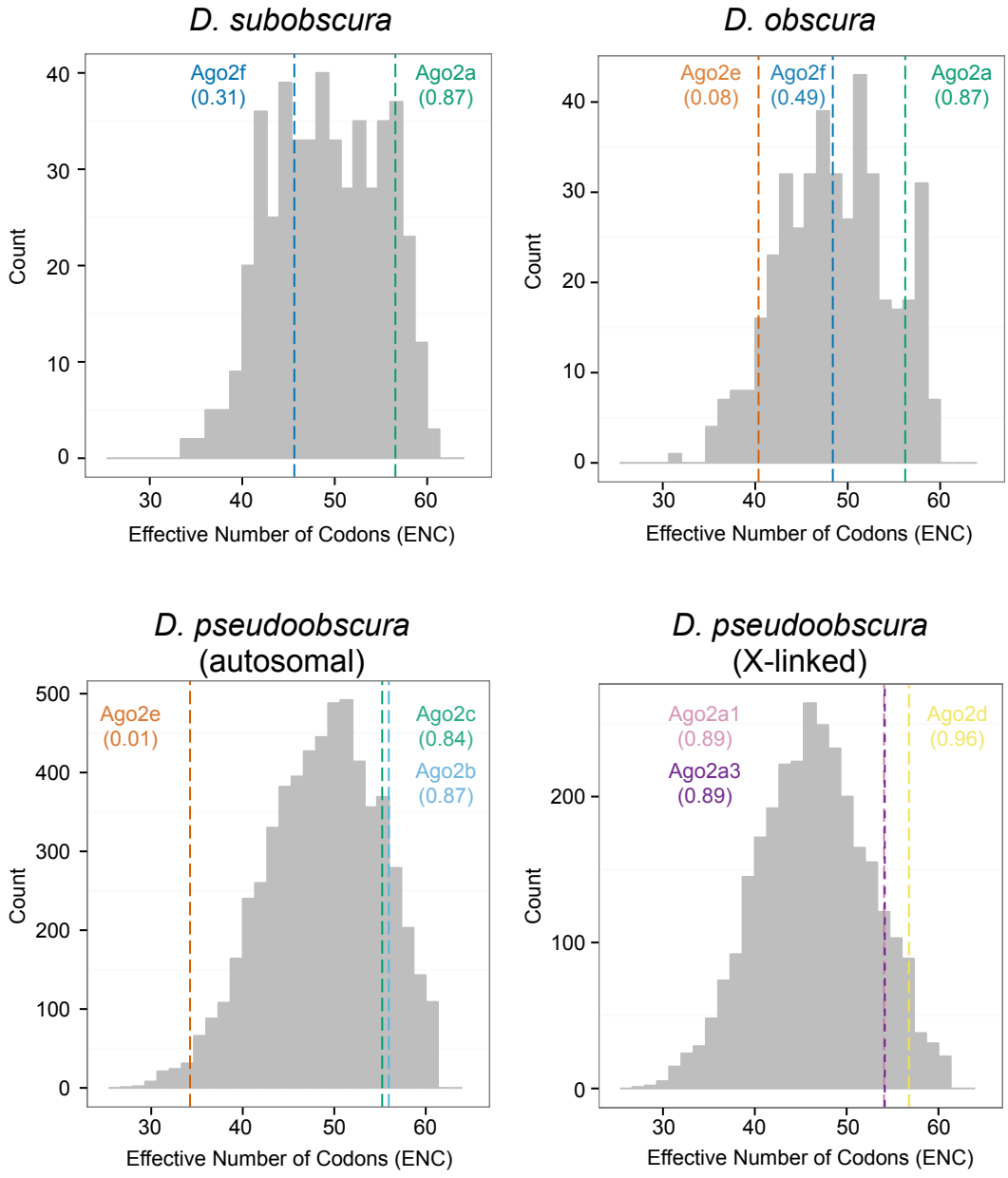


Figure 4.2.: The distribution of codon usage bias, derived from genome (*D. pseudoobscura*) or transcriptome (*D. subobscura* & *D. obscura*) data.
The percentile of the distribution into which each paralogue falls is indicated in brackets under the paralogue name. *Ago2e* has a very low effective number of codons (ENC) compared with the genome as a whole, indicating a high degree of codon usage bias.

sharp troughs in their diversity levels, and large peaks in the likelihood of a sweep which far exceed the significance threshold (p<0.01) (Figure 4.4). These localised reductions in diversity remain when our haplotype data is removed, indicating that these results are robust to demographic differences (Appendix C).There is ambiguous evidence for a sweep at *Ago2d*: there is one significant (p<0.01) likelihood peak just upstream of the paralogue, but two other peaks ~1kb and ~3kb further upstream. Sweepfinder finds no evidence for a sweep at *Ago2e*, with no likelihood peak or diversity trough. However, a SNP at position 1722 of *Ago2e* displays a significant $n\mathrm{S}_L$ value ($n\mathrm{S}_L$ = -1.812379, p<0.001), and has a derived allele frequency of 0.5, possibly indicating an incomplete selective sweep.

4.4.5. Selective constraint is consistent across gene length and protein structure

To quantify differences in selective constraint between different domains and structural components, we mapped polymorphisms along the length of each *Ago2* paralogue, and onto the 3D protein structure of *D. melanogaster* Ago2. We find that polymorphisms are distributed evenly across the gene length and protein structure, with no hotspots of polymorphism in any of the paralogues.

4.5. Discussion

4.5.1. Key findings

Our data give several indications that *Ago2* paralogues in the *obscura* group are evolving under strong selection, and provide strong evidence for selection on numerous paralogues in *D. pseudoobscura*. Firstly, we find low diversity at all *Ago2* paralogues, which is significantly lower than the genome-wide average for the majority of paralogues (Section 4.4.4), suggesting the action of positive selection. Secondly, we identify a clear and convincing signature of recent selective sweeps on *D. pseudoobscura Ago2a1/3*, *Ago2b* and *Ago2c* (Section 4.4.4), and an excess of nonsynonymous fixed differences at *D. pseudoobscura Ago2e* (Section 4.4.3), providing additional evidence for positive selection on *Ago2* paralogues. Finally, we find a general pattern of higher evolutionary rate after specialization to the testis (Section

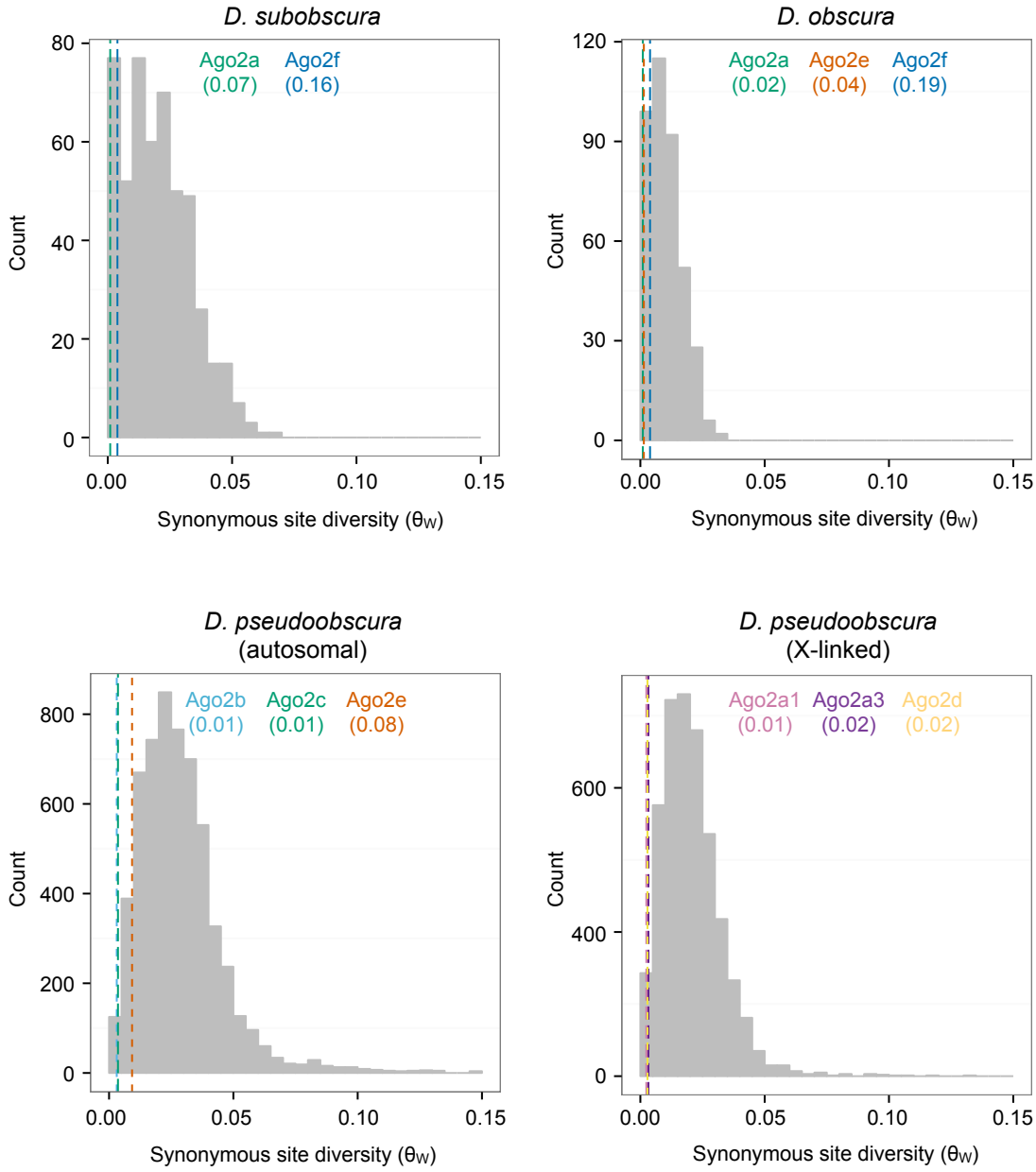


Figure 4.3.: The distribution of synonymous site diversity, derived from genome (*D. pseudoobscura*) or transcriptome (*D. subobscura* & *D. obscura*) data.
The percentile of the distribution into which each paralogue falls is indicated in brackets under the paralogue name. In each species, members of the *Ago2a* and *Ago2e* subclades have very low diversity compared with the genome as a whole.

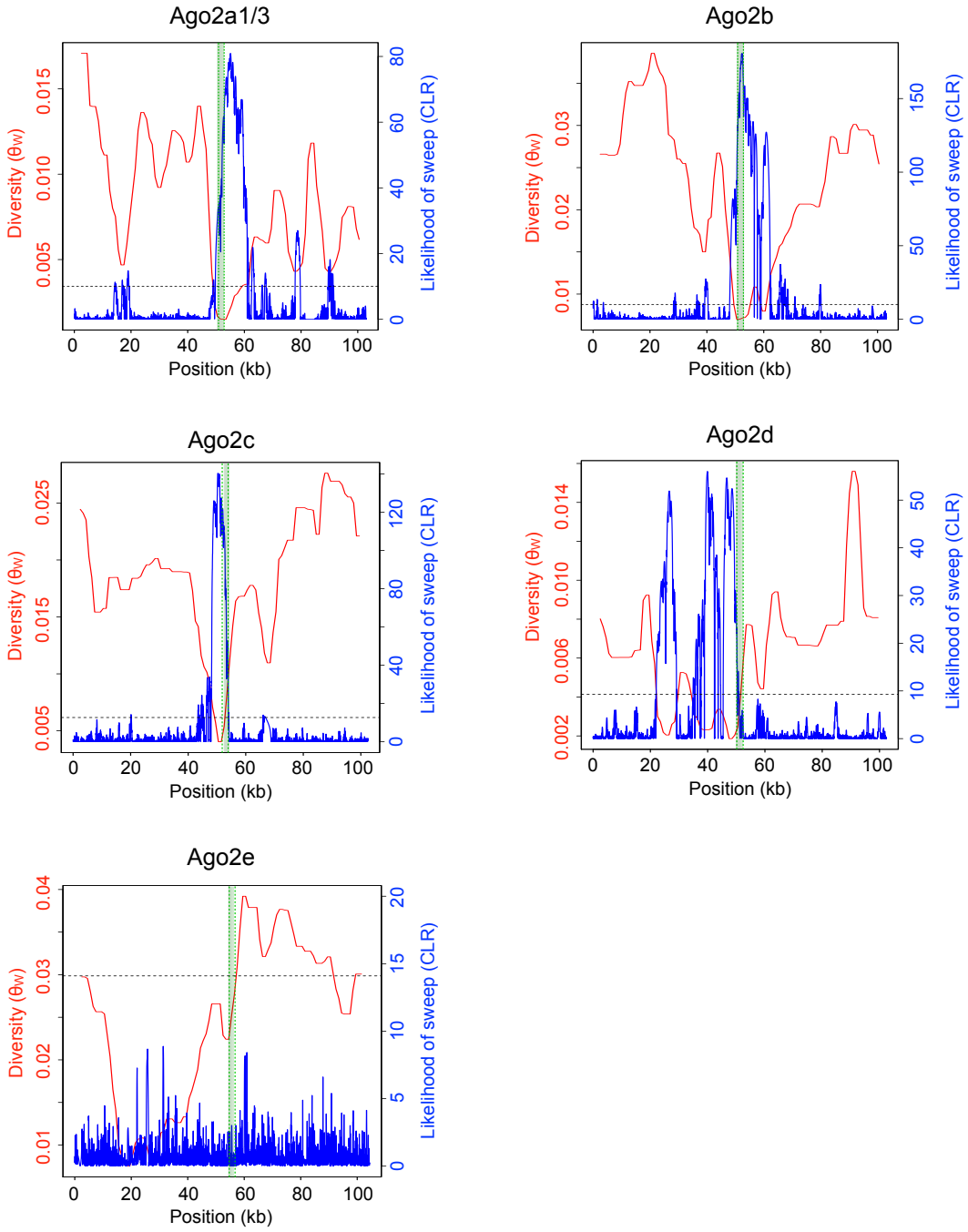


Figure 4.4.: Selective sweeps at *D. pseudoobscura Ago2* paralogues.
For each paralogue, diversity at all sites (Watterson's θ) is displayed in red, and the likelihood of a sweep (composite likelihood ratio, CLR) is displayed in blue. The significance threshold for this likelihood ratio is displayed by the horizontal dotted line (p<0.01, derived from the 10th-highest CLR out of 1000 coalescent simulations, assuming constant recombination rate and Ne as detailed in Section 4.3.3). There is strong evidence for sweeps at *Ago2a*, *Ago2b* and *Ago2c*, indicated by troughs in their diversity levels and peaks in the likelihood of a sweep.

4.4.1); combined with evidence for positive selection and selective sweeps, this indicates that the testis-specific function that has evolved in the majority of *Ago2* paralogues is not only highly derived, but may also be driven by strong selection.

4.5.2. Caveats & considerations

We discover a strong general pattern of reduced diversity in *Ago2* paralogues, which is unusually low for most paralogues. This trend is not pervasive, however, with much higher diversity at *Ago2f* than *Ago2a* & *Ago2e* (although *Ago2f* still falls into the lowest quartile of the distribution in both species). This may be caused by genomic diversity distributions for *D. subobscura* and *D. obscura* being derived from transcriptomic data, potentially enriching the genome-wide distribution for more easily detectable highly expressed genes, which often evolve more slowly and have lower levels of diversity (Pal *et al.* 2001; Lemos *et al.* 2005). This would result in a leftward (i.e. less diverse) skew in the overall distribution, not affecting the validity of any genes identified as having unusually low diversity (e.g. *Ago2a* & *D. obscura Ago2e*), but making other genes with low diversity harder to identify. The significance of the difference between *Ago2f* and the other *Ago2* subclades therefore remains unclear.

Diversity can be reduced by codon usage bias (CUB), which is caused by drift, mutational bias or selection for increased translational efficiency (Sharp and Li 1989; Pal *et al.* 2001; Lemos *et al.* 2005), and which is prevalent across *Drosophila* (Heger and Ponting 2007) and other insects (reviewed in Behura and Severson 2013). This does not appear to be driving the low diversity in the *Ago2a* subclade (all members of which show low levels of CUB), but may have contributed to the low diversity of *Ago2e* in *D. obscura* and *D. pseudoobscura*, both of which show extremely high levels of CUB (Table 4.1, Figure 4.2). In the case of *D. obscura*, *Ago2e* has higher GC content in its exons (57.93%) than its introns (54.84%), which may indicate that its high CUB is driven by selection. In contrast, there is similar GC content for both exons (62.70%) and introns (61.54%) of *D. pseudoobscura Ago2e*, suggesting that its high CUB may be driven by mutational bias rather than selection. Furthermore, the GC content of the exons of *Ago2e* is similarly high in the close relative *D. persimilis* (62.34%), but is lower (52.65% - 59.78%) in the more distant relatives *D. subsilvestris*, *D. tristis*, *D. azteca*, *D. affinis* and *D. lowei*. This may suggest that *Ago2e* may have translocated to a new location with higher mutational bias in the *D. persimilis-D. pseudoobscura* ancestor, although the lack of genomic data for most of these species prevents firm conclusions about the cause of this difference.

Another potential confounding factor is the recent origin of *D. pseudoobscura Ago2a1/3*, *Ago2b*, *Ago2c* and *Ago2d*, which may be expected to place an upper limit on diversity if the origin is later than the expected time of allele coalescence. Based on dates from our time-scaled gene tree (Figure 4.1) and assuming 10 generations per year (e.g. Cutter 2008), we estimate that *Ago2c* & *Ago2d* originated 18 million generations ago, *Ago2b* 36 million generations ago, and *Ago2a1* & *Ago2a3* fewer than 100,000 generations ago. In comparison, we expect the alleles of each paralogue to coalesce in 4Ne generations, or 4 million generations (assuming Ne=10^6). This suggests that it is only the diversity of *Ago2a1* & *Ago2a3* that may be limited by their recent origin. In this case, however, these paralogues have become fixed in the population despite an extremely recent origin, which is highly unlikely to have occurred by drift, and is in itself evidence of positive selection.

Finally, our analyses of selective sweeps localised some likelihood peaks slightly upstream (*Ago2c*) or downstream (*Ago2a1/3*) of the *Ago2* paralogue position. This could be caused by selection on cis-acting regulatory elements flanking the paralogue, selection on neighbouring loci, chance recombination events during a sweep, or local fluctuations in recombination rate, which shows considerable variation across the *D. pseudoobscura* genome (McGaugh *et al.* 2012). In the case of *Ago2c*, there is one locus (*FBgn0071384*) immediately upstream that could contribute to the 5' shift in the likelihood peak, meaning that selection cannot be unequivocally attributed to this paralogue. In contrast, there are no neighbouring loci around *Ago2a1/3*, making the shift in its likelihood peak more likely due to recombination rate variation or selection on regulatory elements. Combined with the unambiguous positioning of the highest likelihood peak within the coding region of *Ago2b*, this indicates that selective sweeps are likely to have acted on at least two of the *Ago2* paralogues in *D. pseudoobscura*.

4.5.3. Implications

Our finding of strong selection on *Ago2* paralogues in the *obscura* group mirrors the results of previous analyses of *D. melanogaster* (Obbard *et al.* 2006; Obbard *et al.* 2009b; Obbard *et al.* 2011) and *Drosophila* generally (Kolaczkowski *et al.* 2011). In these species, however, *Ago2* is generally present in a single copy (Chapter 2), in contrast to the frequent duplications seen in the *obscura* group (Figure 4.1). Selection therefore appears to be retaining paralogues of *Ago2* in the *obscura* group, as well as driving rapid evolution of these paralogues after they have been fixed.

After duplication, *Ago2* paralogues have either retained their (presumably ancestral) ubiquitous expression pattern, as seen for *D. subobscura Ago2a*, *D. obscura Ago2a* and *D. pseudoobscura Ago2c*, or specialized to the testis, as seen for *D. pseudoobscura Ago2a* & *Ago2b*, and for the *Ago2e* & *Ago2f* clades (Chapter 3). Our data suggest that both of these expression patterns appear to be under strong selection. The ubiquitously-expressed paralogues all have low diversity compared with the genome as a whole (Figure 4.3), and *D. pseudoobscura Ago2c* also shows evidence of a recent selective sweep (Figure 4.4). This is consistent with the patterns of selection on ubiquitously expressed *Ago2* in other species, which also show evidence for recent selection (Obbard *et al.* 2006; Obbard *et al.* 2009b) and selective sweeps (Obbard *et al.* 2011). This suggests that the selection pressures acting on ubiquitously expressed Ago2 paralogues in the *obscura* group may be similar to those previously suggested for *Ago2* genes in other *Drosophila* species, such as an arms race with viruses (Obbard *et al.* 2006; Kolaczkowski *et al.* 2011).

Positive selection also appears to be acting on some testis-specific *Ago2* paralogues, as evidenced by the low diversity of *D. obscura Ago2e* and *D. pseudoobscura Ago2a1*, *Ago2a3* & *Ago2b*, the selective sweeps identified on *D. pseudoobscura Ago2a* & *Ago2b*, and the positive selection on *D. pseudoobscura Ago2e*. This is consistent with functional divergence, possibly to specialise to a pre-existing function (see Chapter 3), or even a novel testis-specific function, as seen in other testis-specific genes in *D. melanogaster*, which evolve at a faster rate than the genome-wide average (Haerty *et al.* 2007). However, these evolutionary patterns are not consistent across all testis-specific paralogues, with the diversity of *D. subobscura* & *D. obscura Ago2f* and *D. pseudoobscura Ago2e* not significantly lower than the genome-wide level, and no selective sweep identified on *D. pseudoobscura Ago2e*. This could indicate the evolution of contrasting functions, which impose different selection pressures; alternatively, testis-specificity may be driven by the same selection pressure, which is subsequently relaxed in some lineages.

In conclusion, we find low diversity at the majority of *obscura* group *Ago2* paralogues, and identify selective sweeps on *D. pseudoobscura Ago2a*, *Ago2b* and *Ago2c*. Combined with the general increase in evolutionary rate after specialization to the testis, this suggests that some, if not all, testis-specific paralogues are evolving under positive selection.

5. CRISPR/Cas9-mediated knockout of *Ago2a1-Ago2e* in *Drosophila pseudoobscura*

5.1. Introduction

Previously, we have documented the discovery of paralogues of *Ago2* in the *D. obscura* group (Chapter 3). We have also found that these paralogues have frequently specialised to the testis (Chapter 3), and are under strong positive selection (Chapter 4). This selection may be imposed by an anti-TE role, because *Ago2* has a well-characterized anti-TE role in the soma and germline of *D. melanogaster* (Czech *et al.* 2008; Chung *et al.* 2008), and because some TEs are more active in the testis (e.g. *copia*; Pasyukova *et al.* 1997; Morozova *et al.* 2009). We sought to create knockouts of the *Ago2* paralogues in *D. pseudoobscura*, in order to characterize the nature of their functions. In addition, several of these paralogues show evidence of functional divergence while still retaining very close sequence homology to each other (Chapter 3). We therefore aimed to generate knockouts to test for functional redundancy between these recent paralogues.

A variety of transgenic techniques have been developed in the last 25 years, and have provided a valuable insight into the functions of *Argonaute* genes in *D. melanogaster*. One of the earliest techniques involved the use of TEs, either to knock down a gene by inserting and disrupting its open reading frame, or to insert an exogenous gene (Rubin and Spradling 1982). For example, TE insertion-mediated knockout revealed that *Ago1* functions downstream of sRNA production during embryo development (Williams and Rubin 2002), and TE-induced deletion demonstrated a key role for *Ago2* in antiviral defence (Rij *et al.* 2006). A more precise method was introduced by RNA interference (RNAi), which allows spe-

cific knockdown of gene expression by using complementary small RNA guides to target transcripts (Fire *et al.* 1998), and which has revealed a key role for *Ago2* in somatic TE suppression (Chung *et al.* 2008). More recently, zinc-finger nucleases (ZFNs) (Bibikova *et al.* 2002) and transcription activator-like effector nucleases (TALENs) (Miller *et al.* 2011) have been employed to induce specific, heritable changes to the genome, with a TALEN-based library of miRNA knockouts revealing key characteristics of miRNA-target interactions (Kim *et al.* 2013b).

Despite the enormous potential of these techniques, several factors have limited their use to model species. Firstly, most techniques use microinjection to deliver a reagent mix to embryos: the viability of these embryos after injection often depends on a precise balance of reagents, which has only been optimised for frequently used species such as *D. melanogaster* and *Mus musculus*. Secondly, the identification of transformants often requires the use of a marker gene driven under specific promoters, the correct combinations of which have only been characterised for a handful of species (e.g. Berghammer *et al.* 1999). Thirdly, most transgenic techniques either have low specificity at a low cost (e.g. TE-mediated insertion), necessitating the screening of thousands of potential transformants, or high specificity at a high cost (e.g. ZFN & TALEN), limiting the number of attempts that can be made at transformation. Finally, many transgenic experiments require several generations of backcrossing and full-sib mating to produce homozygous transformants, which is logistically challenging in species with long generation times.

Recently, however, CRISPR/Cas9 (clustered regularly interspersed short palindromic repeats / CRISPR-associated 9) has emerged as a technique that overcomes many of these limitations. This technique is based on the CRISPR/Cas bacterial genome defence mechanism, which provides a defence against bacteriophages and other invading genetic elements through a small RNA-guided mechanism that is analogous to RNAi (Makarova *et al.* 2006; Barrangou *et al.* 2007). The mechanism is triggered by the invasion of a foreign genetic element, which is recognised and cleaved by a Cas protein (Garneau *et al.* 2010) to generate a small DNA termed a spacer (Barrangou *et al.* 2007). This spacer is integrated into a CRISPR array (Barrangou *et al.* 2007), which consists of multiple spacers (each of which is homologous to a specific phage, plasmid or other foreign element) separated by short palindromic repeats (Makarova *et al.* 2006), and often also contains one or multiple *Cas* genes. These CRISPR arrays are then transcribed and processed into CRISPR RNAs (crRNAs), which bind Cas proteins and guide them to invading genetic elements through complementary base pairing, upon which the Cas protein cleaves the element and prevents its invasion (reviewed in Oost *et al.* 2014).

This mechanism has now been adapted and refined into a highly specific, low cost transgenic technology (Cong *et al.* 2013). All CRISPR-mediated techniques are based on the core interaction between crRNAs (which mediate specific targeting) and Cas9 (which cleaves the target region). Each crRNA is designed to bind to a 17-20nt region in the target locus, immediately upstream of an NGG trinucleotide (the protospacer adjacent motif or PAM). If one crRNA is used, Cas9 induces a double-stranded break (DSB) in the target locus which is repaired by the non-homologous end-joining repair pathway, often resulting in a frame shift or deletion. If two chiRNAs are used, Cas9 can induce DSBs on either side of the target locus, resulting in its complete removal from the genome (Figure 5.1, top). To replace the target locus with another sequence, this technique can be combined with a donor DNA strand, which consists of a replacement sequence (usually a marker gene or a modified version of the target locus) flanked by sequences that are homologous to the neighbouring regions of the target locus. When the target locus is cut out of the genome, the homologous flanking regions in the donor DNA induce the endogenous homology-directed repair mechanism to repair the gap, with the replacement sequence used as the template (Figure 5.1, bottom).

CRISPR/Cas9 has numerous advantages over other transgenic technologies, which make it more tractable in non-model organisms. Firstly, CRISPR/Cas9 has high efficiency, with one study in *Danio rerio* transforming embryos with >50% efficiency; this means that fewer embryos need to be injected and fewer potential transformants screened, and therefore allows the technique to be used in species from which only small numbers of embryos can be gathered. Secondly, CRISPR/Cas9 components can be delivered in numerous different ways, ranging from microinjection of the chiRNAs and *Cas9* mRNA, to injection of plasmids encoding these components driven under conserved promoters, through to the use of transgenic lines which have been previously engineered to express some components (e.g. *D. melanogaster* lines expressing *Cas9*; Port *et al.* 2014). This variety of delivery options allows the technique to be applied to a wide range of different organisms. Finally, the fact that target specificity is determined by short RNAs, which can be rapidly designed and cheaply manufactured, removes much of the financial constraint that precludes the application of other techniques such as ZFN and TALEN, and also allows chiRNA design to be tested and optimised quickly and efficiently.

These characteristics have allowed CRISPR to be applied beyond conventional genetic model species to a wide range of organisms. In the cynomolgus monkey *Macaca fascicularis* (Niu *et al.* 2014) and the domestic pig *Sus scrofa domesticus* (Hai *et al.* 2014), coinjection of crRNA and *Cas9* mRNA was used to create indels in single target genes. In the sea squirt *Ciona intestinalis* (Stolfi *et al.* 2014) and the goat *Capra aegagrus hircus* (Ni *et al.* 2014), vectors carrying the *Cas9* coding sequence (CDS) and crRNAs

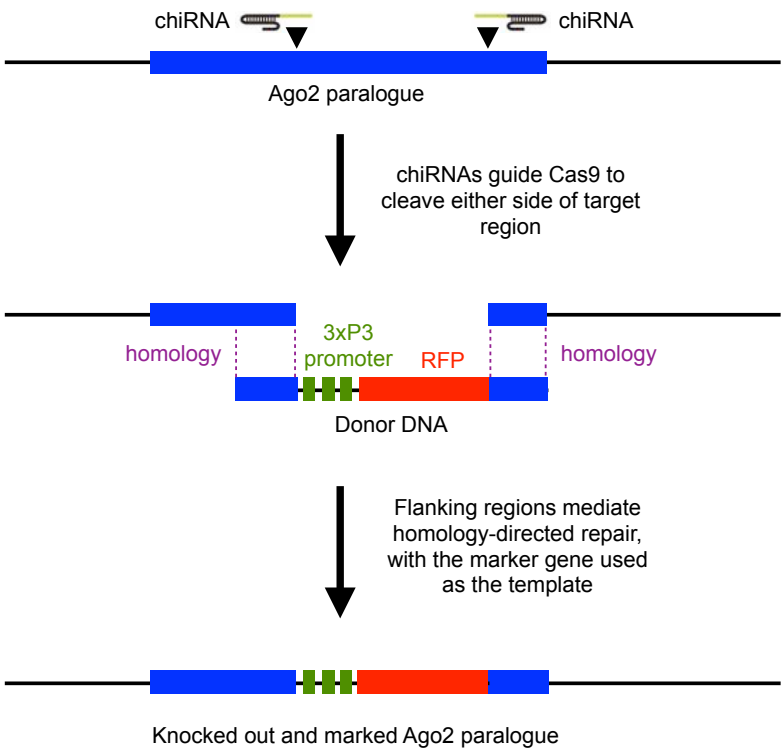


Figure 5.1.: *Ago2* paralogue knockout by CRISPR/Cas9 and homology-directed repair.
Firstly, chimeric RNAs (chiRNAs) guide Cas9 to form double-stranded breaks either side of each *Ago2* paralogue. Secondly, regions of homology flanking the donor double-stranded DNA guide homology-directed repair of these breaks, with *RFP* used as the template. This results in a repaired region that has *RFP* in place of the target region, marking all transformant individuals with eye-specific fluorescence.

were used to create indels in target loci, with the latter study finding very high efficiency of transformation (Ni *et al.* 2014). Finally, in the protozoan parasite *Trypanosoma cruzi*, plasmids carrying the *Cas9* CDS and three redundant crRNAs were used to induce indels in the 65-member β-*Galactofuranosyl glycosyltransferase* (β-*GalGT*) gene family (Peng *et al.* 2014).

5.2. Aims

We have previously found that *Ago2* paralogues in the *obscura* group have undergone extensive functional divergence (Chapter 3), carrying out a testis-specific function that may be evolving under strong positive selection (Chapter 4). Until recently, transgenic technologies did not exist for the *obscura* group; however, the CRISPR/Cas9 technique now allows the highly specific production of knockouts in non-model organisms. We therefore aimed to use CRISPR/Cas9 to create knockouts of the *Ago2*

paralogues in *D. pseudoobscura*, with the following goals:

1. Test for a possible function of *Ago2* paralogues in TE suppression, by comparing transposition levels in the soma and germline between wild type and knockout individuals.

2. Test whether *Ago2* paralogues suppress meiotic drive, by comparing the fertility, fecundity and sex ratios of offspring between wild type and knockout individuals.

3. Quantify the extent of redundancy and functional divergence, by comparing the phenotypic effects of single and combinatorial *Ago2* paralogue knockouts.

5.3. Methods

5.3.1. Details of constructs carrying CRISPR components

We used the *Cas9* CDS from *Streptococcus pyogenes* (plasmid MLM3613, Addgene, Cambridge, MA, USA). We aimed to mark transformants with *red fluorescent protein* (*RFP*) driven under the highly conserved eye-specific promoter 3xP3 (plasmid pM{3xP3-RFPattP}, Bischof *et al.* 2007), as described in Berghammer *et al.* 1999 and demonstrated in *D. pseudoobscura* (our focal species) by Holtzman *et al.* 2010. Both *Cas9* and *RFP* were carried on plasmids which also harboured an ampicillin reporter and PCR priming sites for CDS amplification. The pM{3xP3-RFPattP} plasmid was transformed into an *E.coli* vector using the heat-shock method, whereas the *Cas9* plasmid was already transformed into *E.coli*.

To target *D. pseudoobscura Ago2a1-d* and *Ago2e* separately, we designed four chimeric RNAs (chiRNAs): two targeting conserved regions at the 5' and 3' ends of *Ago2a1-d*, and two targeting the 5' and 3' ends of *Ago2e*. The 5' and 3' *Ago2a1-d* chiRNAs were perfectly complementary to the target sites in *Ago2a1-d*, and had 6 and 5 mismatches respectively to the corresponding region in *Ago2e*. Similarly, the *Ago2e* 5' and 3' chiRNAs had 7 and 5 mismatches respectively to the corresponding region in *Ago2a1-d*, reducing the possibility of unintended off-target effects. We targeted conserved regions in *Ago2a1-d* because of their high degree of sequence similarity (which reduces the number of unique target sites in

each paralogue), and because a previous study successfully targeted conserved regions in a large gene family with a small number of crRNAs (Peng *et al.* 2014). Each chiRNA consisted of a 17-19bp guide RNA (gRNA) complementary to the target region, a 77bp RNA scaffold, a T7 promoter and stop motif, and forward and reverse priming sites (this architecture was based on a chiRNA originally designed by the Church Lab, Harvard, MA, USA).

5.3.2. Synthesis of CRISPR components

To amplify the *Cas9* CDS for PCR, we incubated *E. coli* carrying the *Cas9* plasmid overnight on ampicillin-agar plates, and incubated single colonies overnight in LB broth. We extracted plasmid DNA from these incubations using the QIAprep kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions, and amplified the *Cas9* CDS by PCR using Platinum Taq (Invitrogen). We transcribed mRNA from this PCR product using the mMessage mMachine transcription kit (Ambion), and added a polyA tail to these transcripts using the PolyA Tailing kit (Applied Biosystems, Foster City, CA, USA). We purified polyA-tailed transcripts using a phenol:chloroform extraction with an isopropanol precipitation, and verified successful polyA tailing by agarose gel electrophoresis.

To synthesize the chiRNAs, each chiRNA was first synthesized as a DNA oligonucleotide by Sigma-Aldrich (St. Louis, MO, USA). We then amplified each oligo by PCR using high fidelity Platinum Taq (Invitrogen) (see Table 5.1 for PCR primers). We transcribed RNA from the PCR products using the mMessage mMachine transcription kit (Ambion), purified RNA using the MEGAclear kit (Ambion), and verified successful transcription by agarose gel electrophoresis.

To synthesize the *RFP* donor DNA, we incubated *E. coli* carrying the *RFP* plasmid overnight on ampicillin-agar plates, and incubated single colonies overnight in LB broth. We extracted plasmid DNA from these incubations using the QIAprep kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions, amplified *RFP* by PCR using Platinum Taq (Invitrogen), and verified successful amplification by agarose gel electrophoresis. We used primers that were designed to incorporate 60-128nt of flanking sequence onto the 5' and 3' ends of *RFP*, homologous to the sequence upstream and downstream of the excised region in each of *Ago2a1-e* (see Table 5.1 for PCR primers). Due to the extremely high sequence similarity of the flanking regions around *Ago2a1* and *Ago2a3*, we used the same pair of primers to synthesize donor DNA for both loci, resulting in a single *Ago2a1/Ago2a3* donor

DNA which should guide homology-directed repair of both loci with equivalent efficiency.

Table 5.1.: *D. pseudoobscura* CRISPR primers

Name	**Sequence (5' - 3')**
gRNA_F	TGTACAAAAAAGCAGGCTTTAAAG
gRNA_R	TAATGCCAACTTTGTACAAGAAAG
MLM3613_Cas9_CMV_F	CAAATGGGCGGTAGGCGTG
MLM3613_Cas9_BGH_R	GCAACTAGAAGGCACAGTCGAGG
pM_3xP3_RFP_128overlap_a_F	CGCAAGGATGGAGCTAATCAGGTGGCGGCCAT GATTAAGTATGCTGCCACTTCCACCAACGAGAG GAAGGCCAAGATCATCCGCTTGATGGAATATTT CAGGCACAATTTAGACCCGACCATCAGCCACGA AGTTATGTCGACGAATTCG
pM_3xP3_RFP_60overlap_ab_R	CAGTGAGGTAGACACGTCCACGTGCCGCAGCC AAATGTGCCAAATATGCCGGAGCCGGATGGCG AGCTCGAATTCGATGG
pM_3xP3_RFP_128overlap_b_F	CGCAAGGATACTGCTGCTCGGGTGGCGGCCATA CTTAAGTTTGCTGCCACGTCAACCAACGAGAGG AAGGCCAAGATCGTCCGCCTACTGGAATATTTC AAGCACAATTTAGACCCGACCATCAGCCACGA AGTTATGTCGACGAATTCG
pM_3xP3_RFP_60overlap_ab_R	CAGTGAGGTAGACACGTCCACGTGCCGCAGCC AAATGTGCCAAATATGCCGGAGCCGGATGGCG AGCTCGAATTCGATGG
pM_3xP3_RFP_128overlap_c_F	CGCAAGGATGGAGCTAATCAGGTGGCGGCCAT GATTAAGTATGCTGCCACTTCCACCAACGAGAG GAAGGCCAAGATCATCCACTTGCTCGAATATTT CAAACACAACTTAGATCCGACCATCAGCCACGA AGTTATGTCGACGAATTCG
pM_3xP3_RFP_60overlap_cd_R	CAGTGAGATAGACACGTCCACGTGCCGCAGCC AAATGTGCCAAATATGCCGGAGCCGGATGGCG AGCTCGAATTCGATGG
pM_3xP3_RFP_97overlap_d_F	TGATTAAGTATGCTGCCACTCCCACCAACGAGA GGAAGGCCAAGATCATCCGCTTGATGGAATATT TCAGGCACAATTTAGACCCGACCATCAGCCACG AAGTTATGTCGACGAATTCG
pM_3xP3_RFP_60overlap_cd_R	CAGTGAGATAGACACGTCCACGTGCCGCAGCC AAATGTGCCAAATATGCCGGAGCCGGATGGCG AGCTCGAATTCGATGG
pM_3xP3_RFP_97overlap_e_F	TTTGGGCTGTTCCAGGCGCTGGTGCTCGGCGAT CGGCCGTTCGTGAACGTGGACATAACGCACAA GTGCTTCCACCTGGCGATGCCGGTCGTCGAGTC GAAGTTATGTCGACGAATTCG
pM_3xP3_RFP_60overlap_e_R	CAGTGAGGTACACGCGTCCGCGAGCGGCGGCC AGGTGAGCCAGGTAGGCCGGGGCCGGGTGGCG AGCTCGAATTCGATGG

5.3.3. Microinjection of CRISPR mixtures

To deliver the synthesised CRISPR components to *D. pseudoobscura*, we injected a mixture of *Cas9* mRNA, chiRNAs and *RFP* donor DNA into embryos from *D. pseudoobscura* genome strain individuals using a microinjector kept under positive pressure (injections were carried out by Dr Sang Chan at the Fly Facility, University of Cambridge, UK). We injected two mixtures separately, the first to target *Ago2a1-d* and the second to target *Ago2e*. The *Ago2a1-d* mixture was composed of 0.75ug/ul *Cas9*, 0.25ug/ul of each chiRNA and 0.0625ug/ul of each donor DNA. The *Ago2e* mixture was composed of 0.75ug/ul *Cas9*, 0.25ug/ul of each chiRNA and 0.25ug/ul donor DNA. For each mixture, 500 embryos were injected.

5.3.4. Design of crosses

To generate large numbers of heterozygote knockouts in the F1 generation (in case of inviability of homozygous knockouts), we backcrossed each injected adult (F0) to an uninjected genome strain individual. We pooled the offspring (F1) of these crosses for 10 days, and then separated offspring into individual vials. We grouped males with multiple genome-strain virgin females, and kept females alone to lay eggs, and then screened for transformants 21 days later. Assuming Mendelian segregation of modified *Ago2* loci, we expect half of the offspring (F2) of each F1 transformant to be heterozygous transformant, and the other half to be homozygous WT. To increase the number of heterozygotes, we backcrossed each of the F2 offspring to an uninjected genome strain individual for 21 days, before screening for transformants. For the offspring (F3) of potential transformants in the F2 generation, we paired full-sibs for 14 days of mating. Assuming Mendelian segregation, we expect 25% of these pairs to be made up of two heterozygous transformant individuals, 25% of the offspring (F4) of which are expected to be homozygous transformants (see Figure 5.2 for a diagram of the crossing scheme).

5.3.5. Identification of transformants

We attempted to identify transformants using four methods. Firstly, we assayed individuals for *RFP* activity in the eyes using a Dark Reader DR89X Transilluminator (Clare Chemical, Dolores, CO). Sec-

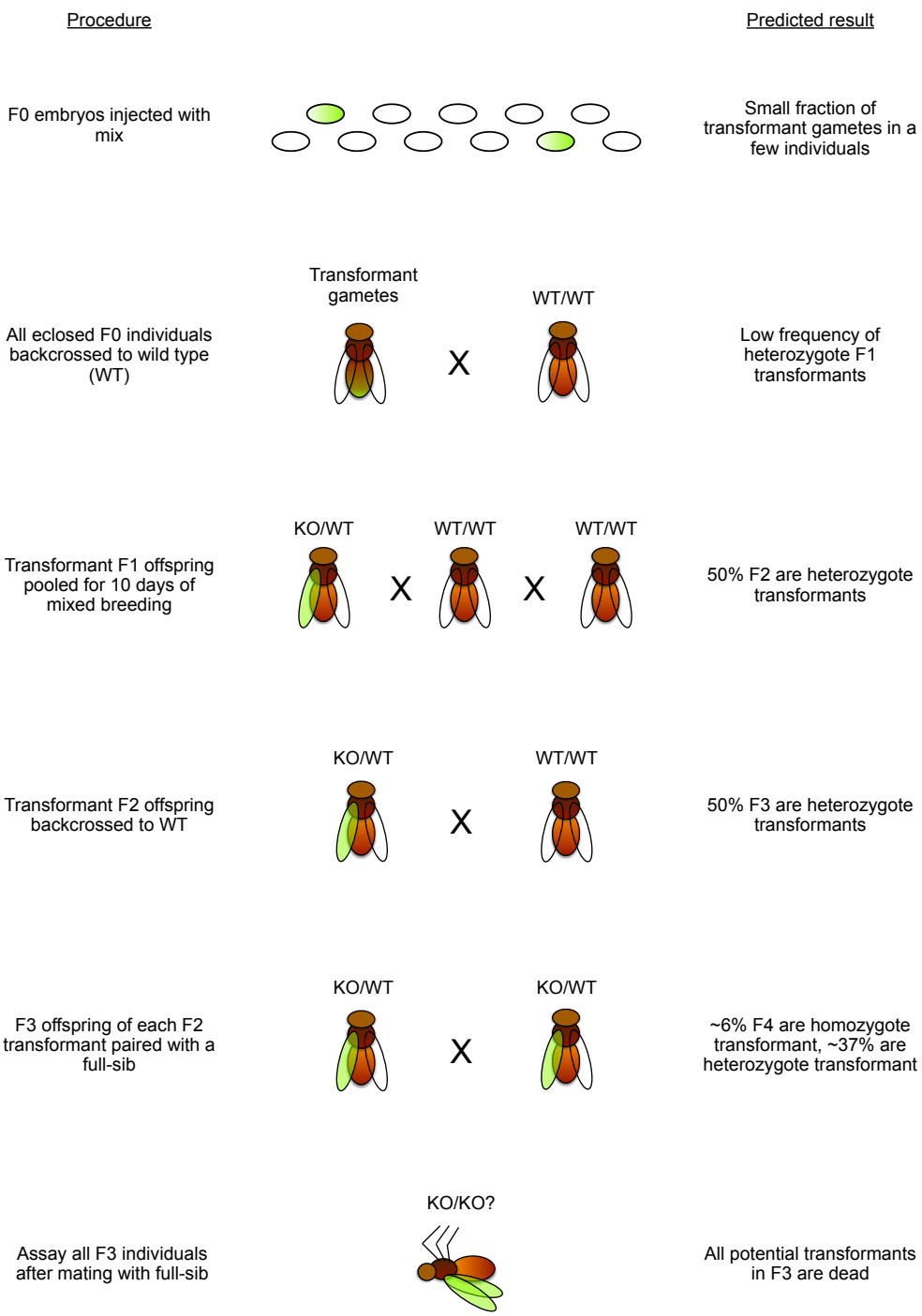


Figure 5.2.: Design of crosses.
In the F1 generation there is a very low expected proportion of transformants. Two generations of backcrossing are then employed, in order to increase the number of transformant *Ago2* paralogue loci while keeping them in a heterozygote state. In the F3 generation full-sib pairs are set up, with the aim of pairing some heterozygote transformants, and therefore producing homozygote transformant offspring.

ondly, we used PCR to screen for the presence of the *RFP* gene. Thirdly, we attempted to confirm that *RFP* insertion had occurred at or near an *Ago2* locus by using a "bridging" primer pair, consisting of one primer that binds the 5' end of *RFP* and another that binds the 3' end of all *Ago2* paralogues. Finally, we tested whether the *Ago2* gene region length had altered in line with expectation, given a deletion of ~1365bp (*Ago2a1-d*) or ~2170bp (*Ago2e*) and an insertion of ~975bp (*RFP*), using primers that spanned the intended modification sites of each paralogue (primers and cycling conditions as in Chapter 3). Due to the extremely high sequence similarity between *Ago2a1* and *Ago2a3*, we used a redundant primer pair that would amplify both loci, and refer to these loci as "*Ago2a*" when referring to PCR-based assays. For *Ago2a1-d*, we expected deleted but unmarked loci to be ~1365bp shorter than wild-type, and deleted and *RFP*-marked loci to be ~330bp shorter. For *Ago2e*, we expected deleted but unmarked loci to be ~2170bp shorter than wild-type, and deleted and *RFP*-marked loci to be ~1200bp shorter. DNA extractions from F1 offspring were carried out using the Chelex method (Walsh *et al.* 1991), and all other DNA extractions were performed using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

Due to mispriming and ambiguity of markers (see Sections 5.4.1 & 5.4.2), we adjusted our method of identifying transformants in different generations. In the F1 generation, the frequency of transformants was predicted to be very low: we therefore used visual fluorescence as the primary indicator, and used the bridging PCR to confirm *RFP* insertions in fluorescent individuals. In the F2 generation the expected frequency of transformants was much higher, as all individuals were offspring of at least one putative transformant. Additionally, validation of the F1 generation had revealed that the *RFP* mispriming sites may be segregating in the population used for microinjection. We therefore used the bridging PCR as the primary indicator in the F2 generation, and used the *Ago2* PCRs to identify which *Ago2* paralogue (if any) had been knocked out in these potential transformants. The F2 generation also revealed that the segregating *RFP* priming sites may interfere with the bridging PCR; we therefore used only the *Ago2* PCRs to identify transformants in the F3 generation.

5.3.6. Sampling strategy

The financial and logistical constraints of switching to PCR-based markers (see Section 5.3.5) necessitated the subsampling of individuals from each generation. We therefore froze all individuals from each generation after a period of mating, but extracted DNA and carried out PCR assays on a subset of

individuals. In the F1 generation, we carried out the *RFP* PCR on all individuals showing fluorescence, the bridging PCR on a subset of 26 individuals showing fluorescence, and the *Ago2* PCRs on a subset of 18 fluorescent individuals which had produced offspring. In the F2 generation, we carried out the bridging PCR on a randomly selected subset of 59 individuals, and the *Ago2* PCRs on the 13 potential transformants identified in this subset. In the F3 generation, we used the *Ago2* PCRs to screen for transformants in 4-7 F3 pairs each from 4 F2 pairs (18 F3 pairs in total), as an initial estimation of which F2 pairs were the most likely to produce transformant offspring. For F2 pairs identified as producing at least one transformant offspring in this screen, we used the *Ago2* PCRs to screen all remaining pairs of F3 offspring.

5.4. Results

5.4.1. Frequency and reliability of the fluorescent marker in the F1 generation

We observed a very low (~10%) proportion of successful eclosion of injected individuals: 31 females and 20 males for the *Ago2a1-d* mix, and 37 females and 15 males for the *Ago2e* mix. This was significantly female biased for the *Ago2e* mix (Fisher's Exact Test, p=0.0442), but not for the *Ago2a1-d* mix (Fisher's Exact Test, p=0.3198). In the F1 progeny of these individuals, we found no individuals with red fluorescence in the eye (the expected site of *RFP* expression). However, we did observe fluorescence on the dorsal face of the wings and the left, right and dorsal sides of the thorax, with several individuals harbouring large fluorescent tissue or other material on the thorax (Figure 5.3). We also observed a difference between injection mixes in the overall number of fluorescent individuals, with individuals injected with the *Ago2e* mix producing more fluorescent F1 individuals than the *Ago2a1-d* mix (Chi-squared Test, p=0.016). The injection mixes also differed in the relative proportions of the location of fluorescence: F1 individuals from the *Ago2a1-d* mix primarily fluoresced in the thorax only or wings only, whereas a larger proportion of individuals from the *Ago2e* mix fluoresced at both locations (Fisher's Exact Test, p<0.0001). Finally, there was also a sex-biased difference in the overall number of fluorescent individuals, with both mixes producing more fluorescent males than females; while this

was not a significant difference (Fisher's Exact Test, p=0.08), there was a significant difference between the sexes in the location of fluorescence (Fisher's Exact Test, p<0.0001), with females fluorescing more frequently in the wings or wings and thorax, and males in the thorax only (Figure 5.4). The presence of fluorescence implies that knock-in of the *RFP* marker locus was successful, but does not confirm the knock-out of the *Ago2* paralogues.

To validate this visual fluorescence marker, we assayed fluorescence in a random sample of uninjected

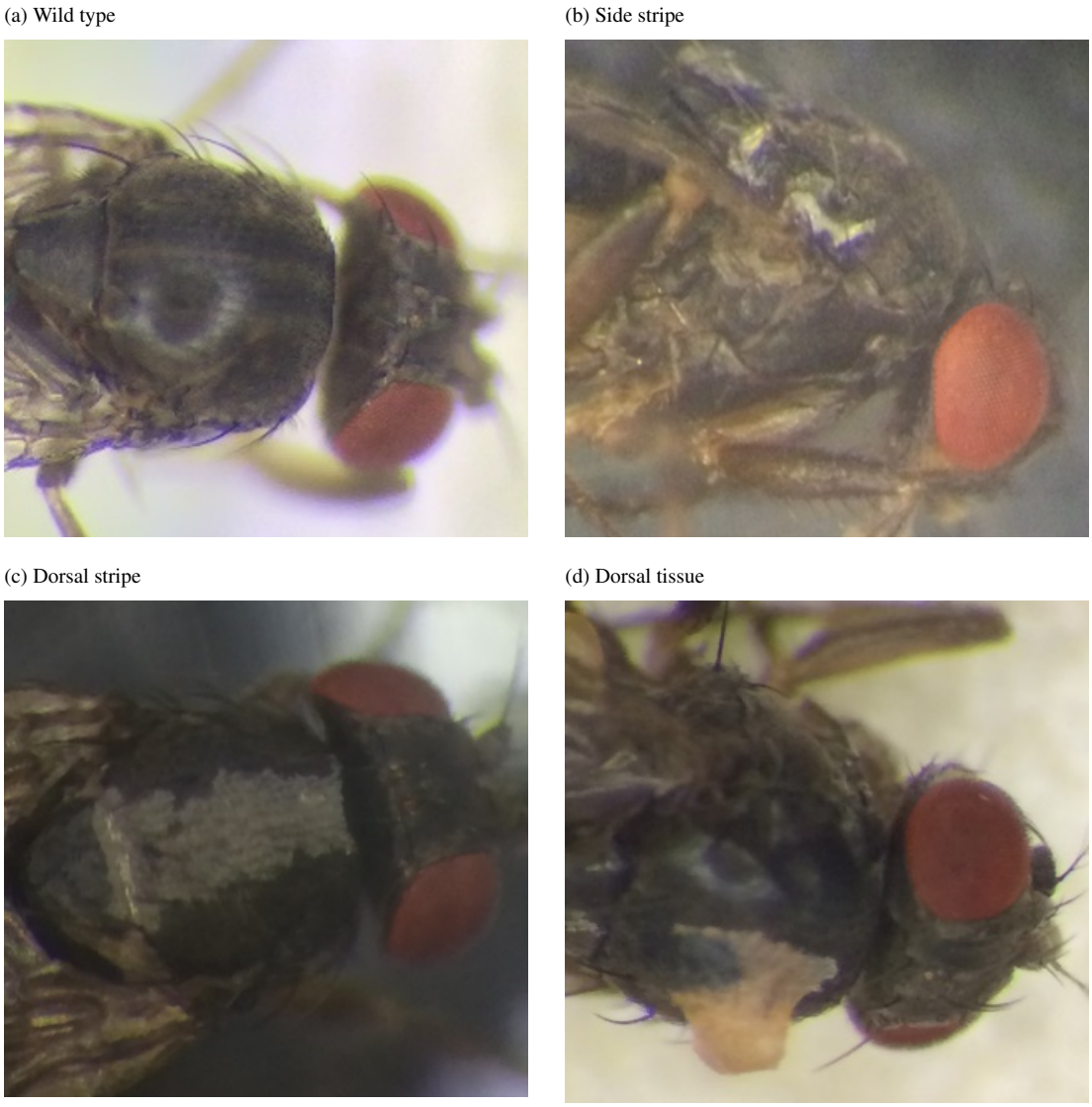


Figure 5.3.: Thoracic markings of potential transformants.
The side stripe, dorsal stripe and dorsal tissue were all fluorescent, and were not seen in a subsample of uninjected genome strain individuals.

genome strain individuals. We found no fluorescence in these individuals, indicating that the fluorescence seen in the injected F1s was unlikely to be a by-product of autofluorescence from another source

(e.g. food or excretory matter). To estimate the reliability of this visual fluorescence marker, we amplified *RFP* from a random subset of these fluorescent F1 individuals (after they had produced offspring), and found that 42% (19/45) of these individuals appear to carry an *RFP* gene of the expected length (~650bp). Weak amplification of this product prevented us from confirming its identity by sequencing; we therefore assessed the frequency of false-positives by carrying out the *RFP* PCR on a sample of 18 uninjected genome strain individuals, and found that a high proportion (14/18) of these individuals also produce a PCR product from the *RFP* primer pair. Given this high proportion of false positives by PCR, and the ambiguous nature of the visual fluorescence marker, we discounted visual fluorescence and amplification of *RFP* as indicators of potential transformants in subsequent generations.

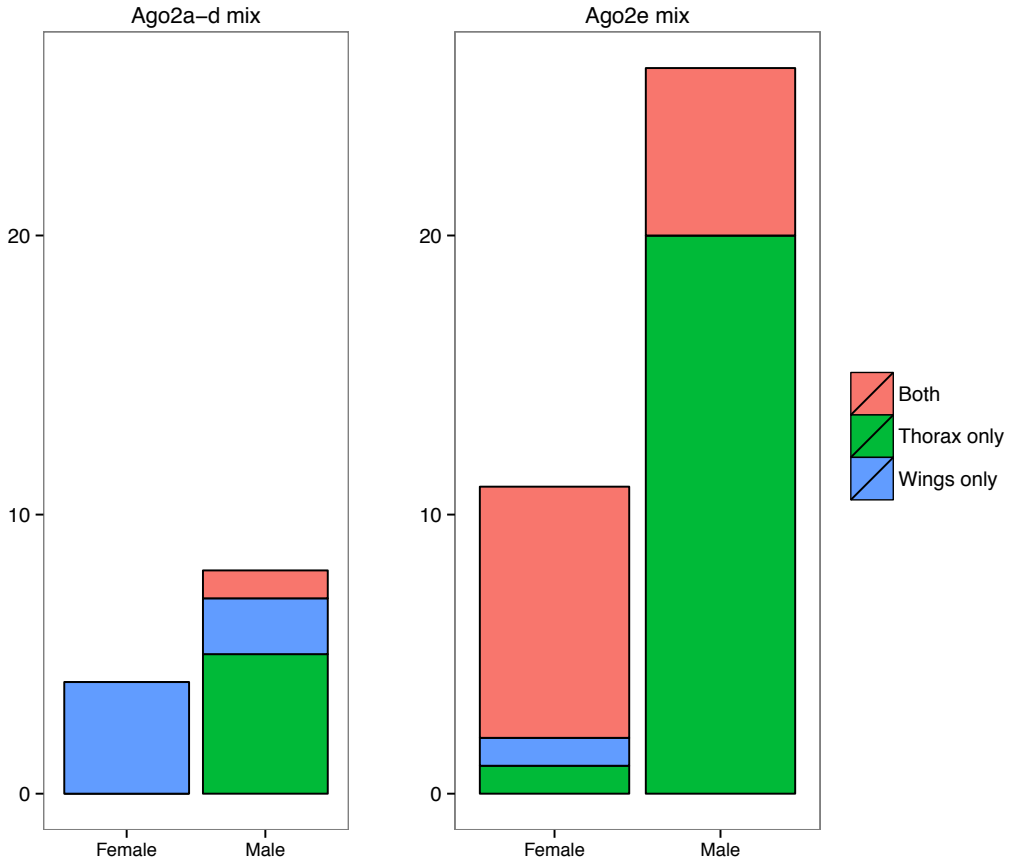


Figure 5.4.: The number of fluorescent individuals in the F1 generation.
More fluorescent F1 individuals were produced by individuals injected with the *Ago2e* mix (p=0.016). There were also differences in the location of fluorescence for the two mixes (p<0.0001), with more individuals with fluorescent wings from the *Ago2a1-d* mix, and more individuals with fluorescent thoraxes and wings from the *Ago2e* mix.

5.4.2. Frequency and reliability of the *RFP-Ago2* PCR marker in the F2 generation

Using the *Ago2-RFP* bridging PCR, the amplification of which indicates insertion of *RFP* into any of the *Ago2* paralogues in *D. pseudoobscura*, we detected 5 transformant females and 1 transformant male in the F1 generation, and 5 transformant females and 8 transformant males in the F2 generation. In terms of the overall numbers that were sampled in each generation, this equates to a female bias in both generations, which is more pronounced in the F1 generation (Figure 5.6). However, given the apparent segregation of *RFP* priming sites in the population that our injected individuals came from (Section 5.4.1), we checked whether this primer pair also amplified a product in a pooled sample of uninjected genome strain individuals. We find that this sample also produced a PCR product from the *RFP* bridging primer pair, indicating that the segregating *RFP* mispriming sites were potentially interfering with this marker. We therefore discounted the *RFP-Ago2* bridging PCR, and used shorter products in the *Ago2* PCRs to identify transformants in all generations.

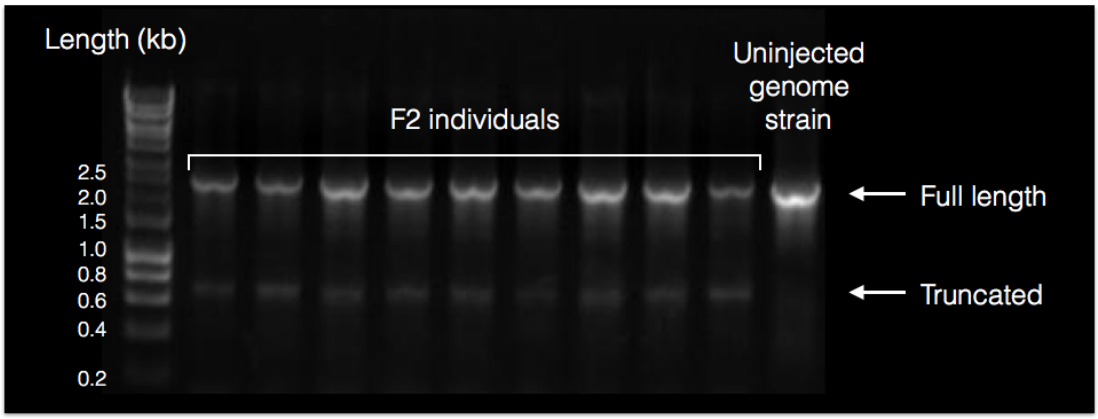


Figure 5.5.: A representative gel showing the truncated *Ago2* PCR products.
Each F2 individual shown has a truncated and a full length *Ago2* PCR product, in comparison to the uninjected genome strain, where only a single full length PCR product is observed.

5.4.3. Frequency and segregation patterns of putative truncated *Ago2* paralogues

Based on short products in the *Ago2* PCR (which may indicate truncated *Ago2* loci, illustrated in Figure 5.5), we detected the following potential transformants: 2 females and 1 male in the F1 generation; 5 females and 7 males in the F2 generation; and 11 females and 1 male in the F3 generation. Scaling these

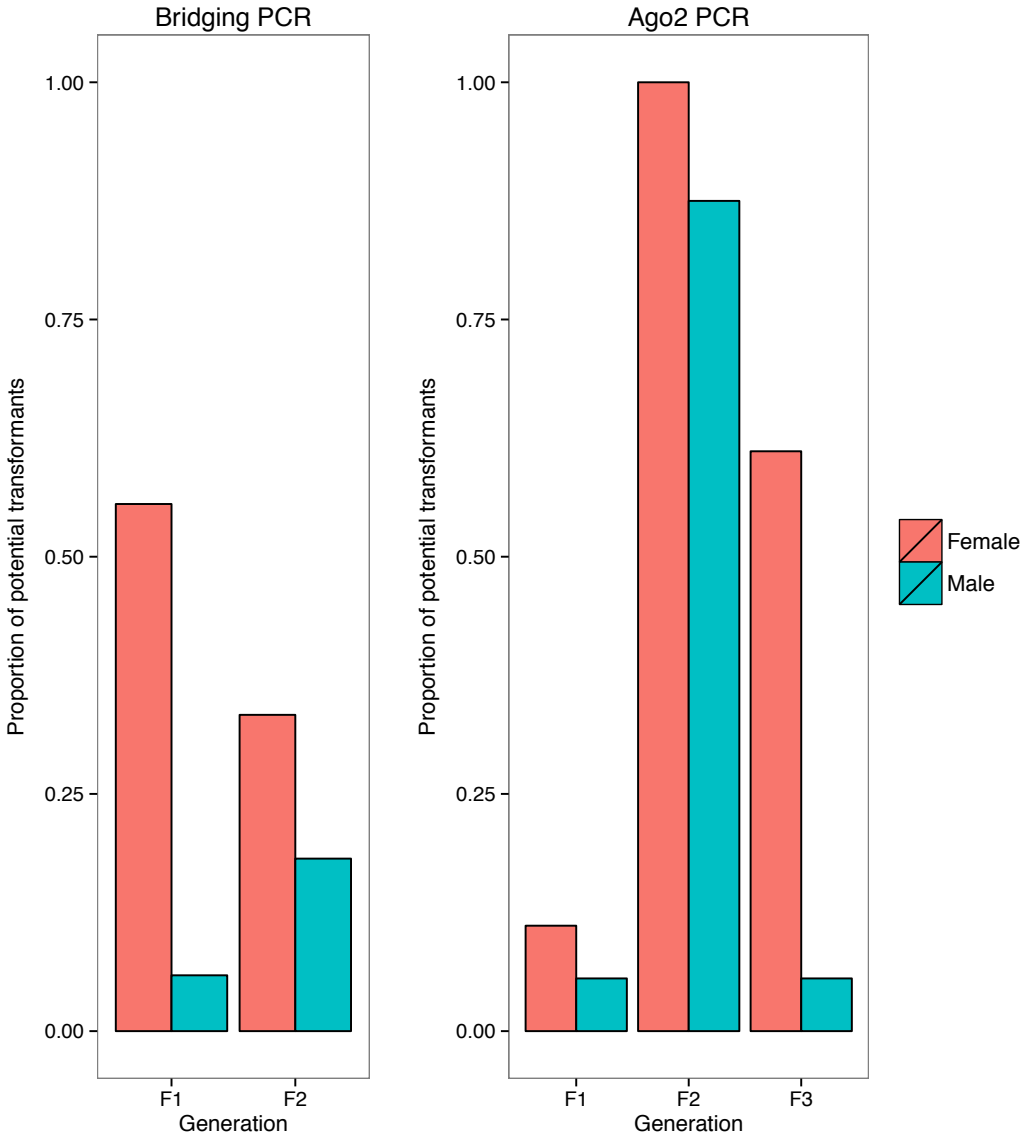


Figure 5.6.: The proportion of potential transformants by the *RFP-Ago2* bridging PCR and the *Ago2* length PCR.
By both assays, a higher proportion of potential transformants were female in all generations (Fisher's Exact Test, p>0.05). Caution should be exercised when comparing proportions between generations and assays, given the different sampling strategies used.

numbers to the total number of individuals sampled reveals dramatic variation between generations in the proportion of potential transformants; however, the different sampling strategies used in each generation make such comparisons difficult to interpret. Valid comparisons can, however, be made between the proportion of transformant females and males within each generation. This reveals a female-biased sex ratio in each generation, which is most dramatic in the F3 generation (Figure 5.6), but non-significant in all generations (Fisher's Exact Test, p>0.05).

If these short *Ago2* PCR products represent genuine *Ago2* truncations induced by CRISPR, and if they do not induce sterility or death, we would expect them to segregate in a Mendelian fashion. Our sampling strategy precludes this analysis for the F1 generation; however, we can estimate the transmission of short *Ago2* products from the F2 to F3 generations. We therefore used the *Ago2* PCRs to screen the F3 offspring of four F2 pairs, all of which were composed of one WT individual and one potential transformant individual which displayed short products in the *Ago2a* and *Ago2b* PCRs. For two of the F2 pairs sampled, we detected no F3 offspring with short products in any *Ago2* PCR. For the other two F2 pairs, however, several F3 individuals displayed short products in both the *Ago2d* and *Ago2e* PCRs (Table 5.2). Notably, in both of the F2 pairs that produced F3 individuals with short products, it was the F2 male that displayed short products. Contrastingly, in the two F2 pairs that produced no F3 individuals with short products, it was the F2 female that displayed short products. This may suggest that *Ago2* paralogue deletion has different effects in males and females, possibly linked to gamete formation or viability, although without confirming the identity of these short products this remains speculative.

Unfortunately, all of the F3 individuals in which we identified short *Ago2* products were dead on collection. Sanger sequencing of the single short product displayed by 11 of these individuals produced no recoverable sequence. We therefore ended the experiment at the F3 generation, due to the lack of F4 offspring from any potential transformants.

Table 5.2.: Frequency of the female F3 offspring of potential F2 transformants producing short products in *Ago2* PCR

F2 pair	***Ago2a***	***Ago2b***	***Ago2c***	***Ago2d***	***Ago2e***
1	0/6	0/6	0/6	0/6	0/6
2	0/15	0/15	0/15	6/15	7/15
3	0/4	0/4	0/4	0/4	0/4
4	0/14	1/14	0/14	2/14	3/14

5.5. Discussion

Transgenic technologies have enabled the testing of functional hypotheses, and recently the CRISPR/Cas9 technique has emerged as a powerful tool for use in non-model species. Here we aimed to use this technique to characterize the functions of *Ago2* paralogues in *D. pseudoobscura*, in order to test the existence and degree of functional divergence and redundancy between these paralogues; however, we failed to generate knockout individuals. Our lack of success was most probably caused by the technique working at a very low efficiency (or failing entirely), but might also indicate that *Ago2* paralogue knockouts are infertile or lethal.

Our data provide conflicting indications regarding the efficacy of CRISPR/Cas9 knockout of the *Ago2* paralogues. We gained a high number of fluorescent individuals (Section 5.4.1), which were not present in the parental stock, suggesting that the *RFP* marker gene inserted into the genome with relatively high efficiency. However, the mispriming of the *RFP* PCR primers in the parental genome, combined with the weak amplification of this potential *RFP* product, prevented us from validating these visual markers. This meant that no reliable estimate of efficacy could be drawn from fluorescence, and forced us to rely solely on PCR-based screening. This identified a small number of individuals with (potentially) truncated *Ago2* loci in each generation, which were not found in the parental stock or our previous population genetic survey of *Ago2* paralogues (Chapter 4), and are therefore likely to indicate the successful cleavage of these loci by Cas9. However, we were prevented from confirming the identity of these short products by sequencing, due to the presence of multiple overlapping sequences (in dead homozygotes), or the presence of larger additional products (in heterozygotes). This means that we cannot rule out another source, such as bacterial contamination, which is common in laboratory populations of *D. melanogaster* (Staubach *et al.* 2013). This would be expected to increase after bacterial colonization of the cadavers of homozygous individuals, leading to preferential amplification of the bacterial product. The ambiguity of these markers therefore makes it difficult to draw firm conclusions regarding the efficacy of *Ago2* paralogue knockout by CRISPR/Cas9.

There are two main factors that may have contributed to this apparent lack of success. Firstly, we observed low (~10%) survival of embryos after the initial microinjection stage of the procedure. While a previous comparison of 12 *Drosophila* species found that *D. pseudoobscura* had one of the lowest survival rates after microinjection for another transgenic procedure (Holtzman *et al.* 2010), our survival rate is similar to that of *D. melanogaster* after microinjection with CRISPR/Cas9 components (3-10%,

Bassett *et al.* 2013), suggesting that *D. pseudoobscura* is not particularly sensitive to CRISPR/Cas9 microinjection. Secondly, previous studies have rarely targeted more than one locus with the same gRNA pair, in contrast to our targeting of five loci (*Ago2a1-Ago2d*) with one gRNA pair. However, a previous study which used 3 redundant gRNAs to target all 65 members of the *β-GalGT* gene family found that 63% of these loci had been successfully truncated (Peng *et al.* 2014), demonstrating the efficacy of this approach.

Despite this uncertainty regarding the success of the technique, two patterns in the marker loci suggest that our lack of knockouts may have been driven by a vital function carried out by *Ago2* paralogues (although these suggestions remain necessarily speculative). Firstly, F3 homozygotes for truncated *Ago2* loci were all dead: while the potential for bacterial contamination and the lack of sequence from these individuals means we cannot be certain which (if any) *Ago2* locus was truncated, this may suggest that one or more *Ago2* paralogues carry out a vital function. Secondly, the *Ago2* length assay indicated a female bias in transformation success (Figure 5.6). This could be caused by more successful transformation in females, possibly because each female has two copies of each X-linked *Ago2* locus while each male has only one. Alternatively, if the testis-specific duplicates have evolved a key role in gamete formation, any knockout male may have had reduced fertility and therefore left very few knockout offspring. This effect would be expected to be more pronounced in homozygous knockouts, which we would have expected to see only in the F4 generation, but may still have reduced fertility in heterozygous knockouts due to dosage effects. If this reduction in fertility was due to inviability of Y-bearing sperm (e.g. by a meiotic drive element that is normally suppressed by the knocked-out *Ago2* paralogue), this would increase the proportion of X-bearing sperm, leading to an excess of female transformants in the next generation (as we observe). However, the ambiguity of markers and lack of sequence data prevent us from moving beyond speculation regarding biological function, and mean that the most likely explanation for our failure to produce knockouts is failure of the technique.

Had we been successful in producing homozygous knockouts, we would have tested the function of the testis-specific *Ago2* paralogues in the suppression of meiotic drive, by measuring the effect of paralogue knockout on sex ratio, fecundity and fertility. Additionally, we would have quantified the effect of paralogue knockout on rates of transposition in the soma and germline, thereby exploring the hypothesis that the testis-specificity of some paralogues is driven by an anti-TE role. Finally, combinatorial knockouts would have allowed us to explore the level of functional divergence and redundancy between closely-related paralogues, and therefore infer the speed with which functional specialization evolved.

In conclusion, we failed to generate viable *Ago2* paralogue knockout lines. Homozygous knockouts would have enabled us to test the function of *Ago2* paralogues, and infer the speed and extent of functional divergence. However, the success or otherwise of the CRISPR/Cas9 technique remains uncertain, given ambiguity around marker phenotypes and molecular assays.

6. General Discussion

6.1. Summary of the field

Argonaute genes are found in both prokaryotes and eukaryotes (Swarts *et al.* 2014b) and have been conserved across the eukaryotic tree of life (Cerutti and Casas-Mollano 2006). Eukaryotic Argonaute proteins are effectors in the RNAi mechanism, where they manipulate nucleic acid targets in partnership with small RNAs (Ding 2010). Through this mechanism, Argonautes carry out a diverse range of functions, including gene regulation, antiviral defence, dosage compensation and the suppression of TEs (Meister 2013). This functional diversity is associated with dynamic selection pressures, which variously constrain the rate of evolution (*D. melanogaster Ago1*; Obbard *et al.* 2009b, Chapter 2) or drive rapid evolution, possibly indicative of an arms race with viruses (*D. melanogaster Ago2*; Obbard *et al.* 2006; Obbard *et al.* 2011; Chapters 2 & 4).

Argonautes have duplicated frequently throughout their evolution, with ancient and recent duplications leading to wide variation in the size of the *Argonaute* gene family in different taxa (Cerutti and Casas-Mollano 2006; Mukherjee *et al.* 2013; Swarts *et al.* 2014b). Additionally, some *Argonautes* have undergone rapid lineage-specific expansions, leading to the evolution of highly derived functions (Morazzani *et al.* 2012; Schnettler *et al.* 2013b; Schnettler *et al.* 2014; Leebonoi *et al.* 2015). This suggests that duplication has played an important role in the evolution of *Argonautes*, and the RNAi pathways in general. However, previous work has largely focused on easily-amenable model organisms, leaving large gaps in our knowledge of *Argonaute* evolution.

6.2. Summary of findings

6.2.1. Frequent duplication of Dipteran *Argonautes* drives functional divergence

Gene duplication has exerted a major influence on the evolution of the *Argonaute* genes (Mukherjee *et al.* 2013; Swarts *et al.* 2014b), permitting evolutionary expansions which have often led to functional divergence; however, a comprehensive large-scale analysis of *Argonaute* duplication is yet to be undertaken. To measure the turnover rate of Dipteran *Argonaute* genes and the effect this has on their evolution, I carried out an *in silico* search for *Argonaute* homologues in 86 Dipteran species (Chapter 2). For each *Argonaute* subclade, I quantified the rate of gene turnover, the rate of protein evolution before and after duplication, and mapped rapidly evolving sites onto each Argonaute protein structure.

I found that gene turnover rate varies widely between different *Argonautes* and in different lineages. Additionally, I found an increase in evolutionary rate after duplication of *Ago2*, *Ago3* and *Piwi/Aub*, which is short-lived in *Piwi/Aub* but prolonged in *Ago2* and *Ago3*. Finally, I identified a cluster of rapidly-evolving residues at the mouth of the RNA-binding pocket of Ago2, suggesting that Ago2 paralogues could be binding different sRNA guides, and therefore cleaving different targets.

Combined with the characterisation of Argonautes with novel functions in *G. morsitans* (Chapter 2) and *A. aegypti* (Vodovar *et al.* 2012; Morazzani *et al.* 2012; Schnettler *et al.* 2013b; Miesen *et al.* 2015), these results suggest that the canonical view of Argonaute function derived from molecular studies of *D. melanogaster* is unlikely to apply across the Diptera and the insects as a whole. This work also advances our understanding of *Argonaute* gene turnover, which was previously based on isolated reports from individual organisms or sparsely-sampled gene trees; my densely-sampled order-wide analysis reveals that *Argonaute* genes can be gained and lost frequently and rapidly, potentially resulting in functional differences in the RNAi pathways of closely related species.

6.2.2. *Ago2* paralogues in the *obscura* group have repeatedly specialized to a novel, testis-specific function

Duplication of *Ago2* has previously been noted in *D. pseudoobscura*; combined with my finding of increased evolutionary rate after *Argonaute* duplication (Chapter 2), this suggests that these *Ago2* paralogues may have functionally diversified in *D. pseudoobscura*. To validate the existence of these paralogues, and quantify their age and distribution in related species, I identified *Ago2* homologues in other *obscura* group species, and carried out phylogenetic analysis to characterize their relationships. To test for functional divergence after duplication, I quantified the tissue-specific and antiviral expression patterns of *Ago2* paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*.

I found that *Ago2* has duplicated frequently throughout the evolution of the *obscura* group, producing between two and six *Ago2* paralogues in different species (Chapter 3). Additionally, I found that the *Ago2* paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura* have diverged functionally, with only one paralogue in each species retaining the ancestral ubiquitous expression pattern. Almost all of the other paralogues have specialized to a testis-specific role, as they are expressed only in the testis and are not induced by viral challenge. I attempted to characterize this function by CRISPR/Cas9-mediated gene knockout, but the resulting data were too ambiguous to draw any strong conclusions (Chapter 5). In spite of this, phylogenetic analysis revealed that testis-specificity has evolved repeatedly and been retained over long time periods (Chapter 3), suggesting that it has an adaptive basis.

This work documents the repeated emergence of a derived testis-specific expression pattern for *Ago2*, which has previously been studied overwhelmingly in *D. melanogaster* (but see Leebonoi *et al.* 2015), and which has therefore been assumed to have a ubiquitous function in invertebrates. These results also provide valuable data on the speed with which new paralogues functionally diverge, which has previously been understudied and poorly understood, and which I show can occur rapidly after duplication.

6.2.3. Testis-specific *Ago2* paralogues in the *obscura* group are under strong selection

When present as a single copy in members of the *melanogaster* group of *Drosophila*, *Ago2* evolves under strong positive selection (Obbard *et al.* 2006; Obbard *et al.* 2009a; Obbard *et al.* 2011). Ad-

ditionally, recent *Ago2* paralogues in the *obscura* group have evolved a novel testis-specific function (Chapter 3), which may be driven by positive selection. To quantify the rate of evolution and level of positive selection acting on each *Ago2* paralogue in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, I gathered intraspecies polymorphism data for each paralogue, analysed publicly available population genomic data for evidence of selective sweeps on the *Ago2* paralogues in *D. pseudoobscura*, and explicitly tested all *Ago2* paralogues for an increase in evolutionary rate after the evolution of testis-specificity.

I found that the vast majority of *Ago2* paralogues are evolving under strong selection, evidenced by positive selection on *D. pseudoobscura Ago2e*, signatures of selective sweeps at *D. pseudoobscura Ago2a*, *Ago2b* & *Ago2c*, and significantly lower diversity at most paralogues compared with the genome as a whole. In addition to these signatures of positive selection on many testis-specific paralogues (e.g. *D. pseudoobscura Ago2a*, *Ago2b* and *Ago2e*), I found that testis-specific paralogues across the entire *obscura* group generally evolve more rapidly than ubiquitously expressed paralogues. These results suggest an adaptive basis for the derived testis specificity that has repeatedly evolved after duplication of *Ago2*, and support my finding that *Argonaute* genes across the Diptera evolve more rapidly after duplication due to selection (Chapter 2)

These results provide further evidence for the importance of duplication in permitting selectively advantageous evolutionary innovation in the RNAi pathway, and suggest that the evolution of derived functions under strong positive selection is a common fate for young *Argonaute* paralogues. This work also provides a rare insight into the relative importance of lack of constraint and positive selection in the early stages of paralogue evolution, and suggest that positive selection can act soon after duplication, potentially driving the evolution of novel functions.

6.3. Implications and future directions

6.3.1. The role of gene duplication in the functional diversity of *Argonaute* genes

The functions and mechanisms of *Argonaute* genes in large phylogenetic groups are mainly (and necessarily) based on results from a few model species, such as *D. melanogaster* (arthropods), *M. musculus* & *H. sapiens* (vertebrates), and *A. thaliana* (plants). However, expansions or losses of *Argonaute* genes have been reported for other species in these groups, as seen in centipedes (Chipman *et al.* 2014) (arthropods) and rice (Cerutti and Casas-Mollano 2006) (plants). My analysis of Dipteran *Argonaute* genes suggests that these expansions are not rare, but instead happen frequently in different lineages (Chapter 2). Moreover, many studies have uncovered functional divergence after *Argonaute* duplication, with derived functions found in tsetse flies (Chapter 2), tiger shrimp (Leebonoi *et al.* 2015), nematodes (reviewed in Buck and Blaxter 2013), planarians (Palakodeti *et al.* 2008) and rice (Wu *et al.* 2015). This indicates that frequent change in copy number and function is a common pattern during *Argonaute* evolution, and suggests that generalizations based on *Argonaute* function in model species will have limited relevance over large phylogenetic distances, especially for taxa in which *Argonaute* has duplicated.

A prevalent driver of *Argonaute* divergence appears to be the separation of the soma and germline. Specialization to the germline occurred early in the evolution of the Metazoan Piwi subfamily (reviewed in Ross *et al.* 2014), and much more recently after duplication of *Ago2* in the tiger shrimp *Penaeus monodon* (Leebonoi *et al.* 2015) and the *obscura* subgroup (Chapter 3). Additionally, some Piwi subfamily paralogues have lost this specialization and evolved (or reverted to) somatic expression, as seen in *A. aegypti* (Vodovar *et al.* 2012; Morazzani *et al.* 2012; Schnettler *et al.* 2013b; Miesen *et al.* 2015) and *G. morsitans* (Chapter 2). It has even been suggested that somatic roles for Piwi subfamily genes are so prevalent that the canonical view of germline-specificity for these genes should be revised, or at least limited to a subset of taxa (Ross *et al.* 2014).

The role of the germline-soma divide in *Argonaute* evolution could be underpinned by the role for Argonaute proteins in the suppression of exogenous (viruses) and endogenous (TEs) parasites, which was apparently present in the prokaryotic ancestral Argonaute (Swarts *et al.* 2014a; Swarts *et al.* 2015) and has since been retained in the majority of eukaryotic taxa (Chapter 1). While they are both sup-

pressed by Argonautes, TEs and viruses show contrasting activity levels in the soma and germline: TEs are expected to be active primarily in the germline (Charlesworth and Langley 1989), whereas viruses replicate in the germline and soma, and in some cases preferentially in the soma (e.g. arboviruses in salivary glands; Lambrechts and Scott 2009). Additionally, many viruses encode viral suppressor of RNAi (VSRs) to inhibit Argonaute proteins (reviewed in Bronkhorst and Rij 2014), whereas suppressors of RNAi encoded by TEs have been reported only rarely (but see Nosaka *et al.* 2012). Viruses and TEs therefore impose contrasting selection pressures on the Argonautes suppressing them, even though in some cases they are both suppressed by the same Argonaute (e.g. *D. melanogaster* Ago2). This introduces adaptive conflict that can be alleviated by duplication, resulting in functional divergence between the paralogues as they specialize to either an ubiquitous (antiviral) or germline-specific (anti-TE) role.

To explore the role of the germline-soma divide in *Argonaute* evolution, further investigation into two aspects of the germline and somatic expression patterns of *Argonaute* genes would be highly valuable. Firstly, to compare the prevalence of germline specificity across the *Argonautes*, a fruitful approach may be further comparison of expression patterns of Ago and Piwi subfamily genes in the germline and soma, and accompanying analysis of the prevalence of siRNAs and piRNAs in these tissues. The only explicit comparison of these patterns to date unexpectedly discovered piRNA-like sRNAs and the expression of Piwi subfamily genes in the soma, despite a limited sample of three species (Yan *et al.* 2011). The evolution of both somatic (Vodovar *et al.* 2012; Morazzani *et al.* 2012; Schnettler *et al.* 2013b) and germline-specific (Cox *et al.* 1998; Vagin *et al.* 2004; Kalmykova *et al.* 2005) functions for Piwi subfamily genes could be caused by the evolution of novel functions after the relaxation of selection pressures by duplication (Ohno 1970). Alternatively, it could be the result of inherent functional promiscuity of Piwi, as suggested by the piRNAs in *D. melanogaster*, which were initially thought to be produced exclusively from TEs (Aravin *et al.* 2001; Brennecke *et al.* 2007), but have since been found to originate from host gene 3' UTRs (Robine *et al.* 2009) and viruses in ovarian stem cells (Wu *et al.* 2010). To discern between these possibilities, it would be especially informative to compare the expression patterns and sRNA binding partners of Piwi subfamily genes in species which have a single copy and multiple copies of *Piwi*. Secondly, to assess the relevance of the paradigm of germline-specificity for Piwi subfamily genes, and potentially to date their specialization to the germline, somatic and germline *Piwi* expression patterns could be compared in a phylogenetically diverse sample of species, with a particular focus on basal Metazoa.

6.3.2. The evolution of RNAi

The range of functions that *Argonaute* genes have adopted after duplication (Chapters 2 & 3, Swarts *et al.* 2014a) have not evolved in isolation, but rather as part of the entire machinery of the RNAi pathways. There are numerous reasons to expect evolutionary change in *Argonaute* genes to co-occur with change in the other RNAi genes. Firstly, these genes may be under similar selection pressures as *Argonaute* genes and may have undergone similar diversification, as demonstrated by the *Dicer* genes of *A. thaliana*, which have duplicated and specialized to provide immune defence against distinct viral types (Blevins *et al.* 2006). Secondly, evolutionary change in *Argonaute* genes may impose an additional selection pressure on these genes to maintain optimum chemical and physical interactions with the Argonaute protein, or alternatively, the evolutionary change in *Argonaute* may be a response to a preceding change in another gene. Lastly, in many pathways duplication of one member triggers duplication of other pathway members, as seen in the co-duplication of the *insulin-NGF* gene family and their receptors (Fryxell 1996). This pattern is evident in two sRNA-directed mechanisms. Firstly, the gene regulatory and antiviral RNAi pathways of *D. melanogaster*, each of which has a distinct RNA sensor (Dicer1 & Dicer2 respectively), sRNA loading protein (R3D1 & R2D2 respectively) and effector (Ago1 & Ago2 respectively) (Jiang *et al.* 2005; Ding 2010). Secondly, the RNA Polymerases involved in the plant RNA-directed DNA methylation (RdDM) pathway (reviewed in Matzke and Mosher 2014), which are made up of numerous subunits that arose from duplications early in plant evolution, at approximately the same time as duplications that produced their interacting cofactors DCL3 and AGO4 (Huang *et al.* 2015).

These factors suggest that the functional divergence of *Argonaute* genes may serve as an indicator of evolutionary change in other RNAi genes. To test the possibility that other genes co-duplicate with *Argonaute*, the copy number for genes that directly interact with Argonautes (e.g. Dicer & R3D1/R2D2) could be quantified, with a focus on species in which *Argonautes* have duplicated (e.g. *T. whitei*, *S. deflexa* and other species identified in Chapter 2). Such an approach has been successfully applied to six gene families involved in chemosensing and digestion in *Drosophila*, and revealed parallel expansions that correlate with the evolution of novel diets (Wu *et al.* 2011). Where tissue-specific transcriptome data is available, it would be particularly informative to explore patterns of differential expression in any identified paralogues, which would be expected to mirror the *Argonaute* paralogues with which these genes are interacting.

Another emerging theme in RNAi evolution is the overlap between what were previously considered as functionally discrete RNAi pathways. This is demonstrated by the interaction between AGO1, AGO18 and miRNAs in *O. sativa*: by sequestering the microRNA miR168, AGO18 prevents the downregulation of the key antiviral Argonaute AGO1, thus upregulating the antiviral defence mechanism (Wu *et al.* 2015). A second example is provided by the *D. melanogaster* miRNA pathway, which is canonically involved in the regulation of host genes, but also contributes to somatic TE suppression by an uncharacterized mechanism (Mugat *et al.* 2015). Further evidence is provided by the binding of both siRNAs and some miRNAs by *D. melanogaster* Ago2 (Tomari *et al.* 2007), despite its canonical role in siRNA-mediated silencing (Ding 2010). Finally, the distinction between the siRNA and piRNA pathways is being increasingly blurred, due to the discovery of mixed siRNA-piRNA "Ping-Pong" pairs derived from TEs (Shpiz *et al.* 2014), and virally-derived piRNAs in the soma of *D. melanogaster* (Wu *et al.* 2010) and *A. aegypti* (Vodovar *et al.* 2012; Morazzani *et al.* 2012; Schnettler *et al.* 2013b; Miesen *et al.* 2015). These discoveries suggest that there is a large degree of interplay between the RNAi pathways (discussed in Fablet 2014), and highlight the importance of investigating the impact of novel sRNA classes or RNAi genes on all RNAi pathways, rather than focusing narrowly on canonical sRNAs, binding partners and target loci.

6.3.3. The evolution of gene families

My results can also be viewed more widely in the context of gene duplication, which has produced a high degree of functional diversity in many gene families. The testis has previously been implicated in the evolution of this diversity through the derived, testis-specific expression of young paralogues (Kaessmann 2010; Assis and Bachtrog 2013) and orphan genes (Palmieri *et al.* 2014; Zhao *et al.* 2014). Some previous studies have attributed a non-adaptive basis to this role (Kaessmann *et al.* 2009; Kaessmann 2010) basis, while others have inferred that testis-specificity is driven by selection (Zhao *et al.* 2014) . My finding that positive selection is acting on many testis-specific *Ago2* paralogues (Chapters 3 &4) builds the case for the testis as a crucible of novel gene function, and raises the possibility that testis-specificity observed in other young paralogues also has an adaptive basis.

Another key influence on the evolution of gene families is parasite-mediated selection, which has been linked to the expansion of *prophenoloxidase* genes in *Ae. aegytpi* and *An. gambiae* (Waterhouse *et al.* 2007), *Toll-like Receptors* in mammals (Leulier and Lemaitre 2008) and *DSCAM* in arthropods

(Palmer and Jiggins 2015). Parasites also appear to have exerted strong selection on their hosts to retain Argonaute paralogues, as evidenced by the expansions of *Ago2* (putatively antiviral) and *Piwi/Aub* (putatively anti-TE) in the Diptera (Chapter 2). Where functional change has occurred in these clades, it has also been linked to adaptation against parasites. Selection imposed by exogenous parasites appears to have driven the expansion of antiviral *Piwi* genes of mosquitoes (Vodovar *et al.* 2012; Morazzani *et al.* 2012; Schnettler *et al.* 2013b), and may be involved in the upregulation of *Ago3b* in *G. morsitans* salivary glands infected with *T. brucei* (Chapter 2). Similarly, endogenous parasites such as TEs may be involved in the retention of *Ago2* duplicates in the *obscura* group, and their subsequent specialization to the testis (Chapters 3 & 4). This suggests that gene duplication is a prevalent influence in the evolutionary arms race between host and parasite, and may provide a mechanism by which a host can balance the conflicting selection pressures imposed by a wide variety of pathogens.

6.4. Conclusions

I have found that *Argonaute* genes duplicate frequently in the Diptera, and evolve more rapidly after duplication, indicative of frequent functional divergence. This process is exemplified by the *Ago2* paralogues in the *obscura* group, the majority of which have specialized to a highly-derived testis-specific function. This function has evolved numerous times independently, and is driven by strong positive selection, possibly imposed by a novel role in the suppression of TEs or meiotic drive. I conclude that gene duplication is likely to have driven the emergence of novel and adaptive functions throughout *Argonaute* evolution.

Bibliography

K. A. Al-Mukhtar and A. C. Webb, "An ultrastructural study of primordial germ cells, oogonia and early oocytes in xenopus laevis.," *Journal of Embryology and Experimental Morphology*, vol. 26, no. 2, pp. 195–217, 1971.

R. Aliyari, Q. Wu, H.-W. Li, X.-H. Wang, F. Li, L. D. Green, C. S. Han, W.-X. Li, and S.-W. Ding, "Mechanism of induction and suppression of antiviral immunity directed by virus-derived small rnas in drosophila.," *Cell Host & Microbe*, vol. 4, no. 4, pp. 387–97, 2008.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

S. Anders, P. T. Pyl, and W. Huber, "Htseq - a python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 2, pp. 166–169, 2015.

A. A. Aravin, N. M. Naumova, A. V. Tulin, V. V. Vagin, Y. M. Rozovsky, and V. A. Gvozdev, "Double-stranded rna-mediated silencing of genomic tandem repeats and transposable elements in the d. melanogaster germline.," *Current Biology*, vol. 11, no. 13, pp. 1017–27, 2001.

L. Aravind, "Guilt by association: contextual information in genome analysis," *Genome Research*, vol. 10, no. 8, pp. 1074–1077, 2000.

R. Assis and D. Bachtrog, "Neofunctionalization of young duplicate genes in drosophila.," *Proceedings of the National Academy of Sciences*, vol. 110, no. 43, pp. 17 409–14, 2013.

M. J. Axtell, J. O. Westholm, and E. C. Lai, "Vive la différence: biogenesis and evolution of microrna s in plants and animals," *Genome Biology*, vol. 12, no. 4, p. 221, 2011.

R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath, "Crispr provides acquired resistance against viruses in prokaryotes.," *Science*, vol. 315, no. 5819, pp. 1709–1712, 2007.

A. R. Bassett, C. Tibbit, C. P. Ponting, and J.-L. Liu, "Highly efficient targeted mutagenesis of drosophila with the crispr/cas9 system.," *Cell Reports*, vol. 4, no. 1, pp. 220–8, 2013.

D. J. Begun, H. a. Lindfors, A. D. Kern, and C. D. Jones, "Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade.," *Genetics*, vol. 176, no. 2, pp. 1131–7, 2007.

S. K. Behura and D. W. Severson, "Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes," *Biological Reviews*, vol. 88, no. 1, pp. 49–61, 2013.

J. B. Benoit, G. M. Attardo, V. Michalkova, T. B. Krause, J. Bohova, Q. Zhang, A. A. Baumann, P. O. Mireji, P. Takáč, D. L. Denlinger, J. M. Ribeiro, and S. Aksoy, "A novel highly divergent protein family identified from a viviparous insect by rna-seq analysis: a potential target for tsetse fly-specific abortifacients," *PLoS Genetics*, vol. 10, no. 4, pp. 6–10, 2014.

A. J. Berghammer, M. Klingler, E. A. Wimmer, P. Sutovsky, R. D. Moreno, C. Simerly, and G. Schatten, "A universal marker for transgenic insects," *Nature*, vol. 402, no. November, pp. 1996–1997, 1999.

S. A. Bernhardt, M. P. Simmons, K. E. Olson, B. J. Beaty, C. D. Blair, and W. C. Black, "Rapid intraspecific evolution of mirna and sirna genes in the mosquito aedes aegypti.," *PLoS One*, vol. 7, no. 9, e44198, 2012.

E. Betrán, K. Thornton, M. Long, M. D. Vibranovski, and Y. Zhang, "Retroposed new genes out of the x in drosophila," *Genome Research*, vol. 12, pp. 1854–1859, 2002.

M. Bibikova, M. Golic, K. G. Golic, and D. Carroll, "Targeted chromosomal cleavage and mutagenesis in drosophila using zinc-finger nucleases," *Genetics*, vol. 161, no. 3, pp. 1169–1175, 2002.

C Biémont and G Cizeron, "Distribution of transposable elements in drosophila species.," *Genetica*, vol. 105, no. 1, pp. 43–62, 1999.

R. B. Billmyre, S. Calo, M. Feretzaki, X. Wang, and J. Heitman, "Rnai function, diversity, and loss in the fungal kingdom," *Chromosome Research*, vol. 21, pp. 561–572, 2013.

J. Bischof, R. K. Maeda, M. Hediger, F. Karch, and K. Basler, "An optimized transgenesis system for drosophila using germ-line-specific phic31 integrases.," *Proceedings of the National Academy of Sciences*, vol. 104, no. 9, pp. 3312–7, 2007.

T. Blevins, R. Rajeswaran, P. V. Shivaprasad, D. Beknazariants, A. Si-Ammour, H. S. Park, F. Vazquez, D. Robertson, F. Meins, T. Hohn, and M. M. Pooggin, "Four plant dicers mediate viral small rna biogenesis and dna virus induced silencing," *Nucleic Acids Research*, vol. 34, no. 21, pp. 6233–6246, 2006.

K. Bohmert, I. Camus, C. Bellini, D. Bouchez, M. Caboche, and C. Benning, "Ago1 defines a novel locus of arabidopis controlling leaf development," *The EMBO Journal*, vol. 1, no. 1, pp. 170–180, 1998.

J. Brennecke, A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon, "Discrete small rna-generating loci as master regulators of transposon activity in drosophila.," *Cell*, vol. 128, no. 6, pp. 1089–103, 2007.

A. W. Bronkhorst and R. P. van Rij, "The long and short of antiviral defense: small rna-based immunity in insects.," *Current Opinion in Virology*, vol. 7C, pp. 19–28, 2014.

A. W. Bronkhorst, K. W. R. V. Cleef, N. Vodovar, Ä. A. Ářnce, H. Blanc, J. M. Vlak, M.-C. Saleh, and R. P. V. Rij, "The dna virus invertebrate iridescent virus 6 is a target of the drosophila rnai machinery," *Proceedings of the National Academy of Sciences*, vol. 109, no. 51, E3604–E3613, 2012.

J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki, K. H. Wan, C. Yu, D. Zhang, J. W. Carlson, L. Cherbas, B. D. Eads, D. Miller, K. Mockaitis, J. Roberts, C. A. Davis, E. Frise, A. S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A. A. Samsonova, R. Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P. J. Bickel, P. Carninci, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, E. C. Lai, B. Oliver, N. Perrimon, B. R. Graveley, and S. E. Celniker, "Diversity and dynamics of the drosophila transcriptome," *Nature*, pp. 1–7, 2014.

A. H. Buck and M. Blaxter, "Functional diversification of argonautes in nematodes: an expanding universe.," *Biochemical Society Transactions*, vol. 41, no. 4, pp. 881–6, 2013.

A. M. Burroughs, L. M. Iyer, and L Aravind, "Two novel piwi families: roles in inter-genomic conflicts in bacteria and mediator-dependent modulation of transcription in eukaryotes.," *Biology Direct*, vol. 8, no. 1, p. 13, 2013.

A. M. Burroughs, Y. Ando, M. J. L. de Hoon, Y. Tomaru, H. Suzuki, Y. Hayashizaki, and C. O. Daub, "Deep-sequencing of human argonaute-associated small rnas provides insight into mirna sorting and reveals argonaute association with rna fragments of diverse origin.," *RNA Biology*, vol. 8, no. 1, pp. 158–177, 2011.

C. L. Campbell, W. C. Black, A. M. Hess, and B. D. Foy, "Comparative genomics of small rna regulatory pathway components in vector mosquitoes.," *BMC Genomics*, vol. 9, p. 425, 2008.

C. Casola and M. W. Hahn, "Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods.," *Journal of Molecular Evolution*, vol. 68, no. 6, pp. 679–87, 2009.

D. M. Castillo, J. C. Mell, K. S. Box, and J. P. Blumenstiel, "Molecular evolution under increasing transposable element burden in drosophila: a speed limit on the evolutionary arms race.," *BMC Evolutionary Biology*, vol. 11, no. 1, p. 258, 2011.

J Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, pp. 540–552, 2000.

S. E. Celniker, L. A. L. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. Macalpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston, "Unlocking the secrets of the genome," *Nature*, vol. 459, no. June, pp. 927–930, 2009.

H. Cerutti and J. A. Casas-Mollano, "On the origin and functions of rna-mediated silencing: from protists to man.," *Current Genetics*, vol. 50, no. 2, pp. 81–99, 2006.

S.-S. Chang, Z. Zhang, and Y. Liu, "Rna interference pathways in fungi: mechanisms and functions," *Annual Review of Microbiology*, vol. 66, no. 1, pp. 305–323, 2012.

D Charif and J. Lobry, "Seqinr 1.0-2: a contributed package to the r project for statistical computing devoted to biological sequences retrieval and analysis," in *Structural approaches to sequence evolution: Molecules, networks, populations*, U Bastolla, M Porto, H. Roman, and M Vendruscolo, Eds., New York: Springer Verlag, New York, NY, 2007, pp. 207–232.

B Charlesworth and C. H. Langley, "The population genetics of drosophila transposable elements.," *Annual Review of Genetics*, vol. 23, pp. 251–87, 1989.

J. Charlesworth and A. Eyre-Walker, "The mcdonald-kreitman test and slightly deleterious mutations.," *Molecular Biology and Evolution*, vol. 25, no. 6, pp. 1007–15, 2008.

S. S. Chaudhari, B. Moussian, C. A. Specht, Y. Arakane, K. J. Kramer, R. W. Beeman, and S. Muthukrishnan, "Functional specialization among members of knickkopf family of proteins in insect cuticle organization.," *PLoS Genetics*, vol. 10, no. 8, e1004537, 2014.

A. D. Chipman, D. E. K. Ferrier, C. Brena, J. Qu, D. S. T. Hughes, R. Schröder, M. Torres-Oliva, N. Znassi, H. Jiang, F. C. Almeida, C. R. Alonso, Z. Apostolou, P. Aqrawi, W. Arthur, J. C. J. Barna, K. P. Blankenburg, D. Brites, S. Capella-Gutiérrez, M. Coyle, P. K. Dearden, L. Du Pasquier, E. J. Duncan, D. Ebert, C. Eibner, G. Erikson, P. D. Evans, C. G. Extavour, L. Francisco, T. Gabaldón, W. J. Gillis, E. a. Goodwin-Horn, J. E. Green, S. Griffiths-Jones, C. J. P. Grimmelikhuijzen, S. Gubbala, R. Guigó, Y. Han, F. Hauser, P. Havlak, L. Hayden, S. Helbing, M. Holder, J. H. L. Hui, J. P. Hunn, V. S. Hunnekuhl, L. Jackson, M. Javaid, S. N. Jhangiani, F. M. Jiggins, T. E. Jones, T. S. Kaiser, D. Kalra, N. J. Kenny, V. Korchina, C. L. Kovar, F. B. Kraus, F. Lapraz, S. L. Lee, J. Lv, C. Mandapat,

G. Manning, M. Mariotti, R. Mata, T. Mathew, T. Neumann, I. Newsham, D. N. Ngo, M. Ninova, G. Okwuonu, F. Ongeri, W. J. Palmer, S. Patil, P. Patraquim, C. Pham, L.-L. Pu, N. H. Putman, C. Rabouille, O. M. Ramos, A. C. Rhodes, H. E. Robertson, H. M. Robertson, M. Ronshaugen, J. Rozas, N. Saada, A. Sánchez-Gracia, S. E. Scherer, A. M. Schurko, K. W. Siggens, D. Simmons, A. Stief, E. Stolle, M. J. Telford, K. Tessmar-Raible, R. Thornton, M. van der Zee, A. von Haeseler, J. M. Williams, J. H. Willis, Y. Wu, X. Zou, D. Lawson, D. M. Muzny, K. C. Worley, R. A. Gibbs, M. Akam, and S. Richards, "The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede strigamia maritima," *PLoS Biology*, vol. 12, no. 11, e1002005, 2014.

W.-J. Chung, K. Okamura, R. Martin, and E. C. Lai, "Endogenous rna interference provides a somatic defense against drosophila transposons.," *Current Biology*, vol. 18, no. 11, pp. 795–802, 2008.

A. G. Clark, "Invasion and maintenance of a gene duplication.," *Proceedings of the National Academy of Sciences*, vol. 91, no. April, pp. 2950–2954, 1994.

L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, "Multiplex genome engineering using crispr/cas systems.," *Science*, vol. 339, no. 6121, pp. 819–23, 2013.

T. Conrad and A. Akhtar, "Dosage compensation in drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription," *Nature Reviews Genetics*, vol. 13, no. 2, pp. 123–134, 2012.

D. N. Cox, a. Chao, J. Baker, L. Chang, D. Qiao, and H. Lin, "A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal," *Genes & Development*, vol. 12, no. 23, pp. 3715–3727, 1998.

A. D. Cutter, "Divergence times in caenorhabditis and drosophila inferred from direct estimates of the neutral mutation rate," *Molecular Biology and Evolution*, vol. 25, no. 4, pp. 778–786, 2008.

B. Czech, C. D. Malone, R. Zhou, A. Stark, C. Schlingeheyde, M. Dus, N. Perrimon, M. Kellis, J. A. Wohlschlegel, R. Sachidanandam, G. J. Hannon, and J. Brennecke, "An endogenous small interfering rna pathway in drosophila.," *Nature*, vol. 453, no. 7196, pp. 798–802, 2008.

R Dawkins and J. R. Krebs, "Arms races between and within species.," *Proceedings of the Royal Society B*, vol. 205, no. 1161, pp. 489–511, 1979.

D. L. Des Marais and M. D. Rausher, "Escape from adaptive conflict after duplication in an anthocyanin pathway gene.," *Nature*, vol. 454, no. 7205, pp. 762–765, 2008.

S.-W. Ding, "Rna-based antiviral immunity.," *Nature Reviews Immunology*, vol. 10, no. 9, pp. 632–44, 2010.

C. Dong, J. Zhang, J. Qiao, and G. He, "Positive selection and functional divergence after melanopsin gene duplication.," *Biochemical Genetics*, vol. 50, no. 3-4, pp. 235–48, 2012.

E. B. Dopman and D. L. Hartl, "A portrait of copy-number polymorphism in drosophila melanogaster.," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19 920–19 925, 2007.

C. Dostert, E. Jouanguy, P. Irving, L. Troxler, D. Galiana-Arnoux, C. Hetru, J. A. Hoffmann, and J.-L. Imler, "The jak-stat signaling pathway is required but not sufficient for the antiviral response of drosophila.," *Nature Immunology*, vol. 6, no. 9, pp. 946–953, 2005.

I. A. Drinnenberg, D. E. Weinberg, K. T. Xie, J. P. Mower, K. H. Wolfe, G. R. Fink, and D. P. Bartel, "Rnai in budding yeast.," *Science*, vol. 326, no. 5952, pp. 544–50, 2009.

I. A. Drinnenberg, G. R. Fink, and D. P. Bartel, "Compatibility with killer explains the rise of rnai-deficient fungi.," *Science*, vol. 333, no. 6049, p. 1592, 2011.

A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian phylogenetics with beauti and the beast 1.7," *Molecular Biology and Evolution*, vol. 29, no. 8, pp. 1969–1973, 2012.

A. Dueck, C. Ziegler, A. Eichner, E. Berezikov, and G. Meister, "Micrornas associated with the different human argonaute proteins," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9850–9862, 2012.

N. A. Dyer, S. P. Lawton, S Ravel, K. S. Choi, M. J. Lehane, A. S. Robinson, L. M. Okedi, M. J. R. Hall, P Solano, and M. J. Donnelly, "Molecular phylogenetics of tsetse flies (diptera: glossinidae) based on mitochondrial (coi, 16s, nd2) and nuclear ribosomal dna sequences, with an emphasis on the palpalis group.," *Molecular Phylogenetics and Evolution*, vol. 49, no. 1, pp. 227–39, 2008.

T. H. Eickbush and D. G. Eickbush, "Finely orchestrated movements: evolution of the ribosomal rna genes," *Genetics*, vol. 175, no. February, pp. 477–485, 2007.

E. Elkayam, C. D. Kuhn, A. Tocilj, A. D. Haase, E. M. Greene, G. J. Hannon, and L. Joshua-Tor, "The structure of human argonaute-2 in complex with mir-20a," *Cell*, vol. 150, no. 1, pp. 100–110, 2012.

A. C. English, S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley, and R. A. Gibbs, "Mind the gap: upgrading genomes with pacific biosciences rs long-read sequencing technology," *PLoS One*, vol. 7, no. 11, pp. 1–12, 2012.

A. Eulalio, E. Huntzinger, and E. Izaurralde, "Getting to the root of mirna-mediated gene silencing," *Cell*, vol. 132, no. 2005, pp. 9–14, 2008.

A. Eyre-Walker, "The genomic rate of adaptive evolution," *Trends in Ecology and Evolution*, vol. 21, no. 10, pp. 569–575, 2006.

M. Fablet, "Host control of insect endogenous retroviruses: small rna silencing and immune response," *Viruses*, vol. 6, no. 11, pp. 4447–4464, 2014.

C. R. Faehnle, E. Elkayam, A. D. Haase, G. J. Hannon, and L. Joshua-Tor, "The making of a slicer: activation of human argonaute-1," *Cell Reports*, vol. 3, no. 6, pp. 1901–1909, 2013.

E. H. Feinberg and C. P. Hunter, "Transport of dsrna into cells by the transmembrane protein sid-1.," *Science*, vol. 301, no. 5639, pp. 1545–1547, 2003.

A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, "On detecting incomplete soft or hard selective sweeps using haplotype structure," *Molecular Biology and Evolution*, vol. 31, no. 5, pp. 1275–1291, 2014.

R. D. Finn, J. Mistry, J. Tate, P. Coggill, a. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and a. Bateman, "The pfam protein families database," *Nucleic Acids Research*, vol. 38, no. November 2009, pp. D211–D222, 2009.

A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans," *Nature*, vol. 391, no. February, pp. 806–811, 1998.

A. Force, M. Lynch, F. B. Pickett, A. Amores, Y.-l. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, pp. 1531–1545, 1999.

F. Frank, N. Sonenberg, and B. Nagar, "Structural basis for 5'-nucleotide base-specific recognition of guide rna by human ago2.," *Nature*, vol. 465, no. 7299, pp. 818–822, 2010.

K. J. Fryxell, "The coevolution of gene family trees," *Trends in Genetics*, vol. 12, pp. 364–369, 1996.

N. Funayama, M. Nakatsukasa, K. Mohri, Y. Masuda, and K. Agata, "Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges," *Evolution and Development*, vol. 12, no. 3, pp. 275–287, 2010.

D. A. Galbraith, X. Yang, E. L. Niño, S. Yi, and C. Grozinger, "Parallel epigenomic and transcriptomic responses to viral infection in honey bees (apis mellifera)," *PLoS Pathogens*, vol. 11, no. 3, e1004713, 2015.

J. S. Garbutt and S. E. Reynolds, "Induction of rna interference genes by double-stranded rna; implications for susceptibility to rna interference," *Insect Biochemistry and Molecular Biology*, vol. 42, pp. 621–628, 2012.

J. E. Garneau, M.-È. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadán, and S. Moineau, "The crispr/cas bacterial immune system cleaves bacteriophage and plasmid dna.," *Nature*, vol. 468, no. 7320, pp. 67–71, 2010.

S. L. Gell and R. A. Reenan, "Mutations to the pirna pathway component aubergine enhance meiotic drive of segregation distorter in drosophila melanogaster," *Genetics*, vol. 193, no. 3, pp. 771–784, 2013.

S Gershenson, "A new sex-ratio abnormality in drosophila obscura.," *Genetics*, vol. 13, no. 6, pp. 488–507, 1928.

J. G. Gibbons, A. T. Branco, S. A. Godinho, S. Yu, and B. Lemos, "Concerted copy number variation balances ribosomal dna dosage in human and mouse genomes," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, p. 201 416 878, 2015.

E. González-González, P. P. López-Casas, and J. del Mazo, "The expression patterns of genes involved in the rnai pathways are tissue-dependent and differ in the germ and somatic cells of mouse testis," *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1779, no. 5, pp. 306–311, 2008.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from rna-seq data without a reference genome.," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–52, 2011.

S. T. Grivna, B. Pyhtila, and H. Lin, "Miwi associates with translational machinery and piwi-interacting rnas (pirnas) in regulating spermatogenesis.," *Proceedings of the National Academy of Sciences*, vol. 103, no. 36, pp. 13 415–13 420, 2006.

M. Ha and V. N. Kim, "Regulation of microrna biogenesis.," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 8, pp. 509–524, 2014.

P. R. Haddrill, L. Loewe, and B. Charlesworth, "Estimating the parameters of selection on nonsynonymous mutations in drosophila pseudoobscura and d. miranda.," *Genetics*, vol. 185, no. 4, pp. 1381–96, 2010.

W. Haerty, S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. R. Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh, "Evolution in the fast lane: rapidly evolving sex-related genes in drosophila," *Genetics*, vol. 177, no. November, pp. 1321–1335, 2007.

M. W. Hahn, "Distinguishing among evolutionary models for the maintenance of gene duplicates.," *The Journal of Heredity*, vol. 100, no. 5, pp. 605–17, 2009.

T. Hai, F. Teng, R. Guo, W. Li, and Q. Zhou, "One-step generation of knockout pigs by zygote injection of crispr/cas system.," *Cell Research*, vol. 24, no. 3, pp. 372–5, 2014.

D. Hain, B. R. Bettencourt, K. Okamura, T. Csorba, W. Meyer, Z. Jin, J. Biggerstaff, H. Siomi, G. Hutvagner, E. C. Lai, M. Welte, and H.-A. J. Müller, "Natural variation of the amino-terminal glutamine-rich domain in drosophila argonaute2 is not associated with developmental defects.," *PLoS One*, vol. 5, no. 12, e15264, 2010.

J. B. S. Haldane, *The causes of evolution*. 1932, p. 222.

B. W. Han, W. Wang, C. Li, and Z. Weng, "Pirna-guided transposon cleavage initiates zucchini-dependent, phased pirna production," *Science*, vol. 348, no. 6236, pp. 817–821, 2015.

M. Han, S. Qin, X. Song, Y. Li, P. Jin, L. Chen, and F. Ma, "Evolutionary rate patterns of genes involved in the drosophila toll and imd signaling pathway.," *BMC Evolutionary Biology*, vol. 13, no. 1, p. 245, 2013.

M. V. Han, G. W. C. Thomas, J. Lugo-Martinez, and M. W. Hahn, "Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using cafe 3," *Molecular Biology and Evolution*, vol. 30, pp. 1987–1997, 2013.

P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, "Mechanisms of change in gene copy number.," *Nature Reviews Genetics*, vol. 10, no. 8, pp. 551–564, 2009.

X. He and J. Zhang, "Gene complexity and gene duplicability.," *Current Biology*, vol. 15, no. 11, pp. 1016–21, 2005.

A. Heger and C. P. Ponting, "Variable strength of translational selection among 12 drosophila species," *Genetics*, vol. 177, no. 3, pp. 1337–1348, 2007.

T. J. A. J. Heinen, F. Staubach, D. Häming, and D. Tautz, "Emergence of a new gene from an intergenic region.," *Current Biology*, vol. 19, no. 18, pp. 1527–31, 2009.

C. T. Hittinger and S. B. Carroll, "Gene duplication and the adaptive evolution of a classic genetic switch.," *Nature*, vol. 449, no. 7163, pp. 677–681, 2007.

S. Holtzman, D. Miller, R. C. Eisman, H. Kuwayama, T. Niimi, and T. C. Kaufman, "Transgenic tools for members of the genus drosophila with sequenced genomes," *Fly*, vol. 4, no. 4, pp. 349–362, 2010.

M. D. Horwich, C. Li, C. Matranga, V. Vagin, G. Farley, P. Wang, and P. D. Zamore, "The drosophila rna methyltransferase, dmhen1, modifies germline pirnas and single-stranded sirnas in risc," *Current Biology*, vol. 17, no. 14, pp. 1265–1272, 2007.

S. Houwing, L. M. Kamminga, E. Berezikov, D. Cronembold, A. Girard, H. van den Elst, D. V. Filippov, H. Blaser, E. Raz, C. B. Moens, R. H. A. Plasterk, G. J. Hannon, B. W. Draper, and R. F. Ketting, "A role for piwi and pirnas in germ cell maintenance and transposon silencing in zebrafish," *Cell*, vol. 129, no. 1, pp. 69–82, 2007.

Y. Huang, T. Kendall, E. S. Forsythe, A. Dorantes-Acosta, S. Li, J. Caballero-Perez, X. Chen, M. Arteaga-Vazquez, M. A. Beilstein, and R. A. Mosher, "Ancient origin and recent innovations of rna polymerase iv and v," *Molecular Biology and Evolution*, pp. 1–12, 2015.

R. R. Hudson, "Generating samples under a wright-fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.

R. R. Hudson, M. Kreitman, and M. Aguadé, "A test of neutral molecular evolution based on nucleotide data," *Genetics*, vol. 116, pp. 153–159, 1987.

T. Hughes and D. A. Liberles, "The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation," *Journal of Molecular Evolution*, vol. 65, pp. 574–588, 2007.

E. Huntzinger and E. Izaurralde, "Gene silencing by microrna: contributions of translational repression and mrna decay.," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 99–110, 2011.

J. K. Hur, M. K. Zinchenko, S. Djuranovic, and R. Green, "Regulation of argonaute slicer activity by guide rna 3' end interactions with the n-terminal lobe," *Journal of Biological Chemistry*, vol. 288, no. 11, pp. 7829–7840, 2013.

G Hutvágner, J McLachlan, A. E. Pasquinelli, E Bálint, T Tuschl, and P. D. Zamore, "A cellular function for the rna-interference enzyme dicer in the maturation of the let-7 small temporal rna.," *Science*, vol. 293, no. 5531, pp. 834–8, 2001.

H. Ito, "Small rnas and regulation of transposons in plants.," *Genes & Genetic Systems*, vol. 88, no. 1, pp. 3–7, 2013.

J. Jaenike, "Sex chromosome meiotic drive," *Annual Review of Ecology and Systematics*, vol. 32, no. May, pp. 25–49, 2001.

S. Jaubert-Possamai, C. Rispe, S. Tanguy, K. Gordon, T. Walsh, O. Edwards, and D. Tagu, "Expansion of the mirna pathway in the hemipteran insect acyrthosiphon pisum," *Molecular Biology and Evolution*, vol. 27, no. 5, pp. 979–987, 2010.

F. Jiang, X. Ye, X. Liu, L. Fincher, D. McKearin, and Q. Liu, "Dicer-1 and r3d1-l catalyze microrna maturation in drosophila," *Genes and Development*, vol. 19, no. 14, pp. 1674–1679, 2005.

G. Jordan and N. Goldman, "The effects of alignment error and alignment filtering on the sitewise detection of positive selection.," *Molecular Biology and Evolution*, vol. 29, no. 4, pp. 1125–39, 2012.

H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes.," *Genome Research*, vol. 20, no. 10, pp. 1313–26, 2010.

H. Kaessmann, N. Vinckenbosch, and M. Long, "Rna-based gene duplication: mechanistic and evolutionary insights.," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 19–31, 2009.

A. I. Kalmykova, M. S. Klenov, and V. A. Gvozdev, "Argonaute protein piwi controls mobilization of retrotransposons in the drosophila male germline.," *Nucleic Acids Research*, vol. 33, no. 6, pp. 2052–9, 2005.

Y. Kataoka, M. Takeichi, and T. Uemura, "Developmental roles and molecular characterization of a drosophila homologue of arabidopsis argonaute1, the founder of a novel gene superfamily," *Genes to Cells*, vol. 6, pp. 313–325, 2001.

V. Katju and U. Bergthorsson, "Copy-number changes in evolution: rates, fitness effects and adaptive significance," *Frontiers in Genetics*, vol. 4, no. DEC, pp. 1–12, 2013.

K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft : a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond, "Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data," *Bioinformatics*, vol. 28, no. 12, pp. 1647–1649, 2012.

M. C. Keays, D. Barker, C. Wicker-Thomas, and M. G. Ritchie, "Signatures of selection and sex-specific expression variation of a novel duplicate during the evolution of the drosophila desaturase gene family.," *Molecular Ecology*, vol. 20, no. 17, pp. 3617–30, 2011.

K. M. Keene, B. D. Foy, I. Sanchez-Vargas, B. J. Beaty, C. D. Blair, and K. E. Olson, "Rna interference acts as a natural antiviral response to o'nyong-nyong virus (alphavirus; togaviridae) infection of anopheles gambiae.," *Proceedings of the National Academy of Sciences*, vol. 101, no. 49, pp. 17 240–5, 2004.

D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.," *Genome Biology*, vol. 14, no. 4, R36, 2013.

V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small rnas in animals.," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–39, 2009.

Y.-K. Kim, G. Wee, J. Park, J. J.-S. Kim, D. Baek, and V. N. Kim, "Talen-based knockout library for human microrna.," *Nature Structural & Molecular Biology*, vol. 20, no. 12, pp. 1458–64, 2013.

K. C. Kleene, "A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells.," *Mechanisms of Development*, vol. 106, no. 1-2, pp. 3–23, 2001.

R. Kofler, P. Orozco-terWengel, N. De Maio, R. V. Pandey, V. Nolte, A. Futschik, C. Kosiol, and C. Schlötterer, "Popoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals.," *PLoS One*, vol. 6, no. 1, e15925, 2011.

B. Kolaczkowski, D. N. Hupalo, and A. D. Kern, "Recurrent adaptation in rna interference genes across the drosophila phylogeny," *Molecular Biology and Evolution*, vol. 28, no. Ml, pp. 1033–1042, 2011.

S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost, "Automated phylogenetic detection of recombination using a genetic algorithm," *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1891–1901, 2006.

S. Kuramochi-Miyagawa, T. Kimura, T. W. Ijiri, T. Isobe, N. Asada, Y. Fujita, M. Ikawa, N. Iwai, M. Okabe, W. Deng, H. Lin, Y. Matsuda, and T. Nakano, "Mili, a mammalian member of piwi family gene, is essential for spermatogenesis.," *Development*, vol. 131, no. 4, pp. 839–849, 2004.

P. B. Kwak and Y. Tomari, "The n domain of argonaute drives duplex unwinding during risc assembly," *Nature Structural & Molecular Biology*, vol. 19, no. 2, pp. 145–151, 2012.

L. Lambrechts and T. W. Scott, "Mode of transmission and the evolution of arbovirus virulence in mosquito vectors.," *Proceedings of the Royal Society B*, vol. 276, no. 1660, pp. 1369–78, 2009.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome.," *Genome Biology*, vol. 10, no. 3, R25, 2009.

Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim, "The nuclear rnase iii drosha initiates microrna processing.," *Nature*, vol. 425, no. 6956, pp. 415–9, 2003.

Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "Microrna genes are transcribed by rna polymerase ii.," *The EMBO Journal*, vol. 23, no. 20, pp. 4051–4060, 2004.

Y. S. Lee, K. Nakahara, J. W. Pham, K. Kim, Z. He, E. J. Sontheimer, and R. W. Carthew, "Distinct roles for drosophila dicer-1 and dicer-2 in the sirna/mirna silencing pathways," *Cell*, vol. 117, pp. 69–81, 2004.

W. Leebonoi, S. Sukthaworn, S. Panyim, and A. Udomkit, "A novel gonad-specific argonaute 4 serves as a defense against transposons in the black tiger shrimp penaeus monodon," *Fish & Shellfish Immunology*, vol. 42, no. 2, pp. 280–288, 2015.

B. Lemos, B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl, "Evolution of proteins and gene expression levels are coupled in drosophila and are independently associated with mrna abundance, protein length, and number of protein-protein interactions," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1345–1354, 2005.

F. Leulier and B. Lemaitre, "Toll-like receptors–taking an evolutionary approach.," *Nature Reviews Genetics*, vol. 9, no. 3, pp. 165–78, 2008.

M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression.," *Proceedings of the National Academy of Sciences*, vol. 103, no. 26, pp. 9935–9, 2006.

P L'Héritier and G Teissier, "Une anomalie physiologique héréditaire chez la drosophile," *CR Acad. Sci. Paris*, vol. 231, pp. 192–194, 1937.

C. Li, V. V. Vagin, S. Lee, J. Xu, S. Ma, H. Xi, H. Seitz, M. D. Horwich, M. Syrzycka, B. M. Honda, E. L. W. Kittler, M. L. Zapp, C. Klattenhoff, N. Schulz, W. E. Theurkauf, Z. Weng, and P. D. Zamore, "Collapse of germline pirnas in the absence of argonaute3 reveals somatic pirnas in flies.," *Cell*, vol. 137, no. 3, pp. 509–21, 2009.

F. Li and S.-W. Ding, "Virus counterdefense: diverse strategies for evading the rna-silencing immunity.," *Annual Review of Microbiology*, vol. 60, pp. 503–531, 2006.

H. Li and W. Stephan, "Inferring the demographic history and rate of adaptive substitution in drosophila," *PLoS Genetics*, vol. 2, no. 10, pp. 1580–1589, 2006.

H. Li, W. X. Li, and S. W. Ding, "Induction and suppression of rna silencing by an animal virus.," *Science*, vol. 296, no. May, pp. 1319–1321, 2002.

Y. Li, J. Lu, Y. Han, X. Fan, and S.-W. Ding, "Rna interference functions as an antiviral immunity mechanism in mammals.," *Science*, vol. 342, no. 6155, pp. 231–4, 2013.

P. Librado and J Rozas, "Dnasp v5: a software for comprehensive analysis of dna polymorphism data.," *Bioinformatics*, vol. 25, no. 11, pp. 1451–2, 2009.

K. van der Linde, D. Houle, G. S. Spicer, and S. J. Steppan, "A supermatrix-based molecular phylogeny of the family drosophilidae.," *Genetics Research*, vol. 92, no. 1, pp. 25–38, 2010.

J. Liu, Y. Zhang, X. Lei, and Z. Zhang, "Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective.," *Genome Biology*, vol. 9, no. 4, R69, 2008.

B. Longdon and F. M. Jiggins, "Vertically transmitted viral endosymbionts of insects: do sigma viruses walk alone?" *Proceedings of the Royal Society B*, vol. 279, no. August, pp. 3889–3898, 2012.

A. Löytynoja and N. Goldman, "An algorithm for progressive multiple alignment of sequences with insertions.," *Proceedings of the National Academy of Sciences*, vol. 102, no. 30, pp. 10 557–10 562, 2005.

H.-L. Lu, S. Tanguy, C. Rispe, J.-P. Gauthier, T. Walsh, K. Gordon, O. Edwards, D. Tagu, C.-c. Chang, and S. Jaubert-Possamai, "Expansion of genes encoding pirna-associated argonaute proteins in the pea aphid: diversification of expression profiles in different plastic morphs.," *PLoS One*, vol. 6, no. 12, e28051, 2011.

M. J. Luteijn and R. F. Ketting, "Piwi-interacting rnas: from generation to transgenerational epigenetics.," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 523–34, 2013.

M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

M Lynch and T. J. Crease, "The analysis of population survey data on dna sequence variation.," *Molecular Biology and Evolution*, vol. 7, no. 4, pp. 377–394, 1990.

J.-B. Ma, Y.-R. Yuan, G. Meister, Y. Pei, T. Tuschl, and D. J. Patel, "Structural basis for 5'-end-specific recognition of guide rna by the a. fulgidus piwi protein.," *Nature*, vol. 434, no. 7033, pp. 666–670, 2005.

P. V. Maillard, C Ciaudo, a Marchais, Y Li, F Jay, S. W. Ding, and O. Voinnet, "Antiviral rna interference in mammalian cells.," *Science*, vol. 342, no. 6155, pp. 235–8, 2013.

K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin, "A putative rna-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic rnai, and hypothetical mechanisms of action.," *Biology Direct*, vol. 1, p. 7, 2006.

K. S. Makarova, Y. I. Wolf, J. van der Oost, and E. V. Koonin, "Prokaryotic homologs of argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements.," *Biology Direct*, vol. 4, p. 29, 2009.

T. Marcussen, B. Oxelman, A. Skog, and K. S. Jakobsen, "Evolution of plant rna polymerase iv/v genes: evidence of subneofunctionalization of duplicated nrpd2/nrpe2-like paralogs in viola (violaceae)," *BMC Evolutionary Biology*, vol. 10, no. 1, pp. 45–60, 2010.

M. A. Matzke and R. A. Mosher, "Rna-directed dna methylation: an epigenetic pathway of increasing complexity.," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 394–408, 2014.

J. H. McDonald and M. Kreitman, "Adaptive protein evolution at the adh locus in drosophila," *Nature*, vol. 351, pp. 652–654, 1991.

S. E. McGaugh, C. S. S. Heil, B. Manzano-Winkler, L. Loewe, S. Goldstein, T. L. Himmel, and M. A. F. Noor, "Recombination modulates how selection affects linked sites in drosophila.," *PLoS Biology*, vol. 10, no. 11, e1001422, 2012.

A. McLysaght, T. Makino, H. M. Grayton, M. Tropeano, K. J. Mitchell, E. Vassos, and D. A. Collier, "Ohnologs are overrepresented in pathogenic copy number mutations.," *Proceedings of the National Academy of Sciences*, vol. 111, no. 1, pp. 361–6, 2014.

G. A. T. McVean and B. Charlesworth, "The effects of hill-robertson interference between weakly selected mutations on patterns of molecular evolution and variation," *Genetics*, vol. 155, pp. 929–944, 2000.

G. Meister, "Argonaute proteins: functional insights and emerging roles.," *Nature Reviews Genetics*, vol. 14, no. 7, pp. 447–59, 2013.

D. U. Menon and V. H. Meller, "A role for sirna in x-chromosome dosage compensation in drosophila melanogaster.," *Genetics*, vol. 191, no. 3, pp. 1023–8, 2012.

W. J. Meyer, S. Schreiber, Y. Guo, T. Volkmann, M. A. Welte, and H. A. J. Müller, "Overlapping functions of argonaute proteins in patterning and morphogenesis of drosophila embryos," *PLoS Genetics*, vol. 2, no. 8, pp. 1224–1239, 2006.

J. T. van Mierlo, A. W. Bronkhorst, G. J. Overheul, S. A. Sadanandan, J.-O. Ekström, M. Heestermans, D. Hultmark, C. Antoniewski, and R. P. van Rij, "Convergent evolution of argonaute-2 slicer antagonism in two distinct insect rna viruses.," *PLoS Pathogens*, vol. 8, no. 8, e1002872, 2012.

J. T. van Mierlo, G. J. Overheul, B. Obadia, K. W. R. van Cleef, C. L. Webster, M.-C. Saleh, D. J. Obbard, and R. P. van Rij, "Novel drosophila viruses encode host-specific suppressors of rnai.," *PLoS Pathogens*, vol. 10, no. 7, e1004256, 2014.

P. Miesen, E. Girardi, and R. P. van Rij, "Distinct sets of piwi proteins produce arbovirus and transposon-derived pirnas in aedes aegypti mosquito cells," *Nucleic Acids Research*, pp. 1–12, 2015.

J. C. Miller, S. Tan, G. Qiao, K. A. Barlow, J. Wang, D. F. Xia, X. Meng, D. E. Paschon, E. Leung, S. J. Hinkley, G. P. Dulay, K. L. Hua, I. Ankoudinova, G. J. Cost, F. D. Urnov, H. S. Zhang, M. C. Holmes, L. Zhang, P. D. Gregory, and E. J. Rebar, "A tale nuclease architecture for efficient genome editing.," *Nature Biotechnology*, vol. 29, no. 2, pp. 143–148, 2011. arXiv: 5.

B. Misof, S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, O. Niehuis, M. Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspock, H. Aspock, D. Bartel, A. Blanke, S. Berger, A. Bohm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermiin, A. Y. Kawahara,

L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schutte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich, M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walzl, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. F. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, X. Xu, H. Yang, J. Wang, K. M. Kjer, and X. Zhou, "Phylogenomics resolves the timing and pattern of insect evolution," *Science*, vol. 346, no. 6210, pp. 763–767, 2014.

F. Mohn, D. Handler, and J. Brennecke, "Pirna-guided slicing specifies transcripts for zucchini-dependent, phased pirna biogenesis," *Science*, vol. 348, pp. 812–817, 2015.

C. Morandin, H. Havukainen, J. Kulmuni, K. Dhaygude, K. Trontti, and H. Helanterä, "Not only for egg yolk-functional and evolutionary insights from expression, selection, and structural analyses of formica ant vitellogenins," *Molecular Biology and Evolution*, vol. 31, no. 8, pp. 2181–2193, 2014.

E. M. Morazzani, M. R. Wiley, M. G. Murreddu, Z. N. Adelman, and K. M. Myles, "Production of virus-derived ping-pong-dependent pirna-like small rnas in the mosquito soma.," *PLoS Pathogens*, vol. 8, no. 1, e1002470, 2012.

T. V. Morozova, E. A. Tsybulko, and E. G. Pasyukova, "Regularory elements of the copia retrotransposon determine different levels of expression in different organs of males and females of drosophila melanogaster," *Genetika*, vol. 45, no. 2, pp. 169–177, 2009.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq.," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

B. R. Morton, "Chloroplast dna codon use: evidence for selection at the psb a locus based on trna availability.," *Journal of Molecular Evolution*, vol. 37, no. 3, pp. 273–280, 1993.

A.-K. Mueller, C. Hammerschmidt-Kamper, and A. Kaiser, "Rnai in plasmodium.," *Current Parmaceutical Design*, vol. 20, no. 2, pp. 278–83, 2014.

B. Mugat, A. Akkouche, V. Serrano, C. Armenise, B. Li, C. Brun, T. A. Fulga, D. Van Vactor, A. Pélisson, and S. Chambeyron, "Microrna-dependent transcriptional silencing of transposable elements in drosophila follicle cells," *PLoS Genetics*, vol. 11, no. 5, e1005194, 2015.

K. Mukherjee, H. Campos, and B. Kolaczkowski, "Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral rna binding in plants.," *Molecular Biology and Evolution*, vol. 30, no. 3, pp. 627–41, 2013.

A. Nagao, T. Mituyama, H. Huang, D. Chen, M. C. Siomi, and H. Siomi, "Biogenesis pathways of pirnas loaded onto ago3 in the drosophila testis.," *RNA*, vol. 16, no. 12, pp. 2503–15, 2010.

K. Nakanishi, D. E. Weinberg, D. P. Bartel, and D. J. Patel, "Structure of yeast argonaute with guide rna.," *Nature*, vol. 486, no. 7403, pp. 368–74, 2012.

A. Nayak, B. Berry, M. Tassetto, M. Kunitomi, A. Acevedo, C. Deng, A. Krutchinsky, J. Gross, C. Antoniewski, and R. Andino, "Cricket paralysis virus antagonizes argonaute 2 to modulate antiviral defense in drosophila.," *Nature Structural & Molecular Biology*, vol. 17, no. 5, pp. 547–554, 2010.

D.-Q. Nguyen, C. Webber, J. Hehir-kwa, R. Pfundt, J. Veltman, and C. P. Ponting, "Reduced purifying selection prevails over positive selection in human copy number variant evolution," *Genome Research*, vol. 18, pp. 1711–1723, 2008.

W. Ni, J. Qiao, S. Hu, X. Zhao, M. Regouski, M. Yang, I. A. Polejaeva, and C. Chen, "Efficient gene knockout in goats using crispr/cas9 system," *PLoS One*, vol. 9, no. 9, e106718, 2014.

M. G. Nielsen, S. R. Gadagkar, and L. Gutzwiller, "Tubulin evolution in insects: gene duplication and subfunctionalization provide specialized isoforms in a functionally constrained gene family.," *BMC Evolutionary Biology*, vol. 10, p. 113, 2010.

R. Nielsen, "Molecular signatures of natural selection.," *Annual Review of Genetics*, vol. 39, pp. 197–218, 2005.

R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante, "Genomic scans for selective sweeps using snp data.," *Genome Research*, vol. 15, no. 11, pp. 1566–75, 2005.

K. M. Nishida, K. Saito, T. Mori, Y. Kawamura, T. Nagami-Okada, S. Inagaki, H. Siomi, and M. C. Siomi, "Gene silencing mechanisms mediated by aubergine pirna complexes in drosophila male gonad.," *RNA*, vol. 13, pp. 1911–1922, 2007.

Y. Niu, B. Shen, Y. Cui, Y. Chen, J. Wang, L. Wang, Y. Kang, X. Zhao, W. Si, W. Li, A. Xiang, J. Zhou, X. Guo, Y. Bi, C. Si, B. Hu, G. Dong, H. Wang, Z. Zhou, T. Li, T. Tan, X. Pu, F. Wang, S. Ji, Q. Zhou, X. Huang, W. Ji, and J. Sha, "Generation of gene-modified cynomolgus monkey via cas9/rna-mediated gene targeting in one-cell embryos," *Cell*, pp. 1–8, 2014.

M. Nosaka, J. I. Itoh, Y. Nagato, A. Ono, A. Ishiwata, and Y. Sato, "Role of transposon-derived small rnas in the interplay between genomes and parasitic dna in rice," *PLoS Genetics*, vol. 8, no. 9, 2012.

D. I. Nurminsky, M. V. Nurminskaya, D. de Aguiar, and D. L. Hartl, "Selective sweep of a newly evolved sperm-specific gene in drosophila," *Nature*, vol. 396, pp. 572–575, 1998.

D. J. Obbard, F. M. Jiggins, D. L. Halligan, and T. J. Little, "Natural selection drives extremely rapid evolution in antiviral rnai genes.," *Current Biology*, vol. 16, no. 6, pp. 580–5, 2006.

D. J. Obbard, J. J. Welch, K.-W. Kim, and F. M. Jiggins, "Quantifying adaptive evolution in the drosophila immune system.," *PLoS Genetics*, vol. 5, no. 10, e1000698, 2009.

D. J. Obbard, K. H. J. Gordon, A. H. Buck, and F. M. Jiggins, "The evolution of rnai as a defence against viruses and transposable elements.," *Philosophical Transactions of the Royal Society of London Biological Sciences*, vol. 364, no. 1513, pp. 99–115, 2009.

D. J. Obbard, F. M. Jiggins, N. J. Bradshaw, and T. J. Little, "Recent and recurrent selective sweeps of the antiviral rnai gene argonaute-2 in three species of drosophila.," *Molecular Biology and Evolution*, vol. 28, no. 2, pp. 1043–56, 2011.

D. J. Obbard, J. MacLennan, K. W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins, "Estimating divergence dates and substitution rates in the drosophila phylogeny," *Molecular Biology and Evolution*, vol. 29, no. 11, pp. 3459–3473, 2012.

S. Ohno, *Evolution by gene duplication*. Springer Verlag, New York, NY, 1970, p. 160.

I. Olovnikov, K. Chan, R. Sachidanandam, D. K. Newman, and A. A. Aravin, "Bacterial argonaute samples the transcriptome to identify foreign dna.," *Molecular Cell*, vol. 51, no. 5, pp. 594–605, 2013.

J. van der Oost, E. R. Westra, R. N. Jackson, and B. Wiedenheft, "Unravelling the structural and mechanistic basis of crispr-cas systems.," *Nature Reviews Microbiology*, vol. 12, no. 7, pp. 479–92, 2014.

J. Pak and A. Fire, "Distinct populations of primary and secondary effectors during rnai in c. elegans.," *Science*, vol. 315, no. 5809, pp. 241–244, 2007.

C. Pal, B. Papp, and L. D. Hurst, "Highly expressed genes in yeast evolve slowly," *Genetics*, vol. 158, pp. 927–931, 2001.

M. Pal-Bhadra, B. A. Leibovitch, S. G. Gandhi, M. Rao, U. Bhadra, J. A. Birchler, and S. C. R. Elgin, "Heterochromatic silencing and hp1 localization in drosophila are dependent on the rnai machinery.," *Science*, vol. 303, no. January 2004, pp. 669–672, 2004.

D. Palakodeti, M. Smielewska, Y.-C. Lu, G. W. Yeo, and B. R. Graveley, "The piwi proteins smedwi-2 and smedwi-3 are required for stem cell function and pirna expression in planarians.," *RNA*, vol. 14, pp. 1174–1186, 2008.

W. J. Palmer and F. M. Jiggins, "Comparative genomics reveals the origins and diversity of arthropod immune systems," *Molecular Biology and Evolution*, vol. 32, no. 8, pp. 2111–2129, 2015.

N. Palmieri, C. Kosiol, and C. Schlotterer, "The life cycle of drosophila orphan genes," *ELife*, vol. 3, e01311–e01311, 2014.

G. Palumbo, S. Bonaccorsi, L. G. Robbins, and S. Pimpinelli, "Genetic analysis of stellate elements of drosophila melanogaster," *Genetics*, vol. 138, no. Livak 1990, pp. 1181–1197, 1994.

A. R. Panchenko, Y. I. Wolf, L. A. Panchenko, and T. Madej, "Evolutionary plasticity of protein families: coupling between sequence and structure variation," *Proteins: Structure, Function and Genetics*, vol. 61, no. May, pp. 535–544, 2005.

J.-S. Parent, A. E. Martínez de Alba, and H. Vaucheret, "The origin and effect of small rna signaling in plants," *Frontiers in Plant Science*, vol. 3, no. August, pp. 1–9, 2012.

E. Pasyukova, S. Nuzhdin, W. Li, and A. J. Flavell, "Germ line transposition of the copia retrotransposon in drosophila melanogaster is restricted to males by tissue-specific control of copia rna levels," *Molecular and General Genetics*, vol. 255, pp. 115–124, 1997.

C. A. Paulding, M. Ruvolo, and D. A. Haber, "The tre2 (usp6) oncogene is a hominoid-specific gene.," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2507–11, 2003.

J. Peden, "Analysis of codon usage bias," PhD thesis, The University of Nottingham, 1995.

D. Peng, S. P. Kurup, P. Y. Yao, T. A. Minning, and R. L. Tarleton, "Crispr-cas9-mediated single-gene and gene family disruption in trypanosoma cruzi," *MBio*, vol. 6, no. 1, e02097–14, 2014.

G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, and A. C. Stone, "Diet and the evolution of human amylase gene copy number variation.," *Nature Genetics*, vol. 39, no. 10, pp. 1256–1260, 2007.

J. Pfaff and G. Meister, "Argonaute and gw182 proteins: an effective alliance in gene silencing.," *Biochemical Society transactions*, vol. 41, no. 4, pp. 855–60, 2013.

H. Philippe, H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain, "Resolving difficult phylogenetic questions: why more sequences are not enough," *PLoS Biology*, vol. 9, no. 3, 2011.

F. Port, H.-M. Chen, T. Lee, and S. L. Bullock, "Optimized crispr/cas tools for efficient germline and somatic genome engineering in drosophila.," *Proceedings of the National Academy of Sciences*, 2014.

A. Prachumwat and W. H. Li, "Protein function, connectivity, and duplicability in yeast," *Molecular Biology and Evolution*, vol. 23, no. 1, pp. 30–39, 2006.

P. Rajasethupathy, I. Antonov, R. Sheridan, S. Frey, C. Sander, T. Tuschl, and E. R. Kandel, "A role for neuronal pirnas in the epigenetic control of memory-related synaptic plasticity," *Cell*, vol. 149, no. 3, pp. 693–707, 2012.

S. Rastogi and D. A. Liberles, "Subfunctionalization of duplicated genes as a transition state to neofunctionalization.," *BMC Evolutionary Biology*, vol. 5, p. 28, 2005.

P. W. Reddien, N. J. Oviedo, J. R. Jennings, J. C. Jenkin, and A. Sánchez Alvarado, "Smedwi-2 is a piwi-like protein that regulates planarian stem cells.," *Science*, vol. 310, pp. 1327–1330, 2005.

R. P. van Rij, M.-C. Saleh, B. Berry, C. Foo, A. Houk, C. Antoniewski, and R. Andino, "The rna silencing endonuclease argonaute 2 mediates specific antiviral immunity in drosophila melanogaster.," *Genes & Development*, vol. 20, no. 21, pp. 2985–95, 2006.

N. Robine, N. C. Lau, S. Balla, Z. Jin, K. Okamura, S. Kuramochi-Miyagawa, M. D. Blower, and E. C. Lai, "A broadly conserved pathway generates 3'utr-directed primary pirnas," *Current Biology*, vol. 19, no. 24, pp. 2066–2076, 2009.

F. Ronquist and J. P. Huelsenbeck, "Mrbayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

R. J. Ross, M. M. Weiner, and H. Lin, "Piwi proteins and piwi-interacting rnas in the soma.," *Nature*, vol. 505, pp. 353–9, 2014.

S. Roush and F. J. Slack, "The let-7 family of micrornas," *Trends in Cell Biology*, vol. 18, no. 10, pp. 505–516, 2008.

N. V. Rozhkov, A. A. Aravin, E. S. Zelentsova, N. G. Schostak, R. Sachidanandam, W. R. McCombie, G. J. Hannon, and M. B. Evgen'ev, "Small rna-based silencing strategies for transposons in the process of invading drosophila species.," *RNA*, vol. 16, no. 8, pp. 1634–45, 2010.

G. M. Rubin and A. C. Spradling, "Genetic transformation of drosophila with transposable element vectors.," *Science*, vol. 218, no. 4570, pp. 348–353, 1982.

P. Sarkies and E. A. Miska, "Small rnas break out: the molecular cell biology of mobile small rnas," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 8, pp. 525–535, 2014.

P. Sarkies, M. E. Selkirk, J. T. Jones, V. Blok, T. Boothby, B. Goldstein, B. Hanelt, A. Ardila-Garcia, N. M. Fast, P. M. Schiffer, C. Kraus, M. J. Taylor, G. Koutsovoulos, M. L. Blaxter, and E. A. Miska, "Ancient and novel small rna pathways compensate for the loss of pirnas in multiple independent nematode lineages," *PLoS Biology*, vol. 13, e1002061, 2015.

S. W. Schaeffer, A. Bhutkar, B. F. McAllister, M. Matsuda, L. M. Matzkin, P. M. O'Grady, C. Rohde, V. L. S. Valente, M. Aguadé, W. W. Anderson, K. Edwards, A. C. L. Garcia, J. Goodman, J. Hartigan, E. Kataoka, R. T. Lapoint, E. R. Lozovsky, C. A. Machado, M. A. F. Noor, M. Papaceit, L. K. Reed, S. Richards, T. T. Rieger, S. M. Russo, H. Sato, C. Segarra, D. R. Smith, T. F. Smith, V. Strelets, Y. N. Tobari, Y. Tomimura, M. Wasserman, T. Watts, R. Wilson, K. Yoshida, T. A. Markow, W. M. Gelbart, and T. C. Kaufman, "Polytene chromosomal maps of 11 drosophila species: the order of genomic scaffolds inferred from genetic and physical maps.," *Genetics*, vol. 179, no. 3, pp. 1601–55, 2008.

N. T. Schirle and I. J. Macrae, "The crystal structure of human argonaute2," *Science*, vol. 336, pp. 1037–1040, 2012.

E. Schnettler, C. L. Donald, S. Human, M. Watson, R. W. C. Siu, M. McFarlane, J. K. Fazakerley, A. Kohl, and R. Fragkoudis, "Knockdown of pirna pathway proteins results in enhanced semliki forest virus production in mosquito cells.," *The Journal of General Virology*, vol. 94, no. Pt 7, pp. 1680–9, 2013.

E. Schnettler, M. Ratinier, M. Watson, A. E. Shaw, M. McFarlane, M. Varela, R. M. Elliott, M. Palmarini, and A. Kohl, "Rna interference targets arbovirus replication in culicoides cells.," *Journal of Virology*, vol. 87, no. 5, pp. 2441–54, 2013.

E. Schnettler, H. Tykalová, M. Watson, M. Sharma, M. G. Sterken, D. J. Obbard, S. H. Lewis, M. McFarlane, L. Bell-Sakyi, G. Barry, S. Weisheit, S. M. Best, R. J. Kuhn, G. P. Pijlman, M. E. Chase-Topping, E. A. Gould, L. Grubhoffer, J. K. Fazakerley, and A. Kohl, "Induction and suppression of tick cell antiviral rnai responses by tick-borne flaviviruses.," *Nucleic Acids Research*, vol. 42, no. 14, pp. 1–11, 2014.

D. R. Schrider, D. Houle, M. Lynch, and M. W. Hahn, "Rates and genomic consequences of spontaneous mutational events in drosophila melanogaster," *Genetics*, vol. 194, no. 4, pp. 937–954, 2013.

J. G. Scott, W. C. Warren, L. W. Beukeboom, D. Bopp, A. G. Clark, S. D. Giers, M. Hediger, A. K. Jones, S. Kasai, C. A. Leichter, M. Li, R. P. Meisel, P. Minx, T. D. Murphy, D. R. Nelson, W. R. Reid, F. D. Rinkevich, H. M. Robertson, T. B. Sackton, D. B. Sattelle, F. Thibaud-Nissen, C. Tomlinson, and L. V. D. Zande, "Genome of the house fly, musca domestica l., a global vector of diseases with adaptations to a septic environment," *Genome Biology*, vol. 15, pp. 466–482, 2014.

C. Segarra and M. Aguadé, "Molecular organization of the x chromosome in different species of the obscura group of drosophila," *Genetics*, vol. 130, no. 3, pp. 513–521, 1992.

K. Seipel, N. Yanze, and V. Schmid, "The germ line and somatic stem cell gene cniwi in the jellyfish podocoryne carnea," *International Journal of Developmental Biology*, vol. 48, no. 1, pp. 1–7, 2004.

S. A. Shabalina and E. V. Koonin, "Origins and evolution of eukaryotic rna interference," *Trends in Ecology and Evolution*, vol. 23, no. 10, pp. 578–587, 2008.

A. K. Sharma, M. C. Nelson, J. E. Brandt, M. Wessman, N. Mahmud, K. P. Weller, and R. Hoffman, "Human cd34+ stem cells express the hiwi gene, a human homologue of the drosophila gene piwi," *Blood*, vol. 97, no. 2, pp. 426–434, 2001.

P. M. Sharp and W. H. Li, "On the rate of dna sequence evolution in drosophila.," *Journal of Molecular Evolution*, vol. 28, no. 5, pp. 398–402, 1989.

S. Shpiz, S. Ryazansky, I. Olovnikov, Y. Abramov, and A. Kalmykova, "Euchromatic transposon insertions trigger production of novel pi- and endo-sirnas at the target sites in the drosophila germline," *PLoS Genetics*, vol. 10, no. 2, 2014.

G. Sienski, D. Dönertas, and J. Brennecke, "Transcriptional silencing of transposons by piwi and maelstrom and its impact on chromatin state and gene expression," *Cell*, vol. 151, pp. 964–980, 2012.

A. Simkin, A. Wong, Y. P. Poh, W. E. Theurkauf, and J. D. Jensen, "Recurrent and recent selective sweeps in the pirna pathway," *Evolution*, vol. 67, pp. 1081–1090, 2013.

R. K. Singh, K. Gase, I. T. Baldwin, and S. P. Pandey, "Molecular evolution and diversification of the argonaute family of proteins in plants," *BMC Plant Biology*, vol. 15, no. 1, pp. 1–16, 2015.

D. E. Skinner, G. Rinaldi, U. Koziol, K. Brehm, and P. J. Brindley, "How might flukes and tapeworms maintain genome integrity without a canonical pirna pathway?" *Trends in Parasitology*, vol. 30, no. 3, pp. 123–129, 2014.

J. M. Smith and J Haigh, "The hitch-hiking effect of a favourable gene.," *Genetical Research*, vol. 23, no. 1, pp. 23–35, 1974.

J.-J. Song, J. Liu, N. H. Tolia, J. Schneiderman, S. K. Smith, R. A. Martienssen, G. J. Hannon, and L. Joshua-Tor, "The crystal structure of the argonaute2 paz domain reveals an rna binding motif in rnai effector complexes.," *Nature Structural Biology*, vol. 10, no. 12, pp. 1026–1032, 2003.

J.-J. Song, S. K. Smith, and G. J. Hannon, "Crystal structure of argonaute slicer activity," *Science*, vol. 305, no. September, pp. 1434–1437, 2004.

F. Staubach, J. F. Baines, S. Künzel, E. M. Bik, and D. A. Petrov, "Host species and environmental effects on bacterial communities associated with drosophila in the laboratory and in the natural environment.," *PLoS One*, vol. 8, no. 8, e70749, 2013.

F. W. Stearns, "One hundred years of pleiotropy: a retrospective.," *Genetics*, vol. 186, no. 3, pp. 767–73, 2010.

M Stephens, N. J. Smith, and P Donnelly, "A new statistical method for haplotype reconstruction from population data.," *American Journal of Human Genetics*, vol. 68, no. 4, pp. 978–989, 2001.

A Stolfi, S Gandhi, F Salek, and L Christiaen, "Tissue-specific genome editing in ciona embryos by crispr/cas9," *Development*, vol. 9, pp. 4115–4120, 2014.

A. H. Sturtevant and T Dobzhansky, "Geographical distribution and cytology of "sex ratio" in drosophila pseudoobscura and related species.," *Genetics*, vol. 21, no. 4, pp. 473–490, 1936.

D. C. Swarts, J. W. Hegge, I. Hinojo, M. Shiimori, M. A. Ellis, J. Dumrongkulraksa, R. M. Terns, M. P. Terns, and J. van der Oost, "Argonaute of the archaeon pyrococcus furiosus is a dna-guided nuclease that targets cognate dna," *Nucleic Acids Research*, pp. 1–10, 2015.

D. C. Swarts, M. M. Jore, E. R. Westra, Y. Zhu, J. H. Janssen, A. P. Snijders, Y. Wang, D. J. Patel, J. Berenguer, S. J. J. Brouns, and J. van der Oost, "Dna-guided dna interference by a prokaryotic argonaute," *Nature*, 2014.

D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, and J. van der Oost, "The evolutionary journey of argonaute proteins," *Nature Structural & Molecular Biology*, vol. 21, no. 9, pp. 743–753, 2014.

F. Tajima, "Statistical method for testing the neutral mutation hypothesis by dna polymorphism," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

O. H. Tam, A. A. Aravin, P. Stein, A. Girard, E. P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R. M. Schultz, and G. J. Hannon, "Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes.," *Nature*, vol. 453, no. 7194, pp. 534–538, 2008.

Y. Tao, J. P. Masly, L. Araripe, Y. Ke, and D. L. Hartl, "A sex-ratio meiotic drive system in drosophila simulans. i: an autosomal suppressor," *PLoS Biology*, vol. 5, no. 11, pp. 2560–2575, 2007.

Y. Tao, L. Araripe, S. B. Kingan, Y. Ke, H. Xiao, and D. L. Hartl, "A sex-ratio meiotic drive system in drosophila simulans. ii: an x-linked distorter," *PLoS Biology*, vol. 5, no. 11, pp. 2576–2588, 2007.

E. L. Telleria, J. B. Benoit, X. Zhao, A. F. Savage, S. Regmi, T. L. A. e Silva, M. O'Neill, and S. Aksoy, "Insights into the trypanosome-host interactions revealed through transcriptomic analysis of parasitized tsetse fly salivary glands," *PLoS Neglected Tropical Diseases*, vol. 8, no. 4, 2014.

K. L. Tkaczuk, A. Obarska, and J. M. Bujnicki, "Molecular phylogenetics and comparative modeling of hen1, a methyltransferase involved in plant microrna biogenesis.," *BMC Evolutionary Biology*, vol. 6, p. 6, 2006.

Y. Tomari, T. Du, and P. D. Zamore, "Sorting of drosophila small silencing rnas," *Cell*, vol. 130, no. 2, pp. 299–308, 2007.

R. L. Unckless, A. M. Larracuente, and A. G. Clark, "Sex-ratio meiotic drive and y-linked resistance in drosophila affinis," *Genetics*, vol. Early Onli, 2015.

V. V. Vagin, M. S. Klenov, A. D. Stolyarenko, R. N. Kotelnikov, and V. A. Gvozdev, "The rna interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of drosophila melanogaster," *RNA Biology*, vol. 1, no. June, pp. 54–58, 2004.

B. Vicoso and B. Charlesworth, "Evolution on the x chromosome: unusual patterns and processes.," *Nature Reviews Genetics*, vol. 7, no. 8, pp. 645–653, 2006.

D. Vlad, D. Kierzkowski, M. I. Rast, F. Vuolo, R. D. Ioio, C. Galinha, X. Gan, M. Hajheidari, A. Hay, R. S. Smith, P. Huijser, C. D. Bailey, and M. Tsianis, "Leaf shape evolution through duplication, regulatory diversification and loss of a homeobox gene," *Science*, vol. 343, pp. 780–783, 2014.

N. Vodovar, A. W. Bronkhorst, K. W. R. van Cleef, P. Miesen, H. Blanc, R. P. van Rij, and M.-C. Saleh, "Arbovirus-derived pirnas exhibit a ping-pong signature in mosquito cells.," *PLoS One*, vol. 7, no. 1, e30861, 2012.

O. Voinnet, "Origin, biogenesis, and activity of plant micrornas," *Cell*, vol. 136, no. 4, pp. 669–687, 2009.

K. Voordeckers, C. A. Brown, K. Vanneste, E. van der Zande, A. Voet, S. Maere, and K. J. Verstrepen, "Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication.," *PLoS Biology*, vol. 10, no. 12, e1001446, 2012.

P. S. Walsh, D. A. Metzger, and R. Higuchi, "Chelex 100 as a medium for simple extraction of dna for pcr-based typing from forensic material," *BioTechniques*, vol. 10, no. 3, pp. 506–513, 1991.

D. Wang, Z. Zhang, E. O'Loughlin, T. Lee, S. Houel, D. O'Carroll, A. Tarakhovsky, N. G. Ahn, and R. Yi, "Quantitative functions of argonaute proteins in mammalian development," *Genes and Development*, vol. 26, no. 7, pp. 693–704, 2012.

X.-H. Wang, R. Aliyari, W.-X. Li, H.-W. Li, K. Kim, R. Carthew, P. Atkinson, and S.-W. Ding, "Rna interference directs innate immunity against viruses in adult drosophila.," *Science*, vol. 312, pp. 452–454, 2006.

M. Wassenegger and G. Krczal, "Nomenclature and functions of rna-directed rna polymerases," *Trends in Plant Science*, vol. 11, no. 3, pp. 142–151, 2006.

R. M. Waterhouse, E. V. Kriventseva, S. Meister, Z. Xi, K. S. Alvarez, L. C. Bartholomay, C. Barillas-Mury, G. Bian, S. Blandin, B. M. Christensen, Y. Dong, H. Jiang, M. R. Kanost, A. C. Koutsos, E. A. Levashina, J. Li, P. Ligoxygakis, R. M. Maccallum, G. F. Mayhew, A. Mendes, K. Michel, M. A. Osta, S. Paskewitz, S. W. Shin, and D. Vlachou, "Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes," *Science*, vol. 316, no. June, pp. 1738–1743, 2007.

B. M. Wiegmann, M. D. Trautwein, I. S. Winkler, N. B. Barr, J.-W. Kim, V. Blagoderov, J. Caravas, S. Narayanan, U. Schmidt-Ott, G. E. Kampmeier, R. Meier, and D. K. Yeates, "Episodic radiations in the fly tree of life," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 5690–5695, 2011.

R. W. Williams and G. M. Rubin, "Argonaute1 is required for efficient rna interference in drosophila embryos.," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6889–6894, 2002.

S. Woods, A. Coghlan, D. Rivers, T. Warnecke, S. J. Jeffries, T. Kwon, A. Rogers, L. D. Hurst, and J. Ahringer, "Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses," *PLoS Genetics*, vol. 9, no. 5, 2013.

F. Wright, "The 'effective number of codons' used in a gene," *Gene*, vol. 87, pp. 23–29, 1990.

S. I. Wright and B. Charlesworth, "The hka test revisited: a maximum-likelihood-ratio test of the standard neutral model.," *Genetics*, vol. 168, no. 2, pp. 1071–6, 2004.

C. I. Wu and A. T. Beckenbach, "Evidence for extensive genetic differentiation between the sex-ratio and the standard arrangement of drosophila pseudobscura and d. persimilis and identification of hybrid sterility factors.," *Genetics*, vol. 105, no. 1, pp. 71–86, 1983.

D.-D. Wu, D. M. Irwin, and Y.-P. Zhang, "Correlated evolution among six gene families in drosophila revealed by parallel change of gene numbers.," *Genome Biology and Evolution*, vol. 3, pp. 396–400, 2011.

J. Wu, Z. Yhang, Y. Wang, L. Zheng, R. Ye, Y. Ji, S. Zhao, S. Ji, R. Liu, L. Xu, H. Zheng, Y. Zhou, X. Zhang, X. Cao, L. Xie, Z. Wu, Y. Qi, and L. Yi, "Viral-inducible argonaute18 confers broad-spectrum resistance in rice by sequestering a host microrna," *ELife*, no. February, 2015.

Q. Wu, Y. Luo, R. Lu, N. Lau, E. C. Lai, W.-X. Li, and S.-W. Ding, "Virus discovery by deep sequencing and assembly of virus-derived small silencing rnas.," *Proceedings of the National Academy of Sciences*, vol. 107, no. 4, pp. 1606–11, 2010.

Z. Yan, H. Y. Hu, X. Jiang, V. Maierhofer, E. Neb, L. He, Y. Hu, H. Hu, N. Li, W. Chen, and P. Khaitovich, "Widespread expression of pirna-like molecules in somatic tissues," *Nucleic Acids Research*, vol. 39, no. 15, pp. 6596–6607, 2011.

Z Yang, "Paml: a program package for phylogenetic analysis by maximum likelihood.," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–6, 1997.

M. Yi, F. Chen, M. Luo, Y. Cheng, H. Zhao, H. Cheng, and R. Zhou, "Rapid evolution of pirna pathway in the teleost fish: implication for an adaptation to transposon diversity," *Genome Biology and Evolution*, vol. 6, no. 6, pp. 1393–1407, 2014.

E. Yigit, P. J. Batista, Y. Bei, K. M. Pang, C.-C. G. Chen, N. H. Tolia, L. Joshua-Tor, S. Mitani, M. J. Simard, and C. C. Mello, "Analysis of the c. elegans argonaute family reveals that distinct argonautes act sequentially during rnai.," *Cell*, vol. 127, no. 4, pp. 747–57, 2006.

R. A. Zambon, M. Nandakumar, V. N. Vakharia, and L. P. Wu, "The toll pathway is important for an antiviral response in drosophila.," *Proceedings of the National Academy of Sciences*, vol. 102, no. 20, pp. 7257–62, 2005.

B. Zhang, Y. H. Liu, W. X. Wu, and Z. L. Wang, "Molecular phylogeny of bactrocera species (diptera: tephritidae: dacini) inferred from mitochondrial sequences of 16s rdna and coi sequences," *BioOne*, vol. 93, no. 3, pp. 369–377, 2010.

H. Zhang, F. A. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz, "Single processing center models for human dicer and bacterial rnase iii," *Cell*, vol. 118, no. 1, pp. 57–68, 2004.

J. Zhang, Y.-p. Zhang, and H. F. Rosenberg, "Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey.," *Nature Genetics*, vol. 30, no. 4, pp. 411–5, 2002.

L. Zhao, P. Saelao, C. D. Jones, and D. J. Begun, "Origin and spread of de novo genes in drosophila melanogaster populations.," *Science (New York, N.Y.)*, vol. 343, no. 6172, pp. 769–72, 2014.

J. Zong, X. Yao, J. Yin, D. Zhang, and H. Ma, "Evolution of the rna-dependent rna polymerase (rdrp) genes: duplications and possible losses before and after the divergence of major eukaryotic groups," *Gene*, vol. 447, no. 1, pp. 29–39, 2009.

Appendices

A. Chapter 2

Table A.1.: Log likelihood values, ω estimates and Akaike weights for M0 (branch) models, comparing ω estimates between subclades.

Subclade	**Model**	**lnL (2dp)**	ω **(3dp)**	**Akaike weight**
Ago1	M0	-48132.17	0.006 ($\pm$0.0002)	1
Ago1	M0 (ω fixed at Ago2 ω)	-52340.60	0.140	0
Ago1	M0 (ω fixed at Ago3 ω)	-51700.12	0.118	0
Ago1	M0 (ω fixed at Piwi/Aub ω)	-50702.80	0.086	0
Ago2	M0	-141349.38	0.140 ($\pm$0.0015)	0.99
Ago2	M0 (ω fixed at Ago1 ω)	-150433.94	0.006	0
Ago2	M0 (ω fixed at Ago3 ω)	-141407.13	0.118	2.2^{-25}
Ago2	M0 (ω fixed at Piwi/Aub ω)	-141787.09	0.086	2.2^{-190}
Ago3	M0	-90817.41	0.118 ($\pm$0.0015)	1
Ago3	M0 (ω fixed at Ago1 ω)	-95795.68	0.006	0
Ago3	M0 (ω fixed at Ago2 ω)	-90860.30	0.140	6.4^{-19}
Ago3	M0 (ω fixed at Piwi/Aub ω)	-90948.07	0.086	4.9^{-57}
Piwi/Aub	M0	-194367.89	0.086 ($\pm$0.0010)	1
Piwi/Aub	M0 (ω fixed at Ago1 ω)	-201332.12	0.006	0
Piwi/Aub	M0 (ω fixed at Ago2 ω)	-195019.63	0.140	2.4^{-238}
Piwi/Aub	M0 (ω fixed at Ago3 ω)	-194629.29	0.118	8.1^{-114}

Table A.2.: Log likelihood values for M8 and M8a (sites) models, and p-values for a likelihood ratio test between the M8 and M8a models for each subclade.

Subclade	**Model**	**lnL (2dp)**	**p-value**
Ago1	M8	-47761.76	
Ago1	M8a	-47759.73	>0.1
Ago2	M8	-135792.60	
Ago2	M8a	-135792.09	>0.1
Ago3	M8	-87230.63	
Ago3	M8a	-87206.97	>0.1
Piwi/Aub	M8	-188055.00	
Piwi/Aub	M8a	-188055.14	>0.1

Table A.3.: Log likelihood values and ω estimates for "Immediate" and "All descendants" models.

Subclade	**Model**	**lnL (2dp)**	ω **(3dp) (post-duplication)**	ω **(3dp) (pre-duplication)**
Ago2	"Immediate"	-141347.25	0.150 ($\pm$0.0058)	0.138 ($\pm$0.0019)
Ago2	"All descendants"	-141579.48	0.165 ($\pm$0.0038)	0.125 ($\pm$0.0021)
Ago3	"Immediate"	-90816.75	0.106 ($\pm$0.0104)	0.118 ($\pm$0.0016)
Ago3	"All descendants"	-90811.18	0.142 ($\pm$0.0082)	0.115 ($\pm$0.0016)
Piwi/Aub	"Immediate"	-194334.11	0.116 ($\pm$0.0045)	0.082 ($\pm$0.0011)
Piwi/Aub	"All descendants"	-194364.84	0.086 ($\pm$0.0010)	0.053 ($\pm$0.0143)

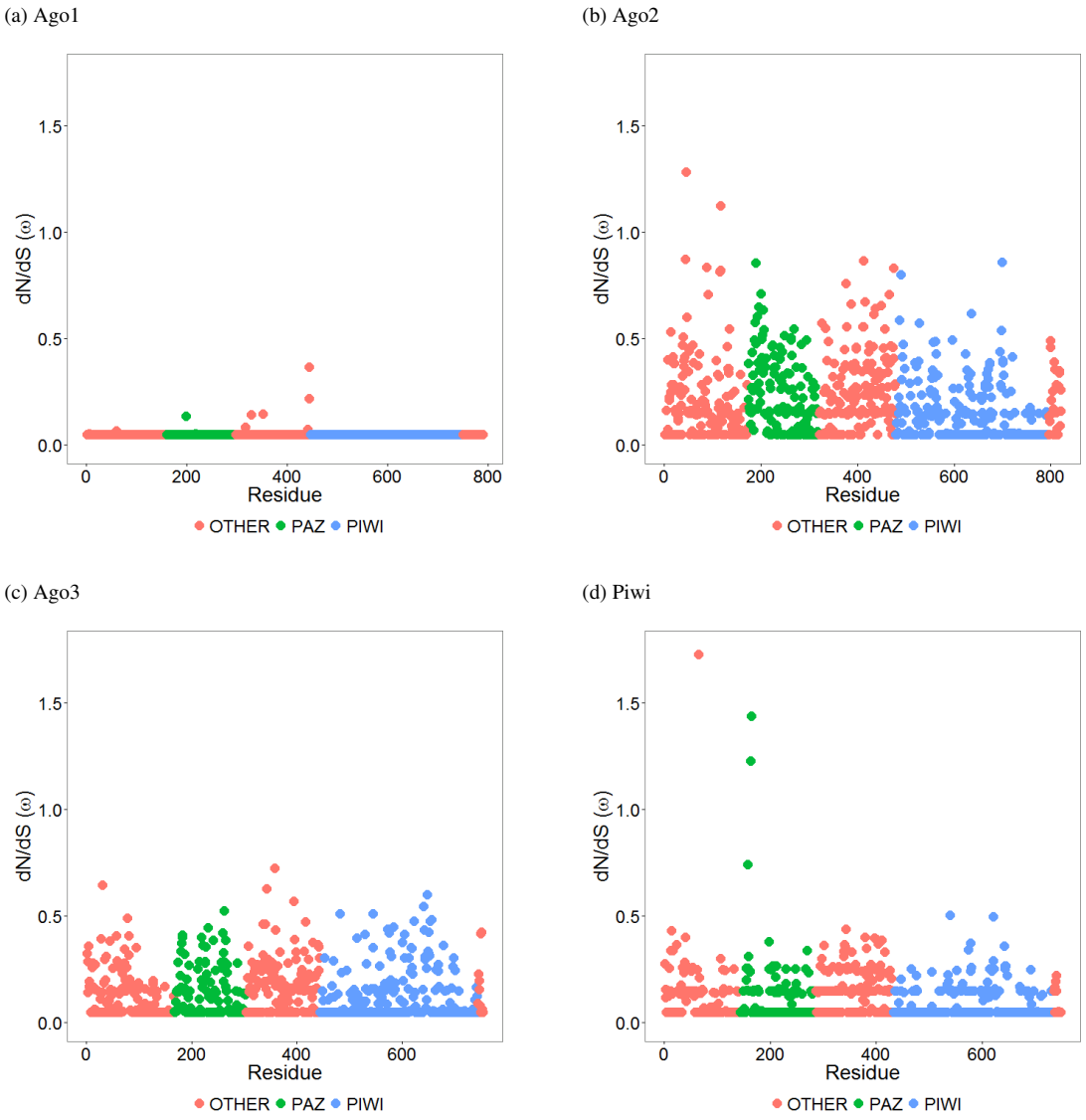


Figure A.1.: Evolutionary rates mapped onto the domain architecture of Dipteran Argonautes.
In each gene, rapidly evolving residues do not cluster in a particular location, but instead are spread across all domains.

B. Chapter 3

Table B.1.: *D. subobscura* Ago2 paralogue PCR primers

Paralogue	**Name**	**Sequence**
Ago2a	Dsubob_Ago2_1229_5_F	CCAAGAAGTGAAAGTAACAGATCG
Ago2a	Dsubob_Ago2_1229_M_F	CCACTGAATCGCAAGGATTCTGC
Ago2a	Dsubob_Ago2_1229_M_R	GGCGAATACCAAAGCGACTGATGG
Ago2a	Dsubob_Ago2_1229_3_R	CTTTGGGGAACGGAACTTGGTGAC
Ago2f	Dsubob_Ago2_21203_5_F	CACGCCTTTGAGGTGTACAGAAAGC
Ago2f	Dsubob_Ago2_21203_3_R	CACCAAATGTGCCAGATAGACCG

Table B.2.: *D. subobscura* Ago2a sequencing primers

Name	**Sequence**
Dsubob_Ago2_1229_5_F	CCAAGAAGTGAAAGTAACAGATCG
Dsubob_s_800_Ago2a_F	CTAGACTTCAGGCGTAACGATATCG
Dsubob_Ago2_1229_M_F	CCACTGAATCGCAAGGATTCTGC
Dsubob_s_1635_Ago2a_F	GCAATATGGCATTCTCACACAATG
Dsubob_s_2085_Ago2a_F	CTGCTGCAAGATGCACATTAAGC
Dsubob_Ago2_1229_3_R	CTTTGGGGAACGGAACTTGGTGAC
Dsubob_s_1965_Ago2a_R	GTGCTCAAGGGTTATTGACTCCATGTC
Dsubob_s_1420_Ago2a_R	GTACTGCATCTTACCGTACAGTATGG
Dsubob_Ago2_1229_M_R	GGCGAATACCAAAGCGACTGATGG
Dsubob_s_710_Ago2a_R	CCACATTGACAAATGGACGATCACC

Table B.3.: *D. subobscura* Ago2f sequencing primers

Name	Sequence
Dsubob_Ago2_21203_5_F	CACGCCTTTGAGGTGTACAGAAAGC
Dsubob_s_440_Ago2f_F	GCTGGTCGCTCCTTCTTCAAGC
Dsubob_s_1480_Ago2f_F	GCTGCAGCACGGCATACTGAC
Dsubob_s_1935_Ago2f_F	CGTGTTGCAAGAAGCACATTCG
Dsubob_Ago2_21203_3_R	CACCAAATGTGCCAGATAGACCG
Dsubob_s_1820_Ago2f_R	GCAAGTGCTCCAGCGTGATGG
Dsubob_s_1500_Ago2f_R	GTCAGTATGCCGTGCTGCAG
Dsubob_s_624_Ago2f_R	CGTGATATCGCTCAATGTACTCCAC

Table B.4.: *D. subobscura* Ago2 paralogue qPCR primers

Name	Sequence
Dsubob_Ago2a_q_F_2_1	CCAACGAGAGGAAGGCCAAGATTATAC
Dsubob_Ago2a_q_R_2	CCAGGCGAATACCAAAGCGACT
Dsubob_Ago2f_q_F_3	GATTTCAAGCGGCTCCAATGTG
Dsubob_Ago2f_q_R_3_1	GTTTGCGCTGCACCGTAAACAG
Obs_group_RpL32_q_F	CTTAGTTGTCGCACAAATGG
Obs_group_RpL32_q_R	TGCGCTTGTTGGAACCGTAAC

Table B.5.: *D. obscura* Ago2 paralogue PCR primers

Paralogue	Name	Sequence
Ago2a	Dobs_Ago2_2809_5_F	GGACAAGTATCTGTCAATTATCTCGACG
Ago2a	Dobs_Ago2_2809_18680_3_R	CTTGGGGAGAACGGAACTTGG
Ago2f	Dobs_Ago2_18680_5_F	CCTTTGAGCTGTTCAGAGTGGAAC
Ago2f	Dobs_Ago2_18680_M_F_2	GTAAATTGAGCCCCCAGTGTGTGTTGA
Ago2f	Dobs_Ago2_18680_M_R	CCAGCTGAGTGCGCGGGTTATC
Ago2f	Dobs_Ago2_all_3_R	TGGCGCCAGTCAGATAGACACG
Ago2e	Dobs_Ago2_24803_5_F	CGMGGTACACTGGGCAGAATCG
Ago2e	Dobs_947_Ago2_24803_F	GGTGAATCGCAAGGACTCCACGCT
Ago2e	Dobs_Ago2_24803_M_R	CAGTTCGGCTTTCTGTTTCAGTTC
Ago2e	Dobs_2269_Ago2_24803_R	GGCTCCACGTTGTTGTATTTGTTGTG

Table B.6.: *D. obscura* Ago2a sequencing primers

Name	**Sequence**
Dobs_Ago2_2809_5_F	GGACAAGTATCTGTCAATTATCTCGACG
Dobs_s_720_Ago2_2809_18680_F	AATCTTGGCGACGGCTACGAAGCTC
Dobs_s_1215_Ago2_2809_18680_F	CCATGATTAGGTATGCTGCCACATC
Dobs_s_1715_Ago2_2809_18680_F	GGCTGAGCTGCAGTATGGCATTCT
Dobs_s_1984_Ago2_2809_18680_F	GCAATATCGCTTGCAACGCTCTG
Dobs_s_2586_Ago2_2809_F	GGTCAGCCATCAGTCCATTCAGG
Dobs_Ago2_2809_18680_3_R	CTTGGGGAGAACGGAACTTGG
Dobs_s_1922_Ago2_2809_18680_R	CGCTGATCGGGCGATGGATG
Dobs_s_1435_Ago2_2809_18680_R	CCATGCGCCACGAACCGTT
Dobs_s_798_Ago2_2809_18680_R	TCCACATTGACAAACGGACG

Table B.7.: *D. obscura* Ago2f sequencing primers

Name	**Sequence**
Dobs_Ago2_18680_5_F	CCTTTGAGCTGTTCAGAGTGGAAC
Dobs_s_720_Ago2_2809_18680_F	AATCTTGGCGACGGCTACGAAGCTC
Dobs_s_1215_Ago2_2809_18680_F	CCATGATTAGGTATGCTGCCACATC
Dobs_s_1715_Ago2_2809_18680_F	GGCTGAGCTGCAGTATGGCATTCT
Dobs_Ago2_18680_M_F_2	GTAAATTGAGCCCCCAGTGTGTGTTGA
Dobs_s_1984_Ago2_2809_18680_F	GCAATATCGCTTGCAACGCTCTG
Dobs_Ago2_all_3_R	TGGCGCCAGTCAGATAGACACG
Dobs_s_1922_Ago2_2809_18680_R	CGCTGATCGGGCGATGGATG
Dobs_Ago2_18680_M_R	CCAGCTGAGTGCGCGGGTTATC
Dobs_s_1435_Ago2_2809_18680_R	CCATGCGCCACGAACCGTT
Dobs_s_802_Ago2_18680_R	CGCGTATAGCAGTGCTATGAC
Dobs_s_798_Ago2_2809_18680_R	TCCACATTGACAAACGGACG

Table B.8.: *D. obscura* Ago2e sequencing primers

Name	**Sequence**
Dobs_Ago2_24803_5_F	CGMGGTACACTGGGCAGAATCG
Dobs_s_553_Ago2_24803_F	GTGAATGTGGACATCACACACAAGTG
Dobs_s_1044_Ago2_24803_F	GCAGTACTTCAGCCACAACACGG
Dobs_947_Ago2_24803_F	GGTGAATCGCAAGGACTCCACGCT
Dobs_s_1377_Ago2_24803_F	GAGCCTGGATCCGCACTTCAAGG
Dobs_s_1881_Ago2_24803_F	GGTCAGCGATGGGCAGTTCC
Dobs_2269_Ago2_24803_R	GGCTCCACGTTGTTGTATTTGTTGTG
Dobs_s_1812_Ago2_24803_R	GGCTGTTATCGACTCCATGTCCTC
Dobs_Ago2_24803_M_R	CAGTTCGGCTTTCTGTTTCAGTTC
Dobs_s_1064_Ago2_24803_R	GTGTTGTGGCTGAAGTACTGCAG
Dobs_s_578_Ago2_24803_R	CACTTGTGTGTGATGTCCACATTCAC

Table B.9.: *D. obscura* Ago2 paralogue qPCR primers

Name	**Sequence**
Dobs_Ago2a_q_F_2	GCTTTCCAAGGTTTCAGCAAGCTC
Dobs_Ago2a_q_R_2_1	CCAACATGCAAGCATAGAAGGT
Dobs_Ago2f_q_F_3_1	GCACTCCGTCCACGTACG
Dobs_Ago2f_q_R_3	CTCATTCGGATGGACAATGATCCT
Dobs_Ago2e_q_F_4	CAACTACAACAAGATGCGGGACCTTG
Dobs_Ago2e_q_R_4	GAAGTGCGGATCCAGGCTCT
Obs_group_RpL32_q_F	CTTAGTTGTCGCACAAATGG
Obs_group_RpL32_q_R	TGCGCTTGTTGGAACCGTAAC

Table B.10.: *D. pseudoobscura* Ago2 paralogue PCR primers

Paralogue	**Name**	**Sequence**
Ago2a1&3	Dper.mir.A_F	TGGAGGTTGTGTTGGCAGtA
Ago2a1&3	DpseAgo2A_R	CTANACGAARTACATAGGRTTCGTCTTC
Ago2a3	Dpse_GA22965_3_out_F_2	CCAAGAGGACGAAAACACTGATTGG
Ago2a3	DpseAgo2A_R	CTANACGAARTACATAGGRTTCGTCTTC
Ago2b	Dper.mir.D_F	CAGTACGATGTGAAGATCACGTCAGTAT
Ago2b	Dpse_Ago2_UnivR	GCCAGTRAGRTAGACACGTCC
Ago2c	Dpse_Ago2c_5_F	GGCCGTACCCTGACTTACACTGTGGAAC
Ago2c	Dpse_Ago2c_M_F	CTTGGAAAACGACTTCATTGTGGTGC
Ago2c	Dpse_Ago2c_M_R	CTGTAATTCCGTATATCGGCCTTCTG
Ago2c	Dpse_Ago2c_3_R	CACGAAATACATGGGGTTCGTTTTCAT
Ago2d	DpseAgo2B_MF	ATGCCAGCTGTGGCCTACCA
Ago2d	Dpse_Ago2d_M_F	CTGGATGGGAAGCAAACGACGG
Ago2d	qrtD_R	GAAGTCAGTGCCCAGGCGT
Ago2d	Dpse_Ago2d_3_R	GGAACTCTGGAACAATCAACCGCTTTT
Ago2e	pse_Ago2E_5_F	CGAGGTGGCTGTGAACTACCTGCAG
Ago2e	pse_Ago2E_3_R	CATGGGGTTCCTGCTGGACAGG

Table B.11.: *D. pseudoobscura* Ago2a1/Ago2a3 sequencing primers

Name	**Sequence**
Dper.mir.A_F	TGGAGGTTGTGTTGGCAGtA
Dpse_s_1561_Ago2ab_F	CTGAARCACATTTAYTTGCCTATCG
Dpse_s_2011_Ago2ab_F	GTCAATCTGTGCCTGGATRCCAARG
Dpse_s_2487_Ago2ab_F	GTCGATAACCCTKGAGCACTTGCGTG
DpseAgo2A_R	CTANACGAARTACATAGGRTTCGTCTTC
Dpse_s_2808_Ago2ab_R	GTTGTATTGCGATGGARCTCCGYTCG
Dpse_s_2220_Ago2ab_R	CTCGACTGTGATCTGCTTGAKGC
Dpse_s_1611_Ago2ab_R	CTGYCCATCSTCAATGCGACATAG

Table B.12.: *D. pseudoobscura* Ago2a3 sequencing primers

Name	Sequence
Dper.mir.A_F	TGGAGGTTGTGTTGGCAGtA
Dpse_s_1561_Ago2ab_F	CTGAARCACATTTAYTTGCCTATCG
Dpse_s_2011_Ago2ab_F	GTCAATCTGTGCCTGGATRCCAARG
Dpse_s_2487_Ago2ab_F	GTCGATAACCCTKGAGCACTTGCGTG
DpseAgo2A_R	CTANACGAARTACATAGGRTTCGTCTTC
Dpse_s_2808_Ago2ab_R	GTTGTATTGCGATGGARCTCCGYTCG
Dpse_s_2220_Ago2ab_R	CTCGACTGTGATCTGCTTGAKGC
Dpse_s_1611_Ago2ab_R	CTGYCCATCSTCAATGCGACATAG
Dpse_GA22965_3_out_F_2	CCAAGAGGACGAAAACACTGATTGG

Table B.13.: *D. pseudoobscura* Ago2b sequencing primers

Name	Sequence
Dper.mir.D_F	CAGTACGATGTGAAGATCACGTCAGTAT
Dpse_s_912_Ago2b_F	CAGGAAGACGCAGGAATCGGAAG
Dpse_s_1561_Ago2ab_F	CTGAARCACATTTAYTTGCCTATCG
Dpse_s_2011_Ago2ab_F	GTCAATCTGTGCCTGGATRCCAARG
Dpse_s_2487_Ago2ab_F	GTCGATAACCCTKGAGCACTTGCGTG
Dpse_Ago2_UnivR	GCCAGTRAGRTAGACACGTCC
Dpse_s_2808_Ago2ab_R	GTTGTATTGCGATGGARCTCCGYTCG
Dpse_s_2220_Ago2ab_R	CTCGACTGTGATCTGCTTGAKGC
Dpse_s_1611_Ago2ab_R	CTGYCCATCSTCAATGCGACATAG
Dpse_s_1087_Ago2b_R	CAACCTCCAGACACTGCAAGGCTC

Table B.14.: *D. pseudoobscura* Ago2c sequencing primers

Name	Sequence
Dpse_Ago2c_5_F	GGCCGTACCCTGACTTACACTGTGGAAC
Dpse_s_1248_Ago2_c_F	GACTACAGGCGTTACGATATCGAATC
Dpse_s_1683_Ago2_c_F	CCAAGATCATCCACTTGCTCG
Dpse_Ago2c_M_R	CTGTAATTCCGTATATCGGCCTTCTG
Dpse_s_1476_Ago2_c_R	CAACATGCAAGCAGAGCAGGTTC
Dpse_Ago2c_M_F	CTTGGAAAACGACTTCATTGTGGTGC
Dpse_s_2508_Ago2_c_F	CATGGAGTCGATAACTCTTGAGCACTTTC
Dpse_Ago2c_3_R	CACGAAATACATGGGGTTCGTTTTCAT
Dpse_s_2533_Ago2_c_R	GTGCTCAAGAGTTATCGACTCCATGTC
Dpse_s_1943_Ago2_c_R	CGACCACACTCATGTAGTTAATCTTACC

Table B.15.: *D. pseudoobscura* Ago2d sequencing primers

Name	**Sequence**
DpseAgo2B_MF	ATGCCAGCTGTGGCCTACCA
Dpse_s_612_Ago2d_F	GTCAGAGCCCGGTAAAGCCTTTG
qrtD_R	GAAGTCAGTGCCCAGGCGT
Dpse_s_754_Ago2d_R	CAATGATCGTCATGGCCTTCGGAAAG
Dpse_Ago2d_M_F	CTGGATGGGAAGCAAACGACGG
Dpse_s_1443_Ago2_d_F	GTCAATGTGTGCCTGAATGACAACG
Dpse_s_1933_Ago2_d_F	GAGCACTTGCGTGTCTATCATCAGTACC
Dpse_Ago2d_3_R	GGAACTCTGGAACAATCAACCGCTTTT
Dpse_s_1911_Ago2_d_R	CCTGTATCTCCTCCAAGTCAGAGC
Dpse_s_1348_Ago2_d_R	CGACCACTCTTATGTAGTCAATCTTACC

Table B.16.: *D. pseudoobscura* Ago2e sequencing primers

Name	**Sequence**
pse_Ago2E_5_F	CGAGGTGGCTGTGAACTACCTGCAG
Dpse_s_926_Ago2e_F	CAACTGTGATGGCACGAAGGTGAC
pse_Ago2E_M_F	GGCTTGTGGCACATCGACAGGTC
Dpse_s_1921_Ago2e_F	GCGTCCTACAACATGCAGTACCG
pse_Ago2E_3_R	CATGGGGTTCCTGCTGGACAGG
Dpse_s_2117_Ago2e_R	CATCCCTCGCAGCTCCTCGTTCC
Dpse_s_1575_Ago2e_R	CTTGGGTCCAGGCTCTTGGCGTC
Dpse_s_1055_Ago2e_R	GCACAGCTCAATGGGCAGATAGACGG

Table B.17.: *D. pseudoobscura* qPCR primers

Gene	Name	Sequence
Ago1	Obsgroup_Ago1_q_F_2	GTGAAGTTCACCAAGGAGATCAAGG
Ago1	Obsgroup_Ago1_q_R_2	GGTTACATTGCAGACACGATACTTGC
Ago2a	Dpse_Ago2a_q_F_3	ATGGTTATTCAGAAGAGTCGCAAAG
Ago2a	Dpse_Ago2a_q_R_3	CTAGTTCACGTTCATCCTTGTAGTACAG
Ago2b	Dpse_Ago2b_q_F	GGGAAAGGAAAATAAATATAAACCGAA
Ago2b	Dpse_Ago2b_q_R	CGCACCTGTAGCTTTTAGTTGA
Ago2c	Dpse_Ago2c_q_F	AAGGAGGCGGACAACAGAG
Ago2c	Dpse_Ago2c_q_R	TGTGCTTGCTGACCCTGAG
Ago2d	Dpse_Ago2d_q_F	TCAGATTGAGTACAAAAACAAGTTG
Ago2d	Dpse_Ago2d_q_R	CCCTGAAAATCGACCACTCTTA
Ago2e	qrtE_F_3	GAACTACAACAAGATGCGGGACTTCG
Ago2e	qrtE_R_2	GCTTGGGTCCAGGCTCTTGG
Ago3	Dpse_Ago3_q_F_1	CGAAAGCAGTTCGATCCTTCATGTCC
Ago3	Dpse_Ago3_q_R_1	CGTCACAGCAGAGCATTAATCCTCC
Aub	Dpse_Aub_q_F_1	GCATTCAACAAGCGCTTGCAATC
Aub	Dpse_Aub_q_R_1	ACGCGAGCTGGGATGTCCAC
Piwi	Dpse_Piwi_q_F_1	GCGTATGGGCATATTGTCAAATCACG
Piwi	Dpse_Piwi_q_R_1	GGCCACACAGCACAATTGAATC
RpL32	Fly rp49 qPCR F-a	TGCCAAGTTGTCGCACAAATGG
RpL32	Fly rp49 qPCR R-i	TACGCTTGTTGGAGCCGTAAC

Table B.18.: Cycling conditions for PCR amplification of Ago2 paralogues

Location	Sequencing
1. 94C 02:00 2. 94C 00:10 3. 55C 00:30 4. 68C 03:00 5. Go to 2 19 times 6. 94C 00:15 7. 55C 00:30 8. 68C 03:00 +00:20 each cycle 9. Go to 6 14 times 10. 68C 07:00 11. 4C forever	1. 94C 05:00 2. 94C 00:15 3. 62C 00:30 -1C per cycle 4. 72C 03:30 5. Go to 2 9 times 6. 94C 00:15 7. 52C 00:30 8. 72C 03:30 9. Go to 6 29 times 10. 72C 05:00 11. 4C forever

Table B.19.: Cycling conditions for sequencing of Ago2 paralogues

Location	Sequencing
1. 94C 02:00 2. 94C 00:10 3. 55C 00:30 4. 68C 03:00 5. Go to 2 19 times 6. 94C 00:15 7. 55C 00:30 8. 68C 03:00 +00:20 each cycle 9. Go to 6 14 times 10. 68C 07:00 11. 4C forever	1. 94C 05:00 2. 94C 00:15 3. 62C 00:30 -1C per cycle 4. 72C 03:30 5. Go to 2 9 times 6. 94C 00:15 7. 52C 00:30 8. 72C 03:30 9. Go to 6 29 times 10. 72C 05:00 11. 4C forever

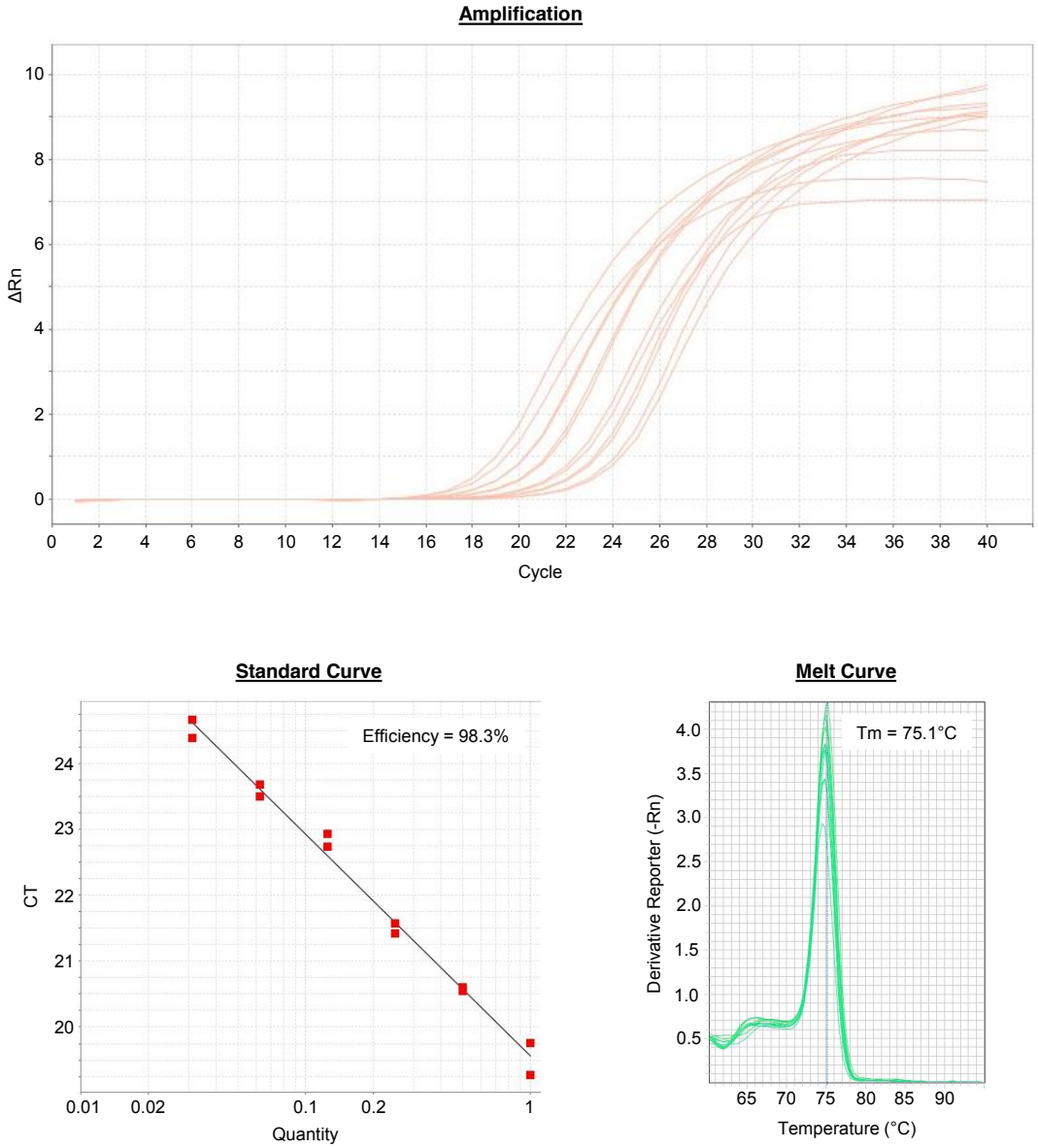


Figure B.1.: A representative efficiency profile of a custom-designed qPCR primer pair.
Displayed are the three criteria used to test primer pair efficiency: exponential amplification (Amplification Plot), 95% - 105% efficiency (Standard Curve), and a single melting point peak (Melt Curve).

Table B.20.: Lewis food recipe

Ingredient	Amount
Water	3.75L
Agar	25.8g
Sugar	351.6g
Maize	259.4g
Yeast	70.3g
Nipagin	56ml

The first five ingredients were mixed together, brought to the boil, and simmered for 5 minutes while stirring continuously. The mixture was allowed to cool to below 70°C, at which point the Nipagin was added. The pH was then neutralised using NaOH.

Table B.21.: Banana food recipe

Ingredient	Amount
Water	3.75L
Agar	45g
Malt powder	130g
Yeast	115g
Golden syrup	95mL
Bananas (blended)	8
Tergosept	30mL
Propionic acid	18mL

The first six ingredients were mixed together, brought to the boil, and simmered for 10 minutes while stirring continuously. The mixture was allowed to cool to below 70°C, at which point the tergosept and propionic acid were added. The pH was then neutralised using NaOH.

C. Chapter 4

Table C.1.: Isofemales lines sequenced

Species	**Sex**	**Line**	**Origin**
D. subobscura	Female	DPG7	Blackford Hill, Scotland
D. subobscura	Female	DPG8	Blackford Hill, Scotland
D. subobscura	Female	DPG9	Blackford Hill, Scotland
D. subobscura	Female	DPG10	Blackford Hill, Scotland
D. subobscura	Female	DPG11	Blackford Hill, Scotland
D. subobscura	Female	DPG12	Blackford Hill, Scotland
D. subobscura	Male	DPG1	Blackford Hill, Scotland
D. subobscura	Male	DPG2	Blackford Hill, Scotland
D. subobscura	Male	DPG3	Blackford Hill, Scotland
D. subobscura	Male	DPG4	Blackford Hill, Scotland
D. subobscura	Male	DPG5	Blackford Hill, Scotland
D. subobscura	Male	DPG6	Blackford Hill, Scotland
D. obscura	Female	DPG7	Blackford Hill, Scotland
D. obscura	Female	DPG8	Blackford Hill, Scotland
D. obscura	Female	DPG9	Blackford Hill, Scotland
D. obscura	Female	DPG10	Blackford Hill, Scotland
D. obscura	Female	DPG11	Blackford Hill, Scotland
D. obscura	Female	DPG12	Blackford Hill, Scotland
D. obscura	Male	DPG1	Blackford Hill, Scotland
D. obscura	Male	DPG2	Blackford Hill, Scotland
D. obscura	Male	DPG3	Blackford Hill, Scotland
D. obscura	Male	DPG4	Blackford Hill, Scotland
D. obscura	Male	DPG5	Blackford Hill, Scotland
D. obscura	Male	DPG6	Blackford Hill, Scotland
D. pseudoobscura	Female	MV15	Mesa Verde, CO, USA
D. pseudoobscura	Female	MV21	Mesa Verde, CO, USA
D. pseudoobscura	Female	MV25	Mesa Verde, CO, USA
D. pseudoobscura	Female	MV26	Mesa Verde, CO, USA
D. pseudoobscura	Female	MV28	Mesa Verde, CO, USA
D. pseudoobscura	Female	MV32	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV2	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV8	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV10	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV11	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV13	Mesa Verde, CO, USA
D. pseudoobscura	Male	MV19	Mesa Verde, CO, USA

Table C.2.: McDonald-Kreitman test results with alternative outgroups.
All values are displayed to 2dp

Paralogue	**pN**	**pS**	**Outgroup**	**dN**	**dS**	α	$\omega(\alpha)$	**p value**
D. subobscura Ago2a	2	9	*D. subsilvestris* Ago2a	58	120	0.54	0.0014	0.51
D. subobscura Ago2f	5	11	*D. subsilvestris* Ago2f ancestor	165	190	0.48	0.0008	0.31
D. obscura Ago2a	3	5	*D. tristis* Ago2a	20	33	0.01	0.00	1.00
D. obscura Ago2f	20	15	*D. tristis* Ago2f	19	23	-0.22	-0.0008	0.71
D. obscura Ago2e	1	5	*D. tristis* Ago2e	90	82	0.76	0.0105	0.38
D. pseudoobscura Ago2a1	5	1	*D. pseudoobscura* Ago2b	113	80	-2.54	-0.0095	0.40
D. pseudoobscura Ago2a3	5	4	*D. pseudoobscura* Ago2b	110	82	0.07	-0.0002	1.00
D. pseudoobscura Ago2b	3	0	*D. pseudoobscura* Ago2a1	100	73	N/A	N/A	N/A
D. pseudoobscura Ago2c	8	8	*D. pseudoobscura* Ago2d	60	45	0.25	0.0016	0.60
D. pseudoobscura Ago2d	5	3	*D. pseudoobscura* Ago2c	62	50	-0.34	-0.0021	0.73
D. pseudoobscura Ago2e	0	17	*D. lowei* Ago2e	66	133	1.00	0.0024	0.00

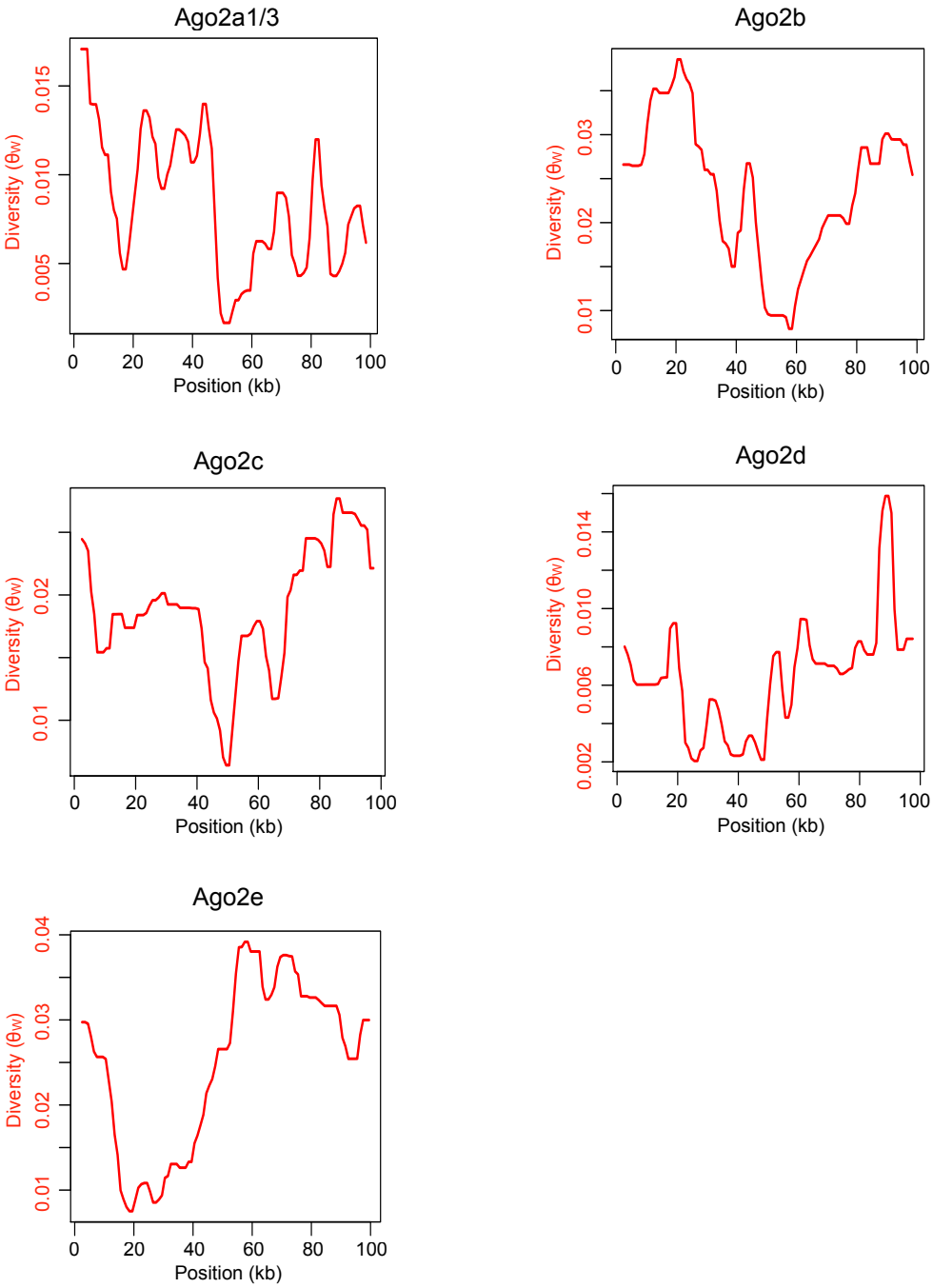


Figure C.1.: Selective sweeps at *D. pseudoobscura* Ago2 paralogues, with Ago2 paralogue haplotype sequences removed.
After specifying Ago2 paralogue sequence data as missing information, sharp troughs in diversity remain at Ago2a, Ago2b and Ago2c, indicating a selective sweep.

D. Other publications that are not part of this thesis

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Recent insights into the evolution of innate viral sensing in animals

Samuel H Lewis and Darren J Obbard

The evolution of viral sensors is likely to be shaped by the constraint imposed through high conservation of viral Pathogen-Associated Molecular Patterns (PAMPs), and by the potential for 'arms race' coevolution with more rapidly evolving viral proteins. Here we review the recent progress made in understanding the evolutionary history of two types of viral sensor, RNA helicases and Toll-like receptors. We find differences both in their rates of evolution, and in the levels of positive selection they experience. We suggest that positive selection has been the primary driver of the rapid evolution of the RNA helicases, while selective constraint has been a stronger influence shaping the slow evolution of the Toll-like receptors.

Addresses
Institute of Evolutionary Biology, and Centre for Immunity, Infection and Evolution, University of Edinburgh, Kings Buildings, EH9 3JT, United Kingdom

Corresponding author: Lewis, Samuel H (S.H.Lewis@sms.ed.ac.uk)

Current Opinion in Microbiology 2014, **20**:170–175

This review comes from a themed issue on **Host–microbe interactions: viruses**

Edited by **Maria-Carla Saleh**

For a complete overview see the Issue and the Editorial

Available online 18th July 2014

http://dx.doi.org/10.1016/j.mib.2014.05.010

Introduction

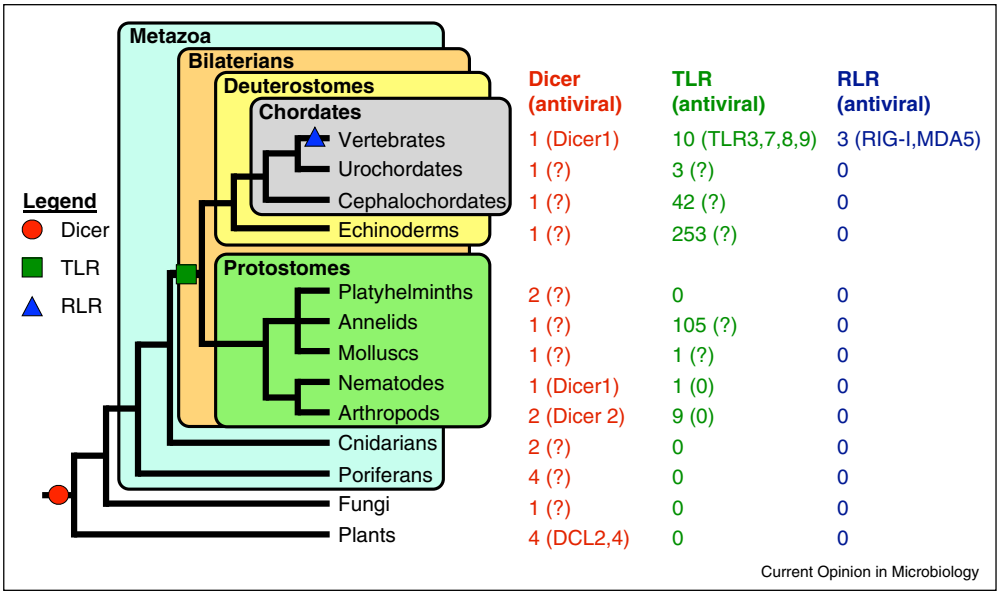
Pathogens reduce host fitness, and thereby exert a strong and ubiquitous selective pressure on hosts that has led to the evolution of a range of immune responses. Immune responses are elicited when sensors detect the presence of pathogens through Pathogen-Associated Molecular Patterns (PAMPs) or through markers of pathogen-associated damage. However, viruses may be uniquely difficult to sense because they use the host's own machinery to replicate, and therefore present fewer exogenous elicitors to immune surveillance mechanisms. Innate antiviral responses are therefore often triggered by conserved signatures of viral nucleic acids, such as dsRNA or CpG dinucleotides, which lead to the activation of multiple downstream immune responses, such as the RNA interference pathway or the vertebrate interferon response.

The conserved nature of these viral PAMPs leads to contrasting predictions regarding the evolution of antiviral genes. On the one hand, sensing these ancient and conserved molecular signatures might be expected to constrain the evolution of viral sensors. On the other hand, viral suppression of the antiviral immune system may lead to rapid evolution of viral sensors, as is seen in some antiviral genes of *Drosophila* [1]. Such rapid evolution may be driven by a host-virus arms race, as viruses escape the host immune response by cleaving or blocking antiviral genes [2]. Mechanisms of viral sensing have recently been reviewed elsewhere [3]; here we summarise the recent progress that has been made in understanding how two important viral sensing mechanisms have evolved, focussing on both phylogenetic history and the ongoing natural selection that shapes antiviral responses of extant populations. We finish by weighing the relative contributions of positive selection and evolutionary constraint during the evolution of viral sensing.

The phylogenetic distribution of viral sensing mechanisms

Although multiple protein families are known to act as viral sensors, many recent evolutionary studies have focussed on the Toll-like receptors (TLRs) and on receptors related to the RNA helicases, such as the Dicers and the RIG-I-like receptors (RLRs). Dicers act as sensors in the RNA interference (RNAi) pathway, binding dsRNA derived from the viral genome, replication intermediates or subgenomic products, and cleaving it into small RNAs that are ultimately used to target the virus or its transcripts for degradation. This is an ancient mechanism that probably arose prior to the most recent eukaryotic common ancestor over 1.5 billion years ago, and has since been conserved in all major eukaryotic lineages, including plants, fungi, ecdysozoa and vertebrates (illustrated in Figure 1) [4]. The helicase domain of the RLRs probably shares a common ancestor with that of Dicer [5], but on sensing viral dsRNA or other PAMPs, RLRs instead activate transcription factors such as nuclear factor-kappa B (NF-κB), and thereby induce the interferon pathway [6]. The RLRs also have a much more recent origin than Dicers, being present only in vertebrates, although homologues to their characteristic CAspase Recruitment Domains (CARDs) and RNA helicase domains are found in more basally branching deuterostomes, such as the tunicate *Ciona intestinalis* and the purple sea urchin

Figure 1



Phylogenetic distribution of viral sensing mechanisms. Gene family sizes are given, with validated antiviral genes in parentheses (0 = no antiviral genes, ? = antiviral function unknown). The three viral sensing mechanisms vary widely in their evolutionary ages: Dicer arose in the early Eukaryotes, whereas TLRs evolved in the early Bilateria, and RLRs first appeared in the vertebrates.

Strongylocentrotus purpuratus [5,7]. At present, direct viral sensing and immune induction functions have only been shown in vertebrates for two of the three RLRs, retinoic acid inducible gene I (RIG-I) [6] and melanoma differentiation associated gene 5 (MDA5) [8]. The third RLR, laboratory of genetics and physiology 2 (LGP2), binds viral RNA but cannot itself induce an immune response, instead triggering interferon production indirectly by signalling to MDA5 [9]. In contrast to the vertebrate-specific RLRs, the antiviral role of Dicer-like genes is much more widespread, being present in plants [10], fungi [11] and animals [12].

The Toll receptors were initially discovered in *Drosophila*, where they are involved in regulating the antibacterial and antifungal immune response [13]. The phylogenetic distribution (Figure 1) of Toll-like receptors (TLRs) suggests that they originated in the early Bilateria, before the divergence of protostomes and deuterostomes. In *Drosophila*, Toll-7 directly binds viruses and activates the autophagy response [14••]. In mammals, four TLRs (TLR3, 7, 8 and 9) play a pivotal role in sensing viral nucleic acids [15–18], subsequently activating the innate and adaptive immune responses through IRF-3, IRF-7 and NF-κB [19]. Other mammalian TLRs recognise different PAMPs, including lipids (TLR1, 2, 4 and 6) [20–22] and proteins (TLR5) [23]. This phylogenetic distribution of antiviral function suggests that TLRs are likely to have evolved a viral sensing role early in animal evolution, before the divergence of the protostomes and deuterostomes.

The evolution of RNA helicases

The most ancient conserved viral sensors are related to RNA helicases present in Archaea and Eukaryotes [5]. Two families of sensing helicases have been the subject of recent evolutionary study: the Dicers [24,25••] and the Rig-I-like receptors (RLRs) [5,7]. Two of the three RLRs (RIG-I and MDA5) each harbour two CARD domains that are integral in triggering the interferon response [6]. Despite this shared function, the two CARD domains appear to have substantially different histories [5], and it has therefore been suggested that the CARDs were gained by RIG-I and MDA5 in two separate events, with the first domain being acquired before the duplication that formed RIG-I and MDA5, and the second domain gained after they diverged [5]. Consistent with this, two CARD domains are found at separate loci in the sea anemone *Nematostella vectensis*, suggesting that the proposed grafting of these CARDs onto RLR may have occurred from these loci after the divergence of the chordates [7]. In contrast to the CARD domains, however, the order of divergence of RIG-I, MDA5 and LGP2 themselves remains unresolved. A neighbour-joining approach suggested that RIG-I diverged in the early deuterostomes, with LGP2 and MDA5 diverging later in the vertebrates

[7], while Bayesian and Maximum Likelihood methods find that LGP2 diverged in the early chordates, with RIG-I and MDA5 diverging later in the tetrapods [5].

It is highly likely that the last eukaryotic common ancestor possessed one Dicer, which was duplicated to produce two paralogues in the early Metazoa soon after their divergence from the other eukaryotes [24,25••]. However, the timing and extent of paralogue loss, and therefore the age of the two well-studied insect Dicer paralogues (Dicer1 & Dicer2), remains unresolved. It is possible that one of the paralogues was lost in the early Metazoa soon after the divergence of the Placazoa, and therefore Dicer1 and Dicer2 are relatively recent duplicates formed from a lineage-specific duplication in the ancestral arthropod [24]. Alternatively, large-scale lineage-specific loss of one of these paralogues may have left only the Placazoa and the arthropods with the two ancient paralogues [25••]. Reconstruction and rooting of this tree is made challenging by the extreme difference in evolutionary rate between Dicer1 and Dicer2, and by the high divergence to non-animal Dicers. Wider taxon sampling may mitigate these problems, and if so, then an ancient origin for Dicer1 and Dicer2 may be more likely [25••]. Accurate reconstruction of this phylogeny would help to determine the extent to which Dicer has retained its presumably ancestral antiviral role, which has been confirmed in plants, fungi, arthropods, and most recently mammals [26,27].

Population-genetic approaches can be used to detect departures from a standard neutral model of evolution, and thus infer the action of recent or ongoing natural selection. These methods have been widely applied to Dicers and RLRs, and have utilised both within-species genetic diversity [28•,29–31] and between-species divergence [1,28•,31,32] to understand the role of positive selection in shaping these genes. In humans, RIG-I appears to be tightly constrained [31], possibly due to the broad range of viruses it detects [33]. In contrast, positive selection has been detected on human LGP2 and MDA5 [31], and may have driven selective sweeps of MDA5, with one variant fixing in Europe and Asia and an alternative variant selected in South America [30]. Across the mammals, positive selection has been detected at individual sites in all domains of RIG-I and MDA5, but only in the helicase domain of LGP2 [34]. Evidence for positive selection has also been found for *Drosophila* Dicer2, which evolves extremely rapidly [1] under strong positive selection [32]. Despite this, it remains challenging to confidently attribute these patterns of RLR evolution to virus-mediated natural selection, as there may be some other shared trait common to all members of the RLR gene family that may predispose them to evolve in this way. Nevertheless, as neither rapid evolution nor positive selection are detected for insect Dicer1 [32], a Dicer2-homologue in the microRNA pathway that lacks a major antiviral role, it seems likely that the rapid evolution of Dicer2 may be driven specifically by its viral sensing function.

The evolution of the Toll-like receptors

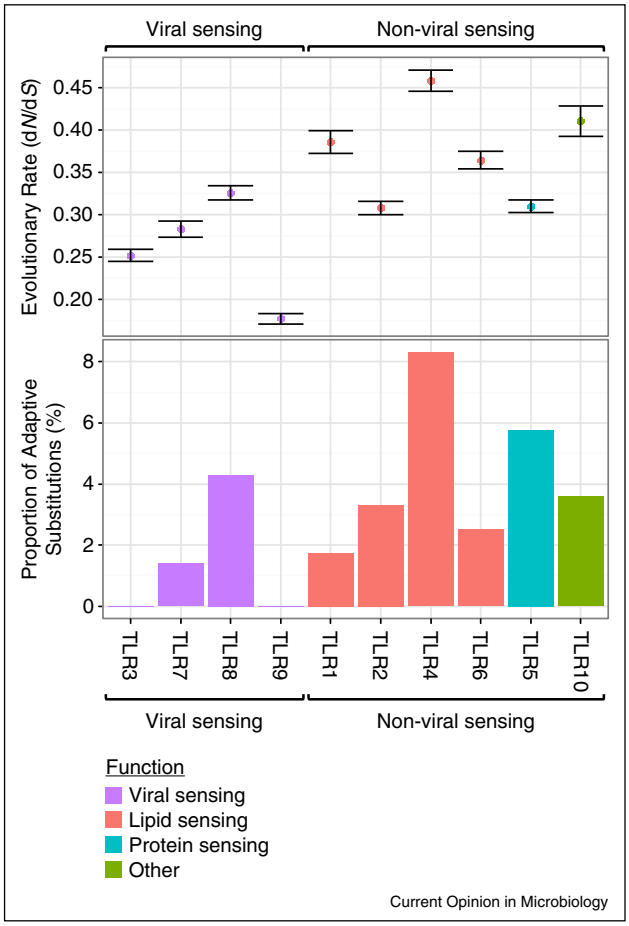
All TLRs have characteristic leucine-rich repeat (LRR) and Toll/interleukin-1 receptor (TIR) domains, which function in PAMP recognition and cell signalling, respectively. These domains appear to have evolved separately in the early Metazoa, as a vertebrate-like TIR is present in the Cnidaria [35]. However, the combination of TIR and LRR domains is seen after the divergence of the Bilateria from basal Metazoa, but before the divergence of the protostomes and deuterostomes [35]. A similar age has been estimated for the TLR adaptor MyD88, which was identified in both vertebrates and invertebrates [36], and for the interaction between TLRs, MyD88 and NF-κB, which has been reported in the oyster *Crassostrea gigas* (Lophotrochozoa) [37•]. However, the full TLR signalling pathway appears to have been acquired slowly, as the other adaptors TIR domain-containing adaptor molecule (TICAM) and TIR domain-containing adaptor protein (TIRAP) appear first in the early chordates [38] following duplication of MyD88 [36].

Direct sensing of viral PAMPs also appears to have evolved in TLRs before the divergence of the protostomes and deuterostomes, being found in both *Drosophila* [14••] and vertebrates. Intriguingly, differential expression of TLRs occurs on exposure of *C. gigas* to different PAMPs [37•], suggesting that specialisation of TLR paralogues to specific classes of pathogens may also have occurred early in the Bilateria. Since its divergence from other deuterostomes, a dramatic expansion of the TLR gene family in the basal deuterostome *S. purpuratus* has produced 253 paralogues, some of which appear to have specialised to a larval-specific or antibacterial role [39]. However, whether any of these paralogues has an antiviral function, and therefore how viral sensing has influenced their evolution, remains unknown.

Studies of TLR molecular evolutionary dynamics have revealed that selective pressures vary between domains, between different levels in the TLR signalling pathway, and between TLRs with different functions. At the domain level, the LRR domain evolves much faster than the TIR domain [39–42], consistent with the role of the latter in signalling to cytoplasmic adaptor molecules that are constrained by their interactions with multiple different TLRs. At the pathway level, a negative relationship between evolutionary rate and pathway position has been found in both *Drosophila* [43] and the Metazoa as a whole [44], suggesting that downstream components are under stronger purifying selection, possibly because of their interactions with multiple different upstream factors [44].

At the level of TLR function, four studies have explicitly compared the molecular evolutionary patterns of viral and non-viral TLRs in humans [45], rodents [46], primates [41], and mammals generally [47••]. These studies have used interspecific divergence at nonsynonymous and synonymous sites (d*N* and d*S*, respectively) to quantify the rate of protein evolution relative to the neutral expectation, with some studies going on to infer positive selection by testing for the existence of individual codon positions showing a d*N*/d*S* ratio greater than one. Comparisons that average d*N*/d*S* across the whole gene have all found that viral sensing TLRs evolve more slowly than TLRs that sense other pathogens; however, the magnitude of this difference in rates varies between focal lineages. In humans, viral sensing TLRs evolve much less rapidly than other TLRs, with average d*N*/d*S* values of 0.25 (viral) and 0.81 (non-viral) [45]. Far more modest differences have been found in rodents [46], primates [41], and birds [48]. Viral sensing TLRs may evolve more slowly because of stronger purifying selection, which has been detected using intraspecific polymorphism data from birds [48], humans [45] and primates as a whole [41]. Alternatively, the higher d*N*/d*S* ratio seen in TLRs that sense other PAMPs may reflect higher rates of positive selection, with a higher proportion of codons experiencing frequent adaptive substitutions.

Figure 2



The evolutionary rate (d*N*/d*S* — upper panel) and the proportion of codons inferred to be positively selected (lower panel) in viral sensing and non-viral sensing TLRs across eight rodent and ten primate species. Sequences were obtained from GenBank, and their phylogeny reconstructed using the Bayesian phylogenetic analysis program MrBayes [49] (see Supplemental File 1 for alignment). Evolutionary rate was estimated under the M0 model in PAML [50] (error bars represent one S.E.), and the proportion of adaptive substitutions represents the estimated proportion of sites with d*N*/d*S* > 1 under the M8 model. Overall, it appears that the primate and rodent viral sensing TLRs evolve more slowly and have a lower proportion of adaptive substitutions than other TLRs.

Adaptive substitutions have been inferred both at the TIR and LRR domains and the TLR sequence as a whole. There is wide variation in the proportion of positively selected codons that are located in the PAMP-binding LRR region: this domain harboured all adaptive substitutions in rodents [46] and the majority in mammals [47••], but in primates this region contained none in viral sensing TLRs, and only a small minority in non-viral TLRs [41]. Across the whole sequence, a mammal-wide study failed to find a significant difference in the proportion of positively selected codons between viral and non-viral TLRs [47••]. However, individual studies of primates [41], rodents [46] and birds [48] identified fewer positively selected codons in viral sensing compared with non-viral TLRs. This may indicate that host-virus arms race dynamics exert a weak or negligible effect on viral sensing TLRs, perhaps because their membrane-bound location limits viral interference. Instead, their evolution may simply be constrained by the conserved nature of viral PAMPs, resulting in low rates of adaptation and few positively selected codons (illustrated in Figure 2).

Conclusion

Viral sensors evolve under contrasting selective pressures: the conserved nature of viral PAMPs may tend to constrain evolution, whereas antagonistic host-virus coevolution may drive rapid evolution. The rapid evolution of RNA helicases could indicate that coevolution with other pathogen proteins (such as immune suppressors) is a major selective pressure on these sensors. In contrast, the slow evolution of TLRs may suggest the absence of a host-virus arms race acting directly on the sensor. In the future, this could be tested by further investigation of viral immune suppression strategies, and the overall importance of such strategies in shaping evolution could be informed by comparative studies of the evolution of viral sensors in a broader phylogenetic range of taxa.

Acknowledgements

We apologise to all authors whose work could not be included due to space constraints. We thank Ronald van Rij and Brian Lazzaro for their comments

on an earlier version of the manuscript, and Maria-Carla Saleh for the invitation and encouragement to write this review and for her comments on the manuscript. SHL is supported by a Natural Environment Research Council Doctoral Training Grant (NERC DTG NE/J500021/1) and work in DJO's lab is supported by a Wellcome Trust RCD Fellowship (085064/Z/08/Z) and a fellowship from the University of Edinburgh.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.mib.2014.05.010.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Obbard DJ, Welch JJ, Kim K-W, Jiggins FM: **Quantifying adaptive evolution in the *Drosophila* immune system**. *PLoS Genet* 2009, **5**:e1000698.
2. Bivalkar-Mehla S, Vakharia J, Mehla R, Abreha M, Kanwar JR, Tikoo A, Chauhan A: **Viral RNA silencing suppressors (RSS): novel strategy of viruses to ablate the host RNA interference (RNAi) defense system**. *Virus Res* 2011, **155**:1-9.
3. Braciale TJ, Hahn YS: **Immunity to viruses**. *Immunol Rev* 2013, **255**:5-12.
4. Cerutti H, Casas-Mollano JA: **On the origin and functions of RNA-mediated silencing: from protists to man**. *Curr Genet* 2006, **50**:81-99.
5. Sarkar D, Desalle R, Fisher PB: **Evolution of MDA-5/RIG-I-dependent innate immunity: independent evolution by domain grafting**. *Proc Natl Acad Sci U S A* 2008, **105**:17040-17045.
6. Yoneyama M, Kikuchi M, Natsukawa T, Shinobu N, Imaizumi T, Miyagishi M, Taira K, Akira S, Fujita T: **The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses**. *Nat Immunol* 2004, **5**:730-737.
7. Zou J, Chang M, Nie P, Secombes CJ: **Origin and evolution of the RIG-I like RNA helicase gene family**. *BMC Evol Biol* 2009, **9**:85.
8. Yoneyama M, Kikuchi M, Matsumoto K, Imaizumi T, Miyagishi M, Taira K, Foy E, Loo Y-M, Gale M Jr, Akira S *et al.*: **Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity**. *J Immunol* 2005, **175**:2851-2858.
9. Deddouche S, Goubau D, Rehwinkel J, Chakravarty P, Begum S, Maillard PV, Borg A, Matthews N, Feng Q: **Identification of an LGP2-associated MDA5 agonist in picornavirus-infected cells**. *eLife* 2014, **3**:e01535.
10. Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O: **Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense**. *Science* 2006, **313**:68-71.
11. Drinnenberg IA, Fink GR, Bartel DP: **Compatibility with killer explains the rise of RNAi-deficient fungi**. *Science* 2011, **333**:1592.
12. MacKay CR, Wang JP, Kurt-Jones EA: **Dicer's role as an antiviral: still an enigma**. *Curr Opin Immunol* 2014, **26**:49-55.
13. Lemaitre B, Nicolas E, Michaut L, Reichhart J-M, Hoffmann JA: **The dorsoventral regulatory gene cassette *spätzle/Toll/cactus* controls the potent antifungal response in *Drosophila* adults**. *Cell* 1996, **86**:973-983.
14. •• Nakamoto M, Moy RH, Xu J, Bambina S, Yasunaga A, Shelly SS, Gold B, Cherry S: **Virus recognition by Toll-7 activates antiviral autophagy in *Drosophila***. *Immunity* 2012, **36**:658-667.

This paper shows that *Drosophila* Toll-7 binds Vesicular Stomatitis Virus, providing evidence that TLRs sense viruses directly in the protostomes, and therefore that direct sensing is likely to be an ancestral TLR function that arose in the early Bilateria.

15. Alexopoulou L, Holt AC, Medzhitov R, Flavell RA: **Recognition of double-stranded RNA and activation of NF-κB by Toll-like receptor 3**. *Nature* 2001, **413**:732-738.
16. Hemmi H, Kaisho T, Takeuchi O, Sato S, Sanjo H, Hoshino K, Horiuchi T, Tomizawa H, Takeda K, Akira S: **Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway**. *Nat Immunol* 2002, **3**:196-200.
17. Lund J, Sato A, Akira S, Medzhitov R, Iwasaki A: **Toll-like receptor 9-mediated recognition of Herpes Simplex Virus-2 by plasmacytoid dendritic cells**. *J Exp Med* 2003, **198**:513-520.
18. Heil F, Hemmi H, Hochrein H, Ampenberger F, Kirschning C, Akira S, Lipford G, Wagner H, Bauer S: **Species-specific recognition of single-stranded RNA via Toll-like receptor 7 and 8**. *Science* 2004, **303**:1526-1529.
19. Akira S, Uematsu S, Takeuchi O: **Pathogen recognition and innate immunity**. *Cell* 2006, **124**:783-801.
20. Hoshino K, Takeuchi O, Kawai T, Sanjo H, Ogawa T, Takeda Y, Takeda K, Akira S: **Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the *Lps* gene product**. *J Immunol* 1999, **162**:3749-3752.
21. Takeuchi O, Kawai T, Mühlradt PF, Morr M, Radolf JD, Zychlinsky A, Takeda K, Akira S: **Discrimination of bacterial lipoproteins by Toll-like receptor 6**. *Int Immunol* 2001, **13**:933-940.
22. Takeuchi O, Sato S, Horiuchi T, Hoshino K, Takeda K, Dong Z, Modlin RL, Akira S: **Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins**. *J Immunol* 2002, **169**:10-14.
23. Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM, Aderem A: **The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5**. *Nature* 2001, **410**:1099-1103.
24. de Jong D, Eitel M, Jakob W, Osigus H-J, Hadrys H, Desalle R, Schierwater B: **Multiple Dicer genes in the early-diverging metazoa**. *Mol Biol Evol* 2009, **26**:1333-1340.
25. •• Mukherjee K, Campos H, Kolaczkowski B: **Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants**. *Mol Biol Evol* 2013, **30**:627-641.

This paper reconstructs the phylogenetic history of the eukaryotic Dicers, and suggests that antiviral Dicer2 is the result of an ancient rather than recent gene duplication event.

26. Maillard PV, Ciaudo C, Marchais A, Li Y, Jay F, Ding SW, Voinnet O: **Antiviral RNA interference in mammalian cells**. *Science* 2013, **342**:235-238.
27. Li Y, Lu J, Han Y, Fan X, Ding S-W: **RNA interference functions as an antiviral immunity mechanism in mammals**. *Science* 2013, **342**:231-234.
28. • Vasseur E, Boniotto M, Patin E, Laval G, Quach H, Manry J, Crouau-Roy B, Quintana-Murci L: **The evolutionary landscape of cytosolic microbial sensors in humans**. *Am J Hum Genet* 2012, **91**:27-37.

This paper compares the intraspecific diversity and evolutionary rate of human TLRs, RLRs and NRLs, finding both measures highly variable between viral sensing gene families.

29. Obbard DJ, Jiggins FM, Bradshaw NJ, Little TJ: **Recent and recurrent selective sweeps of the antiviral RNAi gene *Argonaute-2* in three species of *Drosophila***. *Mol Biol Evol* 2011, **28**:1043-1056.
30. Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, Comi GP, Bresolin N, Clerici M, Sironi M: **Population genetics of *IFIH1*: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes**. *Mol Biol Evol* 2010, **27**:2555-2566.
31. Vasseur E, Patin E, Laval G, Pajon S, Fornarino S, Crouau-Roy B, Quintana-Murci L: **The selective footprints of viral pressures at the human RIG-I-like receptor family**. *Hum Mol Genet* 2011, **20**:4462-4474.

32. Obbard DJ, Jiggins FM, Halligan DL, Little TJ: **Natural selection drives extremely rapid evolution in antiviral RNAi genes**. *Curr Biol* 2006, **16**:580-585.

33. Loo Y-M, Gale M: **Immune signaling by RIG-I-like receptors**. *Immunity* 2011, **34**:680-692.

34. Cagliani R, Forni D, Tresoldi C, Pozzoli U, Filippi G, Rainone V, De Gioia L, Clerici M, Sironi M: **RIG-I-like receptors evolved adaptively in mammals, with parallel evolution at LGP2 and RIG-I**. *J Mol Biol* 2014, **426**:1351-1365.

35. Wu B, Huan T, Gong J, Zhou P, Bai Z: **Domain combination of the vertebrate-like TLR gene family: implications for their origin and evolution**. *J Genet* 2011, **90**:401-408.

36. Roach JM, Racioppi L, Jones CD, Masci AM: **Phylogeny of Toll-like receptor signaling: adapting the innate response**. *PLOS ONE* 2013, **8**:e54156.

37. • Zhang Y, He X, Yu F, Xiang Z, Li J, Thorpe KL, Yu Z: **Characteristic and functional analysis of Toll-like receptors (TLRs) in the lophotrocozoan, *Crassostrea gigas*, reveals ancient origin of TLR-mediated innate immunity**. *PLOS ONE* 2013, **8**:e76464.

This paper shows that interaction between the TLRs, MyD88 and NF-κB was already established in the early Bilateria. It provides evidence that specialization of different TLRs to specific pathogens occurred early in the Bilateria.

38. Wu B, Xin B, Jin M, Wei T, Bai Z: **Comparative and phylogenetic analyses of three TIR domain-containing adaptors in metazoans: implications for evolution of TLR signaling pathways**. *Dev Comp Immunol* 2011, **35**:764-773.

39. Buckley KM, Rast JP: **Dynamic evolution of Toll-like receptor multigene families in echinoderms**. *Front Immunol* 2012, **3**:1-16.

40. Nakajima T, Ohtani H, Satta Y, Uno Y, Akari H, Ishida T, Kimura A: **Natural selection in the TLR-related genes in the course of primate evolution**. *Immunogenetics* 2008, **60**:727-735.

41. Wlasiuk G., Nachman MW.: **Adaptation and constraint at Toll-like receptors in primates**. *Mol Biol Evol* 2010, **27**:2172-2186.

42. Mikami T, Miyashita H, Takatsuka S, Kuroki Y, Matsushima N: **Molecular evolution of vertebrate Toll-like receptors: evolutionary rate difference between their leucine-rich repeats and their TIR domains**. *Gene* 2012, **503**:235-243.

43. Han M, Qin S, Song X, Li Y, Jin P, Chen L, Ma F: **Evolutionary rate patterns of genes involved in the *Drosophila* Toll and Imd signaling pathway**. *BMC Evol Biol* 2013, **13**:245.

44. Song X, Jin P, Qin S, Chen L, Ma F: **The evolution and origin of animal Toll-like receptor signaling pathway revealed by network-level molecular evolutionary analyses**. *PLOS ONE* 2012, **7**:e51657.

45. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B *et al.*: **Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense**. *PLoS Genet* 2009, **5**:e1000562.

46. Fornůsková A, Vinkler M, Pagès M, Galan M, Jousselin E, Cerqueira F, Morand S, Charbonnel N, Bryja J, Cosson J-F: **Contrasted evolutionary histories of two Toll-like receptors (*Tlr4* and *Tlr7*) in wild rodents (MURINAE)**. *BMC Evol Biol* 2013, **13**:194.

47. •• Areal H, Abrantes J, Esteves PJ: **Signatures of positive selection in Toll-like receptor (TLR) genes in mammals**. *BMC Evol Biol* 2011, **11**:368.

This paper explicitly compares the levels of positive selection on viral sensing and non-viral TLRs from a variety of mammals, and finds similar proportions of positively selected codons in both types.

48. Alcaide M, Edwards SV: **Molecular evolution of the Toll-like receptor multigene family in birds**. *Mol Biol Evol* 2011, **28**:1703-1715.

49. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylo genetic inference under mixed models**. *Bioinformatics* 2003, **19**:1572-1574.

50. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci* 1997, **13**:555-556.